Activity of Complex Multifunctional Organic Compounds in Common Solvents

By

Bruce Moller

[B.Sc. (Eng.), M.Sc. (Eng.)]

In fulfilment of the degree Doctor of Philosophy,

Chemical Engineering, University of KwaZulu-Natal

18 December 2009

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

PREFACE

The work presented in this thesis was performed at the University of KwaZulu-Natal, Durban from January 2008 to December 2009. The work was supervised by Prof. D. Ramjugernath and Prof. Dr. J. Rarey.

As the candidate's supervisor I, Prof. D. Ramjugernath, agree/do not agree to the submission of this thesis.

Signed: …….

DECLARATION

I ..... declare that:

(i) The research reported in this thesis, except where otherwise indicated, is my original work.

(ii) This thesis has not been submitted for any degree or examination at any other university.

(iii) This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a) their words have been re-written but the general information attributed to them has been referenced;

b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.

(vi) This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation/thesis and in the References sections.

Signed: …….

ACKNOWLEDGEMENTS

I wish to acknowledge the following people and organizations for their contribution to this work:

- My co-supervisor, Prof. Dr Jürgen Rarey for being a really great supervisor. Thanks for always being willing to give comments and criticism on my ideas, the enjoyment that I have gleaned from this work is, for a large part, down to your input. I will never forget the times we spent in your office discussing ideas and life in general. Working with you has been a fantastic learning experience.

- My supervisor, Prof. Dr Deresh Ramjugernath for initially sparking my interest in the MSc. and ultimately the PhD project. Thanks for always keeping me on track with progress report meetings and the occasional office visits. Thank you for providing the opportunities to travel the globe and have an all round good time while pursuing my degree.

- DDBST GmbH for providing data and software support for this project. Special thanks must go to Wilfried Cordes, thank you for tirelessly (and timelessly!) implementing all the various requests I had of the DDBST software. The time that you spent implementing the various methods developed through the course of my MSc. and PhD is greatly appreciated.

- NRF International Science Liaison, NRF-Thuthuka Programme and BMBF (WTZ-Project) for financial support.

- My parents Erik and Jeanette, my sister Teresa, my grandmother and late grandparents for all their years of wisdom and guidance. Thanks for affording me the opportunity to pursue this degree and keeping me sane during the more taxing times throughout the course of my studies.

- Thanks to all the friends that I have made as a result of pursing this degree. From Durban, thanks to all the post grads (you know who you are, too many to list) for making UKZN a great place to be and especially the bloukrans road trip guys, that trip was super kiff and thanks for convincing me to take the plunge! From Germany, thanks to Michael Döker for the fun times (sorry Germany never won Euro 2008!) and always keeping me in the loop about Tuesday soccer. Also thanks to Anja Diedrichs for showing me the ropes when I arrived and always having good ideas on what to do on the weekends. To the friends that I made at the IBC (International Baptist Church) in Bremen thanks for helping to make the time in Germany awesome, thanks for making a difference for Christ!

- Above all I would like to give praise to my Lord and Saviour Jesus Christ. Without whom my studies would have had no meaning. "*[24] Now all glory to God, who is able to keep you from falling away and will bring you with great joy into his glorious presence without a single fault. [25] All glory to him who alone is God, our Saviour through Jesus Christ our Lord. All glory, majesty, power, and authority are his before all time, and in the present, and beyond all time! Amen.*" Jude 1:24-25 (New Living Translation).

ABSTRACT

The models used in the prediction of activity coefficients are important tools for designing major unit operations (distillation columns, liquid-liquid extractors etc). In the petrochemical and chemical industry, well established methods such as UNIFAC and ASOG are routinely employed for the prediction of the activity coefficient. These methods are, however, reliant on binary group interaction parameters which need to be fitted to reliable experimental data. It is for this reason that these methods are often not applicable to systems which involve complex molecules. In these systems, typically solid-liquid equilibria are of interest where the solid is some pharmaceutical product or intermediate or a molecule of similar complexity (the term complex here refers to situations where molecules contain several functional groups which are either polar, hydrogen bonding, or lead to mesomeric structures in equilibrium). In many applications, due to economic and environmental considerations, a list of no more than 20 solvents is usually considered.

It is for this reason that the objective of this work is to develop a method for predicting the activity coefficient of complex multifunctional compounds in some common solvents. The segment activity coefficient approaches proposed by Hansen, MOSCED and the NRTL-SAC models show that it should be possible to “interpolate” between solvents if suitable reference solvents are available (e.g. non-polar, polar and hydrogen bonding). Therefore it is useful to classify the different solvents into suitable categories inside which analogous behaviour should be observed. To accomplish this, a significant amount of data needs to be collected for the common solvents.

Data with water as a solvent was freely available and multiple sources were found with suitable data. Both infinite dilution activity coefficient (γ^{∞}) and SLE (Solid-Liquid Equilibrium) data were used for model development. The γ^{∞} data were taken from the DDB (Dortmund Data Bank) and SLE data were taken from Beilstein, Chemspider and DDB. The limiting factor for the usage of SLE data was the availability of fusion data (heat of fusion and melting temperature) for the solute. Since γ^{∞} in water is essentially a pure component property it was modelled as such, using the experience gained previously by this group. The overall RMD percentage (in $\ln \gamma^{\infty}$) for the training set was 7.3 % for 630 compounds. For the test set the RMD (in $\ln \gamma^{\infty}$) was 9.1 % for 25 fairly complex compounds.

Typically the temperature dependence of γ^{∞} data is ignored when considering model development such as this. Nevertheless, the temperature dependence was investigated and it was found that a very simple general correlation showed moderate accuracy when predicting the temperature dependence of compounds with low solubility.

Data for solvents other than water were very scarce, with insufficient data to develop a model with reasonable accuracy. A novel method is proposed for the alkane solvents, which allows the values in any alkane solvent to be converted to a value in the solvent hexane. The

method relies on a first principles application of the solution of groups concept. Quite unexpectedly throughout the course of developing the method, several shortfalls were uncovered in the combinatorial expressions used by UNIFAC and mod. UNIFAC. These shortfalls were empirically accounted for and a new expression for infinite dilution activity coefficient is proposed. This expression is however not readily applicable to mixtures and therefore requires some further attention.

The method allows for the extension of the data available in hexane (chosen since it is a common solvent for complex compounds). In the same way as the γ^{∞} data in water, the γ^{∞} data in hexane were modelled as a pure component property. The overall RMD percentage (in $\ln \gamma^{\infty}$) for the training set was 21.4 % for 181 compounds. For the test set the RMD (in $\ln \gamma^{\infty}$) was 11.7 % for 14 fairly complex compounds. The great advantage of both these methods is that, since they are treated as pure component properties, the number of model parameters grows linearly with the number of groups, unlike with mixture models (UNIFAC, ASOG, etc.) where it grows quadratically. For both the water and the hexane method the predictions of the method developed in this work were compared to the predictions of UNIFAC, mod. UNIFAC, COSMO-RS(OL) and COSMO-SAC.

Since water and hexane are not the only solvents of practical interest, a method was developed to interpolate the alcohol behaviour based on the water and hexane behaviour. The ability to predict the infinite dilution activity coefficient in various solvents allowed for the prediction of various other properties, *viz*. air-water partition coefficient, octanol-water partition coefficient, and water-alcohol cosolvent mixtures. In most cases the predictions of these properties were good, even for the fairly complex compounds tested.

TABLE OF CONTENTS

LIST OF FIGURES

Appendices

LIST OF TABLES

Appendices

NOMENCLATURE

a_i	-	Activity of component *i*.
A_w	-	van der Waals surface area (cm^2/mol).
c	-	Cohesive pressure (MPa).
C_p	-	Molar heat capacity (J/mol·K) at constant pressure.
F	-	Degrees of freedom.
f	-	Fugacity (kPa).
$\hat{f}_i$	-	Fugacity in solution of component *i* (kPa).
G	-	Molar Gibbs free energy (J/mol).
H	-	Molar Enthalpy (J/mol).
ln	-	Natural logarithm.
log	-	Logarithm to the base 10.
n	-	Number of moles (mol).
N_A	-	Avogadro's number (= 6.023×10^{23} molecules/mol).
P	-	Pressure (kPa).
q	-	UNIQUAC surface parameter.
r	-	UNIQUAC volume parameter.
R	-	Ideal gas constant (= 8.314472 J/mol·K).
S	-	Molar entropy (J/mol·K).
T	-	Absolute temperature (K).
U	-	Molar internal energy (J/mol).
V	-	Molar volume (cm^3/mol).
w_i	-	Weight fraction of component *i*.
x_i	-	Mole fraction of component *i*.

Greek symbols

β	-	Selectivity.
δ	-	Hildebrand solubility parameter ($MPa^{0.5}$).
ϕ_i	-	Volume fraction of component *i.*
Φ_i	-	Volume fraction of component *i.*
γ_i	-	Activity coefficient of component *i*.
μ_i	-	Chemical potential of component *i*.
π	-	Internal pressure (MPa).
θ_i	-	Surface fraction of component *i*.

Subscripts

b	-	Normal boiling point.
d	-	Dispersive.
fus	-	Fusion.
h	-	Hydrogen bonding.
m	-	Mixing.
p	-	Polar.
v	-	Vaporization.

Superscript

*	-	van der Waals volume.
∞	-	Infinite dilution.
0	-	Standard state.
C	-	Combinatorial.
E	-	Excess property.
FV	-	Free-volume.

L	-	Liquid.
r	-	Reference state.
R	-	Residual.
S	-	Solid.

Abbreviations

DDB	-	Dortmund Data Bank.
FH	-	Flory Huggins.
GLC	-	Gas liquid chromatography.
IDAC	-	Infinite dilution activity coefficient.
LLE	-	Liquid-liquid equilibria.
MM	-	Molar mass (g/mol)
RMD	-	Relative mean deviation.
SG	-	Staveman-Guggenheim.
SLE	-	Solid-liquid equilibria.
VLE	-	Vapour-liquid equilibria.

All other symbols used have been explained in the text and unless otherwise stated SI units are used.

1. INTRODUCTION

In non-ideal solutions the interactions between the solute and solvent(s) has a significant effect on the activity of the solute. Knowledge of the solute activity is required for the solution of a large variety of technical and environmental problems. Solute activity strongly influences the solubility of a liquid or solid solute, its partial vapour pressure above the solution, and the chemical equilibrium composition and reaction rates in the liquid phase. Activities in two immiscible solvents allow the calculation of distribution coefficients required for extraction processes and the distribution of chemicals between environmental compartments. Solute activities usually depend on solute concentration, but may also be nearly constant over a significant concentration range. Besides their effect on these equilibrium properties, solute activities are also required for the calculation of transport properties; they determine the driving force in diffusional mass transfer.

It is estimated that there are over 100 000 chemicals in use in the chemical industry. Every year there are an estimated 700 – 3000 compounds entering the market [1] with between 200 – 1000 of these being produced at over one ton per annum [2]. Unfortunately, experimental physical property data for these compounds and their behaviour in mixtures with other components are typically unavailable. Experimental measurements are generally costly and time consuming and practically impossible to undertake for all pure compounds and their mixtures. Furthermore some mixtures react in very rapidly, which makes determination of physical properties impossible. For these reasons, predictive thermodynamic models are of great practical importance.

Current estimation techniques include simple group contribution approaches for e.g. water solubility, octanol-water partition coefficients [3-8], or thermodynamic models like UNIFAC [9-11] (based on group interactions) or COSMO-RS/SAC [12, 13] (based on quantum-mechanical calculations). In group contribution methods the number of parameters increases quadratically with the number of groups and the available experimental information usually does not allow a strong group differentiation. COSMO-RS/SAC methods on the other hand contain only a few parameters tuned to experimental results but are significantly less reliable [14].

In the case of complex multifunctional compounds like pharmaceuticals and their intermediates, pesticides or food components, all these methods generally lead to rather large errors. A great need exists for a quantitative estimation technique with a broad application.

In contrast to the situation for smaller volatile molecules, where knowledge about the mixture behaviour of many binary combinations of components is required, complex multifunctional compounds are mostly present in diluted solutions in solvents or solvent mixtures. Even with the potentially huge number of possible solvents most chemical companies only use a select list (private communication of a large pharmaceutical company revealed a list of 16 preferred

solvents). The aim of this work is therefore to develop a group contribution method to predict the activities of complex multifunctional organic compounds in various common solvents. This tool will especially support chemical engineers in both small and large companies dealing with the design, simulation and optimization of chemical processes. The method should be especially suitable for complex multifunctional compounds (biological compounds, pharmaceuticals), which are not covered well by existing models.

Literature offers a large amount of experimental data such as solubility information for complex solutes in common solvents (water, ethanol, acetone, benzene, etc.), that can be used for the implementation of an estimation method. Modelling the infinite dilution activity coefficient of a wide range of compounds in a single solvent mainly requires only single group contribution parameters (i.e. it is treated as a pure component property since the concentration dependence falls away at infinite dilution) and allows for sufficient group differentiation. Methods of this type, although very useful and required for solving a wide range of technical and environmental problems, have only been developed to a minor extend up to now (e.g. for water solubility).

The chapters following will cover the following topics:

- Classification of solvents into groups which suitably describe the various effects present in the molecule.
- The selection of sufficient and suitable experimental information on solute activities for complex molecules from various commercial sources available to the partner institutions (Dortmund Data Bank, Beilstein etc.).
- Novel methods for the expansion of the amount of data available in alkane solvents.
- Compilation of the required auxiliary information like melting points, heat of fusion, etc. for the components of interest from the available sources.
- Investigations into the various possible heat capacity approximations.
- Development a group contribution method for solute activities in several common solvents based on the experimental information available.
- Investigation into the temperature dependence of the activity coefficient.
- To verify, whether the activity in further solvents can be estimated by interpolation of results for similar solvents (e.g. activities in ethanol from activities in hexane and water).
- To compare results of the new method with available methods from literature (UNIFAC, COSMO-RS, etc.).

- Various applications of the methods developed.

The methodology has been based on that developed during previous work, in which a variety of computer tools have been built up for the fast and efficient development of group contribution estimation methods.

This work had resulted in the development of group contribution estimation methods for

- Normal boiling temperatures (Nannoolal et al. [15])
- Vapour pressure as function of temperature (Nannoolal et al. [16, 17])
- Critical property data (Nannoolal et al. [17, 18])
- Viscosities of liquids (Nannoolal et al. [17, 19])
- Vapour pressure as function of temperature (improved, Moller et al. [20])

Compared to available methods from literature, all these developments resulted in both a significantly broader range of applicability and greater reliability, as well as smaller estimation errors.

2. SOLVENTS

A solution can be defined as a mixture of at least 2 substances which has uniform chemical and physical properties throughout [21]. There are 2 parts to every solution, the solvent and the solute. The component(s) in excess is(are) usually called the solvent(s) while the component in deficiency is usually referred to as the solute. When a solvent can no longer dissolve a solute it is said to be saturated with the solute. Solvents are often used to convert substances to a form which is suitable for a particular use. Many substances exhibit the greatest usefulness when they are in solution (i.e. lower viscosities, etc.).

When the solute (solid, liquid or gas) dissolves it can either break up into separate molecules or clusters of strongly associated molecules when it enters solution (as happens when a solid sugar is dissolved in water) or the molecules can dissociate into ions (as happens when sodium chloride is dissolved in water). In both cases the process is a physical one since the solid can be reclaimed from the solution (boiling the NaCl solution and crystallizing the sugar out the sugar solution). The tendency of the solute to dissolve in the solvent depends both on how the solute molecules interact in the pure solute and on how the solute and the solvent interact in the solution. Solvents therefore need to be classified in order to understand how a solvent may interact with a prospective solute. This interaction is represented by the activity coefficient (see section 3.1). The lower the activity coefficient the better the solute will dissolve (as seen in section 5.1, for solid solutes this is not the only factor).

2.1. Applications/effects/selection

Some examples of the broad range of solvent applications and desirable properties are listed in the following paragraphs. The list is not meant to be an exhaustive one, but rather to give an idea of the broad scope of solvent application (the list contains desirable properties for the solvent, the ability to dissolve the active ingredient is taken as a given - and is given much attention in later chapters).

2.1.1. Viscosity

The viscosity of a substance can be defined as a measure of the fluids resistance to flow. For most applications it is useful to have a fluid which has a low viscosity because of the operational costs associated with using a viscous fluid (high pumping heads needed). Solvents are very often used to lower the viscosity of other substances thus converting them into a usable form. Some examples of this are:

- polymer adhesives which contain polymers which need to be dissolved in a solvent so that they can be applied to the surfaces,
- solid pharmaceuticals which need to be dispensed in liquid form (capsule or syrup),

- pesticides which need to be applied to crops in a (usually) liquid form and
- when a reactant is very viscous, suitable solvents can be used as a reaction medium (the purpose of this is very often twofold, firstly to alter the viscosity and diffusional mass transfer and secondly to alter the reaction equilibrium)

For this reason it is almost always desirable to have a solvent with a low viscosity.

2.1.2. Volatility

The volatility of a substance can be defined as a measure of the tendency of a substance to evaporate. The volatility of a substance is usually gauged from the vapour pressure. A substance with a high vapour pressure would have a high volatility and *vice versa*. High volatility can either be a desirable or an undesirable property. A solvent with a high relative volatility (i.e. relative to the solute) is useful when the solvent needs to be removed (i.e. evaporate off) so that the substance can dry. Examples of this are paints, coatings (varnishes etc.) and adhesives. When solvent extraction is used, it is useful to have a solvent with a sufficiently different relative volatility so that downstream separation is simplified. Higher volatility allows solvent removal at lower temperature, which reduces thermal stress on the solute while a lower volatility allows for the solute to be removed without having to vaporize the whole feed.

Highly volatile solvents are undesirable when the solvent is harmful to the environment. It is for this reason that more and more attention has been given to green solvents which are more eco-friendly. An extreme case is the widely discussed and researched ionic liquids, which have practically no vapour pressure. It is obviously undesirable to have a highly volatile solvent when the solvent is flammable. The flash point is used in design calculations and HAZOP (HAZard and OPerability) studies to determine how safe the solvent is. The flash point is the lowest temperature at which there will be sufficient vapours above the mixture to ignite.

2.1.3. Toxicity and environmental impact

Toxicity is a measure of how harmful (toxic) a substance is. Two common measures of toxicity are LC_{50} (lethal concentration which causes a 50% mortality rate) and LD_{50} (lethal dose which causes a 50% mortality rate). In most applications it is desirable to have a solvent with a low toxicity. For applications in the pharmaceutical industries the toxicity of the solvents used is obviously of paramount importance; for both the preparation and dispensing of the pharmaceuticals the toxicity of the solvent needs to be known. Due to the importance of solvent toxicity, clinical trials are often required to be carried out with substances synthesized with the same solvent(s) as those later produced in large scale. This means that the solvent selection has to be made at a very early stage and cannot be modified later. Another issue

which is strongly linked with volatility is the presence of VOC's (volatile organic compounds) in end products such as paints and fabrics. In some cases the maximum allowable concentration is in the ppm (part per million) range with prolonged exposure very often leading to serious medical conditions (e.g. acetone [22])

In recent times the toxicity of solvents (as well as other chemicals) has been given special attention and the branch of green chemistry has been formed to find alternatives to hazardous substances currently in use. Green solvents are solvents which have very little environmental impact. As mentioned above ionic liquids are often touted as green solvents due to their negligible vapour pressures; this is however not always true since they can still be toxic. Another consideration is the corrosivity/reactivity of a solvent, since this determines what the solvent can and can't be used with.

2.1.4. Selectivity and capacity

The selectivity of a solvent is defined as the ratio of the "affinity" of the solvent to two different solutes (often a desired and some undesired compound). The selectivity at infinite dilution (β^{∞}) can be defined in terms of the infinite dilution activity coefficients (the reciprocal of the infinite dilution activity coefficient is used as a measure of the affinity - see section 3.1) as:

$$\beta_{ij}^{\infty} = \frac{\gamma_{i,S}^{\infty}}{\gamma_{j,S}^{\infty}} \tag{2-1}$$

where subscript *j* refers to the solute, *i* to some undesired compound and *S* to the solvent. In the case where *i* and *j* are both liquids the selectivity is sufficient in selecting the desired solutes. In applications where one or both of *i* and *j* are solids then fusion data ($\Delta_{fus}H$ and T_m) are needed to make any final conclusions about the solvent selection. An everyday application where a selective solvent is especially desirable is detergents where the cleaning product must dissolve the dirt and grime but not damage what it is cleaning.

In applications involving liquid-liquid extraction, another useful quantity is the solvent capacity. The capacity of the solvent (m) relative to the other carrier phase is:

$$m_{S,A}^{\infty} = \frac{\gamma_{i,A}^{\infty}}{\gamma_{i,S}^{\infty}} \tag{2-2}$$

where the superscript *i* refers to the solute, *A* to the carrier phase and *S* to the solvent.

2.1.5. Liquidus range

The liquidus range is the temperature range for which a fluid is a liquid. The 2 bounds of the liquidus range are the melting (lower bound) and the boiling point at process pressure (upper

bound). The liquidus range is usually considered in crystallization applications. The liquidus range is important because the solvent needs to be liquid when the solute is dissolved at higher temperatures and also when the solute is crystallized out at lower temperatures (for cooling crystallization).

2.1.6. Computer Aided Molecular Design (CAMD)

CAMD is the process whereby a compound is selected based on a certain property. The properties are not measured but are obtained only from the molecular structure. CAMD is basically the reverse of property estimation, the main property estimation technique used with CAMD is group contribution methods. A wide range of group contribution methods (of varying accuracy) are available for the properties mentioned above [3, 4, 8, 10, 12, 14, 15]. Of all the properties discussed above the solubility/selectivity/capacity are the most difficult to determine since they rely on the way the solute interacts with a variable solvent. The other properties are generally simples since solvents are generally simple molecules where estimation methods are typically strong.

CAMD has received much attention in literature for crystallization [23] and liquid extraction [24, 25] solvent selection. There has also been some attention given in the literature to green solvent selection [26], whereby special attention is given to the environmental impact of the solvent. When using CAMD it is usually desirable to specify a set of groups that can be used in order to reduce the problem variables. For example one may specify that only saturated aliphatic compounds with oxygen groups may be used. When using certain groups there should obviously be some restrictions placed on the group selection (e.g. when using aromatic groups enough groups should be present to form the rings). When running the algorithm, specific values are not usually set, rather ranges are specified (for example a viscosity < 1 cP).

Therefore, in theory, it is possible to obtain a structure of a hypothetical solvent which has all of the desirable properties. The problem is that even if a hypothetical structure could be found it would more than likely not be a simple, readily available compound and would therefore need to be synthesised which is time consuming, expensive and impractical. It is therefore very often that, for large scale use, the economic considerations limit the solvent selection to a list of probably less than 20 different solvents.

2.2. Classification of solvents

Solvents are usually liquids, however supercritical fluids have received some attention since they have no surface tension, a rather high density and diffusion coefficients as well as low viscosity. They will however, not be considered in this work. Water is by far the most common solvent. Solvents can be broadly classified into 3 groups [27]:

- molecular liquids (e.g. water, hexane etc)
- atomic liquids (e.g. liquid mercury)
- ionic liquids (e.g. 1-butyl-3-methylimidazolium tetrafluoroborate)

Solvents can either fall into one or a mixture of the three classes. Of the three classes, atomic liquids are the least common solvents, especially in “everyday use”. Ionic liquids have received some attention in recent times due to their environmental friendliness (virtually no vapour pressure) and ability to almost be designed specifically for any purpose [27]. Molecular liquids are by far the most common solvents in use; they find application in many areas from adhesives to pharmaceuticals.

Solvents can be classified in terms of their physical behaviour. For example, they can be split into solvents which are high, medium and low boilers or solvents with high, medium or low viscosity. These classifications make a lot of sense when considering operational costs and design specification but make little sense when trying to classify solute-solvent interactions. It is clear that in order to account for these a chemical classification is necessary.

The basic principle when determining solubility is “like dissolves like”, where molecules with similar intermolecular forces will generally dissolve. The intermolecular forces which make up molecules are dispersive forces (van der Waals forces); polar forces (dipole-dipole) and hydrogen bonding. Dispersive interactions of varying strength exist between all molecules. They depend mainly on the polarizability of the outer shell electrons of the different atoms in the molecule. The polarizability is usually small for small strongly electronegative atoms (“hard” electrons) like fluorine and strong for large atoms where the orbitals are far from the nucleus and electrostatic attraction is shielded by the inner electron shells (“soft” electrons) as in the case of bromine or iodine. In addition, molecules may interact by polar and hydrogen bonding forces, which are both usually much stronger than dispersive forces. Hydrogen bonds are a special form of polar interactions, whereby the hydrogen has to be bonded to an electronegative atom like fluorine, oxygen or nitrogen; the hydrogen bond acceptor is either fluorine, oxygen or nitrogen and the donor is the hydrogen (this discussion applies to intermolecular forces so the bond between an O and H is not hydrogen bonding but the bond between 2 OH atoms from different molecules is).

The hydrogen bond is strongest when the two negative hetero atoms and the positive hydrogen are precisely lined up with the hydrogen atom in the middle. Any deviations from the straight lines lead to a significant weakening of the interaction. This is similar to the behaviour of covalent bonds.

The problem is therefore to determine which intermolecular forces predominate in the molecule. The various approaches for determining these forces are given in the following paragraphs.

2.2.1. Trouton's rule

Trouton's rule states that the entropy of vaporization ($\Delta_b S$) at the normal boiling point (T_b) has a constant value of 88 J.mol^{-1}.K^{-1} for all compounds.

$$\Delta_b S = \frac{\Delta_b H}{T_b} \approx 88\ J.mol^{-1}.K^{-1} \tag{2-3}$$

This rule holds for a fair number of compounds but deviates especially for hydrogen bonding compounds (see Table 2-1). As shown in Table 2-1 there is a positive deviation from Trouton's rule for almost all compounds which hydrogen bond. The reason for this is that the hydrogen bonding increases the order in the liquid phase (i.e. reduces the entropy) and therefore the entropy change upon vaporization is larger than expected (these hydrogen bonds are not present in the vapour phase). This deviation means that depending on the entropy of vaporization the compounds can either be classified as hydrogen bonding or non-hydrogen bonding.

Unfortunately this deviation is not always so clear cut, for example acetic acid, which is hydrogen bonding, has an entropy of vaporization of 60.5 J.mol^{-1}.K^{-1} [28]. The reason for this is widely attributed to dimer formation in the vapour phase, however dimers form in both the liquid and vapour phase. It is therefore probable that this dimer formation increases the entropy in the liquid phase (since dimers behave as non-polar molecules and these non-polar molecules disrupt the structure of the monomer fluid) and decreases it in the vapour phase leading to a smaller than expected change. The presence of the dimers in a acetic acid is shown as a function of temperature (using the dimerization equilibrium constants for the vapour and the liquid [29]) in Fig. 2-1. The plot clearly shows that at the normal boiling point the concentrations of the dimers in the liquid and vapour phase are almost equal and this is the reason why the enthalpy of vaporization (and therefore the entropy of vaporization) is lower than expected.

Large molecules also tend to deviate from Trouton's rule, for example the entropy of vaporization of eicosane is 93.2 J.mol^{-1}.K^{-1} [30]. This is because larger molecules have a more ordered liquid phase owing to their size restricting free movement. Therefore the entropy change upon vaporization is larger than expected. Thus Trouton's rule is a useful means of classification but cannot be blindly applied.

Table 2-1 Examples of the entropy of vaporization (at the normal boiling point) data from Myrdal et al. [30].

Name	$\Delta_b S$ (J.mol^{-1}.K^{-1})
Ethanol	110.1
Water	109.0
Pentanoic acid	96.2
Acetone	88.4
Benzene	87
Diethyl ether	86.5
n-Octane	86.3

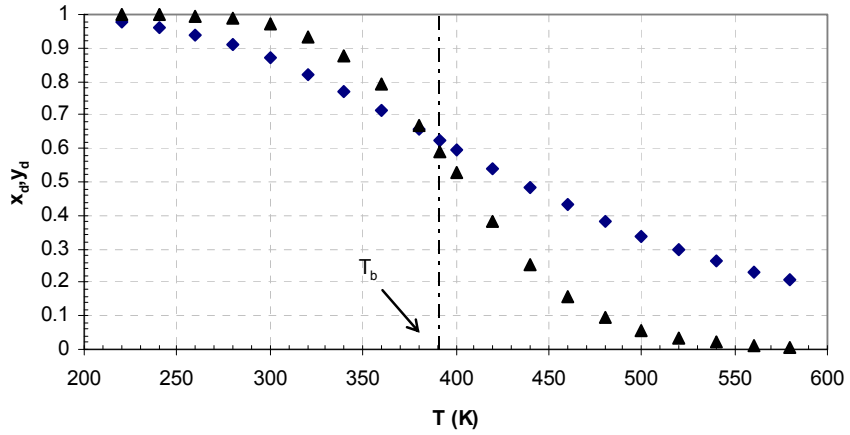


Fig. 2-1 Dimer mole fractions in the liquid (x) and vapour (y) phase as a function of temperature at 1 atm (◆ - liquid dimer mole fraction x_d, ▲ – vapour dimer mole fraction y_d).

2.2.2. Cohesive pressure

The cohesive pressure (also called the cohesive energy density) of a substance is defined by the following relationship:

$$c = \frac{\Delta_v U}{V} \approx \frac{\Delta_v H - RT}{V} \tag{2-4}$$

The cohesive pressure is used as a measure of the total intermolecular forces between molecules. The reason for this is that in order to vaporize a fluid all the intermolecular forces need to be broken, and hence the heat of vaporization will contain this information. Large values of cohesive pressure indicate compounds which are polar/H-bonding, while lower values indicate compounds which are non-polar. The problems associated with dumping all these intermolecular forces into a single parameter are discussed in section 3.2.3.1.

2.2.3. Internal pressure

The internal pressure of a substance is defined by the following relationship:

$$\pi = \left(\frac{\partial U}{\partial V}\right)_T \tag{2-5}$$

The internal pressure is used as a measure of the non-hydrogen-bonding forces present (i.e. polar and dispersive forces). This is possibly due to the fact that polar and dispersive forces are “always on” while hydrogen bonding forces are more “on/off” (i.e. polar and dispersive forces may get stronger or weaker depending on molecule proximity while hydrogen bonds either exist or they don’t i.e. on or off). Therefore, when the molar volume changes the

hydrogen bonding forces will not be disrupted, while the others will. The internal and cohesive pressures are often used together as a measure of the types of forces present in a molecule:

$$n = \frac{c}{\pi} \tag{2-6}$$

where values of n much greater than unity indicate hydrogen bonding molecules and those which are close to unity are non-polar molecules. Consider the examples shown in Table 2-2, water clearly has a large amount of hydrogen bonding while toluene does not.

Table 2-2 Internal and cohesive pressure for some sample compounds, data from Reichardt [27].

Name	π (MPa)	c (MPa)	n
Perfluoroheptane	215	151	0.70
Toluene	355	337	0.95
Ethanol	293	687	2.34
Water	151	2294	15.19

2.2.4. Acid-base classification

A common method to classify solvents is by their acid-base behaviour. According to the Brønsted/Lowry acid-base theory, acids are proton donors and bases are proton acceptors. This is more or less what happens with hydrogen bonding, except that the proton is not transferred but shared:

$$\text{R-X-H + :Y-R'} \rightarrow \text{R-X-H---Y-R'}$$

where XH is some acidic group and Y is some basic group. The strength of an acid is measured by the pK_a, where for some acid HA:

$$pK_a = -\log\frac{[A^-][H^+]}{[HA]} \tag{2-7}$$

Therefore using the pK_a substances can be classified as protic or non-protic.

According to the Lewis acid-base theory, an acid is an electron pair acceptor while a base is an electron donor. A Lewis base is a Brønsted/Lowry base but a Lewis acid is not necessarily a Brønsted/Lowry acid. For further classification the acids and bases are classified as hard or soft, where a hard acid or base is usually derived from small atoms with high electronegativities and low polarizabilities. These are typically polar or hydrogen bonding compounds, for example water. A soft acid or base is usually derived from larger atoms with low electronegativities and are usually polarizable. These are typically non-polar or slightly polar compounds for example alkanes.

2.2.5. Solvent classification

The advantage of a good solvent classification is that solvents in the same categories should behave in a similar way; qualitative predictions of solvent behaviour can therefore be gleaned from analogous solvents. Organic solvents can be very generally classified into 3 categories:

- Apolar, aprotic
- Polar, aprotic
- Protic

This classification is however probably too broad to observe any meaningful trends. Examples of two more complex classifications are shown in Table 2-3.

Table 2-3 Examples of two more complex solvent classifications.

Chastrette et al. [31] classification	Gramatica et al. [32] classification
POLAR	
aprotic dipolar (AD) aprotic highly dipolar (AHD) aprotic highly dipolar and polarizable (AHDP)	aprotic polar (AP)
APOLAR	
aromatic apolar (ARA) aromatic relatively polar (ARP)	aromatic apolar or lightly polar (AALP)
electron pair donors (EPD)	electron pair donors (EPD)
PROTIC	
hydrogen bonding (HB) hydrogen bonding strongly associated (HBSA)	hydrogen bonding donors (HBD)
MISC	
miscellaneous (MISC)	aliphatic aprotic apolar (AAA)

The solvent classification should be able to explain most of the extraordinary phase splitting that occurs. For instance the famous four phase mixture of Hildebrand [33], as shown in Fig. 2-2 in which all four fluids are sufficiently different to form 4 distinct phases.

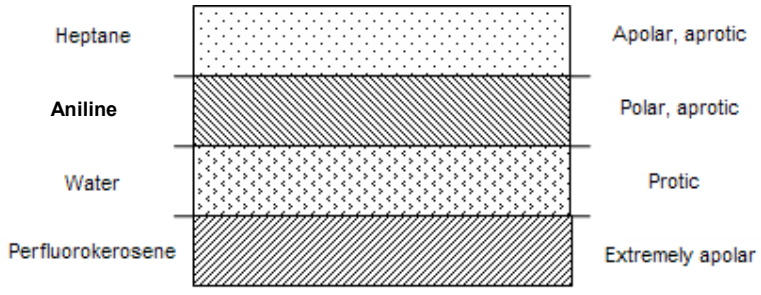


Fig. 2-2 The four phase splitting example of Hildebrand [33].

3. THEORETICAL APPROACHES TO SOLUTE ACTIVITY

3.1. Thermodynamic framework

For a single phase fluid in a closed system where no reaction occurs, the Gibbs energy can be related to temperature and pressure by (associating systems are not considered here):

$$d(nG) = (nV)dP - (nS)dT \tag{3-1}$$

where

$$\left[\frac{\partial(nG)}{\partial P}\right]_{T,n} = nV \tag{3-2}$$

$$\left[\frac{\partial(nG)}{\partial T}\right]_{P,n} = -nS \tag{3-3}$$

A more general case is where there is a single phase open system where mass can be transferred across the system boundary, the Gibbs energy for this case is then not only a function of temperature and pressure but also of the number of moles of each species (for m species):

$$nG = f(P,T,n_1,...,n_m) \tag{3-4}$$

Taking the total differential of Eqn. (3-4) and inserting Eqn's (3-2) & (3-3) the following is obtained:

$$d(nG) = (nV)dP - (nS)dT + \sum_{i=1}^{m}\left[\frac{\partial(nG)}{\partial n_i}\right]_{P,T,n_j} dn_i \tag{3-5}$$

where the chemical potential of species *i* is defined as (the over bar represents a partial molar property):

$$\mu_i \equiv \left[\frac{\partial(nG)}{\partial n_i}\right]_{P,T,n_j} = \bar{G}_i \tag{3-6}$$

Now consider a closed system of two phases (α and β) in equilibrium, where each phase is an open system:

$$d(nG)^{\alpha} = (nV)^{\alpha} dP - (nS)^{\alpha} dT + \sum_{i=1}^{m} \mu_i^{\alpha} dn_i^{\alpha} \tag{3-7}$$

$$d\left(nG\right)^{\beta} = \left(nV\right)^{\beta} dP - \left(nS\right)^{\beta} dT + \sum_{i=1}^{m} \mu_i^{\beta} dn_i^{\beta} \tag{3-8}$$

In Eqn's (3-7) and (3-8) thermal and mechanical equilibrium are assumed and therefore temperature and pressure are uniform throughout the whole system. The Gibbs energy differential for the whole system is simply the sum of the Gibbs energy in each phase (where $\left(nV\right) = \left(nV\right)^{\alpha} + \left(nV\right)^{\beta}$ and $\left(nS\right) = \left(nS\right)^{\alpha} + \left(nS\right)^{\beta}$):

$$d\left(nG\right) = \left(nV\right)dP - \left(nS\right)dT + \sum_{i=1}^{m} \mu_i^{\alpha} dn_i^{\alpha} + \sum_{i=1}^{m} \mu_i^{\beta} dn_i^{\beta} \tag{3-9}$$

Since the whole system is closed, Eqn. (3-1) holds, also since the system is closed mass lost by one phase must be equal to mass gained by another and therefore at equilibrium:

$$\sum_{i=1}^{m} \left(\mu_i^{\alpha} - \mu_i^{\beta}\right) dn_i^{\alpha} = 0 \tag{3-10}$$

Therefore the chemical potentials in each phase must be equal at equilibrium; this relation can be easily extended to systems with π phases to yield:

$$T^{(1)} = T^{(2)} = ... = T^{(\pi)} \tag{3-11}$$

$$P^{(1)} = P^{(2)} = ... = P^{(\pi)} \tag{3-12}$$

$$\mu_i^{(1)} = \mu_i^{(2)} = ... = \mu_i^{(\pi)} \quad i = 1..m \tag{3-13}$$

For each phase there will be m+2 (T, P and μ's) intensive variables, however only m+1 of these variables are independent. Thus, for the total system, there will be $\pi\left(m+1\right)$ independent variables. The number of equilibrium relations given by Eqn's (3-11)-(3-13) is $\left(\pi-1\right)\left(m+2\right)$. Therefore the number of intensive variables that can be set (known as the degrees of freedom of the system) can be given by:

$$F = \pi\left(m+1\right) - \left(\pi-1\right)\left(m+2\right) = m + 2 - \pi \tag{3-14}$$

The chemical potential of a pure species is related to the temperature and pressure through an equation analogous to Eqn. (3-1):

$$d\mu_i = v_i dP - s_i dT \tag{3-15}$$

This equation can be integrated by using some reference state (superscript r):

$$\mu_i\left(T,P\right) = \mu_i\left(T^r,P^r\right) - \int_{T^r}^{T} s_i dT + \int_{P^r}^{P} v_i dP \tag{3-16}$$

The reference state, often called the standard state, can be arbitrarily specified, but is usually specified so as to simplify the required calculations (see section 5.1 for an example).

The chemical potential is somewhat of an abstract concept and doesn't have a direct equivalent in the real world [34]. It is therefore desirable to express the chemical potential in terms of some function which does have a real world equivalent. Assuming a pure ideal gas ($Pv = RT$), and integrating Eqn. (3-15) under isothermal conditions (using superscript 0 to denote the standard state):

$$\mu_i - \mu_i^0 = RT \ln \frac{P}{P^0} \tag{3-17}$$

The advantage of Eqn. (3-17) is that it simply relates temperature and pressure to the more abstract chemical potential. The disadvantage is that it is only suitable for pure ideal gases. This problem was overcome by Lewis [35] by defining a function called fugacity (f, when used in solution it is given the symbol $\hat{f}_i$), where he described the fugacity as a measure of the tendency of a molecule to escape from the phase in which it is [35]. The fugacity replaces pressure for real solids, liquids and gases:

$$\mu_i - \mu_i^0 = RT \ln \frac{\hat{f}_i}{f_i^0} \tag{3-18}$$

The ratio $\hat{f}_i / f_i^0$ was called the activity (a_i) by Lewis [36] (the fugacity in solution is not used for the standard state since it is more convenient to chose the standard state as the pure liquid, the superscript 0 is therefore often left off). Eqn. (3-18) represents an isothermal change in chemical potential when going from f_i^0 to $\hat{f}_i$. Therefore the standard states for all components in all phases must be at the same temperatures; the pressure and composition of the standard states, however, do not necessarily have to be the same. The phase equilibrium criterion given in Eqn. (3-13) can therefore be rewritten in terms of fugacity as:

$$\hat{f}_i^{(1)} = \hat{f}_i^{(2)} = \ldots = \hat{f}_i^{(\pi)} \quad i = 1..m \tag{3-19}$$

The activity coefficient is defined as the ratio of the fugacity in solution to the ideal fugacity in solution; it can also be defined in terms of the activity relative to some measure of composition:

$$\gamma_i \equiv \frac{\hat{f}_i}{\hat{f}_i^{id}} = \frac{a_i}{x_i} = \frac{\hat{f}_i}{x_i f_i^0} \tag{3-20}$$

The activity coefficient is a measure of the deviation from ideality. For ideal solutions the activity coefficient is unity. The two important limits when working with activity coefficients are:

$$\lim_{x_i \to 1} \gamma_i \equiv 1 \tag{3-21}$$

$$\lim_{x_i \to 0} \gamma_i \equiv \gamma_i^{\infty} \tag{3-22}$$

where γ_i^{∞} is called the infinite dilution activity coefficient. The activity coefficient can be defined in terms of Gibbs energy by using the following procedure. Taking the difference for the real and the ideal chemical potential (Eqn. (3-18), partial molar Gibbs energy is used for the sake of notation – see Eqn. (3-6)).

$$\bar{G}_{i(real)} - \bar{G}_{i(ideal)} = RT\left[\ln \hat{f}_{i(real)} - \ln \hat{f}_{i(ideal)}\right] \tag{3-23}$$

This can be rewritten in terms of excess partial molar Gibbs energy (where an excess property is simply the difference between the real and the ideal solution denoted by superscript *E*).

$$\bar{G}_i^E = RT \ln\left[\frac{\hat{f}_i}{\hat{f}_{i(ideal)}}\right] \tag{3-24}$$

To simplify the notation the subscript *real* was dropped. Substituting Eqn. (3-20) (recalling that $a_i = \hat{f}_i / f_i^0$) into Eqn. (3-24) results in the relationship between excess Gibbs energy and activity coefficient:

$$\frac{\bar{G}_i^E}{RT} = \ln \gamma_i \tag{3-25}$$

Therefore if an analytical expression for the concentration dependence of the excess Gibbs energy is known then it is a fairly simple process to find an analytical expression for the activity coefficient. Some of the more popular expressions for excess Gibbs energy are shown in the sections that follow.

The temperature and pressure dependence of the activity coefficient can be shown using the following procedure. Taking the total differential of the excess partial Gibbs energy yields the following:

$$d\left(\frac{\bar{G}_i^E}{RT}\right) = \left(\frac{\partial \bar{G}_i^E}{\partial P}\right)_{T,n} \frac{dP}{RT} + \left(\frac{\partial\left(\bar{G}_i^E / T\right)}{\partial T}\right)_{P,n} \frac{dT}{R} + \sum_i \left(\frac{\partial \bar{G}_i^E}{\partial n}\right)_{P,T,n_{j \neq i}} \frac{dn_i}{RT} \tag{3-26}$$

Eqn. (3-5) can be rewritten in terms of excess properties as:

$$d\left(\frac{nG^E}{RT}\right) = \frac{nV^E}{RT} dP - \frac{nH^E}{RT^2} dT + \sum_i \ln \gamma_i dn_i \tag{3-27}$$

Taking the partial derivative of Eqn. (3-27) and matching coefficients with Eqn. (3-26) leads to the following temperature and pressure dependencies of the activity coefficient:

$$\left(\frac{\partial \ln \gamma_i}{\partial P}\right)_{T,x} = \frac{\bar{V}_i^E}{RT} \tag{3-28}$$

$$\left(\frac{\partial \ln \gamma_i}{\partial T}\right)_{P,x} = -\frac{\bar{H}_i^E}{RT^2} \tag{3-29}$$

At temperatures far removed from the critical point the pressure dependence is usually ignored, however the temperature dependence is very often non-trivial and needs to be given some attention when developing activity coefficient models.

3.2. Empirical and semi-empirical models for G^E

The following section outlines the main models that are available in the literature for the representation of excess Gibbs free energy. Polynomial type expansions of G^E such as Redlich-Kister and Margules and empirical correlations like van Laar are not covered in the following sections since they are fairly well known (and widely covered in the literature [34, 37, 38]) and the local composition models have been shown to be more suitable [39]. The models covered in the following sections are by no means an exhaustive list; rather they are the ones which find the widest use in the prediction/modelling of complex molecule solubility.

3.2.1. Local composition concept (Wilson, NRTL, UNIQUAC)

Wilson [40, 41] is widely credited with developing the local composition concept even though Guggenheim [42, 43] was probably the first; nevertheless Wilson was the first to provide a practical application of the concept (Guggenheim's model requires the simultaneous solution of chemical equilibria equations while Wilson provides an explicit approximation). The concept of local composition assumes that the composition in the region of some molecule is different from the composition in the bulk liquid. Wilson expressed the local compositions (ξ_{ii} is the volume fraction of molecules *i* around a molecule of *i*, similarly ξ_{ji} is the volume fraction of molecule *j* around molecule *i*) for a system of *m* components as:

$$\xi_{ii} = \frac{x_i}{\sum_{j=1}^{m} x_j \left(\frac{v_j}{v_i}\right) \exp\left(-\frac{\lambda_{ji} - \lambda_{ii}}{RT}\right)} \tag{3-30}$$

$$\sum_{j=1}^{m} \xi_{ji} = 1 \tag{3-31}$$

where λ_{ji} is the interaction between component *j* and *i*. Then by using an equation analogous to the Flory-Huggins expression for athermal solutions he defined the excess Gibbs energy of mixing, given for *m* components as:

$$\frac{G^E}{RT} = \sum_{i=1}^{m} x_i \ln\left(\frac{\xi_{ii}}{x_i}\right) \tag{3-32}$$

By substitution of Eqn. (3-30) into Eqn. (3-32) the following results:

$$\frac{G^E}{RT} = -\sum_{i=1}^{m} x_i \ln\left(\sum_{j=1}^{m} x_j A_{ji}\right) \tag{3-33}$$

where

$$A_{ji} = \frac{v_j}{v_i}\exp\left(-\frac{\lambda_{ji}-\lambda_{ii}}{RT}\right) = \frac{v_j}{v_i}\exp\left(-\frac{\Delta\lambda_{ji}}{RT}\right) \tag{3-34}$$

(Note: this derivation is the one given by Wilson [40]; there are however different ways to arrive at the same result - see Tsuboka and Katayama [44]). The great advantage of the local composition equations is that they can predict multi-component behaviour from binary data. The Wilson equation is still widely used today, its biggest flaw is that it cannot account for 2 liquid phases in binary systems (since the second derivative of the activity coefficient with respect to composition always results in a positive value and therefore only a single stable phase [38]). The general form of the activity coefficient for the Wilson equation of some component *i* in a mixture of *m* components is (since this work is involved quite heavily with infinite dilution data the binary infinite dilution data expression is shown for this and further models):

$$\ln\gamma_i = -\ln\left(\sum_{j=1}^{m} x_j A_{ji}\right) + 1 - \sum_{k=1}^{m}\frac{x_k A_{ik}}{\sum_{j=1}^{m} x_j A_{jk}} \tag{3-35}$$

$$\ln\left(\gamma_2^{\infty}\right) = 1 - A_{12} - \ln\left(A_{21}\right) \tag{3-36}$$

Since the initial work of Wilson, many local composition models have been developed (including modifications to the Wilson equation which allow phase splitting [44, 45]). The two most popular and widely used are the NRTL (Non-Random Two Liquid) and UNIQUAC (UNIversal QUAsi Chemical) equations. The NRTL [46] equation introduced a 3rd parameter call the non-randomness parameter (α_{ij} , where $\alpha_{ij} = \alpha_{ji}$), local mole fractions (as with local volume fractions, local mole fractions sum to 1) are used and the excess Gibbs energy is defined differently:

$$x_{ij} = \frac{x_i G_{ij}}{\sum_{i=1}^{m} x_i G_{ij}} \tag{3-37}$$

$$G_{ij} = \exp\left(-\alpha_{ij}\tau_{ij}\right) \tag{3-38}$$

$$\tau_{ij} = \frac{g_{ij} - g_{jj}}{RT} \tag{3-39}$$

$$\frac{G^E}{RT} = \sum_{i=1}^{m} x_i \left(\sum_{j}^{m} x_{ji} \left(\frac{g_{ji}}{RT} - \frac{g_{ii}}{RT} \right) \right) \tag{3-40}$$

The NRTL equation is widely used and it is superior to the Wilson equation in that it can account for 2 (or more) liquid phases. A disadvantage of the NRTL equation is that for binary LLE (liquid-liquid equilibria) α_{12} must be set and cannot be solved for and ternary and higher LLE requires large amount of data if one wishes to regress values of α_{12}. Another disadvantage of the NRTL equation is that it cannot sufficiently account for asymmetric mixtures [47] (i.e. mixtures with components of vastly different sizes). This can be simply solved by considering volume fractions [47, 48]. The general form of the activity coefficient for the NRTL equation of some component *i* in a mixture of *m* components is:

$$\ln \gamma_i = \frac{\sum_{j=1}^{m} \tau_{ji} G_{ji} x_j}{\sum_{l=1}^{m} G_{li} x_i} + \sum_{j=1}^{m} \frac{x_j G_{ij}}{\sum_{l=1}^{m} G_{lj} x_l} \left(\tau_{ij} - \frac{\sum_{r=1}^{m} x_r \tau_{rj} G_{rj}}{\sum_{l=1}^{m} G_{lj} x_l} \right) \tag{3-41}$$

$$\ln\left(\gamma_2^{\infty}\right) = \tau_{12} + G_{12}\tau_{21} \tag{3-42}$$

The UNIQUAC [49] equation is somewhat of a compromise between Wilson and NRTL in that it has only 2 model parameters but can account for phase splitting. The derivation of the UNIQUAC equation is given by Maurer [50], the basic idea is that the excess Gibbs free energy is split into 2 parts. One part for the energetic interactions between molecules (called the residual) and one for the effect of size/shape interaction (called the combinatorial – see section 3.2.4):

$$\frac{G^E}{RT} = \left(\frac{G^E}{RT}\right)_{\text{Residual}} + \left(\frac{G^E}{RT}\right)_{\text{Combinatorial}} \tag{3-43}$$

Following the derivation given by Maurer [50] the following expressions result for the residual and combinatorial expressions:

$$\left(\frac{G^E}{RT}\right)_{\text{Residual}} = -\sum_{i=1}^{m} q_i x_i \left(\sum_{j=1}^{m} \theta_j \tau_{ji} \right) \tag{3-44}$$

$$\left(\frac{G^E}{RT}\right)_{\text{Combinatorial}} = \sum_{i=1}^{m} x_i \ln\frac{\Phi_i}{x_i} + \frac{z}{2}\sum_{i=1}^{m} q_i x_i \ln\frac{\theta_i}{\Phi_i} \tag{3-45}$$

$$\psi_{ij} = \exp\left[-\frac{u_{ij} - u_{jj}}{RT}\right] = \exp\left[-\frac{\Delta u_{ij}}{RT}\right] \tag{3-46}$$

where r and q are the size and surface area parameters (defined in terms of van der Waals volume and surface area) respectively, ψ_{ij} is the binary interaction parameter, Φ is the volume fraction and θ is the surface fraction (r and q and defined by Eqn's (3-60) and (3-61)):

$$\Phi_i = \frac{x_i r_i}{\sum_j x_j r_j} \tag{3-47}$$

$$\theta_i = \frac{x_i q_i}{\sum_j x_j q_j} \tag{3-48}$$

Taking the partial derivative of Eqn. (3-43) yields the following relationships for the activity coefficient (assuming that the coordination number is 10 for all molecules):

$$\ln\gamma_i = \ln\gamma_i^R + \ln\gamma_i^C \tag{3-49}$$

$$\ln\gamma_i^R = q_i\left[1 - \ln\left(\sum_j \theta_j\psi_{ji}\right) - \sum_j \frac{\theta_j\psi_{ij}}{\sum_k \theta_k\psi_{kj}}\right] \tag{3-50}$$

$$\ln\left(\gamma_2^{\infty,R}\right) = q_2\left(1 - \ln\left(\psi_{12}\right) - \psi_{21}\right) \tag{3-51}$$

$$\ln\gamma_i^C = 1 - \frac{\Phi_i}{x_i} + \ln\frac{\Phi_i}{x_i} - 5q_i\left(1 - \frac{\Phi_i}{\theta_i} + \ln\frac{\Phi_i}{\theta_i}\right) \tag{3-52}$$

$$\ln\left(\gamma_2^{\infty,C}\right) = 1 - \frac{r_2}{r_1} + \ln\left(\frac{r_2}{r_1}\right) - 5q_2\left(1 - \frac{r_2 q_1}{r_1 q_2} + \ln\left(\frac{r_2 q_1}{r_1 q_2}\right)\right) \tag{3-53}$$

where the superscripts C and R refer to the combinatorial and residual parts respectively.

In the Wilson, NRTL and UNIQUAC models, for the binary case, the composition around molecule type 1 is treated as independent from the composition around molecule type 2. However it has been noted in literature [51] that local compositions cannot be independent since there is some overall concentration which must be obeyed. Therefore the interaction parameters (Δu_{ij}, A_{ij} & g_{ij}) and the NRTL non-randomness parameter cannot be concentration independent. The UNIQUAC coordination number (z) cannot be assumed *a priori* to be 10 for all molecules. In practice these systematic errors are accounted for by the interaction

parameters but can never be zero. These inconsistencies are not present in the COSMO-models (discussed in section 3.2.3.4 below) since they account for local activity and not local composition.

3.2.2. Functional group activity coefficients

Probably the most widely known and used of the activity coefficient prediction methods are the group contribution methods. The prediction of activity coefficients by the concept of functional group interactions (also known as the solution of groups concept) is generally attributed to the work of Wilson and Deal [41]. As the name suggests solution of groups entails breaking a molecule up in to functional groups and considering mixtures as solutions of these groups and not compounds. For example consider a mixture of ethanol and hexane, the mixture would be made up of only alkane (CH_2) and alcohol (OH) groups (see Table 3-1).

Table 3-1 Group frequency illustration for the mixture hexane – ethanol.

Groups	Ethanol (1)	Hexane (2)	Mixture
CH_2	2	6	$2 \cdot x_1 + 6 \cdot x_2$
OH	1	0	x_1

Deal and Derr [52, 53] were the first to develop a model for the solution of groups; it was called the ASOG (Analytical Solution Of Groups) method. Later Fredenslund et al. [9] developed the UNIFAC (UNIversal Functional group Activity Coefficient) model; both models are the same in principle but differ in the details. The UNIFAC model has subsequently had many modifications, often developed for specific applications [10, 54-57].

The great advantage of the solution of groups approach is that the number of possible functional groups is much less than the number of different compounds, this means that the required number of binary interaction parameters is much lower. For example if the CH_2-OH interactions (model parameters in the ASOG and UNIFAC equations) are known then the behaviour of any alkane - aliphatic alcohol (also aliphatic alcohol – aliphatic alcohol and alkane - alkane) mixture can be predicted, drastically reducing the number of experimental measurements needed.

As mentioned above, ASOG was the first method to use the principle of the solution of groups. In a similar way to the UNIQUAC equation the activity coefficient is assumed to be made up of a part for molecular forces interactions ($\ln \gamma_i^G$) and a part for size/shape interactions ($\ln \gamma_i^{FH}$).

$$\ln \gamma_i = \ln \gamma_i^G + \ln \gamma_i^{FH} \tag{3-54}$$

The size/shape term is calculated from an equation which is analogous to the Flory-Huggins equation for athermal solutions (see section 3.2.4):

$$\ln \gamma_i^{FH} = \ln \frac{v_i^{FH}}{\sum_j v_j^{FH} x_j} + 1 - \frac{v_i^{FH}}{\sum_j v_j^{FH} x_j} \tag{3-55}$$

where v_i^{FH} is the (total) number of size groups in molecule *i*, and x_i is the mole fraction. Combinatorial expressions are always based on the size and shape of the molecules (as opposed to the functional groups). In contrast, residual expressions are calculated for group activity coefficient in the pure components ($\Gamma_k^{(i)}$) and the mixture (Γ_k) respectively. The component activity coefficients are calculated from the group activity coefficients by the following equation:

$$\ln \gamma_i^G = \sum_k v_{ki} \left(\ln \Gamma_k - \ln \Gamma_k^{(i)} \right) \tag{3-56}$$

where v_{ki} is the frequency of group k in molecule *i*, Γ_k is the activity coefficient of group k in the mixture and $\Gamma_k^{(i)}$ is the activity coefficient of group *k* in the pure component *i*. For compounds with only one group (e.g. hexane), $\Gamma_k^{(i)}$ will be unity, however for compounds with multiple groups (e.g. ethanol) it will have a deviation from ideal behaviour. The group activity coefficients are calculated from the Wilson equation (Eqn. (3-35)):

$$\ln \Gamma_k = -\ln \sum_m X_m a_{mk} + \left(1 - \sum_m \frac{X_m a_{km}}{\sum_n x_n a_{nm}} \right) \tag{3-57}$$

$$X_m = \frac{\sum_j x_j v_{mj}}{\sum_j x_j \sum_n v_{nj}} \tag{3-58}$$

where X_m is the group fraction and a_{km} is given by equations similar to Eqn. (3-34), but for the interaction between groups and not compounds. $\Gamma_k^{(i)}$ is calculated from the same equation but instead of X_m, $X_m^{(i)}$ is used, where:

$$X_m^{(i)} = \frac{v_m^{(i)}}{\sum_n v_n^{(i)}} \tag{3-59}$$

UNIFAC follows the same basic scheme as ASOG (Eqn.(3-54) – (3-59)). Depending on the UNIFAC version (original, modified Dortmund etc.) different combinatorial expressions are used (see section 3.2.4). For original UNIFAC the combinatorial expression of UNIQUAC (Eqn. (3-52)) is employed, where the surface and volume parameters are given by:

$$r_i = \sum_k v_k^{(i)} R_k \quad \text{and} \quad q_i = \sum_k v_k^{(i)} Q_k \tag{3-60}$$

$$R_k = \frac{V_{wk}}{15.17\,cm^3.mol^{-1}} \quad \text{and} \quad Q_k = \frac{A_{wk}}{2.5\times10^9\,cm^2.mol^{-1}} \tag{3-61}$$

where R_k and Q_k are the group volume and surface parameters respectively and V_{mk} & A_{mk} are the group van der Waals volume and surface area as given by Bondi [58]. The normalization factors 15.17 and 2.5 x 10^9 are determined by the volume and external surface area of a CH_2 unit of polymethylene [37] (see Appendix D for further discussion on the reference surface and volume). The expression for the residual is analogous to Eqn. (3-56):

$$\ln\gamma_i^R = \sum_k v_k^{(i)}\left(\ln\Gamma_k - \ln\Gamma_k^{(i)}\right) \tag{3-62}$$

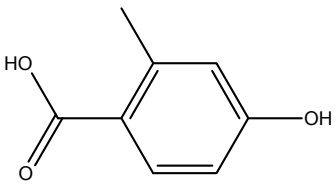
Unlike with ASOG group activity coefficients are calculated from the UNIQUAC residual term:

$$\ln\Gamma_k = Q_k\left[1 - \ln\left(\sum_m \theta_m\Psi_{mk}\right) - \sum_m \frac{\theta_m\Psi_{km}}{\sum_n \theta_n\Psi_{nm}}\right] \tag{3-63}$$

$$\theta_m = \frac{Q_m X_m}{\sum_n Q_n X_n} \tag{3-64}$$

where Ψ_{mk} is given by an equation analogous to Eqn. (3-46) but is for the interaction between groups and not molecules, and as with ASOG, $\Gamma_k^{(i)}$ is calculated from the same expression with X_m being replaced by $X_m^{(i)}$. As explained above with both methods the residual is calculated from the groups while the combinatorial is calculated from the molecule properties.

As mentioned above, these group contribution methods have found wide use in industry. The cases for which they work best are mixtures containing molecules of similar size. The reason for this is that most of the data which were used to fit the binary interaction parameters are for systems made up of such components. Nevertheless, they can still be applied to the prediction of activity coefficients of solutions containing large molecules. This is typically the case when a solid is dissolved in a solvent (see section 5.1 for the calculation details). As a means of comparison, the methods in this and the following sections will be used to generate predictions of the solubility of methylparaben and 1,2-dihydro-acenaphthylene in a variety of solvents at 298.15 K. The molecular structures of methylparaben and 1,2-dihydro-acenaphthylene and the various thermo physical properties required are shown in Fig. 3-1. These 2 compounds were chosen on the basis of available data, and also because of the nature of the solutes (one polar and one non-polar as a means of broadly analysing the methods).



Methylparaben

PROPERTIES

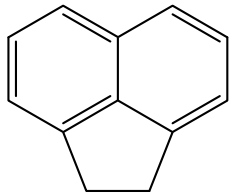
Molar mass : 152.15 g/mol

Enthalpy of fusion: 24.29 kJ/mol

Melting temperature: 404.15 K

Liquid molar volume: 118.87 cm^3/mol

Fig. 3-1 The molecular formula and thermophysical properties required for methylparaben (all data taken from DDB [28]).



1,2-Dihydro-acenaphthylene

PROPERTIES

Molar mass : 154.21 g/mol

Enthalpy of fusion: 21.52 kJ/mol

Melting temperature: 367.15 K

Liquid molar volume: 150.60 cm^3/mol

Fig. 3-2 The molecular formula and thermophysical properties required for 1,2-dihydro-acenaphthylene (all data taken from DDB [28]).

The results for the test set data are shown in Table 3-2. The results for the ASOG method are not included as there are many missing groups and application to SLE is not very common. For UNIFAC and mod. UNIFAC there are two different sets of parameters available, the published and the consortium values. The published values are freely available in the literate while the consortium values are only available to the members of the UNIFAC consortium. Since the consortium matrices are fuller (have more binary interaction parameters) the consortium values were used throughout this work (unless otherwise stated). Predictions for methylparaben are quite good while the results for 1,2-dihydro-acenaphthylene are poor. As mentioned above, the great power in the group contribution type methods is that the results are purely predictive and do not require any raw data once the group interaction values have been determined (fusion data are obviously still required) but still provide reasonable predictions.

Table 3-2 The experimental and predicted solubilities for methylparaben and 1,2-dihydro-acenaphthylene in various solvents at 298 K using UNIFAC and mod. UNIFAC (experimental data from Beilstein [59]).

Methylparaben			
	Solubility (g/l)		
Solvent	**Exp.**	**UNIFAC**	**mod. UNIFAC**
1-Propanol	259.46	333.02	-
2-Ethoxyethanol	352.48	-	-
Ethyl acetate	203.61	367.32	-
Hexane	0.07	0.45	-
Methanol	369.83	445.39	-
Tetrahydrofuran	453.10	227.36	-
Water	2.12	23.53	-

1,2-Dihydro-acenaphthylene			
	Solubility (g/l)		
Solvent	**Exp.**	**UNIFAC**	**mod. UNIFAC**
1,4-Dioxane	21.82	36.18	382.32
1-Propanol	2.60	37.75	60.78
2-Butanone	20.16	264.70	281.56
Acetic acid butyl ester	21.13	215.78	198.90
Acetonitrile	44.42	113.69	213.76
Dibutyl ether	17.21	129.22	135.18
Ethyl acetate	16.75	268.64	167.36
Hexane	8.01	118.44	78.84
N-Methylformamide	38.58	-	-
Tetrahydrofuran	30.43	271.02	546.22
Water	0.004	0.001	0.004

For the prediction of methylparaben there are only UNIFAC results and this is because mod. UNIFAC doesn't have all the binary interaction parameters available. This is a typical problem when trying to predict solubilities. The reason why many more "exotic" interactions are not present is that there is simply no VLE data with the appropriate groups. [Possible "fixes" for this problem are discussed in Appendix A]. Fig. 3-3 shows the state of the UNIFAC matrix as of September 2008, the common binary interactions (left upper part of the triangle) are mostly available while the others are fairly sparse.

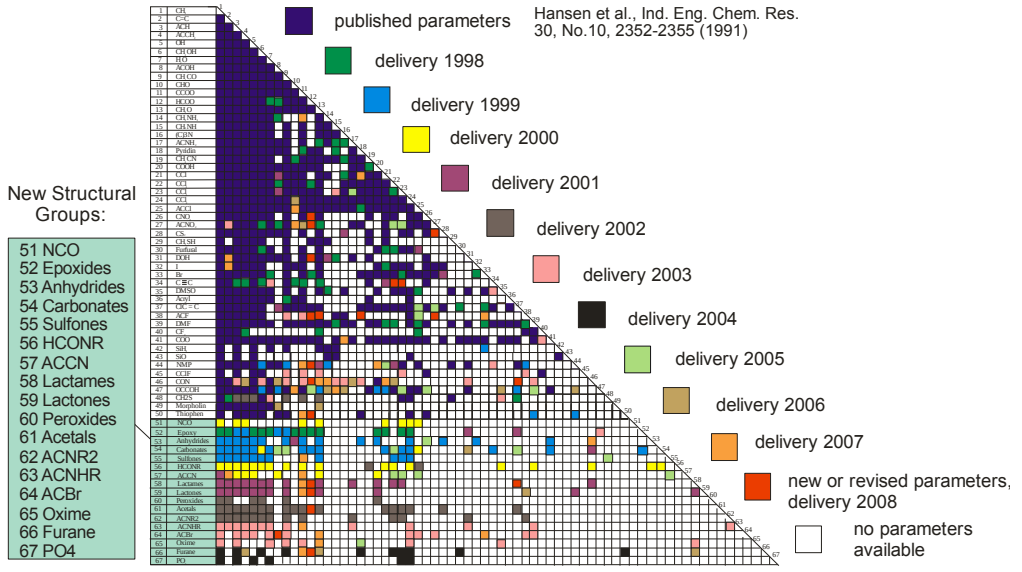


Fig. 3-3 The state of the UNIFAC matrix as of September 2008 available for members of the UNIFAC consortium, figure obtained from DDBST [28].

3.2.3. Segment activity coefficients

The following sections outline some methods which assign segments (e.g. polar, dispersive etc.) to a molecule and are all considered to be segment activity coefficient models.

3.2.3.1. Hansen solubility parameters

Thermodynamically, two substances will be mutually soluble when the change in Gibbs energy upon mixing is negative. The change in Gibbs energy upon mixing ($\Delta_m G$) is defined in terms of the enthalpy ($\Delta_m H$) and entropy ($\Delta_m S$) change upon mixing as follows:

$$\Delta_m G = \Delta_m H - T\Delta_m S \tag{3-65}$$

Since in the mixture there will usually be a greater degree of randomness in the molecule distribution, $\Delta_m S$ will usually be positive. Therefore if we consider Eqn. (3-65), substances will be mutually soluble if $\Delta_m H < T\Delta_m S$. Hildebrand and Scatchard [33, 60] proposed the Scatchard-Hildebrand equation for the change in enthalpy upon mixing (Eqn. (3-66)). This equation was developed by using regular solution theory, where a regular solution has a zero change of volume and entropy upon mixing but unlike an ideal solution it has a positive enthalpy of mixing.

$$\Delta_m H = \left(x_1 V_1 + x_2 V_2\right)\phi_1\phi_2\left(\delta_1 - \delta_2\right)^2 \tag{3-66}$$

$$\phi_i = \frac{x_i V_i}{\sum_j x_j V_j} \tag{3-67}$$

where V is the liquid molar volume, x is the mole fraction, ϕ is the volume fraction and δ is the Hildebrand solubility parameter which is defined as the square root of the cohesive pressure (see section 2.2.2):

$$\delta = \sqrt{c} = \left[\frac{\Delta_v H - RT}{V}\right]^{0.5} \tag{3-68}$$

It is therefore clear from Eqn. (3-66) that the more similar the Hildebrand solubility parameters are (i.e. minimizing $\Delta_m H$) the more the substances will be mutually soluble. The Hildebrand solubility parameter is only suitable for apolar or slightly polar compounds (i.e. approaching regular solution behaviour).

One of the reasons why the Hildebrand solubility parameter doesn't work for polar and hydrogen bonding compounds is that all the molecular interactions are simply "dumped" into one parameter and no differentiation is made between them. This problem was noted by many authors and many multi-parameter solubility parameters were developed [61]. The

approach which has found the widest use is the one developed by Hansen [62] where he split the cohesive pressure (and therefore the solubility parameter) into a dispersive (δ_d), polar (δ_p) and hydrogen bonding (δ_h) part, given mathematically as:

$$\delta_t^2 = \delta_d^2 + \delta_p^2 + \delta_h^2 \tag{3-69}$$

where δ_t should equal the Hildebrand solubility parameter but may differ slightly in some cases. The three parameters can be calculated in various means; these are outlined by Barton [61] and Hansen [62]. The two methods for the calculation of the parameters which have found the widest use are those of Hansen and Hoy.

If the parameters ($\delta_d, \delta_p, \delta_h$) are considered as three mutually orthogonal axes then each compound can be represented by a point in space. It is then the distance between these points that will indicate whether compounds will be mutually soluble or not. The distance equation is similar to the normal equation for the distance between vectors but the dispersive axis is doubled to give:

$$R_{12} = \left[4\left(\delta_{d,1} - \delta_{d,2}\right)^2 + \left(\delta_{p,1} - \delta_{p,2}\right)^2 + \left(\delta_{h,1} - \delta_{h,2}\right)^2 \right]^{0.5} \tag{3-70}$$

Every solute (1) is experimentally assigned a sphere of solubility with radius R_1 and therefore if $R_{12} < R_1$ the compounds will be mutually completely soluble. Consider the example of sucrose shown in Table 3-3 and Fig. 3-4, the sphere of solubility only encompasses water and therefore we would expect water to be a good solvent and the others to be poor (which can be verified in the literature [63]).

Table 3-3 Hansen solubility parameters for sucrose and some solvents, data from Barton [61].

Solute $(J/cm^3)^{0.5}$				
Name	δ_d	δ_p	δ_h	R_j
Sucrose	21.7	26.3	29.6	20.4

Solvent $(J/cm^3)^{0.5}$				
Name	δ_d	δ_p	δ_h	R_{ij}
Water	15.6	16.0	42.3	17.5
Benzene	18.4	0.0	2.0	38.3
Acetone	15.5	10.4	7.0	28.3
Ethanol	15.8	8.8	19.4	21.1
Cyclohexanol	17.4	4.1	13.5	27.8

The Hansen model is not only useful for a qualitative estimation of solubility but it can also be used to provide a quantitative one. The change in the enthalpy upon mixing in terms of the Hansen solubility parameters is analogous to Eqn. (3-66) and is given by:

$$\Delta_m H = \left(x_1 V_1 + x_2 V_2\right)\phi_1\phi_2 \left[\left(\delta_{d,1} - \delta_{d,2}\right)^2 + b_1\left(\delta_{p,1} - \delta_{p,2}\right)^2 + b_1\left(\delta_{h,1} - \delta_{h,2}\right)^2 \right] \tag{3-71}$$

Since it is not desirable to only be able to calculate properties of regular solutions we can assume that the entropy change upon mixing is given by the Flory-Huggins term:

$$\Delta_m S = -R\sum_{i=1}^{m} x_i \ln \phi_i \tag{3-72}$$

The excess change in Gibbs energy upon mixing is therefore given by:

$$G^E = \Delta_m H - T\Delta_m S - \Delta_m G^{id} \tag{3-73}$$

where

$$\Delta_m G^{id} = RT\sum_{i=1}^{m} x_i \ln x_i \tag{3-74}$$

Substituting in all the terms on the right hand side of Eqn. (3-73) and taking the partial derivative with respect to number of moles of component 2 the following results:

$$\ln \gamma_2 = \frac{V_2}{RT}\phi_1^2\left[\left(\delta_{d,2} - \delta_{d,1}\right)^2 + b_2\left(\delta_{p,2} - \delta_{p,1}\right)^2 + b_2\left(\delta_{h,2} - \delta_{h,1}\right)^2\right] + \ln\frac{\phi_2}{x_2} + 1 - \frac{\phi_2}{x_2} \tag{3-75}$$

This is frequently written in a simplified form as follows:

$$\ln \gamma_2 = \frac{V_2}{RT}\left[\left(\delta_{d,2} - \bar{\delta}_d\right)^2 + b_2\left(\delta_{p,2} - \bar{\delta}_p\right)^2 + b_2\left(\delta_{h,2} - \bar{\delta}_h\right)^2\right] + \ln\frac{\phi_2}{x_2} + 1 - \frac{\phi_2}{x_2} \tag{3-76}$$

where $\bar{\delta}_z$ is the "effective" solubility parameter for either the dispersive, polar or hydrogen bonding part. Eqn. (3-76) given in general form for *m* components is:

$$\ln \gamma_i = \frac{V_i}{RT}\left[\left(\delta_{d,i} - \bar{\delta}_d\right)^2 + b_i\left(\delta_{p,i} - \bar{\delta}_p\right)^2 + b_i\left(\delta_{h,i} - \bar{\delta}_h\right)^2\right] + \ln\frac{\phi_i}{x_i} + 1 - \frac{\phi_i}{x_i} \tag{3-77}$$

$$\bar{\delta}_{z,i} = \sum_{i}^{m} \phi_i \delta_{z,i} \tag{3-78}$$

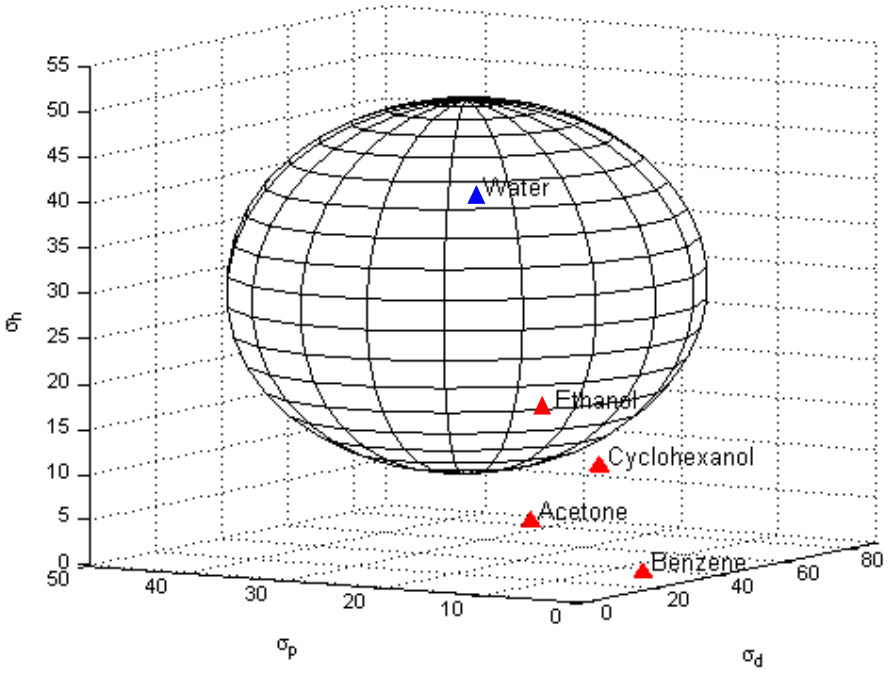


Fig. 3-4 Hansen solubility sphere for sucrose (▲ – position of solvents relative to the solubility sphere, all solubility parameters taken from Barton [61] and all have units of $(J/cm^3)^{0.5}$. Red symbolizes a poor solvent and blue a good one).

When modelling the test solubility data there were two different methods of approach. Firstly if it is assumed that no solubility data are known, then the solubility parameters can be predicted from group contribution methods (e.g. Beerbower's [62]). On the other hand if it is assumed that solubility data is known in three or four solvents then the parameters can be fitted to the data. When the model parameter in this and the following sections were fitted to the data the following objective function was used (where subscript *2* refers to the solute):

$$O.F. = \left(\ln \gamma_2^{\exp} - \ln \gamma_2^{model}\right)^2 \tag{3-79}$$

When fitting the model parameters to the data it is important that the solvents are sufficiently different so that there is sufficient data for the segments used. For both the polar and non-polar test data the same 4 solvents were used as the reference compounds for solvents

Table 3-4 shows both the fitted and the predicted results for the test data solubility in mixed solvents. The predicted results are fairly poor in the case of methylparaben they almost always under-predicted the solubility. For 1,2-dihydro-acenaphthylene the predictions are very poor, in most cases giving errors about an order of magnitude off. For the fitted approach the four solvents that were used were: ethyl acetate, hexane, tetrahydrofuran and water. The fitted results are in most cases very good, and in most cases reproducing the experimental data very well. In some cases there are two values for a single compound. This is because when large deviations were found the parameters of Hoy were tried. Remarkably, with no

further fitting the results were improved in both cases. This suggests that there is some optimum set of parameters which are a mix of both the Hansen and the Hoy parameters. This would however require further investigation.

Table 3-4 The experimental, fitted and predicted solubilities for methylparaben and 1,2-dihydro-acenaphthylene in various solvents at 298 K using Hansen's parameters [61, 62] (* - data fitted to all solvents, ** - data fitted to 4 reference solvents, * - the 4 reference solvents, [a] – the Hoy solubility parameters used, experimental data from Beilstein [59]).**

Methylparaben								
Solvent	**V (cm^3/mol)**	**δ_d**	**δ_p**	**δ_h**	**S_{exp} (g/l)**	**$^{***}S_{fit}$ (g/l)[I]**	**$^{**}S_{fit}$ (g/l)[II]**	**S_{pred} (g/l)[III]**
1-Propanol	75.1	16.0	7.8	17.4	259.46	241.31	239.55	106.84
2-Ethoxyethanol	96.9	16.2	9.2	14.3	352.48	172.63	171.48	78.93
Ethyl acetate*	98.5	15.8	5.3	7.2	203.61	31.53	30.86	30.52
Hexane*	131.6	14.9	0	0	0.07	0.37	0.35	0.65
Methanol	40.7	15.1	12.3	22.3	369.83	526.35	526.73	115.76
Tetrahydrofuran*	81.6	16.8	5.7	8	453.10	58.71	57.75	84.46
Water*	18.1	15.5	16	42.3	2.12	1.50	1.55	0.11
1,2-Dihydro-acenaphthylene								
Solvent	**V (cm^3/mol)**	**δ_d**	**δ_p**	**δ_h**	**S_{exp} (g/l)**	**$^{***}S_{fit}$ (g/l)[IV]**	**$^{**}S_{fit}$ (g/l)[V]**	**S_{pred} (g/l)[VI]**
1,4-Dioxane	85.7	19.0	1.8	7.4	21.82	0.73	1.58	466.78
1,4-Dioxane[a]	85.7	16.3	10.1	7.9	21.82	41.99	82.19	44.56
1-Propanol	75.1	16	7.8	17.4	2.60	3.95	5.38	1.77
2-Butanone	90.2	16	9	5.1	20.16	58.07	104.33	86.66
Acetic acid butyl ester	132.5	14.5	8	13.5	21.13	12.01	12.98	1.62
Acetonitrile	52.9	15.3	18	6.1	44.42	727.45	959.96	3.16
Acetonitrile[a]	52.9	10.3	11.1	19.6	44.42	35.59	15.40	0.00
Dibutyl ether	170.4	15.2	3.4	4.2	17.21	11.62	13.20	52.01
Ethyl acetate*	98.5	15.8	5.3	7.2	16.75	19.34	26.72	98.26
Hexane*	131.6	14.9	0	0	8.01	7.40	6.92	67.41
N-Methylformamide	59.1	17.4	18.8	15.9	38.58	9.59	27.67	0.14
Tetrahydrofuran*	81.6	16.8	5.7	8	30.43	13.36	22.99	224.77
Water*	18.1	15.5	16	42.3	0.004	0.005	0.004	0.00

3.2.3.2. MOSCED

As with the Hansen model, MOSCED (Modified Separation of Cohesive Energy Density Model) is based on regular solution theory. The main difference is in the way that the cohesive density is treated in both. MOSCED is only applicable at infinite dilution. Combining Eqn. (3-73), (3-72), (3-66), (3-74), (3-68) and (3-25) at infinite dilution of component 2 yields:

$$\ln \gamma_2^{\infty} = \frac{V_2}{RT}\left[c_1 + c_2 - 2\sqrt{c_1 c_2}\right] + \ln\left(\frac{V_2}{V_1}\right) + 1 - \left(\frac{V_2}{V_1}\right) \tag{3-80}$$

[I] (δ_d, δ_p, δ_h, b_2) = (17.09, 10.34, 16.94, 0.318) [=] $(J/cm^3)^{0.5}$ @ 298.15 K – b_2 is dimensionless

[II] (δ_d, δ_p, δ_h, b_2) = (17.124, 10.34, 16.95, 0.318) [=] $(J/cm^3)^{0.5}$ @ 298.15 K – b_2 is dimensionless

[III] Beerbower's method was used: (δ_d, δ_p, δ_h, b_2) = (20.31, 6.83, 14.24, 0.25) [=] $(J/cm^3)^{0.5}$ @ 298.15 K – b_2 is dimensionless and was set at 0.25, V_m was calculated as 143.4 cm^3/mol

[IV] (δ_d, δ_p, δ_h, b_2) = (11.06, 17.07, 0.145, 0.156) [=] $(J/cm^3)^{0.5}$ @ 298.15 K – b_2 is dimensionless

[V] (δ_d, δ_p, δ_h, b_2) = (12.66, 18.57, 0.00, 0.163) [=] $(J/cm^3)^{0.5}$ @ 298.15 K – b_2 is dimensionless

[VI] Beerbower's method was used: (δ_d, δ_p, δ_h, b_2) = (19.71, 1.55, 1.55, 0.25) [=] $(J/cm^3)^{0.5}$ @ 298.15 K – b_2 is dimensionless and was set at 0.25, V_m was calculated as 175 cm^3/mol

The root term (geometric mean of c_1 and c_2) arises out of some approximations [60] made in Eqn. (3-66) and should rather be replaced with a more general term as follows:

$$\ln\gamma_2^{\infty} = \frac{V_2}{RT}\left[c_1 + c_2 - 2c_{12}\right] + \ln\left(\frac{V_2}{V_1}\right) + 1 - \left(\frac{V_2}{V_1}\right) \tag{3-81}$$

As stated above the main difference with the Hansen model is the way in which the cohesive energy density is treated. As with Hansen's model it is assumed that the cohesive energy density is equivalent to the sum of the contribution of each force, however unlike with Hansen's model, 5 parameters and not 3 are used to describe these interactions. For the binary case the 2 pure component and mixture values are given as:

$$c_1 = \lambda_1^2 + \tau_1^2 + \sigma_1\tau_1 + \alpha_1\beta_1 \tag{3-82}$$

$$c_2 = \lambda_2^2 + \tau_2^2 + \sigma_2\tau_2 + \alpha_2\beta_2 \tag{3-83}$$

$$c_{12} = \lambda_1\lambda_2 + \tau_1\tau_2 + \frac{1}{2}\left(\sigma_1\tau_2 + \sigma_2\tau_1 + \alpha_1\beta_2 + \alpha_2\beta_1\right) \tag{3-84}$$

Where the parameters λ, q, τ, α and β (all have units of $(J/cm^3)^{0.5}$) have the following definitions (any discussion of the estimation of these parameters is based on the molecules of interest in this work. For simple molecules the predictions are fairly good):

- The dispersion parameter (λ) is analogous to the dispersion solubility parameter (δ_d) used by Hansen. Attempts have been made to develop correlations which are generally applicable to all species [64]. These generally fail for multifunctional molecules [65] and therefore λ is frequently treated as an adjustable parameter. The exception to this is the alkanes which use the value of δ_d for λ.

- The induction parameter (q) represents the interaction between the polar and the non-polar part of a molecule. A value of 0.9 is assigned for aromatic compounds, 1 for most saturated aliphatic compounds and the following correlation for unsaturated aliphatics (where $n_{C=C}$ is number of double bonds and n_C is number of carbon atoms):

$$q = 1.0 - \left(\frac{n_{C=C}}{n_C}\right) \tag{3-85}$$

- The polar parameter (τ) is a measure of the fixed dipole of a molecule in solution. There are some correlations to predict this value but these are usually very poor and therefore it is generally treated as an adjustable parameter.

- The acidity (α) and basicity (β) parameters account for the effects of hydrogen bond formation. There is no satisfactory way of correlating these parameters and therefore they are treated as adjustable parameters.

Substituting these equations back into Eqn. (3-81) results in what is generally termed a multicomponent solubility parameter equation [64]:

$$\ln\gamma_2^{\infty} = \frac{V_2}{RT}\left[\left(\lambda_1-\lambda_2\right)^2+\left(\tau_1-\tau_2\right)^2+\left(\sigma_1-\sigma_2\right)\left(\tau_1-\tau_2\right)+\left(\alpha_1-\alpha_2\right)\left(\beta_1-\beta_2\right)\right]+\ln\left(\frac{V_2}{V_1}\right)+1-\left(\frac{V_2}{V_1}\right) \quad (3\text{-}86)$$

In a simplified form (with slight modification [64]) Eqn. (3-86) is:

$$\ln\gamma_2^{\infty} = \frac{V_2}{RT}\left[\left(\lambda_1-\lambda_2\right)^2+q_1^2q_2^2\left(\tau_1-\tau_2\right)^2+\left(\alpha_1-\alpha_2\right)\left(\beta_1-\beta_2\right)\right]+d_{12} \quad (3\text{-}87)$$

$$d_{12} = \ln\left(\frac{V_2}{V_1}\right)+1-\left(\frac{V_2}{V_1}\right) \quad (3\text{-}88)$$

It was observed that Eqn. (3-87) was insufficient for non-symmetrical systems (which result from polar and hydrogen bond interactions, for example ethanol-hexane). Therefore it was empirically modified as follows:

$$\ln\gamma_2^{\infty} = \frac{V_2}{RT}\left[\left(\lambda_1-\lambda_2\right)^2+\frac{q_1^2q_2^2\left(\tau_1-\tau_2\right)^2}{\psi_1}+\frac{\left(\alpha_1-\alpha_2\right)\left(\beta_1-\beta_2\right)}{\xi_1}\right]+d_{12} \quad (3\text{-}89)$$

where ψ and ξ are considered to be functions of the model parameters and are correlated as follows:

$$\psi = POL+0.002629\alpha_1\beta_1 \quad (3\text{-}90)$$

$$\xi = 0.68\left(POL-1\right)+\left[3.24-2.4\exp\left(-0.002687\left(\alpha_1\beta_1\right)^{1.5}\right)\right]^{\left(293/T\right)^2} \quad (3\text{-}91)$$

$$POL = q_1^4\left(1.15-1.15\exp\left(-0.002337\tau_1^3\right)\right)+1 \quad (3\text{-}92)$$

The Flory-Huggins term d_{12} was also modified:

$$d_{12} = \ln\left(\frac{V_2}{V_1}\right)^{aa}+1-\left(\frac{V_2}{V_1}\right)^{aa} \quad (3\text{-}93)$$

$$aa = 0.953-0.002314\left(\tau_2^2+\alpha_2\beta_2\right) \quad (3\text{-}94)$$

The temperature dependence for the model parameters α, β and τ is given as follows:

$$\alpha_T = \alpha_{293}\left(\frac{293}{T}\right)^{0.8} \tag{3-95}$$

$$\beta_T = \beta_{293}\left(\frac{293}{T}\right)^{0.8} \tag{3-96}$$

$$\tau_T = \tau_{293}\left(\frac{293}{T}\right)^{0.4} \tag{3-97}$$

where all the model parameters are reported at 293 K. Table 3-5 shows the results for the test data. The fitted results reproduce the experimental data fairly well; overall the error between the model and the experimental data is less than in the case of the Hansen method even though there is no concentration dependence in MOSCED. When examining the molar volumes in Table 3-5 it is clear that the value for water is incorrect (should be $18 cm^3.mol^{-1}$). This is not an error, it was found that the activity coefficient was greatly under predicted when using the correct molar volume [65] and therefore the water molar volume was used as an adjustable parameter. The explanation given for this larger volume is that because of the extensive hydrogen bonding present in water the molecules could behave like a larger molecule. When the same correction was applied to Hansen's model the fit was worse in all cases.

Table 3-5 The experimental and fitted solubilities for methylparaben and 1,2-dihydro-acenaphthylene in various solvents at 298 K using MOSCED [65] (* - data fitted to all solvents, ** - data fitted to 4 reference solvents, * - the 4 reference solvents, experimental data from Beilstein [59]).**

Methylparaben									
Solvent	**V_1 (cm^3/mol)**	**λ**	**τ**	**q**	**α**	**β**	**S_{exp} (g/l)**	*****S_{fit} (g/l)[I]**	****S_{fit} (g/l)[II]**
1-Propanol	75.1	14.93	1.39	1	11.97	10.35	259.46	108.51	78.37
2-Ethoxyethanol	97.3	15.12	7.39	1	3.77	16.84	352.48	528.42	311.36
Ethyl acetate *	98.6	14.51	5.74	1	0	7.25	203.61	392.98	211.70
Hexane*	131.4	14.90	0	1	0	0	0.07	0.12	0.11
Methanol	40.6	14.43	3.77	1	17.43	14.49	369.83	254.09	207.21
Tetrahydrofuran*	81.9	15.78	4.41	1	0	6.61	453.1	198.49	110.04
Water*	36.0	10.58	10.48	1	52.78	15.86	2.12	9.88	11.49
1,2-Dihydro-acenaphthylene									
Solvent	**V_1 (cm^3/mol)**	**λ**	**τ**	**q**	**α**	**β**	**S_{exp} (g/l)**	*****S_{fit} (g/l)[III]**	****S_{fit} (g/l)[IV]**
1,4-Dioxane	85.7	16.96	6.72	1	0	10.39	21.82	13.81	1.05
1-Propanol	75.1	14.93	1.39	1	11.97	10.35	2.60	3.18	1.82
2-Butanone	90.2	14.74	6.64	1	0	9.7	20.16	76.61	10.32
Acetic acid butyl ester	132.0	15.22	4.16	1	0	6.4	21.13	29.97	6.54
Acetonitrile	52.9	13.78	11.51	1	3.49	8.98	44.42	67.27	8.87
Dibutyl ether	170.4	15.13	1.73	1	0	5.29	17.21	11.11	6.79

[I] (λ, τ, q, α, β) = (8.416, 0, 0.9, 24.10, 3.92) [=] $(J/cm^3)^{0.5}$ @ 298.15 K

[II] (λ, τ, q, α, β) = (8.139, 0, 0.9, 22.99, 4.17) [=] $(J/cm^3)^{0.5}$ @ 298.15 K

[III] (λ, τ, q, α, β) = (9.412, 6.177, 0.9, 0, 0) [=] $(J/cm^3)^{0.5}$ @ 298.15 K

[IV] (λ, τ, q, α, β) = (7.431, 2.276, 0.9, 0, 0) [=] $(J/cm^3)^{0.5}$ @ 298.15 K

Ethyl acetate*	98.6	14.51	5.74	1	0	7.25	16.75	80.32	15.23
Hexane*	131.4	14.90	0	1	0	0	8.01	6.69	9.18
N-Methylformamide	59.1	15.55	8.92	1	8.07	22.01	38.58	3.77	0.61
Tetrahydrofuran*	81.9	15.78	4.41	1	0	6.61	30.43	35.73	8.87
Water*	36.0	10.58	10.48	1	52.78	15.86	0.004	0.006	0.009

3.2.3.3. NRTL-SAC

Just as UNIFAC and ASOG use the concept of the solution of groups, NRTL-SAC [66, 67] (NRTL – Segmented Activity Coefficient) uses the principle of the "solution of surfaces". Instead of considering molecules as been broken down into functional groups, the molecules are broken down into surface segments. The four surfaces that are used are hydrophobic (X), hydrophilic (Z), polar positive (Y+) and polar negative(Y-). So for example hexane would only have a hydrophobic surface, whereas acetic acid would have a bit of all 4 surfaces.

As with UNIFAC and ASOG it is assumed that the activity coefficient consists of a combinatorial and residual part. The combinatorial expression used is analogous to the Flory-Huggins expression used in ASOG.

$$\ln \gamma_I^C = \ln \frac{\phi_I}{x_I} + 1 - r_I \sum_J \frac{\phi_J}{r_J} \tag{3-98}$$

$$r_I = \sum_i r_{i,I} \tag{3-99}$$

$$\phi_I = \frac{r_I x_I}{\sum_J r_J x_J} \tag{3-100}$$

where $r_{i,I}$ is the number (analogous to group frequency but a real number rather than a whole number) of segments *i* in molecule *I*. The lower case indices (*i,j,k,m* and *m'*) run from 1 to 4 where 1 to 4 are assigned segments (usually 1 = X; 2 = Y-; 3 = Y+ and 4 = Z) and the uppercase indices (*I* and *J*) are for the number of components (for examples of the types of values found for the segment numbers, see Table 3-7). The expression for the residual contribution to the activity coefficient is analogous to those of UNIFAC and ASOG:

$$\ln \gamma_I^R = \sum_m r_{m,I} \left[\ln \Gamma_m - \ln \Gamma_m^I \right] \tag{3-101}$$

where Γ_m is the activity coefficient of segment *m* in the mixture and Γ_m^I is the activity coefficient of segment *m* in the pure component *I*. As the name suggests, NRTL is used to compute the segment activity coefficients:

$$\ln \Gamma_m = \frac{\sum_j x_j G_{jm} \tau_{jm}}{\sum_k x_k G_{km}} + \sum_{m'} \frac{x_{m'} G_{mm'}}{\sum_k x_k G_{km'}} \left(\tau_{mm'} - \frac{\sum_j x_j G_{jm'} \tau_{jm'}}{\sum_k x_k G_{km'}} \right) \tag{3-102}$$

$$x_j = \frac{\sum_J x_J r_{j,J}}{\sum_I \sum_i x_I r_{i,I}} \tag{3-103}$$

The segment activity coefficient in the pure component (Γ_m^I) is calculated from the same equation except $x_{j,I}$ is used in place of x_j, where:

$$x_{j,I} = \frac{r_{j,I}}{\sum_i r_{i,I}} \tag{3-104}$$

The NRTL-SAC segment interaction parameters were obtained by considering large amounts of VLE and LLE data [67]. The NRTL-SAC interaction parameters are shown in Table 3-6. The values for α_{ij} were set by considering the typical values that are found in systems containing the appropriate surfaces.

Table 3-6 Segment interaction parameters for NRTL-SAC taken from Chen et al. [67].

	τ_{ij}				α_{ij}			
i \ j	**X**	**Y-**	**Y+**	**Z**	**X**	**Y-**	**Y+**	**Z**
X	0	1.643	1.643	6.547	0	0.2	0.2	0.2
Y-	1.834	0	0	-2	0.2	0	0	0.3
Y+	1.834	0	0	2	0.2	0	0	0.3
Z	10.949	1.787	1.787	0	0.2	0.3	0.3	0

Table 3-7 shows the results for the test data. For methylparaben NRTL-SAC reproduces the data excellently when fitted to all the solvents, but when fitted to the 4 reference solvents the predictions for propanol and methanol are quite far off. For 1,2-dihydro-acenaphthylene the results are poor for both fits. This is more than likely due to water being a reference solvent since 1,2-dihydro-acenaphthylene is very hydrophobic this would cause the X parameter to be too high and therefore reduce the accuracy of the other fits.

The rationale behind using these surface segment equations to fit to some reference solvents to predict the general behaviour is that during the manufacture of drugs some solubility measurements are undertaken anyway. Thus if a large set of data can be predicted from a couple of measurements this greatly reduces the labour intensiveness of the synthesis. Nevertheless this does mean that in the absence of experimental data no pure predictions can be undertaken (in the case of NRTL-SAC and MOSCED; there are predictive methods available for Hansen parameters as shown above but these are frequently very poor).

Table 3-7 The experimental and fitted solubilities for methylparaben and 1,2-dihydro-acenaphthylene in various solvents at 298 K using NRTL-SAC [67] (* - the 4 reference solvents, experimental data from Beilstein [59], * - data fitted to all solvents, ** - data fitted to 4 reference solvents).**

Methylparaben							
Solvent	**r_X**	**r_{Y-}**	**r_{Y+}**	**r_Z**	**S_{exp} (g/l)**	*****S_{fit} (g/l)[I]**	****S_{fit} (g/l)[II]**
1-Propanol	0.374	0.013	0	0.53	259.46	232.78	559.12
2-Ethoxyethanol	0.179	0.121	0.106	0.323	352.48	344.12	418.10
Ethyl acetate*	0.339	0.058	0.441	0	203.61	234.62	205.19
Hexane*	1	0	0	0	0.07	0.07	0.07
Methanol	0.09	0.139	0	0.594	369.83	473.52	848.84
Tetrahydrofuran*	0.235	0.04	0.32	0	453.1	445.19	439.74
Water*	0	0	0	1	2.12	2.13	2.12
1,2-Dihydro-acenaphthylene							
Solvent	**r_X**	**r_{Y-}**	**r_{Y+}**	**r_Z**	**S_{exp} (g/l)**	*****S_{fit} (g/l)[III]**	****S_{fit} (g/l)[IV]**
1,4-Dioxane	0.154	0.086	0.401	0	21.82	443.24	472.31
1-Propanol	0.374	0.013	0	0.53	2.60	676.24	712.65
2-Butanone	0.261	0.095	0.463	0	20.16	333.92	349.24
Acetic acid butyl ester	0.317	0.03	0.33	0	21.13	296.70	305.08
Acetonitrile	0.018	0.131	0.833	0	44.42	87.00	88.50
Dibutyl ether	-	-	-	-	17.21	-	-
Ethyl acetate*	0.339	0.058	0.441	0	16.75	284.14	287.30
Hexane*	1	0	0	0	8.01	16.62	8.70
N-Methylformamide	-	-	-	-	38.58	-	-
Tetrahydrofuran*	0.235	0.04	0.32	0	30.43	562.17	584.16
Water*	0	0	0	1	0.004	0.007	0.005

3.2.3.4. COSMO-RS and COSMO-SAC

The huge potential of quantum chemical methods such as COSMO-RS [12] (COnductor like Screening MOdels for Real Solvents) and COSMO-SAC [13, 68] (COSMO – Segmented Activity Coefficient) is that they could supply *ab initio* predictions of activity coefficients (and other properties) eliminating any need to undertake measurements. Both methods are based on the so called sigma profiles that are obtained from quantum chemical calculations. These profiles are generated by constructing a molecule shaped cavity into a solvent continuum by assuming a certain van der Waals radius for the atoms in the molecule. A large number of charges are then positioned on the surface of the cavity and their charges are optimized during the quantum chemical calculations in order to minimise the total energy. The shielding charge density is denoted by σ, and the frequency of each shielding charge density of component *i* is given by $p_i(\sigma)$:

$$p_i(\sigma) = \frac{A_i(\sigma)}{A_i} \tag{3-105}$$

[I] (X, Y^-, Y^+, Z) = (0.406, 0.507, 1.518, 0.524)

[II] (X, Y^-, Y^+, Z) = (0.780, 1.096, 0.209, 0.779)

[III] (X, Y^-, Y^+, Z) = (1.183, 0.091, 0, 0.665)

[IV] (X, Y^-, Y^+, Z) = (1.229, 0.070, 0, 0.786)

where $A_i(\sigma)$ is the surface area of all segments with the charge σ and A_i is the total surface area of the molecule. Examples of some sigma profiles are shown in Fig. 3-5, a fair amount of qualitative data can be gleaned by examining the sigma profiles. For example hexane is clearly non-polar as all its segments lie in the low charge density range, while acetone is clearly a polar molecule since it has a much wider profile.

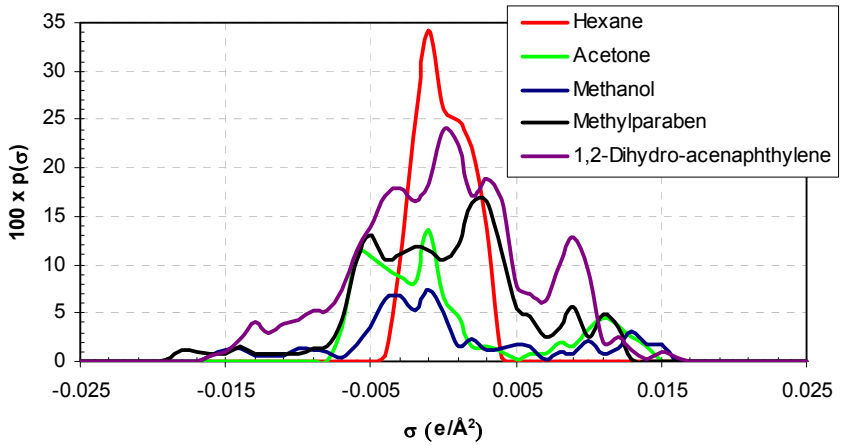


Fig. 3-5 Sigma profiles for 3 solvents and the example solutes (profiles from the VT database [69]).

Where COSMO-RS and COSMO-SAC differ is in the calculation of the segment chemical potential. Lin and Sandler [13] report certain anomalies in the method of COSMO-RS which are not present in the method of COSMO-SAC. This is however disputed by Klamt [70] and there still seems to be much disagreement on the matter [71]. Both methods use the Staverman-Guggenheim (denoted by superscript SG) combinatorial term (see section 3.2.4) in the calculation of the activity coefficient. The COSMO-SAC expression is:

$$\ln\gamma_{i/S} = n_i \sum_{\sigma_m} p_i(\sigma_m)\left[\ln\Gamma_S(\sigma_m) - \ln\Gamma_i(\sigma_m)\right] + \ln\gamma_{i/S}^{SG} \qquad (3\text{-}106)$$

where $\ln\Gamma_S(\sigma_m)$ is the segment activity coefficient of segment σ_m in the mixture and $\ln\Gamma_i(\sigma_m)$ is the activity coefficient of segment σ_m in pure component *i* (the exact details of the calculations are not in the scope of this work). The expression for the COSMO-RS activity coefficient is:

$$\ln\gamma_{i/S} = n_i \sum_{\sigma_m} p_i(\sigma_m)\exp\left[\frac{\mu'_S(\sigma_m) - \mu'_i(\sigma_m)}{RT}\right] - \lambda\ln\left(\frac{A_S}{A_i}\right) + \ln\gamma_{i/S}^{SG} \qquad (3\text{-}107)$$

where $\mu'_S(\sigma_m)$ is the chemical potential of segment σ_m in the mixture, $\mu'_i(\sigma_m)$ is the chemical potential of segment σ_m in pure component *I*, A_S is the mole fraction weighted surface area of all the species in the mixture, similarly A_i is for pure component *i* and λ is a solvent specific adjustable parameter. While it can be shown that the terms in the square brackets in Eqn's (3-106) and (3-107) are equivalent the calculation procedures differ. This is the fundamental

difference between COSMO-SAC and COSMO-RS, but as mentioned above, the exact details of these calculations are not within the scope of this work.

As mentioned above the huge advantage of using COSMO calculations for calculating phase equilibria is that results are obtained from first principles. Table 3-8 shows the results for the prediction of the test set data. The results for both methods are fairly poor. This is to be expected as out of all the methods analysed in this section these have the "seen" the least amount of experimental data. In almost every instance tried the COSMO-RS model is superior to that of COSMO-SAC. (Note: any references to COSMO-RS from this point onwards refer to COSMO-RS(OL) [72] which is a slightly modified version of the original COSMO-RS [12]).

The results for all the methods covered in this chapter unsurprisingly show that the approaches which use reference solvents are almost always superior to the purely predictive methods. Of the predictive methods it seems that UNIFAC or mod. UNIFAC should be considered as the more reliable methods

Table 3-8 Comparison of the experimental and predicted data for methylparaben and 1,2-dihydro-acenaphthylene in various solvents at 298 K using COSMO-SAC and COSMO-RS(OL) (experimental data from Beilstein [59], COSMO profiles from the DDB [28]).

	Methylparaben		
	Solubility (g/l)		
Solvent	**Exp.**	**COSMO-RS(OL)**	**COSMO-SAC**
1-Propanol	259.46	355.41	772.84
2-Ethoxyethanol	352.48	268.16	444.99
Ethyl acetate	203.61	341.60	660.26
Hexane	0.07	0.94	0.10
Methanol	369.83	706.83	1528.20
Tetrahydrofuran	453.10	198.52	1145.58
Water	2.12	43.37	114.56
	1,2-Dihydro-acenaphthylene		
	Solubility (g/l)		
Solvent	**Exp.**	**COSMO-RS(OL)**	**COSMO-SAC**
1,4-Dioxane	21.82	544.80	539.05
1-Propanol	2.60	238.27	241.10
2-Butanone	20.16	520.78	520.40
Acetic acid butyl ester	21.13	374.53	351.71
Acetonitrile	44.42	230.69	245.15
Dibutyl ether	17.21	275.49	234.45
Ethyl acetate	16.75	457.64	463.90
Hexane	8.01	226.93	181.57
N-Methylformamide	38.58	128.18	158.97
Tetrahydrofuran	30.43	687.11	720.60
Water	0.004	0.462	0.493

3.2.4. Combinatorial contributions

The Gibbs energy change upon mixing is given as (where G_m is the Gibbs energy of the mixture and G_i is the Gibbs energy of component *i*):

$$\Delta_m G = G_m - \sum_i x_i G_i \tag{3-108}$$

$$G_m = \sum_i \overline{G}_i \tag{3-109}$$

where $\overline{G}_i$ is the partial Gibbs energy of component *i*. For an ideal mixture the partial Gibbs energy and Gibbs energy of the mixture are given as (where the superscript *id* refers to the ideal mixture):

$$\overline{G}_i^{id} = G_i + RT \ln x_i \tag{3-110}$$

$$G_m^{id} = \sum_i x_i G_i + RT \sum_i x_i \ln x_i \tag{3-111}$$

Combining Eqn. (3-108) and (3-111) gives the ideal Gibbs energy change upon mixing:

$$\Delta_m G^{id} = RT \sum_i x_i \ln x_i \tag{3-112}$$

While Eqn. (3-112) is approximately true for similar molecules (e.g. benzene and toluene) of similar size and type there is a major difference however for systems with molecules of greatly different size and type. It was however observed that for mixtures of monomers and polymers (where the molecule types are the same but the sizes are vastly different) that when the volume fraction (ϕ_i) was used as a measure of composition then the mixture behaved almost ideally, therefore considering Raoult's law (assuming an ideal vapour phase) for such systems:

$$y_i P = \phi_i P_i^s \neq x_i P_i^s \tag{3-113}$$

Using this modification for the Gibbs energy of mixing results in the following expression (using the volume fraction in Eqn. (3-110) and combining with Eqn. (3-109)):

$$\Delta_m G = RT \sum_i x_i \ln \phi_i \tag{3-114}$$

Mixtures such as these (i.e. polymers and monomers) are known as athermal mixtures (i.e. $H^E = \Delta_m H = 0$), for athermal mixtures Eqn. (3-65) reduces to:

$$\left(\frac{\Delta_m G}{RT} \right)_C = -\frac{\Delta_m S}{R} \tag{3-115}$$

where the subscript *C* refers to the combinatorial contribution. One of the simplest expressions for the entropy change upon mixing was proposed by Flory and Huggins (see Eqn. (3-72)). Rearrangement of Eqn. (3-73) yields Eqn. (3-116) which can be combined with

the Flory-Huggins entropy expression to give Eqn. (3-117), where the van der Waals volume fraction (Eqn. (3-47)) is used instead of the molar volume fraction.

$$\frac{G^E}{RT} = \frac{\Delta_m G}{RT} - \sum_i x_i \ln x_i \tag{3-116}$$

$$\frac{G^E}{RT} = \sum_i x_i \ln \frac{\Phi_i}{x_i} \tag{3-117}$$

Taking the partial derivative of Eqn. (3-117) yields the following expression for the Flory-Huggins combinatorial contribution to the activity coefficient:

$$\ln \gamma_i^C = 1 - \frac{\Phi_i}{x_i} + \ln \frac{\Phi_i}{x_i} \tag{3-118}$$

In the case of non-athermal mixtures, an additional effect on the entropy of mixing has to be taken into account (the residual contribution). Consider mixing 2 components at infinite temperature where:

$$\lim_{T\to\infty} \Delta_m G = \lim_{T\to\infty} \left(\Delta_m H - T \Delta_m S \right) = -T \Delta_m S \tag{3-119}$$

This is a strictly athermal mixture since at infinite temperature any changes in enthalpy term become insignificant compared to the entropy term. When cooling this mixture to the system temperature of interest, ordering due to enthalpic preference will result in local compositions different from athermal mixing. This leads to a reduction of the entropy of mixing as enthalpic effects produce a more “ordered” mixture.

In the UNIQUAC equation, the combinatorial contribution was introduced as the boundary condition at $1/T = 0$ which is given by Guggenheim [42, 43, 50, 73] as:

$$\left.\frac{\Delta_m G}{RT}\right|_{1/T\to 0} = \frac{-\Delta_m S}{R} = \sum_i x_i \ln \Phi_i + \frac{1}{2} \sum \left(x_i z q_i \ln \frac{\theta_i}{\Phi_i} \right) \tag{3-120}$$

where z is the coordination number of the lattice (number of neighbouring atoms) also θ_i and Φ_i are defined by Eqn’s (3-48) and (3-47) respectively. Combining Eqn’s (3-115), (3-116), (3-120) and (3-25) yields the Guggenheim-Staverman combinatorial expression:

$$\ln \gamma_i^C = \left(1 - \frac{\Phi_i}{x_i} + \ln \frac{\Phi_i}{x_i} \right) - \frac{z q_i}{2} \left(1 - \frac{\Phi_i}{\theta_i} + \ln \frac{\Phi_i}{\theta_i} \right) \tag{3-121}$$

In many applications (such as UNIFAC and UNIQUAC) the value of z is set to 10. The Guggenheim-Staverman equation reduces to the Flory-Huggins equation when $r_i = q_i$ for all *i*. These two expressions were used among others by Kikic et al. [74] to predict the infinite

dilution activity coefficients for athermal systems (e.g. alkane-alkane systems – see section 6.1 for a different analysis proposed in this work). Their analysis showed that the Flory-Huggins expression actually performs better that the Guggenheim-Staverman (SG) equation in all the examples tried. They proposed the following empirical modification for the SG equation:

$$\ln \gamma_i^C = \ln \frac{\Phi_i'}{x_i} + 1 - \frac{\Phi_i'}{x_i} - \frac{1}{2} z q_i \left(\ln \frac{\Phi_i}{\theta_i} + 1 - \frac{\Phi_i}{\theta_i} \right) \tag{3-122}$$

where Φ_i' is the empirically modified volume fraction, for which two different expressions were proposed:

$$\Phi_i' = \frac{x_i r_i^{2/3}}{\sum_j x_j r_j^{2/3}} \tag{3-123}$$

$$\Phi_i' = \frac{x_i V_i^{2/3}}{\sum_j x_j V_j^{2/3}} \tag{3-124}$$

where r and V are the van der Waals and liquid molar volumes respectively. Both the expressions provided much improved predictions of the infinite dilution activity coefficients. Gmehling et al. [10] proposed another empirical modification of Φ_i' which provided better representation of mixtures of large and small molecules based on a large set of mixture data. The modification is given by Eqn. (3-125) and is used in the combinatorial expression of the modified UNIFAC (Dortmund) equation (from this point onwards this will be referred to as the mod. UNIFAC combinatorial).

$$\Phi_i' = \frac{x_i r_i^{3/4}}{\sum_j x_j r_j^{3/4}} \tag{3-125}$$

When these expressions are applied to systems which have massive differences in component size such as polymer solutions they tend to perform quite poorly. It was theorised that this deviation was due to the “free-volume” in the solution, which is the volume which is not occupied by molecules and is only really noticeable in polymer solutions. The mathematical definition of free-volume is not strict and varies from author to author [60, 75, 76]. The term “free-volume” actually encompasses several distinct definitions these have been given the following nomenclature by Bondi [76]:

- The empty volume (V_f); this is the difference between the macroscopic (liquid) molar volume (V) and the microscopic (van der Waals) molar volume(V_w):

$$V_f = V - V_w \tag{3-126}$$

- The expansion volume (V_E); this is the difference between the macroscopic (liquid) molar volume (V) and the volume of the substance in crystalline state (V_0):

$$V_E = V - V_0 \tag{3-127}$$

- The fluctuation volume (Υ); this is the volume in which the centre of mass of the molecule moves due to the thermal energy of the molecule. The definition of the fluctuation volume is not mathematically distinct and will not be considered any further here.

It is clear that regardless of the free-volume definition used, it is a function of temperature. In this work no consideration will be given to the temperature dependence since this effect is relatively small and for rigorous application requires high precision data. All reference to free-volume from this point forwards refers to the empty volume.

One of the first free-volume equations for applications to polymer solutions, called the ELBRO-FV expression [75], was proposed as follows:

$$\ln \gamma_i^{C-FV} = 1 - \frac{\Phi_i^{FV}}{x_i} + \ln \frac{\Phi_i^{Fv}}{x_i} \tag{3-128}$$

where the superscript C-FV refers to the combinatorial free-volume part and the free-volume fraction is defined as:

$$\Phi_i^{FV} = \frac{x_i V_i^{FV}}{\sum_j x_j V_j^{FV}} \tag{3-129}$$

$$V_i^{FV} = V_i - V_{w,i} \tag{3-130}$$

where V_i^{FV} is the free-volume, V_i is the liquid molar volume and $V_{w,i}$ is the van der Waals volume. The ELBRO-FV expression was modified by Kontogeorgis et al. [75] (named the GK-FV expression) by introducing the free-volume into the Guggenheim-Staverman equation instead of the Flory-Huggins:

$$\ln \gamma_i^{C-FV} = \ln \frac{\Phi_i^{FV}}{x_i} + 1 - \frac{\Phi_i^{FV}}{x_i} - \frac{1}{2} z q_i \left(\ln \frac{\Phi_i}{\theta_i} + 1 - \frac{\Phi_i}{\theta_i} \right) \tag{3-131}$$

These expressions provide much improved results with polymer systems indicating that there is some realism behind the idea of free-volume. These are by no means the only expressions available, many others for example UNIFAC-FV can be found in literature [54, 75, 77, 78].

4. TACKLING THE PROBLEM

As shown in the previous chapters the activity coefficient relies on the following 5 parameters:

1. Concentration
2. Temperature
3. Pressure
4. Solvent
5. Solute

Some of the popular approaches of determining the activity coefficient dependence on these parameters are discussed in the preceding chapter. The objective of this work is to combine the strengths of the methods discussed above and develop a new approach for activity coefficient prediction. The approaches taken for each of the variables are outlined in the sections following:

4.1. Concentration

The concentration dependence of the activity coefficient is non-trivial. Of all the methods discussed above, the NRTL-SAC and UNIFAC type methods have the most rigorous treatment of the concentration dependence of the activity coefficient, despite the aforementioned non self consistency. It is interesting to not that MOSCED provides a good representation of the SLE data, although it only uses the activity coefficient at infinite dilution. The big advantage is that it required only pure component property data. This is ideal since the pure component methods have the huge advantage in that the number of parameters only grows linearly with the number of groups. In the case of mixture methods (i.e. UNIFAC type methods) the number of parameters grows quadratically with the number of groups. For these reasons only infinite dilution will be considered, for SLE a compound will be considered to be at infinite dilution if the mole fraction solubility is below 1 percent (this is discussed again in Chapters 7 and 8, and was suggested in literature [43]).

4.2. Temperature

As with the concentration dependence, the temperature dependence of the activity coefficient is non-trivial. The temperature dependence is given by Eqn. (3-29). Since only infinite dilution is being considered this temperature dependence is somewhat simplified. Nevertheless, the infinite dilution temperature dependence is still non-trivial and is, for the sake of simplicity, generally ignored. Modified UNIFAC accounts for the temperature dependence fairly well; however this requires 2 or 3 parameters per binary interaction which greatly increases the

number of model parameters. The temperature dependence has been given some attention and the results are presented in sections 7.7 and 8.4.

4.3. Pressure

Unlike the concentration and temperature dependence the pressure dependence is typically trivial at low to moderate pressures. The pressure dependence of the activity coefficient is given by Eqn. (3-28), at infinite dilution this equation reduces to:

$$\left(\frac{\partial \ln \gamma_i^{\infty}}{\partial P}\right)_T = \frac{\bar{V}_i^{E,\infty}}{RT} = \frac{\bar{V}_i^{\infty} - V_i^0}{RT} \tag{4-1}$$

where $\bar{V}_i^{\infty}$ is the partial molar volume at infinite dilution and V_i^0 is the liquid molar volume of component *i*. If, as suggested, the pressure dependence is indeed negligible then the isothermal pressure dependence of $\bar{V}_i^{\infty}$ must be small (as it is clear that the V_i^0 is a weak function of pressure). Consider the examples of benzene and toluene shown in Fig. 4-1, it is apparent that the pressure dependence of $\bar{V}_i^{\infty}$ is small. The pressure dependence of the activity coefficient will therefore receive no more consideration in this work. Neglecting the pressure dependence implies the following (which is assumed true from this point forwards):

$$\left(\frac{\partial \ln \gamma_i^{\infty}}{\partial T}\right)_P = \frac{d \ln \gamma_i^{\infty}}{dT} \tag{4-2}$$

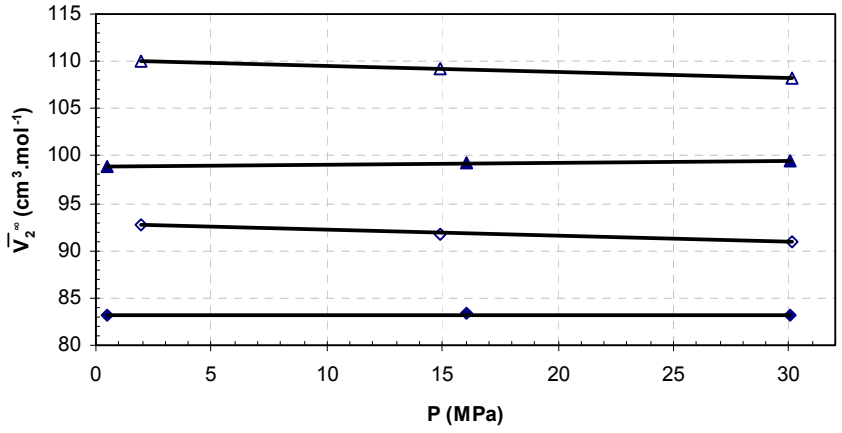


Fig. 4-1 $\bar{V}_2^{\infty}$ vs. P for benzene & toluene (2) in water (1) (△ - toluene at 373.15 K [79], ▲ – toluene at 298.15 K [79], ◇ - benzene at 373.15 K [79], ◆ - benzene at 298.15 K [79], ── lines to aid the eye).

4.4. Solvent

It is clear that the activity coefficient depends on both the choice of solvent[♠] and solute. In the preceding chapter various different treatments for simplifying the solvent were discussed. In the group contribution methods the solvent (and solute) is treated as a mixture of groups. The strength of this approach is, as mentioned before, that the number of functional group interactions is much less than the number of possible compound interactions. Nevertheless, the number of possible group interactions is still large. This is where the segmented activity coefficient methods discussed before are so powerful. They treat the solvent (and in many cases the solute) as a solution of molecule surface fractions that interact. The number of surface interactions is very small and therefore prediction relies not on surface interactions (since they are typically all given), but rather on the allocation of surface segments to the solute and solvent. Another great advantage with an approach such as this is that if the γ^{∞} in some reference solvents are known then the γ^{∞} in the other solvents (for which the size of the surface segments are known) can be interpolated. The disadvantage with this approach is that data in some reference solvents needs to be known. While this may not be a problem in some instances, in many cases it will.

The idea of this work is therefore to develop prediction methods in some reference solvents and then interpolate the behaviour within a solvent class or solvent group to yield predictions for other solvents. There are various options in determining the behaviour of solvents and some were discussed briefly in section 2.2. The objective is to make the interpolation as general as possible so as to avoid having lists of published solvent values. The limiting factor in an approach such as this is the availability of data in the reference solvents. This is given much attention in the chapters that follow.

4.5. Solute

Along with the solvent it is clear that the activity coefficient relies on the choice of solute. Since only infinite dilution considered in this work the infinite dilution activity coefficient in the reference solvents essentially reduces to a pure component property (at a single temperature chosen as 298.15 K). In each reference solvent this property is only a function of the solute molecular structure. Therefore since much work has already been invested into pure component group contribution methods by our group [15-20, 80], the data will be modelled using the experience gained previously. The approach is given in-depth attention in Chapters 7 and 8. The advantage of an approach such as this is that since it is specifically trained to data in a single solvent it is almost guaranteed that prediction methods of the UNIFAC and COSMO type will not have a comparable accuracy.

[♠] Included in the choice of solvent is the pH of the solvent. This is given no attention in this work, all solubility reproduced in this work are for pH neutral solvents.

5. EXPERIMENTAL DETERMINATION OF SOLUTE ACTIVITY

The chemical potential or fugacity of a solute can usually not be measured directly. The experimental techniques for extracting solute fugacity are, among others: VLE (Vapour-Liquid Equilibrium) measurements, LLE (Liquid-Liquid Equilibrium) measurements, SLE (Solid-Liquid Equilibrium) measurements and γ^{∞} measurements. The sources of data for complex multifunctional organic compounds are basically limited to SLE and partition coefficients. VLE and LLE data for such systems are scarce. This however does not mean that these data are useless for model development since these data can be used a "filler" for any methods developed ("filler" - even though the data may not be for complex multifunctional data, they are still useful for fitting group parameters). The three main sources of data are:

1. SLE data
2. Partition coefficient data
3. Infinite dilution data

The following sections will outline the applications and limitations of these 3 sources of data.

5.1. Solid-liquid equilibria

The phase equilibrium criterion given in Eqn. (3-19) for equilibrium between a solid and a liquid phase becomes (superscripts *S* and *L* represent solid and liquid phase respectively):

$$\hat{f}_i^S = \hat{f}_i^L \tag{5-1}$$

$$x_i \gamma_i^L f_i^L = z_i \gamma_i^S f_i^S \tag{5-2}$$

where x_i and z_i are the mole fractions in the liquid and solid phases respectively. If it is assumed that there is negligible mutual solubility in the solid (i.e. pure solid, $z_i \to 1, \gamma_i^S \to 1$) then:

$$x_i = \frac{f_i^S}{\gamma_i^L f_i^L} \tag{5-3}$$

where f_i^S is the fugacity of the pure solid and f_i^L is the fugacity of the (hypothetical) sub-cooled pure liquid. A simple (if the required data are available) way to estimate the f_i^S / f_i^L ratio can be found by plotting the vapour pressure curves for the solid and the liquid compound as shown in Fig. 5-1. Using the extrapolated (hypothetical) sub-cooled liquid vapour pressure (P_i^L) and

the solid vapour pressure (P_i^S) and assuming that the vapour phase behaves ideally the fugacity ratio can be given as:

$$\frac{f_i^S}{f_i^L} = \frac{P_i^S}{P_i^L} \tag{5-4}$$

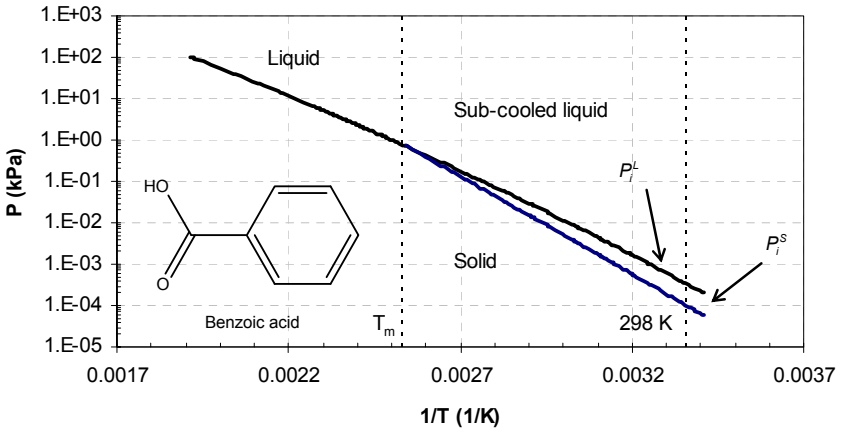


Fig. 5-1 P vs. 1/T for benzoic acid, sub-cooled liquid line extrapolated from liquid vapour pressure data using the Antoine equation (curves regressed to data from the DDB [28]).

Being thermodynamically rigorous, the ratio f_i^S / f_i^L can be related to the change in the Gibbs energy when going from the solid (point a in Fig. 5-2) to the sub cooled liquid (point d in Fig. 5-2).

$$\underset{a\to d}{\Delta G} = -RT \ln \frac{f_i^S}{f_i^L} \tag{5-5}$$

This process can be represented by the thermodynamic cycle given by Prausnitz et al. [34] and shown in Fig. 5-2.

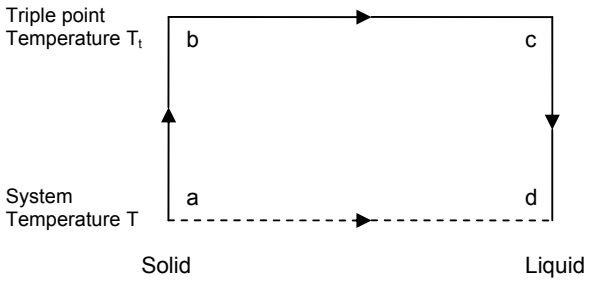


Fig. 5-2 Thermodynamic cycle for calculating the fugacity of a pure sub-cooled liquid, taken from Prausnitz et al. [34].

The Gibbs free energy is related to the enthalpy and the entropy through Eqn. (5-6). Since the enthalpy and the entropy are state functions, and are therefore not dependent on the path taken, the alternate path $a \to b \to c \to d$ can be used.

$$\underset{a\to d}{\Delta G} = \underset{a\to d}{\Delta H} - T \underset{a\to d}{\Delta S} \tag{5-6}$$

The path $a \to b \to c \to d$ can be calculated by using the appropriate relations for changing the temperature of a solid ($a \to b$), melting the solid ($b \to c$) and changing the temperature of the liquid ($c \to d$). This is represented by the following 4 equations:

$$\underset{a\to d}{\Delta H} = \underset{a\to b}{\Delta H} + \underset{b\to c}{\Delta H} + \underset{c\to d}{\Delta H} \tag{5-7}$$

$$\underset{a\to d}{\Delta H} = \Delta_{fus} H\big|_{T_t} + \int_{T_t}^{T} \Delta C_P dT \tag{5-8}$$

$$\underset{a\to d}{\Delta S} = \underset{a\to b}{\Delta S} + \underset{b\to c}{\Delta S} + \underset{c\to d}{\Delta S} \tag{5-9}$$

$$\underset{a\to d}{\Delta S} = \left.\frac{\Delta_{fus} H}{T_t}\right|_{T_t} + \int_{T_t}^{T} \frac{\Delta C_P}{T} dT \tag{5-10}$$

where $\Delta C_P \equiv C_P^l - C_P^s$ and T_t is the triple point temperature. Since data for the triple point is not abundant, and the influence of pressure on the solid-liquid equilibrium is small, we can reasonably assume that the triple point temperature is well approximated by the melting temperature (T_m). Combining Eqn's (5-5), (5-6), (5-8) and (5-10) yields the following:

$$\ln \frac{f_i^L}{f_i^S} = \frac{\Delta_{fus} H}{RT_m}\left(\frac{T_m}{T} - 1\right) + \frac{1}{RT}\int_{T_m}^{T} \Delta C_P dT - \int_{T_m}^{T} \frac{\Delta C_P}{RT} dT \tag{5-11}$$

It is clear from this equation that the last 2 terms are fairly similar and since the sign of each of the terms are opposite they are frequently assumed (under certain circumstances) to cancel each other out. This is given much discussion in the section following. All of the equations shown up to this point have been applicable to substances with a single solid phase over the temperature range $T \to T_t$. This is however not always the case as the solid phase can change; this change is called the transition point. As with the triple point at the transition point there is a certain enthalpy change called the enthalpy of solid-solid transition (from this point on simply referred to as the enthalpy of transition). If a substance with a single solid-solid transition point over the range $T \to T_t$ (as shown in Fig. 5-3), then Eqn's (5-8) and (5-10) are written as [81]:

$$\underset{a\to d}{\Delta H} = \Delta_{fus} H\big|_{T_t} + \int_{T_t}^{T_{trs}} \Delta C_P^{\alpha} dT + \Delta_{trs} H\big|_{T_{trs}} + \int_{T_{trs}}^{T} \Delta C_P^{\beta} dT \tag{5-12}$$

$$\underset{a\to d}{\Delta S} = \frac{\Delta_{fus}H}{T_t}\bigg|_{T_t} + \int_{T_t}^{T_{trs}} \frac{\Delta C_P^{\alpha}}{T} dT + \frac{\Delta_{trs}H}{T_{trs}}\bigg|_{T_{trs}} + \int_{T_{trs}}^{T} \frac{\Delta C_P^{\beta}}{T} dT \tag{5-13}$$

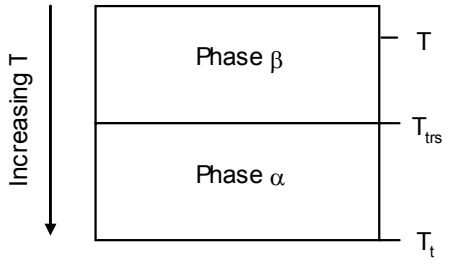


Fig. 5-3 Representation of the solid-solid phase transition of an arbitrary compound.

Substituting the melting point for the triple point and combining Eqn's (5-5), (5-6), (5-12) and (5-13) yields the following:

$$\begin{aligned}\ln\frac{f_i^L}{f_i^S} = \frac{\Delta_{fus}H}{RT_m}\left(\frac{T_m}{T}-1\right) + \frac{1}{RT}\int_{T_m}^{T_{trs}} \Delta C_P^{\alpha} dT - \int_{T_m}^{T_{trs}} \frac{\Delta C_P^{\alpha}}{RT} dT + \\ \frac{\Delta_{trs}H}{RT_{trs}}\left(\frac{T_{trs}}{T}-1\right) + \frac{1}{RT}\int_{T_{trs}}^{T} \Delta C_P^{\beta} dT - \int_{T_{trs}}^{T} \frac{\Delta C_P^{\beta}}{RT} dT\end{aligned} \tag{5-14}$$

For any subsequent transition points Eqn's (5-12) to (5-14) can be modified in a similar way. No consideration is given to transition points in this work since any data available is sparse.

5.1.1. Heat capacity approximations

An example of the liquid and solid heat capacity curves is shown in Fig. 5-4. Assuming that ΔC_P is constant over the temperature range, using the average heat capacity change at the ($\Delta \overline{C}_P$) and combining Eqn's (5-3) and (5-11):

$$\ln\frac{f_i^L}{f_i^S} = -\frac{\Delta_{fus}H}{RT_m}\left(\frac{T_m}{T}-1\right) + \frac{\Delta\overline{C}_P}{R}\left(\frac{T_m}{T}-1\right) - \frac{\Delta\overline{C}_P}{R}\ln\frac{T_m}{T} = \ln x_i\gamma_i \tag{5-15}$$

While this assumption is not totally accurate it is typically quite a good one (Fig. 5-8) and appreciably simplifies the equation and reduces the amount of data needed (solid and liquid heat capacity data are typically unavailable). For temperatures not too remote from the melting point some authors [38, 82, 83] suggest that the heat capacity change is negligible. This results in the following simplified relation:

$$\ln x_i = -\ln\gamma_i - \frac{\Delta_{fus}H}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right) \tag{5-16}$$

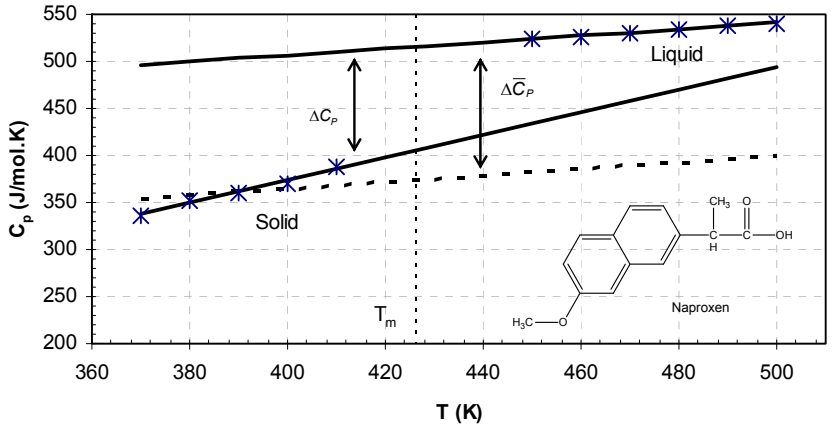


Fig. 5-4 Solid and liquid heat capacity data for naproxen (x – heat capacity data [84], — linear fits to the data, ---- linear fit to the solid heat capacity data with the same slope as the liquid heat capacity data).

Many authors have suggested that this relation represents an oversimplification [33, 60, 84-88] since the heat capacity change upon phase transition is actually very often significant. Since heat capacity data are very scarce in the literature some approximation is necessary to make the equation widely applicable. There are predictive methods [89] available for the heat capacity but these are not suitable for use in this work since this would introduce a 4th variable (the accuracy of the prediction method) which would be hard to judge. An assumption which has been proposed in literature [60] is to assume that the heat capacity at the melting point is approximated by the entropy of fusion (Hildebrand and Scott [60] show a short derivation):

$$\Delta \overline{C}_P \approx \Delta_{fus} S = \frac{\Delta_{fus} H}{T_m} \tag{5-17}$$

This approximation can be quite a poor one but as shown by Neau et al. [84] $\Delta \overline{C}_P$ is generally closer to the entropy of fusion than to zero. Using this approximation in Eqn. (5-15) results in a very simple expression:

$$\ln x_1 = -\ln \gamma_1 - \frac{\Delta_{fus} H}{R T_m} \ln\left(\frac{T_m}{T}\right) \tag{5-18}^{\clubsuit}$$

$^{\clubsuit}$ It is interesting to note that a Taylor expansion with 2 terms of Eqn. (5-18), centred at T_m, yields the following:

$$\ln\left(\frac{T_m}{T}\right) \approx \left(\frac{T_m - T}{T}\right) + \frac{1}{2}\left(\frac{T_m - T}{T}\right)^2$$

If the first term on the RHS is substituted back into Eqn. (5-18) then Eqn. (5-16) results. This explains why both expressions are so similar for temperatures close to T_m. The error (or difference) when using Eqn. (5-16) as opposed to Eqn. (5-18) is below 10% when $T > T_m/1.45$. This means at 298 K the error is below 10% when $T_m < 435$ K. For a fair number of compounds this would hold true but since the input parameters in the 2 equations are identical one may as well use the (seemingly) more accurate one.

This expression also agrees with the observations of Hildebrand [60, 85] who noticed that the solubility of some non-electrolytes were proportional to the logarithm of temperature. For applications where many data for a single solute are available the following expression for the heat capacity change is recommended [86]:

$$\Delta\overline{C}_P \approx \sigma.\Delta_{fus}S = \sigma.\frac{\Delta_{fus}H}{T_m} \tag{5-19}$$

where σ is a solute specific fitted parameter. Fig. 5-5 and Fig. 5-6 show how the ideal solubility ($\gamma_i = 1$) changes with the melting temperature and the heat of fusion of the solute in a solvent at 298.15 K. This shows that the solubility is quite a strong function of the pure component properties of the solute. At first glance both figures seem identical but the shapes are slightly different and there is quite a large difference in the ideal solubility values as is evident from the level curves shown in Fig. 5-7. From both figures it is clear that the ideal solubility (or activity) decreases both as the melting point and heat of fusion increases.

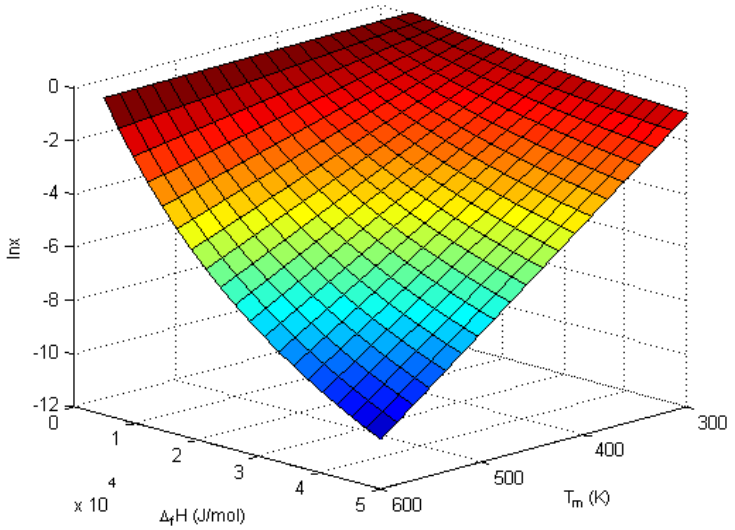


Fig. 5-5 The ideal solubility as a function of the melting temperature and the heat of fusion at 298.15 K using Eqn. (5-16).

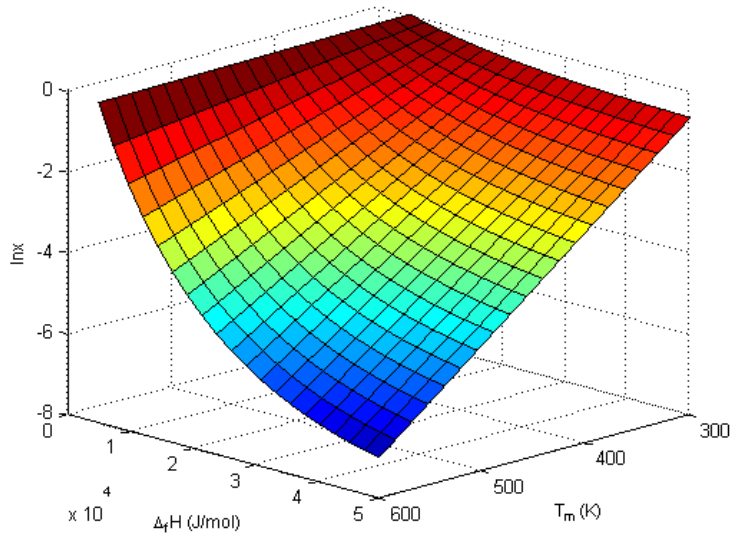


Fig. 5-6 The ideal solubility as a function of the melting temperature and the heat of fusion at 298.15 K using Eqn. (5-18).

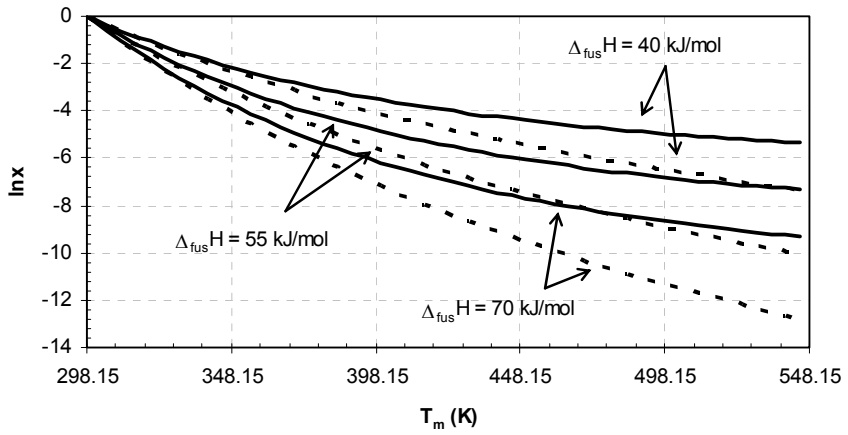


Fig. 5-7 Comparison of the 2 different heat capacity approximations for ideal solutions as a function of T_m and $\Delta_{fus}H$ at 298.15 K (—— Eqn. (5-18), ---- Eqn. (5-16)).

In order to quantify the difference between the various approximations that are made for ΔCp consider the example of naproxen shown in Fig. 5-8. It is apparent from the curves that the mean ΔCp and the full temperature dependent term are more or less equivalent. The other 2 approximations (mean and entropy approximations) are worse with the approximated (entropy of fusion) ΔCp being quite a lot closer than the negligible ΔCp curve. Up to about 50 K from the melting point all the approximations are approximately equivalent.

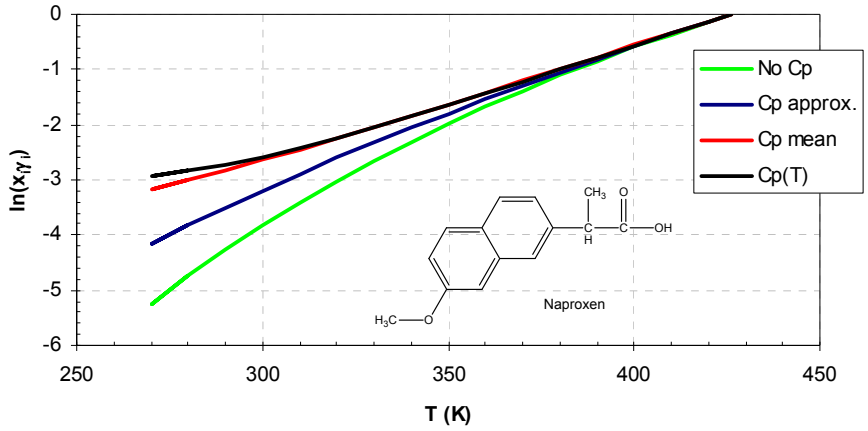


Fig. 5-8 $\ln(x_i\gamma_i)$ vs T for naproxen for various heat capacity approximations (No Cp – assumed ΔCp is negligible, Cp approx. – assumed ΔCp is equivalent to the entropy of fusion, Cp mean – ΔCp value obtained from Fig. 5-4, Cp(T) – temperature dependent ΔCp taken from Neau et al. [84]).

5.2. Partition coefficients

Partition coefficients are widely used in the pharmaceutical and biological sciences. They are used as a measure of how compounds partition (or distribute) between 2 phases or compartments. A partition coefficient is defined as the ratio of the equilibrium concentration of a compound in one compartment with respect to another. Descriptions of two of the most common partition coefficients are given in the sections following.

5.2.1. Octanol-water partition coefficient

Arguably the most commonly available partition coefficient is the octanol-water partition coefficient (K_{ow}). Values of K_{ow} are key parameters for estimating toxicity, bioaccumulation, and sorption into soils and sediments [90]. They do, however, have some meaning on their own as they are a measure of how a compound distributes between an organic phase (e.g. a fish) and an aqueous phase [91]. K_{ow} is defined as a measure of the partitioning between water-saturated octanol and octanol-saturated water phases:

$$K_{ow} = \frac{C_i^{ow}}{C_i^{wo}} \tag{5-20}$$

where C is the concentration ($mol.l^{-1}$) and the superscripts refer to the water saturated octanol (ow) and the octanol saturated water (wo) phases. Using the molar volumes of octanol ($V_{oct} = 6.313\ mol.l^{-1}$) and water ($V_{water} = 55.343\ mol.l^{-1}$) at 298.15 K and the equilibrium relationships (essentially pure water phase; 20.7 mol% water in the octanol phase [92]) the following relationship results (the superscript *wo* is replaced by *w* since the water is essentially pure):

$$K_{ow} = 0.1508 \frac{x_i^{ow}}{x_i^{w}} \tag{5-21}$$

If it is assumed that the solute concentration is small, then using the criterion for phase equilibrium (isofugacity criterion Eqn. (3-19)), the following results:

$$K_{ow} = 0.1508 \frac{\gamma_{i,w}^{\infty}}{\gamma_{i,ow}^{\infty}} \tag{5-22}$$

This relation is only useful if the infinite dilution activity coefficient is known in water saturated octanol as well as pure water. However Tse and Sandler [92] have shown that the octanol-water partition coefficient can be related, fairly simply, to the pure phase infinite dilution activity coefficients:

$$\log K_{ow} = 0.1 + 0.91.\log\left(0.1508 \frac{\gamma_{i,w}^{\infty}}{\gamma_{i,o}^{\infty}}\right) \tag{5-23}$$

where this correlation should only be applied to hydrophobic species ($1.0 < \log K_{ow} < 5.0$). The great advantage of a correlation such as this is that K_{ow} can be calculated from experimental or estimated values of $\gamma_{i,w}^{\infty}$ and $\gamma_{i,o}^{\infty}$. Also if K_{ow} and $\gamma_{i,w}^{\infty}$ are known then the value for the octanol infinite dilution activity can be extracted.

While there is a large amount of K_{ow} data available in the literature, it has been noted that there is a large amount of variability in the data for hydrophobic compounds [90]. This combined with the fact that there are only empirical methods to extract infinite dilution data from K_{ow}, these data will not be used in model development, but will be used for model tests.

5.2.2. Air-water partition coefficient

A partition coefficient which is particularly important in the environmental sciences is the air-water partition coefficient (K_{aw}). As is evident by the name, the air-water partition coefficient determines the partitioning between air and water:

$$K_{aw} = \frac{C_a}{C_w} \tag{5-24}$$

where C_a and C_w are the concentrations ($mol.l^{-1}$) in the air and water phases respectively. K_{aw} is sometimes referred to as the dimensionless Henry's coefficient [93]. Eqn. (5-24) can be written in a more useful form by following the following procedure. The fugacity in solution of some species *i* in the air (a) and water (w) phase are given as follows:

$$\hat{f}_i^{a} = \hat{\phi}_i y_i P \tag{5-25}$$

$$\hat{f}_i^w = \gamma_i x_i f_i^0 \tag{5-26}$$

where y_i is the vapour phase mole fraction, x_i is the liquid phase mole fraction and $\hat{\phi}_i$ is the vapour phase fugacity coefficient. If one assumes that the vapour phase is ideal (i.e. $\hat{\phi}_i \to 1$ and $f_i^0 \to P_i^{sat}$) and that the compound is dilute in the liquid phase (i.e. $\gamma_i = \gamma_i^\infty$) then isofugacity criterion (Eqn. (3-19)) these equations can be combined to yield:

$$\frac{y_i}{x_i} = \frac{\gamma_i^\infty P_i^{sat}}{P} \tag{5-27}$$

Keeping the same assumptions (ideal gas and dilute aqueous phase) allows Eqn. (5-24) to be rewritten as:

$$K_{aw} = \frac{y_i}{x_i}\frac{V_w P}{RT} \tag{5-28}$$

where V_w is the molar volume of water (approx. 0.018 $m^3.mol^{-1}$ at 298.15 K). Combining Eqn's (5-27) and (5-28) yields:

$$K_{aw} = \frac{\gamma_i^\infty P_i^{sat} V_w}{RT} \tag{5-29}$$

Therefore the water infinite dilution activity coefficient can be calculated if K_{aw} and the vapour pressure are known. This is however not a viable option for getting data for complex multifunctional compounds as they typically have very small vapour pressures which are difficult to accurately predict. As with K_{ow} data, this data will not be used in model development, but will be used as a model test.

5.3. Infinite dilution data

Infinite dilution data in this context refers to any data which is not taken from SLE data, i.e. data extracted from VLE, LLE (see below) or 'directly measured' infinite dilution data (e.g. from gas-liquid chromatography). These data are frequently measured (or analysed in the case of LLE) by using the vapour pressure of the compounds. This means that for complex multifunctional compounds which typically have very small vapour pressures, such methods are not applicable. The data for more volatile (smaller) molecules will still be useful in that it will significantly enlarge any training set which is used and therefore mean that there are more data behind each group. The direct measurement (i.e. not using G^E models to extrapolate from VLE or LLE data; or fusion data to extract from SLE data) of infinite dilution activity coefficient data is usually carried out by one of the following methods [94]:

- Gas-liquid chromatography (GLC)

- Differential ebulliometry
- Dew-point method
- Differential pressure
- Inert gas stripping
- Inverse solubility

Reviews of the above methods are widely available in literature [94-97] and therefore, with the exception of GLC and inverse solubility, will not be discussed further.

5.3.1. Gas-liquid chromatography

In order to verify some of the methods presented in the following chapter, some measurements were carried out using GLC with a squalane stationary phase. The basic principle of GLC is that a non-volatile (usually, but doesn't have to be) solvent is coated onto some inert solid support and packed into a column (usually some narrow tubing, except in the case of volatile solvents). A carrier gas (usually helium) is then passed over the coated packing. When the column is at the steady state temperature of interest then the solute is injected into the column and provided the solute is volatile enough the solute partitions relatively quickly into the gas and when the solute containing gas passes over the stationary phase the solute interacts with the solvent. Depending on the nature of the interaction the retention time (how long the solvent stays in the column) of the solute changes. This retention time can be directly correlated with the activity coefficient at infinite dilution.

Since the experimental procedure is not the focus of this work it will not be covered in detail here. An almost identical procedure was used previously for infinite dilution activity coefficient measurements done by members of the University in KwaZulu-Natal Thermodynamics Research Unit (UKZN TRU) [98, 99]. The procedures were different in that squalane is used instead of the ionic liquid and hexane was used instead of dichloromethane to coat the squalane on the column packing. The mass of packing that was used in the column was 6.2469 g and the mass percentage squalane was 28.529 %. In order to test whether the column produces useful measurements some test solutes were used and these results are shown in Table 5-1.

Table 5-1 Comparison of the literature [28] and the measured γ^{∞} for some test compounds in squalane (323K).

Solute (2)	Literature γ_2^{∞}	Measured γ_2^{∞}	RMD%
Acetone	2.85	2.7	5.3
Acetronitrile	11.1	10.39	6.4
Benzene	0.65	0.65	0
Cyclohexane	0.52	0.50	3.8
MEK	2.06	1.98	3.9
2-Propanol	8.39	8.52	1.5

The solutes and temperatures were chosen on the basis of available data in smaller solvents (e.g. hexane) since the purpose was to extrapolate the squalane data to the smaller solvents. Three runs were done for each solute and the results are shown in Table 6-2.

5.3.2. Inverse solubility

Inverse solubility is a very simplified case of extracting infinite dilution activity coefficient data from LLE data. Under certain conditions the infinite dilution activity coefficient of a solute is simply the inverse of the (liquid) solubility in some other immiscible phase. The criteria for this special case are outlined in the paragraph following.

The isofugacity criterion (Eqn. (3-19)) for 2 partially miscible liquid phases (α and β) is:

$$\hat{f}_i^{\alpha} = \hat{f}_i^{\beta} \tag{5-30}$$

$$x_i^{\alpha}\gamma_i^{\alpha} = x_i^{\beta}\gamma_i^{\beta} \tag{5-31}$$

If there is a binary system (of component *1* and *2*), and both phases are sufficiently immiscible then, as an arbitrary example, if phase α is practically pure *1* and phase β is practically pure *2* then $x_2^{\beta} \approx 1$, $\gamma_2^{\beta} \approx 1$ and $\gamma_2^{\alpha} \approx \gamma_2^{\infty}$ and Eqn. (5-31) reduces to (dropping the notation of α and β as it is implied):

$$\gamma_2^{\infty} = \frac{1}{x_2} \tag{5-32}$$

There is no hard and fast rule for determining the solubility threshold; it suffices to say that the solubility in each phase should certainly be below 1 mole percent. What is very important to note, which is sometimes just assumed always true, is that the mutual solubility must be low not just the solubility in one phase. Consider the example of hexylamine(2) and water(1), as one would expect at 303.15 K hexylamine is only slightly soluble in the water (x_2 = 0.00173 [100]) while water is extremely soluble in hexylamine (x_1 = 0.901 [100]). If one simply found a solubility value of x = 0.0017 of hexylamine in water in literature the temptation would be to use the inverse solubility. This would be a large overestimation of γ^{∞} (using Eqn. (5-32) yields $\gamma_2^{\infty} = 578$ which is about 5 times larger than the experimental value $\gamma_2^{\infty} = 114.5$ at 298.15 K found in the DDB [28]).

6. REDUCTION OF INFINITE DILUTION DATA TO A COMMON SOLVENT

Since this work is concerned with the prediction of the activity coefficient of complex compounds in common solvents it is imperative that there are sufficient data for the solvents that are chosen. When considering the infinite dilution activity coefficient database in DDB [28] water was by far the solvent with the largest amount of data (i.e. for different solutes) as shown in Table 6-1. From the table it is clear that there is a significantly greater amount of data in water than in any other solvent; if SLE data are included this difference becomes much more pronounced. The presence of a majority of large compounds in the top 10 is due to the low volatility of such compounds and hence the suitability to being stationary phases in GC columns (see section 5.3).

Table 6-1 Number of solutes with data in each solvent for the infinite dilution activity coefficient data in the DDB [28] (for all temperatures, alkane solvents indicated in bold).

Solvent	No. Solutes
Water	585
Squalane	**207**
Hexadecane	**198**
Sulfolane	137
1-Octanol	128
Phthalic acid dinonyl ester	126
Heptane	**116**
N-Methyl-2-pyrrolidone	115
Octadecane	**103**
19,24-Dioctadecyldotetracontane	**99**

However when trying to compile a set of data it soon became clear that large amounts of data in other single solvents was not that easy to come by. It is for this reason that one of our objectives was to be able to grow our dataset in a simple reliable way; the following sections outline our proposed method for alkanes, alcohols and ketones.

6.1. Alkanes

As is evident from Table 6-1, 5 out of the top 10 solvents are alkanes. Therefore if it is possible to reduce the measurements in any alkane to some reference alkane (e.g. hexane) then it would be possible to expand the available data in hexane for model development. The success of methods like UNIFAC and mod. UNIFAC has shown that the solution of groups principle is able to successfully predict the behaviour of a large variety of liquid mixtures. Since alkanes are all made up of sp^3 carbons and hydrogens (CH_2 UNIFAC main group, i.e. encompassing CH_3-, -CH_2-, >CH- and >C< subgroups) the energetic interactions at infinite dilution should be the same regardless of what alkane is being considered. This principle is indicated in Fig. 6-1 where at infinite dilution there will only be a single solute molecule surrounded by a solution of CH_2 groups. Therefore regardless of the source of these groups (i.e. pentane, hexadecane, squalane etc) the energetic interactions will be identical. This

reasoning is only applicable to infinite dilution of the solute since at infinite dilution the solute molecule will only "see" solvent molecules and not other solute molecules (and hence eliminating the solute-solute energetic interactions).

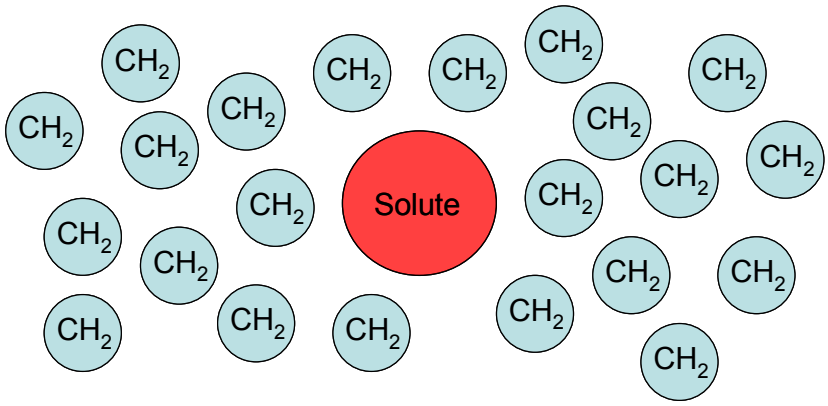


Fig. 6-1 An arbitrary solute molecule at infinite dilution in an alkane solvent.

As mentioned in section 3.2.1, the energetic interactions for the activity coefficient in UNIQUAC are represented by a residual contribution. The size/shape contribution to the activity coefficient is represented by a combinatorial contribution. If one considers an arbitrary solute (*i*) in 2 different alkane solvents (*sol1* and *sol2*) at infinite dilution then, since the residuals are assumed equal, the following would result:

$$\frac{\gamma_{i,sol1}^{\infty}}{\gamma_{i,sol2}^{\infty}} = \frac{\gamma_{i,sol1}^{C,\infty}}{\gamma_{i,sol2}^{C,\infty}} \tag{6-1}$$

where the superscript *C* denotes the combinatorial expression and the subscripts *sol1* and *sol2* differentiate between any two alkane solvents. This expression can therefore be rearranged to give the infinite dilution activity coefficient in any alkane solvent (*sol1*) relative to a solvent for which data are known (*sol2*):

$$\gamma_{i,sol1}^{\infty} = \gamma_{i,sol2}^{\infty} \frac{\gamma_{i,sol1}^{C,\infty}}{\gamma_{i,sol2}^{C,\infty}} \tag{6-2}$$

To the knowledge of this author the form of Eqn. (6-1) or (6-2) is unique to this work, the theory behind it is, however, not. A very similar methodology has been applied in literature when analysing the UNIFAC expression [101, 102]. While in both these cases the authors realised the usefulness of such an approach for analysing combinatorial expressions they seemed to have missed some of the other applications (while many of the conclusions gleaned from this work are the same as those in literature neither propose a solution to the problems found). The paper by Abildskov et al. [101] is especially interesting and proposes some very interesting theories about the source of the error in the combinatorial expressions (their theory that free-volume plays a role is shown to be true – however not in the way they had hoped – in the sections following).

6.1.1. Development of a modified combinatorial expression

What Eqn. (6-2) suggests is that if a good combinatorial expression is known then data in one alkane solvent can be extrapolated to another. Different combinatorial expressions have been discussed in section 3.2.4. The three expressions which will be examined in this work are the Guggenheim-Staverman (SG, Eqn. (6-3)), mod. UNIFAC (Eqn. (6-4)) and GK-FV (Eqn. (6-5)) expressions, which at infinite dilution (for a binary mixture) are given by:

$$\ln \gamma_2^{C-SG,\infty} = 1 - \frac{r_2}{r_1} + \ln\left(\frac{r_2}{r_1}\right) - 5q_2\left(1 - \frac{r_2 q_1}{q_2 r_1} + \ln\left(\frac{r_2 q_1}{q_2 r_1}\right)\right) \tag{6-3}$$

$$\ln \gamma_2^{C-MU,\infty} = 1 - \left(\frac{r_2}{r_1}\right)^{3/4} + \ln\left(\frac{r_2}{r_1}\right)^{3/4} - 5q_2\left(1 - \frac{r_2 q_1}{q_2 r_1} + \ln\left(\frac{r_2 q_1}{q_2 r_1}\right)\right) \tag{6-4}$$

$$\ln \gamma_2^{C-GKFV,\infty} = 1 - \frac{V_2^{FV}}{V_1^{FV}} + \ln\left(\frac{V_2^{FV}}{V_1^{FV}}\right) - 5q_2\left(1 - \frac{r_2 q_1}{q_2 r_1} + \ln\left(\frac{r_2 q_1}{q_2 r_1}\right)\right) \tag{6-5}$$

For all of the equations above, the physically realistic values of the r and q values given by Bondi [58] will be used. This may seem slightly unfair on the mod. UNIFAC expression (and in many instances it is because the use of the mod. UNIFAC r and q values improves the extrapolations) however it was decided that the physically realistic values are a better test of the combinatorial expression (see Appendix B). Some examples of the usage of Eqn. (6-2) are shown in Fig. 6-2 to Fig. 6-4. In all the examples the “biggest solvent“ was chosen as the reference solvent (for no reason other than consistency) and where multiple data points were available GLC (Gas-Liquid Chromatography) measurements were considered as superior. LLC (Liquid-Liquid Chromatography) data were not considered as a reference as it is typically poor [65] but it is included on some of the plots following.

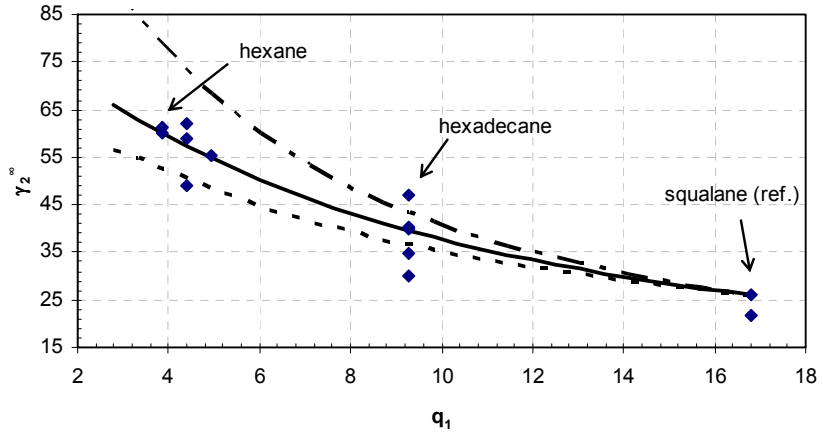


Fig. 6-2 γ_2^∞ vs. q_1 for ethanol(2) in alkane solvents(1) using squalane as the reference solvent at 298.15 K (◆ – data from the DDB, ⁻·⁻·⁻ SG combinatorial, ····· mod. UNIFAC combinatorial, ── GK-FV combinatorial).

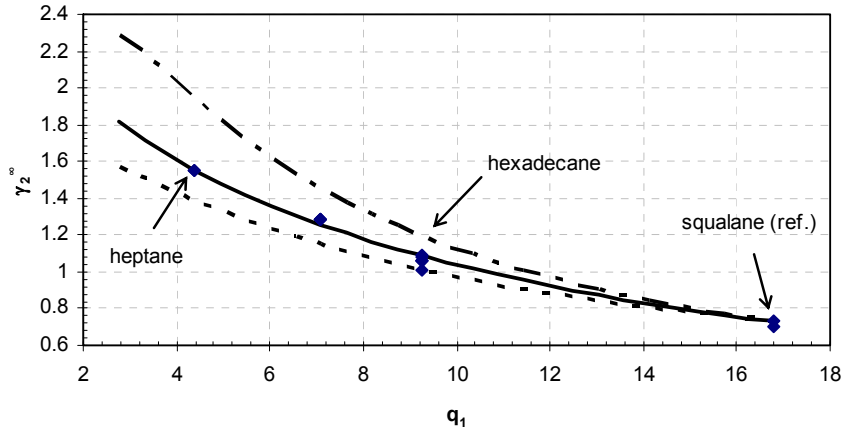


Fig. 6-3 γ_2^∞ vs. q_1 for benzene(2) in alkane solvents(1) using squalane as the reference solvent at 298.15 K (◆ – data from the DDB, ⁻·⁻·⁻ SG combinatorial, ····· mod. UNIFAC combinatorial, ── GK-FV combinatorial).

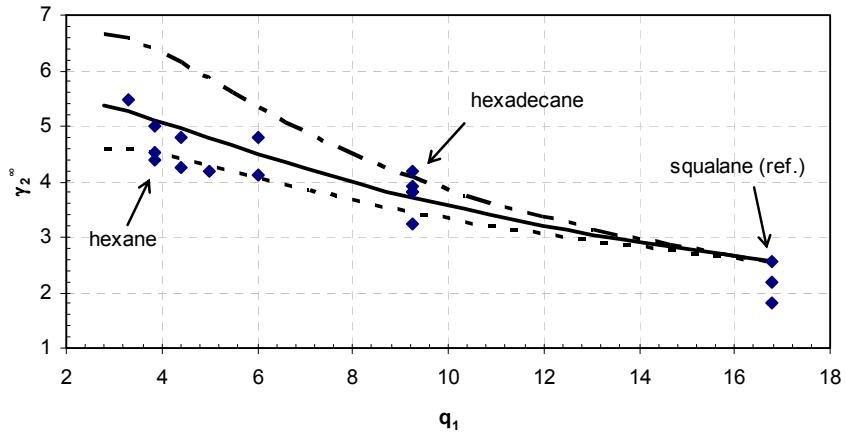


Fig. 6-4 γ_2^∞ vs. q_1 for butanone(2) in alkane solvents(1) using squalane as the reference solvent at 298.15 K (◆ – data from the DDB, ⁻·⁻·⁻ SG combinatorial, ····· mod. UNIFAC combinatorial, ── GK-FV combinatorial).

As was expected the Guggenheim-Staverman expression showed quite a large deviation when going from small to large solvents (see section 3.2.4). The mod. UNIFAC expression performed slightly better but tended to under-predict the activity coefficient when going from a large reference solvent to a small one (and *vice versa*). The free-volume expression provides a very good representation of the experimental data and can go from very big to very small solvents with a fair level of accuracy. This success is somewhat surprising since free-volume is almost never considered in non-polymer applications.

The success of the free-volume term indicates that free-volume does have an impact on the combinatorial contribution and should be accounted for. The presence of a free-volume effect has been shown to play a role when determining the water solubility in alkanes [103]. In a personal communication (E Johansson 2009, pers. comm., 7 October), Dr. Johansson has indicated that molecular simulations that himself and co-workers have carried out have shown

the definite presence of free-volume in alkane solutions. In the simulations, addition of water molecules into a solution of decane molecules had no effect on the density of the decane (i.e. the water free decane solution, since the experiments are molecular simulations it is possible to determine such data). While the use of water as a solute is somewhat of a worst case scenario (since water is one of the smallest solutes available) these observations are still significant. Initially when analysing the different combinatorial expressions, quite accidentally, only solutes which were smaller than all of the solvents used were considered (this is due to the high availability of such data). However when the size of the solute became larger than the size of the solvent, the predictions started to show very large negative deviations. This is quite well illustrated in the example of ethylcyclohexane shown in Fig. 6-5. All three combinatorial expressions fail to reproduce the experimental findings as soon as the size of the solute (q_2) becomes larger than the size of the solvent (q_1). The smaller the ratio q_1/q_2 the larger this error becomes.

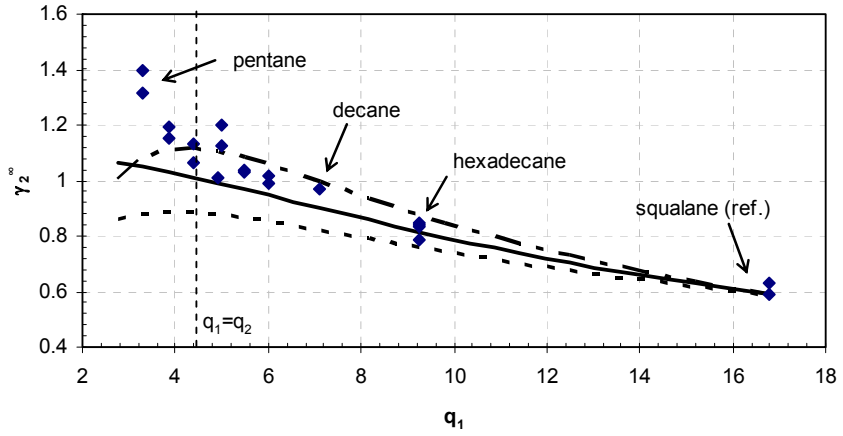


Fig. 6-5 γ_2^∞ vs. q_1 for ethylcyclohexane (2) in alkane solvents(1) using squalane as the reference solvent at 298.15 K (◆ – data from the DDB [28], - · - · - SG combinatorial, ····· mod. UNIFAC combinatorial, — GK-FV combinatorial).

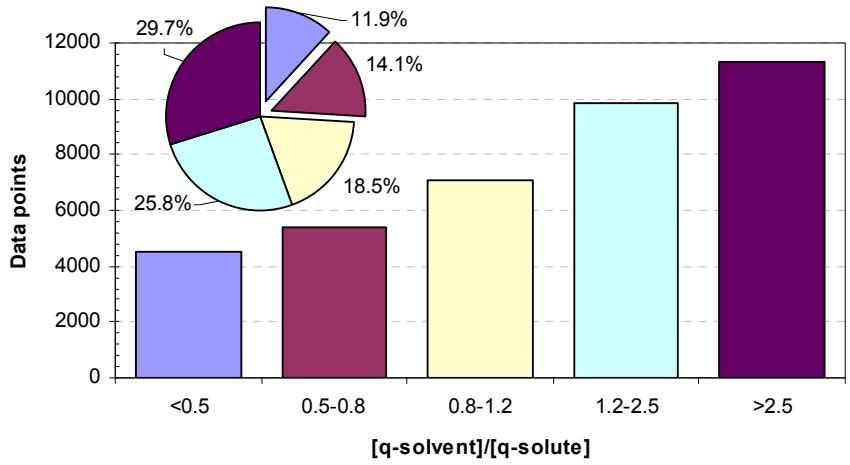


Fig. 6-6 Size ratio's for all points (i.e. at all temperatures) in the in the DDB [28] γ^∞ database for all solvents.

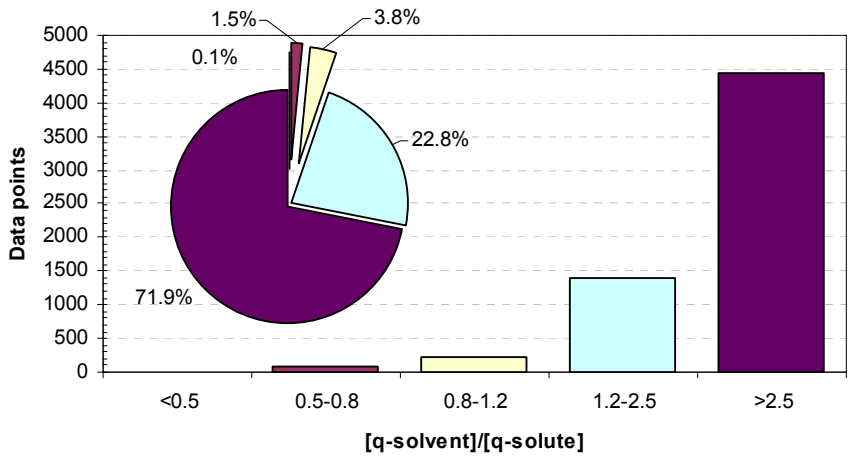


Fig. 6-7 Size ratio's for all points (i.e. at all temperatures) in the in the DDB [28] γ^{∞} database for all alkane solvents.

The reason why this effect probably hasn't been so noticeable before is twofold. Firstly the data for such systems is scarce as evident in Fig. 6-6 and Fig. 6-7. The scarcity of data is especially true for alkane solvents, with only 1.6 % of the data lying in the noticeable range. Secondly for systems where there are an abundance of such data (i.e. systems containing water, acetone etc.) this effect would be absorbed by the interaction parameters in the residual part.

In order to solve the problem of large solutes in small solvents, the GK-FV expression ($\ln\gamma_2^{\text{C-FV},\infty}$ in Eqn. (6-6)) was empirically modified (Eqn. (6-6)) for infinite dilution behaviour prediction based on a variety of data from the DDB [28] (a multi-component, composition dependent expression is discussed below).

$$\ln\gamma_2^{\text{C-Cav},\infty} = \ln\gamma_2^{\text{C-FV},\infty} + \ln\gamma_2^{\text{Cav},\infty} \tag{6-6}$$

$$\ln\gamma_2^{\text{Cav},\infty} = \left(\frac{V_2^{FV}}{V_1^{FV}} - \frac{V'^{FV}_2}{V'^{FV}_1}\right) \tag{6-7}$$

$$V'^{FV}_i = \left(V_i\right)^{2/3} - \left(V_i^*\right)^{2/3} \tag{6-8}$$

where V_i is the liquid molar volume and V_i^* is the van der Waals volume. Eqn. (6-6) provides a far better representation of the behaviour of large molecular size solutes in small solvents (as shown in Fig. 6-11 for the solute ethylcyclohexane which was considered above) while still providing very good results for small solutes in larger solvents.

This modification was initially thought to account for some cavitation contribution. The thinking being that if it is more difficult to create a cavity in, for example, pentane than in octane then this would cause a drop in solubility and therefore an increase in activity coefficient. A

cavitation effect seems quite intuitive and has shown to, at least qualitatively, produce the same behaviour as is needed [104-106]. This is illustrated by Fig. 6-8 where for an increasing cavity size the Gibbs free energy of cavity formation increases. The problem considered here is however not for a changing cavity size in single solvent but rather for a single cavity size in a changing solvent. When considering the energy of cavity formation of a single solute in a changing (alkane) solvent the picture is quite different. One common method for calculating the cavitation energy in literature is the Scaled Particle Theory (SPT) [107]:

$$\Delta G_{cav} = RT\left[-\ln(1-y)+\frac{9}{2}\left(\frac{y}{1-y}\right)^2 r^2+3\frac{y}{1-y}r(1+r)\right]+\frac{yP}{\rho_s}r^3 \tag{6-9}$$

$$y=\frac{\pi}{6}d_s^3\rho_s \tag{6-10}$$

$$r=\frac{d_m}{d_s} \tag{6-11}$$

where R,T & P have their usual meaning and d_s is the effective solvent molecule diameter (i.e. considering the molecular volume to be a sphere and finding the diameter, all molecule volumes used are from Bondi's method [58]), d_m is the effective cavity diameter and ρ_s is the numeral density (N_A / V_m). Fig. 6-9 shows the results for ethylcyclohexane in n-alkane solvents. When the solvents are smaller the energy of cavity formation tends to fall away instead of increasing as one would expect if cavitation was the effect observed here. SPT is however only an approximation, it considers molecules to be spherical, and for large n-alkanes this is certainly not the case. Therefore another approach needs to be tested. If the second approach agrees with SPT (for a large solute in small solvents, at least) then it is probably sufficient to accept both. Ideally this second approach would come from some molecular simulation but no suitable simulations could be found.

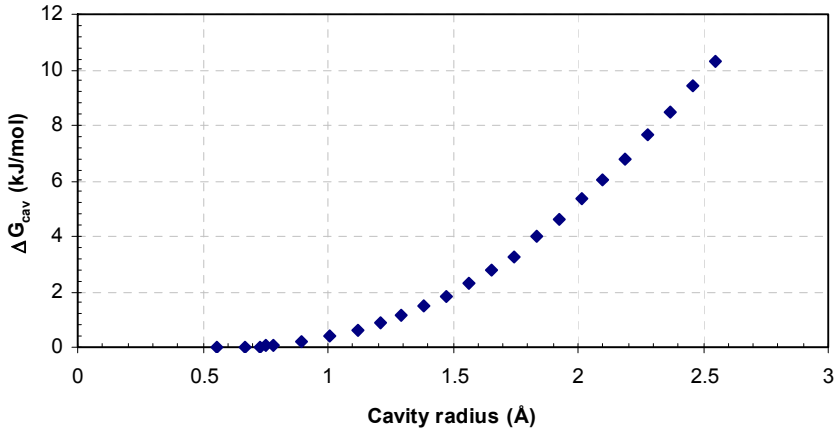


Fig. 6-8 ΔG_{cav} vs. Cavity radius in octanol (◆ – data from molecular simulations [105]).

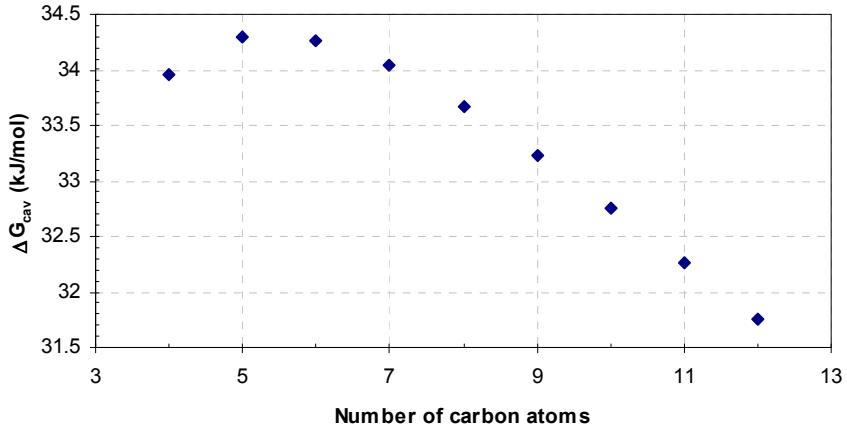


Fig. 6-9 ΔG_{cav} vs. number of carbon atoms for ethylcyclohexane in n-alkanes from SPT [107].

For a second approach, consider the work of Uhlig [108] who approximated the energy of cavity formation as follows:

$$\Delta G_{cav} = 4\pi r^2 \sigma \tag{6-12}$$

where $4\pi r^2$ is from the surface area of a sphere (here the van der Waals surface area from Bondi's [58] method will be used) and σ is the surface tension (data taken from Lange's handbook [109]). Fig. 6-10 shows the results for ethylcyclohexane in n-alkane solvents. Once again the expected behaviour is not observed (there are other approximations in literature [110], these won't be discussed further, it suffices to say they all predict similar behaviour).

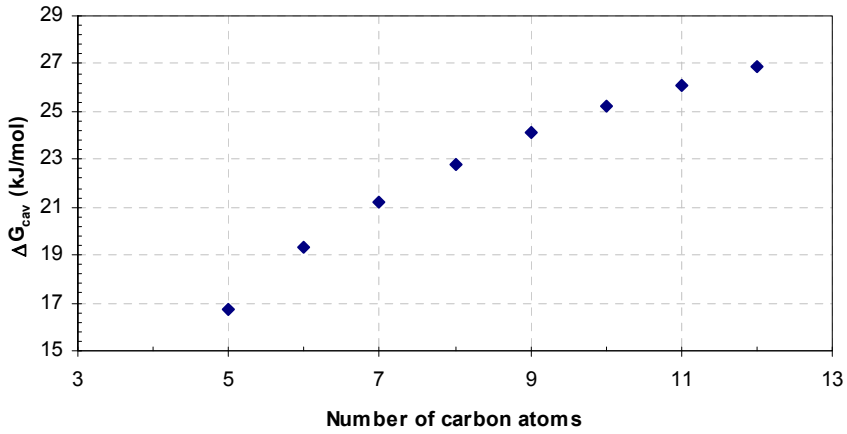


Fig. 6-10 ΔG_{cav} vs. number of carbon atoms for ethylcyclohexane in n-alkanes from the correlation proposed by Uhlig [108].

The obvious question is: *what does cause this effect?* The definitive answer to that question is beyond the scope of this work. However, if one considers that the above correlation of surface tension with cavity formation is correct then the smaller cavitation energy would mean than a

solute could diffuse with less effort in the solvent. If one considers that a molecule need to have a cavity (thermally) formed before it can move (or jump) within a fluid then lower cavitation energy would mean easier diffusion. So with the example of ethylcyclohexane, it can diffuse easier in, for example, pentane than octane. This easier diffusion would mean that it is simpler for ethylcyclohexane to escape from pentane than octane and perhaps this could drive up the volatility in pentane. This is however quite speculative and would require further investigation. Nevertheless since the cavitation term does seem to be (at least indirectly) involved the contribution will be denoted as a cavitation (Cav) contribution.

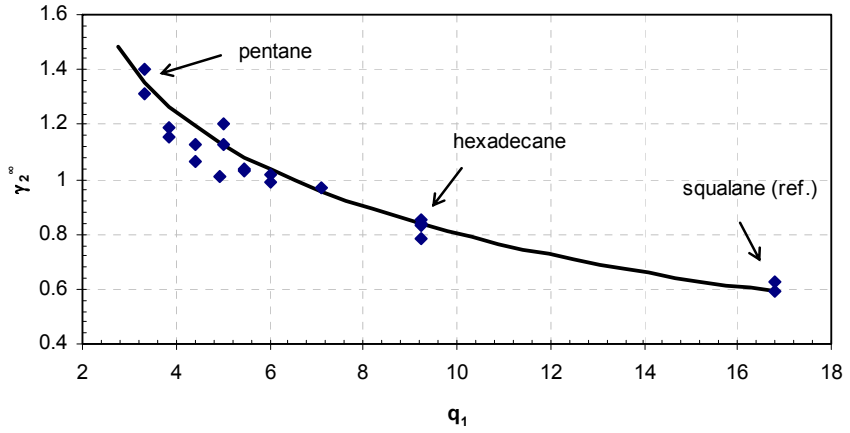


Fig. 6-11 γ_2^∞ vs. q_1 for ethylcyclohexane (2) in alkane solvents(1) using squalane as the reference solvent at 298.15 K (◆ – data from the DDB [28], ── Eqn. (6-6)).

While Eqn. (6-6) does provide better extrapolations it suffers from the fact that the absolute value is under predicted. This under prediction is due to the cavitation contribution and can be very simply (and empirically) solved by multiplying it by some factor to offset the under prediction. To ensure an activity coefficient of unity in mixtures of molecules of identical size and shape a step-like function had to be introduced:

$$\gamma_2^{Cav,\infty}\Big|_{corrected} = \gamma_2^{Cav,\infty} \times \left(1.15 - 0.15 \times \exp\left(-0.5\left(q_1 - q_2\right)^2 - 0.5\left(r_1 - r_2\right)^2\right)\right) \qquad (6\text{-}13)$$

Any reference to Eqn. (6-6) (and to a 'new combinatorial') from this point forwards implies the use of Eqn. (6-13) for the cavitation contribution. Results for this new expression are shown in Fig. 6-12 and Fig. 6-13. Due to the correction by Eqn. (6-13), the calculated curve now shows a local minimum and maximum, which is not in contradiction to the experimental results but should nevertheless be backed up by a molecular based interpretation. Further results for cases where the solute size exceeds that of some or all of the solvents can be found in section 6.1.3.

It should be noted at this point that both the Flory-Huggins and Guggenheim-Staverman expressions were developed on a simplified but sound theoretical basis and both predict a

maximum of the combinatorial contribution in the case of equal molecular size. reproduce the intuitive assumption that the more the molecules differ in size, the stronger the increase of entropy upon mixing and the lower the activity coefficient of the components (the Guggenheim-Staverman expression may predict a different behaviour for r/q ratios far away from unity, but this is usually not encountered in practical cases, see Appendix B). Both the new expression and the experimental findings contradict this behaviour in the case of large molecular weight solutes in small solvents. This is quite definitely not the result of a decreasing entropy of mixing but, as mentioned above, could be attributed to the Gibbs energy of cavity formation (albeit, possibly quite indirectly). This is well illustrated by plotting the different contributions to Eqn. (6-6) as shown in Fig. 6-14.

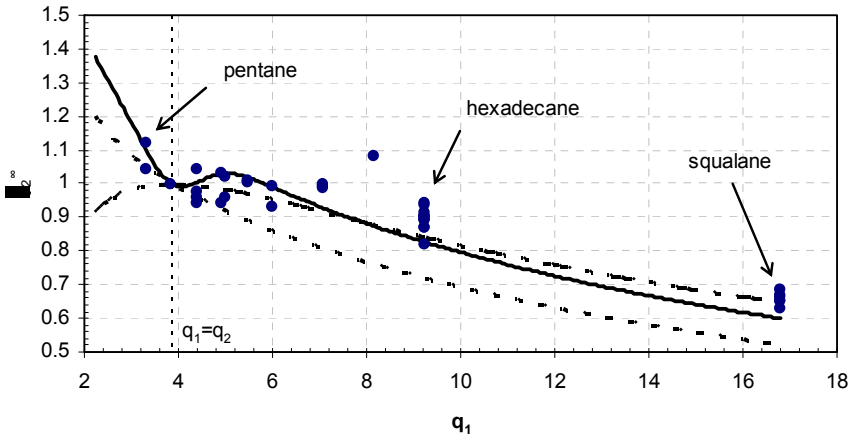


Fig. 6-12 γ_2^∞ vs. q_1 for hexane (2) in alkane solvents (1) using various combinatorial expressions (◆ – data from the DDB [28], ── Eqn. (6-6) with Eqn. (6-13), - - - - Eqn. (6-6) without Eqn. (6-13) , ·-·-·- mod. UNIFAC prediction).

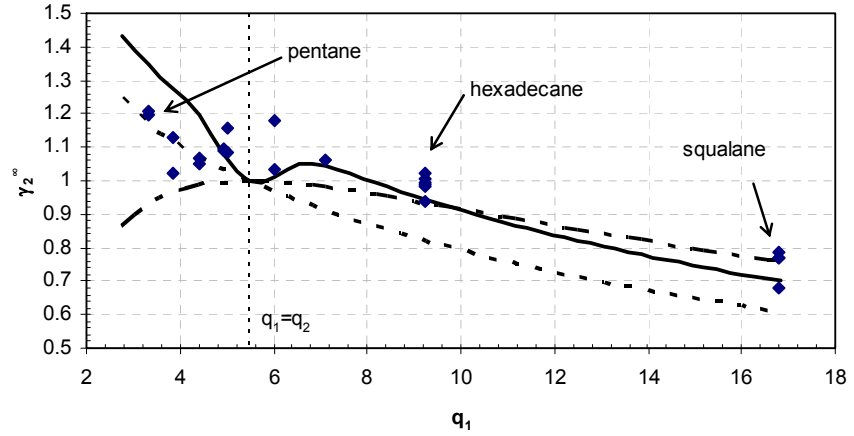


Fig. 6-13 γ_2^∞ vs. q_1 for nonane (2) in alkane solvents (1) using various combinatorial expressions (◆ – data from the DDB [28], ── Eqn. (6-6) with Eqn. (6-13), - - - - Eqn. (6-6) without Eqn. (6-13) , ·-·-·- mod. UNIFAC prediction).

Since during the development of this new combinatorial expression only alkane solvents were considered (which do not differ significantly regarding cavity formation) Eqn. (6-6) should be reliably applicable to all alkane solvents. For simplicity and historic reasons, the new expression will also be denoted as a “combinatorial” in this study although it now covers not only combinatorial effects but also some cavitation contribution.

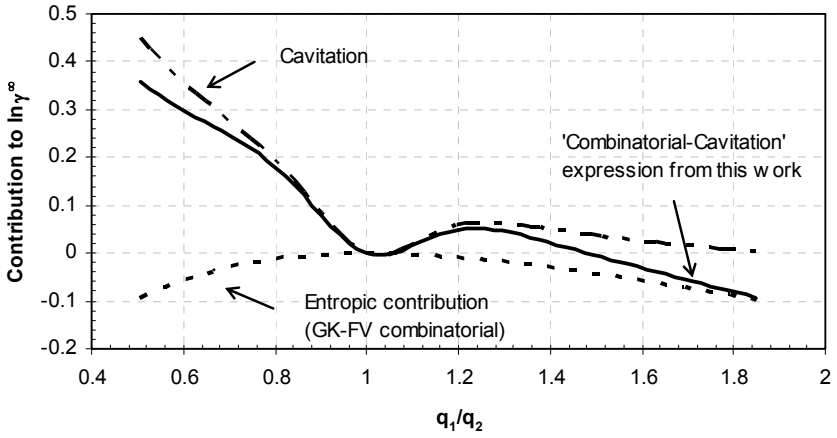


Fig. 6-14 Various contributions to the logarithm of the infinite dilution activity coefficient, as given by Eqn. (6-6) for nonane (2) in alkane solvents (1).

6.1.2. Applications to solutes with “exotic” groups

Eqn. (6-6) is useful only for compounds where the liquid molar volume; van der Waals surface area and volume are known with reasonable accuracy. This is most frequently not the case for complex multifunctional compounds like pharmaceuticals and their intermediates or biological metabolites. While testing the various combinatorial expressions it was noted that the ratio of the combinatorial activity coefficient at infinite dilution of a component in two different alkanes is well represented by the following correlation (r was found to represent the data better than q) :

$$\frac{\gamma_{i,sol1}^{C,\infty}}{\gamma_{i,sol2}^{C,\infty}} = \left(\frac{r_{sol2}}{r_{sol1}}\right)^{0.6} \tag{6-14}$$

Combining Eqn. (6-14) with Eqn. (6-2) results in:

$$\gamma_{i,sol1}^{\infty} = \gamma_{i,sol2}^{\infty}\left(\frac{r_{sol2}}{r_{sol1}}\right)^{0.6} \tag{6-15}$$

As can be seen in Fig. 6-15 , Eqn. (6-15) approximates the new combinatorial expression quite well for either small or large molecular size solutes. For solutes of similar size however this simple expression cannot provide the shape of the combinatorial expression derived

above. This could be fairly easily introduced by taking the ratio of the corrections (Eqn. (6-13)) in each solvent however this would mean that r and q values are also needed for the solute which would defeat the purpose of the correlation. The great advantage of this approach is that only the r values for the solvents are needed; no structural information about the solute is required.

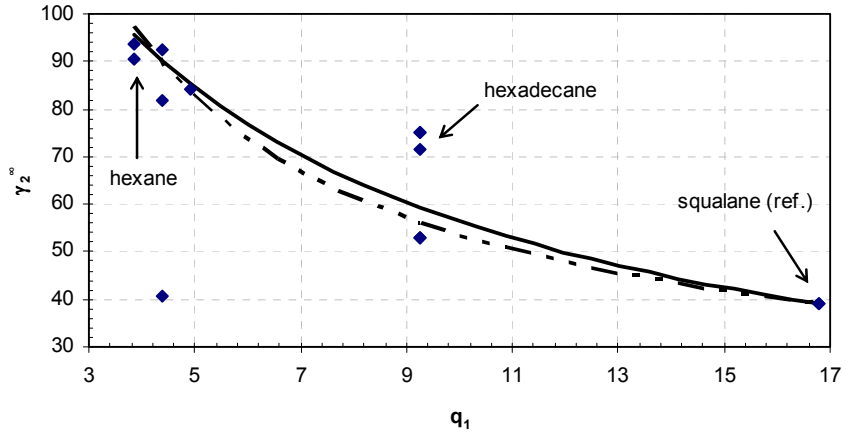


Fig. 6-15 γ_2^∞ vs. q_1 for methanol (2) in alkane solvents(1) using squalane as the reference solvent at 298.15 K (◆ – data from the DDB [28], ── Eqn. (6-6), - - - Eqn. (6-15)).

To further simplify the application of the equations, empirical correlations for the liquid molar volumes and the r and q values of the linear (and slightly branched) alkane solvents have been regressed:

$$r_1 = \frac{6.88 + 10.23 \times n_c}{15.17} \tag{6-16}$$

$$q_1 = \frac{1.54 + 1.35 \times n_c}{2.5} \tag{6-17}$$

$$V_m = 24.026 \times r_1 + 23.102 \tag{6-18}$$

where n_c is the number of carbon atoms, V_m is the liquid molar volume and subscript *1* refers to the alkane solvent.

To incorporate the new combinatorial expressions into a mixture model, values of the combinatorial contribution need to be calculated and therefore the molar volume and relative van der Waals surface and volume of the solute are required. While the van der Waals values are mostly derived from structural group values of the UNIFAC methods or from tables published by Bondi [58] the liquid molar volume may be more difficult to acquire for more complex solutes since these compounds are typically solid at system temperature. Different estimation methods are available in the literature [111-113] but may not be always applicable

to complex molecules or unavailable to the user. In these cases, the molar volume can be estimated from the van der Waals volume by the following correlation:

$$V_m = 1.681 \times V^* \tag{6-19}$$

Where V_m is the liquid molar volume and V* is the van der Waals volume. Fig. 6-16 shows the plot of V_m against V* for 1594 non-electrolyte organic compounds, V* values calculated from UNIFAC r value and Eqn's (3-60) and (3-61). This correlation is biased towards low density compounds and should be used for high density compounds with caution.

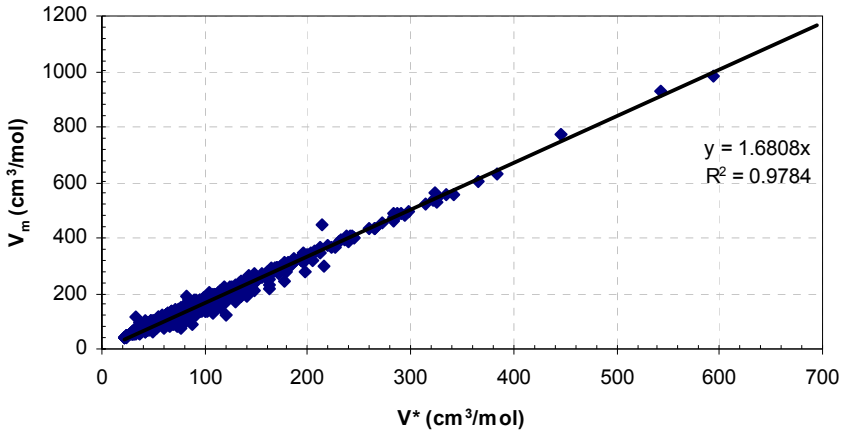


Fig. 6-16 V_m vs. V* for 1594 non-electrolyte organic compounds (◆ - data from the DDB [28], ── linear fit to the data, all molar volume data at approximately 298 K).

6.1.3. Test of the new expression

In order to ensure that the method was not only applicable to simple compounds (few functional groups) several GLC measurements were undertaken to determine the activity coefficient at infinite dilution of some complex solutes in squalane. The solutes were chosen on the basis of availability of infinite dilution data in a lower molecular weight alkane. All data could be predicted well from the new GLC data (Table 6-2).

In addition, extrapolations based on available γ^{∞} data in 19,24-dioctadecyldotetracontane (which is a C78 alkane) to lower molecular weight solvents perform very well in all cases (within approximately 5%) as shown in Table 6-2.

For all examples shown in Table 6-2, Eqn. (6-15) led to unsatisfactory results. This approach should only be applied over a smaller range of alkane solvents, which should not differ by more than 10–15 carbon atoms from the reference solvent.

Table 6-2 Alkane solvent extrapolations for various solutes in large alkane solvents (* - data measured by this author, all other data from the DDB [28], bold values denote best prediction and underline the second best in terms of relative mean deviation percentage).

	T (K)	γ_2^∞				
Solute (2) / Solvent (1) [q_1/q_2]	**(exp)**	**(exp)**	**UNIQUAC**	**mod. UNIFAC**	**Eqn. (6-6)**	**Eqn. (6-15)**
Naphthalene						
Heptane [1.278]	403.15	2.4	2.06	1.23	**2.24**	1.62
19,24-Dioctadecyldotetracontane (ref.) [12.42]	403.15	0.4	0.40	0.40	0.40	0.40
Benzene						
Octadecane [4.307]	373.15	0.89	0.64	0.95	**0.92**	0.64
Squalane [6.997]	373.15	0.58	0.48	**0.63**	0.67	0.47
19,24-Dioctadecyldotetracontane (ref.) [17.80]	373.15	0.27	0.27	0.27	0.27	0.27
Chlorobenzene						
Octadecane [3.634]	373.15	1.03	0.72	1.07	**1.06**	0.74
19,24-Dioctadecyldotetracontane (ref.) [15.02]	373.15	0.313	0.313	0.313	0.313	0.31
Bis(2,2,2-trifluoroethyl)ether						
Dodecane [1.739]	303.15	16	19.02	**15.37**	16.73	19.03
Squalane (ref.)* [4.115]	303.15	11.2	11.20	11.20	11.20	11.20
1-Bromo-2-chloro-1,1,2-trifluoroethane						
Hexane [1.217]	308.15	1.71	2.16	1.55	**1.72**	2.30
Squalane (ref.)* [5.291]	308.15	0.92	0.92	0.92	0.92	0.92
N-Methyl-2-pyrrolidone						
2-Methylbutane [1.035]	363.55	11.23	**10.82**	7.91	10.27	12.31
Squalane (ref.)* [5.248]	363.55	4.47	4.47	4.47	4.47	4.47

As the method in this work is especially aimed at predicting γ^∞ for complex, multifunctional and high boiling solutes in alkane solvents of varying molecular weight and most of these components are solid at ambient temperature, the test of the extrapolation procedure had to be extended to the dependence of solid solubilities on the relative vdW surface area of the solvent. Activity coefficients at infinite dilution were therefore derived from solid-liquid equilibrium data (Eqn. (5-18)). The resulting activity coefficient was considered to be at infinite dilution when the solubility was below 1 mole percent. In all instances the lowest molecular weight solvent was chosen as a reference solvent, since data in these solvents is much more abundant than in higher molecular weight solvents. Fig. 6-17 - Fig. 6-23 show the extrapolations for various solutes (liquid molar volumes were taken from Eqn. (6-19)).

Of the three literature combinatorial expressions tested the GK-FV expression far outperformed the ones of UNIQUAC and mod. UNIFAC so for lucidity only the GK-FV expression is shown. Quite surprisingly, both Eqn. (6-6) and (6-15) consistently over-predicted γ_2^∞ in case of testosterone propionate. The most likely reason is that the value for γ_2^∞ in the solvent hexane is erroneous. A second calculation using Eqn. (6-6) with hexadecane as reference solvent leads to a much better description of all data except for the value in hexane (Fig. 6-18). For 2-hydroxy benzoic acid, Eqn. (6-6) over-predicts to the point of being worse than the GK-FV expression. As carboxylic acids are nearly completely

dimerized in apolar solvents like alkanes, the calculation was repeated after doubling the values for V_m, r and q of 2-hydroxy benzoic acid to represent the molecular parameters of the dimer (Fig. 6-20). This greatly improved the results except for the value in the solvent hexadecane.

Fig. 6-21 is an excellent example of the application of the combinatorial expression extrapolation, the data seems to be scattered at first view but are exceptionally well reproduced by the corrected combinatorial expression.

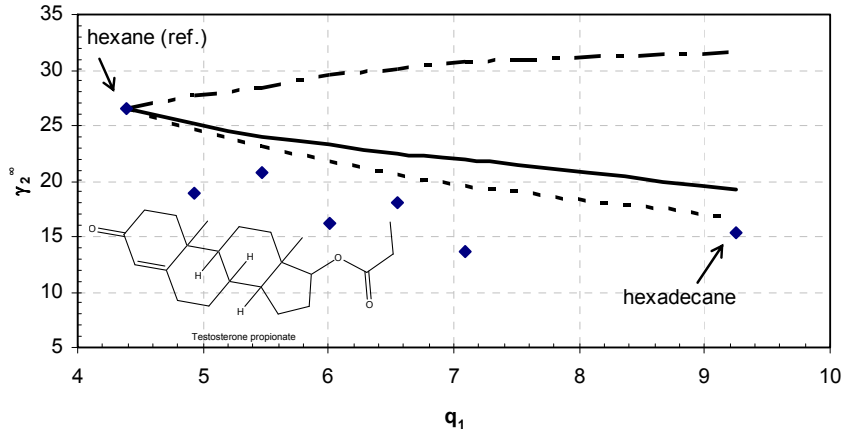


Fig. 6-17 γ_2^∞ vs. q_1 for testosterone propionate (2) in alkane solvents (1) using hexane as the reference solvent at 298.15 K (◆ – data extracted from SLE data from DDB [28], ── Eqn. (6-6), - - - - Eqn. (6-15), - · - · - GK-FV combinatorial).

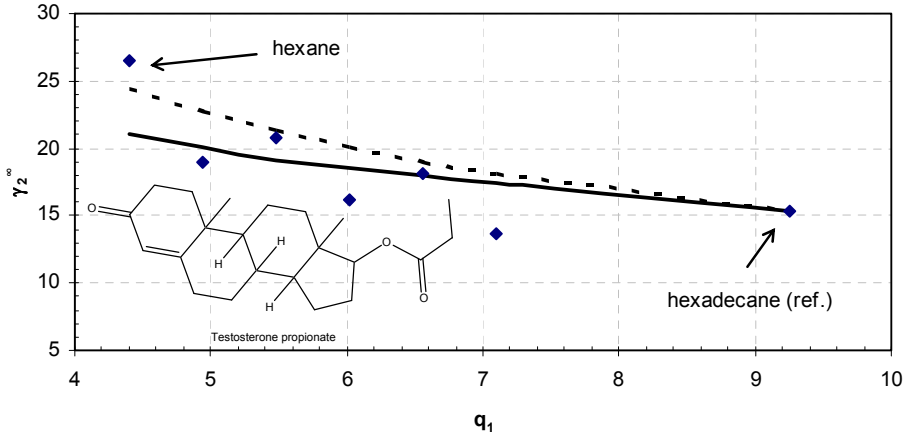


Fig. 6-18 γ_2^∞ vs. q_1 for testosterone propionate (2) in alkane solvents (1) using hexadecane as the reference solvent at 298.15 K (◆ – data extracted from SLE data from DDB [28], ── Eqn. (6-6), - - - - Eqn. (6-15)).

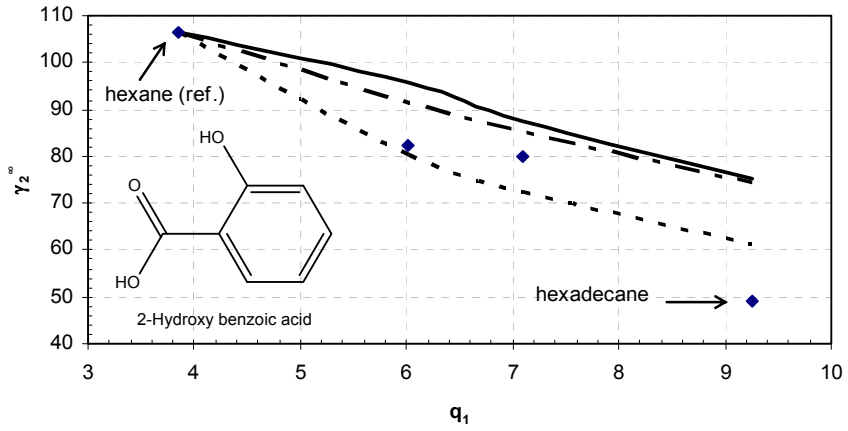


Fig. 6-19 γ_2^∞ vs. q_1 for 2-hydroxy benzoic acid (2) in alkane solvents (1) using hexane as the reference solvent at 298.15 K (◆ – data extracted from SLE data from DDB [28], ── Eqn. (6-6), - - - - Eqn. (6-15), -·-·- GK-FV combinatorial).

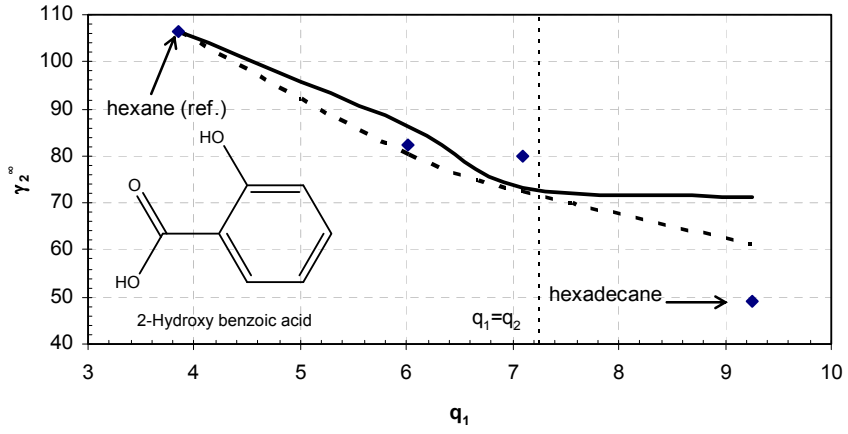


Fig. 6-20 γ_2^∞ vs. q_1 for dimer of 2-hydroxy benzoic acid (2) in alkane solvents (1) using hexane as the reference solvent at 298.15 K (◆ – data extracted from SLE data from DDB [28], ── Eqn. (6-6), - - - - Eqn. (6-15)).

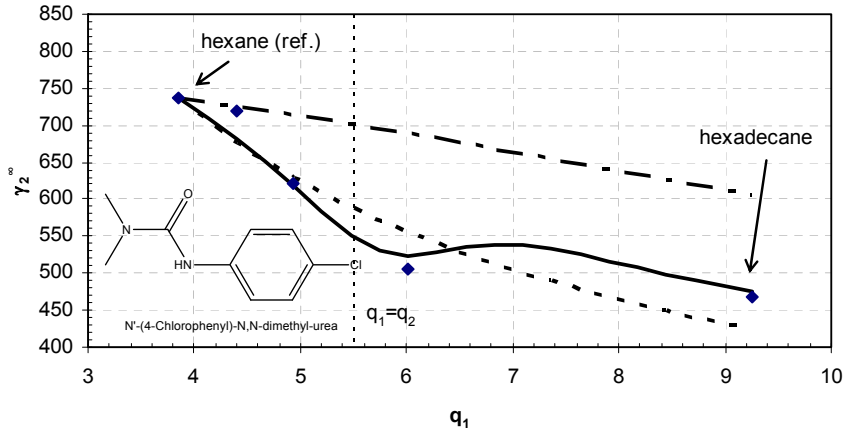


Fig. 6-21 γ_2^∞ vs. q_1 for monuron (2) in alkane solvents (1) using hexane as the reference solvent at 298.15 K (◆ – data extracted from SLE data from DDB [28], ── Eqn. (6-6), - - - - Eqn. (6-15), -·-·- GK-FV combinatorial).

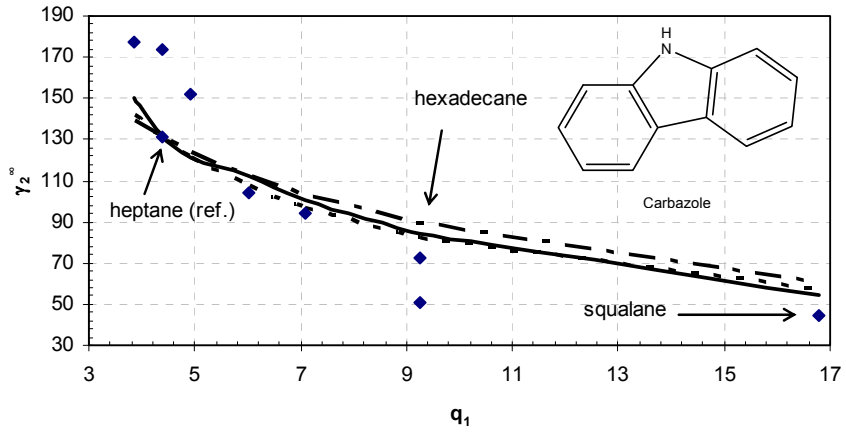


Fig. 6-22 γ_2^∞ vs. q_1 for carbazole (2) in alkane solvents (1) using heptane as the reference solvent at 298.15 K (◆ – data extracted from SLE data from literature [114], ── Eqn. (6-6), - - - - Eqn. (6-15), - · - · - GK-FV combinatorial).

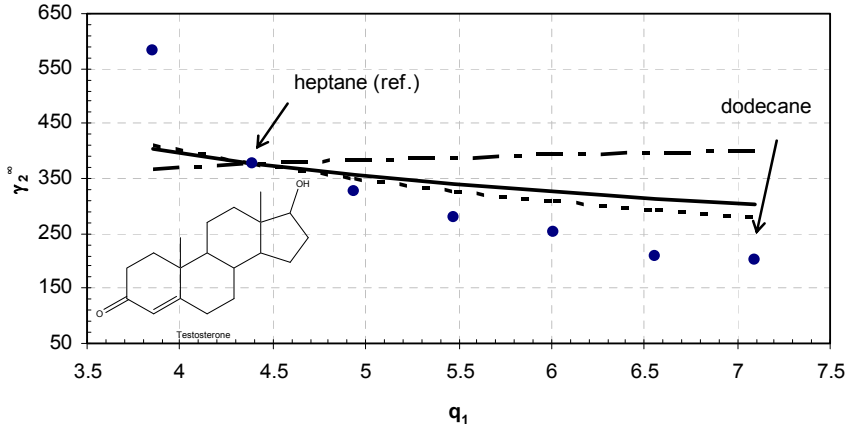


Fig. 6-23 γ_2^∞ vs. q_1 for testosterone (2) in alkane solvents (1) using heptane as the reference solvent at 298.15 K (◆ – data extracted from SLE data from literature [114], ── Eqn. (6-6), - - - - Eqn. (6-15), - · - · - GK-FV combinatorial).

As a test for the application of the absolute value of Eqn. (6-6) to solid solubilities, Eqn. (6-6) was used to predict the infinite dilution activity coefficient for n-hexatriacontane (a 36 carbon n-alkane, which is one of the few alkanes in the test as it has a solubility of less than 1 mole percent in a number of alkane solvents) as shown in Fig. 6-24. For n-hexatriacontane the heat of fusion from the DDB is 89.2 kJ/mol however there is a solid transition point very close to the melting point (within 4 K) with an enthalpy of transition of 30 kJ/mol therefore the heat of fusion was assumed to be 119.2 kJ/mol. Both Eqn's. (6-6) and (6-15) are in agreement with the data from [115] while the data from [116] are quite closely matched by the GK-FV combinatorial. The data from [116] lead to activity coefficients at infinite dilution below unity, which is in contradiction to all cases discussed above. It seems very probable that the data from this source is erroneous.

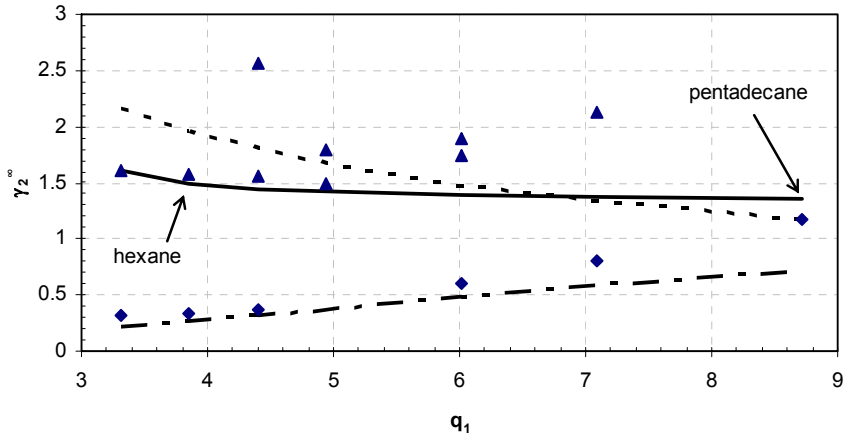


Fig. 6-24 γ_2^∞ vs. q_1 for n-hexatriacontane (2) in alkane solvents (1) at 298.15 K (◆ – data extracted from SLE data from literature [116], ▲ – data extracted from SLE data from literature [115] , ── Eqn. (6-6), - - - - Eqn. (6-15), –·–·– GK-FV combinatorial).

As a final test of the new combinatorial some solubility data of monomers (or very short polymers) in polymers were predicted (these types of systems are athermal and therefore the residual contribution to the activity coefficient falls away). The results for the prediction are shown along with the mod. UNIFAC and GK-FV predictions in Table 6-3. It is interesting to note that even though the GK-FV expression was developed specifically for this type of data the expression developed here (Eqn. (6-6)) outperforms it in all but 2 cases.

Table 6-3 Comparison of the experimental and predicted activity coefficients for some polyethylene (PE) - alkane solutions (data from [117], the number in brackets is the molecular weight of the polymer, Eqn's (6-16) , (6-17) and (6-18) were used to calculate the volume and surface parameters needed).

Polymer (2)	Solvent (1)	γ_1^∞	Eqn. (6-6)	GK-FV	Mod. UNIFAC
PE(35000)	Heptane	0.01489	**0.0145**	0.013	0.0363
PE(7400)	Decane	0.075	0.0794	**0.0741**	0.1417
PE(7400)	Dodecane	0.0882	**0.0893**	0.0837	0.159
PE(84000)	Decane	0.00748	**0.00764**	0.00676	0.024
PE(35000)	Octane	0.01583	**0.0157**	0.014	0.0396
PE(15000)]	Octane	0.02818	0.0356	**0.0324**	0.0738
PE(105000)	Nonane	0.0058	**0.00573**	0.00508	0.019
PE(105000)	Decane	0.006167	**0.00613**	0.00541	0.0203

6.1.4. Extension from infinite dilution

All the discussion up to this point has focused on the application to infinite dilution of the solute. For a combinatorial expression to be widely applicable it must obviously be usable for finite concentrations as well. The cavitation term can readily be extended to finite concentrations but the resulting expression is not thermodynamically consistent. However since the endpoint (infinite dilution) values are known (for alkane mixtures) it is possible to fit this expression to some 2 parameter thermodynamically consistent expression. The Non-Random Two Liquid (NRTL, Eqn. (3-41)) expression was chosen since the third adjustable

parameter is convenient in that it allows the equation to be suitably tuned. The resulting equations which need to be fitted are therefore:

$$\ln \gamma_i^{Cav,\infty} = \tau_{ji} + \tau_{ij} G_{ij} = \left(\frac{V_i^{FV}}{V_j^{FV}} - \frac{V'^{FV}_i}{V'^{FV}_j} \right) \tag{6-20}$$

$$G_{ij} = \exp\left(-\alpha.\tau_{ij}\right) \tag{6-21}$$

where τ_{ij} are the fitted parameters, and α is fixed at a value of 10. The large value of α means that the effect of the cavitation term is most noticeable at infinite dilution and much less pronounced (almost negligible) at higher concentrations. This results in the following expression for finite concentrations:

$$\ln \gamma_i^{C-Cav} = \ln \gamma_i^{C} + \ln \gamma_i^{Cav} \tag{6-22}$$

where $\ln \gamma_i^{C}$ is given by Eqn. (3-131). This expression can be used to predict (when the pure component vapour pressures are known) the P-x diagram of athermal systems, one such example is shown in Fig. 6-25. It is clear that the predictions of Eqn. (6-22) are superior to those of the UNIFAC combinatorial (for the sake of lucidity the mod. UNIFAC predictions are not shown, the results are similar to those of Eqn. (6-22)) but this is primarily due to the combinatorial contribution in Eqn. (6-22). The problem inherent with all VLE examples of alkane systems is that to investigate the effect of the new combinatorial, sufficiently asymmetric systems are needed (since this is when the activity coefficient is noticeably different) however in these systems the effect of the activity coefficient is severely damped by the lower vapour pressure of the larger compound. This effect of the larger compound, while small, is still noticeable at lower concentrations of the larger compound (where the cavitation term starts to become significant) as shown in the exploded view in Fig. 6-25.

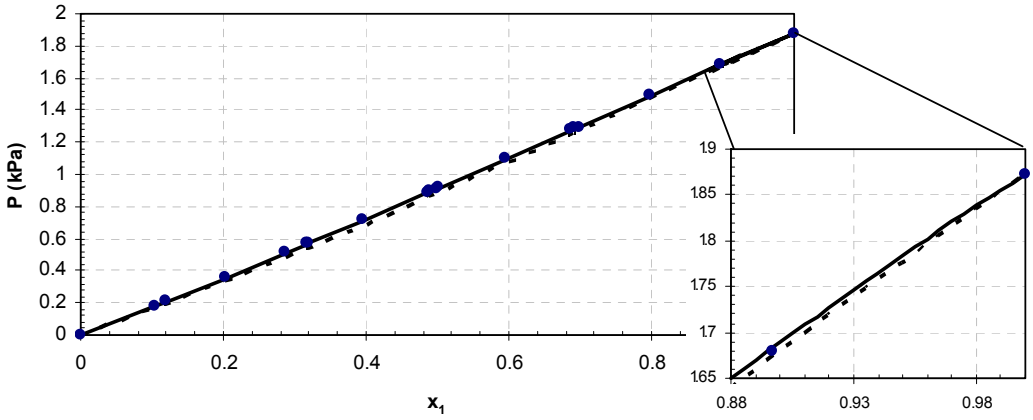


Fig. 6-25 P-x diagram for the system octane (1) – hexadecane (2) (● – data from the DDB [28], —— Eqn. (6-22), ---- UNIQUAC combinatorial).

Even with the exploded view in Fig. 6-25 the difference between the activity coefficients from the 2 expressions in unclear. This difference can be made clearer by plotting the ratio of the activity coefficient as shown in Fig. 6-26. As mentioned above, at low concentrations of the larger compound the difference between the 2 curves is drastic. The shape of the curve proposed by this work seems strange, however since at $x_1 = 0$ and $x_1 = 1$ $\ln(\gamma_1/\gamma_2)$ is negative, the Gibbs-Duhem* equation [118] dictates that there must be a maxima in the curve.

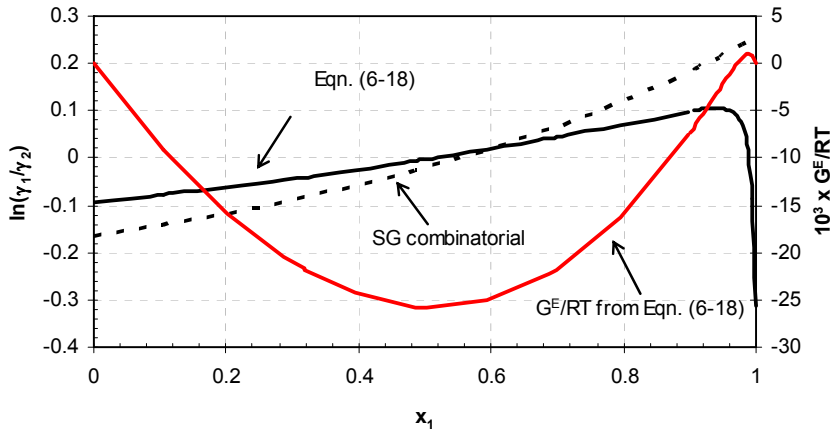


Fig. 6-26 $\ln(\gamma_1/\gamma_2)$ vs. x_1 for the system octane (1) – hexadecane (2), comparison of the SG combinatorial and the combinatorial from this work as given by Eqn. (6-22) (── Eqn. (6-22), - - - - SG combinatorial, ── G^E/RT) .

6.1.5. Conclusion

A method has been developed which allows one convert the infinite dilution activity coefficient in one non-cyclic alkane solvent to the corresponding value in any other non-cyclic alkane based only on the liquid molar volumes, van der Waals volumes and surface areas of the two solvents. For cases where these data are not available, convenient correlations have been developed. This is especially useful for example with GLC (Gas Liquid Chromatography) where the non-polar stationary phase is very often squalane or some other large alkane (e.g. 19,24-dioctadecyldotetracontane or hexadecane). Therefore measurements can be carried

---

* The Gibbs-Duhem expression for the binary case is given as follows [118]:

$$x_1 d\ln\gamma_1 + x_2 d\ln\gamma_2 = 0$$

Some rearrangement yields the following [118]:

$$\int_0^1 \ln\frac{\gamma_1}{\gamma_2} dx_1 = 0$$

what this expression basically means is that the area between the curve $\ln(\gamma_1/\gamma_2)$ and the x-axis must sum to zero. This is commonly used as a consistency test for low pressure isothermal data. It is clear that since both the SG combinatorial and the NRTL are thermodynamically consistent that the plots in Fig. 6-26 obey the Gibbs-Duhem expression. Therefore since $\ln(\gamma_1/\gamma_2)$ is negative for both $x_1 = 0$ and $x_1 = 1$, the curve has to go above the x-axis and then dip below again, as shown in Fig. 6-26.

out with relative ease on the low volatility stationary phase and extrapolated to a more volatile solvent, thereby circumventing the experimental difficulties of measurements in volatile solvents.

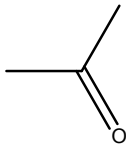
It has been shown that, perhaps contrary to common perception, free-volume does play a role in the determination of the combinatorial contribution and should be accounted for. It could also be shown, that for mixtures in which the molecular size of the solute is sufficiently larger than that of the alkane solvent, the athermal part of a G^E model, which usually only contains an entropic contribution, shows a qualitatively wrong behavior, i.e. γ^{∞} is decreasing with solute size instead of increasing as seen in the different experimental data. The source of this deviation is uncertain but possibly due to the Gibbs energy of cavity formation. The new expression has been verified against numerous experimental data and found, at least qualitatively, to be successfully applicable in every case.

6.1.6. Example calculation

Table 6-4 shows an example calculation of the extrapolation of the infinite dilution data from one solvent to another.

Table 6-4 Example calculation for acetone with alkane solvents.

Solute(2) :



acetone

Temperature : 298.15 K

Solvent(1)	* $\gamma_2^{\infty,C}$	γ_2^{∞} (exp.)
Heptane (1a)	0.8578	7.4
Hexadecane (1b)	0.5969	5.48
Squalane (1c)	0.4080	3.7

* - from Eqn. (6-6)

Using squalane as the reference solvent for heptane:

$$\gamma_{2,1a}^{\infty} = \frac{\gamma_{2,1a}^{\infty,C}}{\gamma_{2,1c}^{\infty,C}} \gamma_{2,1c}^{\infty}$$
$$= \frac{0.8578}{0.4080} \times 3.7$$
$$= 7.78$$

Which is within 5% off the experimental value, similarly for hexadecane, with squalane as the reference:

$$\gamma_{2,1b}^{\infty} = \frac{\gamma_{2,1b}^{\infty,C}}{\gamma_{2,1c}^{\infty,C}} \gamma_{2,1c}^{\infty}$$
$$= \frac{0.5969}{0.4080} \times 3.7$$
$$= 5.41$$

Which is within 2% off the experimental value

6.2. Non-alkane 'monofunctional' homologous series'

As the title suggests this section covers discussion pertaining to interpolation within non-alkane monofunctional homologous series'. A monofunctional homologous series refers to a series with only one other type (other than the CH_2/CH_3 group) of functional group. Examples of such series' are alcohols, ketones and carboxylic acids.

In the previous section a simple method was proposed for the extrapolation of γ^{∞} data in one non-cyclic alkane solvent to any other. For practical applications it is clear that alkane solvents are not the only solvents of interest and therefore it would be useful if there was a similar procedure available for other solvent families. The method for the alkanes worked because, at infinite dilution, from a solution of groups standpoint, the residual contribution for a single solute is the same in any alkane solvent (since the solution contains only CH_2/CH_3 groups). This is obviously not generally true for other homologous series, since depending on the solvent the ratio of the number of the functional groups to alkane groups changes and therefore the residual contribution will change.

However, as previously, we can very simplistically apply the solution of groups concept. If we know how the solute behaves in a (hypothetical) solution of CH_2 groups and *fun* groups (where *fun* is some functional group e.g. OH, CO etc.) and we know the ratio of the frequency of these groups in any solvent, it should be possible to predict the behaviour in any solvent in the series. It is intuitive that the behaviour that must be known in each hypothetical solution is a residual contribution and therefore expressing this mathematically:

$$\ln \gamma_{i,j}^{\infty} = a_{j,fun} \ln \gamma_{i,fun}^{R,\infty} + a_{j,CH_2} \ln \gamma_{i,CH_2}^{R,\infty} + \ln \gamma_{i,j}^{C} \tag{6-23}$$

$$\gamma_{i,j}^{\infty} = \left(\gamma_{i,fun}^{R,\infty}\right)^{a_{j,fun}} \left(\gamma_{i,CH_2}^{R,\infty}\right)^{a_{j,CH_2}} \left(\gamma_{i,j}^{C,\infty}\right) \tag{6-24}$$

where the subscript *i* refers to the solute, *j* refers to the solvent, *fun* refers to the functional group and *CH_2* refers to the hydrocarbon contribution. The coefficients $a_{j,fun}$ and a_{j,CH_2} are the surface contributions of the functional group and the CH_2 group for the solvent respectively. These values should be readily available or easily calculated for each solvent in the homologous series. When testing this expression on some data available for various homologous series' in the DDB [28] it soon became clear that there was absolutely no advantage gained by including the combinatorial expression in Eqn. (6-24). It is assumed that the combinatorial expression is somehow absorbed into the model parameters. Since the inclusion of the combinatorial expression makes it less generally applicable it was decided to use the equation without the combinatorial part. Therefore it was sufficient to use the following expression:

$$\ln \gamma_{i,j}^{\infty} = a_{j,fun} \ln \gamma_{i,fun}^{\infty} + a_{j,CH_2} \ln \gamma_{i,CH_2}^{\infty} \tag{6-25}$$

$$\gamma_{i,j}^{\infty} = \left(\gamma_{i,fun}^{\infty}\right)^{a_{j,fun}} \left(\gamma_{i,CH_2}^{\infty}\right)^{a_{j,CH_2}} \tag{6-26}$$

It is clear that since there are 2 functional groups (i.e. fun and CH_2) there will need to be 2 reference solvents in order to calculate the hypothetical infinite dilution activity coefficient in each solvent group ($\gamma_{i,fun}^{\infty}$ & γ_{i,CH_2}^{∞}).

6.2.1. Alcohols

For the (non-aromatic) alcohols Eqn. (6-26) becomes:

$$\gamma_{i,j}^{\infty} = \left(\gamma_{i,OH}^{\infty}\right)^{a_{j,OH}} \left(\gamma_{i,CH_2}^{\infty}\right)^{a_{j,CH_2}} \tag{6-27}$$

The surface contributions for the alcohols are specified in Table 6-5 (water and hexane are included, the reasons are discussed below). Therefore, if γ^{∞} data in 2 alcohol solvents of sufficiently different size (so as to be able to glean sufficient information about the OH and CH_2 hypothetical solution behaviour) are known then the behaviour in the other alcohols can be predicted. Some examples are shown in Fig. 6-27 to Fig. 6-31. Typically the best results are obtained when interpolating (i.e. using the biggest, e.g. n-decanol, and smallest, e.g. methanol, alcohol in the dataset), extrapolations can be attempted but should be used with caution as they are very dependent on the accuracy of the fitted data. The reason for this is that the smallest and largest (in terms of CH_2 surface) contain the most information about the OH and CH_2 surface respectively.

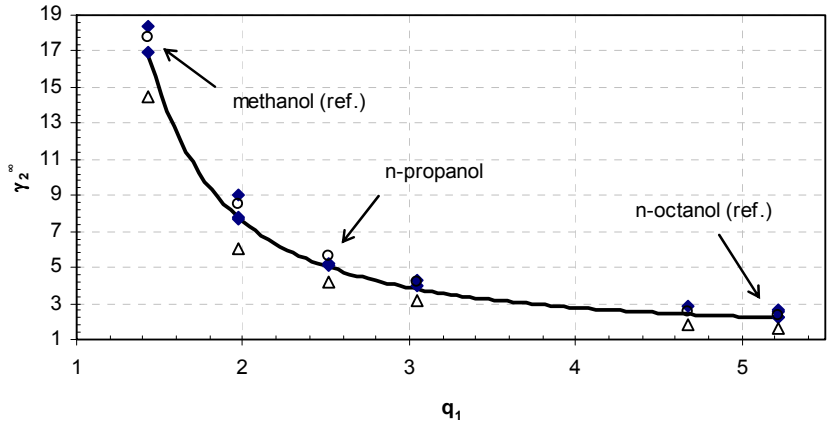


Fig. 6-27 γ_2^{∞} vs. q_1 for pentane (2) in alcohol solvents(1) at 298 K (◆ – data from the DDB [28], ── Eqn. (6-27), Δ – UNIFAC prediction, o – mod. UNIFAC prediction).

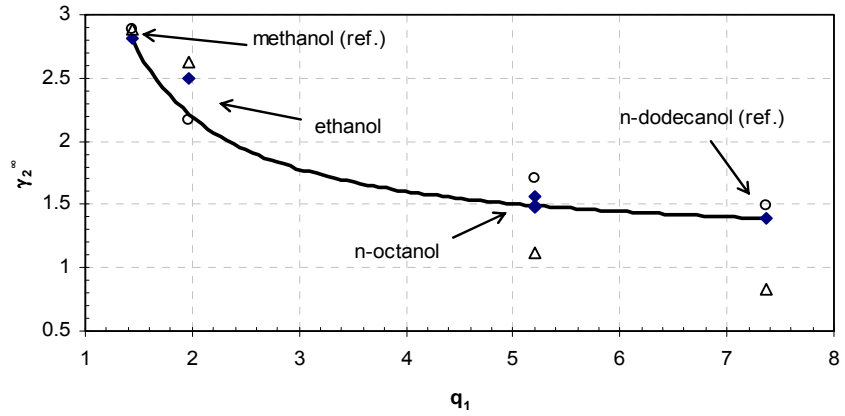


Fig. 6-28 γ_2^∞ vs. q_1 for dichloromethane (2) in alcohol solvents(1) at 298 K (◆ – data from the DDB [28], ── Eqn. (6-27), Δ – UNIFAC prediction, o – mod. UNIFAC prediction).

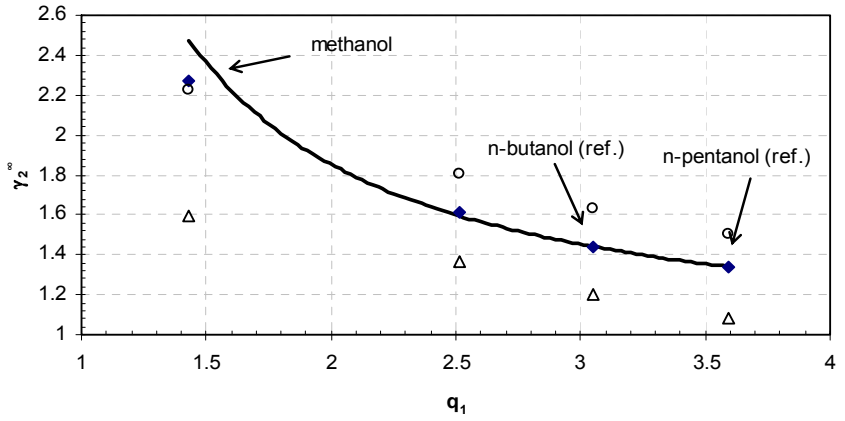


Fig. 6-29 γ_2^∞ vs. q_1 for dimethyl ether (2) in alcohol solvents(1) at 330 K (◆ – data from the DDB [28], ── Eqn. (6-27), Δ – UNIFAC prediction, o – mod. UNIFAC prediction).

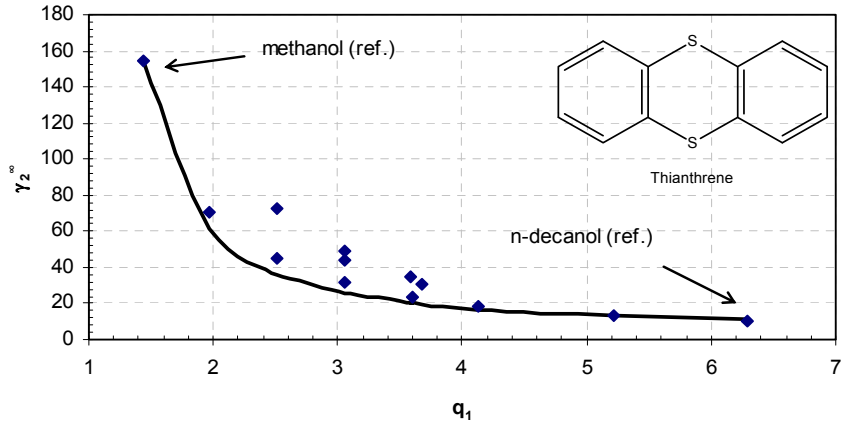


Fig. 6-30 γ_2^∞ vs. q_1 for thianthrene (2) in alcohol solvents(1) at 298 K (◆ – data extracted from SLE data from DDB [28], ── Eqn. (6-27)).

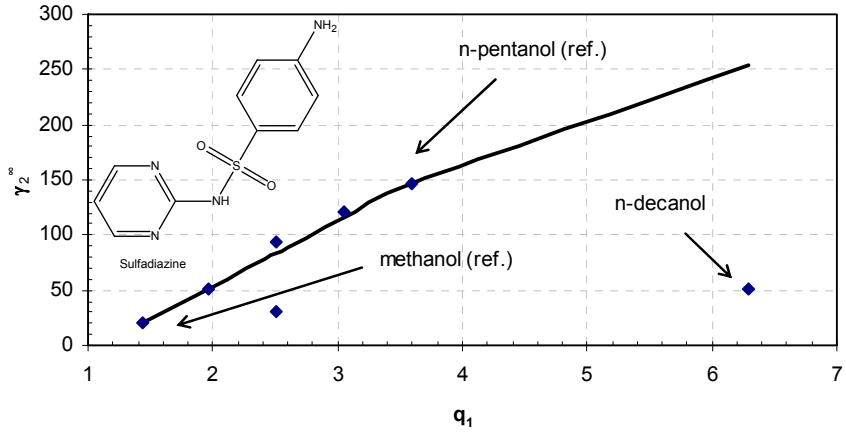


Fig. 6-31 γ_2^{∞} vs. q_1 for sulfadiazine (2) in alcohol solvents(1) at 298 K (◆ – data extracted from SLE data from literature [119], ── Eqn. (6-27)).

Table 6-5 Surface contributions for (non-aromatic) alcohol solvents (j), q_{OH} = 0.584.

Solvent	$a_{j,OH}$	a_{j,CH_2}
Water	1.7	0
Non-cyclic mono alcohols	q_{OH} / q_j	$1 - a_{j,OH}$
Non-cyclic diols	$3 \times q_{OH} / q_j$	$1 - a_{j,OH}$
Cyclic non-aromatic alcohols	$1.3 \times q_{OH} / q_j$	$1 - a_{j,OH}$
Hexane	0	2.2

While the interpolation of γ^{∞} alcohol data between alcohol reference solvents does have some applications, the applications are fairly limited due to the need for 2 reference alcohol solvents, for which one may expect data to be scarce. It would therefore be useful to extend this approach to other solvents.

When considering the strict definition of alcohols as a homologous series, the smallest alcohol would be methanol. If a CH_2 group is added to methanol then ethanol results and if a CH_2 group is removed then water would be left. The uninitiated may therefore expect that the change in physical properties from ethanol to methanol is the same as that of methanol to water. This is clearly not true (as one example consider Fig. 6-32) due to strong hydrogen bonding present in water. Nevertheless water should represent some kind of analogue to a hypothetical solution of OH groups. Therefore it should be possible to somehow estimate the behaviour in the hypothetical OH solution from the behaviour in water. This can be achieved in Eqn. (6-27) by changing the value of $a_{j,OH}$. Initially a value of 2.4 was used (since $q_{H_2O} / q_{OH} = 2.4$ when using the UNIQUAC value for q_{H_2O}) however this value didn't provide any reasonable interpolations when using Eqn. (6-27). Similarly when using a value of 1.55 (since $q_{H_2O} / q_{OH} = 1.55$ when using the q_{H_2O} value given by Voutsas et al. [102]) reasonable interpolations could not be found. A value of 1.7 was found to provide the most satisfactory results. Similarly one would expect large alcohols (monofunctional) to behave in a similar way to the corresponding, structurally similar alkane (since $a_{j,OH} \to 0$, as an analogy

consider the example of the boiling points of n-alcohols and n-alkanes shown in Fig. 6-32). It was found that if hexane was assigned a surface contribution value of 2.2 the infinite dilution activity coefficient in alcohols could be fairly well approximated from only water and hexane data, as shown in Table 6-6. For compounds which are highly associating (e.g. acetic acid) this interpolation between water and hexane will fail since the γ^{∞} in water would be for the monomer while the γ^{∞} in hexane would be for the dimer and therefore any interpolation is meaningless.

Table 6-6 Interpolated and experimental data for various solutes (2) in non-aromatic alcohol solvents (1) (data from the DDB [28], * - 2 reference solvents , ** - data predicted using Eqn. (6-15)).

Solute(2)	**Ethyl iodide**				**2-Butanone**			
Temperature	**298.15 K**				**298.15 K**			
		γ_2^{pred}				γ_2^{pred}		
Solvent(1)	γ_2^{exp}	**Eqn. (6-27)**	**UNIFAC**	**Mod. UNIFAC**	γ_2^{exp}	**Eqn. (6-27)**	**UNIFAC**	**Mod. UNIFAC**
Water	2192.0	2192.0*	-	-	26.2	26.2*	32.0	24.1
1,2-Ethanediol	-	36.3	-	-	8.4	5.3	9.2	8.5
Methanol	8.0	7.5	7.2	8.6	2.4	3.4	2.2	2.3
Ethanol	5.0	4.8	5.3	5.8	2.6	3.0	2.6	2.6
1-Propanol	3.9	3.6	4.0	4.3	2.3	2.8	2.3	2.4
2-Propanol	3.8	3.6	4.0	4.2	-	2.7	2.3	2.3
1-Butanol	3.2	3.0	3.3	3.5	2.1	2.7	2.1	2.3
tert-Butanol	4.0	3.0	3.3	3.6	1.7	2.7	2.1	2.4
1-Pentanol	-	2.7	2.8	3.0	-	2.6	2.0	2.3
3-Methyl-1-butanol	2.3	2.7	2.8	3.0	1.8	2.6	2.0	2.3
1-Octanol	-	2.2	1.9	2.2	2.3	2.5	1.7	2.1
1-Dodecanol	-	1.9	1.4	1.7	-	2.3	1.5	1.9
Hexane	1.9	1.9*	1.9	2.1	5.0	5.0*	5.4	5.5
Solute(2)	**1,4-Dioxane**				**Toluene**			
Temperature	**298.15 K**				**333.15 K**			
		γ_2^{pred}				γ_2^{pred}		
Solvent(1)	γ_2^{exp}	**Eqn. (6-27)**	**UNIFAC**	**Mod. UNIFAC**	γ_2^{exp}	**Eqn. (6-27)**	**UNIFAC**	**Mod. UNIFAC**
Water	5.4	5.4*	4.9	6.8	5109.1	5109.1*	7833.5	4373.7
1,2-Ethanediol	4.8	2.5	2.4	-	50.4	52.4	50.2	47.6
Methanol	3.4	2.2	1.3	2.9	8.8	8.7	7.6	8.9
Ethanol	3.1	2.1	2.6	3.3	5.6	5.1	5.3	5.7
1-Propanol	-	2.0	2.6	2.9	-	3.8	3.8	3.9
2-Propanol	-	2.0	2.6	2.9	-	3.8	3.8	3.7
1-Butanol	2.4	2.0	2.5	2.7	-	3.1	3.0	3.0
tert-Butanol	1.6	2.0	2.4	3.1	-	3.0	3.2	3.2
1-Pentanol	-	2.0	2.5	2.5	-	2.7	2.5	2.5
3-Methyl-1-butanol	2.1	2.0	2.5	2.5	-	2.7	2.5	2.5
1-Octanol	2.1	1.9	2.3	2.1	-	2.1	1.7	1.7
1-Dodecanol	-	1.9	2.1	1.8	1.5	1.8	1.2	1.3
Hexane	3.95**	3.95*	7.0	4.7	1.56**	1.56*	1.6	1.6

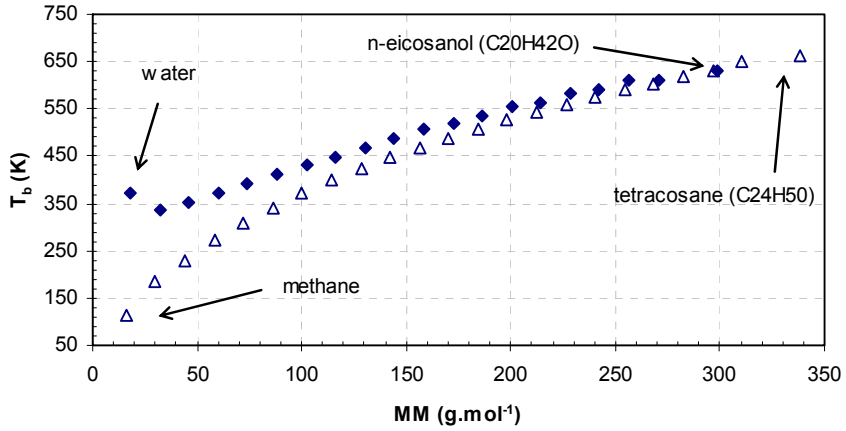


Fig. 6-32 Normal boiling points of n-alcohols (◆, including water) and n-alkanes (△) versus molar mass.

While it is possible to use data for water and hexane as solvents to predict the behaviour of the alcohols, superior representation of the alcohols is achieved if the reference solvents are both alcohols. Nevertheless if predictions can be made about the infinite dilution activity coefficient in water and hexane, then (at the least qualitative) predictions about the infinite dilution activity coefficient in alcohols can be made.

6.2.1.1. Application to mixtures

The methods developed in this section have all been for pure solvents. While these data is useful it is frequently desirable to use a cosolvent to increase the solubility (referred to as an antisolvent when used to decrease the solubility) of some solute. One would expect that the infinite dilution activity coefficient in a binary solvent (i.e. the system is ternary) would not be able to be rigorously represented by a scheme similar to Eqn. (6-27) since this equation is designed for pure solvents where the size/shape interactions do not play a role (since it is clear that the molecules are identical). Nevertheless if the correct measure of composition is selected for the mixing rule, then a similar sort of expression may represent a decent approximation. A mixing rule for the surface contribution parameters would take the following form:

$$a_{m,s} = z_{j_1}\left(a_{j_1,s}\right) + z_{j_2}\left(a_{j_2,s}\right) \tag{6-28}$$

where *s* is any surface segment (i.e. OH or CH_2), z is any composition measure (i.e. volume, mass or mole fraction), *m* is the mixture property and j_1 & j_2 are the 2 solvents (e.g. water and methanol). In order to chose a suitable measure of the composition the infinite dilution activity coefficient data should fall on (or close) to a straight line (which joins the pure component values) when plotted on the appropriate axes. A typical example of such a plot is shown in Fig. 6-33. It is clear that the mass fraction flattens the curve out enough that Eqn. (6-28) (using mass fraction) would be a suitable approximation of the mixture behaviour.

Fig. 6-34 and Fig. 6-35 show the typical application to some more complex compounds. It is interesting to note that in all cases there is a similar deviation from the ideal mixing line shown. The effect of cosolvents has received quite a large amount of attention in literature [120, 121]. However all the discussion in literature has focused on how the solubility changes with the addition of a cosolvent, and the resulting problem often ends up being a curve fitting exercise. Infinite dilution data is a more logical measure of the cosolvent effects, since the solute concentration is always assumed to be negligible and therefore there is one less variable to consider.

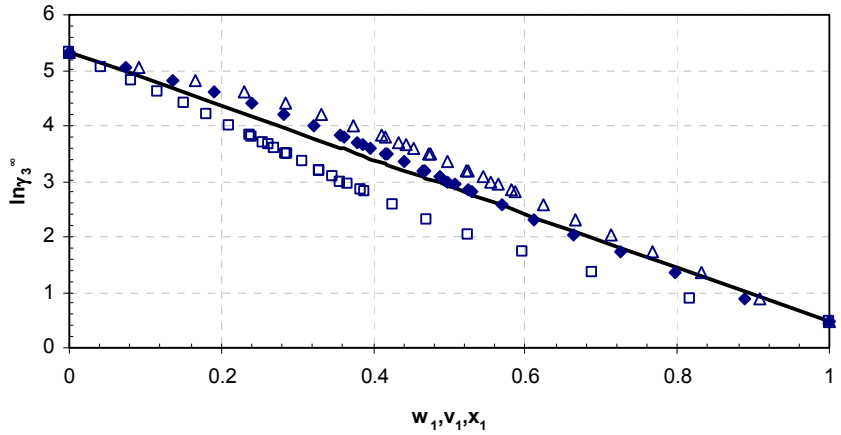


Fig. 6-33 $\ln\gamma_3^\infty$ vs. mass (w_1, ◆), volume (v_1, △) and mole (x_1, □) fraction for 3-methyl-1-butanol (3) in a mixture of water (2) and methanol (1) at 298 K (data from DDB [28], ── Eqn. (6-27) using Eqn. (6-28)).

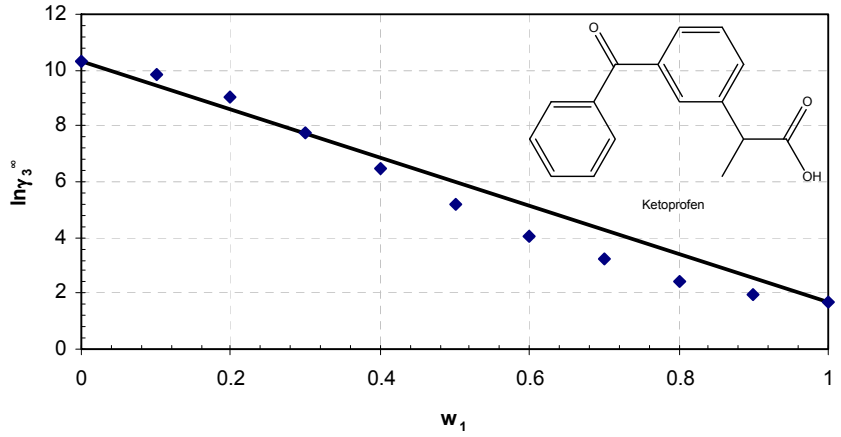


Fig. 6-34 $\ln\gamma_3^\infty$ vs. w_1 for ketoprofen (3) in a mixture of water (2) and ethanol (1) at 298 K (◆ - data extracted from SLE data from literature [122], ── Eqn. (6-27) using Eqn. (6-28)).

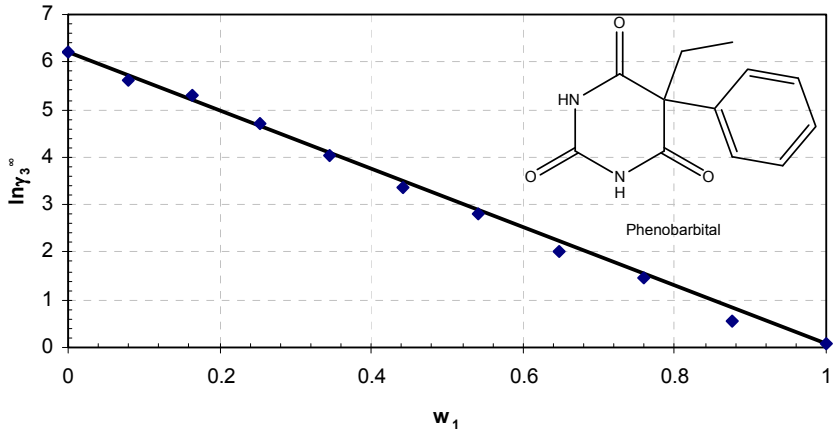


Fig. 6-35 $\ln\gamma_3^{\infty}$ vs. w_1 for phenobarbital (3) in a mixture of water (2) and propylene glycol (1) at 298 K (◆ - data extracted from SLE data from literature [123], ── Eqn. (6-27) using Eqn. (6-28)).

6.2.2. Ketones

As with the n-alcohols it was assumed that the infinite dilution of any solute (*i*) in a (non-cyclic, non-aromatic) ketone solvent (*j*) may be represented by the following equation:

$$\gamma_{i,j}^{\infty} = \left(\gamma_{i,CO}^{\infty}\right)^{a_{j,CO}} \left(\gamma_{i,CH_2}^{\infty}\right)^{a_{j,CH_2}} \tag{6-29}$$

where the superscript CO refers to the ketone contribution, $a_{i,CO}$ and a_{i,CH_2} are the surface fractions of CO and CH_2 groups in the solvent, given as: $a_{i,CO} = q_{CO} / q_i$ and $a_{i,CH_2} = (q_i - q_{CO}) / q_i$ ($q_{CO} = 0.64$ is taken from Bondi [58]). As with the n-alcohols two reference solvents are required since there are two different contributions. Unfortunately, unlike with the n-alcohols, there are not large amounts of data available for detailed model tests, Fig. 6-36 shows the example of hexane for which there was a fair amount of data available. As with the n-alcohols the interpolation is fairly good.

In all the examples that have been shown so far the infinite dilution activity coefficient decreases with the increasing size of the solute. This is, however, not necessarily the case. Fig. 6-37 and Fig. 6-38 show the curves for water and ethanol in ketone solvents; in these examples the size of the infinite dilution activity coefficient increases with increasing solvent size. The reason for this is that as the hydrophobic chain of the solvent increases the solvent becomes more hydrophobic and therefore one would expect the infinite dilution activity of hydrophilic solutes in the mixture to increase. This effect is accounted for in the model by making the hydrophobic contribution to γ_i^{∞} larger (i.e. the regressed value of γ_i^{∞,CH_2} will be larger than the value of $\gamma_i^{\infty,CO}$).

When looking at Fig. 6-38 it seems like the prediction for 2-butanone is incorrect. However looking at Fig. 6-39, the predicted point seems to follow the trend of the other data while the experimental point seems off. Therefore in this case it can be assumed that the measured value was subject to experimental error. Unlike with the alcohols no suitable surface value could be found for hexane. This method also has a much more limited scope of application since solid solubility data in 2 different ketones are fairly rare. When trying to test this method against SLE it was found that there were no compounds which had solubility data in 3 sufficiently different ketone solvents (2-heptanone and 3-heptanone are not sufficiently different).

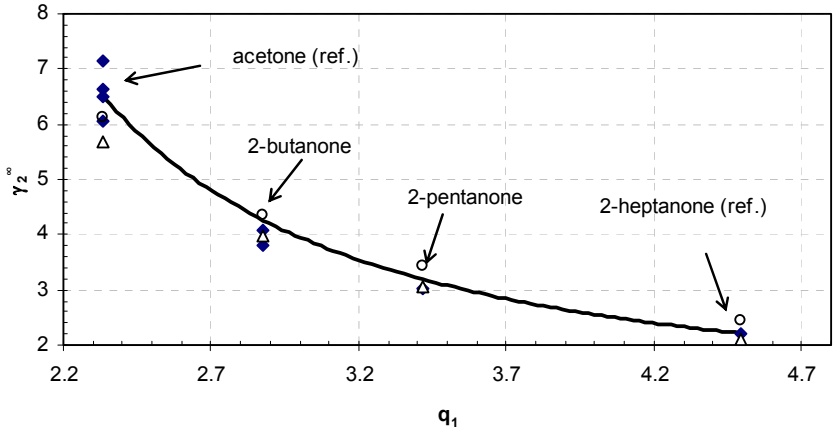


Fig. 6-36 γ_2^∞ vs. q_1 for hexane (2) in ketone solvents(1) at 298.15 K (◆ – data from the DDB [28], ── Eqn. (6-29) , Δ – UNIFAC prediction, o – mod. UNIFAC prediction).

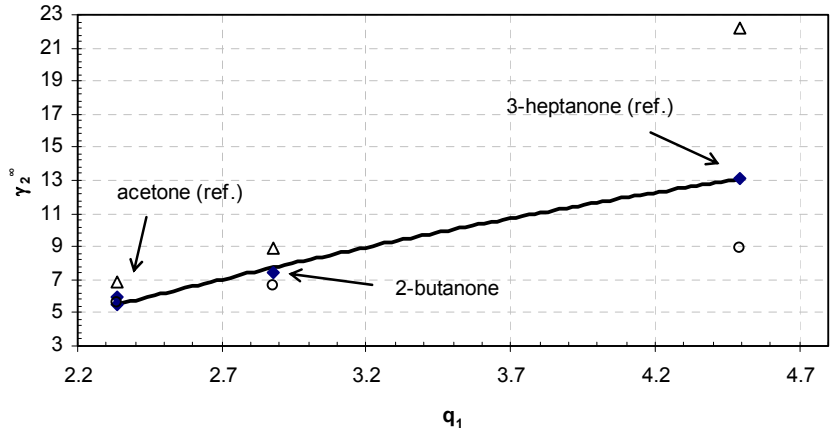


Fig. 6-37 γ_2^∞ vs. r_1 for water (2) in ketone solvents(1) at 333.15 K (◆ – data from the DDB [28], ── Eqn. (6-29), Δ – UNIFAC prediction, o – mod. UNIFAC prediction).

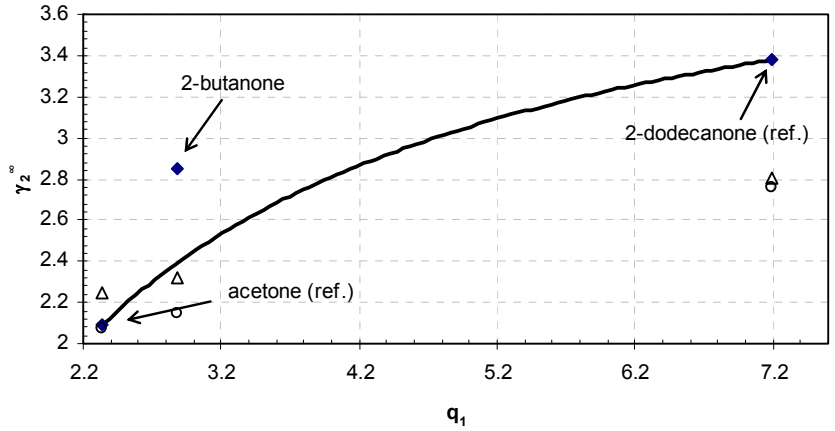


Fig. 6-38 γ_2^∞ vs. r_1 for ethanol (2) in ketone solvents(1) at 313.15 K (◆ – data from the DDB [28], ── Eqn. (6-29) , Δ – UNIFAC prediction, o – mod. UNIFAC prediction).

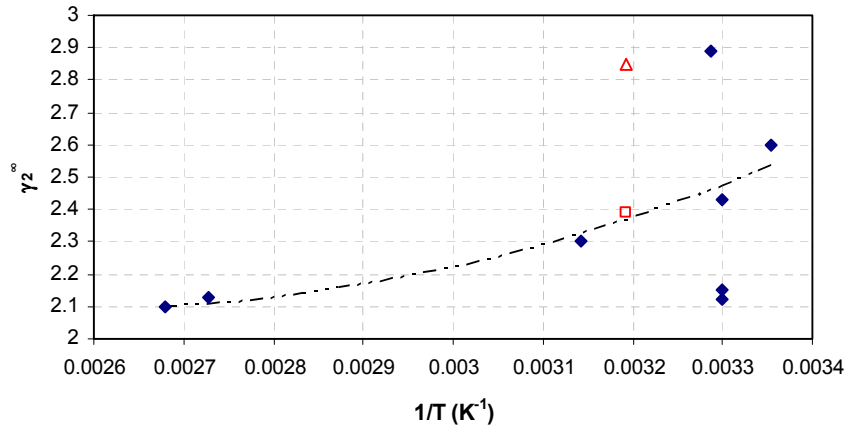


Fig. 6-39 γ_2^∞ vs. 1/T for ethanol (2) – 2-butanone (1) (◆ – data from the DDB [28], △ – experimental data at 313.15 K, □ – value predicted from Eqn. (6-29) at 313.15 K, -..-..- Eqn. (7-13) regressed to the data).

6.2.3. Conclusion

In conclusion, these methods have been developed for situations where conventional prediction methods cannot work or predictions are potentially unreliable. It is evident from the data comparisons carried out above that the UNIFAC and especially mod. UNIFAC methods are excellent for predicting the data of more simple systems. UNIFAC and mod. UNIFAC also have the huge advantage in that they don't require any experimental data. Where they are typically poor are for systems with large multifunctional molecules. These family interpolation methods are applicable to situations where 2 experimental reference points are available or suitable prediction methods are available (see section 7 and 8).

6.3. Interpolation between homologous series'

The success of the "surface contribution" methods described in the previous section suggests that the principle is fairly sound when the appropriate surface segments are chosen and when a suitable chemical family is defined. It is however not always desirable to only be able to interpolate within a single homologous series. It would therefore be useful to be able to interpolate between solvents from different homologous series'.

The simplest possible way to do this is to assume (as Hansen does – see section 3.2.3.1) that molecules are made up of only 3 surface types. These are dispersive, polar and hydrogen bonding surfaces. While this does represent an over simplification of reality the wide application and popularity of the Hansen model is testament to its usefulness. A slightly more rigorous approach is to assume that there are 2 different types of polar interactions (as NRTL-SAC does – see section 3.2.3.3). The 2 different polar surface types take into account the interactions between different polar and hydrophilic molecules. The example given by Chen [67] is acetone (which is a hydrogen bond acceptor) with water. While this approach is more rigorous it is still an over simplification as one would expect that some effects would still not be captured by this approach. Probably the most rigorous "surface contribution" approach is that applied by MOSCED (see section 3.2.3.2) where there are a dispersive, 2 polar and 2 hydrogen bonding (donor and acceptor) surfaces. These 5 surfaces enable MOSCED to accurately represent a large variety of systems with good accuracy.

The differences between these approaches can be illustrated by the examples in Table 6-7. While the examples in the table are somewhat of an unfair comparison (since Hansen and NRTL-SAC were probably never fitted to such data while MOSCED more than likely was) it is still useful. As would be expected, MOSCED is best in almost all cases. The system which really exhibits MOSCED's usefulness is chloroform-acetone. The activity coefficients for this system are much lower than one would expect due to the presence of hydrogen bonds between chloroform and acetone. The data for the system ethyl acetate – acetonitrile is well reproduced by both NRTL-SAC and MOSCED. The reason for this is that Hansen's single polar parameter probably isn't sufficient to capture all the effects present.

Table 6-7 Comparison of the Hansen (parameters taken from [61]), NRTL-SAC (parameters taken from [66]) and MOSCED (implementation by artist [28]) models for the prediction of some infinite dilution activity coefficients (Data from the DDB [28] lowest relative mean deviation in bold and second lowest underlined).

Solvent (1)	Solute (2)	T (K)	γ_2^{∞} exp.	γ_2^{∞} Hansen	γ_2^{∞} NRTL-SAC	γ_2^{∞} MOSCED
Acetonitrile	Ethyl acetate	293.15	1.58	4.10	**1.47**	<u>1.90</u>
Ethyl acetate	Acetonitrile	311.65	1.73	1.96	<u>1.53</u>	**1.73**
Acetone	Chloroform	307	0.50	1.80	<u>1.50</u>	**0.49**
Chloroform	Acetone	305	0.39	1.70	<u>1.70</u>	**0.39**
Benzene	Hexane	298.15	2.10	<u>1.80</u>	1.30	**2.00**
Hexane	Benzene	303.15	1.60	<u>1.50</u>	1.30	**1.70**

The "surface methods" mentioned above could be readily extended to a multisurface model as follows:

$$\ln \gamma_i^{\infty} = a_X \ln \gamma_i^{\infty,X} + a_{Y-} \ln \gamma_i^{\infty,Y-} + a_{Y+} \ln \gamma_i^{\infty,Y+} + a_{Z-} \ln \gamma_i^{\infty,Z-} + a_{Z+} \ln \gamma_i^{\infty,Z+} \tag{6-30}$$

$$\gamma_i^{\infty} = \left(\gamma_i^{\infty,X}\right)^{a_X} \left(\gamma_i^{\infty,Y-}\right)^{a_{Y-}} \left(\gamma_i^{\infty,Y+}\right)^{a_{Y+}} \left(\gamma_i^{\infty,Z-}\right)^{a_{Z-}} \left(\gamma_i^{\infty,Z+}\right)^{a_{Z+}} \tag{6-31}$$

where *a* refers to the surface contribution, *X*, *Y*-, *Y*+, *Z*- & *Z*+ are for dispersive, polar negative, polar positive, hydrogen bond acceptor and hydrogen bond donor respectively. This however means that there needs to be 5 reference solvents in which the solute is sufficiently dilute. This is unrealistic as one would expect that most molecules would be above the infinite dilution threshold (which, as mentioned above, is assumed to be 1 mole percent). Alternately if predictive methods could be developed in 5 representative solvents it would be possible (at least theoretically) to make predictions in any solvent for which the surface parameters are known. The problem is that there is simply not nearly enough data available in literature to develop such methods (this is the reason for looking into reducing alkane solvents to a single solvent). Therefore the focus of the methods developed in this work is for data in water and hexane. The alcohol interpolation method developed will be used to interpolate to alcohol solvents.

7. INFINITE DILUTION ACTIVITY COEFFICIENTS IN WATER

Since, by definition, infinite dilution properties do not depend on the composition of the mixture (in the binary case) it is reasonable to assume that the infinite dilution activity coefficient can be modelled in the same way as many other pure component properties [15-20, 80]. Due to the large amount of solubility data available for aqueous systems, water was the obvious choice for the testing of this idea. This idea is by no means original to this author as there are many instances of methods such as these in the literature [91, 124, 125]. However these methods are frequently QSPR (Quantitative Structure-Property Relationship) type methods or 'chemical family' type methods, where 'chemical family' type methods are those that are homologous series specific and are fitted to certain variables (e.g. MM or n_a). The problem with methods such as these is that, while they can provide very good results for a specific family or group of compounds, they are very often not generally applicable. Also many complex multifunctional organic compounds don't fall into a single family.

Another possibility of obtaining activity coefficient data is to predict the water solubility [3, 5, 126, 127] and use either experimental or predicted heat of fusion and melting temperature to extract γ^{∞} in the case where the solute is a solid at the temperature of interest. In practice this is probably not very common since experimental data for heat of fusion are not so widely available and prediction of heat of fusion and melting temperature is notoriously difficult [86, 128]. As an example a recent method for melting point temperature [129] reports an absolute error of 33.2 K which relative to other methods is considered good.

7.1. Similar methods in literature

As mentioned above the literature on predicting activity coefficients at infinite dilution in water is fairly sparse and is usually limited to QSPR or similar methods (the methods for predicting activity coefficient discussed in section 3.2 will not be detailed here). The literature is however full of methods for predicting the aqueous solubility [3-6, 8, 126]. These methods frequently use the prediction of the activity coefficient via group contribution and some other approximations for the heat of fusion and experimental value for the melting point. The method in literature which seems to have generated the most attention [130] is AQUAFAC [4, 6, 8]. The basic premise of AQUAFAC is that the activity coefficient is assumed to be made up of the sum of group contributions as follows:

$$\log \gamma_w = \sum_i n_i q_i \qquad (7\text{-}1)$$

where n_i is the frequency of the group and q_i is the group contribution value. In the AQUAFAC papers [4, 6, 8] no mention is made about the concentration dependence but this is somewhat

implied in the restrictions placed on the solubility (all data with solubility greater than 1M were rejected). Since only poorly soluble solutes are considered the following relation applies:

$$\log S_w = \log(55.5) - \log \gamma_w - \frac{\Delta_m S(T_m - T)}{2.303RT} \tag{7-2}$$

where the $\log(55.5)$ comes from the conversion of mole fraction to mole per litre solubility (the molarity of water is 55.5 mol/l), R is the ideal gas constant and $\Delta_m S$ is the entropy of melting. The entropy of melting is calculated by the following approximation [129]:

$$\Delta_m S = 50 - R\ln\phi + R\ln\sigma \tag{7-3}$$

where σ is the molecular symmetry and ϕ is the molecular flexibility which is defined as:

$$\phi = 2.435^{[SP3+0.5SP2+0.5RR-1]} \tag{7-4}$$

where the terms in the exponent are the frequency of the corresponding type of atoms (e.g. RR is the number of rigid single or fused conjugated aromatic ring systems). For compounds with melting temperatures below 298.15 K (i.e. liquid at 298.15 K) the last term in Eqn. (7-2) is set to zero and the liquid solubility is simply calculated from the activity coefficient (this can however not just always be assumed true as discussed in section 5.3.2):

$$\log S_w = \log(55.5) - \log \gamma_w \tag{7-5}$$

The great advantage of such a method is that the water solubility can be well predicted with only knowledge of the molecular structure and the melting point. Another big advantage is that it is fitted to a large data set (the most recent fit being 1642 compounds [4]) and very frequently with group contribution methods, the larger the dataset the wider the applicability.

One of the advantages of this approach is also a disadvantage, which is that the entropy of fusion is estimated in the fits. The advantage of this is that users of the method can reproduce the value for the entropy of fusion used in the model fairly simply, however the disadvantage of this is that the groups are therefore not only relying on the possible experimental error in the solubility but also on any failures in the entropy of fusion prediction. These factors mean that in all likelihood the group values would probably not be reliable in reproducing the aqueous activity coefficient and are rather more tailored to use in the solubility equation. As an arbitrary example consider the infinite dilution activity coefficient of benzene in water; experimentally the value is around 2400 [28] but AQUAFAC [8] predicts a value of 4160.

Another possible drawback is that the groups are fairly narrowly defined as for most groups there are 3 different variations: groups attached to 2 non-aromatics (X_2), groups attached to 1 aromatic and 1 non-aromatic (XY) and groups attached to 2 aromatics (Y_2). However since there are a large amount of data relative to the number of groups this may not be a serious

problem (most recently [4] it was reported that 147 groups were used for 1642 compounds meaning that there should still be sufficient data behind many of the groups).

7.2. Training set data

Section 4 gives the possible sources of data as: partition coefficient data, SLE data and infinite dilution activity coefficient data. For the development of a new model for γ^{∞} in water at 298.15 K only SLE and infinite dilution data* were used. All the infinite dilution activity data were taken from the huge Dortmund Data Bank [28]. The SLE data was primarily taken from the DDB [28] and Beilstein [59] but in order to bulk up (put more data behind) some of the groups, data was taken from the free online database Chemspider [131]. Table 7-1 gives the number of data points from the various sources. In order to avoid having the method biased towards low molecular weight data (for which there is an abundance of data) only one data point was used per solute. For the SLE data the activity coefficient was considered to be at infinite dilution when the mole fraction solubility was less that 0.01 (1 mole percent) [43]. Interestingly almost all data fell into this solubility range except for a couple, some of which will be discussed in section 7.8.2.

Table 7-1 Number of data from each source used for the model development.

Source	Number
Gamma infinite - DDB	269
SLE - DDB	106
SLE - Beilstein	197
SLE - Chemspider	58
Total	630

All the heat of fusion and melting point data were taken from Beilstein [59] and DDB [28]. When there was duplicate data (i.e. data from both Beilstein and DDB) the data from DDB was generally considered to be superior.

7.3. Data Validation

When collecting the data for the training set it was clear to see that there is a large amount of scatter in the water solubility data in the literature. This is illustrated by the data in Table 7-2, where for some solutes the data can vary by as much as a factor of 1000. These large deviations are due the difficulties associated with solubility measurements [132, 133]. This large scatter coupled with the fact that many of the compounds in the training set fall into multiple chemical families makes data validation very difficult. For data which does fall into a

* As outlined in section 5.3 all data extracted from VLE and LLE are considered as infinite dilution data. There are however a fair amount of liquid solubility data available in literature. These data (unlike the LLE data used in the DDB) are typically not for mutual solubility and therefore cannot just be blindly used (as shown in section 5.3.2).

chemical family (alkanes, esters etc.) data validation was slightly more simplified since any obvious outliers from the families' general trend were removed.

When questionable data were found they were not simply removed but, in the case of SLE data, the three properties (*viz*. heat of fusion, melting temperature and water solubility) used to calculate the activity coefficients were checked. Water solubility and melting temperature data are fairly widely available and were therefore much simpler to check but heat of fusion data is much scarcer and was therefore almost never checked. As in the previous work [20], VBA (visual basic for applications) was used for the development of graphical interfaces to streamline the whole data validation process.

Table 7-2 Solubility data for a few compounds in water @ 298.15 K (all data obtained from Beilstein [59]).

Solubility in water (g/l)			
Hexachlorobenzene	**Benz(a)anthracene**	**(R)-Flurbiprofen**	**Psoralen**
4.7E-05	1.0E-02	3.3E-02	6.5E-02
5.4E-06	9.4E-06	9.5E-03	4.4E-05
8.0E-06	9.2E-06	2.9E-01	-
1.0E-05	1.3E-05	-	-

7.4. Group contribution and group interaction

The underlying principle in group contribution methods is that a solution of molecules can be adequately represented by a solution of functional groups. The huge advantage of group contribution methods is that there are significantly fewer functional groups than molecules. This principle has been widely applied for a large amount of thermophysical properties with good success. The simplest scheme is given as follows:

$$A = \sum v_i C_i \tag{7-6}$$

where A is some property, v_i is the group frequency and C_i is the group contribution value.

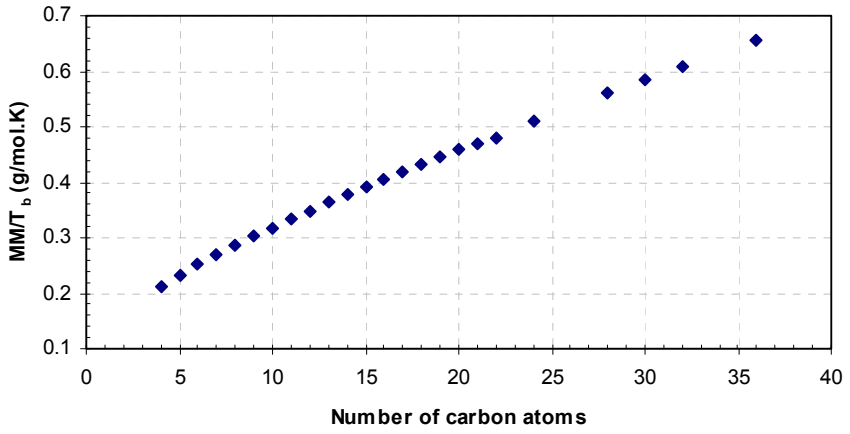


Fig. 7-1 MM/T_b as a function of the number of carbon atoms for the n-alkanes.

So for example if the compound of interest is n-propanol then there would be 1xCH_3, 2xCH_2 and 1xOH group. As mentioned above the basic assumption is that all groups behave additively, i.e. the change in a thermophysical property from ethane to ethanol (addition of an OH group) is assumed to be the same as the change in a thermophysical property from ethanol to 1,2-ethandiol (addition of an OH group). This assumption is generally acceptable but does fail for some groups as illustrated for the OH group in Table 7-3. If the OH group behaved additively then one would expect the difference of MM/T_b (Fig. 7-1 shows this is approximately linear for the n-alkanes) to be similar for the addition of an OH group as it is for the addition of a CH_3 group.

Table 7-3 Illustration of group non-additivity (all data from the DDB[28]).

Addition of an OH group			Addition of a CH_3 group		
Compound	**MM/T_b (g/mol.K)**	**Difference**	**Compound**	**MM/T_b (g/mol.K)**	**Difference**
Propane	0.191		Propane	0.191	
Propanol	0.162	-0.029	Butane	0.213	+0.022
1,3-Propandiol	0.156	-0.006	Pentane	0.233	+0.020
1,2,3-Propantriol	0.163	+0.008	3-Methylpentane	0.256	+0.023

This was noted and accounted for by Nannoolal et al. [15-18] who added a group interaction term which accounts for the non-additivity of these groups. This group interaction term is accounted for in this work as:

$$A = \sum v_i C_i + \frac{1}{2}\sum_i \sum_j \frac{GI_{i-j}}{n \times m} \tag{7-7}$$

where n is the total number of heavy atoms, m is the total number of non-additive groups and GI is the group interaction value. GI_{i-i} is zero since this is accounted for by the group contribution and $GI_{i-j} = GI_{j-i}$ (hence the ½ before the double summations). Consider the following example of a compound with two OH (the numbers in superscript are to differentiate between them) and one C=O group. The double summation term in Eqn. (7-7) results in $(2*GI_{OH-C=O} + 1*GI_{OH-OH})/(3n)$ where m = 3.

	OH (1)	**OH (2)**	**C=O**
OH (1)	0	1	1
OH (2)	1	0	1
C=O	1	1	0

7.5. Group contribution scheme

Due to the large amount of scatter present in the data (discussed in section 7.3) a simplified group contribution approach was applied. The reason being that the fewer groups there are the more data there are per group and therefore the less susceptible the group values are to bad data. For example if a hypothetical group *AA* is used and there is only one compound with that group and the data turns out to be very poor, data for any other compound with that group are going to be incorrectly predicted. Conversely if a more general group is used which encompasses AA this will help absorb the effect of the bad data; this point is further illustrated

in the paragraphs following. Careful attention was given to the removal of the groups since these groups had been added after careful consideration of the data used for the development of previous methods by this group [15-20].

All the structural correction groups that were used in the previous work [20] were removed, not because they are erroneous but simply because there is insufficient good data to back them up (data for 630 compounds were used in this work as opposed to the vapour pressure study [20] where there were data for approximately 2330 compounds). Similarly all of the group interactions, with the exception of the OH, COOH and C=O groups, were removed. These and other simplifications that have been made are discussed in the sections following. Had these groups not been removed the fit for the training set would definitely have been better but this, in many instances, could have resulted in very erroneous predictions for any data external to the training set.

This point is fairly well illustrated by the following example for the boiling point method developed by Nannoolal et al. [15] shown in Fig. 7-2. What this plot shows is very counter intuitive; the number of model parameters is directly proportional to the absolute error in the boiling point. The probable reason for this is that groups are frequently added to account for effects which do not exist and are due to experimental error. This means that when a non-training set compound containing one of these groups is analysed it results in large errors. There is however some optimum number of groups because if too few are used many effects which do exist will not be captured and the method would become less accurate. Therefore careful consideration needs to be made to the selection of the groups and it should not just be a fitting exercise with the purpose of getting the lowest possible training set error.

For pure component group contribution methods it is common to use the group contribution parameter as a variable in another equation and not fit the property directly. For example the boiling point method of Nannoolal [15] uses the following equation:

$$T_b = \frac{\sum v_i C_i}{n^a + b} + c \tag{7-8}$$

where a,b and c are model parameters which are the same for all compounds and $\sum v_i C_i$ is the group contribution value. These “more fitted” expressions work very well for properties such as boiling point where there are large amounts of reasonable data. However to avoid any over-fitting to the training set, an approach such as this was not used but rather a more simple direct fit to the activity coefficient was used:

$$\ln \gamma_2^{\infty} = \sum_i v_i C_i + \frac{1}{2} \sum_i \sum_j \frac{GI_{i-j}}{n \times m} \tag{7-9}$$

where the symbols are the same as those defined in Eqn. (7-7) and the subscript 2 refers to the solute. In Eqn. (7-9), $\ln \gamma_2^{\infty}$ is chosen over γ_2^{∞} since group contributions represent

$\bar{g}^E(group) = RT \ln \Gamma^{\infty}(group)$ and $\bar{g}^E(molecule) = \sum n_i \bar{g}_i^E(group)$; whereby $\ln \Gamma^{\infty}$ of the groups should be additive to yield $\ln \gamma^{\infty}$ of the component. As mentioned above most of the group interactions were dropped. This means that the only groups considered non-additive are the aliphatic OH (group 49 and 163), the ketone groups (group 57 and 58) and the aliphatic carboxylic acid group (group 53). Aside from the practical advantage of dropping many of the group interaction terms there is also a theoretical basis as in a single molecule at infinite dilution; one may expect that the effect of hydrogen bonds should be much less pronounced than in a pure fluid of many interacting molecules. For discussion on the regression of the groups refer to Appendix C.

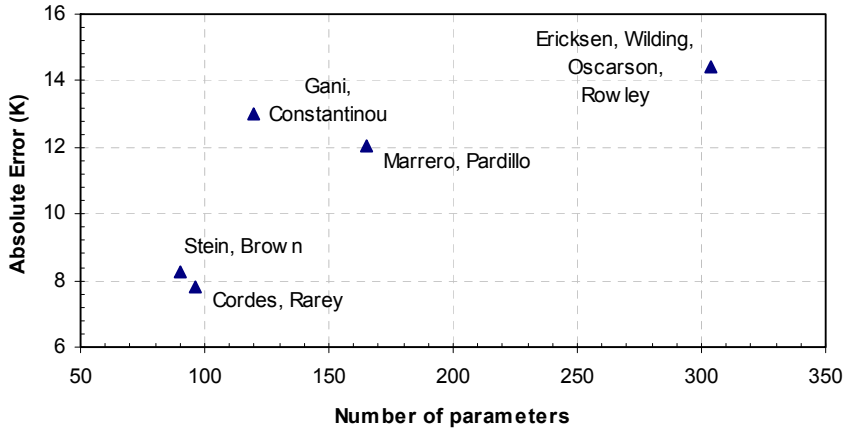


Fig. 7-2 Number of model parameters and average absolute error in the boiling point prediction for the methods indicated (Cordes, Rarey [80]; Stein, Brown [134]; Gani, Constantinou [135]; Marrero, Pardillo [136]; Ericksen et al. [137]).

7.6. Model considerations

Since there was quite extensive work done in examining combinatorial expressions it was initially thought that the best results could be obtained by fitting the model groups to the residual term i.e.:

$$\ln \gamma_2^{R,\infty} = \ln \gamma_2^{\infty} - \ln \gamma_2^{C,\infty} \tag{7-10}$$

where $\ln \gamma_2^{C,\infty}$ is given by Eqn. (6-6). While the use of this expression is theoretically appealing (since UNIFAC, ASOG etc. use group contribution to calculate the residual) there are some practical problems with doing so:

1. The percentage error of the fit was considerably worse. While this work is not just concerned with getting the lowest error at any cost, when there is such a considerable reduction in predictive power the inclusion is simply not practical.

2. The number of compounds for which the method could be applied was greatly reduced (since V_m r and q are needed and are usually not available for compounds with exotic functional groups).

3. The combinatorial expression gave extremely large values for some compounds (it is clear that this point is linked to point 1.) which seem unrealistic.

It is however very probable that the group values contain both the residual and the combinatorial contribution and therefore a more complicated approach is unnecessary (as a case in point consider the results shown in section 7.8.5). There had been an argument in literature that at temperatures around 298 K, the cavitation contribution in water would be exceptionally "cheap" [138]; the unique ability of water to dissolve macromolecules had been attributed to this. If this should be true the new combinatorial expression derived from n-alkanes would not be applicable to H_2O.

7.7. Temperature dependence

As shown above (section 3.1) the temperature dependence of the activity coefficient is given by Eqn. (3-29), for application to infinite dilution data this reduces to:

$$\left(\frac{\partial \ln \gamma_i^{\infty}}{\partial T}\right)_P = -\frac{\overline{H}_i^{E,\infty}}{RT^2} \tag{7-11}$$

If it is assumed that $\overline{H}_i^{E,\infty}$ is constant with respect to temperature, then integration of Eqn. (7-11) results in the following:

$$\ln \gamma_i^{\infty} = a + \frac{b}{T} \tag{7-12}$$

where *a* is a constant of integration and *b* is the constant assumed value of $-\overline{H}_i^{E,\infty} / R$. This assumption of constant $\overline{H}_i^{E,\infty}$ can be quite a good one [139] however for polar solvents such as water where $\overline{H}_i^{E,\infty}$ is very often a non-trivial function of temperature is typically not good. This was confirmed by Hovorka et al. [140] who investigated the temperature dependence in the case of some moderately hydrophobic organic solutes in water and proposed the following expression:

$$\ln \gamma_i^{\infty} = A + \frac{B}{\tau} + C \ln \tau + D\tau \tag{7-13}$$

$$\tau = \frac{T}{298.15K} \tag{7-14}$$

Where A, B, C and D are model parameters. The temptation when dealing with temperature dependence of the activity coefficient is to simply ignore it since data are typically only available at or around 298 K. Also data for $\overline{H}_i^{E,\infty}$ are very scarce and reliable temperature dependent data for the infinite dilution activity coefficient are also hard to obtain. Consider the histogram of the temperature distribution for the γ^{∞} data in the DDB [28] shown in Fig. 7-3. At first glance it appears that there is a large amount of temperature dependent data available however, if one considers that there are data for a total of 585 different solutes it is clear that much of the non-298 K data is for other solutes. For SLE data the situation is much worse, consider the histogram shown in Fig. 7-4. While it again seems that there are a reasonable amount of data, these data are very often poor (missing units, huge variations of values, etc.). Nevertheless even if all the data shown in these histograms were usable, model development over any kind of reasonable temperature range would be exceptionally difficult (almost impossible if reasonable accuracy is required) due to the fairly complex nature of the γ^{∞} temperature behaviour in water.

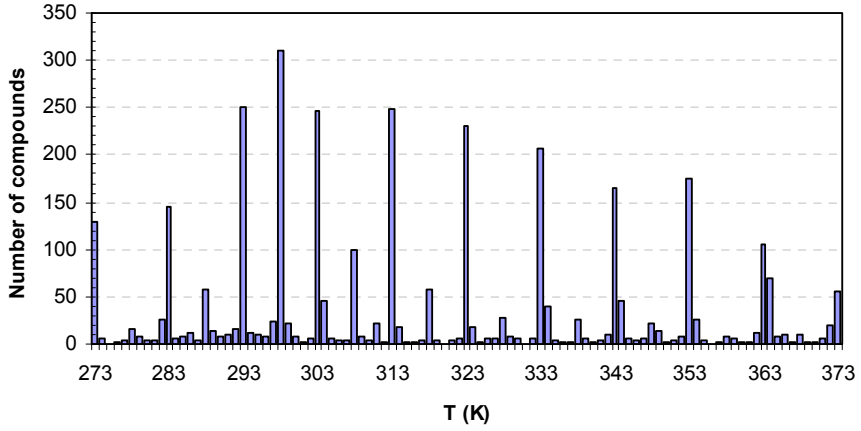


Fig. 7-3 Number of solutes at various temperatures for γ^{∞} data in the DDB [28].

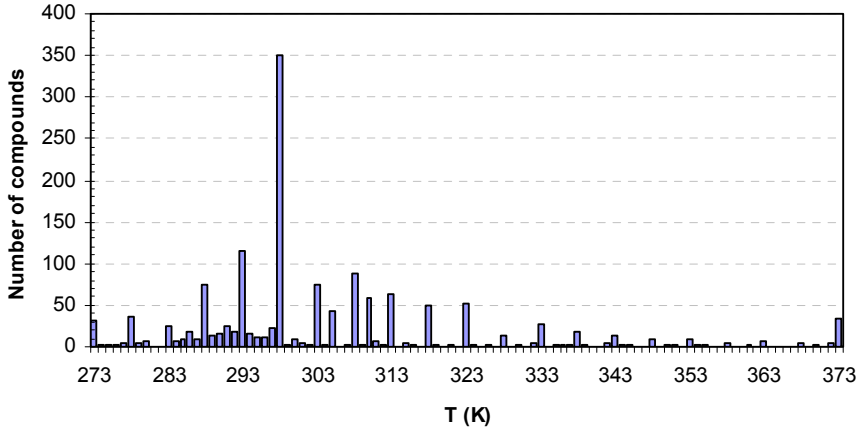


Fig. 7-4 Number of solutes at various temperatures for SLE data (with available fusion data from either DDB or Beilstein) in Beilstein [59].

However since the activity coefficient in this work is only considered at infinite dilution it may be possible that for each solvent, in this case water, it changes in the same way regardless of the solute used. In order to test this assumption the infinite dilution activity coefficient data (from the DDB [28] γ^{∞} database) were plotted in log units against reciprocal temperature. The resulting plots are shown in Fig. 7-5 and Fig. 7-6. It is clear that the lines in these plots are parallel. This is no coincidence, rather they are fitted to the following equations for the polar and non-polar solutes respectively (for the derivation of these equations see Appendix E):

$$\ln\gamma_2^{\infty} = \ln\gamma_{2,298K}^{\infty} - 21.2\frac{(1-\tau)}{\tau} - 17.5\ln\tau \tag{7-15}$$

$$\ln\gamma_2^{\infty} = \ln\gamma_{2,298K}^{\infty} - 24.5\frac{(1-\tau)}{\tau} - 27.5\ln\tau \tag{7-16}$$

where τ is given by Eqn. (7-14). The constants in the second two terms of the equations are a best fit to some representative data (over a sufficiently wide range) from the DDB [28]. What these equations essentially mean is that all compounds have the same value of $\overline{H}_i^{E,\infty}$ for the polar compounds this value is -9.14 kJ/mol and for the non-polar compounds it is 7.42 kJ/mol. For the non-polar compounds this assumption is fairly good for some compounds but rather poor for others as shown in Table 7-4 which contains some $\overline{H}_i^{E,\infty}$ data for various compounds. That Eqn. (7-15) works so well for the compounds shown in Fig. 7-5 is a coincidence rather than fact. A deeper evaluation of the data available in the DDB [28] shows that for many small polar compounds (acetonitrile, propionitrile etc.) similar temperature dependence is not observed.

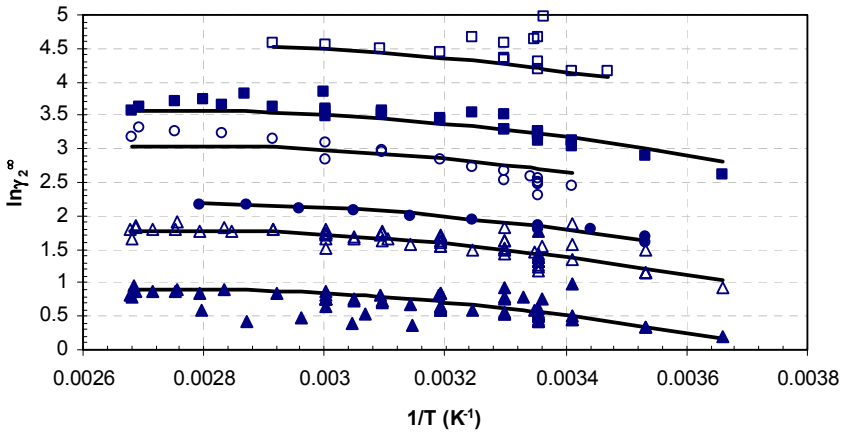


Fig. 7-5 $\ln\gamma_2^{\infty}$ vs. 1/T for various polar solutes (2) in water (1) (Data from the DDB [28], □ - ethyl acetate, ■ - 2-butanol, ○ - tert-butanol, ● - ethylene oxide, △ - ethanol, ▲ - methanol, —— Eqn. (7-15) fitted to the data).

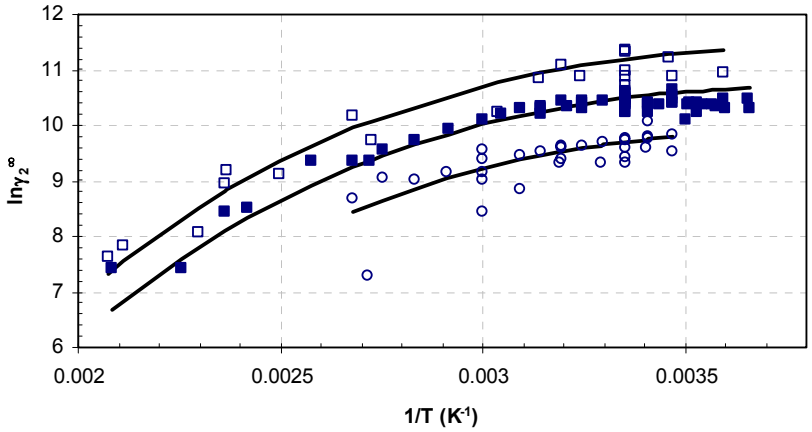


Fig. 7-6 $\ln\gamma_2^\infty$ vs. 1/T for various non-polar (or weakly polar) solutes (2) in water (1) (Data from the DDB [28], □ - cyclohexane, ■ - ethylbenzene, ○ - 1-octanol, —— Eqn. (7-16) fitted to the data).

Table 7-4 Partial molar excess heat of mixing at infinite dilution for some solutes in water (data from Hovorka et al. [141]).

Solute	$\overline{H}_i^{E,\infty}$
1-Butanol	-9.2
2,4-Pentanedione	1.62
2-Butanone	-10.65
2-Ethoxyethyl acetate	-15.98
2-Methoxyethyl acetate	-12.38
2-Methyl-1-propanol	-9.18
Acetonitrile	-1.51
Acrylonitrile	-2.42
Cyclohexanol	-8.46
Cyclohexanone	-8.71
Cyclopentanol	-10.23
Cyclopentanone	-8.74
Ethyl acetoacetate	-5.43
Methyl acetoacetate	-2.58
Methyl methoxyacetate	-9.07
Propionitrile	-3.37

When examining Fig. 7-5 and Fig. 7-6 it is clear that the trends are opposite. For the polar compounds γ_2^∞ increases with increasing T (and therefore decreasing 1/T), however when considering the relationship between activity coefficient and solubility this seems intuitively wrong since an increase in activity coefficient generally means a decrease in solubility[†]. The non-polar compounds show the expected trend. If this non-polar correlation is just blindly

[†] This is generally, however not always the case. For a dilute liquid solute (where the mutual solubility is low) an increase in γ^∞ will always mean a decrease in the liquid solubility in the solvent. For solid solutes the situation is slightly more complex the solubility depends on both γ^∞ (for a dilute solute) and the crystalline nature of the solute (dependent on T, T_m and $\Delta_{fus}H$). Since both are temperature dependent it is possible that there are some combination of solutes and solvents whereby the temperature dependence of γ^∞ is similar to those given in Fig. 7-5 but yet the overall solubility is still increasing with increasing temperature (i.e. the temperature dependence of the crystallinity of the solute is stronger). Such a combination is probably rare and has never been observed by this author.

applied to solubility data (regardless of the polarity of the solute) the results are surprisingly good (see Fig. 7-7 - Fig. 7-17). These examples were chosen based on the complexity of the molecule and the value of the activity coefficient (so as to get a full spectrum of activity coefficient data). All data is extracted from SLE/LLE data and the resulting activity coefficients are considered to be at infinite dilution if x<0.01 (as discussed above). In no examples tried (where data was extracted from SLE/LLE) was the "polar" correlation applicable, the reasons for this are discussed in Appendix F.

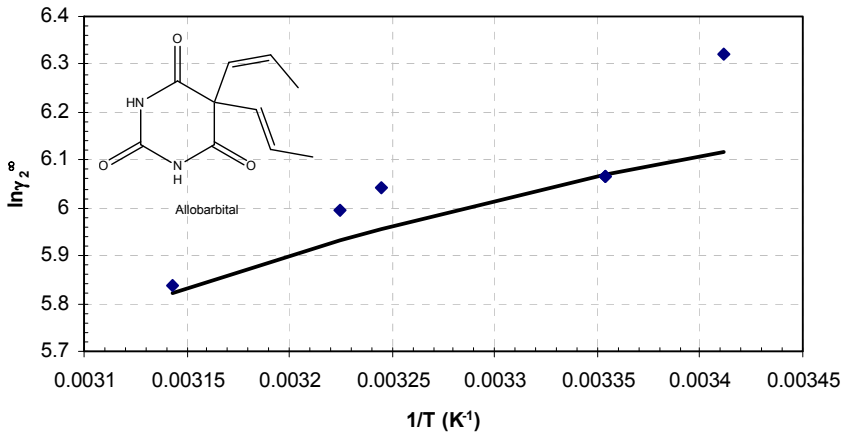


Fig. 7-7 $\ln\gamma_2^\infty$ vs. 1/T for allobarbital (2) in water (1) (◆ - data extracted from SLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point).

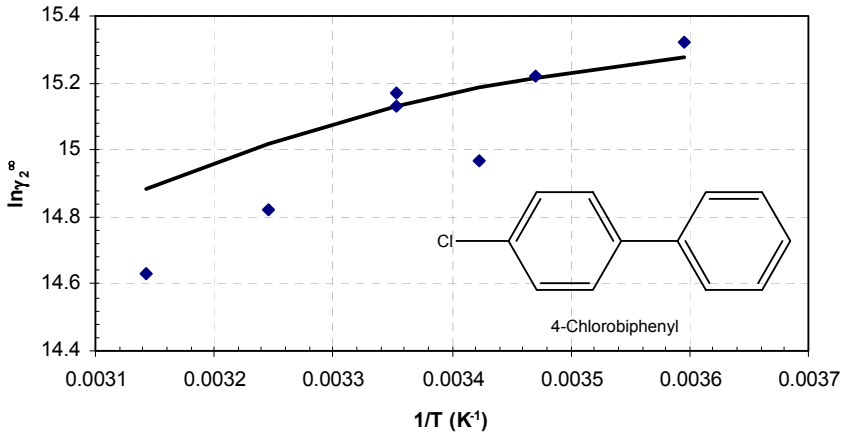


Fig. 7-8 $\ln\gamma_2^\infty$ vs. 1/T for 4-chlorobiphenyl (2) in water (1) (◆ - data extracted from SLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point).

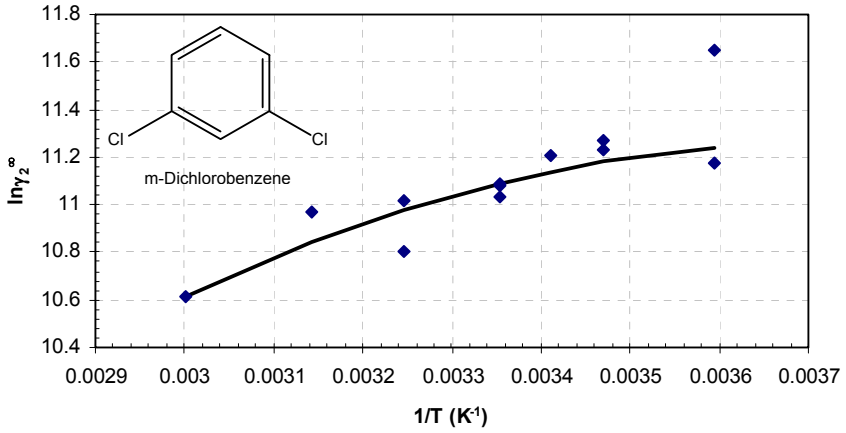


Fig. 7-9 $\ln\gamma_2^\infty$ vs. 1/T for m-dichlorobenzene (2) in water (1) (◆ - data extracted from LLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point).

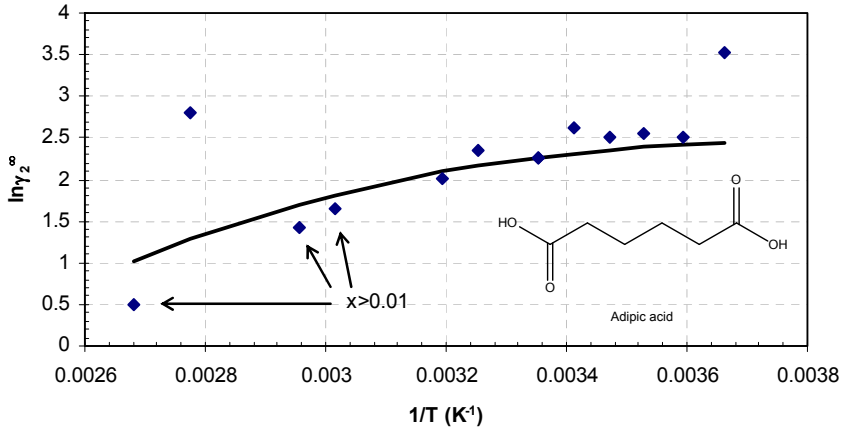


Fig. 7-10 $\ln\gamma_2^\infty$ vs. 1/T for adipic acid (2) in water (1) (◆ - data extracted from SLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point).

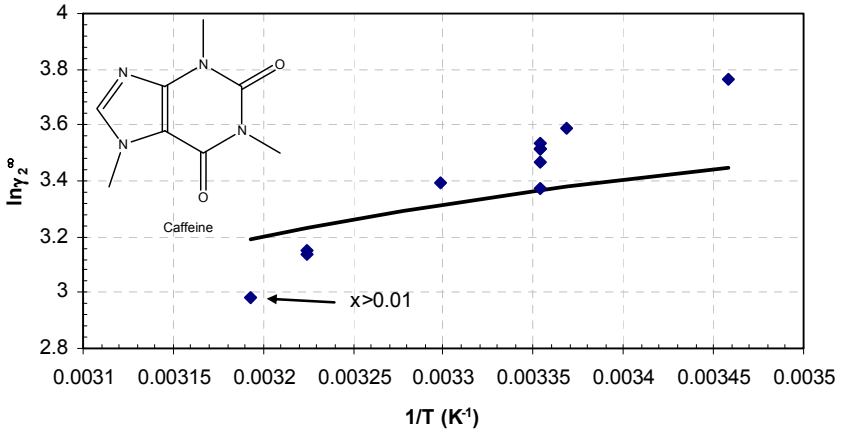


Fig. 7-11 $\ln\gamma_2^\infty$ vs. 1/T for caffeine (2) in water (1) (◆ - data extracted from SLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point).

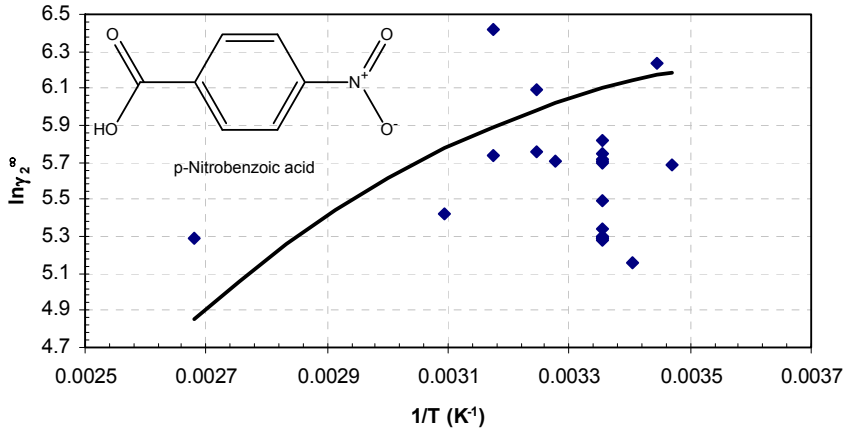


Fig. 7-12 $\ln\gamma_2^{\infty}$ vs. 1/T for p-nitrobenzoic acid (2) in water (1) (◆ - data extracted from SLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point).

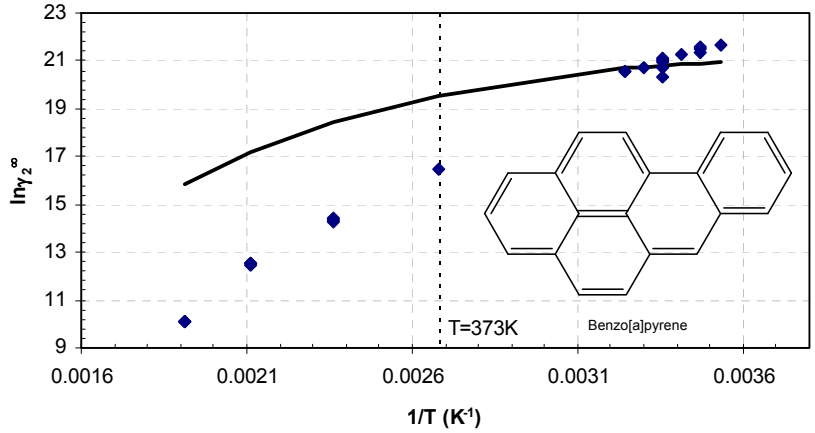


Fig. 7-13 $\ln\gamma_2^{\infty}$ vs. 1/T for benzo[a]pyrene (2) in water (1) (◆ - data extracted from SLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point).

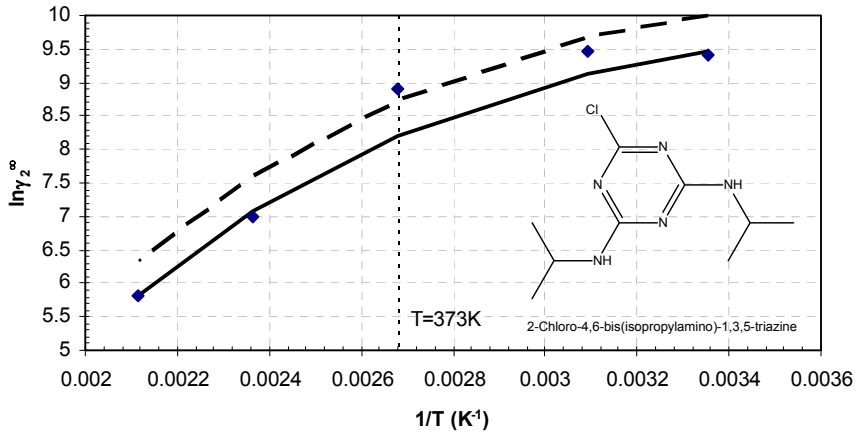


Fig. 7-14 $\ln\gamma_2^{\infty}$ vs. 1/T for 2-chloro-4,6-bis(isopropylamino)-1,3,5-triazine (2) in water (1) (◆ - data extracted from SLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point, ----- Eqn. (7-16) with a guessed (but seemingly more likely) value for the 298 K reference point).

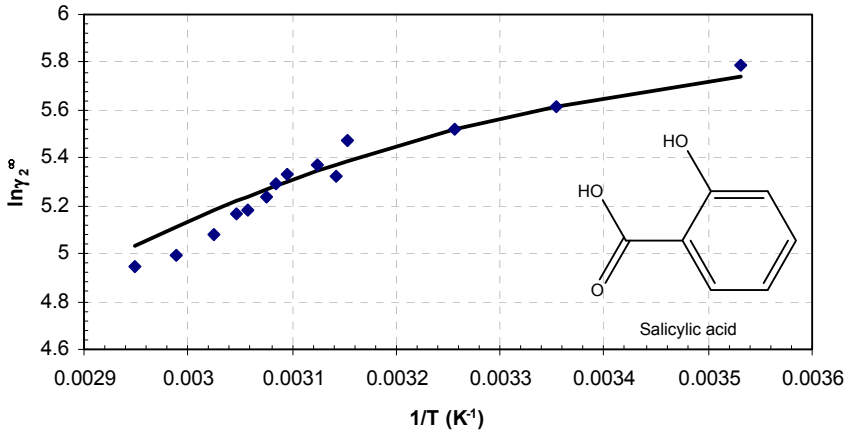


Fig. 7-15 $\ln\gamma_2^\infty$ vs. 1/T for salicylic acid (2) in water (1) (◆ - data extracted from SLE data from the DDB [28], —— Eqn. (7-16) with the 298 K reference experimental data point).

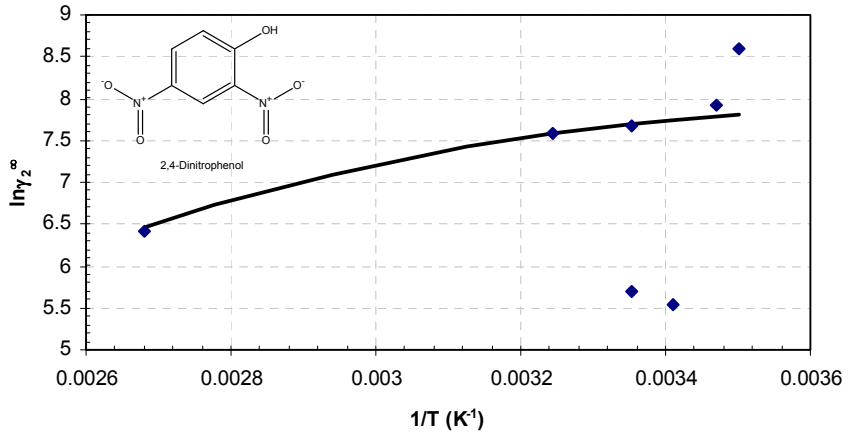


Fig. 7-16 $\ln\gamma_2^\infty$ vs. 1/T for 2,4-dinitrophenol (2) in water (1) (◆ - data extracted from SLE data from Beilstein [59], —— Eqn. (7-16) with the 298 K reference experimental data point).

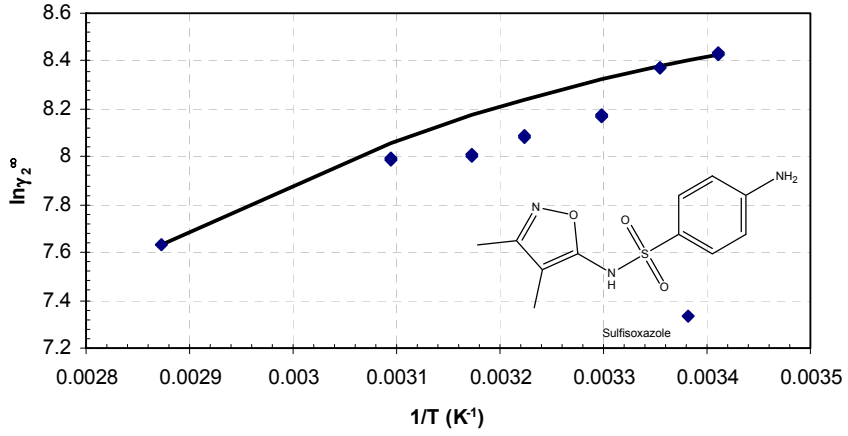


Fig. 7-17 $\ln\gamma_2^\infty$ vs. 1/T for sulfisoxazole (2) in water (1) (◆ - data extracted from SLE data from literature [119, 142], —— Eqn. (7-16) with the 298 K reference experimental data point).

The above examples suggest that the general temperature dependence of the activity coefficient at infinite dilution in water given be Eqn. (7-16) does hold quite well in most instances. In the above examples, the inclusion of adipic (Fig. 7-10) and salicylic (Fig. 7-15) acids may seem strange as one would expect them to be well represented by the polar curve. It is quite likely that these 2 compounds strongly associate and behave as a non-polar dimer. This correlation (Eqn. (7-16)) however is not proposed as something rigorously true in all situations but rather somewhat of a good first approximation. The only fairly large deviations are for caffeine (Fig. 7-11) and benzo[a]pyrene (Fig. 7-13). In the case of caffeine the reason for the deviation is unclear since practically every data point is from a different literature source (if they had all been from the same source, the chance of the data being poor would be much greater), however for benzo[a]pyrene the deviation is probably because the high temperature data are solubilities in water above 373 K (at pressures between 60 and 70 bar [143]) and these higher pressures are not accounted for by this simple correlation. A fairly good prediction is obtained for 2-chloro-4,6-bis(isopropylamino)-1,3,5-triazine (Fig. 7-14) at temperatures above 373, but this is more than likely a coincidence due to an erroneous value at 298 K, a more likely prediction is shown in the plot. The recommended temperature range for the use of these equations is: 280 K – 350 K.

7.8. Results

The following sections give analysis on the errors that were observed for the various chemical families and also the overall errors for the training set and the test set. The error plots will show the errors for this work, mod. UNIFAC (Do) [10] and COSMO-RS(OL) [72] (any reference to COSMO-RS from this point onwards refers to the use of COSMO-RS(OL)) because these were the 2 literature methods with the lowest errors. The error tables however will show the comparison between the predictions of the four common methods available in the literature (i.e. UNIFAC, mod. UNIFAC, COSMO-RS(OL) & COSMO-SAC). The predictions of Kuhne's [5] method are included, even though this is not directly an activity coefficient method but rather a water solubility method used to calculate activity coefficient at infinite dilution *via* Eqn. (5-18). The following sections contain discussions pertaining to the various chemical families. The compounds are split such that, for example, the oxygen compounds section encompasses compounds which contain only oxygen groups (aside from the hydrocarbon backbone). Compounds which contain a mixture of family groups fall into the category of multi-functional while, for example, 1,2-ethandiol would only fall into the oxygenated compounds.

7.8.1. Hydrocarbons

As mentioned above all the structural correction groups introduced previously [20] were dropped because there was not enough good quality data to warrant their inclusion. The error plots for the aromatic hydrocarbons and all hydrocarbons are shown in Fig. 7-18 and Fig.

7-21 respectively. Usually the alkanes are one of the simplest families to predict, but in this case there was quite a large amount of scatter in the data due to the very large values of γ^{∞}, which are difficult to determine experimentally. The errors for some of the most popular methods available are compared to the errors in this work in Table 7-5. The aromatic hydrocarbons have the lowest percentage error; it is not totally clear why this is the case. It is possible that these data are simpler to measure (since the γ^{∞} are smaller than alkanes of similar size) and therefore these data are more reliable. It is clear from Table 7-5 that mod. UNIFAC produces the largest errors. This has been previously observed in the literature [56, 144] and the following correction, for predicting water solubility of alkanes and alkenes, is given by Jakob et al. [56]:

$$\log S_w = 1.104 \log\left(\frac{55.56}{\gamma^{\infty}_{2,298.5K}}\right) + 0.0042T - 2.817 \tag{7-17}$$

where S_w is the water concentration. Since the hydrocarbons and water are practically immiscible at 298 K, Eqn. (5-32) applies and this can be rewritten as:

$$\gamma_2^{\infty} = \frac{55.56}{S_w} \tag{7-18}$$

If Eqn's (7-17) and (7-18) are combined then the following results for liquid solutes:

$$\gamma^{\infty}_{corr} = \frac{\left(\gamma^{\infty}_{Mod.UNIFAC,298K}\right)^{1.104} 10^{2.817}}{\left(55.56\right)^{0.104} 10^{0.0042T}} = \frac{\left(\gamma^{\infty}_{Mod.UNIFAC,298K}\right)^{1.104} 432.061}{10^{0.0042T}} \tag{7-19}$$

When this correction is applied to the mod UNIFAC predictions the results are dramatically improved as shown in Fig. 7-19. The comparisons of the corrected and uncorrected prediction of the alkanes and alkenes are shown in Fig. 7-19 and Fig. 7-20.

Table 7-5 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for hydrocarbons (number of solutes in superscript).

Name	Eqn. (7-9)	Kuhne	UNIFAC	mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
Hydrocarbons	4.5 [103]	10.4 [32]	12.1 [96]	20.6 [96]	13.8 [78]	18.6 [78]
Hydrocarbons (corrected)	4.5 [103]	10.4 [32]	12.1 [96]	10.8 [96]	13.8 [78]	18.6 [78]
Alkanes	7.4 [24]	36.7 [1]	17.4 [24]	23.1 [24]	11.0 [22]	13.4 [22]
Alkanes (corrected)	7.4 [24]	36.7 [1]	17.4 [24]	12.5 [24]	11.0 [22]	13.4 [22]
n-Alkanes	4.6 [9]	36.7 [1]	8.1 [9]	10.3 [9]	16.8 [7]	6.8 [7]
Alkanes (non-cyclic)	6.6 [20]	36.7 [1]	16.8 [20]	19.4 [20]	12.2 [18]	13.2 [18]
Alkanes (cyclic)	9.6 [4]	-	20.7 [4]	41.8 [4]	5.5 [4]	14.4 [4]
Alkenes	6.0 [19]	-	14.7 [19]	45.9 [19]	13.2 [19]	19.3 [19]
Alkenes (corrected)	6.0 [19]	-	14.7 [19]	10.1 [19]	13.2 [19]	19.3 [19]
Alkenes (cyclic)	5.5 [7]	6.8 [1]	10.7 [7]	42.0 [7]	16.1 [7]	21.2 [7]
Alkenes (non-cyclic)	5.3 [13]	13.2 [1]	15.6 [13]	40.0 [13]	13.4 [12]	19.8 [12]
Alkynes	5.6 [7]	-	-	-	20.1 [7]	23.0 [7]
Aromatic hydrocarbons	2.7 [53]	9.5 [31]	8.8 [53]	10.3 [53]	14.8 [30]	21.0 [30]

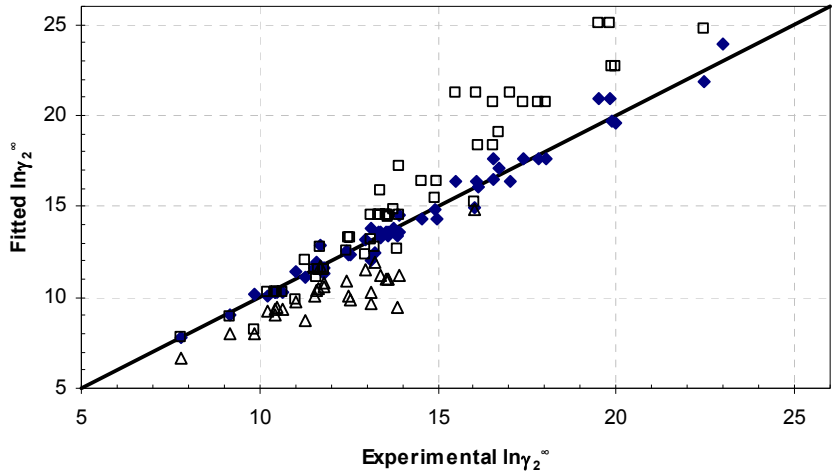


Fig. 7-18 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for the aromatic hydrocarbons (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

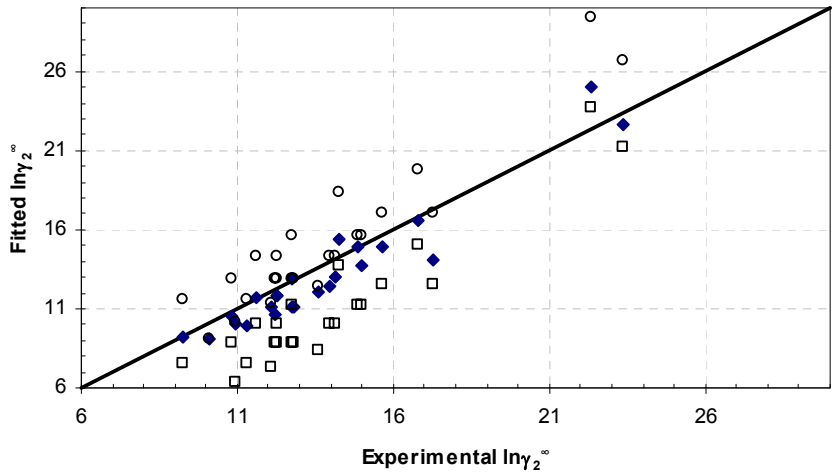


Fig. 7-19 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for all alkanes using the correction given by Eqn. (7-19) (◆ - This work, □ – mod. UNIFAC (Do) uncorrected, ○ – mod. UNIFAC (Do) corrected).

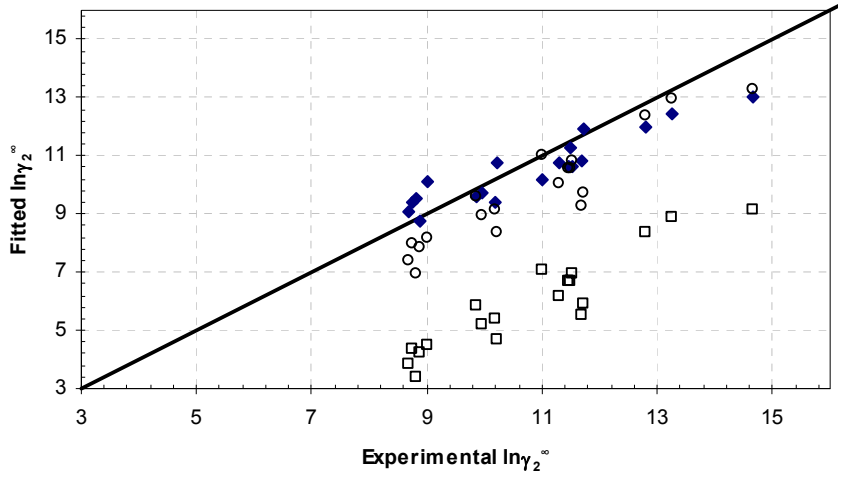


Fig. 7-20 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for all alkenes using the correction given by Eqn. (7-19) (◆ - This work, □ – mod. UNIFAC (Do) uncorrected, ○ – mod. UNIFAC (Do) corrected).

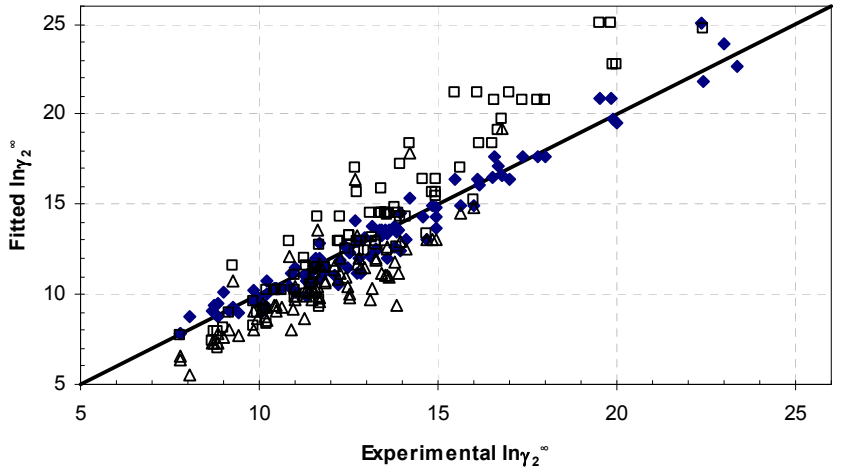


Fig. 7-21 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for all hydrocarbons using the correction for alkanes and alkenes (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

7.8.2. Oxygen compounds

Typically when developing predictive methods one finds that oxygenated compounds have the largest errors. This is largely due to hydrogen bonding which is widely present in the oxygenated compounds. The errors obtained for this work are only slightly higher than the overall average of 7.3 %. Mod. UNIFAC performs remarkably well but this is partially due to the fact that the mod. UNIFAC parameters would have been fitted to a large portion of the infinite dilution activity coefficient data present in the training set for the oxygenated compounds. The comparisons of the various methods are shown in Table 7-6. For ketones and esters the predictions of UNIFAC and mod. UNIFAC respectively are better. This is because the first members of each homologous series (e.g. for ketones: acetone, butanone

etc.) are not well represented by the new method. These compounds are however not the focus of this work so this should not be considered as a large model drawback.

Table 7-6 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for oxygenated compounds (number of solutes in superscript).

Name	Eqn. (7-9)	Kuhne	UNIFAC	mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
All oxygenated compounds	8.1 [231]	23.2 [91]	17.5 [182]	12.2 [167]	25.5 [134]	33.2 [134]
Alcohols	7.7 [69]	28.2 [25]	20.9 [69]	12.7 [68]	26.0 [55]	25.9 [55]
n-Alcohols	5.7 [13]	23.1 [4]	16.1 [13]	9.6 [13]	27.7 [9]	14.9 [9]
Aromatic alcohols	7.9 [16]	15.5 [12]	28.8 [16]	15.9 [16]	25.6 [14]	48.5 [14]
Ethers	12.6 [24]	4.8 [1]	18.5 [23]	12.9 [19]	50.0 [19]	23.3 [19]
Epoxides	8.9 [1]	-	-	-	78.7 [1]	35.8 [1]
Aldehydes	6.6 [8]	-	12.1 [8]	14.2 [7]	16.6 [8]	42.0 [8]
Ketones	11.3 [20]	33.5 [5]	13.5 [16]	7.9 [16]	17.8 [16]	37.4 [16]
Carboxylic acids	8.5 [21]	21.7 [19]	10.1 [19]	17.5 [19]	48.9 [5]	122.8 [5]
n-Carboxyllic acids	8.6 [9]	25.9 [7]	13.4 [9]	23.5 [9]	50.6 [4]	139.2 [4]
Esters	6.8 [31]	12.4 [7]	5.7 [31]	6.9 [31]	9.3 [26]	34.6 [26]

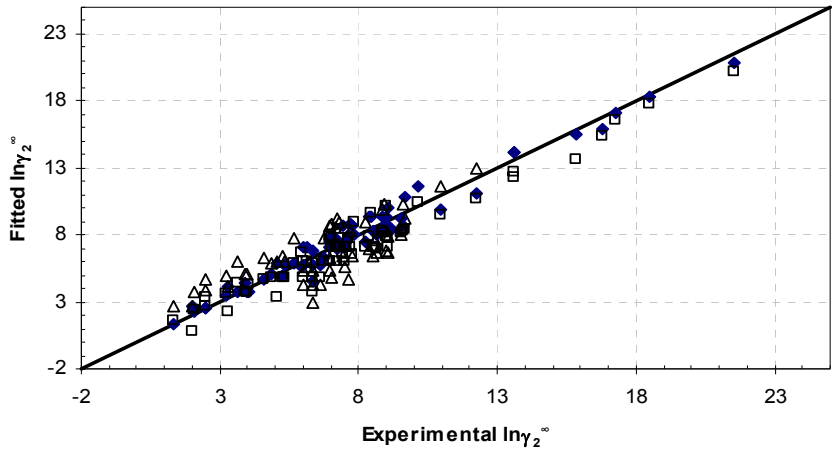


Fig. 7-22 Experimental and fitted lnγ_2^∞ @ 298.15 K for alcohols (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

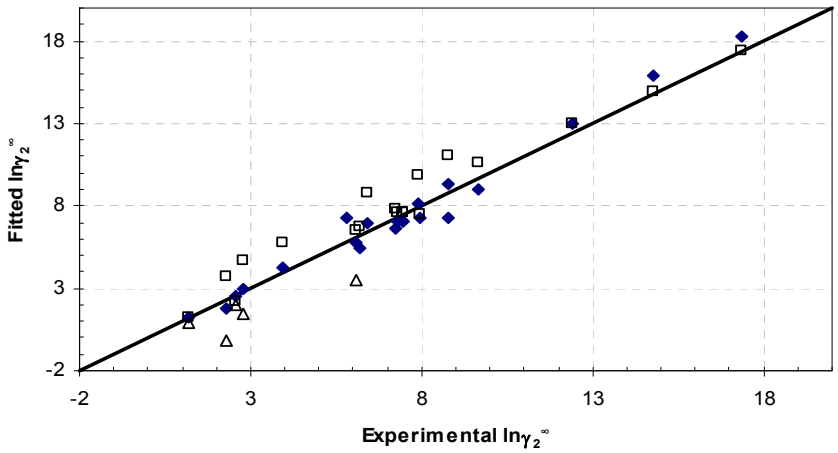


Fig. 7-23 Experimental and fitted lnγ_2^∞ @ 298.15 K for carboxylic acids (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

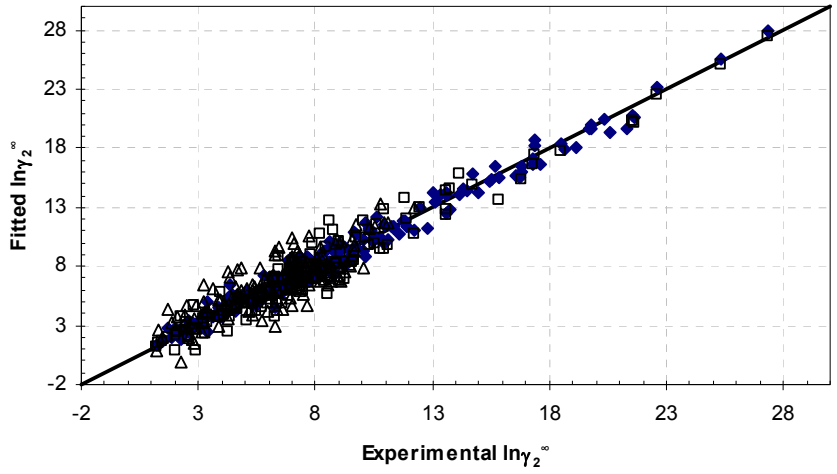
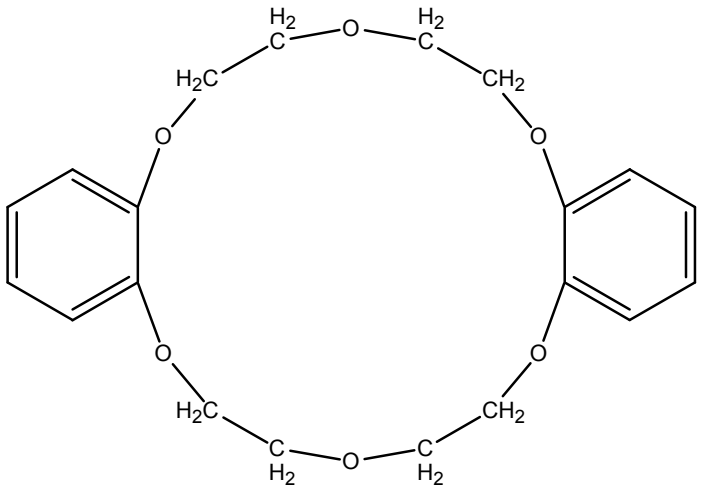


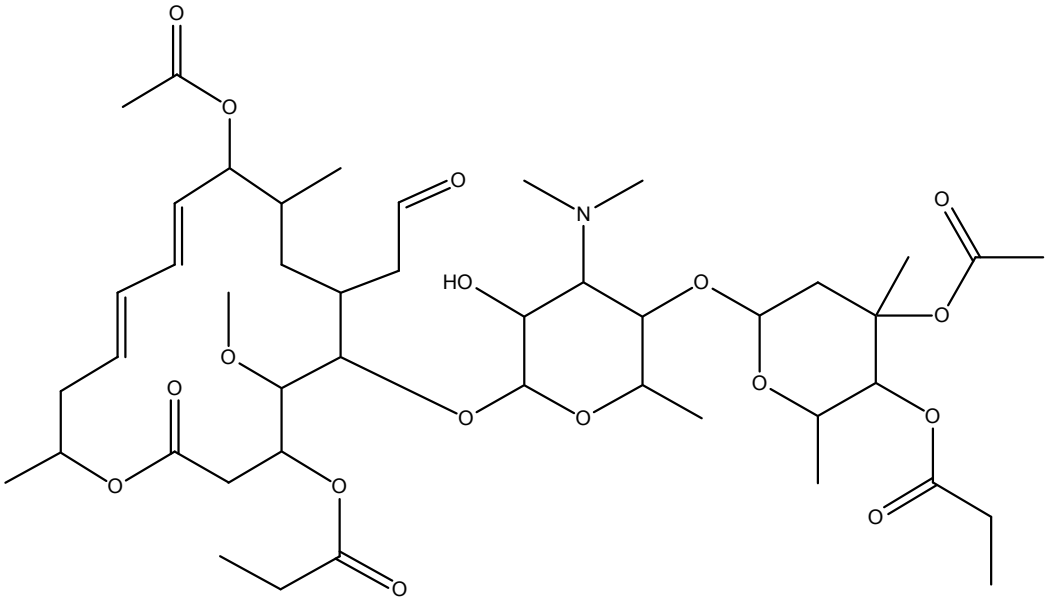
Fig. 7-24 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for oxygenated compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

As mentioned above the aliphatic alcohols, aliphatic carboxylic acids and the ketones were the only 3 groups for which group interactions were introduced to account for non-additivity. This is because there was not enough good data to support the inclusion of further interaction groups. Another reason for this is that for large compounds it was frequently found that group interactions were not needed; consider the examples of dibenzo-18-crown-6 (db18c6) and 9,3''-diacetylmidecamycin (this is discussed above – section 7.4). As shown in Fig. 7-25 and Fig. 7-26 they both contain many groups that were previously considered to be non-additive (i.e. in need of a group interaction term). This doesn't seem to have much of an effect in this case because the experimental and fitted values are fairly close ($\ln\gamma_{exp}^{\infty} = 9.1$ and $\ln\gamma_{fit}^{\infty} = 9.4$ for dibenzo-18-crown-6 (db18c6); $\ln\gamma_{exp}^{\infty} = 8.8$ and $\ln\gamma_{fit}^{\infty} = 9.2$ for 9,3''-diacetylmidecamycin) while the predictions for the mono-functional ethers are still good.



Dibenzo-18-crown-6 (db18c6)

Fig. 7-25 Molecular structure of dibenzo-18-crown-6 (db18c6).



9,3''-diacetylmidecamycin

Fig. 7-26 Molecular structure of 9,3''-diacetylmidecamycin.

The need for the OH (group 49 and 163) and C=O (group 57 and 58) group interaction became clear when examining the predictions for some compounds which did not fall into the infinite dilution assumption (i.e. mole fraction solubility less than 0.01). Typically the predictions of the activity coefficient would result in completely ridiculous values ($< 10^{-3}$) which would result in a mole fraction solubility greater than 1 which is clearly nonsense. The addition of the group interaction terms made the predictions much more realistic and in many cases can provide a good approximation for the solubility. Fig. 7-27 shows the predictions for 3 different compounds with non-additive groups. Even with the group interactions the sucrose solubility prediction fails; compounds with a similar amount of OH groups and of a similar size to sucrose should therefore be predicted with caution. However for fructose and xylitol the solubility predictions are fair considering that they both fall far out of the infinite dilution approximation ($x \approx 0.1$ for both).

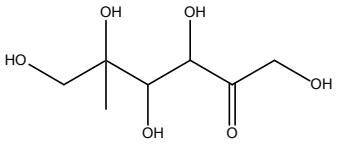
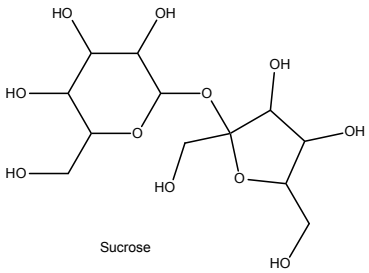
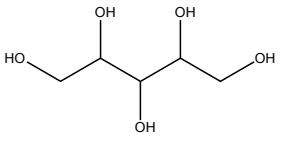
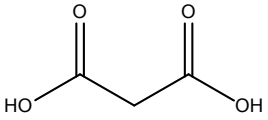
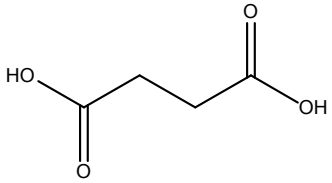
 Fructose	 Sucrose	 Xylitol
Experimental = 3600 g/l	**Experimental = 2100 g/l**	**Experimental = 642 g/l**
Predicted = 1640 g/l	**Predicted = fail (x > 1)**	**Predicted = 357 g/l**

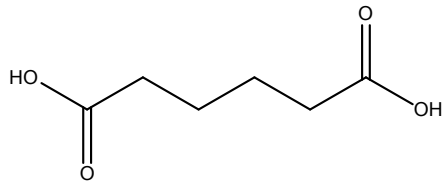
Fig. 7-27 Experimental and predicted values for the water solubility (at 298.15 K) of 3 compounds with non additive groups.

Similarly it was noted for the aliphatic COOH group (group 53) that the groups were non-additive. Initially they were considered to be additive since the only data available for dicarboxylic acids was for large compounds (e.g. decanedioic acid) where the contribution would be small and therefore seemingly unimportant. However since the aliphatic OH and COOH groups had to be treated very differently to the other groups in the previous work [20] it was decided to investigate this further. Very few data were available in the training set (even for $x > 0.01$) for small compounds with multiple COOH groups, however data were available in the literature for some compounds of interest [131, 145-147]. Since these data, in most cases, fell out of the assumed infinite dilution range ($x < 0.01$) they had not been added to the training set. Instead the group interactions were fitted to these data separately, since they lay outside the assumed infinite dilution range and may skew the regression. This is also an illustration of how simply the missing groups or interaction parameters can be estimated.

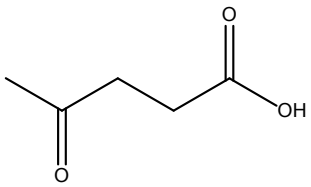
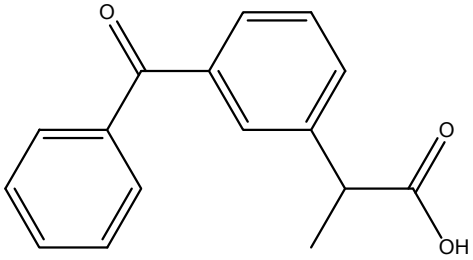
These values, in all but one case, improved the predictions of infinite dilution activity coefficient for the large dicarboxylic acids. The table of the predicted solubilities with and without group interaction is shown in Table 7-7 (as above, solubilities are used as a comparison since a failure in solubility prediction is obvious while a failure in activity coefficient prediction is not so obvious – regressions were done using the activity coefficient and not the solubility). It is clear from the table that the predictions for the substances containing the COOH-CO interaction were sufficiently accurate without the interaction term and therefore it was not included. As with the predictions done above, small compounds with many non-additive functional groups are still very difficult to predict and, as evident in Table 7-7, show moderate to large deviations in almost all cases. The interaction matrix for the non-additive groups is shown in Fig. 7-28.

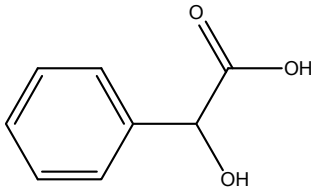
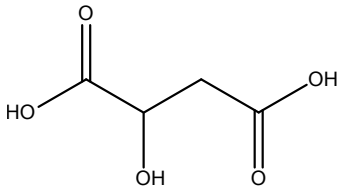
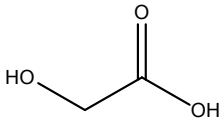
Table 7-7 Experimental and predicted solubilities for compounds containing non-additive groups including COOH (data from various sources [131, 145-147], new refers to prediction with fitted group interaction values while old refers to predictions without COOH group interactions).

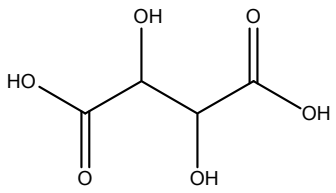
Main interaction	COOH-COOH		
Solute	S_{exp} (g/l)	S_{pred}^{old} (g/l)	S_{pred}^{new} (g/l)
 Malonic acid	1601.7	17294.6	1394.3
 Succinic acid	83.2	307.5	91.5



Adipic acid	30.8	40.4	15.7

Main interaction	**COOH-CO**		
Solute	**S_{exp} (g/l)**	**S^{old}_{pred} (g/l)**	**S^{new}_{pred} (g/l)**
	5224.1	6946.1	-
Levulinic acid			
	0.051	0.055	-
Ketoprofen			

Main interaction	**COOH-OH**		
Solute	**S_{exp} (g/l)**	**S^{old}_{pred} (g/l)**	**S^{new}_{pred} (g/l)**
	110	294.4	58.9
Mandelic acid			
	1390	fail (x>1)	3812.5
DL-Malic acid			
	2466.5	fail (x>1)	2527.2
Glycolic acid			



Tartaric acid

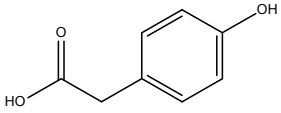
1391.7 fail (x>1) fail (x>1)

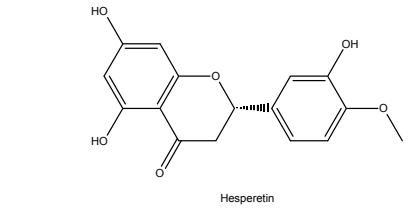
	OH	COOH	C=O
OH	x		
COOH	x	x	
C=O	x	o	x

Fig. 7-28 The Group interaction matrix for the infinite dilution activity coefficient in water (x – available value, o – no value available or in this case value neglected).

In all the cases considered above COOH and OH groups were only considered if they were attached to a non-aromatic carbon. This is because these interactions partially are accounted for, in aromatic compounds, by the *–ortho*, *–meta* and *–para* correction groups discussed below. One may however expect that for compounds which contain a large number of these aromatic groups that that there may need to be some group interaction term. This point is illustrated by considering the predictions for 3 such compounds shown in Table 7-8. The results for silybin indicate that the inclusion of the term does provide a much improved prediction even so the "new" solubility prediction is almost 2 orders of magnitude off which suggests that the data could be suspect (but this seems unlikely) or that predictions for similar compounds should be used with caution. For the other large molecule, hesperetin, the prediction is moderately improved when considering the aromatic groups as non-additive. The solubility prediction for 4-hydroxyphenylacetic acid is adversely effected when considering both groups to be non-additive. For this reason, and considering that all the other activity coefficient predictions for compounds with multiple OH and COOH groups are adequate in the training set, the aromatic OH and COOH groups will not be generally considered as non-additive. However for compounds which are either quite bulky (n > 15) and there are multiple groups it may be beneficial to include the group interaction term between the aromatic groups.

Table 7-8 Comparison of the predictions when including aromatic non-additive group interactions and when neglecting them (new refers to including the group interactions while old refers to neglecting them, data from various sources [131, 148, 149]).

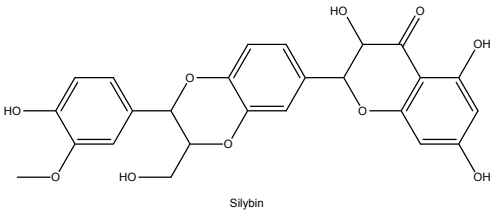
Molecular structure	Experimental/predicted solubility (g/l)	Experimental/predicted activity coefficient
 4-Hydroxyphenylacetic acid	$S_{pred,old} = 42.5$	$\ln \gamma^{\infty}_{pred,old} = 2.493$
	$S_{pred,new} = 8.66$	$\ln \gamma^{\infty}_{pred,new} = 4.071$
	$S_{exp} = 60.7$	$\ln \gamma^{\infty}_{exp} = 2.130$



$S_{pred,old} = 7.47$ $\quad \ln \gamma^{\infty}_{pred,old} = 3.255$

$S_{pred,new} = 0.039$ $\quad \ln \gamma^{\infty}_{pred,new} = 8.505$

$S_{exp} = 0.273$ $\quad \ln \gamma^{\infty}_{exp} = 6.561$



$S_{pred,old} = 65.94$ $\quad \ln \gamma^{\infty}_{pred,old} = 1.524$

$S_{pred,new} = 3.22$ $\quad \ln \gamma^{\infty}_{pred,new} = 4.540$

$S_{exp} = 0.054$ $\quad \ln \gamma^{\infty}_{exp} = 8.498$

An interesting effect was observed in the case of the aromatic rings, for some substituents the activity coefficient changed quite substantially depending on the position of the groups. Examples of this effect are shown in Fig. 7-29; as shown the *para*- position has the lowest activity coefficient while the *ortho*- position has the highest. This only occurs when at least 2 groups on the ring are one of the following: NH_2, OH, COOH or NO_2; this is shown by the 2 counter examples in Fig. 7-30.

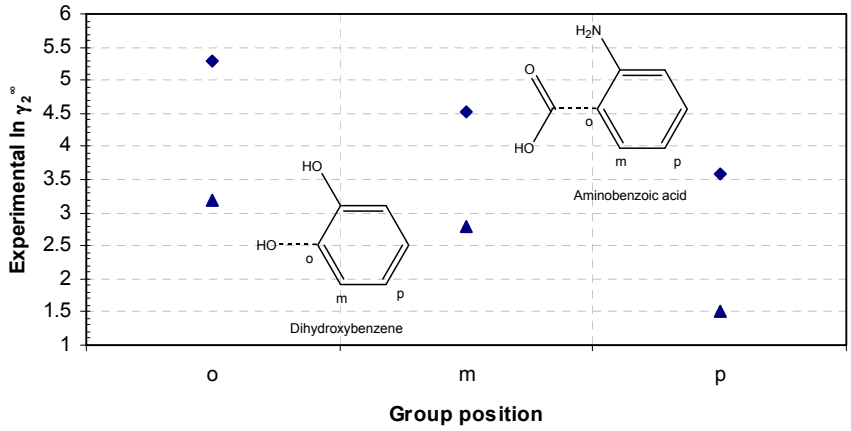


Fig. 7-29 Change in the activity coefficient with the position of the groups on the aromatic ring (◆ - aminobenzoic acid, ▲ – dihydroxybenzene).

There are probably other groups which would exhibit the same effect, such as nitrile or isocyanate groups, but no data was available for these groups. A possible explanation for this is that in the *para*- position the hydrophilic groups provide the maximum amount of shielding for the hydrophobic benzene ring (essentially disguising it) while in the *ortho*- position provides the minimum. Another explanation could be intramolecular hydrogen bonding, the *ortho*- position would be the most susceptible while the *para*- position would be the least. To account for this effect 3 new groups were added: group 156 – *ortho*- position, group 157 –

meta- position and group 158 – *para*- position. The priority♦ of the groups are 158 → 157 → 156 so for example 2,4,6-trinitrophenol would have a *para*- and a *meta*- group. Also once a correction is applied neither molecule must have the correction applied again, for example: 2,4-dinitrophenol would only have a *para*- group.

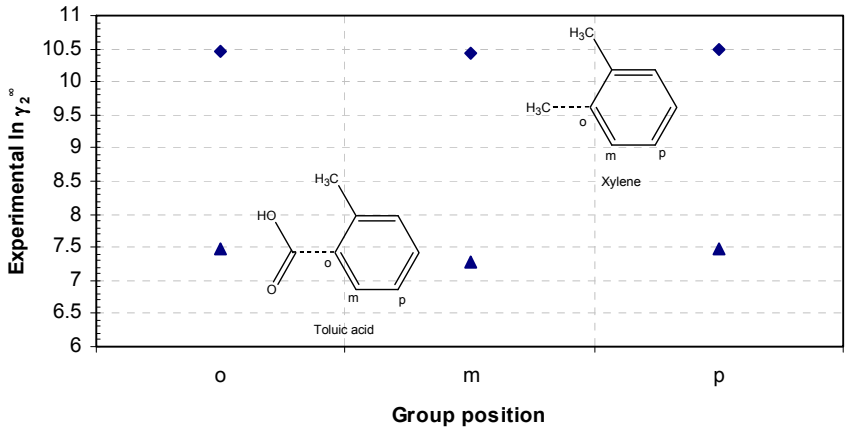
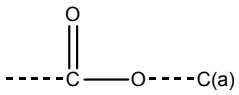
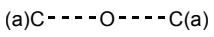
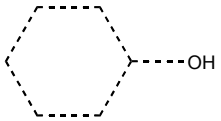


Fig. 7-30 Change in the activity coefficient with the position of the groups on the aromatic ring (◆ - xylene, ▲ – toluic acid).

Besides the *ortho*-, *meta*- and *para*- groups that were added there were 3 new groups added to account for some further observed effects, these groups are shown in Table 7-9. It may seem that group 163 should be applied to all secondary or tertiary carbon atoms but the data does not support this.

Table 7-9 New oxygen groups added.

Group #	Molecular structure	Name/description
159		**An ester attached to an aromatic carbon by the carbon atom, the dashed line can connect to any carbon atom**
160		**An oxygen atom joining 2 aromatic rings**
163		**A hydroxyl attached to a ring carbon**

♦ Group priority is the order in which groups are fragmented. It is necessary in order to make sure that the correct groups are fragmented. Consider the example of acetic acid (CH_3COOH), if the OH group had a higher priority than the COOH group would be fragmented as an OH and a CO group which would produce the wrong results. As a rule of thumb, the more complex the group is, the higher the priority.

7.8.3. Nitrogen compounds

As with the oxygenated compounds, it is typically difficult to predict properties of nitrogen containing compounds with good accuracy. For this work the error of the nitrogen compounds is about 2 percentage points higher than the overall average. The results for the nitrogen containing compounds are shown in Table 7-10. Something which is quite evident from the table is that for some of the chemical families (*viz*. nitrates and cyanides) very good predictions are obtained from the literature methods. This is in most cases because the data available for these groups is primarily infinite dilution data which is typically for smaller molecules which are well represented by many literature methods.

Table 7-10 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for nitrogen compounds (number of solutes in superscript).

Name	Eqn. (7-9)	Kuhne	UNIFAC	mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
All nitrogen compounds	9.2 [57]	55.2 [25]	11.8 [30]	17.7 [23]	37.7 [26]	34.1 [26]
All amines	12.6 [11]	13.0 [3]	16.2 [11]	26.4 [11]	79.2 [9]	31.8 [9]
Primary amines	13.7 [9]	14.8 [2]	16.8 [9]	26.0 [9]	73.5 [8]	24.1 [8]
Secondary amines	7.6 [1]	9.3 [1]	0.6 [1]	11.4 [1]	-	-
Tertiary amines	7.0 [1]	-	25.9 [1]	44.8 [1]	124.8 [1]	93.0 [1]
Cyanides	4.0 [7]	-	4.8 [6]	3.1 [6]	8.1 [7]	43.6 [7]
Amides	26.6 [2]	18.6 [1]	-	-	22.6 [2]	61.9 [2]
Oximes	0.0 [1]	-	-	-	5.3 [1]	9.5 [1]
All nitrates	6.5 [4]	-	3.2 [4]	4.7 [3]	29.0 [4]	17.3 [4]
Barbituates	10.3 [13]	88.9 [13]	-	-	-	-
Pyrrolidinediones	6.1 [4]	-	-	-	-	-

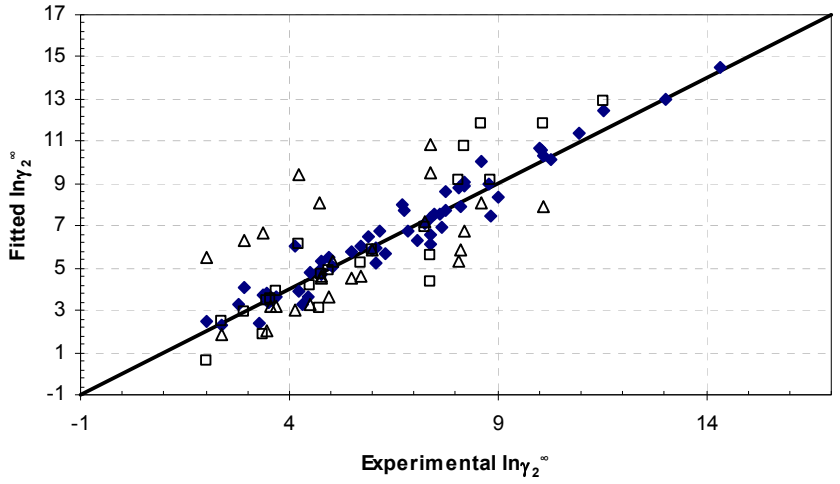


Fig. 7-31 Experimental and fitted lnγ_2^∞ for nitrogen compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

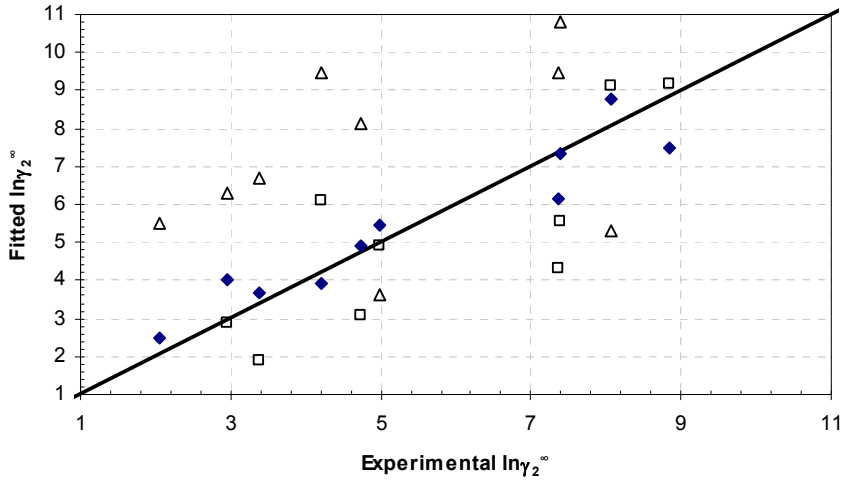
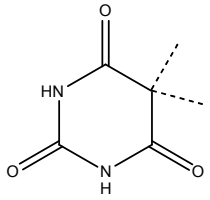
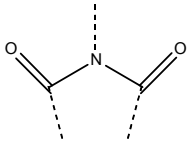
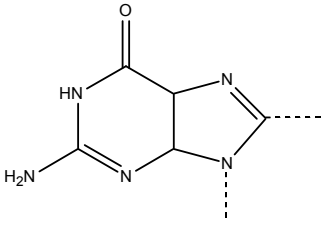


Fig. 7-32 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for amines (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

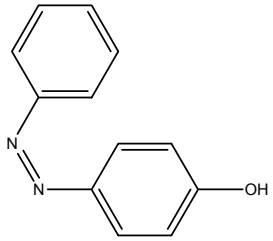
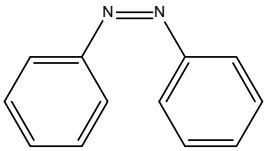
A frequent criticism of group contribution methods such as this one is that there is no allowance for proximity effects. This is partially accounted for with some groups by the inclusion of the *-ortho*, *-meta* and *-para* groups. However in many instances groups in close proximity are not equivalent to the sum of their parts and therefore are assumed to behave as a totally new group. A classic example of this the carboxylic acid group which is not equivalent to a hydroxyl and a ketone group together. Identifying the need for a new functional group is usually a fairly simple procedure since either, all compounds with the same groups tend to have a similar error or the fragmentation fails for the compounds. The new groups are outlined in Table 7-11.

Table 7-11 New nitrogen groups added.

Group #	Molecular structure	Name/description
165		**Barbituate group, the dashed lines can connect to any atom (including hydrogen)**
164		**Pyrrolidinedione group, the dashed lines from the carbons must connect via a "ring bond" while the dashed line from the nitrogen can connect to any atom (including hydrogen)**
162		**Guanine group, the dashed lines can connect to any atom (including hydrogen)**

Much discussion has been given in the preceding sections about the poor quality of data available necessitating the simplification of some groups which are not backed up by sufficient data. This was applied for the nitrogen groups in two instances, where previously there were groups for both ring and chain amines in this work only a single group was used (groups 82 & 84). In some instances it is simply not possible to simplify a group. An example of this is azene; there was only one compound in the training set that contained an azene group (p-hydroxyazobenzene), however when this group is used to predict the activity coefficient of another azene containing compound (azobenzene, data found on Chemspider [131]) the results are very good. One of the reasons why this works so well is that the data for the compound in the training set was obviously good and therefore allowed for a realistic value of the group to be found. What this means is that if there are good data available for some compounds containing more exotic groups the method can be extended to these groups by calculating the group value from the reliable data point.

Table 7-12 Experimental and fitted/predicted activity coefficient @ 298.15 K for two azenes (p-hydroxyazobenzene from Beilstein [59] and azobenzene from Chemspider [131]).

Compound	Experimental – fitted/predicted results
 p-Hydroxyazobenzene	$\ln \gamma^{\infty}_{exp} = 10.362$ $\ln \gamma^{\infty}_{fit} = 10.362$
 Azobenzene	$\ln \gamma^{\infty}_{exp} = 13.241$ $\ln \gamma^{\infty}_{pred} = 13.583$

7.8.4. Sulfur compounds

There were few data available for sulfur compounds in this work (i.e. for hydrocarbons with only sulfur groups). This is not ideal since data of this type is typically the best source of group values as the groups are essentially isolated (assuming, as typically is the case, that the hydrocarbon group contributions are well known and accurate). For most of the sulfur compounds the fits are very good. The percentage error comparisons for the various methods available in literature are shown in Table 7-13.

Table 7-13 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for sulfur compounds (number of solutes in superscript).

Name	Eqn. (7-9)	Kuhne	UNIFAC	mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
All sulfur compounds	7.6 [7]	28.3 [2]	-	-	14.7 [5]	18.6 [5]

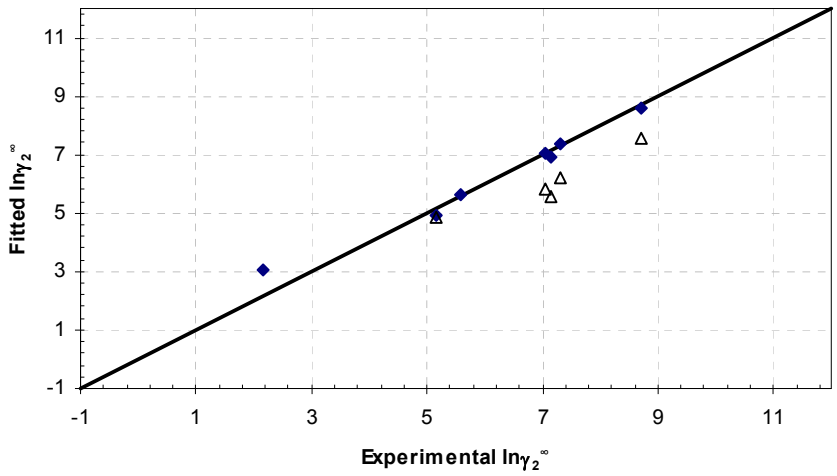
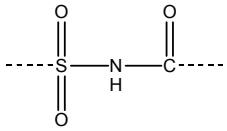
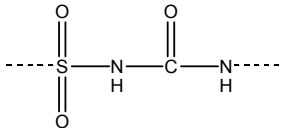


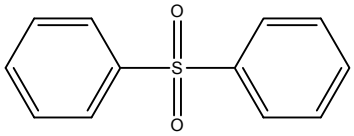
Fig. 7-33 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for sulfur compounds (◆ - This work, △ – COSMO-RS(OL)).

Table 7-14 New sulfur groups added.

Group #	Molecular structure	Name/description
170		**Sulfone-amide group, the dashed lines can connect to any atom (including hydrogen)**
171		**Sulfone-urea, the dashed lines can connect to any atom (including hydrogen)**

Two new groups were added for the sulfur compounds, one for a sulfone joined to an amide group and another for a sulfone joined to a urea group (see Table 7-14). As mentioned above there were very few data available for the sulphur compounds and more specifically for hydrocarbons with only sulphur groups. The literature was searched for at least one test compound for any of the groups with only 1 or 2 data points. The limiting factor was the availability of both fusion and solubility data. One source was found [150] which contained fusion data for a suitable test compound, the result of the prediction is shown in Table 7-15. The results for the prediction are excellent and illustrate the point discussed above for azobenzene, that even if a group is fitted to only one or two data points the predictions can still be good as long as the training data are of a high quality.

Table 7-15 Experimental and fitted/predicted activity coefficient @ 298.15 K for diphenylsulfone [150].

Compound	Experimental – fitted/predicted results
 Diphenylsulfone	$\ln\gamma_{exp}^{\infty} = 11.035$ $\ln\gamma_{fit}^{\infty} = 11.219$

7.8.5. Halogen compounds

As with the previous work [20] the halogens could be fitted very well. This is even though some of the halogen groups used previously have been removed due to the reasons discussed above. The relative mean deviations for the halogens are shown in Table 7-16 and the error plots are shown in the figures following.

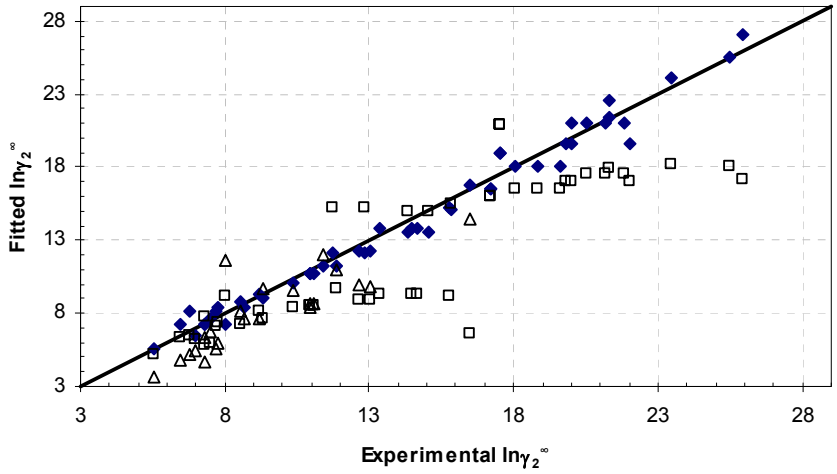


Fig. 7-34 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for chlorine compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

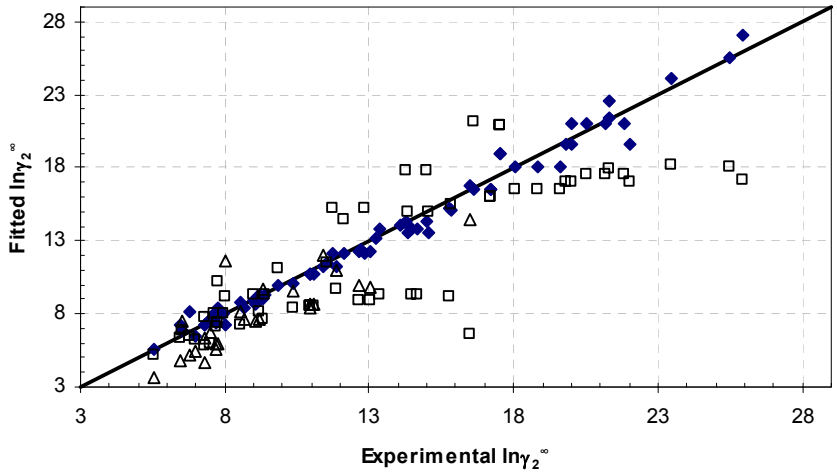


Fig. 7-35 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for halogen compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

Table 7-16 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for halogen compounds (number of solutes in superscript).

Name	Eqn. (7-9)	Kuhne	UNIFAC	mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
Halogenated compounds	3.8 [69]	7.2 [42]	36.8 [62]	17.1 [59]	19.3 [25]	20.8 [25]
Fluorinated	-	-	-	-	-	-
Chlorinated	4.4 [51]	7.4 [33]	43.8 [50]	18.3 [47]	19.2 [23]	20.6 [23]
Brominated	1.8 [11]	8.6 [4]	5.7 [11]	13.6 [11]	-	-
Iodinated	2.5 [4]	5.5 [2]	-	-	20.0 [2]	23.2 [2]

Due to insufficient good data all the halogen double bonded carbon groups were removed. As for the reasons mentioned previously this is not because they are not needed but because there are insufficient data to back them up.

A large drawback with group contribution methods is that they can only be applied to compounds which contain groups that were in the training set. One way to counter this is to keep the method dynamic by continually adding new group values (this is discussed in section 7.8.3). Another way is to predict unknown groups from groups which may be expected to behave similarly. Since the halogens are all part of the same periodic group it is assumed here that they all behave in the same way and only depend on the size of the atom (for the molecular halogens consider the plots of T_m and T_b against molecular size shown in Fig. 7-36). While this assumption may not be totally accurate (one may expect that fluorine could be an outlier due to its ability to form hydrogen bonds) the examples shown in the paragraphs following seem to suggest that this assumption it is sufficiently accurate for this purpose. The case in point is shown in Fig. 7-37, where the group values for the groups 145, 44 and 39 are plotted against the square of the atomic radius of bromine, chlorine and fluorine atoms respectively. The square of the atomic radius is a measure of the surface area of the atom.

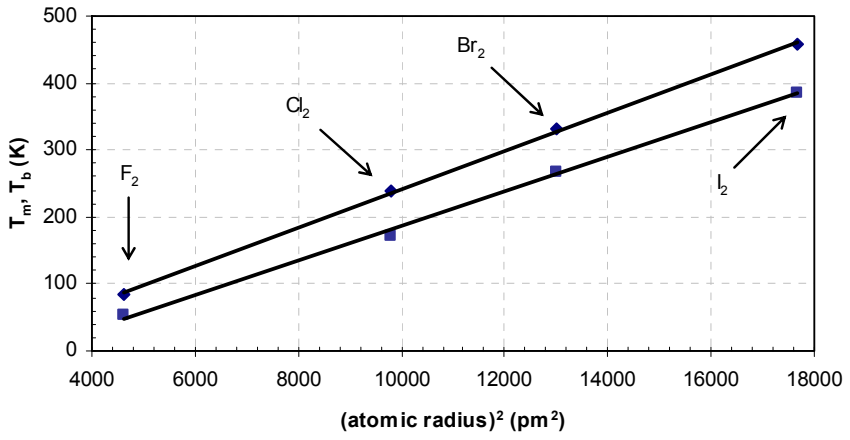


Fig. 7-36 T_m and T_b vs. atomic radius squared for molecular halogens (◆ – T_b (K), ■ – T_m (K), — linear fits to the data, all data from the DDB [28]).

Quite surprisingly there were no fluorinated hydrocarbon (i.e. compounds with only C, H and F atoms – and possibly no H) data available; this means that the fluorine groups are very susceptible to errors. The plots for the aromatic and non-aromatic halogen groups are shown

in Fig. 7-38 and Fig. 7-39 respectively. For the aromatic halogen groups shown in Fig. 7-38, group 37 seems to be the outlier (group 41 and 47 weren't considered as outliers due to the large amount of data behind the groups).

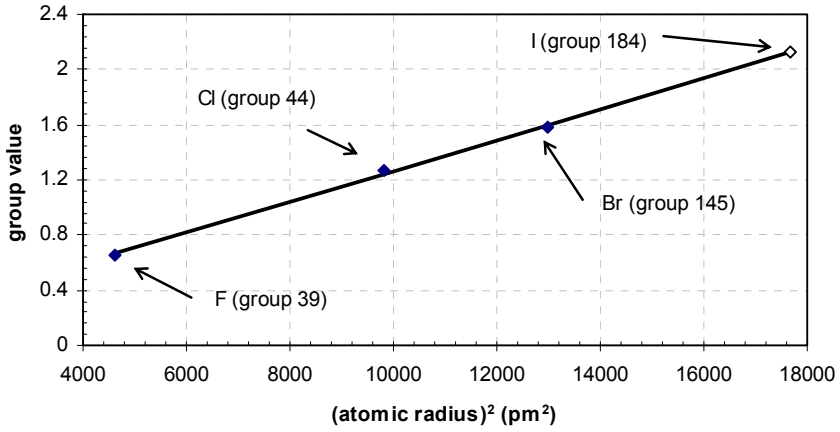


Fig. 7-37 Group value vs. atomic radius squared for groups 145, 44 and 39 (halogen atom attached to an carbon with at least 2 other halogens) (◆- fitted group value, ◇ - new group value).

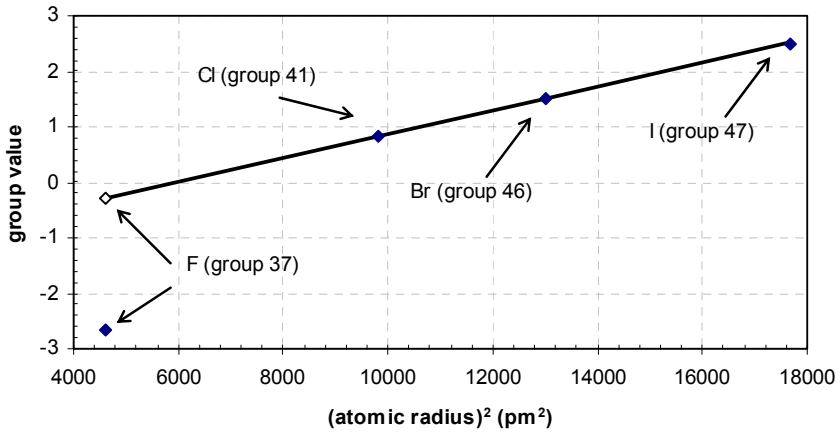


Fig. 7-38 Group value vs. atomic radius squared for groups 47, 46, 41 and 37 (halogen atom attached to an aromatic carbon) (◆ - fitted group value, ◇ - new group value).

In order to test a worst case scenario (prediction for a compound with a large amount of fluorine groups) perfluoronaphthalene and diflunisal were used (flurbiprofen was also used, with fusion data from literature [151]), the results are shown in Table 7-17. The results clearly show that the old values produce nonsense results and therefore the new value is recommended (and indeed used). The application of the new value, however, causes all 3 compounds contained in the training set to give significantly worse errors. Nevertheless since this value has performed so well for the compounds in Table 7-17 (which contain fewer other functional groups than the training set data) it was decided to use the physically realistic value. This same assumption could not be tested with group 35 due to the absence of any data for compounds with more than one group. Due to groups 38 and 39, very few

compounds actually have multiple occurrences of group 35 and therefore the physically realistic value could not be tested. Nevertheless it worked so well for the aromatic group and therefore the physically realistic value was used.

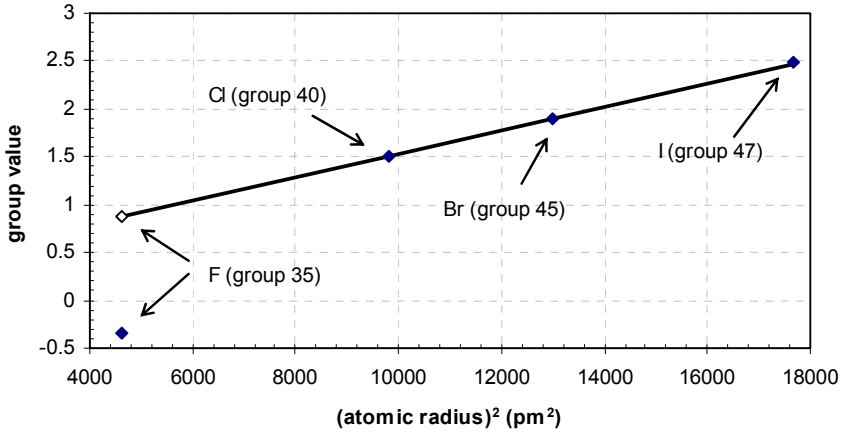
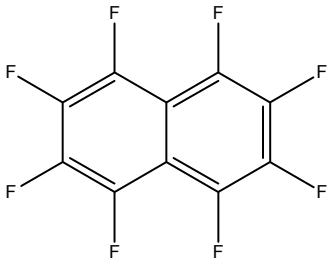
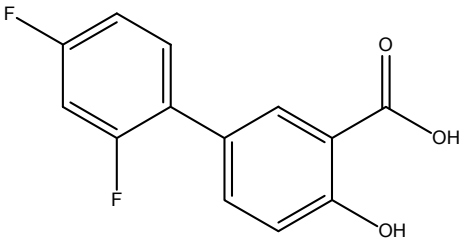
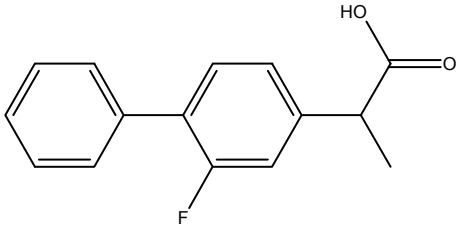


Fig. 7-39 Group value vs. atomic radius squared for groups 47, 45, 40 and 35 (halogen atom attached to a non-aromatic carbon) (◆ - fitted group value, ◇ - new group value).

Table 7-17 Experimental and predicted results for 3 compounds containing group 37 (subscript old refers to the old value of the group and new refers to the new physically realistic value, data for flurbiprofen from Chemspider [131] and diflunisal from literature [152]).

Molecular structure	Experimental/predicted solubility (g/l)	Experimental/predicted activity coefficient
 Perfluoronaphthalene	$S_{pred,old}$ = failed (x>1) $S_{pred,new} = 0.0041$ S_{exp} = N/A	$\ln \gamma^{\infty}_{pred,old} = -4.721$ $\ln \gamma^{\infty}_{pred,new} = 14.067$ $\ln \gamma^{\infty}_{exp}$ = N/A
 Diflunisal	$S_{pred,old} = 1.936$ $S_{pred,new} = 0.0177$ $S_{exp} = 0.0083$	$\ln \gamma^{\infty}_{pred,old} = 4.545$ $\ln \gamma^{\infty}_{pred,new} = 9.243$ $\ln \gamma^{\infty}_{exp} = 10.000$
 Flurbiprofen	$S_{pred,old} = 0.076$ $S_{pred,new} = 0.007$ $S_{exp} = 0.008$	$\ln \gamma^{\infty}_{pred,old} = 9.933$ $\ln \gamma^{\infty}_{pred,new} = 12.280$ $\ln \gamma^{\infty}_{exp} = 12.180$

In previous works of this group [15-18, 20] the iodine groups were all represented by one unified group (group 47) due to the low availability of data for iodinated compounds. When examining Fig. 7-37 and Fig. 7-38 it shows that the values for both the aromatic and non-aromatic groups could be very similar anyway. As this physical realism can be used to reproduce the values of these groups so well, a whole new set of group contributions were added in a similar way to the other halogen groups. Unfortunately there are no data available to test this assumption, but considering that the physical realism has worked so well for the other groups there is a good argument for adding them.

In Fig. 7-37 and Fig. 7-39 the fitted curves had a slope of approximately 1.2×10^{-4} pm^{-2} but when the values for groups 144 and 43 (halogen atom attached to a carbon with 1 other halogen) were plotted a slope of 2.2×10^{-4} pm^{-2} was found. Since there is no plausible explanation for this, the slope was fixed as the average between the other two values (1.17×10^{-4} pm^{-2}) and the intercept was calculated from group 43 (since it has much more data to back it up). This resulted in a new value for group 144 (Br attached to a carbon with another halogen), which improved the fit in the case of one of the compounds while reducing it for the other, and enabled a group contribution value to be assigned to group 38.

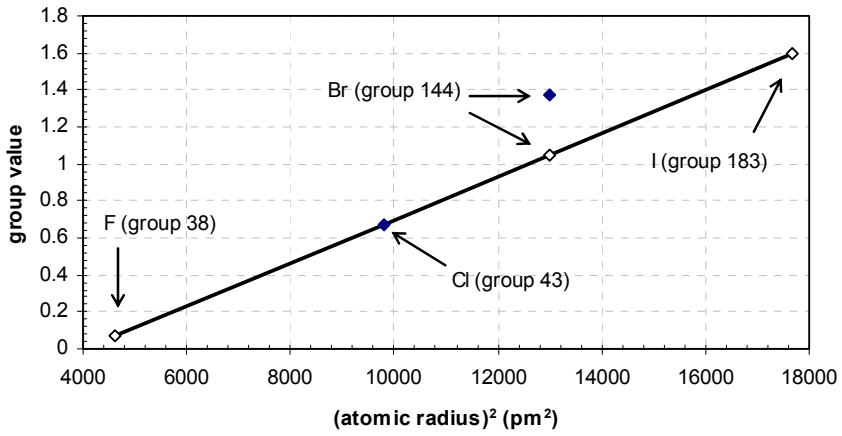


Fig. 7-40 Group value vs. atomic radius squared for groups 144 and 43 (halogen atom attached to a carbon with 1 other halogen) (◆ - fitted group value, ◇ - new group value).

7.8.6. Phosphorus compounds

Unlike the previous sections this section covers not only hydrocarbons which contain a phosphorus group but any compound which contains phosphorus groups. The reason for this is that there is only one compound which would have fallen into the first category. The results for the phosphorus compounds are shown in Fig. 7-41 and Table 7-18. In the previous work [20] the concept of group simplification was introduced, whereby groups should be defined in such a way that they are simple enough to have wide applicability but not so simple so as to

lose any physical meaning. The advantage of this type of approach is that it reduces the variety of groups needed in the training set and also means that more data is behind each group which leads to more physically realistic group values. The simplified phosphorus groups are shown in Table 7-19.

Table 7-18 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for phosphorus compounds (number of solutes in superscript).

Name	Eqn. (7-9)	Kuhne	UNIFAC	mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
All phosphorus compounds	6.3 [12]	23.1 [2]	-	-	-	-

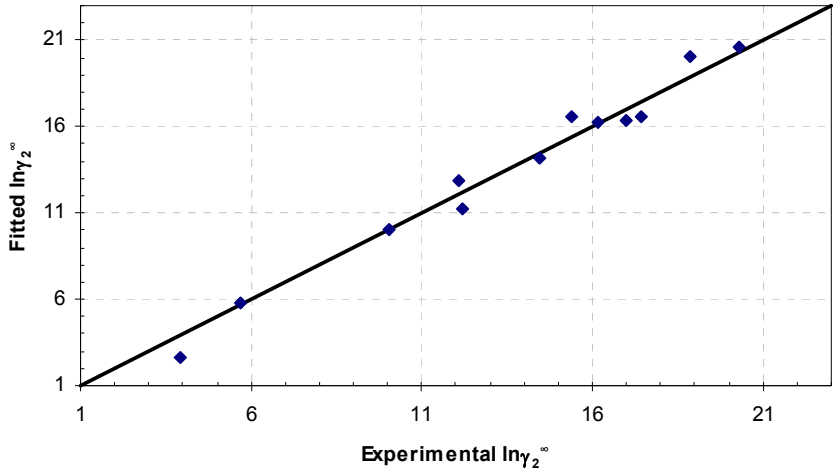
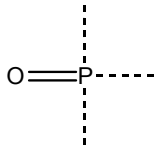
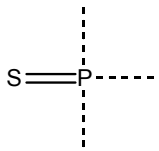
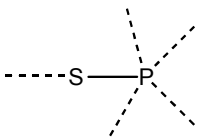
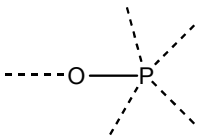


Fig. 7-41 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for phosphorous compounds (◆ - This work).

Table 7-19 Simplification of the phosphorus groups.

Group number	Group structure	Description
97		**Oxygen double bonded to a phosphorus (includes both)**
166		**Sulfur double bonded to a phosphorus (includes both)**
167		**Sulfur bonded to a phosphorus (only includes the sulfur) and another atom**
168		**Oxygen bonded to a phosphorus (only includes the oxygen) and another atom**

7.8.7. Multi-functional compounds

This section is applicable to all compounds which do not fall into only one of the above mentioned chemical families. Any discussion involving multi-functional compounds in this chapter does not refer to compounds containing more than one functional group but rather compounds with at least 2 different functional groups (i.e. not both oxygen or nitrogen etc. groups). This family of compounds exhibits the largest errors of all the literature methods and in the case of UNIFAC and mod. UNIFAC less than half of the compounds could be predicted due to missing group parameters. The results are shown in Fig. 7-42 and Table 7-20 and are testament to the fact that the literature methods perform quite poorly for these compounds.

Table 7-20 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for multifunctional compounds (number of solutes in superscript).

Name	Eqn. (7-9)	Kuhne	UNIFAC	mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
Multi-functional compounds	8.7 [155]	32.6 [85]	46.1 [76]	27.8 [54]	30.6 [30]	52.2 [30]

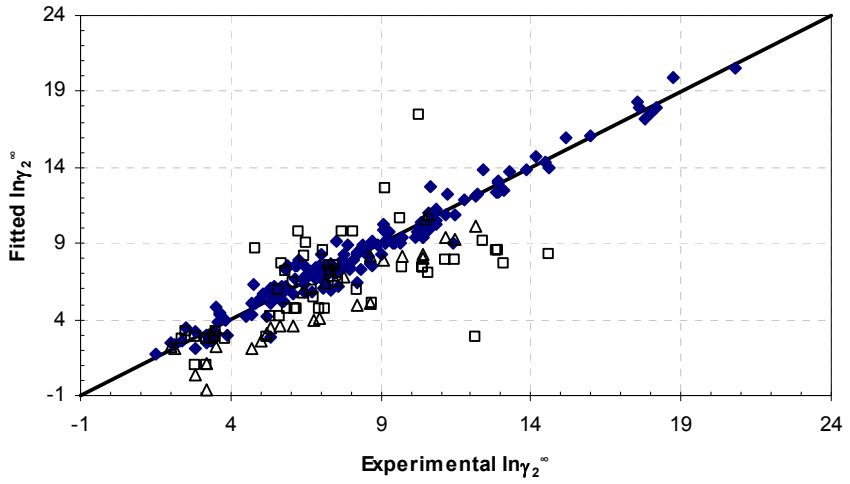


Fig. 7-42 Experimental and fitted lnγ_2^∞ @ 298.15 K for multifunctional compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – COSMO-RS(OL)).

7.8.8. Missing group values

While group contribution methods are very useful in that they can make predictions about any compound for which group contributions are available, the disadvantage is that if groups are unavailable then predictions cannot be made. This problem was partially addressed for the halogen groups in section 7.8.5. The approach used for the halogen groups is very specific and cannot be readily applied to other families of groups. The problem is therefore making this group prediction generally applicable. For non-halogen groups it is assumed that the group value would depend on both the size and the polarity of the group.

Since almost all of these groups are made up of more than one atom (incl. hydrogen) the atomic radius was not used as a measure of the size. In order to keep the calculations simple the molar mass was used as a measure of the size. The group electronegativity [153] was used as a measure of the polarity of the group and some values are shown in Table 7-21. As seen in Fig. 7-43 when group 99 (-SH group) is not considered in the fit there is a moderate deviation. The fact that group 99 is such a large outlier is worrying since it is unclear whether this correlation is generally applicable or not. The data behind group 99 are more than likely good data as they are from different sources and can all be fitted very well. The plot for the groups attached to aromatic carbons (Fig. 7-44) also adds to the uncertainty since there is a general agreement with the non-aromatic trend but the scatter is quite large. However, if a value of a group is needed then this correlation could at least provide a value which would allow for a prediction to be made. Unfortunately, no fusion data could be found for any compound with a group that was missing. Nevertheless, even though this approach is fairly "rough-and-ready" it seems that approximate group values can be found if a compound contains any exotic group.

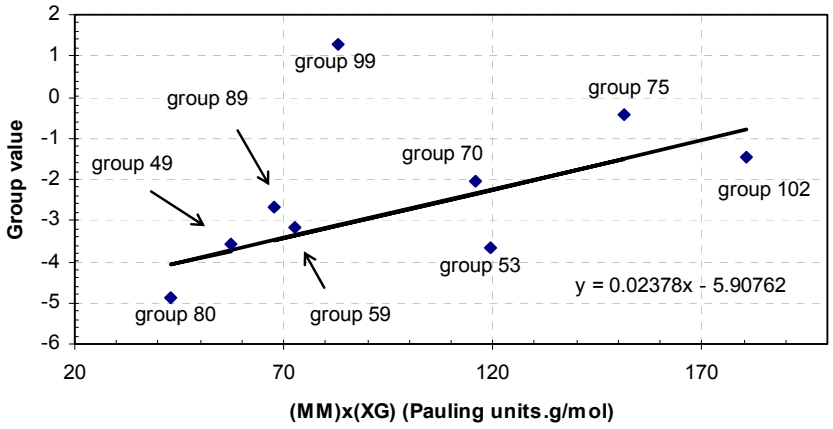


Fig. 7-43 Group value for groups attached to non-aromatic atoms vs. the product of molar mass (MM) and electronegativity (XG) (values of the electronegativities taken from [153]).

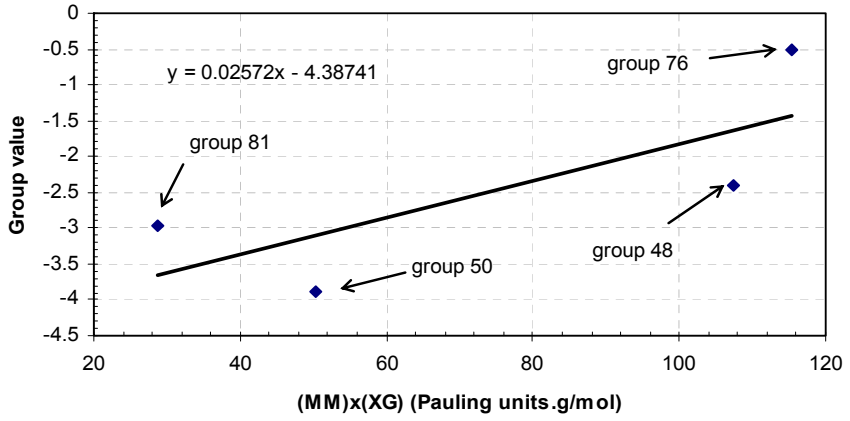


Fig. 7-44 Group value for groups attached to aromatic atoms vs. the product of molar mass (MM) and electronegativity (XG) (values of the electronegativities taken from [153]).

Table 7-21 Group electronegativities from literature [153] for some groups in this work ((a) refers to an aromatic group).

Grp	Formula	XG (Pauling units)	MM (g.mol^{-1})
77	-ONO	3.444	46
73 or 141(a)	-OCN	3.407	42
49 or 50(a)	-OH	3.364	17
75 or 76(a)	-NO2	3.294	46
82	>NH	3.037	15
74	-NCO	2.978	42
106	-NCS	2.91	58
80 or 81(a)	-NH2	2.878	16
102	>S(=O)2	2.825	64
105	>S=O	2.748	48
58 or 57(a)	>C=O	2.706	28
53 or 48(a)	-COOH	2.655	45
70	-CONH2	2.631	44
89	-CN	2.599	26
59 or 60(a)	CHO	2.597	29
99	-SH	2.51	33

7.8.9. Overall results

In total there were 630 compounds that were used in the training set with an overall average percentage error of 7.3 % for the logarithm of the activity coefficient at 298.15 K. In total, when comparing with the other methods available in literature it is fairly obvious that the percentage errors for these methods would be a lot higher than the method developed in this work because much of this data is a blind prediction for these methods while the new method has been directly fitted to these data. Nevertheless, the considerably lower error of this work coupled with the fact that it is more widely applicable are additional justifications of this work. The total number of model parameters used in this work is 96 (1 constant, 92 group contributions and 3 group interactions). 100 group values are reported in Table 7-23 because 4 values were taken from the physically realistic values that were found for the halogen groups (see section 7.8.5).

When looking at the relative and absolute deviations for the compounds in the training set, shown in Fig. 7-45 and Fig. 7-46 respectively, it is apparent that the vast majority of the errors are very low for the training set with only a few data with larger errors. The larger errors are typically for small compounds which are not the objective of this work (for example dioxane is the compound with the highest RMD for this work with 64%). When this is compared to the relative mean deviation histogram for mod. UNIFAC (which performed considerably better than the other 4 literature methods) it is apparent that the errors for mod. UNIFAC are much more spread over the percentage range. One possible complaint with the comparison of these results is that the mod. UNIFAC parameters that were fitted to SLE data were fitted using Eqn. (5-16), while the activity coefficient data for these comparisons is extracted from SLE using Eqn. (5-18). One may therefore expect that using Eqn. (5-16) and comparing with the mod. UNIFAC predictions would give marginally better results. This is not the case; in fact the overall error when using Eqn. (5-18) for all the data in the training set rises to 16.4 %. This

confirms that there is a distinct advantage when using Eqn. (5-18) and this could perhaps be used for future mod. UNIFAC matrix fits.

Table 7-22 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for all compounds in the training set (number of solutes in superscript).

Name	Eqn. (7-9)	Kuhne	UNIFAC	mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
All compounds	7.3 [630]	25.2 [277]	23.7 [442]	14.9 [396]	23.4 [295]	30.0 [295]

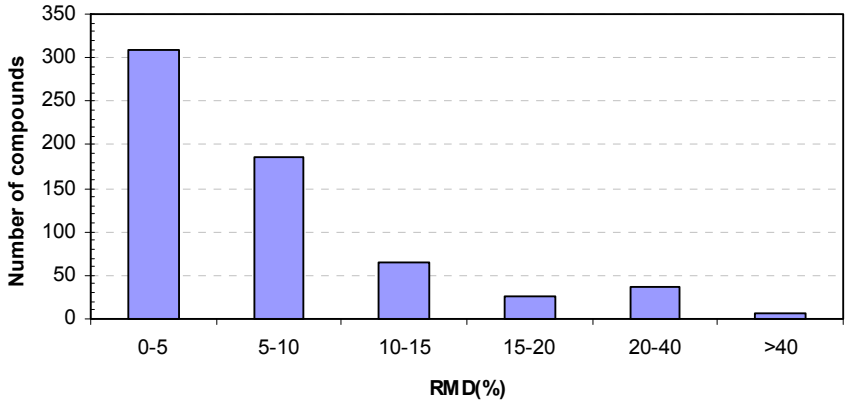


Fig. 7-45 Histogram of this works RMD (%) in $\ln\gamma^{\infty}$ for the compounds in the training set.

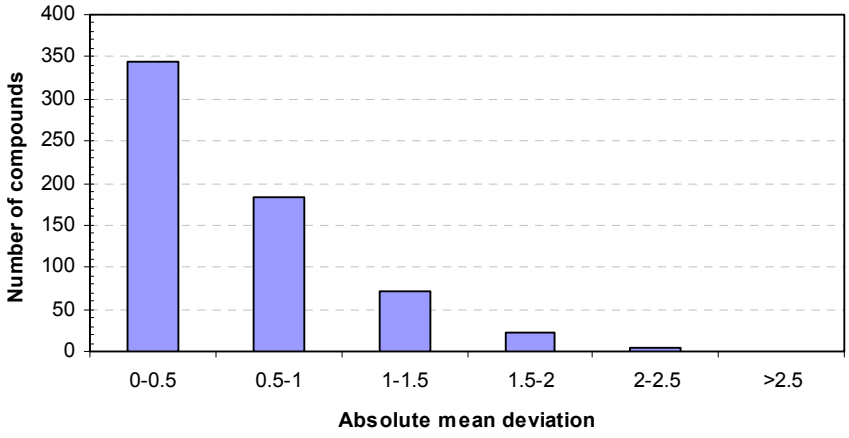


Fig. 7-46 Histogram of this works absolute mean deviations in $\ln\gamma^{\infty}$ for the compounds in the training set.

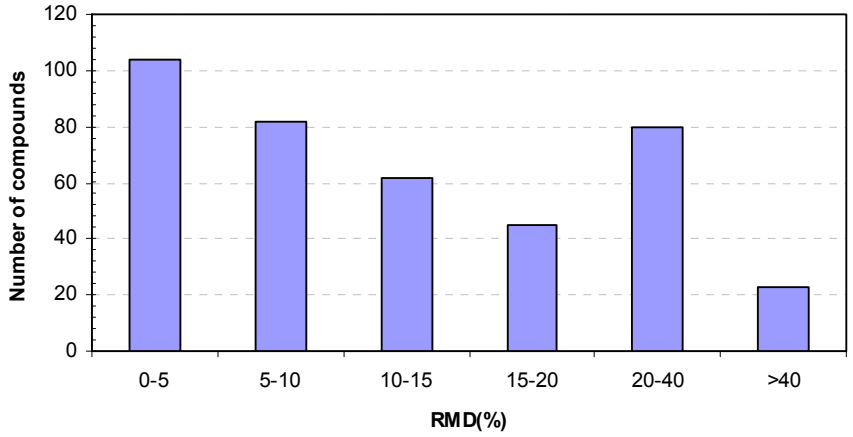


Fig. 7-47 Histogram of the mod. UNIFAC RMD's (%) in $\ln\gamma^{\infty}$ for the compounds in the training set.

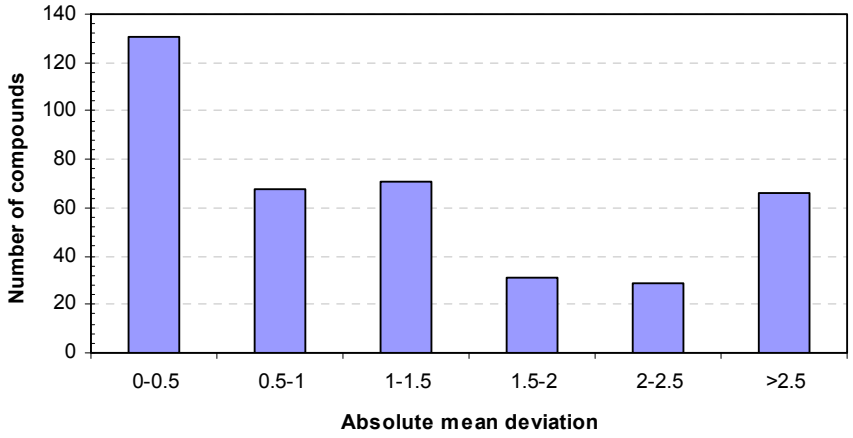


Fig. 7-48 Histogram of the mod. UNIFAC absolute mean deviations in $\ln\gamma^{\infty}$ for the compounds in the training set.

Table 7-23 The group contribution and group interaction values for the infinite dilution activity coefficient in water at 298.15 K using Eqn. (7-9) (lnk No – the number which is used by the fragmentation program to identify the group, NA – number of non-hydrogen atoms in the group, Prty – Priority – order in which the groups are fragmented, [a] – group value only fitted to one data point, [b] – group value only fitted to two data points, [c] – group value taken from the physically realistic value).

$$\ln\gamma_2^{\infty} = \sum_i v_i C_i + \frac{1}{2}\sum_i\sum_j \frac{GI_{i-j}}{n(m)}$$

lnk No	Name	NA	Description	C_i	Example	Prty	Ref No
	Constant	0	Constant	4.21523	All Compounds	0	0
	Aliphatic carbon groups						
1	CH3-(ne)	1	CH3 attached to a non-aromatic non-electronegative atom	0.70225	Hexane	124	101
4	-C(c)H2-	1	CH2 in a chain	1.21445	Decane	130	102
5	>C(c)H-	1	CH in a chain	2.40952	2-Methylbutane	133	103
6	>C(c)<	1	C in a chain	2.34305	2,2,4-Trimethylpentane	136	104
2	CH3-(e)	1	CH3 group attached to a non-aromatic electronegative atom	-0.28185	N,N-Dimethylformamide (DMF)	121	105
7	-CH2(c)-(e)	1	CH2 in a chain attached to an electronegative atom	0.00544	Ethanol	125	106
8	>CH(c)-(e)	1	CH in a chain attached to an electronegative atom	0.16648	2-Butanol	126	107
9	>C(c)<(e)	1	C in a chain attached to an electronegative atom	-0.28270	tert-Butanol	127	108
10	-C(r)H2-	1	CH2 in a ring	0.97748	Cyclohexane	132	110
11	>C(r)H-	1	CH in a ring	-0.02468	Methylcyclohexane	135	111
12	>C(r)<	1	C in a ring	0.64894	alpha-Pinene	134	112

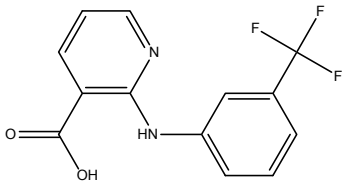
24	-C(r)H2-(en)	1	CH2 in a ring attached to an electronegative carbon	-0.22417	1,4-Dioxane	131	116
14	>CH(r)-(e,c)	1	CH in a ring attached to an electronegative atom	-0.19258	Cyclopentanol	128	117
15	>C(r)<(e,c)	1	C in a ring attached to an electronegative atom	1.76747	Perfluorocyclopentane	129	118
26	C(c)H2=C(na)<	1	Double bonded carbon at the end of a chain/ring	0.67054	beta-Pinene	110	120
20	-C(c)H=C(c)-	1	Double bonded carbon in a chain with only 1 carbon neighbour	1.37481	2-Heptene	113	121
27	-C(c)=C(c)<	1	Double bonded carbon in a chain with 2 carbon neighbours	1.44074	2-Methyl-2-butene	111	122
21	-C(r)H=C(r)H-	1	Double bonded carbon in a ring with 1 carbon neighbours	0.81532	Cyclohexene	114	125
13	-C(r)=C(r)<	1	Double bonded carbon in a ring with 2 carbon neighbours	0.97808	1-Methylcyclohexene	112	126
25	C(c)H#C(c)-	1	Carbon triple bonded to another carbon at the end of a chain	-0.98929	1-Octyne	115	129
22	-C(c)#C(c)-	1	Triple bond between 2 carbons in a chain	1.44056	2-Heptyne	116	130
	Aromatic carbon groups						
3	CH3-(a)	1	CH3 group attached to an aromatic atom	0.90181	Toluene	122	201
16	-CH(a)<	1	CH in an aromatic ring	0.59361	Benzene	119	202
17	>C(a)<	1	C in an aromatic ring	0.96000	m-Xylene	120	203
19	(a)=C(a)<2(a)	1	Aromatic carbon attached to three aromatic neighbours	1.04993	Naphthalene	109	204
18	>C(a)<(e)	1	C in an aromatic ring attached to an electronegative atom	1.26747	Aniline	118	209
	Fluorine groups						
35	F-C(na)	1	F attached to nonaromatic carbon	0.87161[c]	2-Fluoropropane	95	301
38	F-C(na)-1Halo	1	F attached to a carbon with one other halogen atom	0.06948[c]	Perfluoropentane	76	302
39	F-C(na)-2Halo	1	F attached to a carbon with at least two other halogen atoms	0.65063	Perfluorohexane	75	303
37	F-C(a)	1	F attached to aromatic carbon	-0.29836[c]	Fluorobenzene	94	306
	Chlorine groups						
40	Cl-C(na)	1	Cl attached to noaromatic carbon	1.50420	cis-1,2-Dichloroethylene	79	401
43	Cl-C(na)-1Halo	1	Cl attached to a carbon with one other halogen atom	0.67603	Dichloromethane	74	402
44	Cl-C(na)-2Halo	1	Cl attached to a carbon with at least two other halogen atoms	1.27165	Tetrachloromethane	73	403
41	Cl-C(a)	1	Cl attached to aromatic carbon	0.82697	Chlorobenzene	80	406
	Bromine groups						
45	Br-C(na)	1	Br attached to a nonaromatic carbon	1.89460	Ethyl bromide	81	501
144	Br-C(na)-1Halo	1	Br attached to a Carbon with one other halogen atom	1.05036[c]	2-Bromo-2-chloro-1,1,1-trifluoroethane	78	502
145	Br-C(na)-2Halo	1	Br attached to a Carbon with at least two other halogen atom	1.58246[a]	Bromochlorodifluoromethane [R12B1]	77	503
46	Br-C(a)	1	Br attached to aromatic carbon	1.52146	Bromobenzene	82	506
	Iodine groups						
47	I-C(na)	1	I attached to a nonaromatic carbon	2.46805[b]	Ethyl iodide	69	601
183	I-C(na)-1Halo	1	I attached to a Carbon with one other halogen atom	1.60021[c]	Diiodomethane	67	602
184	I-C(na)-2Halo	1	I attached to a Carbon with at least two other halogen atom	2.12684[c]	Perfluoro-n-amyliodide	68	603
169	I-C(a)	1	I attached to an aromatic carbon	2.54158	Iodobenzene	70	606
	Oxygen groups						
53	C(na)-COOH	3	COOH Group attached to a carbon	-3.64344	Acetic acid	34	701
48	C(a)-COOH	3	Aromatic COOH	-2.40491	Benzoic acid	33	703
49	C(c)-OH	1	OH Group attached to a chain carbon	-3.55430	1-Hexanol	97	704
50	C(a)-OH	1	Aromatic OH	-3.89491	Phenol	98	706
51	C(na)-O-C(na)	1	Ether oxygen	-0.27038	Diethyl ether	101	707
52	C(a)-O(a)-C(a)	1	Aromatic oxygen	0.99727	Furfural	99	708
54	C(c)-COO-C(c)	3	Ester in a chain	-0.96841	Ethyl acetate	35	709
55	HCOO-C(c)	3	Formic acid ester	-0.69281	Methyl formate	37	710
56	C(r)-C(r)OO-C(r)	3	Ester in a ring (lactone)	3.30768	epsilon-Caprolactone	36	711
57	O=C(a)<	2	Ketone bonded to aromatic ring	-0.89116	Acetophenone	64	712
58	O=C(na)<	2	Ketone	-3.14373	Acetone	65	713
59	HCO-C(na)	2	Aldehyde in Chain	-3.18973	Acetaldehyde	63	714
60	HCO-C(a)	2	Aldehyde attached to an aromatic ring	0.10968	Benzaldehyde	62	715
63	(-C=O-O-C=O-)r	7	Cyclic anhydrides with double or aromatic bond	-1.30601[b]	Maleic anhydride	19	718
64	>(OC2)<	3	Epoxide	-2.18496	Ethylene oxide	60	719
68	>N-(C=O)-N<	4	Urea	-2.19581	1,1,3,3-Tetramethyl urea	13	723
69	-OCON<	4	Carbamate	-2.47862	Methyldimethylcarbamate	12	724
159	A_Ester	3	Ester attached to an aromatic carbon (via the carbon atom)	1.18291	Benzoic acid methyl ester	6	725
163	C(r)-OH	1	OH Group attached to a ring carbon	-3.02818	Cyclododecanol	96	726
160	C(a)-O(r)-C(a)	1	Ether oxygen (ring) connencting 2 aromatic rings	0.86595	Dibenzo-p-dioxin	100	727
	Nitrogen groups						
70	-CONH2	3	Amide with no substituents	-2.06097	Acetamide	38	801
71	-CONH-	3	Amide with one substituent attached to the nitrogen	-3.36844	N-Methylformamide	21	802
72	-CON<	3	Amide with two substituents attached to the nitrogen	-3.72509	N,N-Dimethylformamide (DMF)	22	803
74	ON=C-	3	Oxime	-4.05705[a]	Methyl ethyl ketoxime	41	806
75	NO2-C(na)	3	Nitro group attached to a nonaromatic carbon	-0.44621	1-Nitropropane	30	807
76	NO2-C(a)	3	Nitro group attached to an aromatic carbon	-0.50698	Nitrobenzene	31	808
78	-ON=C	2	Isoxazole O-N=C	-2.01637	3,5-Dimethylisoxazole	57	810
80	NH2-C(na)	1	Primary amine attached to nonaromatic carbon	-4.86888	Hexylamine	103	812
81	NH2-C(a)	1	Primary amine attached to an aromatic carbon	-2.97323	Benzidine	102	813
82	-N(na)H-	1	Secondary amines	-0.26690	N,N-Diethylamine	106	814
86	(C,Si)=N-	1	Secondary amines attached to one carbons via a double bond	-0.27182	N-Benzylidenemethyl amine	108	815
84	>N(na)-	1	Tertiary amine	-2.43286	Triethylamine	107	818
83	-N=N-	2	Azene N=N	0.89682[a]	Azobenzene	56	821
142	N-N<	1	Hydrazine with 2 carbon neighbours	0.24837[b]	1,1-Dimethylhydrazine	53	824

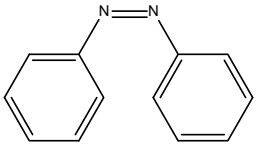
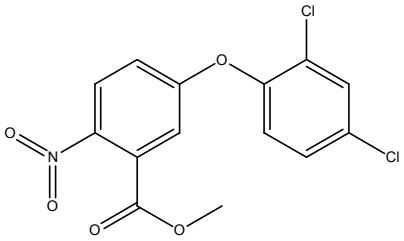
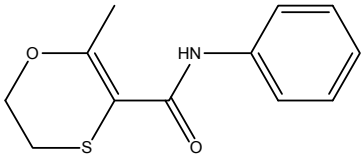
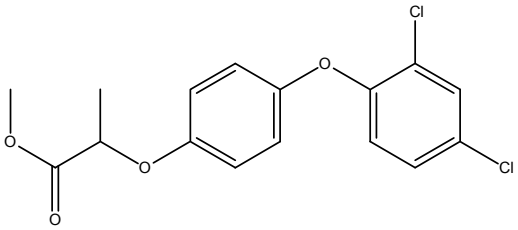
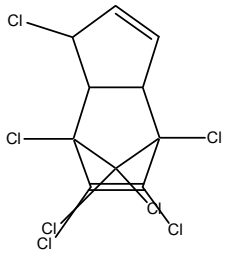
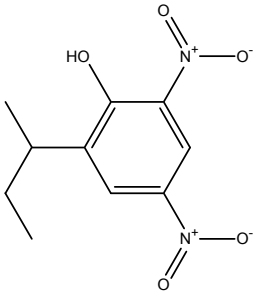
87	=N(a)-(R5)	1	Aromatic nitrogen in a five-membered ring	-0.25141	Pyrrole	105	825
88	=N(a)-(R6)	1	Aromatic nitrogen in a six-membered ring	-0.80148	Pyridine	104	826
89	(C)-CtN	2	CN Group attached to a carbon	-2.65918	Acetonitrile	66	827
165	barbiturate	9	Barbiturate	-4.81217	Butabarbital	1	830
164	-(O=C)N-(C=O)-	5	Pyrrolidinedione group	-0.80275	Theophylline	3	831
162	guanine	6	Guanine	4.73983	Acyclovir	2	832
	Sulfur groups						
99	C(a,na)-SH	1	Thiol/mercaptane attached to carbon	1.25845	1-Propanethiol	83	903
100	-S(na)-	1	Thioether	1.26049	Ethyl methyl sulfide	84	904
101	-S(a)-	1	Aromatic thioether	0.30232	Thiophene	85	905
102	-SO2-	3	Sulfolane O=S=O	-1.46766[b]	Sulfolane	47	906
104	-SO2N<	4	Sulfon amides, attached to N and to S with 2 double bond O	-1.81496	Sulfanilamide	46	908
105	>S=O	2	Sulfoxide	-3.21430[a]	Dimethyl sulfoxide	49	909
170	-(amide)SO2-	6	Sulfolane O=S=O attached to an amide	-2.94510[a]	Sulfacetamide	5	911
171	-(urea)SO2-	7	Sulfolane O=S=O attached to a urea	-6.11700[b]	Gliquidone	4	912
	Phosphorous groups						
97	O=P	2	Oxygen double bonded to a phosphorus	-2.76316	Tributyl phosphate	14	1003
166	S=P	2	Sulfur double bonded to a phosphorus	0.78191	Bromophos	15	1004
167	-S-P	1	Sulfur bonded to a phosphorus and another carbon atom	3.33260	Phosphamide	17	1005
168	-O-P	1	Oxygen bonded to a phosphorus and another carbon atom	0.81616	Triethoxyphosphine	18	1006
	Metal groups						
110	(C)2>Sn<(C)2	1	Stannane with four carbon neighbors	6.06962[b]	Tetramethylstannane	71	1103
Special groups							
156	Ortho	0	(NH2, OH, COOH or NO2) in the ortho possition on the ring	1.67997	o-Nitrophenol	0	1403
157	Meta	0	(NH2, OH, COOH or NO2) in the meta possition on the ring	0.73594	3-Nitroaniline	0	1404
158	Para	0	(NH2, OH, COOH or NO2) in the para possition on the ring	0.33435	2,5-Dinitrophenol	0	1405
	Group interactions						
199	OH-OH	0	OH-OH	68.46555	1,2-Ethanediol	0	3001
195	COOH-OH	0	COOH-OH	34.71902[a]	Malic acid	0	3005
197	OH-C=O	0	OH-C=O	85.51572	3-Hydroxy-2-butanone	0	3009
196	COOH-COOH	0	COOH-COOH	18.81830	Succinic acid	0	3075
194	COOH-C=O	0	COOH-C=O	0.00000	Levulinic acid	0	3079
198	C=O-C=O	0	C=O-C=O	48.34426	2,3-Butanedione	0	3133

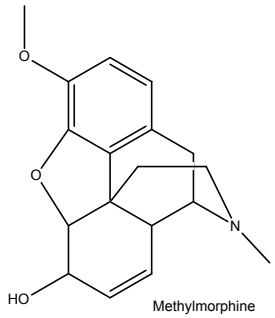
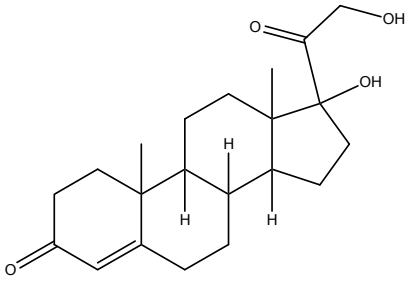
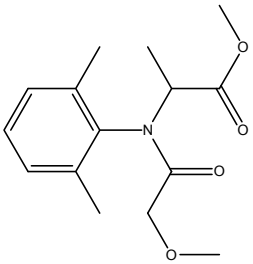
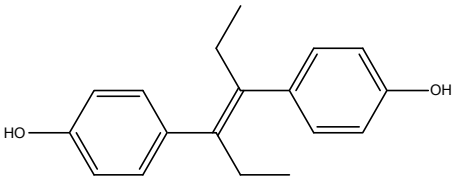
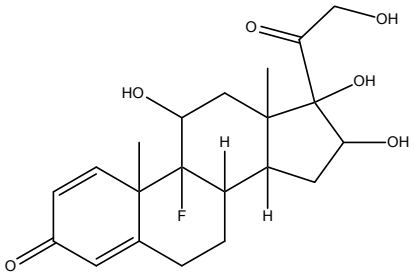
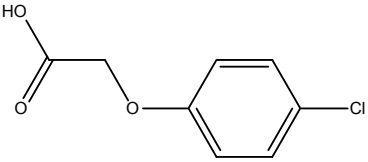
7.8.10. Test set

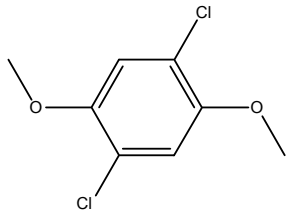
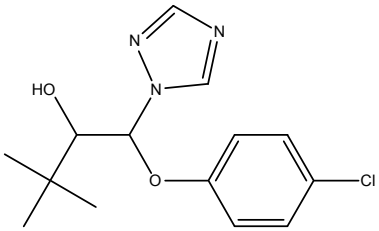
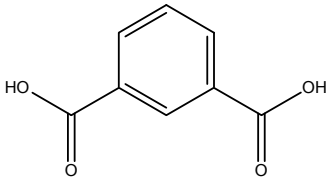
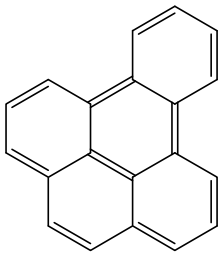
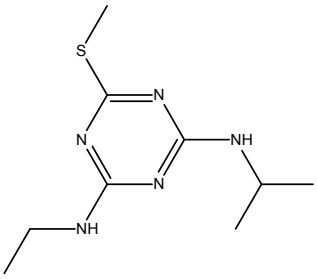
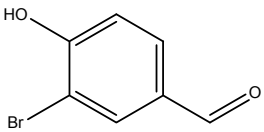
Much discussion has been given above about reducing the number of parameters in the group so that the parameters are not simply highly tuned to the training set but have validity also outside the training set. The data in the test set was chosen on the availability on melting point and heat of fusion data in the database. Compounds with many functional groups were chosen, with the objective being to cover as many of the functional groups as possible. The relative mean percentage deviation for the 25 compounds which were contained in the test set is 9.1 %; this is slightly higher than the training set but still very good considering the especially complicated compounds which are contained in the test set.

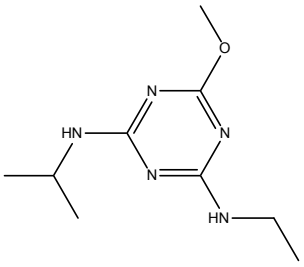
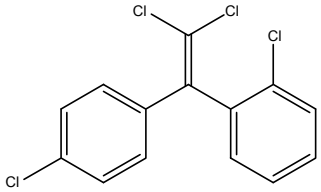
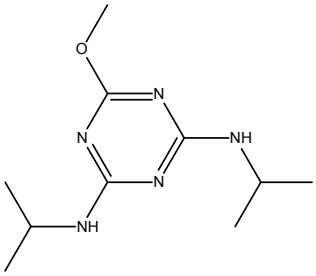
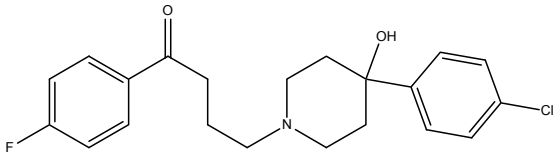
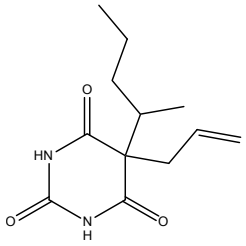
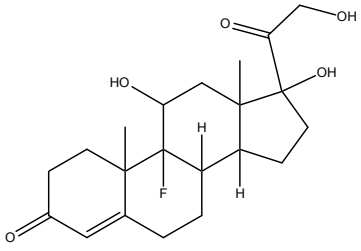
Table 7-24 Compounds with structures used in the test data set, comparison of the experimental activity coefficient with the predictions of this work, UNIFAC and mod. UNIFAC all at 298.15 K

Molecular structure	$\ln\gamma_2^\infty$ exp. 298.15 K	$\ln\gamma_2^\infty$ pred. 298.15 K	UNIFAC	MOD. UNIFAC
 Niflumic acid	10.20	11.02	-	-

Molecular structure	$\ln\gamma_2^{\infty}$ exp. 298.15 K	$\ln\gamma_2^{\infty}$ pred. 298.15 K	UNIFAC	MOD. UNIFAC
 Azobenzene	13.24	13.58	9.39	-
 Bifenox	16.04	16.85	-	-
 Carboxin	9.17	8.28	-	-
 Diclofop-methyl	16.44	15.44	-	-
 Heptox	19.60	21.74	-	-
 Dinoseb	11.99	10.62	8.46	-

Molecular structure	$\ln\gamma_2^{\infty}$ exp. 298.15 K	$\ln\gamma_2^{\infty}$ pred. 298.15 K	UNIFAC	MOD. UNIFAC
 Methylmorphine	5.73	6.39	11.60	-
 Cortexolone	10.10	9.95	-	-
 Metalaxyl	6.17	6.33	-	-
 Diethylstilbestrol	10.61	12.34	8.05	12.36
 Triamcinolone	6.86	6.11	-	-
 4-Chlorophenoxyacetic acid	5.58	6.04	4.70	7.37

Molecular structure	$\ln\gamma_2^{\infty}$ exp. 298.15 K	$\ln\gamma_2^{\infty}$ pred. 298.15 K	UNIFAC	MOD. UNIFAC
 1,4-Dichloro-2,5-dimethoxybenzene	11.68	11.02	5.64	7.06
 Triadimenol	11.27	11.34	-	-
 Isophthalic acid	5.03	4.44	4.73	5.99
 Benzo[E]pyrene	19.69	19.74	21.75	22.74
 Ametryne	9.67	8.34	-	-
 3-Bromo-4-hydroxybenzaldehyde	7.22	7.23	-	-

Molecular structure	$\ln\gamma_2^{\infty}$ exp. 298.15 K	$\ln\gamma_2^{\infty}$ pred. 298.15 K	UNIFAC	MOD. UNIFAC
 Atraton	7.21	6.81	-	-
 1-(2-Chlorophenyl)-1-(4-chlorophenyl)- 2,2-dichloroethylene	17.34	19.31	-	-
 Prometon	8.34	7.67	-	-
 Haloperidol	11.96	14.95	-	-
 Secobarbital	8.89	8.91	-	-
 Fludrocortisone	6.76	9.51	-	-

8. INFINITE DILUTION ACTIVITY COEFFICIENT IN ALKANES

Much consideration is given in the literature to alkanes, and specifically hexane, as a solvent for the extraction of seed oil's and similar products [154-158]. Nevertheless no methods were found in literature specifically for the prediction of the solubility in hexane (or other alkanes). To the knowledge of this author the only methods available for the prediction of the solubility employ the activity coefficient prediction models discussed in section 3.2. The most likely reason for this is that there are insufficient data in the literature to develop or test any predictive method. It is for this exact reason that the methods discussed in section 6.1 were developed. The following sections outline a new group contribution method for the prediction of γ^{∞} in hexane.

8.1. Training set data

The methods presented in section 6.1 facilitated the expansion of the training set to a level where model development was possible. The most suitable method to extrapolate the data was chosen in the following way:

1. If V_m, q and r were available for the solute then Eqn's (6-2) and (6-6) were used.
2. If V_m was unavailable but r and q were available then Eqn's (6-2) and (6-6) were used again but Eqn. (6-19) was used to calculate V_m.
3. If no information about the solute was known then Eqn. (6-15) was used.

The breakdown of the number of data from each source is shown in Table 8-1. It is clear that infinite dilution activity coefficient data make up the vast majority of the data set with data from SLE only comprising about a quarter of the set.

Table 8-1 Sources of data used for the development of the method.

Source	Number
Gamma infinite - DDB	134
SLE - DDB	21
SLE - Beilstein	26
Total	181

In order to avoid skewing the fits to the low molecular weight compounds (which are in abundance in the γ^{∞} database) only one data point was used per solute. Unless otherwise stated all fusion data were taken from DDB and Beilstein. SLE data were considered to be at infinite dilution if $x < 0.01$ [43]. All alkane data were removed from the training set because this data couldn't be fitted very well and there is no need to reproduce this data as Eqn. (6-6) can do it very well.

8.2. Data validation

As with the aqueous infinite dilution activity coefficient there was large scatter present in the literature data. However in this case it was almost impossible to find alternate data in the case of questionable data being found, due to the scarcity of data in the literature. Fortunately most of the fusion data validation had already been done for the water method and therefore the focus was almost solely on the solubility (in the case of SLE).

8.3. Group contribution scheme

Due to the small amount of data in the training set the number of groups was drastically reduced in order to make sure that there is more data behind each group. With the exception of group 163 (OH group attached to a ring carbon) no functional groups were dropped. The *ortho*-, *meta*- and *para*- corrections that were used with the aqueous infinite dilution activity coefficient method were also dropped due to insufficient data to back them up. The same group contribution scheme as the aqueous method was used (see section 7.5) and therefore it will not be covered again here.

8.4. Temperature dependence

As shown above (section 7.4) if $\overline{H}_i^{E,\infty}$ is assumed to be constant with respect to temperature, the following results from Eqn. (7-11):

$$\ln \gamma_i^{\infty} = a + \frac{b}{T} \tag{8-1}$$

where a is a constant of integration b is the constant assumed value of $-\overline{H}_i^{E,\infty} / R$. This assumption is typically a bad one for highly polar solvents such as water; however for the alkanes this equation can provide quite good results. This is evident when examining the γ^{∞} data for some polar solutes in squalane as shown in Fig. 8-1. A similar plot for non-polar (or slightly polar) solutes is shown in Fig. 8-2. It is interesting that in the same way as in the case of the solvent water the polar and non-polar solutes give different trends. The non-polar solutes were found to have more or less constant γ^{∞} data with respect to temperature (an athermal mixture). The polar solutes were correlated with the following relation (where τ is given by Eqn. (7-14)):

$$\ln \gamma_2^{\infty} = \ln \gamma_{2,298K}^{\infty} + 7\frac{(1-\tau)}{\tau} \tag{8-2}$$

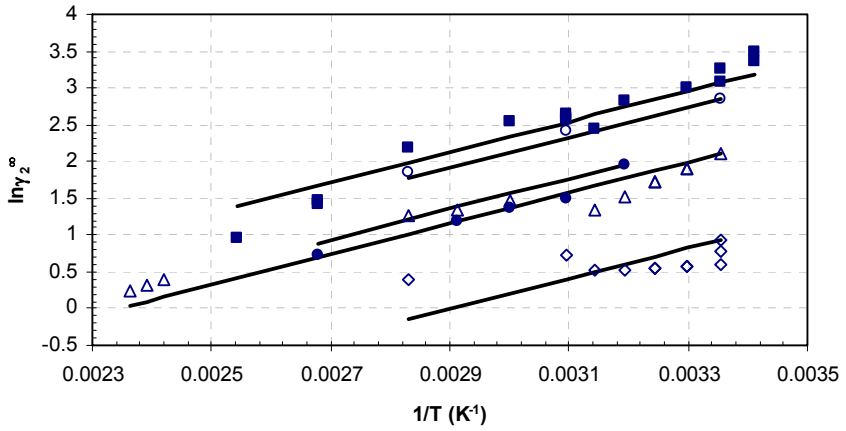


Fig. 8-1 $\ln\gamma_2^\infty$ vs. 1/T for various polar solutes (2) in squalane (1) (Data from the DDB [28], ■ - ethanol, ○ - acetoniltrile, ● - 2-pentanol, △ - 2-methylphenol, ◆ - 2-butanone, —— Eqn. (8-2) fitted to the data)

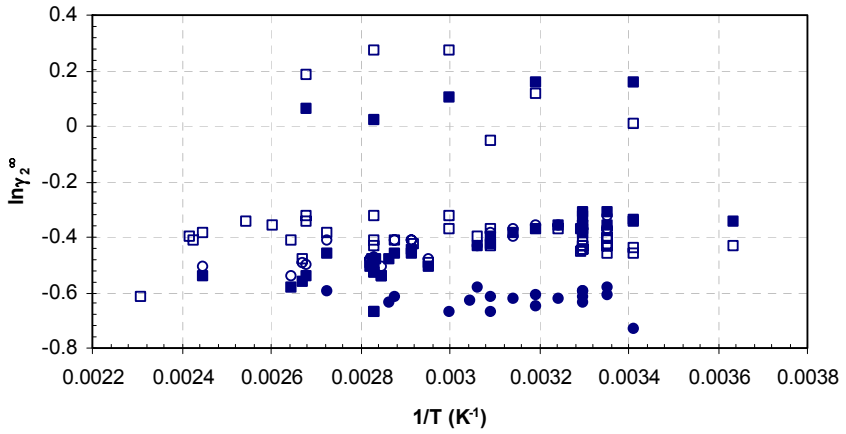


Fig. 8-2 $\ln\gamma_2^\infty$ vs. 1/T for various non-polar solutes (2) in squalane (1) (Data from the DDB [28], □ - hexane, ■ - benzene, ○ - toluene, ● - tetrachloromethane)

The derivation of Eqn. (8-2) is shown in Appendix E. A further discussion of the trends of the data observed for polar and non-polar compounds in alkanes is given in Appendix G. Unlike the data for $\ln\gamma^\infty$ in water (section 7.4) both the trends for the polar and non-polar solvents are plausible. However since there is quite a large amount of scatter in the non-polar data and the trend is uncertain, Eqn. (8-2) will be used as a generalization. Since it has been shown in section 6.1 that the activity coefficient can be accurately extrapolated, at some arbitrary (but constant) temperature, from one alkane solvent to another it holds that the all the alkane solvents should produce the same slope. This is very useful since while there are many data for γ^∞ in squalane the solid solubility data in squalane are very scarce and solubility data at multiple temperatures even more so. Therefore the model will be tested on various alkane data. Fig. 8-3 to Fig. 8-10 show the predictions of the temperature dependence when using Eqn. (8-2) for alkanes as solvents.

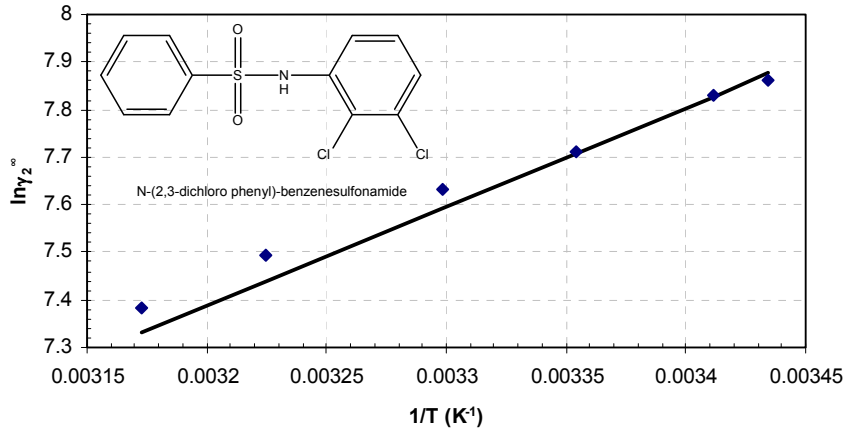


Fig. 8-3 $\ln\gamma_2^\infty$ vs. 1/T for N-(2,3-dichlorophenyl)benzenesulfonamide (2) in hexane (1) (◆ - data extracted from SLE data from literature [159], —— Eqn. (8-2) with the 298 K reference experimental data point).

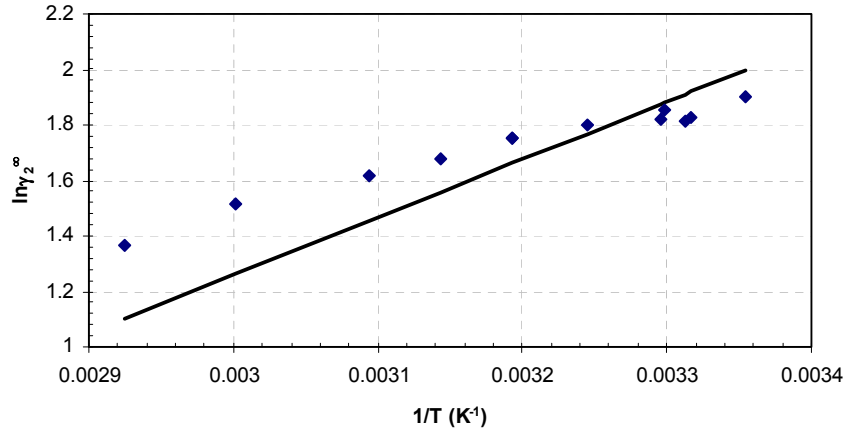


Fig. 8-4 $\ln\gamma_2^\infty$ vs. 1/T for acetone (2) in hexane (1) (◆ - data from the DDB [28], —— Eqn. (8-2) with the 298 K reference experimental data point).

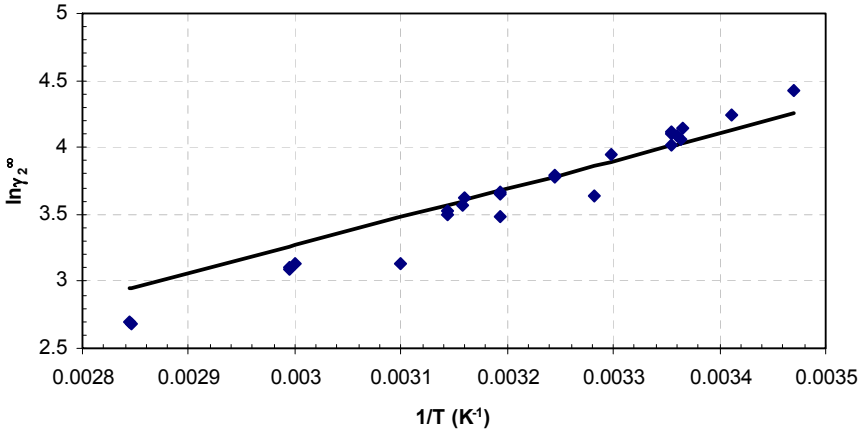


Fig. 8-5 $\ln\gamma_2^\infty$ vs. 1/T for ethanol (2) in hexane (1) (◆ - data from the DDB [28], —— Eqn. (8-2) with the 298 K reference experimental data point).

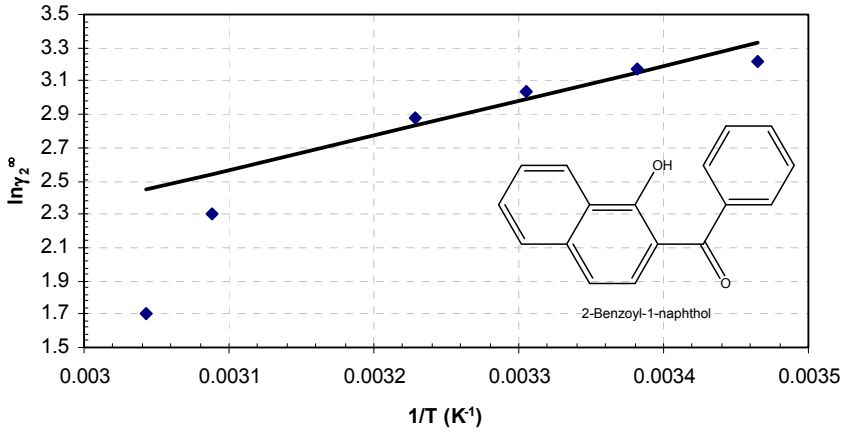


Fig. 8-6 $\ln\gamma_2^\infty$ vs. 1/T for 2-benzoyl-1-naphthol (2) in hexane (1) (◆ - data extracted from SLE data from the DDB [28] – for all data x > 0.01, —— Eqn. (8-2) with the 298 K reference experimental data point).

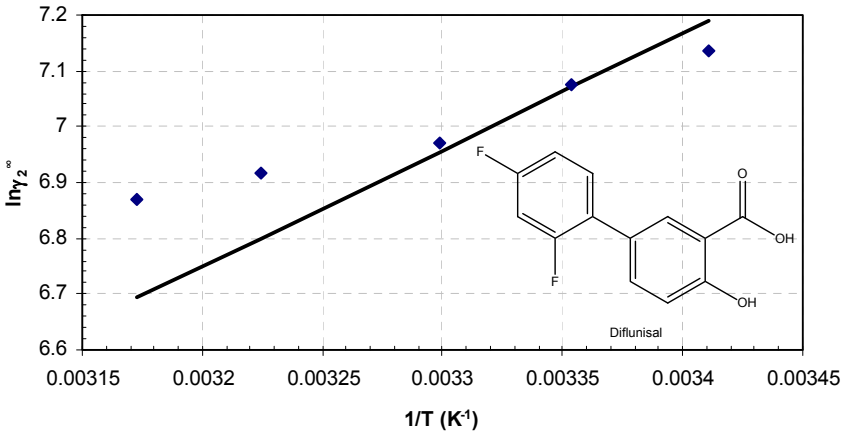


Fig. 8-7 $\ln\gamma_2^\infty$ vs. 1/T for diflunisal (2) in hexane (1) (◆ - data extracted from SLE data from literature [160], – for all data x > 0.01 , —— Eqn. (8-2) with the 298 K reference experimental data point).

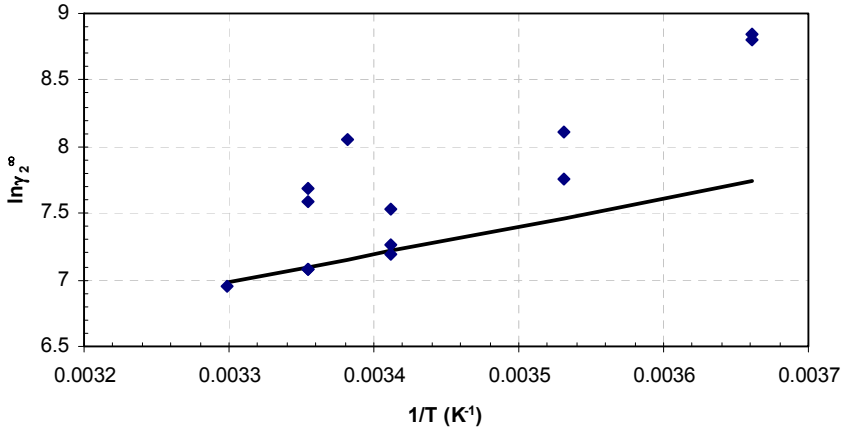


Fig. 8-8 $\ln\gamma_2^\infty$ vs. 1/T for water (2) in heptane (1) (◆ - data from the DDB [28], —— Eqn. (8-2) with the 298 K reference experimental data point).

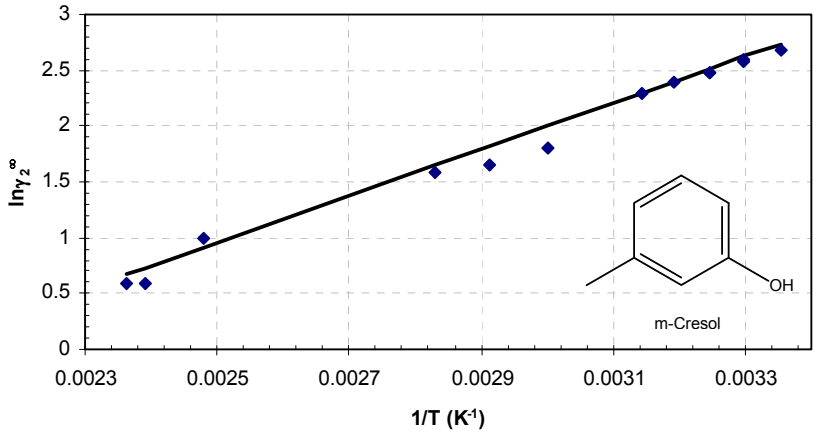


Fig. 8-9 $\ln\gamma_2^\infty$ vs. 1/T for m-cresol (2) in squalane (1) (◆ - data from the DDB [28], ── Eqn. (8-2) with the 298 K reference experimental data point).

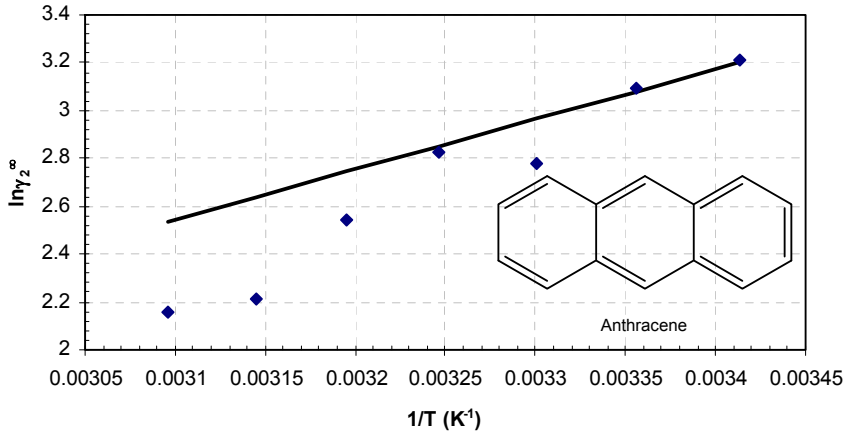


Fig. 8-10 $\ln\gamma_2^\infty$ vs. 1/T for anthracene (2) in heptane (1) (◆ - data extracted from SLE data from literature [142], ── Eqn. (8-2) with the 298 K reference experimental data point).

As with the infinite dilution activity coefficient in water, it seems that the general temperature dependence may hold for alkane solvents. However, due to the scarcity of sufficiently accurate data this correlation is more uncertain than the one used with the water method. Again this method is not proposed as something which is always rigorously correct (it is clear from the above examples that this is not the case) but rather as a good first approximation. The suggested temperature range for the general temperature dependence is 280 K to 330 K.

8.5. Results

As with the aqueous infinite dilution activity coefficient method the results in the training set are compared to both COSMO (SAC and RS, again where RS refers to the use of COSMO-RS(OL)) and both UNIFAC (mod. and original) methods. Due to the scarcity of the available data, very few data were set aside for a test set. The same definition of the families as applied previously is applied here (section 7.8). Since there is limited data for the sulphur compounds

and no data for the phosphorous compounds, sections dedicated to them will not be included in this discussion. The two methods which are used for the error plots are UNIFAC and mod. UNIFAC since they are the ones with the lowest overall errors.

8.5.1. Hydrocarbons

As mentioned above due to the small training set the groups were defined as broadly as possible. All the aliphatic hydrocarbon groups were compressed as much as possible since otherwise they had some unreasonable values (for example group 15 >C< had a value of -8 which is totally unrealistic). As an example groups 4, 5 and 6 (CH2, CH and C in a chain) were compressed into a single group as CH2/CH/C in a chain (new group 4). The groups compressed and the new group numbers are shown in Table 8-2. Similar simplifications were not applied to the aromatic groups since there was sufficient data to back each group up.

Table 8-2 New and old group number for the compressed hydrocarbon groups.

Old no.'s	New no.	New group name
4,5,6	4	CH2/CH/C in a chain
7,8,9	7	CH2/CH/C in a chain attached to an electronegative atom
10,11,12	10	CH2/CH/C in a ring
13,21	13	Double bonded ring carbon with 1 or 2 carbon neighbour(s)
20,27	20	Double bonded chain carbon with 1 or 2 carbon neighbour(s)
14,15,24	24	CH2/CH/C in a ring attached to an electronegative atom

The result breakdown for the hydrocarbons is shown in Table 8-3. As with the aqueous solution data, the hydrocarbons give an error which is much higher than the overall average error of the method. Many of the hydrocarbons are more or less ideally soluble in hexane ($0 < \ln\gamma_2^\infty < 1$) and therefore small deviations in γ_2^∞ result in large deviations in $\ln\gamma_2^\infty$ ‡. The error plots show a fair agreement between the experimental and fitted results despite the large error values. A good example of the wide applicability of this method to large hydrocarbon solutes is beta-carotene (structure shown in Fig. 8-11). The fitted value of $\ln\gamma_2^\infty$ is 3.64 which is very close to the experimental value of 3.50. This value is much better than the predicted value obtained from UNIFAC (-1.98) and mod. UNIFAC (2.06), however considering that mod. UNIFAC was not fitted to this data the latter result is reasonable.

Table 8-3 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in hexane @ 298.15 K for hydrocarbons (number of solutes in superscript).

Name	Eqn. (7-9)	UNIFAC	Mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
Hydrocarbons	32.1 [21]	52.0 [21]	35.8 [21]	53.4 [20]	49.5 [20]
Alkenes	62.0 [3]	113.0 [3]	70.4 [3]	88.0 [2]	88.1 [2]
Aromatic hydrocarbons	27.1 [18]	41.9 [18]	30.0 [18]	49.6 [18]	45.3 [18]

‡ This seems like a poor choice in objective function; however the RMD is not the objective function it is only used to represent the error in a way which is simple. It is for this reason that both RMD and error plots are presented. The regression routines that were used, for both the alkane and the water method, are outlined in Appendix C.

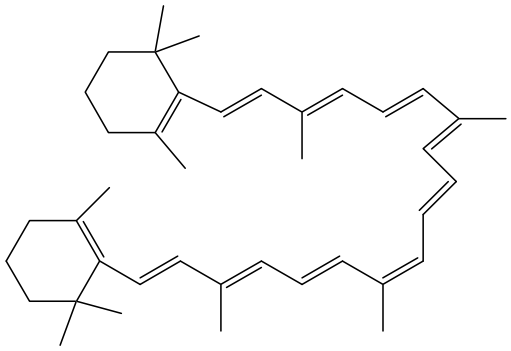


Fig. 8-11 Molecular structure of beta-carotene.

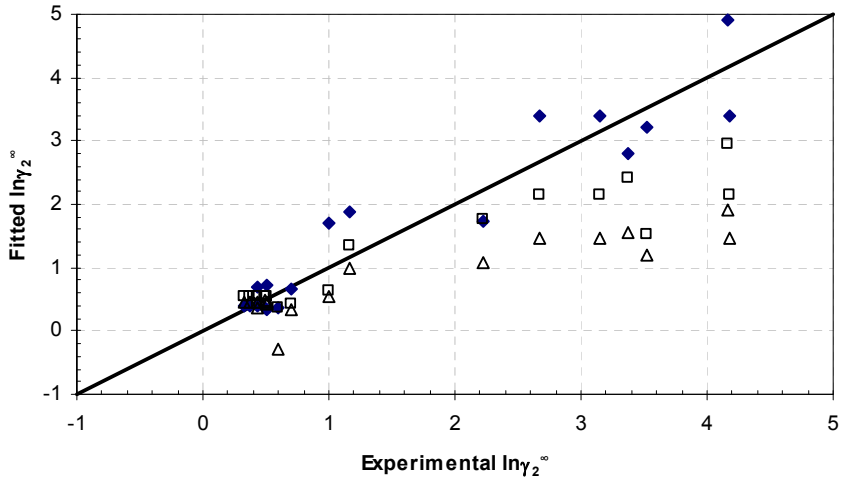


Fig. 8-12 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for aromatic hydrocarbons (◆ - This work, □ – mod. UNIFAC (Do), △ – UNIFAC).

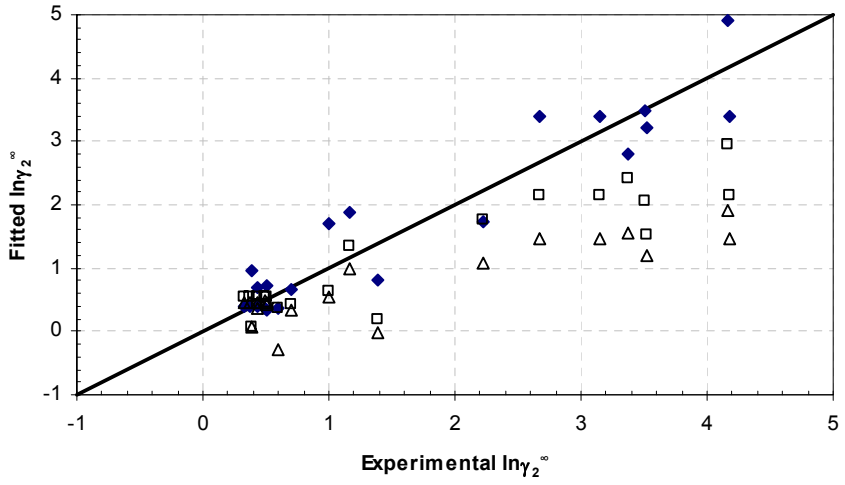


Fig. 8-13 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for hydrocarbons (◆ - This work, □ – mod. UNIFAC (Do), △ – UNIFAC).

8.5.2. Oxygenated compounds

As mentioned above the *ortho*-, *meta*- and *para*- groups were dropped due to there being insufficient data to back each group up. This does not mean that the groups do not apply but rather that if they had been included it would possibly cause any predictions to be far off due to an unreliable parameter value. Another group that was dropped was group 163 (OH attached to a ring carbon), again due to insufficient data. The results for the oxygenated compounds are shown in Table 8-4. It seems from the RMD's in the table that the ethers have large deviations; however the large percentage is similar to that of the hydrocarbons, whereby because the logarithm is small, even small deviations make a large percentage error. This point is illustrated by the error plot shown in Fig. 8-14. The error plot for all oxygenated compounds is shown in Fig. 8-15. Interestingly the ethers are the only family where the RMD for the COSMO methods is superior to the other 3 methods.

Table 8-4 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in hexane @ 298.15 K for oxygenated compounds (number of solutes in superscript).

Name	Eqn. (7-9)	UNIFAC	Mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
All oxygenated compounds	15.2 [85]	29.3 [80]	17.7 [76]	36.5 [81]	27.6 [81]
Alcohols	6.3 [24]	27.5 [24]	9.7 [24]	36.7 [23]	9.6 [23]
n-Alcohols	5.3 [9]	37.4 [9]	9.0 [9]	43.0 [8]	11.3 [8]
Aromatic alcohols	7.1 [5]	25.6 [5]	13.4 [5]	29.1 [5]	7.4 [5]
Ethers	53.0 [12]	50.2 [12]	47.9 [12]	41.3 [12]	43.7 [12]
Aldehydes	9.9 [6]	12.5 [6]	12.5 [6]	41.9 [6]	29.8 [6]
Ketones	10.1 [14]	17.9 [12]	10.7 [12]	46.6 [13]	37.1 [13]
Carboxylic acids	15.4 [6]	44.7 [6]	33.8 [6]	24.6 [6]	46.8 [6]
n-Carboxyllic acids	6.2 [4]	52.5 [4]	39.8 [4]	23.6 [4]	51.5 [4]
Esters	12.2 [15]	20.5 [15]	10.0 [15]	22.0 [14]	30.6 [14]

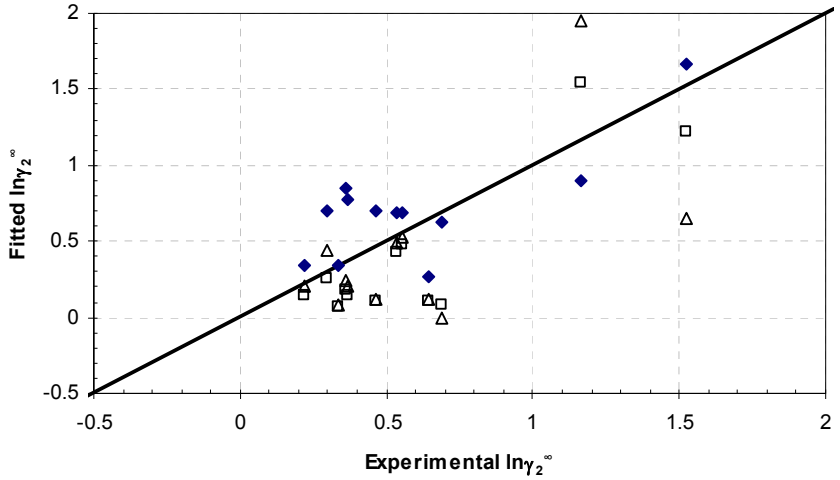


Fig. 8-14 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for the ethers (◆ - This work, □ – mod. UNIFAC (Do), △ – UNIFAC).

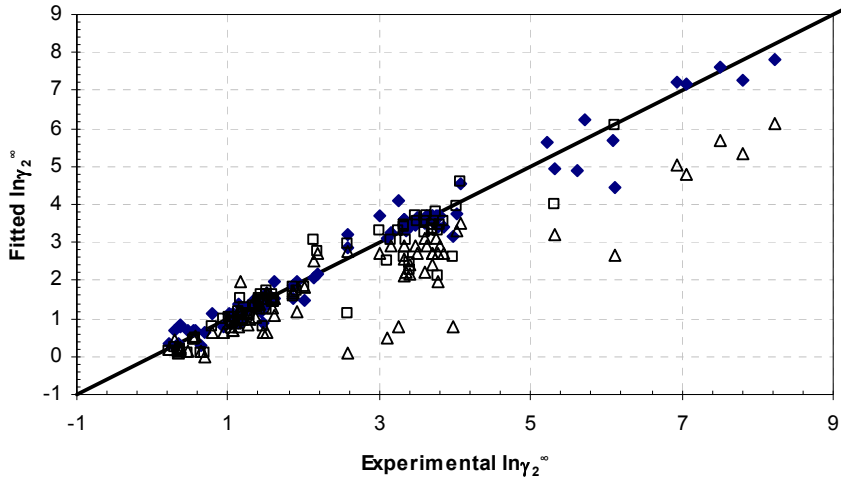


Fig. 8-15 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for the oxygenated compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – UNIFAC).

8.5.3. Nitrogen compounds

The errors for the nitrogen compounds are shown in Table 8-5 and the error plot is shown in Fig. 8-16. As previously the large errors for secondary amines are due to the fact that the logarithm of the infinite dilution activity coefficient are less than 1.

Table 8-5 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in hexane @ 298.15 K for nitrogen compounds (number of solutes in superscript).

Name	Eqn. (7-9)	UNIFAC	Mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
All nitrogen compounds	26.8 [34]	77.3 [23]	96.1 [24]	105.8 [30]	102.2 [30]
All amines	59.6 [10]	155.0 [10]	207.9 [10]	228.0 [10]	229.7 [10]
Primary amines	84.2 [5]	256.6 [5]	372.2 [5]	407.1 [5]	419.5 [5]
Secondary amines	54.6 [2]	40.8 [2]	18.7 [2]	26.2 [2]	11.8 [2]
Tertiary amines	22.0 [3]	61.9 [3]	60.1 [3]	64.1 [3]	58.7 [3]
Cyanides	10.9 [8]	22.4 [6]	12.0 [6]	49.6 [7]	34.9 [7]
Amides	2.9 [4]	18.0 [2]	12.0 [2]	49.4 [4]	30.7 [4]
All nitrates	8.1 [4]	8.5 [4]	21.2 [4]	28.5 [4]	20.8 [4]
Pyrrolidinediones	20.9 [3]	-	-	-	-

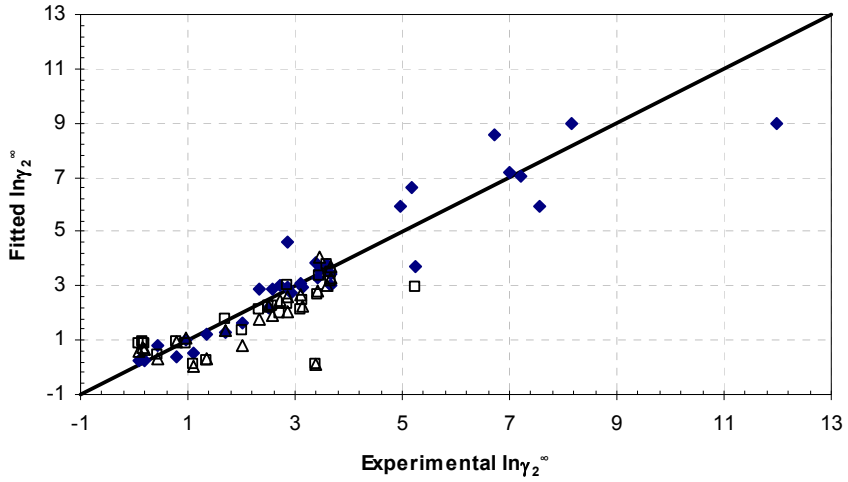


Fig. 8-16 Experimental and fitted $\ln\gamma_2^{\infty}$ @ 298.15 K for the nitrogen compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – UNIFAC).

8.5.4. Halogen compounds

Past experience suggests that along with the hydrocarbons, the halogenated compounds typically follow the group contribution scheme the best. Generally if the groups are defined broadly enough they follow very well and are fairly widely applicable. Table 8-6 and Fig. 8-17 show the percentage errors and the error plots respectively.

Table 8-6 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in hexane @ 298.15 K for halogen compounds (number of solutes in superscript).

Name	Eqn. (7-9)	UNIFAC	Mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
Halogenated compounds	39.6 [21]	40.8 [21]	57.2 [21]	43.6 [19]	40.9 [19]
Fluorinated	0.2 [3]	16.7 [3]	2.2 [3]	77.6 [3]	73.2 [3]
Chlorinated	58.5 [14]	39.8 [14]	70.9 [14]	37.6 [14]	35.9 [14]
Brominated	0.0 [2]	93.2 [2]	52.7 [2]	34.6 [2]	27.7 [2]
Iodinated	6.0 [2]	31.6 [2]	48.2 [2]	-	-

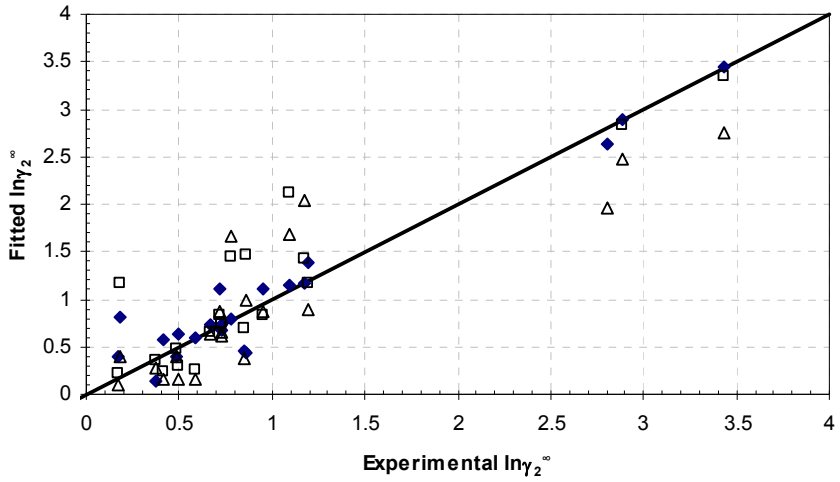


Fig. 8-17 Experimental and fitted $\ln\gamma_2^\infty$ @ 298.15 K for the halogen compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – UNIFAC).

Table 8-6 makes 2 things clear, firstly that the percentage error is fairly poor for all the halogenated compounds and secondly that there are very little data available for the halogen compounds. Due to the small amount of available data it was tempting to reduce the number of halogen groups that were used. This was, however, not done due to the success for the principle of physically realistic groups applied previously (section 7.8.5). The plot for the halogen compounds attached to aromatic carbons is shown in Fig. 8-18. Group 41 and 46 were chosen as the reference compounds since these 2 groups seem to fit the trend of the other group values the best. Also when a fit was made to all groups then this resulted in group values which were much worse for all groups (i.e. the fits in the training set were worse for all compounds) and therefore defeating the whole purpose of physically realistic groups. The halogen groups attached to an aromatic carbon are the only groups which have fitted values for all 4 groups. The group with the largest difference between the physically realistic fit and the direct fit is group 47. Unfortunately, there is no data available to test this new value but there was a test available for the fluorine group as shown in Fig. 8-19. The physically realistic value gives a value which fits the trend of the data much better than the old value.

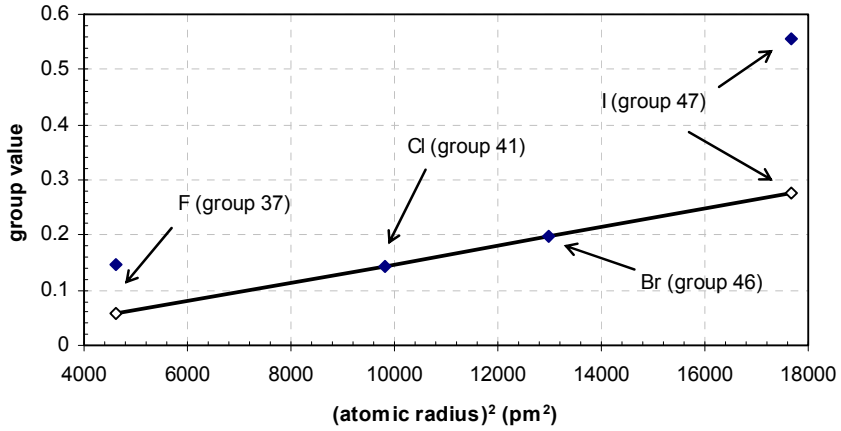


Fig. 8-18 Group value vs. atomic radius squared for groups 47, 46, 41 and 37 (halogen atom attached to an aromatic carbon) (◆ - fitted group value, ◇ - new group value).

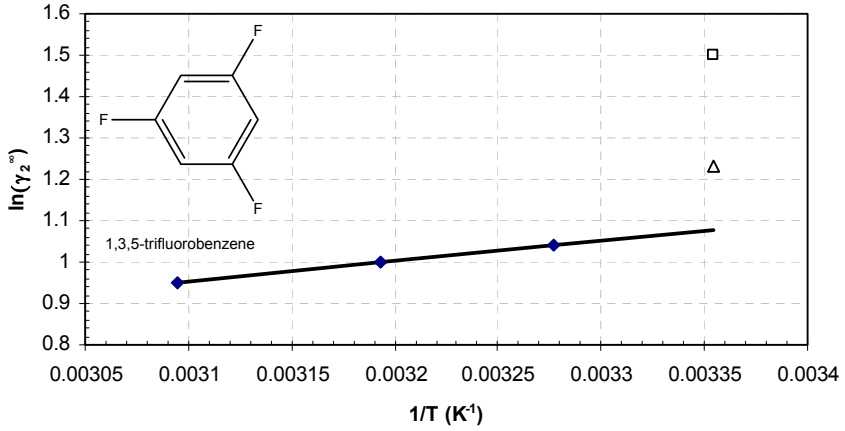


Fig. 8-19 $\ln\gamma_2^{\infty}$ vs. 1/T for 1,3,5-trifluorobenzene(2) in hexane(1) (◆ – data extrapolated from octadecane data from the DDB [28] using Eqn. (6-2), ── straight line fitted to the data, □ - $\ln\gamma_2^{\infty}$ predicted using the old value for group 37, △ - $\ln\gamma_2^{\infty}$ predicted using the physically realistic value for group 37).

The plot for halogen atoms attached to non-aromatic carbons is shown in Fig. 8-20. Initially it was thought that the plot is erroneous since the gradient is negative, which is different to all the previous examples of the physically realistic halogen groups (section 7.8.5) and also different to the group plot for the aromatic groups (Fig. 8-18). However, when examining experimental data it is clear that the aliphatic and aromatic groups should have opposite trends. Two examples of each are shown in Fig. 8-21 and Fig. 8-22. It should be noted that fluorobenzene shows a similar deviation as group 37, however it is not clear whether this is due to experimental difficulties associated with fluorobenzene or if this is indeed a valid trend. Nevertheless, since the physically realistic value seems to be more broadly applicable (Fig. 8-19) and was applied with great success in section 7.8.5, it was decided to use this value.

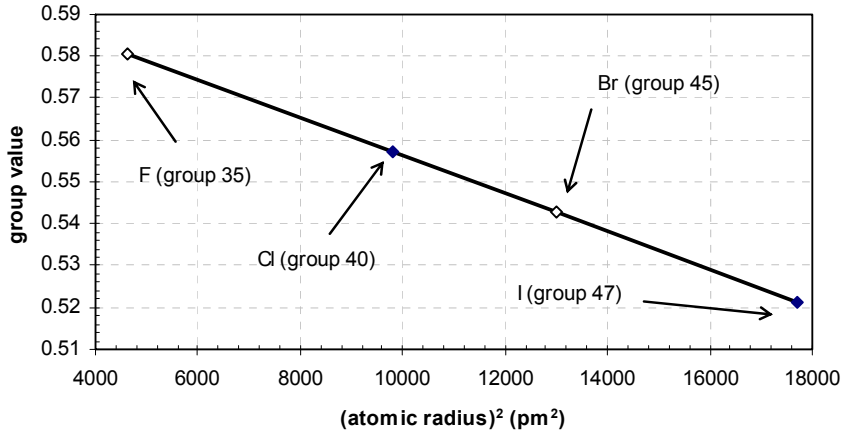


Fig. 8-20 Group value vs. atomic radius squared for groups 35, 40, 45 and 47 (halogen atom attached to a non-aromatic carbon) (◆ - fitted group value, ◇ - new group value).

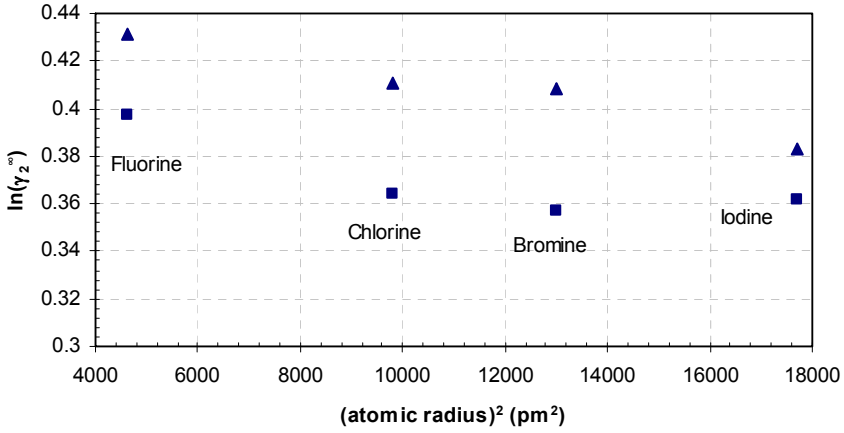


Fig. 8-21 $\ln\gamma_2^{\infty}$ vs. atomic radius squared for 1-halopentane (■) and 1-haloheptane (▲) in 19,24-dioctadecyldotetracontane at 373 K (data from the DDB [28]).

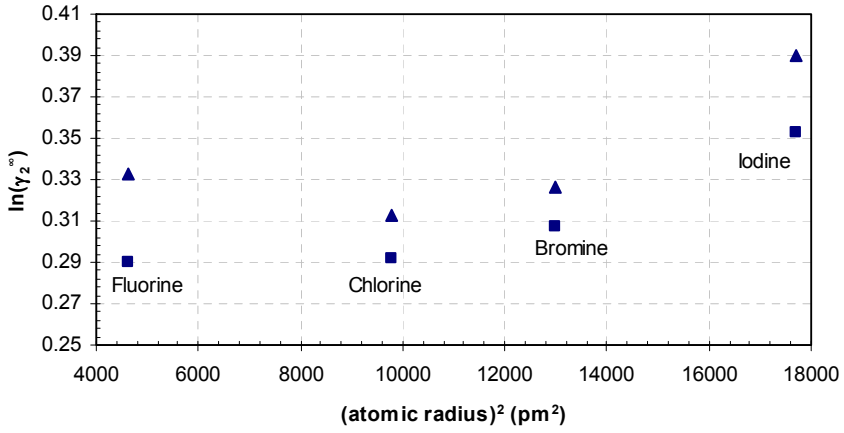


Fig. 8-22 $\ln\gamma_2^{\infty}$ vs. atomic radius squared for halobenzene @ 418 K (■) and halobenzene @ 373 K (▲) in 19,24-dioctadecyldotetracontane (data from the DDB [28]).

Since this principle was applied with such success previously and it seems to hold for the aromatic groups it was decided to apply it to the missing non-aromatic groups. As with the water method, there were only sufficient accurate (assumed accurate if there are sufficient data behind the group) group data to fit curves to 2 classes of groups. The slopes for the lines in Fig. 8-20 and Fig. 8-23 are similar ($-4.546\times10^{-6}\,pm^{-2}$ and $-2.809\times10^{-6}\,pm^{-2}$ respectively), therefore an average was taken and used for the fixed value for the groups in Fig. 8-24. In both cases the group which the line passes through is the group with the most data behind it. Group 43 had data with high scatter behind it and this is probably the reason for the high deviation for the physically realistic fit.

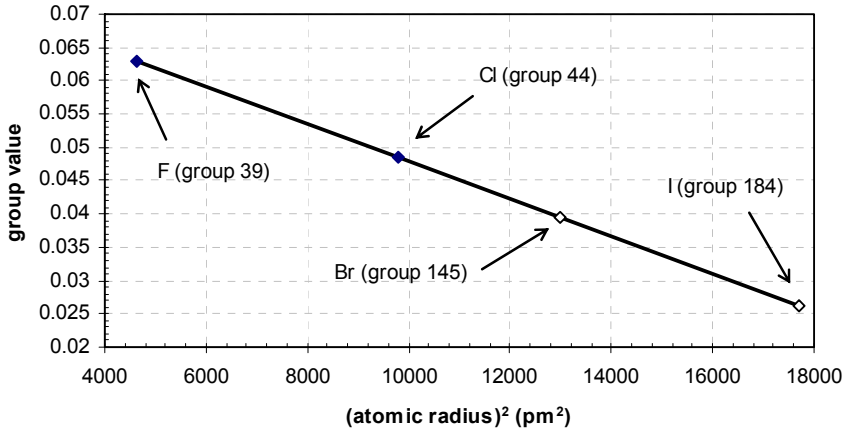


Fig. 8-23 Group value vs. atomic radius squared for groups 44 and 39 (halogen atom attached to an carbon with at least 2 other halogens) (◆ - fitted group value, ◇ - new group value).

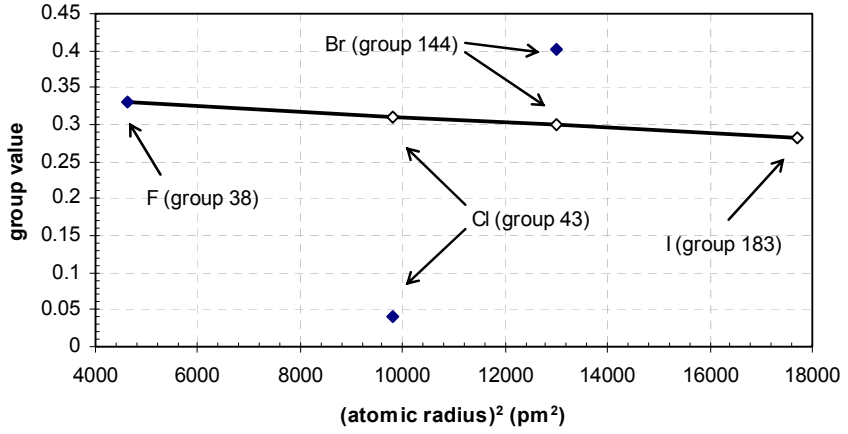


Fig. 8-24 Group value vs. atomic radius squared for groups 144, 43 and 38 (halogen atom attached to an carbon with 1 other halogen) (◆ - fitted group value, ◇ - new group value).

8.5.5. Multifunctional compounds

This section is applicable to all compounds which do not fall into the above mentioned chemical families only. Any discussion involving multi-functional compounds in this chapter does not refer to compounds containing more than one functional group but rather compounds with at least 2 different functional groups (i.e. not both oxygen or nitrogen, etc. groups). This group of compounds had the largest errors from the literature methods and in the case of UNIFAC and mod. UNIFAC less than half of the compounds could be predicted due to missing group parameters. This is where the real usefulness of a method such as this is noticeable. Even though there are no group interaction parameters, the representation of the multifunctional groups is good (quite surprisingly it is more than half of the overall average). The results are shown in Fig. 8-25 and Table 8-7 and are testament to the fact that the literature methods perform quite poorly for these compounds.

Table 8-7 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for multifunctional compounds (number of solutes in superscript).

Name	Eqn. (7-9)	UNIFAC	Mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
Multi-functional compounds	9.3 [17]	28.9 [9]	21.8 [6]	33.6 [12]	26.7 [12]

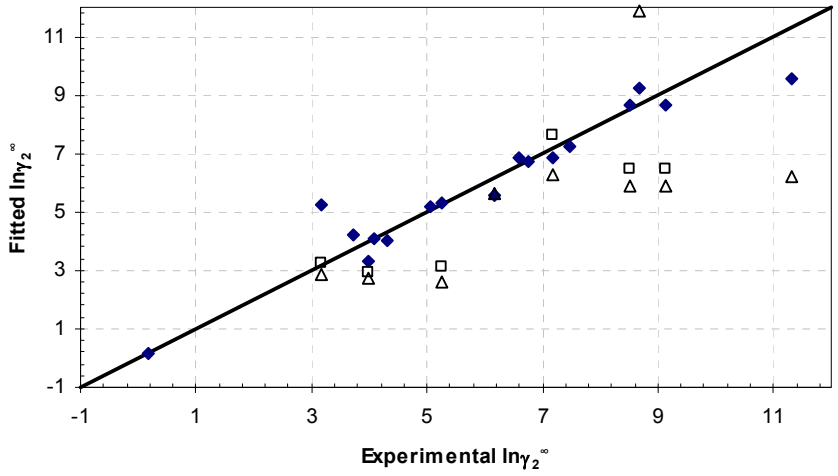


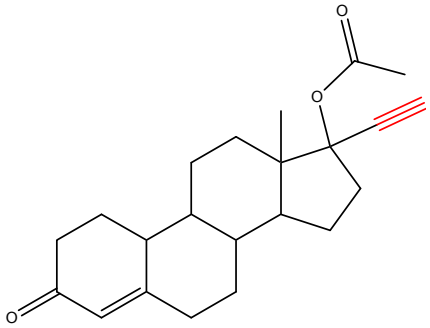
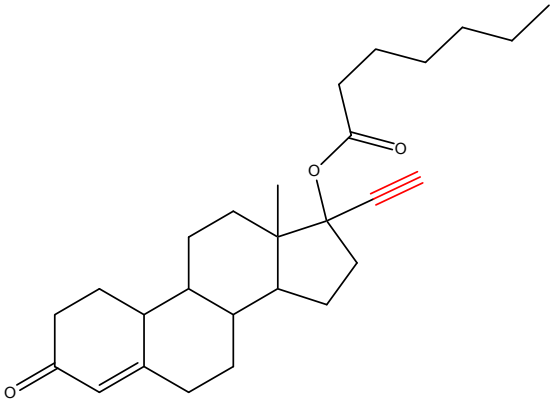
Fig. 8-25 Experimental and fitted lnγ_2^∞ @ 298.15 K for multifunctional compounds (◆ - This work, □ – mod. UNIFAC (Do), △ – UNIFAC).

8.5.6. Missing group values

As frequently mentioned above, the main limiting factor on the development of an activity coefficient model in non-aqueous solvents is the availability of sufficient accurate data for each group. The database used in this work is fairly small and therefore it is only logical that some group values are missing. For many of these values this is a problem because it is very likely that these values are non-trivial (for example the OH group in hexane is certainly non-trivial) and therefore any reasonable estimation is difficult.

For some groups however we may expect that the contribution is very small and therefore (due to the large amount of other groups) may simply ignore the group so that at least a qualitative estimate may be obtained. Consider the two compounds shown in Table 8-8, if the alkyne groups (shown in red) are neglected the predicted results are actually not too far from the experimental values. This data was purposely not included in the training set for two reasons; firstly, the -CCH group is made up of 2 groups (group 22 and group 25) and therefore unique group values cannot be found (since both compounds contain both groups in equal proportions); secondly, since there are many other groups, even if the groups were combined into a single larger group this group value would be very susceptible to error. The under prediction of the activity coefficient is realistic since these groups would probably have small positive values.

Table 8-8 Experimental and predicted activity coefficient and solubility for some substances with missing group values (experimental values from literature [114]).

Molecular structure	Experimental/predicted solubility (g/l)	Experimental/predicted activity coefficient
 Norethindrone acetate	$S_{pred} = 1.87$ $S_{exp} = 1.08$	$\ln \gamma^{\infty}_{pred} = 4.38$ $\ln \gamma^{\infty}_{exp} = 4.96$
 Norethindrone enanthate	$S_{pred} = 17.23$ $S_{exp} = 11.98$	$\ln \gamma^{\infty}_{pred} = 4.20$ $\ln \gamma^{\infty}_{exp} = 4.56$

While an approach such as this can be useful it is obviously only applicable to molecules comprised of only select groups. It is therefore desirable to have some general procedure for estimating group values for groups which will not have small values. One example is the amides; there are 3 different groups for the amides: 0, 1 and 2 substituents on the nitrogen (group 70, 71 and 72 respectively). For this method there are group values for group 71 & 72, there are therefore 2 possible methods for estimating the missing group value.

Firstly, based on the other group (71 & 72) values it may be possible to estimate group 70. The group values for the aqueous method are shown in Fig. 8-26. As hoped the group values lie on a straight line and produce a fairly good fit to the data. Therefore, if it is assumed that this trend is observed for the hexane groups then it would be possible to estimate the group value as shown in Fig. 8-27. The resulting group value is large and seems intuitively wrong.

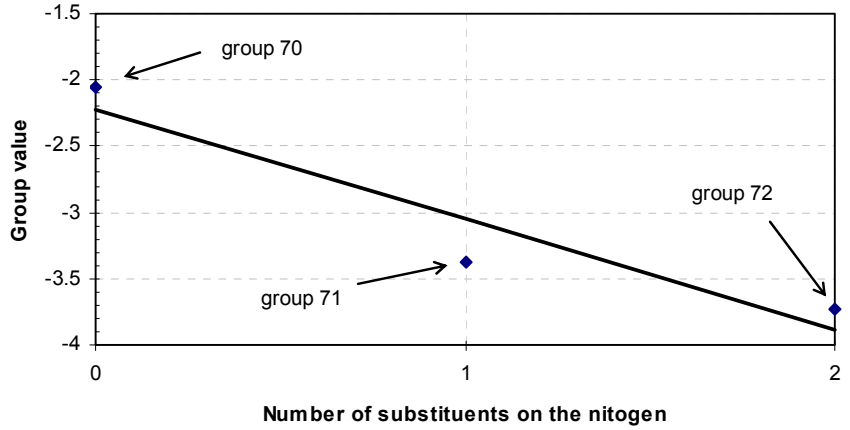


Fig. 8-26 Group values vs. number of substituents on the nitrogen for the amides (groups 70,71 and 72) for the aqueous infinite dilution method.

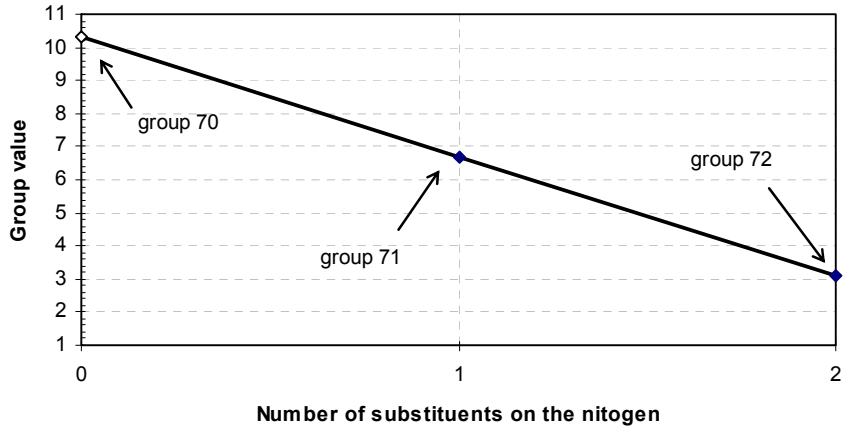


Fig. 8-27 Group values vs. number of substituents on the nitrogen for the amides (groups 71 and 72) for the hexane infinite dilution method (◆ - fitted group value, ◇ - new group value).

Secondly, it was shown above (section 7.8.8) that the non-hydrocarbon groups could be fairly loosely correlated with some measure of the electronegativity and the molecular size. For the water method there was no data available to test the correlation, however in this case there are. The plots for the hexane group values are shown in Fig. 8-28 and Fig. 8-29. Unlike with the previous method, the group values seem a lot more plausible. A comparison of the usefulness of the 2 predicted group values is shown in Table 8-9. It is clear that the value predicted from groups 71 and 72 is totally ridiculous while the other value gives good results. This success suggests that the correlation presented in Fig. 8-28 is useful for predicting missing group values (there is too much scatter present in the aromatic correlation for any

useful results to be derived). The poor results in the predictions from groups 71 and 72 could suggest that one of the values is incorrect. Since group 71 is fairly large this would probably be the value with the most uncertainty. When used to predict some literature data the results are fairly good as shown in Table 8-10.

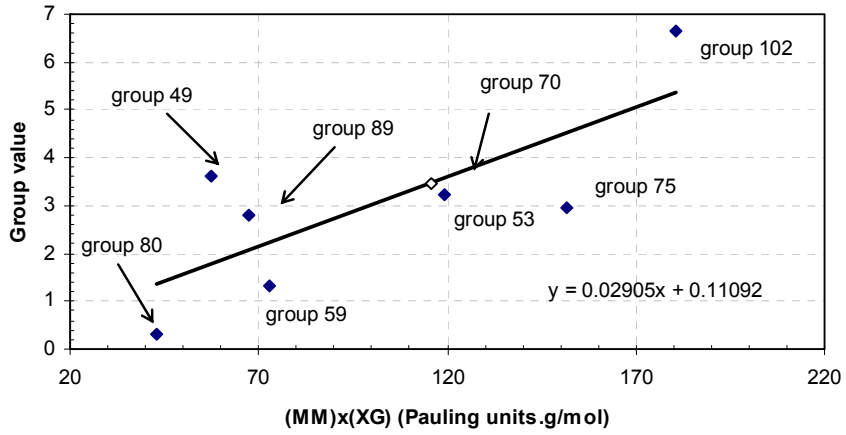


Fig. 8-28 Group value for groups attached to non-aromatic atoms vs. the product of molar mass (MM) and electronegativity (XG) [153] (◆ - fitted group value, ◇ - new group value).

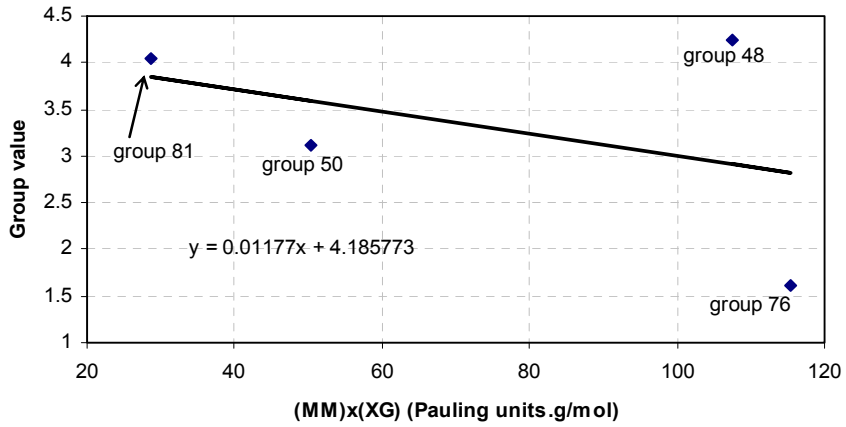


Fig. 8-29 Group value for groups attached to aromatic atoms vs. the product of molar mass (MM) and electronegativity (XG) [153].

Table 8-9 Comparison of the predicted values for group 70 from method 1 (Fig. 8-27) and method 2 (Fig. 8-28) for atenolol (fusion and solubility data from literature [161]).

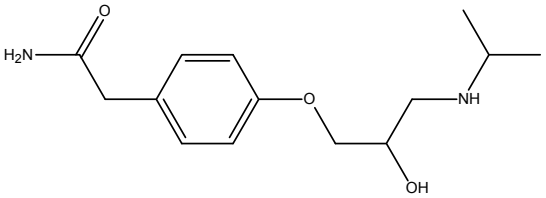
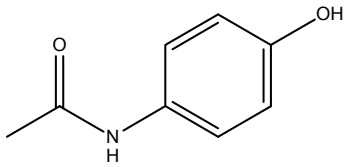
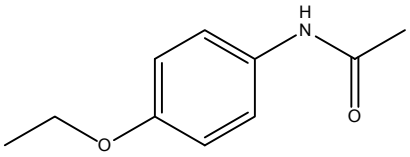
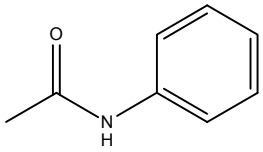
Molecular structure	Experimental/predicted solubility (g/l)	Experimental/predicted activity coefficient
 Atenolol	$S_{pred,1} = 2.8 \times 10^{-7}$ $S_{pred,2} = 0.00259$ $S_{exp} = 0.00076$	$\ln \gamma^{\infty}_{pred,1} = 16.498$ $\ln \gamma^{\infty}_{pred,2} = 9.670$ $\ln \gamma^{\infty}_{exp} = 10.890$

Table 8-10 Comparison of the predicted and experimental values for some compounds containing group 71 (data from literature [162]).

Molecular structure	Experimental/predicted solubility (g/l)	Experimental/predicted activity coefficient
 Paracetamol	$S_{pred} = 0.00188$ $S_{exp} = 0.00473$	$\ln\gamma^{\infty}_{pred} = 10.377$ $\ln\gamma^{\infty}_{exp} = 9.480$
 Phenacetin	$S_{pred} = 0.0257$ $S_{exp} = 0.0270$	$\ln\gamma^{\infty}_{pred} = 7.985$ $\ln\gamma^{\infty}_{exp} = 7.948$
 Acetanilide	$S_{pred} = 0.1575$ $S_{exp} = 0.0133$	$\ln\gamma^{\infty}_{pred} = 7.023$ $\ln\gamma^{\infty}_{exp} = 9.551$

8.5.7. Overall results

In total there were 181 compounds used in the training set, the overall relative mean deviation (%) for the training set was 21.4 % for the logarithm of the activity coefficient in hexane at 298.15 K. Since most of the data used in the training set is made up of data from the infinite dilution databank contained in DDB [28], one would expect UNIFAC and (especially) mod. UNIFAC to perform quite well. This is true in most instances. The overall error comparisons for the various methods tested are shown in Table 8-11. The error histograms for this work are shown in Fig. 8-30 and Fig. 8-31, the RMD histogram is quite spread out across the percentage ranges, but this is because of the small value of the infinite dilution activity coefficient in hexane. The absolute deviation plot is probably a better representation of the error distribution. The plots for mod. UNIFAC are shown in Fig. 8-32 and Fig. 8-33. The plots are quite similar to those of this work but have more data in the higher error ranges. It is also important to remember that mod. UNIFAC errors were not given for many of the SLE data in the training set (since binary interaction parameters were typically unavailable) which would probably offset the mod. UNIFAC fits for the worse. Nevertheless, out of all the literature methods investigated mod. UNIFAC performed the best.

The values for the model parameters are shown in Table 8-12. A total of 59 model parameters were used (1 constant, 55 group contributions and 3 group interactions). As with the aqueous infinite dilution activity coefficient estimation method the table contains 66 group values and this is because many of the physically realistic group values were used.

Table 8-11 Relative mean deviation (%) for the infinite dilution activity coefficient (logarithm) in water @ 298.15 K for all compounds in the training set (number of solutes in superscript).

Name	Eqn. (7-9)	UNIFAC	Mod. UNIFAC	COSMO-RS(OL)	COSMO-SAC
All compounds	21.4 [181]	40.9 [155]	38.5 [150]	51.8 [164]	45.4 [164]

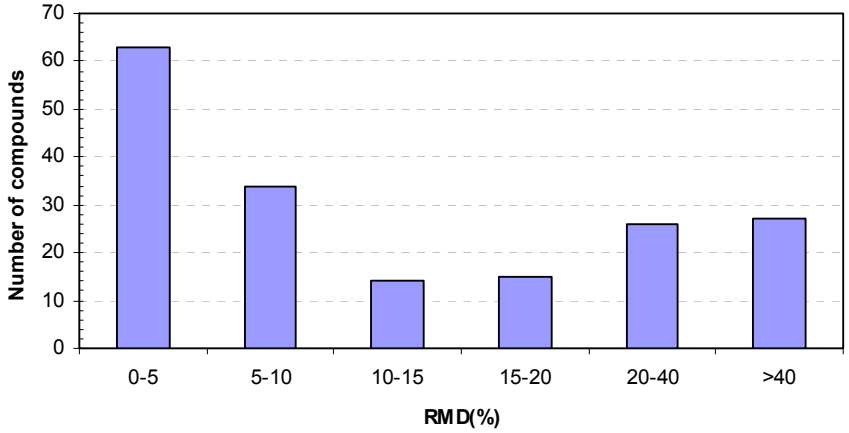


Fig. 8-30 Histogram of this works relative mean deviation (%) in $\ln\gamma^{\infty}$ for the compounds in the training set.

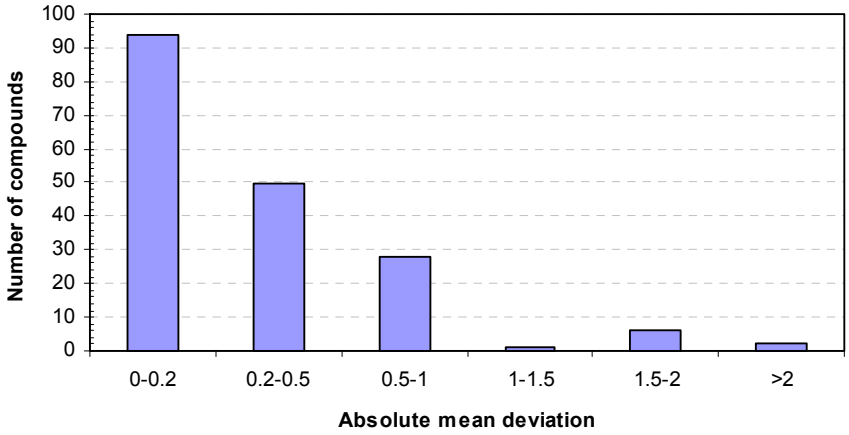


Fig. 8-31 Histogram of this works absolute mean deviation in $\ln\gamma^{\infty}$ for the compounds in the training set.

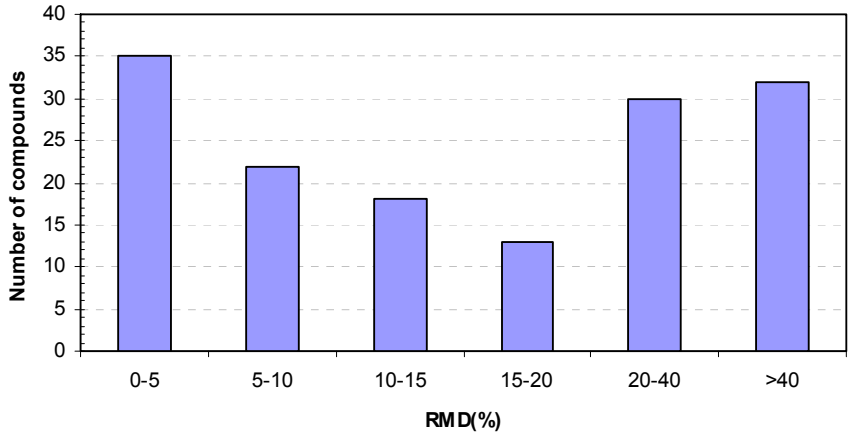


Fig. 8-32 Histogram of the mod. UNIFAC relative mean deviation (%)in $\ln\gamma^{\infty}$ for the compounds in the training set.

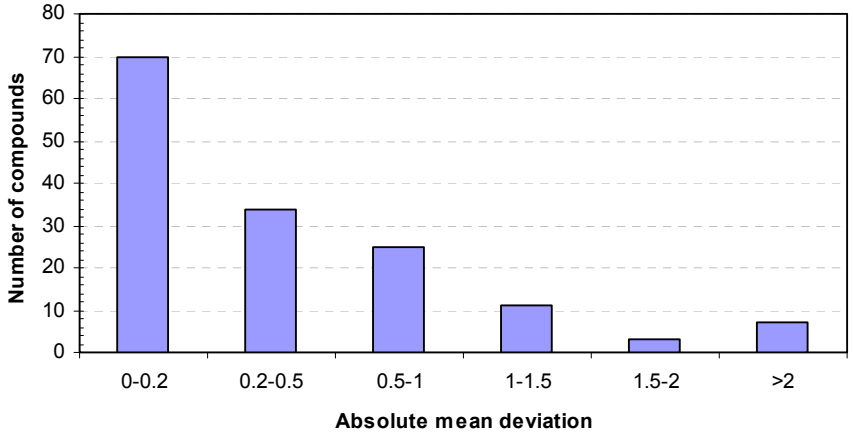


Fig. 8-33 Histogram of the mod. UNIFAC absolute mean deviation in $\ln\gamma^{\infty}$ for the compounds in the training set.

Table 8-12 The group contribution and group interaction values for the infinite dilution activity coefficient in hexane at 298.15 K using Eqn. (7-9) (Ink No – the number which is used by the fragmentation program to identify the group, NA – number of non-hydrogen atoms in the group, Prty – Priority – order in which the groups are fragmented, [a] – group only fitted to one data point, [b] – group only fitted to two data points, [c] – group value taken from the physically realistic value)

$$\ln\gamma_2^{\infty} = \sum_i v_i C_i + \frac{1}{2}\sum_i\sum_j \frac{GI_{i-j}}{n(m)}$$

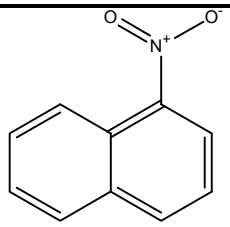
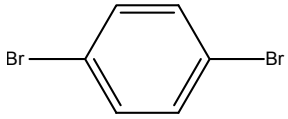
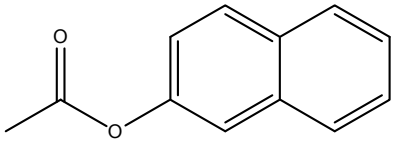
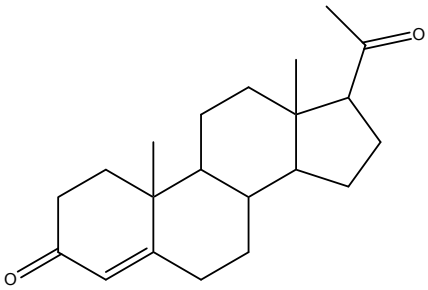
Ink No	Name	NA	Description	C_i	Example	Prty	Ref No
0	Constant	0	Constant	0.47438[c]	All Compounds	0	0
	Aliphatic carbon groups						
1	CH3-(ne)	1	CH3 attached to a non-aromatic non-electronegative atom	-0.25294	Hexane	120	101
4	-C(c)H2-	1	CH2/CH/C in a chain	-0.03642	Decane	122	102
2	CH3-(e)	1	CH3 group attached to a non-aromatic electronegative atom	0.00155	N,N-Dimethylformamide (DMF)	117	105
7	-CH2(c)-(e)	1	CH2/CH/C in a chain attached to an electronegative atom	-0.10233	Ethanol	121	106
10	-C(r)H2-	1	CH2/CH/C in a ring	0.00092	Cyclohexane	124	110
24	-C(r)H2-(en)	1	CH2/CH/C in a ring attached to an electronegative carbon	-0.43428	1,4-Dioxane	123	116
26	C(c)H2=C(na)<	1	Double bonded carbon at the end of a chain/ring	0.76280[b]	beta-Pinene	108	120

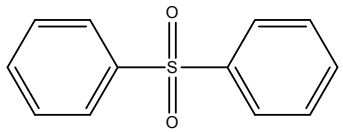
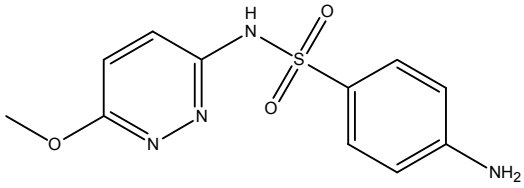
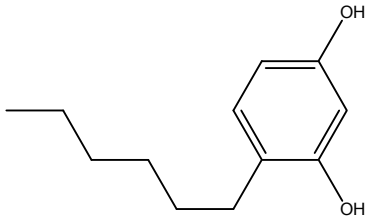
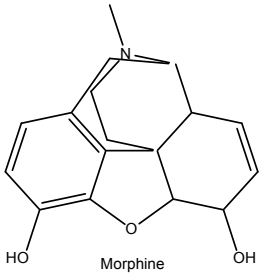
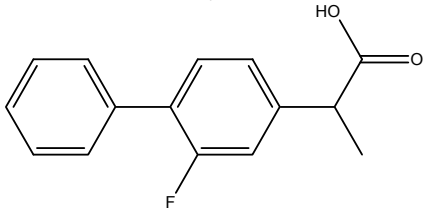
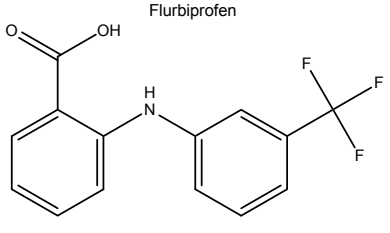
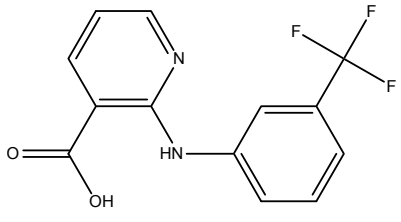
20	-C(c)H=C(c)-	1	Double bonded carbon in a chain with only 1/2 carbon neighbour(s)	0.08228	2-Heptene	110	121
13	-C(r)=C(r)<	1	Double bonded carbon in a ring with 1/2 carbon neighbour(s)	1.01373	1-Methylcyclohexene	109	126
	Aromatic carbon groups						
3	CH3-(a)	1	CH3 group attached to an aromatic atom	-0.65748	Toluene	118	201
16	-CH(a)<	1	CH in an aromatic ring	-0.02079	Benzene	115	202
17	>C(a)<	1	C in an aromatic ring	0.65512	m-Xylene	116	203
19	(a)=C(a)<2(a)	1	Aromatic carbon attached to three aromatic neighbours	0.78194	Naphthalene	107	204
18	>C(a)<(e)	1	C in an aromatic ring attached to an electronegative atom	0.21486	Aniline	114	209
	Fluorine groups						
35	F-C(na)	1	F attached to nonaromatic carbon	0.58074[c]	2-Fluoropropane	93	301
38	F-C(na)-1Halo	1	F attached to a carbon with one other halogen atom	0.33038[b]	Perfluoropentane	74	302
39	F-C(na)-2Halo	1	F attached to a carbon with at least two other halogen atoms	0.06297	Perfluorohexane	73	303
37	F-C(a)	1	F attached to aromatic carbon	0.05847[c]	Fluorobenzene	92	306
	Chlorine groups						
40	Cl-C(na)	1	Cl attached to noaromatic carbon	0.55720	cis-1,2-Dichloroethylene	77	401
43	Cl-C(na)-1Halo	1	Cl attached to a carbon with one other halogen atom	0.31134[c]	Dichloromethane	72	402
44	Cl-C(na)-2Halo	1	Cl attached to a carbon with at least two other halogen atoms	0.00843	Tetrachloromethane	71	403
41	Cl-C(a)	1	Cl attached to aromatic carbon	0.14445	Chlorobenzene	78	406
	Bromine groups						
45	Br-C(na)	1	Br attached to a nonaromatic carbon	0.54268[c]	Ethyl bromide	79	501
144	Br-C(na)-1Halo	1	Br attached to a Carbon with one other halogen atom	0.29959[c]	2-Bromo-2-chloro-1,1,1-trifluoroethane	76	502
145	Br-C(na)-2Halo	1	Br attached to a Carbon with at least two other halogen atom	0.03945[c]	Bromochlorodifluoromethane [R12B1]	75	503
46	Br-C(a)	1	Br attached to aromatic carbon	0.19751[a]	Bromobenzene	80	506
	Iodine groups						
47	I-C(na)	1	I attached to a nonaromatic carbon	0.55634[c]	Ethyl iodide	67	601
183	I-C(na)-1Halo	1	I attached to a Carbon with one other halogen atom	0.28233[c]	Diiodomethane	65	602
184	I-C(na)-2Halo	1	I attached to a Carbon with at least two other halogen atom	0.02627[c]	Perfluoro-n-amyliodide	66	603
169	I-C(a)	1	I attached to an aromatic carbon	0.27545[c]	Iodobenzene	68	606
	Oxygen groups						
53	C(na)-COOH	3	COOH Group attached to a carbon	3.23972	Acetic acid	32	701
48	C(a)-COOH	3	Aromatic COOH	4.24609	Benzoic acid	31	703
49	C(c)-OH	1	OH Group attached to a chain carbon	3.63398	1-Hexanol	94	704
50	C(a)-OH	1	Aromatic OH	3.11812	Phenol	96	706
51	C(na)-O-C(na)	1	Ether oxygen	1.08130	Diethyl ether	99	707
54	C(c)-COO-C(c)	3	Ester in a chain	1.32351	Ethyl acetate	33	709
55	HCOO-C(c)	3	Formic acid ester	1.39393	Methyl formate	35	710
57	O=C(a)<	2	Ketone bonded to aromatic ring	1.29807	Acetophenone	62	712
58	O=C(na)<	2	Ketone	1.48579	Acetone	63	713
59	HCO-C(na)	2	Aldehyde in Chain	1.32114	Acetaldehyde	61	714
60	HCO-C(a)	2	Aldehyde attached to an aromatic ring	1.14886[a]	Benzaldehyde	60	715
63	(-C=O-O-C=O-)r	7	Cyclic anhydrides with double or aromatic bond	5.22703[b]	Maleic anhydride	17	718
68	>N-(C=O)-N<	4	Urea	5.87649[b]	1,1,3,3-Tetramethyl urea	11	723
159	A_Ester	3	Ester attached to an aromatic carbon (via the carbon atom)	3.24164	Benzoic acid methyl ester	4	725
	Nitrogen groups						
70	-CONH2	3	Amide with no substituents	3.47415[c]	Acetamide	36	801
71	-CONH-	3	Amide with one substituent attached to the nitrogen	6.69052[b]	N-Methylformamide	19	802
72	-CON<	3	Amide with two substituents attached to the nitrogen	3.07950[b]	N,N-Dimethylformamide (DMF)	20	803
75	NO2-C(na)	3	Nitro group attached to a nonaromatic carbon	2.94313	1-Nitropropane	28	807
76	NO2-C(a)	3	Nitro group attached to an aromatic carbon	1.60597	Nitrobenzene	29	808
80	NH2-C(na)	1	Primary amine attached to nonaromatic carbon	0.29440	Hexylamine	101	812
81	NH2-C(a)	1	Primary amine attached to an aromatic carbon	4.04079[b]	Benzidine	100	813
82	-N(na)H-	1	Secondary amines	1.06841	N,N-Diethylamine	104	814
84	>N(na)-	1	Tertiary amine	0.41783	Triethylamine	105	818
87	N(a)-(AA14)	1	Aromatic nitrogen in a five-membered ring	0.28286	Pyrrole	103	825
88	N(a)-(AA15)	1	Aromatic nitrogen in a six-membered ring	0.92045[b]	Pyridine	102	826
89	(C)-CtN	2	CN Group attached to a carbon	2.79589	Acetonitrile	64	827
164	-(O=C)N-(C=O)-	5	Pyrrolidinedione group	3.84625	Theophylline	3	831
	Sulfur groups						
100	-S(na)-	1	Thioether	0.00042	Ethyl methyl sulfide	82	904
102	-SO2-	3	Sulfolane O=S=O	6.66546[a]	Sulfolane	45	906
104	-SO2N<	4	Sulfon amides, attached to N and to S with 2 double bond O	2.96712[a]	Sulfanilamide	44	908
105	>S=O	2	Sulfoxide	4.73397[a]	Dimethyl sulfoxide	47	909
	Metal groups						
110	(C)2>Sn<(C)2	1	Stannane with four carbon neighbors	0.71970[a]	Tetramethylstannane	69	1103
	Group interactions						
199	OH-OH	0	OH-OH	-29.18517[b]	1,2-Ethanediol	0	3001
197	OH-C=O	0	OH-C=O	-94.46863[b]	3-Hydroxy-2-butanone	0	3009
198	C=O-C=O	0	C=O-C=O	22.97883	2,3-Butanedione	0	3133

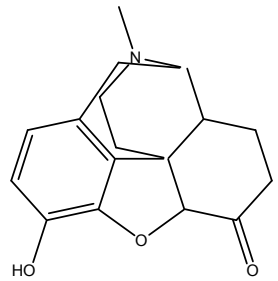
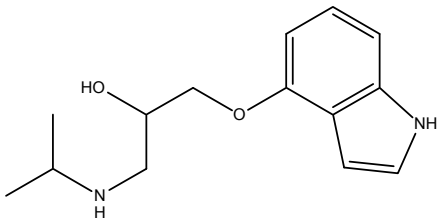
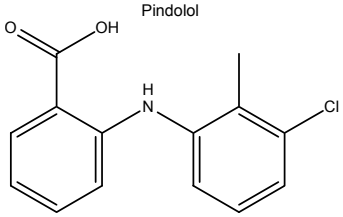
8.5.8. Test set

As mentioned above there were very few data available for the development of the method and therefore as much data as possible was used. For the test set, some data with solubility above (but close to) the chosen infinite dilution limit (x = 0.01) were taken from Beilstein [59] (the first 3 compounds in Table 8-13). The rest of the data was obtained from literature [114, 142, 161, 163] (Special thanks must go to Anja Diedrichs, a chemistry PhD student at the University of Oldenburg, who entered the data from the DECHEMA Data Series [114, 142] into Excel by hand and freely offered these data to me as a test set). The relative mean deviation (%) for the 14 data points in the test set is 11.68% which, strangely enough, is significantly better than the error for the training set data. The errors for the UNIFAC and the mod. UNIFAC methods are not altogether terrible in most cases, but the limiting factor very often is the unavailability of binary interaction parameters.

Table 8-13 Compounds with structures used in the test data set, comparison of the experimental activity coefficient with the predictions of this work, UNIFAC and mod. UNIFAC all at 298.15 K in hexane (data from various sources [59, 114, 142, 161, 163, 164]).

DDB	$\ln\gamma_2^{\infty}$ exp. 298.15 K	$\ln\gamma_2^{\infty}$ pred. 298.15 K	UNIFAC	mod. UNIFAC
 1-Nitronaphthalene	3.51	3.71	2.58	3.02
 p-dibromobenzene	1.20	1.22	2.91	2.32
 2-acetoxynaphthalene	3.32	3.18	2.36	3.04
 Progesterone	6.36	5.23	1.74	4.58

 Diphenylsulfone	6.19	7.36	-	-
 Sulfamethoxypyridazine	12.30	11.14	-	-
 4-Hexyl-1,3-benzenediol	6.69	7.30	4.46	5.55
 Morphine	12.86	11.16	-	-
 Flurbiprofen	5.45	5.50	-	-
 Flufenamic acid	8.36	7.45	-	-
 Niflumic acid	6.62	8.39	-	-

 Hydromorphone	8.04	7.42	-	-
 Pindolol	8.70	7.30	-	-
 Tolfenamic acids	5.85	7.09	-	-

9. APPLICATIONS

The following sections outline some of the various applications of the methods developed in the preceding chapters. All applications which fall into this section are those which are more indirectly applied, so for example water solubility is a direct application of water infinite dilution activity coefficient, while the air-water partition coefficient is an indirect application. The applications shown here are not shown just for the sake of expounding the many applications but rather as another means of model testing. Many of these applications involve the value of γ^{∞} in water, which indicates the importance of being able to predict it with acceptable accuracy. As with the test data presented previously these test data have not been selected based on the model performance but rather on the availability of such data.

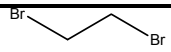
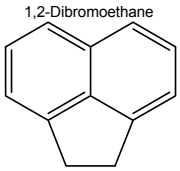
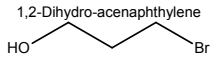
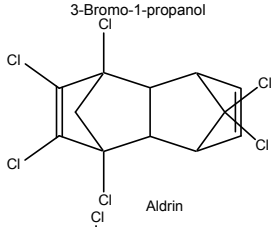
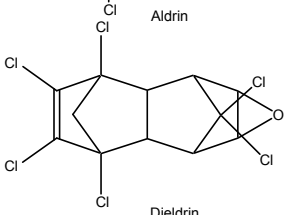
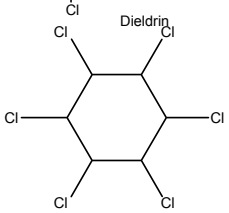
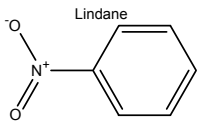
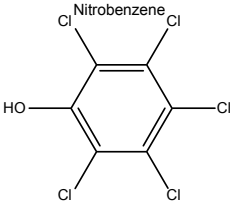
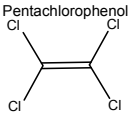
9.1. Air-water partition coefficient

The air-water partition coefficient (K_{aw}) is related to the infinite dilution activity coefficient in water via Eqn. (5-29). Since in a previous work [20] a method for the prediction of the vapour pressure of organic compounds was developed, the air-water partition coefficient is a good way to test the combination of both methods. Since the air-water partition coefficient is temperature dependent it provides a good opportunity to test the application of the general temperature dependence proposed in section 7.7.

As a first test, K_{aw} for some moderately complex compounds at 293.15 K were calculated; the results are shown in Table 9-1. In all cases the boiling point was predicted using the method of Nannoolal et al. [15] (even if there was an experimental value available) therefore the values are purely predicted from the molecular structure. It is important to note that the vapour pressure that is used in Eqn. (5-29) is the liquid vapour pressure. Therefore if the compound is a solid at room temperature then it is the hypothetical liquid vapour pressure. All predictions are within an order of magnitude of the experimental values which is an acceptable accuracy.

As a second test, some temperature dependent data randomly collected from literature were used as a model test. In this case suitable data were available for very few complex molecules (typically available at a single temperature). As mentioned above, these data are a good test of the model performance as their estimation utilizes the general temperature dependence which has been proposed for γ^{∞} in water. The results are shown in Fig. 9-1. The plot is shown in logarithmic scale to save space and not to hide any errors. The results in most cases are very good with the predicted lines following the experimental data fairly well. As an illustration of the uncertainty that is present in the K_{aw} data consider Fig. 9-2, it is clear from this plot that the error bars are very big, in some cases the bars cover an order of magnitude.

Table 9-1 Comparison of the experimental and predicted air-water partition coefficients (K_{aw}) calculated from Eqn. (5-29); γ^{∞} taken from the method developed in chapter 7, vapour pressure from Moller et al. [20] and normal boiling point from Nannoolal et al. [15], all data from literature [91].

Name	K_{aw} exp.	K_{aw} pred.	T_b (K) pred.
 1,2-Dibromoethane	2.80E-02	3.19E-02	400.15
 1,2-Dihydro-acenaphthylene	6.20E-03	3.16E-03	558.06
 3-Bromo-1-propanol	4.60E-06	8.70E-06	446.37
 Aldrin	6.10E-04	3.04E-03	643.68
 Dieldrin	8.90E-06	5.57E-05	665.32
 Lindane	2.20E-05	1.46E-05	596.72
 Nitrobenzene	9.30E-04	5.36E-04	480.28
 Pentachlorophenol	1.50E-04	4.94E-05	580.46
 Tetrachloroethylene	3.40E-01	3.04E-01	387.94

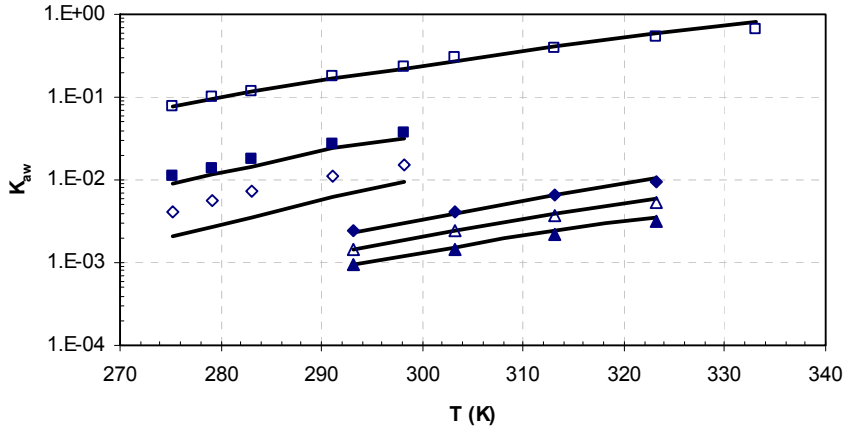


Fig. 9-1 K_{aw} vs. T for some compounds (□ - benzene [165], ■ - 1,1,2-trichloroethane [165], ◇ - methylphenylether [165], ◆ - 1-nitropropane [166], △ - nitroethane [166], ▲ – nitromethane [166], ── data predicted from Eqn. (5-29); γ^{∞} taken from the method developed in chapter 7, vapour pressure from Moller et al. [20] and normal boiling point from the DDB [28]).

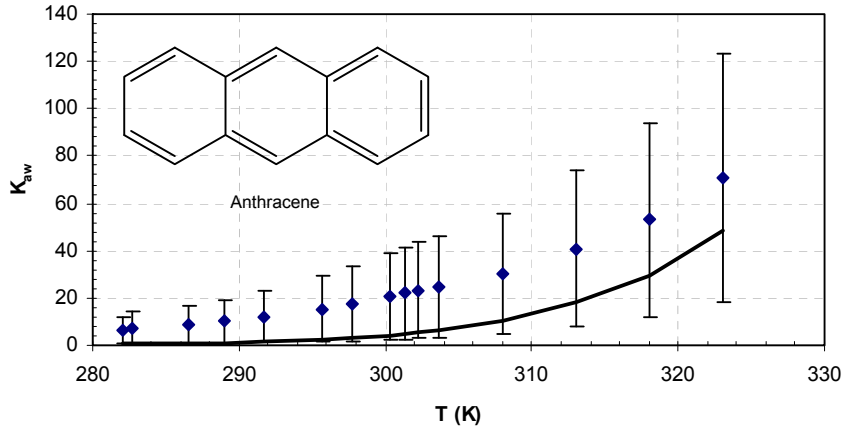


Fig. 9-2 K_{aw} vs. T for anthracene (◆ - data and uncertainty (error bars) from literature [93], ── data predicted from Eqn. (5-29); γ^{∞} taken from the method developed in chapter 7, vapour pressure from Moller et al. [20] and normal boiling point from the DDB [28]).

9.2. Alcohol interpolation

Section 6.2.1 outlined a method which enabled the interpolation of alcohol infinite dilution activity coefficient data for a single solute between solvents. Water and hexane where included as worst case scenarios for alcohols (hexane representing an infinitely long alcohol and water the smallest possible alcohol). Therefore since methods have been developed for the infinite dilution activity coefficient in water and hexane, predictions for infinite dilution activity coefficients in alcohols are possible. Some results are shown in Fig. 9-3 to Fig. 9-6. In most cases the interpolations are fairly good. Table 9-2 shows the interpolations for a selection of compounds; again the interpolations are fairly good in most cases.

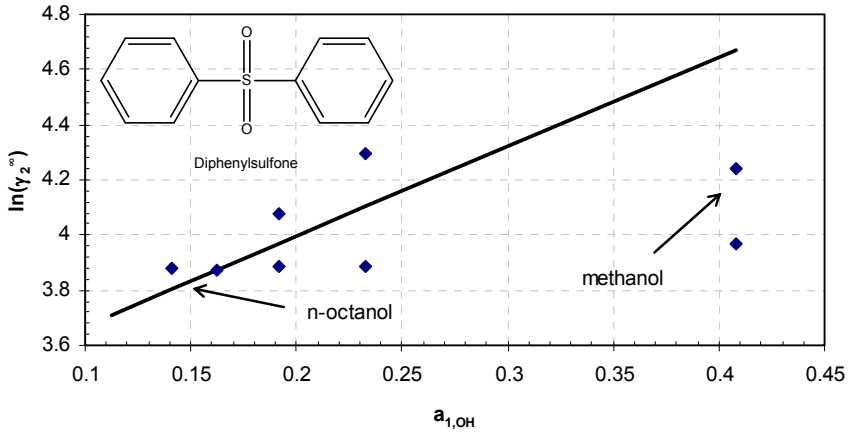


Fig. 9-3 $\ln\gamma_2^{\infty}$ vs. $a_{1,OH}$ for diphenylsulfone (2) in alcohols (1) (◆ - Data extracted from SLE data from Beilstein [59], $a_{1,OH}$ given by Table 6-5).

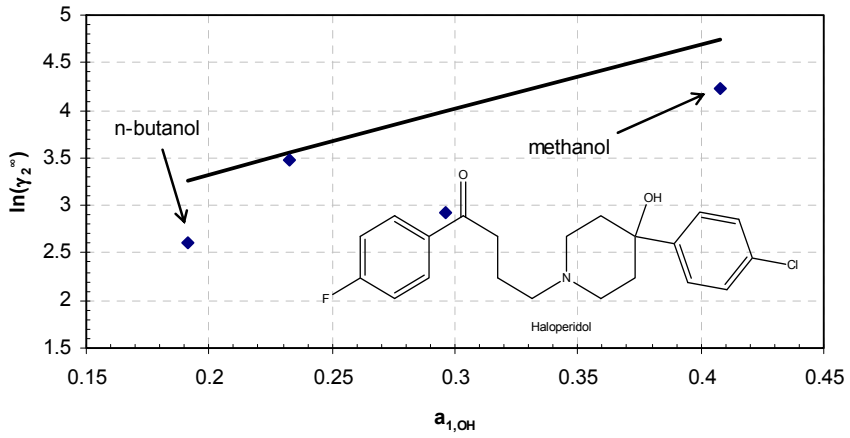


Fig. 9-4 $\ln\gamma_2^{\infty}$ vs. $a_{1,OH}$ for haloperidol (2) in alcohols (1) (◆ - Data extracted from SLE data from literature [114], $a_{1,OH}$ given by Table 6-5).

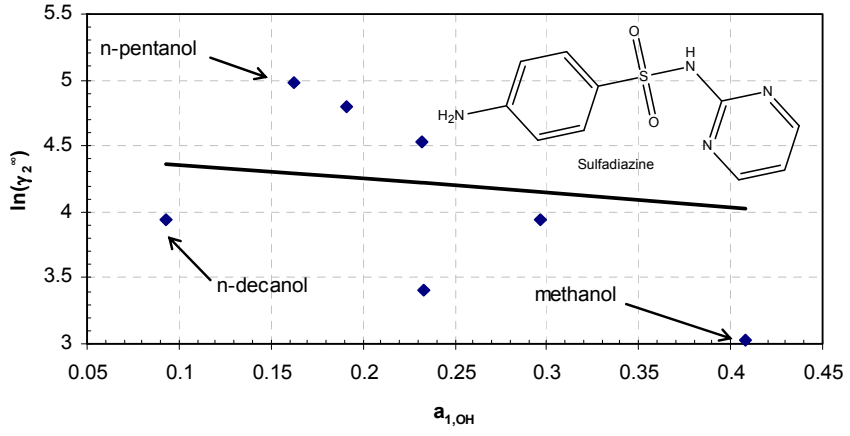


Fig. 9-5 $\ln\gamma_2^{\infty}$ vs. $a_{1,OH}$ for sulfadiazine (2) in alcohols (1) (◆ - Data extracted from SLE data from literature [119], $a_{1,OH}$ given by Table 6-5).

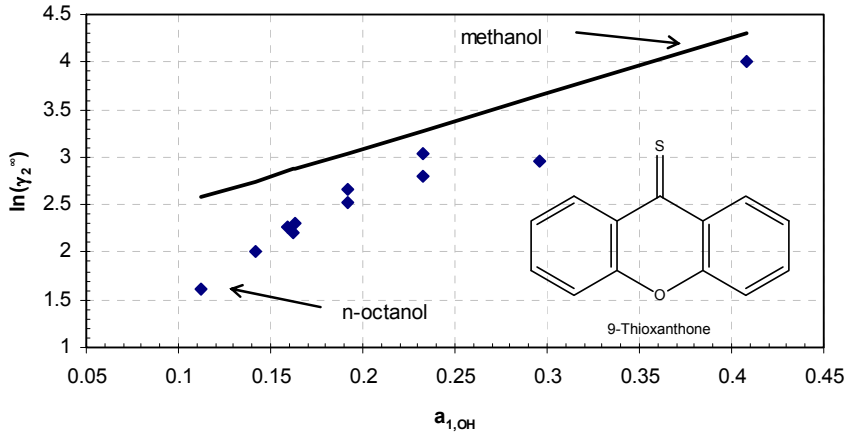
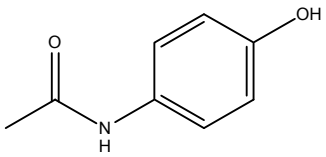


Fig. 9-6 $\ln\gamma_2^{\infty}$ vs. $a_{1,OH}$ for 9-thioxanthone (2) in alcohols (1) (◆ - Data extracted from SLE data the DDB [28], $a_{1,OH}$ given by Table 6-5).

Table 9-2 Predicted results for the infinite dilution activity coefficient of various compounds in alcohol solvents using the water and hexane predictions and Eqn. (6-27).

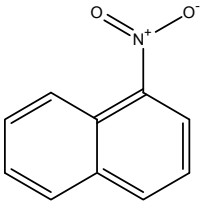
Solute



Paracetamol

Solvent	a_{OH}	a_{CH3}	$\ln\gamma^{\infty}$ exp.	$\ln\gamma^{\infty}$ pred.
Water	1.7	0		2.56
Ethanol	0.296	0.704	2.81	3.77
Hexane	0	2.2		10.38

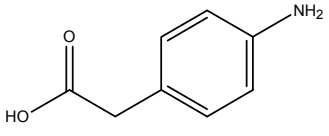
Solute



1-Nitronaphthalene

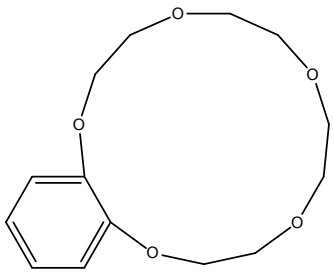
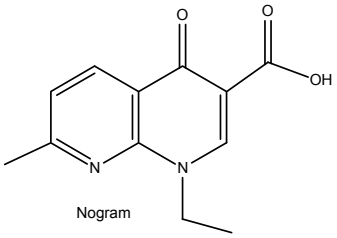
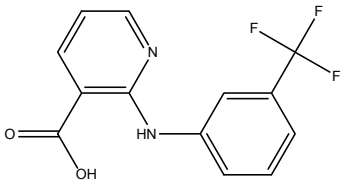
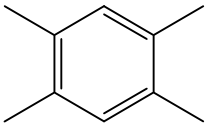
Solvent	a_{OH}	a_{CH3}	$\ln\gamma^{\infty}$ exp.	$\ln\gamma^{\infty}$ pred.
Water	1.7	0		11.23
2-Methyl-1-propanol	0.192	0.808	2.36	2.63
1,2-Ethanediol	0.779	0.221	4.68	5.52
Hexane	0	2.2		3.71

Solute



p-Aminophenylacetic acid

Solvent	a_{OH}	a_{CH3}	$\ln\gamma^{\infty}$ exp.	$\ln\gamma^{\infty}$ pred.
Water	1.7	0		3.41
Methanol	0.408	0.592	1.90	3.11
Ethanol	0.296	0.704	2.50	3.32
Hexane	0	2.2		8.50

Solute				
				
	Benzo-15-crown-5			
Solvent	**a_{OH}**	**a_{CH3}**	**$\ln\gamma^{\infty}$ exp.**	**$\ln\gamma^{\infty}$ pred.**
Water	1.7	0		5.36
Methanol	0.408	0.592	2.50	2.20
Hexane	0	2.2		3.41
Solute				
	Nogram			
Solvent	**a_{OH}**	**a_{CH3}**	**$\ln\gamma^{\infty}$ exp.**	**$\ln\gamma^{\infty}$ pred.**
Water	1.7	0		3.92
Ethanol	0.296	0.704	3.99	3.65
Hexane	0	2.2		9.29
Solute				
	Niflumic acid			
Solvent	**a_{OH}**	**a_{CH3}**	**$\ln\gamma^{\infty}$ exp.**	**$\ln\gamma^{\infty}$ pred.**
Water	1.7	0		11.02
Octanol	0.112	0.888	5.57	4.11
Hexane	0	2.2		8.39
Solute				
	1,2,4,5-Tetramethylbenzene			
Solvent	**a_{OH}**	**a_{CH3}**	**$\ln\gamma^{\infty}$ exp.**	**$\ln\gamma^{\infty}$ pred.**
Water	1.7	0		12.85
Methanol	0.408	0.592	3.33	3.20
1,2-Ethanediol	0.779	0.221	5.99	5.93
Hexane	0	2.2		0.43

9.3. Water-alcohol cosolvents

Following on from the methods presented in the previous section, as shown in section 6.2.1.1, these methods can be extended to cosolvent mixtures. The amount of cosolvent data available in literature is scarce but of the data that was tried the results are acceptable. Some

results are shown in Fig. 9-7 - Fig. 9-10, it is clear that the predictions for cumene and PCB 155 are excellent, however the predictions for the other compounds and not so good. Nevertheless when the water solution is dilute in the cosolvent the results are good as one would expect since the error for the water method is quite low and therefore there is a good reference point. The deviation for ketoprofen is partially due to the data in the ethanol rich mixture being above the infinite dilution limit.

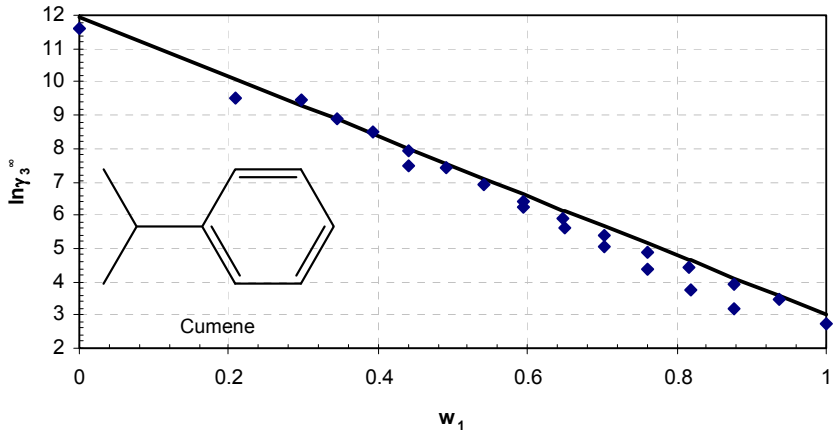


Fig. 9-7 $\ln\gamma_3^\infty$ vs. w_1 for cumene (3) in a mixture of water (2) and methanol (1) at 298 K (◆ - data from the DDB [28], ── Eqn. (6-27) using Eqn. (6-28) with the predictions for water and hexane).

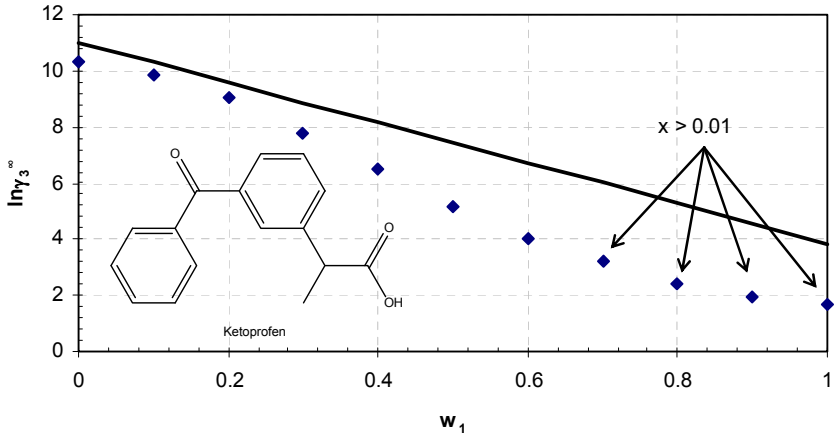


Fig. 9-8 $\ln\gamma_3^\infty$ vs. w_1 for ketoprofen (3) in a mixture of water (2) and ethanol (1) at 298 K (◆ - data extracted from SLE data from literature [122], ── Eqn. (6-27) using Eqn. (6-28) with the predictions for water and hexane).

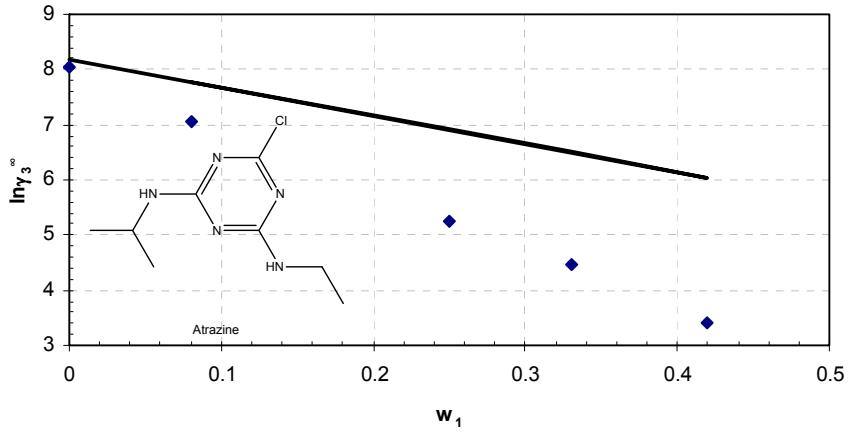


Fig. 9-9 $\ln\gamma_3^\infty$ vs. w_1 for atrazine (3) in a mixture of water (2) and ethanol (1) at 298 K (◆ - data extracted from SLE data from literature [167], ── Eqn. (6-27) using Eqn. (6-28) with the predictions for water and hexane).

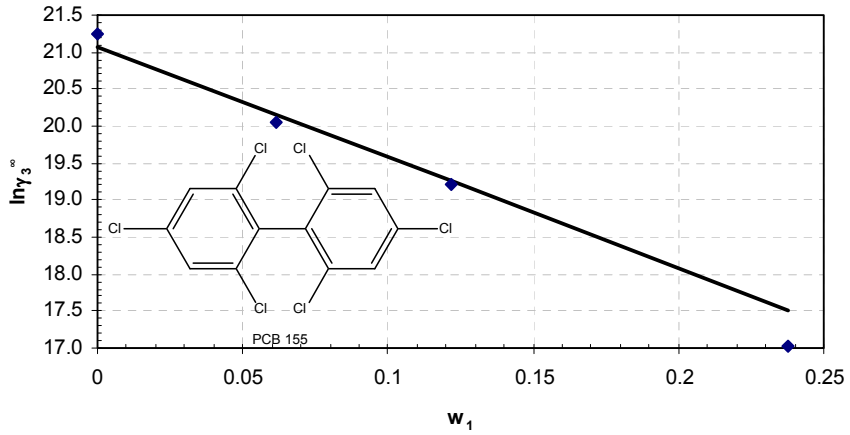


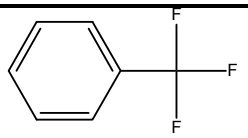
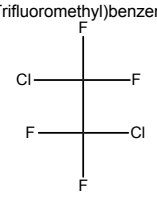
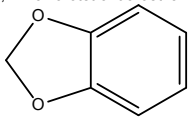
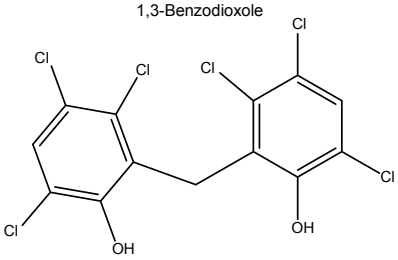
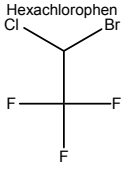
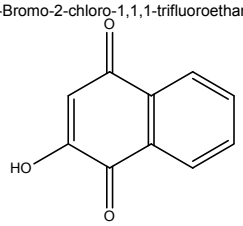
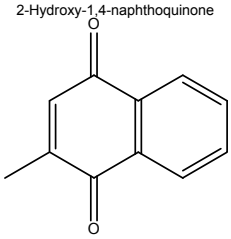
Fig. 9-10 $\ln\gamma_3^\infty$ vs. w_1 for PCB 155 (3) in a mixture of water (2) and n-propanol (1) at 298 K (◆ - data extracted from SLE data from literature [168], ── Eqn. (6-27) using Eqn. (6-28) with the predictions for water and hexane).

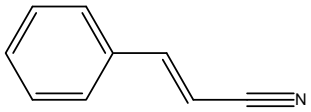
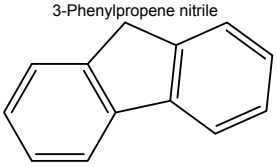
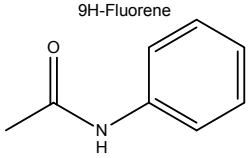
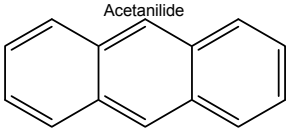
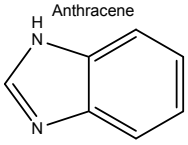
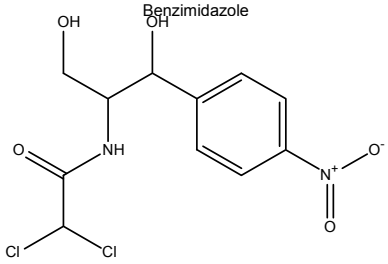
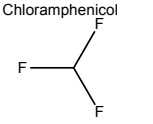
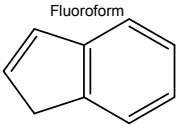
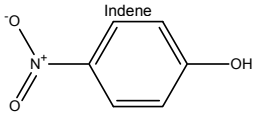
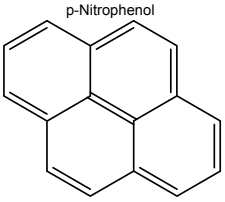
9.4. Octanol-water partition coefficient

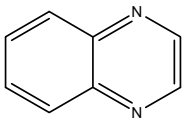
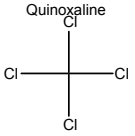
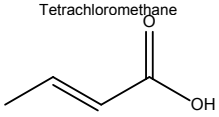
It has been shown above that it is possible to use the predictions of the infinite dilution activity coefficient in water and hexane to calculate the values in alcohols. One obvious application of this is therefore the calculation of the octanol-water partition coefficient. There are 2 different methods for calculating K_{ow}; firstly via the correlation with the pure component γ^∞ given by Eqn. (5-23). Secondly (as shown in section 9.3), since it is possible to calculate γ^∞ in a mixture of octanol and water it is possible to use Eqn. (5-22) directly. The mass fraction solubility of water (2) in octanol (1) is w_2 = 0.035, therefore for the water-saturated octanol phase Eqn. (6-28) yields $a_{ow,OH} = 0.168$ and $a_{ow,CH_3} = 0.857$. The use of the correlation (Eqn. (5-23)) will be denoted as method 1 and the use of Eqn. (5-22) as method 2. The comparison of the results for some test compounds are shown in Table 9-3, both methods perform very similarly and in

all cases (except the 2 nitrates and fluoroform) the predictions are good. There are however a few instances where there is a large failure in the prediction. As mentioned above this is proposed as a test of the method and will probably not have the required accuracy to displace the methods which are fitted directly to K_{ow} [7, 169, 170].

Table 9-3 Comparison of the predicted and experimental logK_{ow} values pred1 refers to method 1 and pred2 referes to method 2, all data from literature [91].

DDB	LogK_{ow} Exp.	LogK_{ow} Pred1.	LogK_{ow} Pred2.
 (Trifluoromethyl)benzene	2.90	2.80	2.83
 1,2-Dichlorotetrafluoroethane	2.82	2.52	2.53
 1,3-Benzodioxole	2.08	1.66	1.63
 Hexachlorophen	3.93	3.52	3.60
 2-Bromo-2-chloro-1,1,1-trifluoroethane	2.30	2.05	2.04
 2-Hydroxy-1,4-naphthoquinone	1.46	2.55	2.56
 2-Methyl-1,4-naphthochinone	2.20	2.33	2.34

 3-Phenylpropene nitrile	1.96	1.76	1.73
 9H-Fluorene	4.12	3.97	4.08
 Acetanilide	1.16	0.37	0.25
 Anthracene	4.45	4.11	4.22
 Benzimidazole	1.34	2.19	2.15
 Chloramphenicol	1.14	-2.73	-3.03
 Fluoroform	0.64	1.60	1.56
 Indene	2.92	2.86	2.9
 p-Nitrophenol	1.91	0.33	0.22
 Pyrene	4.88	4.63	4.78

 Quinoxaline	1.08	1.81	1.78
 Tetrachloromethane	2.83	2.62	2.64
 trans-2-Butenoic acid	0.72	0.26	0.14

9.5. Improvement of group contribution models

As shown in section 6.1, the combinatorial expressions used by UNIFAC and mod. UNIFAC are not sufficient to capture the size/shape affects which are present in alkane rich mixtures. A new expression for infinite dilution which showed the correct behaviour was proposed (Eqn. (6-6)). Therefore it could be possible that simply using the new expression instead of the mod. UNIFAC and UNIFAC combinatorials could lead to an improvement in the predictions for large solutes in small solvents. When analysing the results of the mod. UNIFAC combinatorial expression it became apparent that the values produced were sometimes not physically realistic. This is probably due to the empirical adjustment to the r and q values used in mod. UNIFAC (this is briefly discussed in Appendix B). Analysis was therefore solely focused on the UNIFAC method.

Initially it had been hoped that the data collected for the development of the water method could be used to test the combinatorial replacement. Theoretically this is a perfect application since, as shown in section 6.1, the SG expression severely under-predicts the size/shape interactions for “large in small” and water is smaller that all solutes. The opposite is however true, as using the new expression results in significantly larger errors. The reason for this is the r and q values used in UNIFAC are not realistic (they were fitted) and therefore result in, convenient but, incorrect combinatorial values (this point is discussed in Appendix B).

The data for γ^{∞} in hexane used in chapter 8 was therefore used to test the combinatorial swap. This is also a convenient application as there is some SLE data for large solutes in small solvents which should result in a lower error. The overall percentage error goes from 40.9 % to 36.8 % with the new combinatorial expression (no fitting of parameters). While this seems like a rather small improvement this means that the overall error is lower than that of mod. UNIFAC. Also when one looks at the improvement for “big in small” as shown in Fig. 9-11 it is clear that there is a marked improvement for large in small ($q_1/q_2 < 1$) and only a slight worsening of small in large ($q_1/q_2 > 1$).

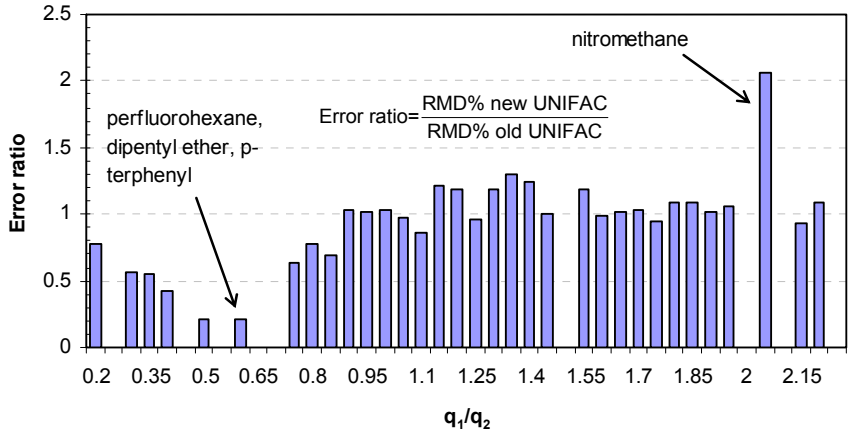


Fig. 9-11 Error ratio in $\ln\gamma_2^{\infty}$ vs. q_1/q_2 (solvent(1)/solute(2)) for the data presented in Chapter 8 (new UNIFAC refers to UNIFAC with Eqn. (6-6) while old UNIFAC refers to UNIFAC with the original combinatorial, an error ratio less than 1 means a better RMD while greater than 1 is a worse RMD).

Since the new combinatorial part was successfully applied to the hexane data set the next step was to apply the combinatorial (again by simply swapping the combinatorials with no parameter fitting) to UNIFAC for the prediction of γ^{∞} in any solvent. As shown in Table 9-4 the new combinatorial results in a significant reduction in the overall RMD for the data in the DDB γ^{∞} database. When analysing the error ratio (ratio of the errors in $\ln\gamma^{\infty}$ for the new and old combinatorials) as shown in Fig. 9-12 it seems that the new combinatorial causes significantly worse errors. This is not the case, rather for such a large set of data an average error ratio for each size ratio probably isn't a good measure of the performance (since one large outlier has a significant effect on the average while an almost zero value has a very small effect). The error ratio plot is just included for completeness as it was used above, although as mentioned it isn't a suitable measure. When analysing the error ratio on a data point by data point basis the improvement becomes more apparent; 17972 data points (or 58.94% of the data) had an error ratio less than 1 (i.e. an improvement in the RMD). What is surprising is that the improvement is not highly focused in the region where $q_1/q_2 < 1$ but rather it is fairly spread across the whole range. There are 2 possible reasons for this:

1. Many of the data for $q_1/q_2 < 1$ is for systems with solutes such as methanol, acetonitrile, acetone, 1,2-ethandiol etc. (and obviously water but these data were not included). Many of these compounds (*viz*. methanol, 1,2-ethandiol, acetonitrile etc) have specific main groups which describe the whole molecule and therefore one may expect that the residual is highly trained to the combinatorial expression. This is what is observed, when the new combinatorial is used. In most cases it results in a larger value of γ^{∞} and for such systems this typically results in an over prediction.

2. The combinatorial developed in this work was used with alkanes which do not differ significantly with regards to size/shape and therefore in many instances the combinatorial values are similar to those of the SG expression. For cyclic, aromatic

and smaller molecules this may not be the case and therefore this difference could be significant.

Table 9-4 Relative mean deviation (RMD) in $\ln\gamma^{\infty}$ when using the UNIFAC parameters with the original and the new combinatorial expression (comparisons made with data from the DDB, all data for water – i.e. water as a solute and solvent – were excluded and all data falling in the range $0.99 < \ln\gamma < 1.01$ were excluded as they greatly skew the percentage errors).

Combinatorial	RMD%	No. data points
Original	101.2	29642
Eqn. (6-6)	82.3	29634

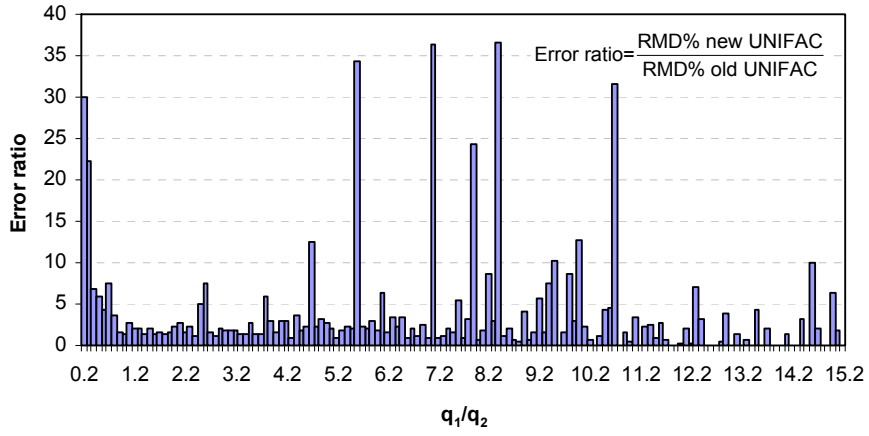


Fig. 9-12 Error ratio in $\ln\gamma_2^{\infty}$ vs. q_1/q_2 (solvent(1)/solute(2)) for all the γ^{∞} data in the DDB [28] (new UNIFAC refers to UNIFAC with Eqn. (6-6) while old UNIFAC refers to UNIFAC with the original combinatorial, an error ratio less than 1 means a better RMD while greater than 1 is a worse RMD).

The success of the application of the combinatorial developed in this work to UNIFAC could suggest that it can be applied in all situations where the SG expression is applied. One such example is the COSMO-RS(OL) method which uses the SG combinatorial. Introducing the new combinatorial instead of the SG expression does not result in improved predictions. The reason is that the residual calculated from COSMO methods isn't constant for a single solute in any alkane solvent as it is in the group contribution methods. Therefore the methodology presented here is only applicable to group contribution methods.

10. CONCLUSIONS

In the first sections of this work some of the literature methods used for the prediction/modelling of the activity coefficient were discussed and briefly evaluated. Even though several models are available, for complex compounds reliable and accurate predictions are almost impossible. The poor performance of these models show the difficulty in describing the real mixture behaviour as a function of temperature, pressure, concentration, solvent and solute. Group contribution methods simplify the vast number of possible solvent-solute interactions by considering a solution of functional groups, whereby the possible number of functional group interactions is significantly lower than the possible solute-solvent interactions.

Since the number of required interaction parameters grows quadratically with the number of functional groups (main groups), they must be kept to a minimum. Hence these methods very frequently represent an oversimplification. Previous work on pure component properties in this group has led to very reliable methods. It was therefore considered to model the activity coefficient at infinite dilution of many solutes in a single solvent in a similar manner. The difference is the solute is now surrounded by a solvent made up of a single different component instead of molecules of its own kind. To further simplify the problem, only data at 298 K were considered. The only problem left was to have a sufficient amount of experimental data for the regression of the group contribution values.

For water the amount of available data is vast and there was sufficient data to develop a group contribution method (see Chapter 7). Infinite dilution data was taken from the DDB γ^{∞} database and extracted from SLE data. In order to extract the γ^{∞} from SLE it was assumed that all solubility data below 1 mole percent was infinitely dilute. When extracting γ^{∞} from SLE data, fusion data are needed; this includes heat capacities, melting temperatures and enthalpies of fusion. While enthalpy of fusion and melting temperature are relatively abundant (melting point temperature data are much more abundant than heat of fusion data) the heat capacity data are very scarce. The heat capacity change upon melting is, however, typically assumed to be negligible [38, 83]. This is an oversimplification, if it is assumed that the heat capacity change upon melting is approximated by the entropy of fusion [33, 84, 85] superior activity coefficients can be extracted. The infinite dilution activity coefficient in water is of great practical interest as it can be used to calculate a large amount of other data (air-water partition coefficient, water solubility etc.).

Even with the relatively large amount of data available in water there were still some functional groups present for which no data were available to regress groups to. This is a widely known weakness of group contribution methods. A method was developed which allows for missing halogen groups to be predicted from current group values. This method was applied to some test compounds with great success and was used to fill the halogen

group values. A general method was proposed for other groups, however no data could be found to validate the correlations presented.

When considering the temperature and pressure dependence of the data available in the training set it was apparent that while the pressure effect was negligible (up to moderate pressures) the temperature dependence certainly wasn't. Quite surprisingly it was found that a very simple correlation could describe the temperature dependence quite well in most instances. The only requirement for this correlation to be applicable is that the solute has a low solubility in water. This general temperature dependence in combination with the vapour pressure model developed previously [20] allowed for the prediction of the air-water partition coefficient (section 9.1).

The overall RMD for the training set was 7.3% for 630 compounds which compared very favourably to mod. UNIFAC and UNIFAC (14.9% for 393 compounds and 23.7% for 442 compounds respectively). The greatest strength of this method is that it can be applied to multifunctional compounds with better accuracy and wider applicability than previous methods. The overall RMD for the multifunctional (i.e. more than one different function group) compounds was 8.7% for 155 compounds which was far superior to mod. UNIFAC and UNIFAC (27.8% for 54 compounds and 46.1% for 76 compounds respectively). As a test set a group of 25 diverse compounds was analysed, the overall error was 9.1% which is only slightly higher than the overall RMD and is quite good bearing in mind the compounds considered.

When looking for other solvents it soon became clear that there is scarce data in any single solvent. What was apparent is that there was a large amount of γ^{∞} in a variety of heavy alkanes (which are popular as stationary phases in gas-liquid chromatography due to their low volatility). Therefore some time was invested in trying to reduce all this data to a common (alkane) solvent. The common solvent was chosen as hexane due to the fair amount of SLE data available in hexane. The method developed (see section 6.1) utilized the solution of groups principle and allows the activity coefficient to be converted from one alkane solvent to another based only on the combinatorial ratio in both solvents.

Quite unexpectedly the method highlighted some interesting shortfalls present in the common combinatorial expressions (*viz*. SG and mod. UNIFAC). Both of these combinatorial expressions were found to be inferior to a free-volume expression found in literature [75]. However when applied to systems involving large solutes in small(er) solvents all methods tried show the incorrect behaviour. It is proposed that this deviation is due to some cavitation contribution (The exact nature of this contribution is not known It is speculated that it could be due to diffusion of the solute due to a decrease in the cavitation energy required in smaller solvents. There is, however, quite a lot of uncertainty and therefore it is simply termed a cavitation contribution) which is not captured by the combinatorial expression. This effect was captured by adding in an empirical correction to the free-volume term.

This new combinatorial expression shows the correct behaviour for both large in small small in large molecules and was empirically extended to finite concentrations. This combinatorial expression now allows for the extrapolation of the easily available data for a solute in some large alkane (which can be a stationary phase on a GC column), to a small solvent. For all compounds tried the results of the extrapolations were good.

γ^{∞} data for a larger number of solutes in alkanes were converted to values in hexane for the development of an estimation method for γ^{∞} in hexane. Even with this procedure, there was still only a limited amount of data and therefore many of the groups were simplified in order to reduce the number of model parameters. The methods developed for estimating missing group values in the water method were applied successfully in the case of the solvent hexane. This is considered an important finding and the methodology should also be applied to other methods in the future. As with the activity coefficient in water method a general temperature dependence was proposed which seems to be adequately applicable.

The overall RMD for the training set was 21.4% for 181 compounds which compared very favourably to mod. UNIFAC and UNIFAC (38.5% for 150 compounds and 40.9% for 155 compounds respectively). The greatest strength of this method is that it can be applied to multifunctional compounds with good accuracy and wide applicability. The overall RMD for the multifunctional (i.e. more than one different function group) compounds was 9.3% for 17 compounds which was far superior to mod. UNIFAC and UNIFAC (21.8% for 8 compounds and 28.9% for 9 compounds respectively). For the test set a group of 14 diverse compounds was analysed, the overall error was 11.7% which is, surprisingly, significantly lower than the training set error even though the compounds considered are fairly complex.

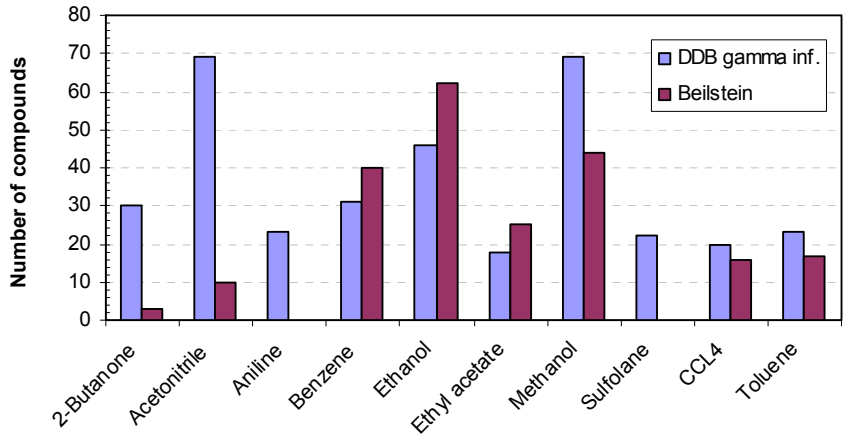


Fig. 10-1 Number of data available in DDB γ^{∞} [28] and Beilstein [59] for some common solvents at 298.15 K (for Beilstein data only compounds for which fusion data are available are shown).

As a result, there are methods available now for predicting γ^{∞} of a large variety of solutes in water and hexane. These are however, obviously, not the only solvents of interest. It would therefore be desirable to add a few other important solvents for which pure component type

methods could be developed. There is however very few data available as is evident in Fig. 10-1 and these are available are very frequently for moderately soluble compounds (outside of the assumed infinite dilution range).

Due to their individual importance the solvents with the most data available are 2 alcohols (communication with a large pharmaceutical company revealed a list of 16 preferred solvents whereby alcohols, water and hexane made up half of the list). It is for this reason that the method presented in section 6.2.1 was developed. This includes water and hexane in the alcohol homologous series and allows for interpolation between water and hexane to predict the behaviour in alcohol solvents. This method is very empirical but nevertheless many of the interpolations tried were of an acceptable accuracy. The ability to predict γ^{∞} in water and alcohols allows for the octanol water partition coefficient to be predicted. The predictions that were shown were good in many cases with very few failures. Nevertheless this method is proposed more as a proof of concept rather than a replacement for specifically fitted K_{ow} methods, as one would expect them to be superior in almost all situations.

With the success of the interpolation between water and hexane it had been hoped that this approach could be extended to further. Owing to the scarcity of the data in any single polar solvent large scale systematic measurements would be required which was outside the scope of this work. Any method that would be developed with a training set of around 100 different solutes would almost certainly be over fitted and therefore would be useless for practical application.

11. RECOMMENDATIONS

11.1. Taking UNIFAC into the 21st century

The great success and wide applicability of the group contribution methods such as UNIFAC and mod. UNIFAC has shown that these methods are of great practical value. Nevertheless they contain many shortfalls which inhibit the user from making predictions about some systems with reasonable confidence (and sometimes any prediction at all). Arguably one of the greatest shortfalls of such methods is that the availability of group interactions is dependent on the availability of training set data. It would seem that there needs to be some investigation into being able to predict the missing group interactions without having to perform tedious and expensive experimental measurements. One possible workaround is briefly discussed in Appendix A.

Another problem which was raised with the group contribution methods by numerous researchers is that for larger multifunctional compounds the assumption of group additivity might be a poor one. While this may be, to a certain extent, true a larger part of the problem could be down to the combinatorial expressions used in these methods. An infinite dilution combinatorial is proposed in this work, this combinatorial, however, suffers from the fact that is not thermodynamically consistent and therefore cannot be easily extended from infinite dilution. Work therefore needs to be invested into an expression which is thermodynamically consistent but still captures the various effects present.

The combinatorial input parameters (i.e. r, q and perhaps V_m) should be physically realistic. In instances where these parameters are simply fitted to the equation (*viz*. water) the combinatorial expression differs greatly from what one may typically expect (see Appendix B). The focus should rather be on using physically realistic values and making sure that the equation reproduces the correct behaviour. What a totally reworked combinatorial probably means is that the binary interaction parameters would need to be refitted. While this is undoubtedly a massive task it does seem that in order for UNIFAC to stay “current” there needs to be a rethinking (and perhaps reworking) of the underlying equations.

In order to take UNIFAC methods into the 21st century it may be possible to combine the great potential of the COSMO-RS type methods with the practicality of the UNIFAC methods. While it is clear that the COSMO-RS type methods have great potential, they are far less reliable in practice. Therefore if one could develop a reliable way to calculate binary interaction parameters from COSMO-RS calculations this would be a huge step forward and would bridge the gap between the wide applicability of COSMO-RS and the good practical applicability of the UNIFAC type methods. However since COSMO-RS calculations are not at the required level of reliability an approach such as that presented in Appendix A is probably more suitable.

11.2. Moving forward with the reference solvent approach

The reference solvent approach proposed in this work does seem to have good practical applications. The advantage of this approach over others is that the interpolations are only considered at infinite dilution and therefore do not need to cover the rather complicated concentration dependence of the activity coefficient. The two solvents that were considered in this work were hexane and water and based on the behaviour in these 2 solvents it is possible to extrapolate to any alkane and interpolate to any alcohol. While this is a very useful toolset it is still somewhat limited as it cannot be applied to other solvents. In order to make this approach more generally applicable one would need a method for at least one further polar solvent. The limiting factor for this is the availability of data as shown above (see Fig. 10-1). Therefore if this approach would be extended there needs to be push to measure more γ^{∞} data for simple compounds in polar solvents as these data are vital for getting reliable group values.

One possible solution to this problem would be to carry out some GLC measurements on a suitable stationary phase (sulfolane is one option as it has a high boiling point (which is desirable for a stationary phase) and it is polar). If the stationary phase is sufficiently stabile the amount of effort required for measurements would be minimal. It is therefore not overly optimistic to assume that 10 solutes could be measured per day. If sufficient number of volatile solutes are available a database of a couple hundred data points could be accumulated in a couple of months with little effort.

While it is convenient, for model development purposes, to only consider infinite dilution data for model development these values are only a guide for applications where higher concentrations are expected. Therefore there may need to be some work invested into extending infinite dilution data to finite concentrations. One possible method of doing this would be to consider the infinite dilution behaviour of the solute in the solvent and solvent in the solute and interpolate between the 2 extremes. This would however require γ^{∞} data for water in a large variety of compounds or some method to predict this data. Unfortunately this data are typically unavailable. It would be useful to be able to predict these values since the knowledge of both make it possible for more accurate conclusions about the mutual solubility.

12. REFERENCES

[1] Hoffmann, G. R., Mutagenicity Testing in Environmental Toxicology. *Environ. Sci. Technol.*, 1982. **16**(10): p. 560A-574A.

[2] Bowen, H. J., Palmer, S. R., Fielder, H. M. P., Coleman, G., Routledge, P. A., Fone, D. L., Community Exposures to Chemical Incidents: Development and Evaluation of the First Environmental Public Health Surveillance System in Europe. *J. Epidemiol. Community Health*, 2000. **54**(11): p. 870-873.

[3] Delaney, J. S., Predicting Aqueous Solubility from Structure. *Drug Discovery Today*, 2005. **10**(4): p. 289-295.

[4] Jain, P., Sepassi, K., Yalkowsky, S. H., Comparison of Aqueous Solubility Estimation from AQUAFAC and the GSE. *Int. J. Pharm.*, 2008. **360**: p. 122-147.

[5] Kuhne, R., Ebert, R. U., Kleint, F., Schmidt, G., Schuurmann, G., Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere*, 1995. **30**(11): p. 2061-2077.

[6] Lee, Y.-C., Myrdal, P. B., Yalkowsky, S. H., Aqueous Functional Group Activity Coefficients (AQUAFAC) 4: Application to Complex Organic Compounds. *Chemosphere*, 1996. **33**(11): p. 2129-2144.

[7] Marrero, J., Gani, R., Group-Contribution-Based Estimation of Octanol/Water Partition Coefficient and Aqueous Solubility. *Ind. Eng. Chem. Res.*, 2002. **41**: p. 6623-6633.

[8] Myrdal, P. B., Ward, G. H., Dannenfelser, R.-M., Mishra, D. S., Yalkowsky, S. H., AQUAFAC 1: Aqueous Functional Group Activity Coefficients Application to Hydrocarbons. *Chemosphere*, 1992. **24**(8): p. 1047-1061.

[9] Fredenslund, A., Jones, R. L., Prausnitz, J. M., Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE J.*, 1975. **21**(6): p. 1086-1099.

[10] Weidlich, U., Gmehling, J., A Modified UNIFAC Model. 1. Prediction of VLE, h^E and $\gamma\infty$. *Ind. Eng. Chem. Res.*, 1987. **26**: p. 1372-1381.

[11] Rarey, J., Gmehling, J., Factual Data Banks and their Application to the Synthesis and Design of Chemical Processes and the Development and Testing of Thermophysical Propertly Estimation Methods. *Pure Appl. Chem.*, 2009. **81**(10): p. 1745-1768.

[12] Klamt, A., Eckert, F., COSMO-RS: A Novel and Efficient Method for the A Priori Prediction of Thermophysical Data of Liquids. *Fluid Phase Equilib.*, 2000. **172**: p. 43-72.

[13] Lin, S.-T., Sandler, S. I., A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model. *Ind. Eng. Chem. Res.*, 2002. **41**: p. 899-913.

[14] Mu, T., Rarey, J., Gmehling, J., Performance of COSMO-RS with Sigma Profiles from Different Model Chemistries. *Ind. Eng. Chem. Res.*, 2007. **46**: p. 6612-6629.

[15] Nannoolal, Y., Rarey, J., Ramjugernath, D., Cordes, W., Estimation of Pure Component Properties, Part 1: Estimation of the Normal Boiling Point of Non-Electrolyte Organic Compounds via Group Contributions and Group Interactions. *Fluid Phase Equilib.*, 2004. **226**: p. 45-63.

[16] Nannoolal, Y., Rarey, J., Ramjugernath, D., Estimation of Pure Component Properties Part 2: Estimation of Critical Data by Group Contribution. *Fluid Phase Equilib.*, 2007. **252**: p. 1.

[17] Nannoolal, Y., *Development and Critical Evaluation of Group Contribution Methods for the Estimation of Critical Properties, Liquid Vapour Pressure and Liquid Viscosity of Organic Compounds*, in *Chemical Engineering*. 2006, University of KwaZulu Natal: Durban.

[18] Nannoolal, Y., Rarey, J., Ramjugernath, D., Estimation of Pure Component Properties Part 3: Estimation of the Vapor Pressure of Non-Electrolyte Organic Compounds via Group Contributions and Group Interactions. *Fluid Phase Equilib.*, 2008. **269**: p. 117-133.

[19] Nannoolal, Y., Rarey, J., Ramjugernath, D., Estimation of Pure Component Properties. Part 4: Estimation of the Saturated Liquid Viscosity of Non-electrolyte Organic Compounds via Group Contributions and Group Interactions. *Fluid Phase Equilib.*, 2009. **281**(97-119).

[20] Moller, B., Rarey, J., Ramjugernath, D., Estimation of the Vapour Pressure of Non-Electrolyte Organic Compounds via Group Contributions and Group Interactions. *J. Mol. Liq.*, 2008. **143**: p. 52-63.

[21] Mellan, I., *Industrial Solvents Handbook*. 2nd ed. 1977, New Jersey: Noyes Data Corporation.

[22] Likhodii, S. S., Serbanescu, I., Cortez, M. A., Murphy, P., Snead, O. C. I., Burnham, W. M., Anticonvulsant Properties of Acetone, a Brain Ketone Elevated by the Ketogenic Diet. *Annals of Neurology*, 2003. **54**(2): p. 219-226.

[23] Karunanithi, A. T., Achenie, L. E. K., Gani, R., A Computer-Aided Molecular Design Framework for Crystallization Solvent Design. *Chem. Eng. Sci.*, 2006. **61**: p. 1247-1260.

[24] Pretel, E. J., Lopez, P. A., Bottini, S. B., Brignole, E. A., Computer-Aided Molecular Design of Solvents for Separation Processes. *AIChE J.*, 1994. **40**(8): p. 1349-1360.

[25] Yang, X., Song, H., Computer Aided Molecular Design of Solvents for Separation Processes. *Chem. Eng. Technol.*, 2006. **29**(1): p. 33-43.

[26] Song, J., Song, H., Computer-Aided Molecular Design of Environmentally Friendly Solvents for Separation Processes. *Chem. Eng. Technol.*, 2008. **31**: p. 177-187.

[27] Reichardt, C., *Solvents and Solvent Effects in Organic Chemistry*. 3rd ed. 2003, Weinheim: Wiley-VCH.

[28] Gmehling, J., Rarey, J., Menke, J., Dortmund Data Bank, Oldenburg (2008) http://www.ddbst.com

[29] Jenkins, J. D., Gibson-Robinson, M., Vapour-Liquid Equilibrium in Systems with Association in Both Phases. *Chem. Eng. Sci.*, 1977. **32**: p. 931-938.

[30] Myrdal, P. B., Krzyzaniak, J. F., Yalkowsky, S. H., Modified Trouton's Rule for Predicting the Entropy of Boiling. *Ind. Eng. Chem. Res.*, 1996. **35**: p. 1788-1792.

[31] Chastrette, M., Rajzmann, M., Chanon, M., Purcell, K. F., Approach to a General Classification of Solvents Using a Multivariate Statistical Treatment of Quantitative Solvent Parameters. *J. Am. Chem. Soc.*, 1985. **107**(1): p. 1-11.

[32] Gramatica, P., Navas, N., Todeschini, R., Classification of Organic Solvents and Modelling of their Physio-Chemical Properties by Chemometric Methods using Different Sets of Molecular Descriptors. *Trends Anal. Chem.*, 1999. **18**(7): p. 461-471.

[33] Hildebrand, J. H., Scott, R. L., *Solubility of Nonelectrolytes*. 3rd ed. 1964, New York: Dover.

[34] Prausnitz, J. M., Lichtenhaler, R. N., de Azevedo, E. G., *Molecular Thermodynamics of Fluid-Phase Equilibria*. 3rd ed. 1999, New Jersey: Prentice-Hall.

[35] Lewis, G. N., The Law of Physico-Chemical Change. *Proc. Amer. Acad.*, 1901. **37**(3): p. 49-69.

[36] Lewis, G. N., Outlines of a New System of Thermodynamic Chemistry. *Proc. Amer. Acad.*, 1907. **43**(7): p. 259-293.

[37] Reid, R. C., Prausnitz, J. M., Poling, B. E., *The Properties of Gases and Liquids*. 4th ed. 1987, New York: McGraw-Hill.

[38] Smith, J. M., Van Ness, H. C., Abbott, M. M., *Introduction to Chemical Engineering Thermodynamics*. 6th ed. 2001: McGraw-Hill.

[39] Frank, T. C., Downey, J. R., Gupta, S. K., Quickly Screen Solvents for Organic Solids. *Chem. Eng. Prog.*, 1999: p. 41.

[40] Wilson, G. M., A New Expression for the Excess Free Energy of Mixing. *J. Am. Chem. Soc.*, 1963. **86**: p. 127-130.

[41] Wilson, G. M., Deal, C. H., Activity Coefficients in Changing Environments - Solutions of Groups. *Ind. Eng. Chem. Fundam.*, 1962. **1**(1): p. 20-23.

[42] Guggenheim, E. A., Statistical Thermodynamics of Mixtures with Zero Energies of Mixing. *Proc. R. Soc. London, Ser. A*, 1944. **183**(993): p. 203-212.

[43] Guggenheim, E. A., *Mixtures: The Theory of the Equilibrium Properties of Some Simple Classes of Mixtures, Solutions and Alloys*. 1952, Oxford: Clarendon Press.

[44] Tsuboka, T., Katayama, T., Modified Wilson Equation for Vapor-Liquid and Liquid-Liquid Equilibria. *J. Chem. Eng. Jpn.*, 1975. **8**(3): p. 181-187.

[45] Huang, J.-F., Lee, L.-S., A Modification of the Wilson Equation With Only Binary Parameters for Ternary Liquid-Liquid Equilibria. *Fluid Phase Equilib.*, 1994. **101**: p. 27-51.

[46] Renon, H., Prausnitz, J. M., Local Composition in Thermodynamic Excess Functions for Liquid Mixtures. *AIChE J.*, 1968. **14**(1): p. 135-144.

[47] Vetere, A., A Simple Modification of the NRTL Equation. *Fluid Phase Equilib.*, 2000. **173**(1): p. 57-64.

[48] Vetere, A., The NRTL Equation as a Predictive Tool for Vapor-Liquid Equilibria. *Fluid Phase Equilib.*, 2004. **218**(1): p. 33-39.

[49] Abrams, D. S., Prausnitz, J. M., Statistical Thermodynamics of Liquid Mixtures: A New Expression for the Excess Gibbs Energy of Partly or Completely Miscible Systems. *AIChE J.*, 1975. **21**(1): p. 116-128.

[50] Maurer, G., Prausnitz, J. M., On the Derivation and Extension of the UNIQUAC Equation. *Fluid Phase Equilib.*, 1978. **2**: p. 91-99.

[51] McDermott, C., Ashton, N., Note on the definition of local composition. *Fluid Phase Equilib.*, 1977. **1**(1): p. 33-35.

[52] Deal, C. H., Derr, E. L., Group Contributions in Mixtures. *Ind. Eng. Chem.*, 1968. **60**(4): p. 28-38.

[53] Kojima, K., Tochigi, K., *Prediction of Vapor-Liquid Equilibria by the ASOG Method*. 1979, Tokyo: Kodansha Ltd.

[54] Bekker, A. Y., Knox, D. E., Sund, S. E., Prediction of Solvent Activities Using the UNIFAC-FV Model. *J. Solution Chem.*, 1986. **16**(8): p. 635-639.

[55] Chen, F., Holten-Andersen, J., Tyle, H., New Developments of the UNIFAC Model for Environmental Application. *Chemosphere*, 1993. **26**(7): p. 1325-1354.

[56] Jakob, A., Grensemann, H., Lohmann, J., Gmehling, J., Further Development of Modified UNIFAC (Dortmund): Revision and Extension 5. *Ind. Eng. Chem. Res.*, 2006. **45**(23): p. 7924-7933.

[57] Magnussen, T., Rasmussen, P., Fredenslund, A., UNIFAC Parameter Table for Prediction of Liquid-Liquid Equilibria. *Ind. Eng. Chem. Proc. Des. Dev.*, 1981. **20**(2): p. 331-339.

[58] Bondi, A., van der Waals Volumes and Radii. *J. Phys. Chem. A*, 1964. **68**(3): p. 441-451.

[59] Beilstein, F. K., CrossFire Beilstein, Frankfurt (2008) http://info.crossfiredatabases.com/

[60] Hildebrand, J. H., Scott, R. L., *Regular Solutions*. 1962, New Jersey: Prentice-Hall.

[61] Barton, A. F. M., *CRC Handbook of Solubility Parameters and Other Cohesion Parameters*. 2nd ed. 1991, London: CRC Press.

[62] Hansen, C. M., *Hansen Solubility Parameters: A User's Handbook*. 2nd ed. 2007: CRC Press.

[63] Macedo, E. A., Solubility of Amino Acids, Sugars, and Proteins. *Pure Appl. Chem.*, 2005. **77**(3): p. 559-568.

[64] Thomas, E. R., Eckert, C. A., Prediction of Limiting Activity Coefficients by a Modified Seperation of Cohesive energy Density Model and UNIFAC. *Ind. Eng. Chem. Proc. Des. Dev.*, 1984. **23**(2): p. 194-209.

[65] Lazzaroni, M. J., Bush, D., Eckert, C. A., Frank, T. C., Gupta, S. K., Olsen, J. D., Revision of MOSCED Parameters and Extension to Solid Solubility Calculations. *Ind. Eng. Chem. Res.*, 2005. **44**(11): p. 4075-4083.

[66] Chen, C.-C., Crafts, P. A., Correlation and Prediction of Drug Molecule Solubility in Mixed Solvent Systems with the Nonrandom Two-Liquid Segment Activity Coefficient (NRTL-SAC) Model. *Ind. Eng. Chem. Res.*, 2006. **45**: p. 4816-4824.

[67] Chen, C.-C., Song, Y., Solubility Modelling with a Nonrandom Two-Liquid Segment Activity Coefficient Model. *Ind. Eng. Chem. Res.*, 2004. **43**: p. 8354-8362.

[68] Lei, Z., Chen, B., Li, C., Lui, H., Predictive Molecular Thermodynamic Models for Liquid Solvents, Solid Salts, Polymers and Ionic Liquids. *Chem. Rev.*, 2008. **108**: p. 1419-1455.

[69] Mullins, E., Oldland, R., Liu, Y. A., Wang, S., Sandler, S. I., Chen, C.-C., Zwolak, M., Seavey, K. C., Sigma-Profile Database for Using COSMO-Based Thermodynamic Methods. *Ind. Eng. Chem. Res.*, 2006. **45**(12): p. 4389-4415.

[70] Klamt, A., Comments on "A Priori Phase Equilibrium Prediction from a Segment Contribution Model". *Ind. Eng. Chem. Res.*, 2002. **41**(9): p. 2330-2331.

[71] Lin, S.-T., Sandler, S. I., Reply to Comments on "A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model". *Ind. Eng. Chem. Res.*, 2002. **41**(9): p. 2332-2334.

[72] Grensemann, H., Gmehling, J., Performance of a Conductor-Like Screening Model for Real Solvents Model in Comparison to Classical Group Contribution Methods. *Ind. Eng. Chem. Res.*, 2005. **44**: p. 1610-1624.

[73] Vera, J. H., Sayegh, S. G., Ratcliff, G. A., A Quasi Lattice Local Composition Model For The Excess Gibbs Free Energy of Liquid Mixtures. *Fluid Phase Equilib.*, 1977. **1**: p. 113-135.

[74] Kikic, I., Alessi, P., Rasmussen, P., Fredenslund, A., On the Combinatorial Part of the UNIFAC and UNIQUAC Models. *Can. J. Chem. Eng.*, 1980. **58**: p. 253-258.

[75] Kontogeorgis, G. M., Fredenslund, A., Tassios, D. P., Simple Activity Coefficient Model for the Prediction of Solvent Activities in Polymer Solutions. *Ind. Eng. Chem. Res.*, 1993. **32**: p. 362-372.

[76] Bondi, A., Free Volumes and Free Rotation in Simple Liquids and Liquid Saturated Hydrocarbons. *J. Phys. Chem.*, 1954. **58**(11): p. 929-939.

[77] Pappa, G. D., Kontogeorgis, G. M., Tassios, D. P., Prediction of Ternary Liquid-Liquid Equilibria in Polymer-Solvent-Solvent Systems. *Ind. Eng. Chem. Res.*, 1997. **36**: p. 5461-5466.

[78] Wibawa, G., Takishima, S., Sato, Y., Masuoka, H., An Improved Prediction Result of Entropic-FV Model for Vapor-Liquid Equilibria of Solvent-Polymer Systems. *J. Appl. Polym. Sci.*, 2005. **97**: p. 1145-1153.

[79] Hynčica, P., Hnědkovský, L., Cibulka, I., Partial Molar Volumes of Organic Solutes in Water. X. Benzene and Toluene at Temperatures From (298 to 573) K and at Pressures up to 30 MPa. *J. Chem. Thermodyn.*, 2003. **35**: p. 1905-1915.

[80] Cordes, W., Rarey, J., A New Method for the Estimation of the Normal Boiling Point of Non-Electrolyte Organic Compounds. *Fluid Phase Equilib.*, 2002. **201**: p. 409-433.

[81] Choi, P. B., Mclaughlin, E., Effect of a Phase Transition on the Solubility of a Solid. *AIChE J.*, 1983. **29**(1): p. 150-153.

[82] Mishra, D. S., Yalkowsky, S. H., Ideal Solubility of a Solid Solute: Effect of Heat Capacity Assumptions. *Pharm. Res.*, 1992. **9**(7): p. 958-959.

[83] Yalkowsky, S. H., Solubility and Partitioning V: Dependence of Solubility on Melting Point. *J. Pharm. Sci.*, 1981. **70**(8): p. 971-973.

[84] Neau, S. H., Bhandarkar, S. V., Hellmuth, E. W., Differential Molar Heat Capacities to Test Ideal Solubility Estimations. *Pharm. Res.*, 1997. **14**(5): p. 601-605.

[85] Hildebrand, J. H., The Temperature dependence of the Solubility of Solid Nonelectrolytes. *J. Chem. Phys.*, 1952. **20**(1): p. 190-191.

[86] Nordstrom, F. L., Rasmussen, A. C., Prediction of Solubility Curves and melting Properties of Organic and Phatmaceutical Compounds. *Eur. J. Pharm. Sci.*, 2009. **26**: p. 330-344.

[87] Gracin, S., Brinck, T., Rasmussen, A. C., Prediction of Solubility of Solid Organic Compounds in Solvents by UNIFAC. *Ind. Eng. Chem. Res.*, 2002. **41**(20): p. 5114-5124.

[88] Nti-Gyabaah, J., Chmielowski, R., Chan, V., Chiew, Y. C., Solubility of Lovastatin in a Family of Six Alcohols: Ethanol, 1-Propanol, 1-Butanol, 1-Pentanol, 1-Hexanol and 1-Octanol. *Int. J. Pharm.*, 2008. **359**: p. 111-117.

[89] Domalski, E. S., Hearing, E. D., Estimation of the Thermodynamic Properties of C-H-N-O-S-Halogen Compounds at 298.15 K. *J. Phys. Chem. Ref. Data*, 1993. **22**(4): p. 805-1159.

[90] Renner, R., The Kow Controversy. *Environ. Sci. Technol.*, 2002. **36**(21): p. 411A-413A.

[91] Lyman, W. L., Reehl, W. F., Rosenblatt, D. H., *Handbook of Chemical Property Estimation Methods Environmental Behavior of Organic Compounds*. 1982, New York: McGraw-Hill.

[92] Tse, G., Sandler, S. I., Determination of Infinite Dilution Activity Coefficients and 1-Octanol/Water Partition Coefficients of Volatile Organic Pollutants. *J. Chem. Eng. Data*, 1994. **39**: p. 354-357.

[93] Reza, J., Trejo, A., Temperature Dependence of the Infinite Dilution Activity Coefficient and Henry's Law Constant of Polycyclic Aromatic Hydrocarbons in Water. *Chemosphere*, 2004. **56**: p. 537-547.

[94] Harris, R. A., *Monoethanolamine: Suitability as an Extractive Solvent*, in *Chemical Engineering*. 2000, University of KwaZulu Natal: Durban.

[95] Eckert, C. A., Newman, B. A., Nicolaides, G. L., Long, T. C., Measurement and Application of Limiting Activity Coefficients. *AIChE J.*, 1981. **27**(1): p. 33-40.

[96] Krummen, M., Gruber, D., Gmehling, J., Measurement of Activity Coefficients at Infinite Dilution in Solvent Mixtures Using the Dilutor Technique. *Ind. Eng. Chem. Res.*, 2000. **39**(6): p. 2114-2123.

[97] Sandler, S. I., Infinite Dilution Activity Coefficients in Chemical, Environmental and Biochemical Engineering. *Fluid Phase Equilib.*, 1996. **116**: p. 343-353.

[98] Gwala, N. V., Deenadayalu, N., Tumba, K., Ramjugernath, D., Activity Coefficients at Infinite Dilution for Solutes in the Trioctylmethylammonium Bis(trifluoromethylsulfonyl)imide Ionic Liquid Using Gas Liquid Chromatography. *J. Chem. Thermodyn.*, 2009.

[99] Olivier, E., Letcher, T. M., Naidoo, P., Ramjugernath, D., Activity Coefficients at Infinite Dilution of Organic Solutes in the Ionic Liquid 1-Ethyl-3-Methylimidazolium Trifluoromethanesulfonate Using Gas–Liquid Chromatography at T = (313.15, 323.15, and 333.15) K *J. Chem. Thermodyn.*, 2009.

[100] Stephenson, R. M., Mutual Solubility of Water and Aliphatic Amines. *Journal of Chemical & Engineering Data*, 1993. **38**(4): p. 625-629.

[101] Abildskov, J., Gani, R., Rasmussen, P., O'Connel, J. P., Analysis of Infinite Dilution Activity Coefficients of Solutes in Hydrocarbons from UNIFAC. *Fluid Phase Equilib.*, 2001. **181**: p. 163-186.

[102] Voutsas, E. C., Tassios, D. P., Analysis of the UNIFAC-Type Group-Contribution Models at the Highly Dilute Region. 1. Limitations of the Combinatorial and Residual Expressions. *Ind. Eng. Chem. Res.*, 1997. **36**(11): p. 4965-4972.

[103] Johansson, E., Bolton, K., Theodorou, D. N., Ahlström, P., Monte Carlo Simulations of Equilibrium Solubilities and Structure of Water in n-Alkanes and Polyethylene. *J. Chem. Phys.*, 2007. **126**(22): p. 224902.

[104] Benzi, C., Cossi, M., Improta, R., Barone, V., Building Cavities in a Fluid of Spherical or Rod-Like Particles: A Contribution to the Solvation Free Energy in Isotropic and Anisotropic Polarizable Continuum Model. *J. Comput. Chem.*, 2005. **26**(11).

[105] Höfinger, S., Zerbetto, F., The Costly Process of Creating a Cavity in n-Octanol. *Theor. Chem. Acc.*, 2004. **112**: p. 240-246.

[106] Prevost, M., Oliveira, I. T., Kocher, J.-P., Wodak, S. J., Free Energy of Cavity Formation in Liquid Water and Hexane. *J. Phys. Chem.*, 1996. **100**(7): p. 2738-2743.

[107] Pierotti, R. A., A Scaled Particle Theory of Aqueous and Nonaqueous Solutions. *Chem. Rev.*, 1976. **76**(6): p. 717-726.

[108] Uhlig, H. H., The Solubilities of Gases and Surface Tension. *J. Phys. Chem.*, 1937. **41**(9): p. 1215-1225.

[109] Lange, N. A., Dean, J. A., *Lange's Handbook of Chemistry*. 14th ed. 1992, New York: McGraw-Hill.

[110] Tomasi, J., Persico, M., Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.*, 1994. **94**(7): p. 2027-2094.

[111] Ihmels, E. C., Gmehling, J., Extension and Revision of the Group Contribution Method GCVOL for the Prediction of Pure Compound Liquid Densities. *Ind. Eng. Chem. Res.*, 2003. **42**(2): p. 408-412.

[112] Rackett, H. G., Equation of State for Saturated Liquids. *J. Chem. Eng. Data*, 1970. **15**(4): p. 514-517.

[113] Tsibanogiannis, I. N., Kalospiros, N. S., Tassios, D. P., Extension of the GCVOL Method and Application to Some Complex Compounds. *Ind. Eng. Chem. Res.*, 1994. **33**(6): p. 1641-1643.

[114] Marrero, J., Abildskov, J., *Part 1 Organic Solutes Ranging from C_4 to C_{40}*. DECHEMA Chemistry Data Series: Solubility and Related Properties of Large Complex Chemicals. Vol. XV. 2003.

[115] Lundager Madsen, H. E., Boistelle, R., Solubility of Octacosane and Hexatriacontane in Different n-Alkane Solvents. *J. Chem. Soc., Faraday Trans.*, 1979. **75**: p. 1254-1258.

[116] Haulait-Pirson, M.-C., Huys, G., Vanstraelen, E., New Predictive Equation for the Solubility of Solid n-Alkanes in Organic Solvents. *Ind. Eng. Chem. Res.*, 1987. **26**: p. 447-452.

[117] Kontogeorgis, G. M., Voutsas, E. C., Tassios, D. P., A Molecular Simulation-Based Method for the Estimation of Activity Coefficients for Alkane Solutions. *Chem. Eng. Sci.*, 1996. **51**(12): p. 3247-3255.

[118] Jackson, P. L., Wilsak, R. A., Thermodynamic Consistency Tests Based on the Gibbs-Duhem Equation Applied to Isothermal, Binary Vapor-Liquid Equilibrium Data: Data Evaluation and Model Testing. *Fluid Phase Equilib.*, 1995. **103**: p. 155-197.

[119] Regosz, A., Pelplinska, T., Kowalski, P., Thiel, Z., Prediction of Solubility of Sulfonamides in Water and Organic Solvents Based on the Extended Regular Solution Theory. *Int. J. Pharm.*, 1992. **88**: p. 437-442.

[120] Jouyban, A., Review of the Cosolvency Models for Predicting Solubility of Drugs in Water-Cosolvents Mixtures. *J. Pharm. Pharmaceut. Sci.*, 2008. **11**(1): p. 32-58.

[121] Millard, J. W., Alvarez-Núnez, F. A., Yalkowsky, S. H., Solubilization by Cosolvents Establishing Useful Constants for the Log-Linear Model. *Int. J. Pharm.*, 2002. **245**: p. 153-166.

[122] Gantiva, M., Yurquina, A., Martínez, F., Solution Thermodynamics of Ketoprofen in Ethanol + Water Cosolvent Mixtures. *Journal of Chemical & Engineering Data*, 2009.

[123] Shayanfar, A., Acree, W. E., Jouyban, A., Solubility of Lamotrigine, Diazepam, Clonazepam and Phenobarbital in Propylene Glycol + Water Mixtures at 298.15 K. *J. Chem. Eng. Data*, 2009. **54**(3): p. 1153-1157.

[124] Estrada, E., Diaz, G. A., Delgado, E. J., Predicting Infinite Dilution Activity Coefficients of Organic Compounds in Water by Quantum-Connectivity Descriptors. *J. Comput.-Aided Mol. Des.*, 2006. **20**: p. 539-548.

[125] Giralt, F., Espinosa, G., Arenas, A., Ferre-Gine, J., Girones, X., Amat, L., Carbo-Dorca, R., Cohen, Y., Estimation of Infinite Dilution Activity Coefficients of Organic Compounds in Water with Neural Classifiers. *AIChE J.*, 2004. **50**(6): p. 1315-1343.

[126] Votano, J. R., Parham, M., Hall, L. H., Kier, L. B., Hall, L. M., Prediction of Aqueous Solubility Based on Large Datasets Using Sevel QSPR Models Utilizing Topological Structure Representation. *Chem. Biodivers.*, 2004. **1**: p. 1829 - 1841.

[127] Yan, A., Gasteiger, J., Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.*, 2003. **43**(2): p. 429-434.

[128] Hughes, L. D., Palmer, D. S., Nigsch, F., Mitchell, J. B. O., Why Are Some Properties More Difficult to Predict than Others? A Study of QSPR Models of Solubility, Melting Point and LogP. *J. Chem. Inf. Model.*, 2008. **48**(1): p. 220-232.

[129] Jain, A., Yang, G., Yalkowsky, S. H., Estimation of Melting Points of Organic Compounds. *Ind. Eng. Chem. Res.*, 2004. **43**(23): p. 7618-7621.

[130] Boethling, R. S., Mackay, D., *Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences*. 2000: CRC Press.

[131] Chemspider, ChemZoo (2009) www.chemspider.com

[132] Jorgensen, W. L., Duffy, E. M., Prediction of Drug Solubility from Structure. *Adv. Drug Delivery Rev.*, 2002. **54**: p. 355-366.

[133] Llinas, A. L., Glen, R. C., Goodman, J. M., Solubility Challenge: Can you Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements?, *J. Chem. Inf. Model.*, 2008. **48**(7): p. 1289-1303.

[134] Stein, S. E., Brown, R. L., Estimation of Normal Boiling Points from Group Contributions. *J. Chem. Inf. Comput. Sci.*, 1994. **34**(3): p. 581-587.

[135] Constantinou, L., Gani, R., New Group Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.*, 1994. **40**(10): p. 1697-1710.

[136] Marrero-Morejón, J., Pardillo-Fontdevilla, E., Estimation of Pure Component Properties Using Group-Interaction Contributions. *AIChE J.*, 1999. **45**(3): p. 615-621.

[137] Ericksen, D., Wilding, W. V., Oscarson, J. L., Rowley, R. L., Use of the DIPPR Database for Development of QSPR Correlations: Normal Boiling Point. *Journal of Chemical & Engineering Data*, 2002. **47**(5): p. 1293-1302.

[138] Moelwyn-Hughes, E. A., *The Chemical Statics and Kinetics of Solutions* 1971, London: Academic Press.

[139] Poling, B. E., Prausnitz, J. M., O'Connel, J. P., *The Properties of Gases and Liquids*. 5th ed. 2001, Singapore: McGraw-Hill.

[140] Hovorka, S., Dohnal, V., Roux, A. H., Roux-Desgranges, G., Determination of Temperature Dependence of Limiting Activity Coefficients for a Group of Moderately Hydrophobic Organic Solutes in Water. *Fluid Phase Equilib.*, 2002(201): p. 135-164.

[141] Hovorka, S., Roux, A. H., Roux-Desgranges, G., Dohnal, V., Limiting Partial Molar Excess Enthalpies of Selected Organic Compounds in Water at 298.15 K. *J. Chem. Eng. Data*, 2002. **47**: p. 954-959.

[142] Marrero, J., Abildskov, J., *Part 2 Organic Solutes Ranging from C_2 to C_{41}*. DECHEMA Chemistry Data Series: Solubility and Related Properties of Large Complex Chemicals. Vol. XV. 2005.

[143] Miller, D. J., Hawthorne, S. B., Method for Determining the Solubilities of Hydrophobic Organics in Subcritical Water. *Anal. Chem.*, 1998. **70**(8): p. 1618-1621.

[144] Banerjee, S., Calculation of Water Solubility of Organic Compounds with UNIFAC-Derived Parameters. *Environ. Sci. Technol.*, 1985. **19**: p. 369-370.

[145] Apelblat, A., Manzurola, E., Solubility of Oxalic, Malonic, Succinic, Adipic, Maleic, Malic, Citric, and Tartaric Acids in Water from 278.15 to 338.15 K. *J. Chem. Thermodyn.*, 1987. **19**: p. 317-320.

[146] Apelblat, A., Manzurola, E., Solubility of Ascorbic, 2-Furancarboxylic, Glutaric, Pimelic, Salicylic, and o-Phthalic Acids in Water from 279.15 to 242.15 K, and Apparent Molar Volumes of Ascorbic, Glutaric, and Pimelic Acids in water at 298.15 K. *J. Chem. Thermodyn.*, 1989. **21**: p. 1005-1008.

[147] Apelblat, A., Manzurola, E., Solubility of Suberic, Azelaic, Levulinic, Glycolic, and Diglycolic Acids in Water from 278.25 to 361.35 K. *J. Chem. Thermodyn.*, 1990. **22**: p. 289-292.

[148] Zhang, H. L., Bai, T. C., Yan, G. B., Hu, J., Solubility of Silybin in Aqueous Poly(vinylpyrrolidone) Solution. *Fluid Phase Equilib.*, 2005. **238**: p. 186-192.

[149] Wishart, D., Knox, C., Guo, A. C., Cheng, D., Shirvastava, S., Tzur, D., Gautam, B., Hassanali, M., DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.*, 2008. **36**: p. 901-906.

[150] Mackle, H., O'Hare, P. A. G., Studies in the Thermochemistry of Sulphones. Part 6. - Heats of Combustion, Fusion, Vaporization and Atomization of Six Aromatic and Two Allylic Sulphones. *Trans. Faraday Soc.*, 1961. **57**: p. 1521-1526.

[151] Domanska, U., Pobudkowska, A., Pelczarska, A., Gierycz, P., pKa and Solubility of Drugs in Water, Ethanol, and 1-Octanol. *J. Phys. Chem. B*, 2009.

[152] Rytting, E., Lentz, K. A., Chen, X.-Q., Qian, F., Venkatesh, S., Aqueous and Cosolvent Solubility Data for Drug-like Organic Compounds. *The AAPS Journal*, 2005. **7**(1): p. E78-E105.

[153] Garner-O'Neale, L. D., Bonamy, A. F., Meek, T. L., Patrick, B. G., Calculating Group Electronegativities Using the Revised Lewis-Langmuir Equation. *J. Mol. Struct.*, 2003. **639**: p. 151-156.

[154] Fiorini, D., Fiselier, K., Biedermann, M., Ballini, R., Coni, E., Grob, K., Contamination of Grape Seed Oil with Mineral Oil Paraffins. *J. Agric. Food. Chem.*, 2008. **56**(23): p. 11245-11250.

[155] Friedrich, J. P., List, G. R., Characterization of Soybean Oil Extracted by Supercritical Carbon Dioxide and Hexane. *J. Agric. Food. Chem.*, 1982. **30**(1): p. 192-193.

[156] Gonzlez, C., Resa, J. M., Ruiz, A., Gutirrez, J. I., Densities of Mixtures Containing n-Alkanes with Sunflower Seed Oil at Different Temperatures. *J. Chem. Eng. Data*, 1996. **41**(4): p. 796-798.

[157] Knowles, R. E., Taylor, K. W., Kohler, G. O., Goldblatt, L. A., Industrial Oils from Seeds, Hydroxy-Unsaturated Oils and Meal from Dimorphotheca and Lesuerella Seed *J. Agric. Food. Chem.*, 1964. **12**(5): p. 390-392.

[158] Proctor, A., Bowen, D. J., Ambient-Temperature Extraction of Rice Bran Oil with Hexane and Isopropanol. *J. Am. Oil Chem. Soc.*, 1996. **73**(6): p. 811-813.

[159] Perlovich, G. L., Strakhova, N. N., Kazachenko, V. P., Volkova, T. V., Tkachev, V. V., Schaper, K.-J., Raevsky, O. A., Studying Thermodynamic Aspects of Sublimation, Solubility and Solvation Processes and Crystal Structure Analysis of Some Sulfonamides. *Int. J. Pharm.*, 2007. **334**: p. 115-124.

[160] Kurkov, S. V., Perlovich, G. L., Thermodynamic Studies of Fenbufen, Diflunisal and Flurbiprofen: Sublimation, Solution and Solvation of Biphenyl Substituted Drugs. *Int. J. Pharm.*, 2008. **357**: p. 100-107.

[161] Perlovich, G. L., Volkova, T. V., Bauer-Brandl, A., Thermodynamic Study of Sublimation, Solubility, Solvation and Distribution Precosses of Atanolol and Pindolol. *Mol. Pharm.*, 2007. **4**(6): p. 929-935.

[162] Perlovich, G. L., Volkova, T. V., Bauer-Brandl, A., Towards an Understanding of the Molecular Mechanism of Solvation of Drug Molecules: A Thermodynamic Approach by Crystal Lattice Energy, Sublimation, and Solubility Exemplified by Paracetamol, Acetanilide and Phenacetin. *J. Pharm. Sci.*, 2006. **95**(10): p. 2158-2169.

[163] Surov, A. O., Szterner, P., Zielenkiewicz, W., Perlovich, G. L., Thermodynamic and Structural Study of Tolfenamic Acid Polymorphs. *J. Pharm. Biomed. Anal.*, 2009. **50**: p. 831-840.

[164] Perlovich, G. L., Surov, A. O., Bauer-Brandl, A., Thermodynamic Properties of Flufenamic and Niflumic Acids - Specific and Non-Specific Interactions in Solution and In Crystal Lattices, Mechanism of Solvation, Partitioning and Distribution. *J. Pharm. Biomed. Anal.*, 2007. **45**: p. 679-687.

[165] Görgényi, M., Dewulf, J., Van Langenhove, H., Temperature Dependence of Henry's Law Constant in an Extended Temperature Range. *Chemosphere*, 2002. **48**: p. 757-762.

[166] Beneš, M., Dohnal, V., Limiting Activity Coefficients of Some Aromatic and Aliphatic Nitro Compounds in Water. *Journal of Chemical & Engineering Data*, 1999. **44**(5): p. 1097-1102.

[167] Curren, M. S. S., King, J. W., Solubility of Triazine Pesticides in Pure and Modified Subcritical Water. *Anal. Chem.*, 2001. **73**(4): p. 740-745.

[168] Li, A., Andren, A. W., Solubility of Polychlorinated Biphenyls in Water/Alcohol Mixtures. 1. Experimental Data. *Environ. Sci. Technol.*, 1994. **28**(1): p. 47-52.

[169] Meylan, W. M., Howard, P. H., Atom/Fragment Contribution Method for Estimating Octanol-Water Partition Coefficients. *J. Pharm. Sci.*, 1994. **84**(1): p. 83-92.

[170] Klopman, G., Li, J.-Y., Wang, S., Dimayuga, M., Computer Automated log P Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.*, 1994. **34**(4): p. 752-781.

[171] Burkhard, L. P., Armstrong, D. E., Andren, A. W., Henry's Law Constants for the Polychlorinated Biphenyls. *Environ. Sci. Technol.*, 1985. **19**(7): p. 590-596.

[172] Kuramochi, H., Noritomi, H., Hoshino, D., Kato, S., Nagahama, K., Application of UNIFAC Models to Partition Coefficients of Biochemicals between Water and n-Octanol or n-Butanol. *Fluid Phase Equilib.*, 1998. **144**: p. 87-95.

[173] Kan, A. T., Tomson, M. B., UNIFAC Prediction of Aqueous and Nonaqueous Solubilities of Chemicals with Environmental Interest. *Environ. Sci. Technol.*, 1996. **30**(4): p. 1369-1376.

[174] Gonzalez, H. E., Abildskov, J., Gani, R., A Method for Prediction of UNIFAC Group Interaction Parameters. *AIChE J.*, 2007. **53**(6): p. 1620-1632.

[175] Marrero, J., Gani, R., Group-Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilib.*, 2001. **183**: p. 183-208.

[176] Alvarez Julia, J., Barrero, C. E., Corso, M. E., Grande, M. C., Marschoff, C. M., On the Applicability of the UNIQUAC Method to Ternary Liquid-Liquid Equilibria. *J. Argent. Chem. Soc.*, 2004. **92**(4): p. 81-90.

[177] Marcilla, A., Gomis, V., Esteban, A., Letter to the Editor. *AIChE J.*, 1995. **41**(4): p. 1044.

[178] Kannan, D. C., Duda, J. L., Danner, R. P., A Free-Volume Term Based on the van der Waals Partition Function for the UNIFAC Model. *Fluid Phase Equilib.*, 2005. **228-229**: p. 321-328.

[179] Edward, J. T., Molecular Volumes and the Stokes-Einstein Equation. *J. Chem. Educ.*, 1970. **47**(4): p. 261-270.

[180] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., *Numerical recipes in FORTRAN The Art of Scientific Computing*. Second ed. 1992, New York: Cambridge Press.

[181] Riley, K. F., Hobson, M. P., Bence, S. J., *Mathematical Methods for Physics and Engineering*. Second ed. 2003, Cape Town: Cambridge Press.

[182] Dallos, A., Hajós-Szikszay, É., Horváth, A., Liszi, J., Barczyńska, J., Bald, A., Enthalpies of Solution and Crystallization of Fumaric Acid in Aqueous Solution. *J. Chem. Thermodyn.*, 2000. **32**: p. 587-595.

[183] Dang, L., Du, W., Black, S., Wei, H., Solubility of Fumaric Acid in Propan-2-ol, Ethanol, Acetone, Propan-1-ol and Water. *Journal of Chemical & Engineering Data*, 2009.

[184] Watson, K. M., Thermodynamics of the Liquid State - Generalized Prediction of Properties. *Ind. Eng. Chem.*, 1943. **35**(4): p. 398-406.

[185] O'Shea, S. J., Stokes, R. H., Activity Coefficients and Excess Partial Molar Enthalpies for (Ethanol + Hexane) from 283 to 318 K. *J. Chem. Thermodyn.*, 1986. **18**: p. 691-696.

[186] Stokes, R. H., Burfitt, C., Enthalpies of Dilution and Transfer of Ethanol in Non-Polar Solvents. *J. Chem. Thermodyn.*, 1973. **5**: p. 623-631.

APPENDICES

A. EXTENDING THE UNIFAC MATRIX

As mentioned much in the sections above it is generally assumed [55, 57, 171, 172] (for a counter example consider the work of Kan and Tomson [173] where very good results were found using UNIFAC) that the UNIFAC and mod. UNIFAC predictions are not suitably accurate for large complex multifunctional compounds for the following reasons:

1. The majority of UNIFAC type model parameters are fitted using only VLE data and therefore these parameters are somehow not suited to calculations involving LLE and SLE.

2. The assumption of group additivity fails for large multifunctional compounds.

3. Very frequently calculations involving multifunctional compounds require knowledge of unavailable binary interaction parameters.

Point number one, while being useful for practical applications, isn't very reasonable from a theoretical point of view as an activity coefficient shouldn't change depending on the solutes standard state. This discrepancy is a result of one of two things, either an incorrect residual expression or the result of the incorrect combinatorial expressions in the UNIFAC models (see section 6.1 and Appendix B). Point 2 is valid for some functional groups (e.g. COOH and OH groups) but for a large number of groups the pure component methods developed here (Chapters 7 and 8) suggest that group interactions (which account for the non-additivity of groups) may not always be so vital for large molecules. The source of this error may also be (at least partially) down to the incorrect combinatorial expression. The final point is a valid one since some interactions can be very difficult to get data for.

In order to fill these gaps with plausible numbers there would need to be large amounts of measurements on fairly exotic systems which in some cases could even be reactive (e.g. COOH and NH_2). This is both time consuming and very expensive and in some cases even impossible. It is for this reason that a method is needed that enables the prediction of the missing parameters either from the existing parameters or from some prediction technique. The need for interaction parameters sometimes results in parameters being predicted from COSMO-RS results. Even though these results are often unreliable [14] the need for the interaction parameters is so great that even this unreliability is accepted.

Another approach is the one taken by Gani [174] which is to fit "connectivity indices" to the binary interaction parameters. In essence the binary interaction parameters are treated as pure component properties and the model parameters are fitted to them. The idea of connectivity indices has been used quite extensively by Gani [7, 174, 175], with moderate success [175], to predict pure component properties. The problem, however, with using the similar procedure with the UNIFAC parameters is that it is assumed that the parameters have

some physical significance. This assumption is, in many cases, not true and therefore any predicted interaction parameters could be drastically affected. Interaction parameters are generally strongly intercorrelated [176, 177] and are therefore only physically significant when considering them simultaneously.

While the interaction parameters may not be physically realistic, the hypothetical infinite dilution activity coefficients of one group in another in many cases should be. The idea was therefore to combine the physical realism of the infinite dilution values with the "molecular surface interpolation" type idea of NRTL-SAC (see section 3.2.3.3). The surface segments for the groups were assigned by considering hypothetical situations of an infinite dilution of one group in another (so for example the infinite dilution activity coefficient of CH_2 group in a pure solution of OH groups). The obvious advantage of this is that if the binary interactions for three or four (depending on whether a hydrogen bonding group is assigned or not) groups are known then all the other interactions could be interpolated.

What a procedure such as this could facilitate is the introduction of more detailed groups into the UNIFAC matrix. The current UNIFAC main groups are very broadly defined, usually encompassing many subgroups. However the pure component estimation methods developed by this group [15, 16, 18-20, 80] have shown that there is probably a need for more detailed groups in some applications.

Table A-1 shows some of the UNIFAC group-in-group γ^{∞} data to which the surface segments were fitted to. The r and q values were taken as the same as the first subgroup that appears in the main group (i.e. the one with the lowest subgroup number).

Table A-1 Matrix of the infinite dilution activity coefficients of one mod. UNIFAC main group in another at 298.15 K (Grp ID – main group number, Grp – main group name).

	Grp ID	1	5	9	11	14
Grp ID	Grp	CH_3	CH_2=CH	ACH	$ACCH_3$	OH (P)
1	CH_3	-	1.39	1.22	1.76	1158.81
5	CH_2=CH	1.15	-	0.76	1.14	116.71
9	ACH	1.48	0.65	-	0.82	56.83
11	$ACCH_3$	1.58	1.51	0.87	-	73.45
14	OH (P)	28.58	63.23	2.76	-	-

As a proof of concept the surface segments were found by using the same parameters (i.e. τ values) as the original NRTL-SAC [67]. The objective function (OF) used in the regression was:

$$OF = \sum_{i=1}^{m} \left| \frac{\gamma_i^{\infty,UNIFAC} - \gamma_i^{\infty,NRTL-SAC}}{m \times \gamma_i^{\infty,UNIFAC}} \right| \qquad \text{(A-1)}$$

where *m* is the total number of UNIFAC group-in-group infinite dilution activity coefficients. There were no constraints on the surface segments except that CH_3 was assumed to only have an X segment and water was assumed to only have a Z segment (in accordance with

the NRTL-SAC model development). Table A-2 shows the regressed surface segment values for some of the UNIFAC groups. Even though there were no constraints placed on the groups they seem to be quite physically realistic. The hydrocarbon groups have predominantly a hydrophobic (X) surface while for example the OH group has mainly a hydrophilic (Z) surface which is what one would expect. The physical realism of the groups is found for most cases.

Table A-2 Regressed surface segments for some of the UNIFAC groups.

		Surface segments			
Grp num	**Grp**	**X**	**Y-**	**Y+**	**Z**
1	CH_3	0.336	-	-	-
5	CH_2=CH	0.326	0.025	0.166	-
9	ACH	0.166	0.015	0.141	-
11	$ACCH_3$	0.300	0.019	0.058	-
14	OH (P)	0.001	0.063	0.000	1.071

The objective function for the first 27 main groups in the UNIFAC matrix was found to be 0.39. This moderate error and the good physical realism observed by the groups seem to indicate that this method is based on sound principles. The problems encountered with this rough and ready approach is that some of the more complex groups cannot be satisfactorily reproduced. The reason for this could be that there are some effects which are not captured by the NRTL-SAC groups or model parameters. This problem therefore requires some in-depth investigation which is beyond the scope of this work.

B. BEHAVIOUR OF THE SG COMBINATORIAL EXPRESSION

When considering the discussions in section 6.1 it is clear that the Staverman-Guggenheim (SG) combinatorial expression (and the mod. UNIFAC expression) greatly under predicts the infinite dilution extrapolation behaviour of large solutes in small solvents. It was also shown in section 9.5 that by simply replacing the combinatorial expression in UNIFAC and mod. UNIFAC the predictions for infinite dilution activity coefficients in hexane can be greatly improved. Since the SG expression corrects the typical under-prediction of larger solutes in hexane one would expect this expression to greatly improve the predictions in water. Water is going to be smaller than any solute used and therefore it seems intuitive that the SG and mod. UNIFAC expressions should under-predict the combinatorial expression quite drastically. This does not happen; the reasons for this are discussed in the paragraphs following.

The SG combinatorial contribution is given by Eqn. (3-52), for the binary case at infinite dilution of the solute(2) the following expression results:

$$\ln \gamma_2^{C,\infty} = 1 - \frac{r_2}{r_1} + \ln \frac{r_2}{r_1} - 5q_2 \left(1 - \frac{r_2 q_1}{r_1 q_2} - \ln \frac{r_2 q_1}{r_1 q_2} \right) \tag{B-1}$$

Where all the variables used are defined in section 3.2.1. Some trivial rearrangement yields the following:

$$\ln \gamma_2^{C,\infty} = 1 - \left(\frac{r_2}{r_1} \right) + \ln \left(\frac{r_2}{r_1} \right) - 5r_1 \left(\frac{r_2}{r_1} \right) \left(\frac{q_2}{r_2} \right) \left(1 - \left(\frac{r_2}{q_2} \right) \left(\frac{q_1}{r_1} \right) - \ln \left[\left(\frac{r_2}{q_2} \right) \left(\frac{q_1}{r_1} \right) \right] \right) \tag{B-2}$$

Therefore the combinatorial contribution to the activity coefficient in a fixed solvent (i.e. r_1 & q_1 fixed) for any solute will depend on the 2 variables: r_2 and r_2/q_2. The two surfaces for hexane and water are shown in Fig. B-1 and Fig. B-2 respectively. If one considers that the average r_2/q_2 value for the 13801 compounds with r and q values in the DDB was 1.25 then the combinatorial contribution decreases with increasing size in hexane (as observed earlier) but it actually increases with increasing size in water. What is also clear from these two plots is that as the size of the solute gets larger the more the combinatorial is sensitive to the shape (q_2) of the solute with the combinatorial expression changing 60 natural log units when the ratio r_2/q_2 doubles at an r_2 of 50.

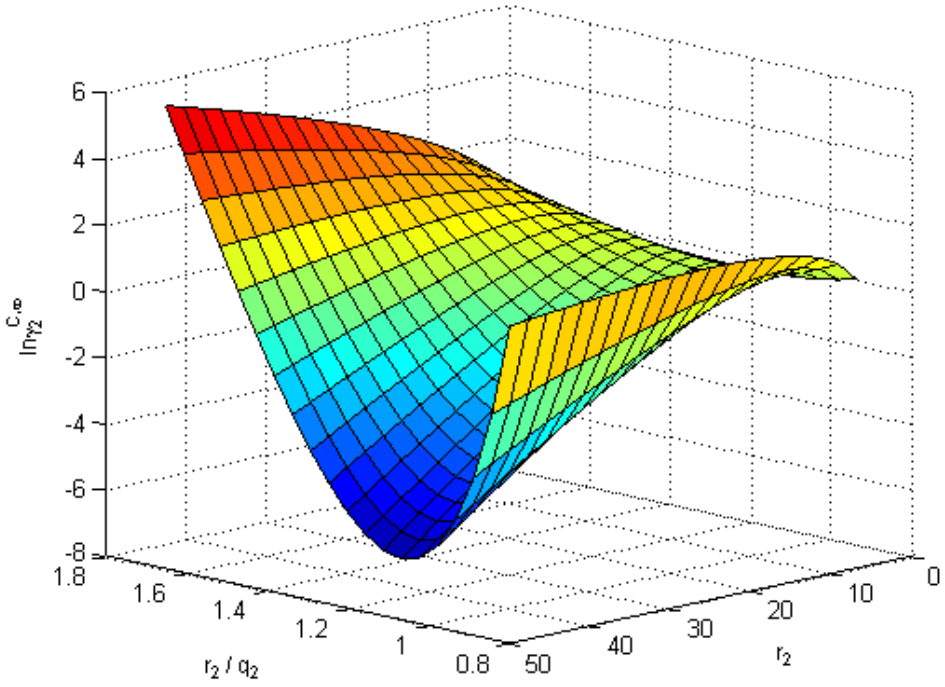


Fig. B-1 $\ln\gamma_2^{C,\infty}$ vs. r_2/q_2 and r_2 for an arbitrary solute (2) in hexane (1) using the SG combinatorial expression (r_1 = 4.4998, q_1 = 3.8560).

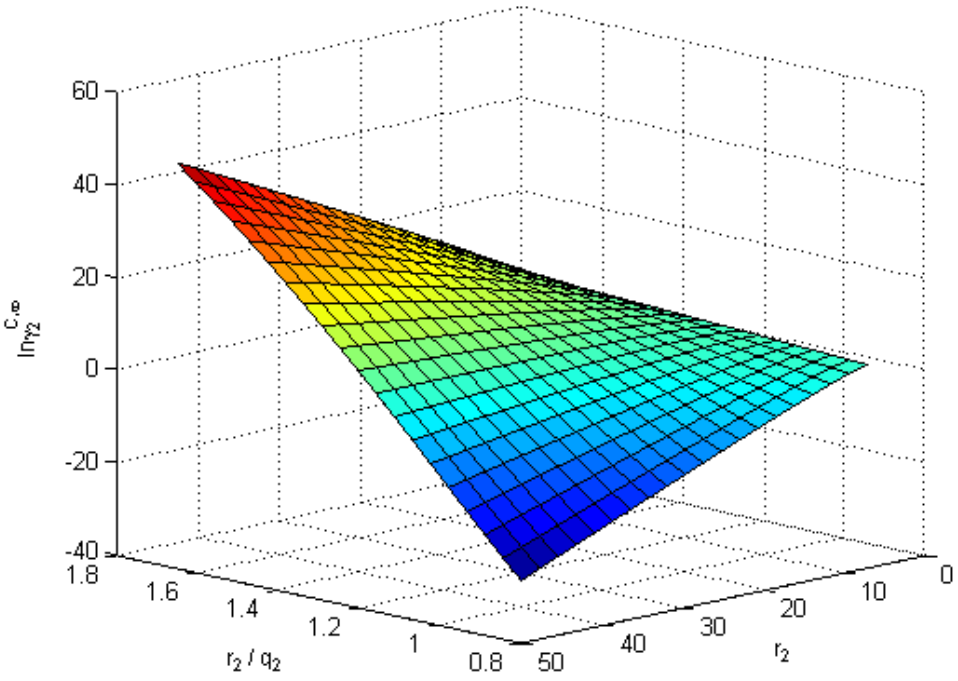


Fig. B-2 $\ln\gamma_2^{C,\infty}$ vs. r_2/q_2 and r_2 for an arbitrary solute (2) in water (1) using the SG combinatorial expression (r_1 = 0.92, q_1 = 1.40).

The reason for this vastly different behaviour in hexane and water is due to r_1/q_1 for water. Whereas most solvents have a ratio of around 1.3 water has a ratio of 0.66 (see Table B-1 for values for some common solvents). The r & q values for water presented in the original UNIQUAC paper [49] were in fact regressed [178]; they are therefore not representative of the van der Waals volume and surface areas. When using the van der Waals volume and surface areas the r and q values for water are 0.8154 and 0.904 respectively [178, 179]. As shown in Fig. B-3 the combinatorial contribution produced with the realistic values is more inline with the trend produced in other solvents (i.e. decreasing combinatorial contribution with an increasing solute size).

When the fitted values are used in the SG expression (and mod. UNIFAC combinatorial expression) they are not giving a correct combinatorial value since they operate outside of the "safe" operating zone of the equation (evident by the quantitative and qualitative differences in Fig. B-1 and Fig. B-2). While one would expect UNIFAC and mod. UNIFAC to fail spectacularly for water (bearing in mind the findings in section 6.1) yet they don't. This is also the reason why when using these values in the "new combinatorial" presented in this work produces huge overestimations. Nevertheless even though, strictly speaking, the SG expression doesn't produce a combinatorial value in water (and therefore by implication the residual contribution is not strictly speaking correct) the results for UNIFAC and mod. UNIFAC in water are still acceptable.

A similar kind of problem is observed in mod. UNIFAC whereby the group r and q values were treated as adjustable parameters and therefore the combinatorial is, strictly speaking, not actually giving good combinatorial values (this is evident in the large difference in Fig. B-1 and Fig. B-4, where Fig. B-4 uses the fitted values of r and q, while these values produce the desired behaviour it is not a combinatorial). Some instances have been observed where there are some ridiculously large values of the combinatorial contribution. Nevertheless the success (relative to the other methods tested) of mod. UNIFAC for the complex compounds considered in this work shows that while in some instances the combinatorial and residual values will not be correct the overall activity coefficient still is.

Table B-1 r_1/q_1 values for some common solvents.

Solvent	r_1/q_1
1-Butanol	1.132
2-Butanone	1.129
Acetic acid	1.063
Acetone	1.102
Acetonitrile	1.085
Benzene	1.328
Chloroform	1.191
Cyclohexane	1.249
Ethanol	1.068
Ethyl acetate	1.116
Hexane	1.167
Methanol	0.999
Toluene	1.322
Water (UNIQUAC)	0.657
Water (exp)	0.902

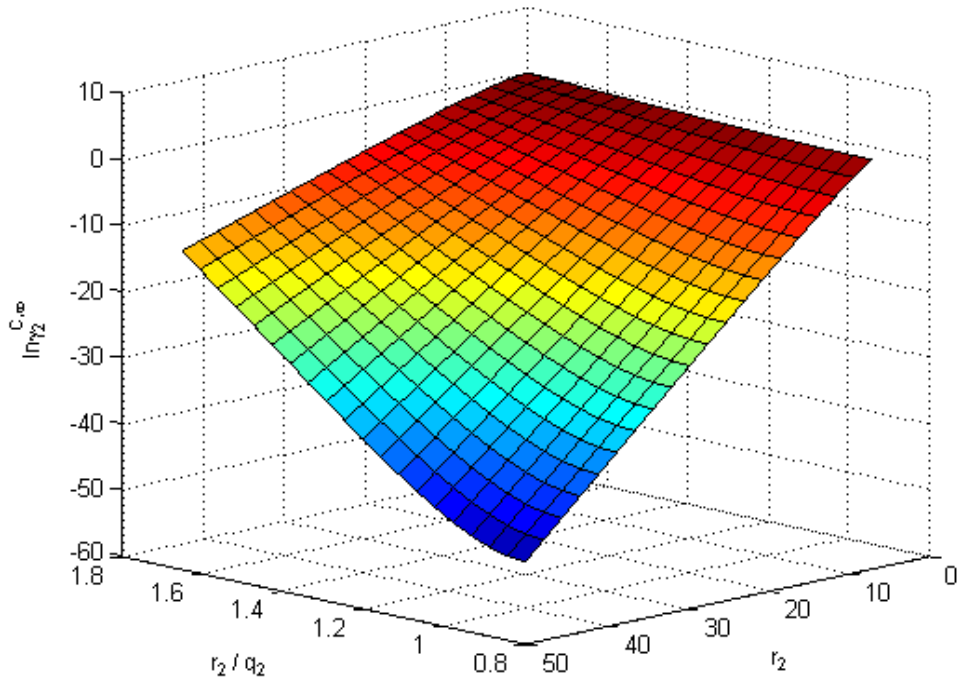


Fig. B-3 $\ln\gamma_2^{C,\infty}$ vs. r_2/q_2 and r_2 for an arbitrary solute (2) in water (1) with the realistic r and q values using the SG combinatorial expression (r_1 = 0.8154, q_1 = 0.904).

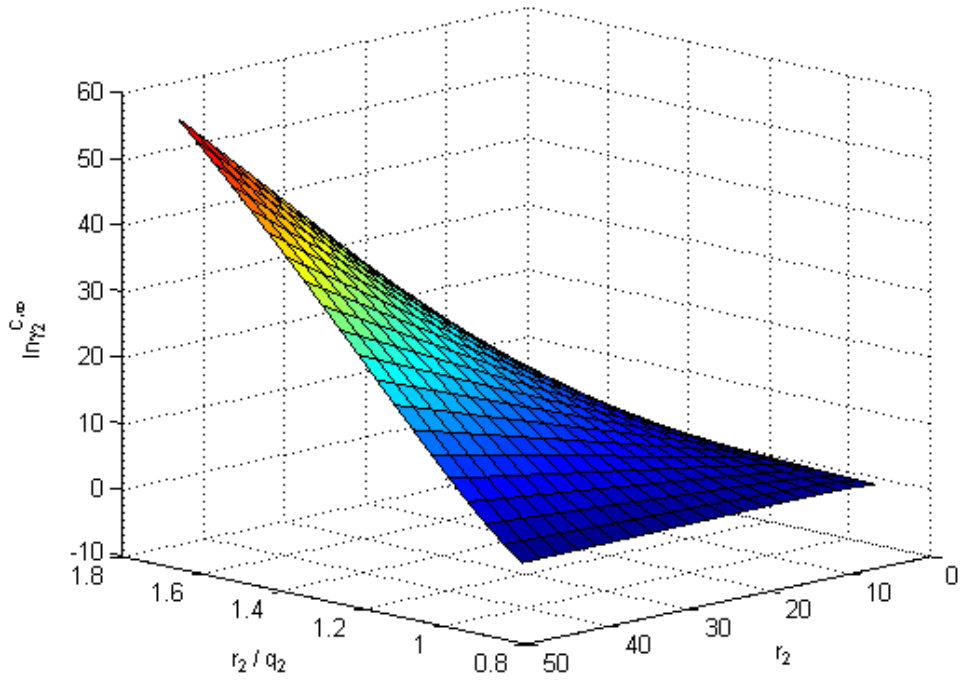


Fig. B-4 $\ln\gamma_2^{C,\infty}$ vs. r_2/q_2 and r_2 for an arbitrary solute (2) in hexane (1) using the mod. UNIFAC combinatorial expression with the mod. UNIFAC r&q values (r_1 = 3.795, q_1 = 4.945).

C. REGRESSION

Since this work is concerned with model development there is naturally quite a large amount of regression which is required. The simplest case is the linear least squared algorithm which is only applicable to linear systems of equations. Since the group contributions and interactions are all linear functions of the group frequency (section 7.5) this was the method that was applied. The algorithm (not the code) was taken from Press et al. [180] and is outlined in the section following. This algorithm was found to be sufficient for the previous work [20] but for this work it very often failed. Whenever the linear problem is underspecified the linear least squared algorithm fails because the alpha matrix becomes singular (or numerically close).

A group contribution problem is underspecified if two (or more) groups are only contained in one compound (this definition of underspecified will be used from here onwards). When this happens there are only sufficient data to fix the sum of the values and not the individual values. This is very tedious to find "by hand" since this involves looking at the parameter matrix column by column. Fortunately, however, there is an algorithm for this exact problem; it is called singular value decomposition (SVD). The problem with using the SVD routine blindly is that it will give values for all the groups but if the problem is underspecified the values will be nonsense. The SVD algorithm outputs values which can be used to analyse the suitability of the solutions. One solution is to therefore only use SVD when needed (i.e. when the linear least squared routine fails) and to make sure that any unrealistically large values are removed. As with the linear least squared algorithm the SVD algorithm (and for SVD, the code) was taken from Press et al. [180] and Riley et al. [181], the algorithm is outlined in the sections following.

C.1. Linear Least Squared Regression

The general form of an equation which is linear in its parameters (a_k) is:

$$y(x) = \sum_{k=1}^{M} a_k X_k(x) \tag{C-1}$$

As the name suggests this routine finds the minimum of the square of the difference between the experimental (not necessarily experimental data but in this work it was) and the model values. The objective function (which is the function which needs to be minimized) is therefore defined as follows:

$$F = \sum_{i=1}^{N} \left[y_i - \sum_{k=1}^{M} a_k X_k(x_i) \right]^2 \tag{C-2}$$

From simple calculus it is known that the minimum of a function is found where the derivative disappears, applying this to Eqn. (C-2):

$$\frac{\partial F}{\partial a_k} = -2\sum_{i=1}^{N} X_k(x_i)\left[y_i - \sum_{j=1}^{M} a_j X_j(x_i) \right] = 0 \tag{C-3}$$

In order to simplify the notation in Eqn. (C-3) the following substitutions are made:

$$A_{ij} = X_j(x_i) \tag{C-4}$$

$$b_i = y_i \tag{C-5}$$

$$\boldsymbol{\alpha} = \mathbf{A}^T \cdot \mathbf{A} \tag{C-6}$$

$$\boldsymbol{\beta} = \mathbf{A}^T \cdot \mathbf{b} \tag{C-7}$$

Using these simplifications and slightly manipulating Eqn. (C-3) this results in:

$$\boldsymbol{\alpha} \cdot \mathbf{a} = \boldsymbol{\beta} \tag{C-8}$$

Therefore the parameter matrix **a** can be found as follows:

$$\mathbf{a} = \boldsymbol{\alpha}^{-1}\boldsymbol{\beta} \tag{C-9}$$

The only thing which poses a problem, so-far as computing time is concerned, is the computation of the inverse of the alpha matrix. This can be fairly simply achieved by using the Gauss-Jordan elimination.

C.2. Singular Value Decomposition (SVD)

As mentioned above, in some cases the linear least squared algorithm fails. This typically occurs when the problem is underspecified or the coefficient matrix is badly scaled and therefore the alpha matrix becomes singular (or in the case of a badly scaled **A**, numerically singular). Any mxn (m is number of simultaneous equations and n is number of unknowns and where) matrix **A** can be broken down by the SVD theorem as follows (for the case where $m \geq n$ which is applicable to this work):

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^T \tag{C-10}$$

where:

- **U** is an mxn orthogonal matrix (i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{I}$)

- **W** is an nxn diagonal matrix (i.e. $W_{ij} = 0$ when $i \neq j$). The diagonal elements are denoted by w_i for $i = 1,2,...,n$ and are termed the singular values of **A**.

- **V** is a nxn orthogonal matrix (i.e. $\mathbf{V}^T\mathbf{V} = \mathbf{I}$)

The problem is therefore how to determine the elements of the 3 matrices **U**,**W** and **V** in terms of the elements of the given matrix A. Two square matrices $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ can be constructed from Eqn. (C-10) as follows:

$$\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{W}^T\mathbf{U}^T\mathbf{U}\mathbf{W}\mathbf{V}^T = \mathbf{V}\mathbf{W}^T\mathbf{W}\mathbf{V}^T \tag{C-11}$$

$$\mathbf{A}\mathbf{A}^T = \mathbf{U}\mathbf{W}\mathbf{V}^T\mathbf{V}\mathbf{W}^T\mathbf{U}^T = \mathbf{U}\mathbf{W}\mathbf{W}^T\mathbf{U}^T \tag{C-12}$$

where $\mathbf{W}^T\mathbf{W}$ and $\mathbf{W}\mathbf{W}^T$ are both equal and are diagonal matrices with elements w_i^2 for $i = 1,2,...,n$. Therefore (recalling that for orthogonal matrices $\mathbf{V}^{-1} = \mathbf{V}^T$):

$$\mathbf{W}^T\mathbf{W} = \mathbf{V}^T\left(\mathbf{A}^T\mathbf{A}\right)\mathbf{V} \tag{C-13}$$

$$\mathbf{W}\mathbf{W}^T = \mathbf{U}^T\left(\mathbf{A}\mathbf{A}^T\right)\mathbf{U} \tag{C-14}$$

From Eqn. (C-13) and (C-14) it can be shown that the columns of **V** must be the normalised eigenvectors v^i of the matrix $\mathbf{A}^T\mathbf{A}$ and similarly, the columns of U must be the normalised eigenvectors u^i of the matrix $\mathbf{A}\mathbf{A}^T$ (see Riley et al. Ch. 8.16 [181] for the proof of this). The singular values (w_i) must also satisfy $w_i = \sqrt{\lambda_i}$ where λ_i are the eigenvalues of $\mathbf{A}^T\mathbf{A}$. Keeping with the notation used in C.1, the general form of a linear equation is:

$$\mathbf{b} = \mathbf{A} \cdot \mathbf{a} \tag{C-15}$$

where

$$\mathbf{a} = \mathbf{A}^{-1}\mathbf{b} \tag{C-16}$$

The simplest case is for a square matrix **A**, Eqn. (C-15) can then simply be solved by Gauss-Jordan elimination or some other matrix inversion technique. However when the problem is over specified (more equations than unknowns) then all normal matrix inversion techniques fail to solve the inverse of **A**. This is where SVD becomes so powerful, since the SVD of **A** is given by Eqn. (C-10) the inverse will be:

$$\mathbf{A}^{-1} = \mathbf{V} \cdot \mathbf{W}^{-1} \cdot \mathbf{U}^T \tag{C-17}$$

where $\mathbf{W}^{-1}$ is a diagonal matrix with elements $1/w_i$. This is very often called the pseudoinverse. Analysis of the singular values enables one to determine which group values

are erroneous; typically a very small value produces a excessively large (absolute) value of the group and is generally easy to spot. For this work, compounds containing such groups were either removed or the literature was searched for compounds containing the same groups.

D. UNIQUAC REFERENCE SURFACE

As discussed in section 3.2.2 the r and q values used in the UNIQUAC equation are van der Waals volume and surface area relative to a unit of polymethylene (basically an n-alkane). The choice of a standard segment is fairly arbitrary [49] if it considered to be a sphere then:

$$V_{w_s} = \frac{4}{3}\pi R_{w_s}^3 \tag{D-1}$$

$$A_{w_s} = 4\pi R_{w_s}^2 \tag{D-2}$$

where V_{w_s} and A_{w_s} are the volume and surface area of the standard segment and R_{w_s} is the radius. The van der Waals volume and surface area of an *n*-mer of polymethylene (C_nH_{2n+2}) when calculated from Bondi's method [58] are (N_A is Avogadro's number):

$$V_{w_i} = \frac{nV_{CH_2}}{N_A} = \frac{n \times 10.23}{6.023 \times 10^{23}} cm^3 \text{ / } molecule \tag{D-3}$$

$$A_{w_i} = \frac{nA_{CH_2}}{N_A} = \frac{n \times 1.35 \times 10^9}{6.023 \times 10^{23}} cm^2 \text{ / } molecule \tag{D-4}$$

For straight or branched open chain molecules Guggenheim [42] showed that the following holds:

$$\left(\frac{z}{2}\right)(r-q) = r-1 \tag{D-5}$$

where z is the coordination number which is the number of neighbours each molecule has ($6 \le z \le 12$). In general form Eqn's (3-60) and (3-61) can be written as:

$$r_i = \frac{V_{w_i}}{V_{w_s}} \tag{D-6}$$

$$q_i = \frac{A_{w_i}}{A_{w_s}} \tag{D-7}$$

Combining Eqn's (D-1) to (D-7) yields the following:

$$\frac{4\pi N_A}{n} R_{w_s}^3 - \frac{zA_{CH_2}}{2} R_{w_s} + 3V_{CH_2}\left(\frac{z}{2}-1\right) = 0 \tag{D-8}$$

Taking the limit of Eqn. (D-8) as the length of the polymethylene molecule tends to infinity ($n \to \infty$) the cubic term vanishes yielding the following equation:

$$R_{w_s} = \frac{3V_{CH_2}(z-2)}{zA_{CH_2}} \tag{D-9}$$

Eqn. (D-9) shows that the radius (and hence the volume and surface area) of the standard segment depends on the coordination number and the choice of the segment molecule. Examining the UNIQUAC equations (Eqn. (3-50) and (3-52)) it is clear that the choice of the surface segment is not arbitrary but will change the value of both the residual and combinatorial contributions. It is likely that any discrepancy would be absorbed by the model parameters [49]. If the coordination number is changed for an athermal mixture of hexane(1) and decane(2) the change is slight but not negligible as shown in Fig. D-1.

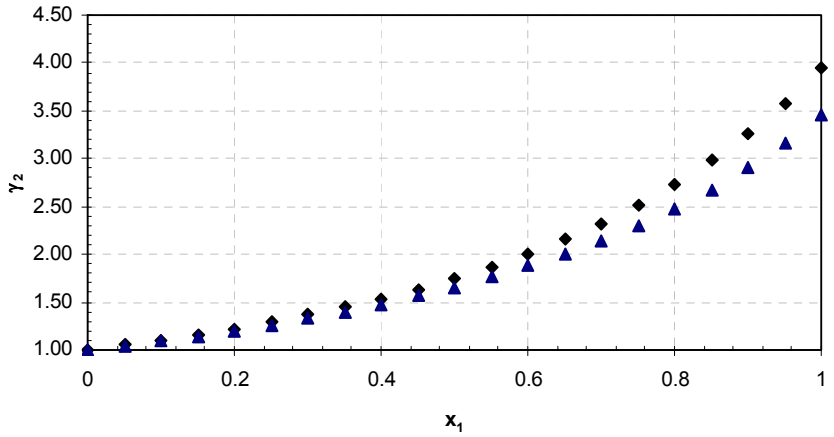


Fig. D-1 γ_2 vs. x_1 for a mixture of hexane(1) and decane(2) using the UNIQUAC combinatorial expression (◆ - z = 10, ▲ – z = 8).

E. IDAC TEMPERATURE DEPENDENCE

As shown above the temperature dependence of the activity coefficient at infinite dilution in water is given by Eqn. (7-11). If it is assumed that $-\overline{H}_i^{E,\infty} / R$ is a linear function of temperature (Eqn. (E-1)) and this in substituted into Eqn. (7-11) and integrated from a reference temperature of 298.15 K to some arbitrary temperature T then Eqn. (E-2) results.

$$\frac{-\overline{H}_i^{E,\infty}}{R} = b + cT \tag{E-1}$$

$$\ln\gamma_2^{\infty} - \ln\gamma_{2,298.15K}^{\infty} = -\frac{b}{T} + c\ln T + \frac{b}{298.15} - c\ln(298.15) \tag{E-2}$$

This equation can be easily rearranged to yield the following:

$$\ln\gamma_2^{\infty} = \ln\gamma_{2,298.15K}^{\infty} + \left(\frac{-b}{298.15}\right)\left(\frac{1-\tau}{\tau}\right) + c\ln\tau \tag{E-3}$$

$$\tau = \frac{T}{298.15} \tag{E-4}$$

Similarly for the infinite dilution activity coefficient in alkanes the general temperature dependence was given by Eqn. (8-2). If it is assumed that $-\overline{H}_i^{E,\infty} / R$ is constant with respect to temperature (Eqn. (E-5)) and this is substituted into Eqn. (7-11) and integrated as above then Eqn. (E-6) results:

$$\frac{-\overline{H}_i^{E,\infty}}{R} = b \tag{E-5}$$

$$\ln\gamma_2^{\infty} - \ln\gamma_{2,298.15K}^{\infty} = -\frac{b}{T} + \frac{b}{298.15} \tag{E-6}$$

Rearrangement yields the following (τ is given by Eqn. (E-4)):

$$\ln\gamma_2^{\infty} = \ln\gamma_{2,298.15K}^{\infty} + \left(\frac{-b}{298.15}\right)\left(\frac{1-\tau}{\tau}\right) \tag{E-7}$$

F. ANOMALOUS BEHAVIOUR OF WATER

F.1. General Temperature Dependence

As discussed in section 7.4 there were quite a few exceptions noted, among the non-polar molecules, for the general temperature dependence that was proposed. As shown in that section the infinite dilution activity coefficient of hydrophilic solutes had the opposite temperature dependence when compared to hydrophobic solutes. The first case that was considered was that the data for these systems could be erroneous. This is however not the case since these data are from various different sources which use different methods to measure the infinite dilution activity coefficient.

Therefore, since the temperature dependence is given by the partial molar excess heat of mixing at infinite dilution, the H^E data were analysed to see why there is such a difference in temperature dependence. A negative value of $\overline{H}_2^{E,\infty}$ (in keeping with the notation used above the first component is the solvent and the second the solute) corresponds to a negative slope at a zero solute concentration (see the left hand side of Fig. F-1). When examining the H^E data in the DDB [28] it quickly became apparent that for almost all systems involving water at low temperatures this trend (a negative slope at $x_2 = 0$) is observed (For systems not involving water this effect is rather less common and a more typical example is shown in Fig. F-2). However what became apparent when analysing the data is that for systems which have a miscibility gap (which is noticeable by a flat top of the curve – see Fig. F-6) the trend is the opposite (a negative slope at $x_2 = 0$) regardless of the temperature. This is, at least qualitatively, the behaviour predicted by the non-polar general temperature dependence equation (Eqn. (7-16)).

What is also clear from the data is that at higher temperatures the shape of the H^E curves for the water-solute system are similar to those we find in other more simple solvents and for systems which have a miscibility gap. A possible reason for these 2 different trends is that at low temperatures the structure of water is highly ordered due to the presence of a large amount of hydrogen bonding. This ordered structure can be disrupted in 2 ways, firstly if the temperature is raised and secondly if a sufficiently hydrophobic molecule is introduced into the water. This would explain why compounds like hexane, benzene and almost all of the compounds tested in section 7.4 (at most temperatures these compounds were sparingly soluble in water, i.e. $x < 0.01$) followed the general trend sufficiently well. The criteria necessary for the compound to be 'sufficiently hydrophobic' seems to be immiscibility.

One example H^E data for a low solubility solid, fumaric acid, is shown in Fig. F-3. The $\overline{H}_2^{E,\infty}$ value calculated from these data is 32 kJ/mol at 298.15 K, this is almost 5 times higher than the value of 7.4 kJ/mol predicted from the general temperature dependence. When examining the activity coefficient data (Fig. F-4) it is clear that this makes quite a large difference to the

data below 298.15 K. It is clear that the slope of the predicted curve needs to be significantly larger at this point which is what one would expect from the $\overline{H}_2^{E,\infty}$ value. Nevertheless the data above 298.15 K is still very well represented. What this indicates is that even if this equation cannot reproduce $\overline{H}_2^{E,\infty}$ data with good accuracy at all temperatures, it does still have good practical application.

In the case of nitromethane (2) in water (1), there is a large miscibility gap (as shown in Fig. F-5) and the excess heat of mixing curves follow the expected (positive) trend, as shown in Fig. F-6. However, more complex immiscibility examples (with an upper and a lower critical solution temperature) such as tetrahydrofuran (2) and water (1) (as shown in Fig. F-7) do not hold (see Fig. F-8), however such types of LLE behaviour are fairly rare and this system shows quite a narrow miscibility gap.

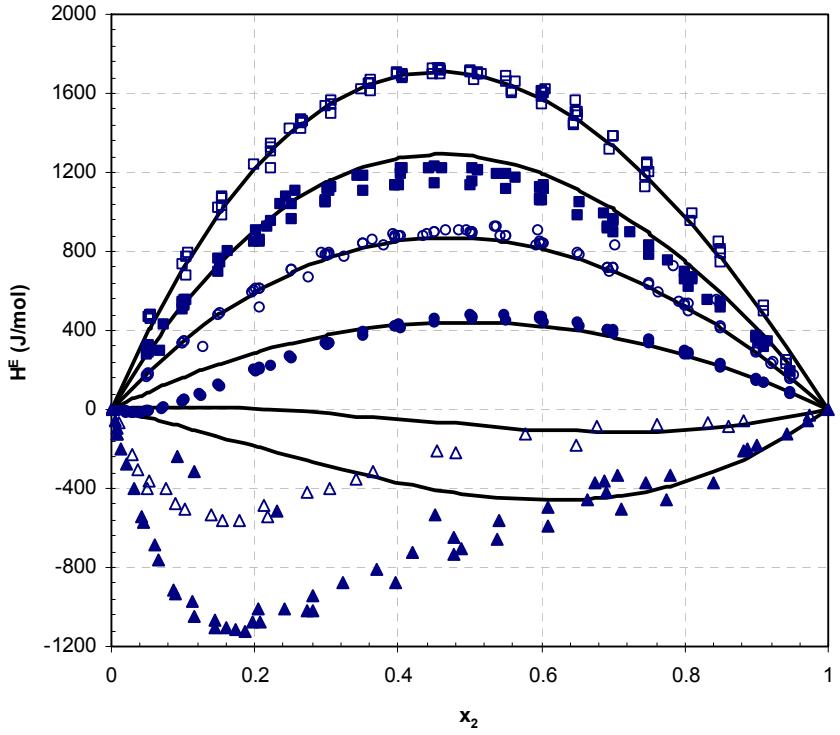


Fig. F-1 H^E vs. x_2 at various temperatures for the system water(1) – ethanol(2) all data from the DDB [28] (□ - 523.15 K, ■ - 473.15 K, ○ - 423.15 K, ● - 373.15 K, △ - 313.15 K, ▲ - 273.15, —— Eqn. (F-1) fitted to the high temperature data).

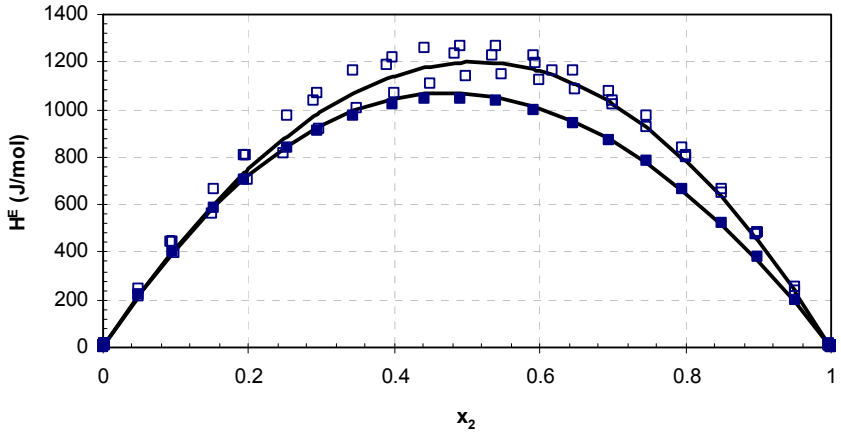


Fig. F-2 H^E vs. x_2 at various temperatures for the system acetone(1) – ethanol(2) all data from the DDB [28] (□ - 323.15 K, ■ - 283.15 K, —— Eqn. (F-1) fitted to the data).

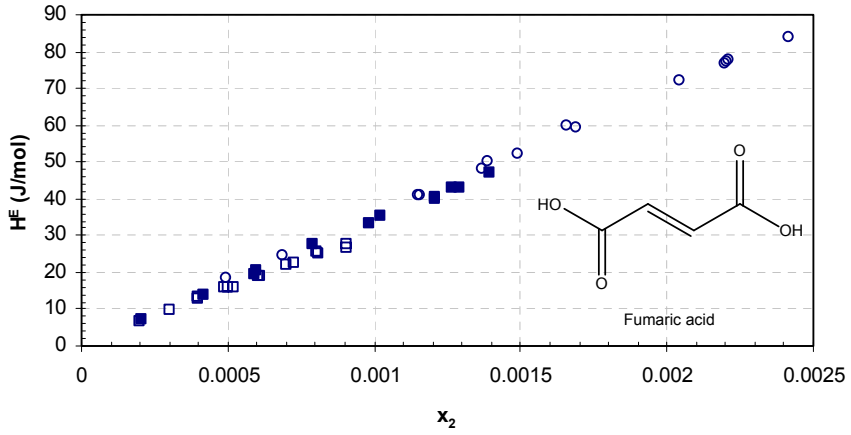


Fig. F-3 H^E vs. x_2 at various temperatures for the system water(1) – fumaric acid (2) all data from literature [182] (□ - 298.15 K, ■ - 313.15 K, ○ - 333.15 K).

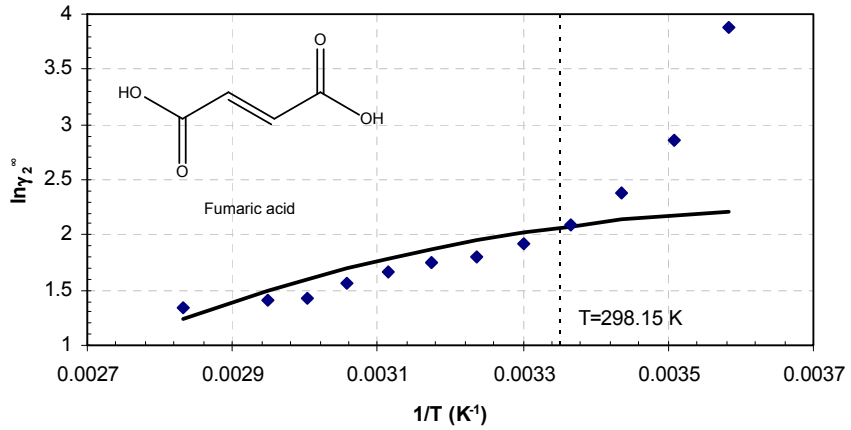


Fig. F-4 $\ln\gamma_2^\infty$ vs. 1/T for fumaric acid (2) in water (1) (◆ - data extracted from SLE data from literature [183], —— Eqn. (7-16) with the 298 K reference experimental data point).

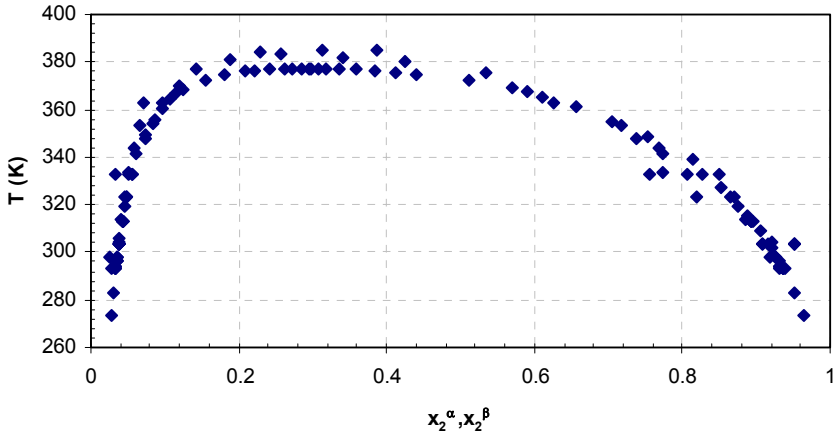


Fig. F-5 T vs. x_2 in the 2 phases α and β for the system water (1) – nitromethane (2) (data from the DDB [28]).

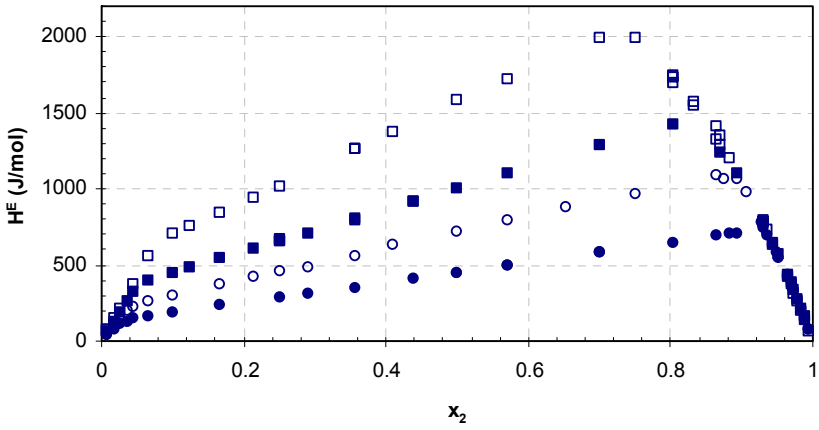


Fig. F-6 H^E vs. x_2 at various temperatures for the system water(1) – nitromethane(2) all data from the DDB [28] (□ - 353.15.15 K, ■ - 333.15 K, ○ - 313.15 K, ● - 293.15 K).

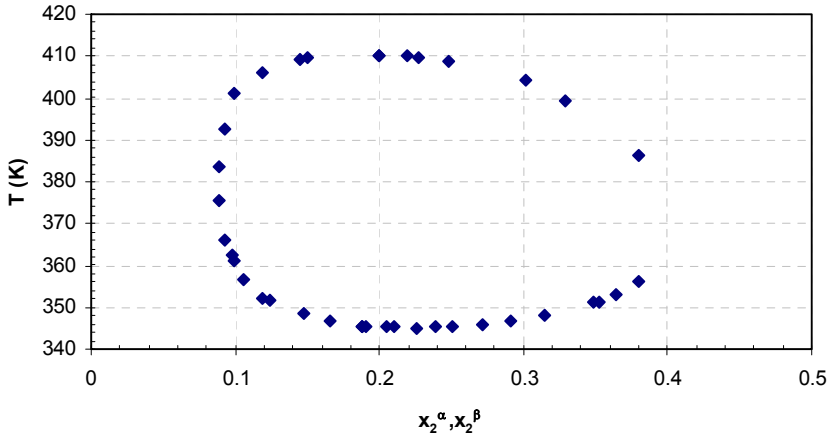


Fig. F-7 T vs. x_2 in the 2 phases α and β for the system water (1) – tetrahydrofuran (2) (data from the DDB [28]).

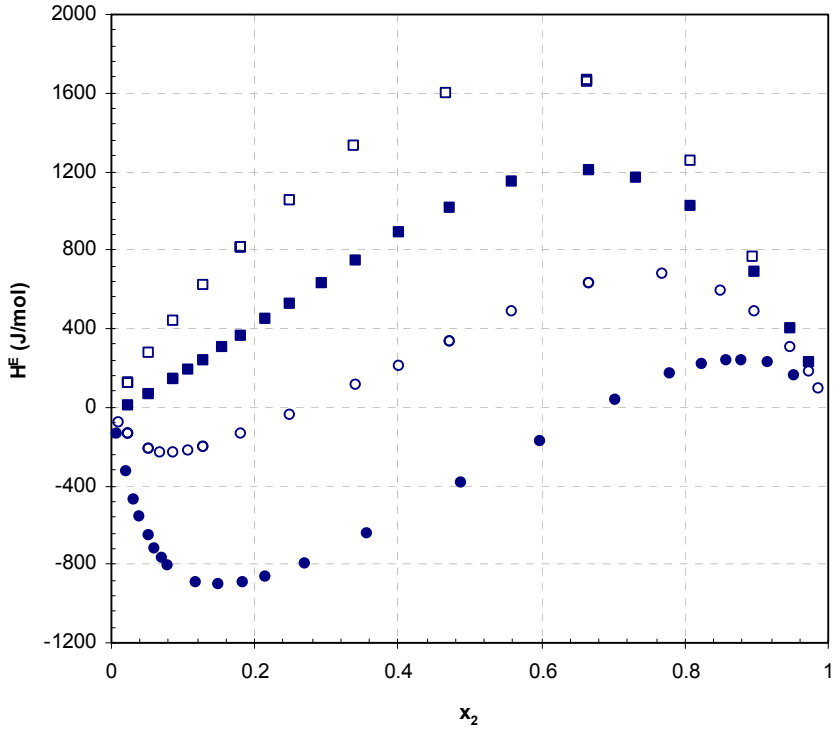


Fig. F-8 H^E vs. x_2 at various temperatures for the system water(1) – tetrahydrofuran (2) all data from the DDB [28] (□ - 416.29 K, ■ - 383.15 K, ○ - 343.15 K, ● - 283.15 K).

F.2. Correlation with a Pure Component Property

As discussed above, water is fairly unique in that it is one of the few solvents which cause negative $\overline{H}_i^{E,\infty}$ for the solute in water (this is not a property which is unique to water, an alternate example is the system methanol – 1,2-butandiol, however it is certainly the solvent for which this effect is the most prevalent). This deviation is due to the structure and behaviour of lower temperature water. Intuitively, if this deviation is solely due to the behaviour of water then this behaviour should be able to be correlated with some temperature dependent pure component property of water. The most logical choice would be the heat of vaporization since excess enthalpy of mixing is being considered.

In order to quantify this behaviour the high temperature data for the water (1) – ethanol (2) system was fitted to a van Laar type equation:

$$H^E = \frac{A(T)B(T)x_1x_2}{A(T)x_1 + B(T)x_2} \tag{F-1}$$

$$A(T) = A_0 + A_1T \tag{F-2}$$

$$B(T) = B_0 + B_1T \tag{F-3}$$

where A_0, A_1, B_0 and B_1 are the model parameters. This high temperature behaviour was assumed to represent the 'simple solvent' behaviour of water, as shown in Fig. F-1. This behaviour was therefore extrapolated to lower temperatures and the difference between this 'simple solvent' (or "unstructured" water) behaviour and the actual (or "structured" water) behaviour is therefore the deviation of the low temperature water. These differences are shown in Fig. F-9.

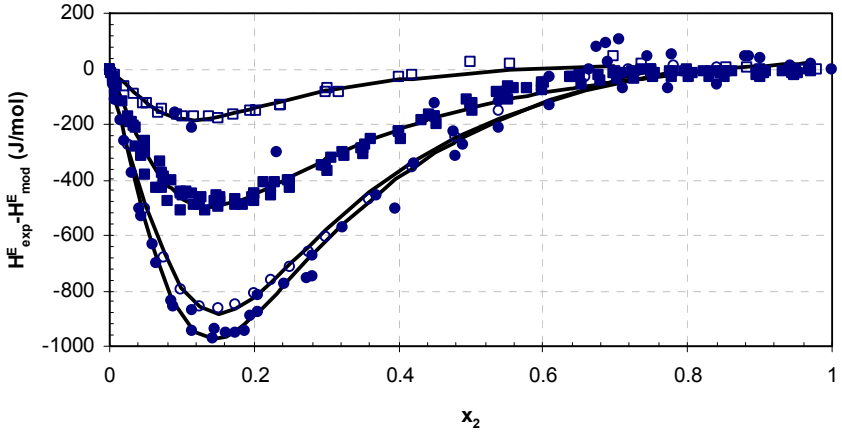


Fig. F-9 Difference between the experimental and modelled data in Fig. F-1 for the system water(1) – ethanol(2) all data from the DDB [28] (□ - 363.15 K, ■ - 323.15 K, ○ - 285.65 K, ● - 273.15 K, —— Eqn. (F-4) fitted to the data).

Rational functions were fitted to the difference isotherms in Fig. F-9 (a_0, a_1, a_2, b_1 and b_2 are all model parameters and $F(x_2)$ will be referred to as the error function from this point):

$$F(x_2) = \frac{a_0 + a_1 x_2 + a_2 x_2^2}{1 + b_1 x_2 + b_2 x_2^2} \tag{F-4}$$

To quantify this difference between the "structured" and "unstructured" water behaviour the equation was integrated in the following way:

$$Area(T) = \int_0^{x_c} \frac{F(x_2)}{1 - x_2} dx_2 \tag{F-5}$$

where Area(T) is the area under the error functions on a per mol of water basis and x_c is the root of the error function between 0 and 1. In the integral the error function ($F(x_2)$) is divided by $1-x_2$ so that the area is on a per mol water basis (needed because the heat of vaporization is obviously also on a per mole water basis). The integration was done numerically using the simple trapezoidal rule. These results are presented in Table F-1.

If one assumes that this integral enthalpy effect appears when converting "structured" water into a mostly unstructured normal solvent diluted with another component, it should relate to the difference between the "structured" water enthalpy of vaporization and the hypothetical

"unstructured" water. If one considers the heat of vaporization behaviour shown for some simple solvents (such as acetone, benzene, heptane etc.) they all follow very similar trends (see Fig. F-10 for two examples). These solvents are easily and accurately represented by the Watson equation [184]:

$$\Delta_{vap}H = \Delta_{vap}H_{ref}\left[\frac{T_c - T}{T_c - T_{ref}}\right]^{0.38} \qquad \text{(F-6)}$$

where the subscript *ref* refers to some reference point, for convenience this is usually chosen as the normal boiling point however it need not be. As shown in Fig. F-10 (and noted by Watson [184]) this equation deviates appreciably for water at temperatures below the normal boiling point ($T_r \approx 0.576$). However for higher temperatures (where water is more "unstructured") the Watson equation can represent the data quite well (if a suitable reference point is chosen). The reference point for water was chosen at the smallest possible reduced temperature which still gives good representation at higher reduced temperatures. The Watson equation is therefore assumed to represent the hypothetical "unstructured' water at lower temperatures and the difference between the curve and the data (Δ_{vap}H diff.) is the difference in enthalpy due to the water structuring at lower temperatures

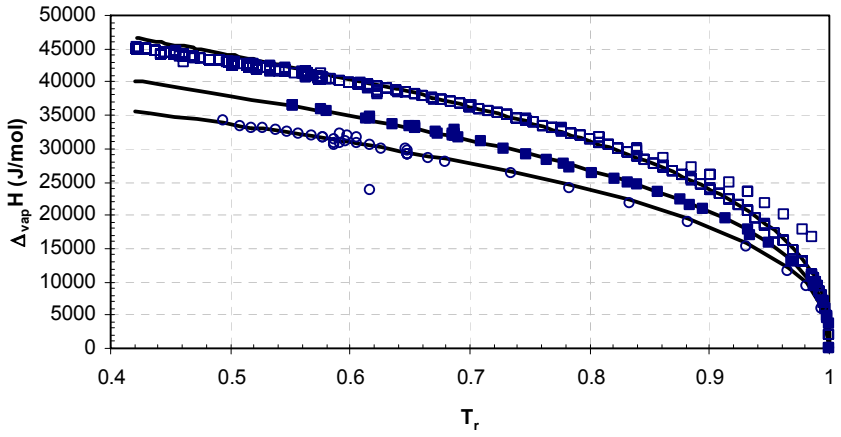


Fig. F-10 Δ_{vap}H vs. T_r for various compounds, all data from the DDB [28] (□ - water, ■ - heptane, ○ - acetone, ── Eqn. (F-6)).

The similarities between the deviation of the excess enthalpy and the heat of vaporization between the "structured" and "unstructured" solvent behaviour are shown in Table F-1. In both cases the real behaviour is lower than the simple one and difference increases with decreasing temperature. The ratios of the values are practically identical for 273.15 K – 298.15 K however at the higher temperatures the ratio deviated quite sharply. Nevertheless, since this is quite a "quick-and-dirty" means to quantify the deviation the relation between the two is still plausible. When testing other water systems it was found that the area functions for each system (e.g. methanol-water & ethanol-water) are different. Had they been the same it

would have indicated that this effect is solely due to the water and that it behaves in a predictable way. It is possible that the maximum value of the area for some molecule which is "invisible" (i.e. it does not disrupt the water structure) to water is given by the heat of vaporization difference. This will however not be evaluated any further as it is beyond the scope of this work.

Table F-1 Comparison if the difference between the simple and the real solvent behaviour for water heat of mixing data with ethanol and pure component water heat of vaporization (Area(T) and $\Delta_{vap}H$ both have units of J/mol).

T (K)	Area(T)	$\Delta_{vap}H$ diff.	Ratio
273.15	-465.9	-1519.4	3.26
285.65	-442.9	-1445.6	3.26
298.15	-409.8	-1328.6	3.24
323.15	-270.2	-1181.2	4.37
363.15	-61.8	-808.4	13.07

G. SIMPLE BEHAVIOUR OF ALKANE SOLVENTS

As shown in section 8.4 the infinite dilution activity coefficients for many compounds can be adequately represented by a generally applicable simple linear temperature dependence given by Eqn. (8-2). What this equation means is that all solutes (there are some exceptions as discussed later) at all temperatures should have an $\bar{H}_2^{E,\infty}$ of 17.35 kJ/mol. Since there are not many data available for the temperature dependence of the infinite dilution activity coefficient in hexane (or any other alkane) it would be desirable to test this assumption by extracting $\bar{H}_2^{E,\infty}$ from some excess heat of mixing data.

Theoretically this is a very simple procedure: fit a suitable curve to the data, numerically or analytically find the derivative and take the limit as one concentration tends to zero, in practice however this is a little more difficult as illustrated in the following example. Excess enthalpies of mixing data at 298.15 K for 2 different data sets are shown in Fig. G-1; one dataset runs over a very wide concentration range ($0.0322 < x_2 < 0.9879$) while the other is focused in the dilute region ($0.0005 < x_2 < 0.0581$ & $0.6914 < x_2 < 0.9879$). Excess enthalpy was fitted to a Redlich-Kister type equation:

$$H^E = x_1 x_2 \left[\sum_{k=0}^{n} a_k \left(x_2 - x_1 \right)^k \right] \tag{G-1}$$

Where a_k are the model parameters and n=5 provided a sufficiently good fit. The partial molar excess enthalpy is related to the excess enthalpy via the definition:

$$\bar{H}_i^{E,\infty} = \lim_{x_i \to 0} \left[\frac{\partial \left(n H^E \right)}{\partial n_i} \right]_{T,P,n_j} \tag{G-2}$$

Combining Eqn's (G-1) and (G-2) results in the following general relation:

$$\bar{H}_2^{E,\infty} = \sum_{k=0}^{n} a_k \left(-1 \right)^k \tag{G-3}$$

where the subscript 2 refers to the solute (i.e. in this case ethanol). Another alternative is to plot $H^E / x_1 x_2$ against x_2 (or x_1) and then the endpoints of the curve ($x_2 = 0$ and $x_1 = 0$) correspond to the $\bar{H}_2^{E,\infty}$ and $\bar{H}_1^{E,\infty}$ values. This is essentially the same procedure as some function needs to be fitted to the data to extrapolate to the endpoints. All the plots in this section will use these axes for the sake of simplicity.

When Eqn. (G-1) was fitted to each dataset and combined with Eqn. (G-3) the resulting $\bar{H}_2^{E,\infty}$ was 11.1 kJ/mol for the data over the full range and 22.9 kJ/mol for the dilute range data. The difference between both values is considerable; this illustrates that $\bar{H}_2^{E,\infty}$ cannot be found with any kind of certainty unless the data are sufficiently dilute ($x \approx 0.001$). Neither the dilute value nor the full range value correspond well with the general value of 17.35 kJ/mol. The amount of sufficiently dilute data available in the literature is quite small and so no rigorous testing could be done. However as shown in Appendix F.1 even if $\bar{H}_2^{E,\infty}$ cannot be reproduced very accurately the water correlation still had good applicability and therefore the same is assumed true here.

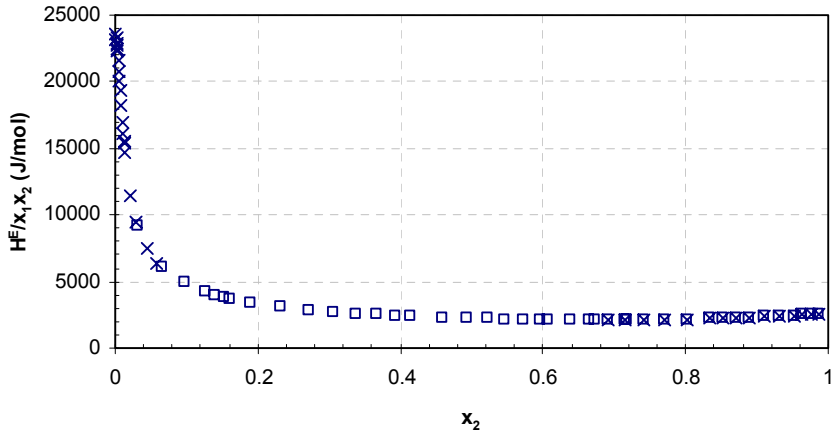


Fig. G-1 H^E/x_1x_2 vs. x_2 at 298.15 K for the system hexane(1) – ethanol (2) (□ - data from O'Shea and Stokes [185], x – data from Stokes and Burfitt [186]).

For some solutes shown in section 8.4 the activity coefficient seemed to be approximately constant over the temperature range. This corresponds to a zero excess heat of mixing at infinite dilution, this seems unlikely, what is more likely is that the slope (i.e. $-\bar{H}_2^{E,\infty} / (R \times 298.15)$) is small enough to be only slightly noticeable over the temperature ranges considered. Consider the examples of cyclohexane (Fig. G-2) and heptane (Fig. G-3) excess heat of mixing with hexane. It is almost certain that $\bar{H}_2^{E,\infty}$ for heptane will be very close to zero, the value for cyclohexane is certainly non-zero, the fitted value for cyclohexane is 0.58 kJ/mol (the data goes down to a dilution of x_2 = 0.0099, while this may not be sufficiently dilute for accurate $\bar{H}_2^{E,\infty}$ data, it suffices for an approximation) which corresponds to a slope in Eqn. (8-2) of 0.22 which is probably more or less what is observed in the data in Fig. 8-2.

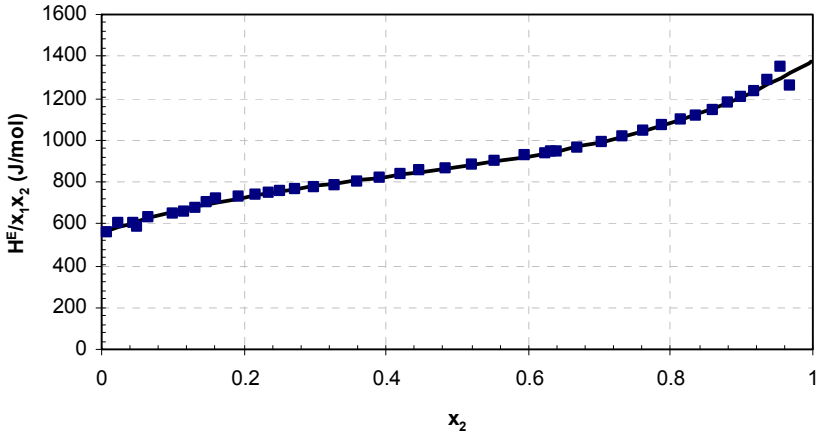


Fig. G-2 H^E/x_1x_2 vs. x_2 at 298.15 K for the system hexane(1) – cyclohexane (2) (■ - data from the DDB [28], —— Eqn. (G-1) with n = 5).

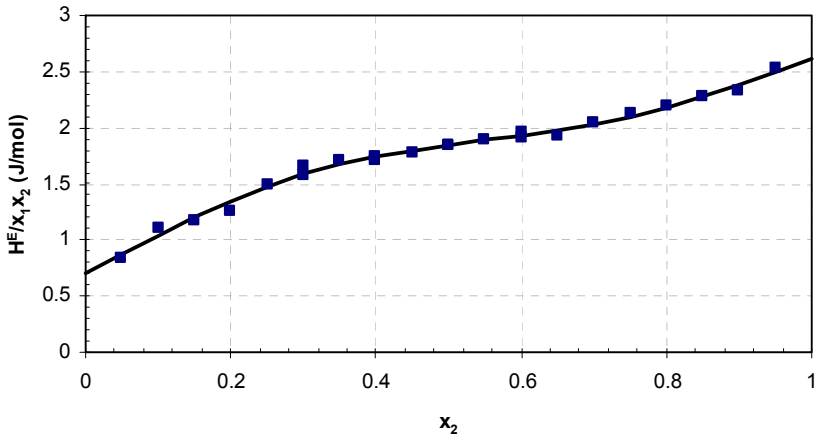


Fig. G-3 H^E vs. x_2 at 298.15 K for the system hexane(1) – heptane (2) (■ - data from the DDB [28], —— Eqn. (G-1) with n = 5) .

H. PUTTING IT ALL TOGETHER

In the chapters above much discussion has been made about the prediction of the activity of complex multifunctional organic compounds. As is often unavoidable, the work is somewhat spread across the thesis and while the conclusion does thread the various bits together there is still not a clear framework available. This section is therefore included to make a final, visual, roundup of the discussions presented above. The flow sheet following outlines the general procedure that should be followed when predictions for some solute (X) need to be made at some temperature, T.

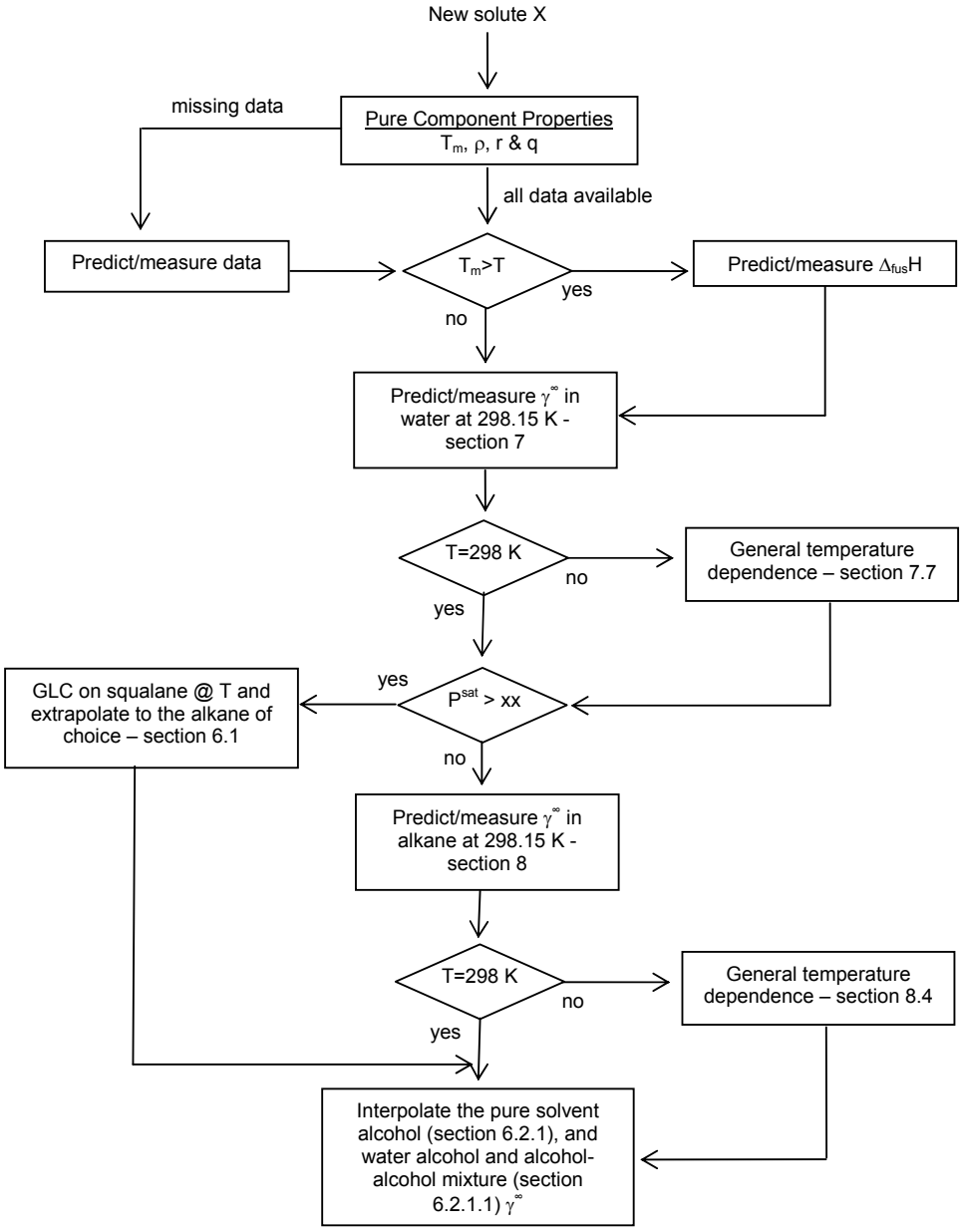


Fig. H-1 Procedure for the determination of γ^{∞} of solute X in alcohols, water and alkanes at temperature T.