Systematic Prediction of Activity Coefficients using Molecular Mechanics and Global Optimisation Techniques

Aileen May Ling Cheung

A thesis submitted for the degree of Doctor of Philosophy
of the University of London and for the
Diploma of Membership of Imperial College

Centre for Process Systems Engineering
Department of Chemical Engineering and Chemical Technology
Imperial College of Science Technology and Medicine
London SW7 2BY
United Kingdom

BIBL.
LONDIN.
UNIV.

Abstract

This thesis presents a novel systematic methodology of determining activity coefficients for phase equilibria prediction. Phase equilibria are most accurately determined experimentally; however experiments are expensive. Group contribution activity coefficient models have been developed but these techniques are limited by the lack of certain group interaction parameters. In the last few decades, there have been advances in computational chemistry techniques but although promising, computational chemistry simulations are currently limited in terms of molecular size due to the high computational expense.

Binary interaction parameters for the activity coefficient models are usually obtained through a fitting technique using experimental data. Due to the lack of these parameters for many systems, new methods have been developed to estimate those parameters by interpretating them as interaction energies between the molecular species. The methods proposed vary from the computational chemistry method used to the number of molecules and the averaging technique. While these different strategies have led to very good results for simple systems, results for more complex systems have been disappointing.

Our method is a consistent, optimal method that extends the technique towards more complex solutes typical of the agrochemical and pharmaceutical industries. Our work has shown that a cluster representation of the solvation shell is important to model these systems. The technique also incorporates a systematic technique of building the solvation shell. The system is locally minimised and the pairwise interaction energies are averaged using the Boltzmann distribution equation. Case studies are shown for mono- and bi-substituted benzene derivatives in solvents of different polarity. This method is also shown to predict the stability of two paracetamol polymorph forms within a range of temperature in various solvents.

A specially developed global optimisation algorithm based on the branch-and-bound framework using Molecular Mechanics is developed to find the specific interactions between molecular pairs. Tight convex underestimators have been derived for the nonconvex terms in the objective function and constraints are convexified using interval arithmetic and symbolic reformulation techniques. Examples are shown for a single molecule and pairs.

Acknowledgements

I am indebted to my supervisor, Dr. Claire Adjiman, for her patience, guidance and support. Thank you for letting me learn and make my own mistakes before bringing me back on track, and for always being ready to help when I ran into problems.

I would also like to thank Mitsubishi Chemicals Corporation for their financial support and occasional trips, and especially to Dr. Petr Kolar for his help and invaluable advice.

A big thank you too to Graham and Terence for helping me with installations and debugging, and to Ruth for making my paperwork so much easier.

Thanks to all my friends for giving me a social life outside the centre by dragging me away from the computer and ignoring my limp excuses, for bright chapters filled with laughter in their company and for lending an ear during my more difficult moments. "I can truly say I had a great time in London because of you."

Lastly, heartfelt gratitude to my family for their unconditional love, and especially to my parents for encouraging me to follow my dreams, for believing in me and for always being there when I needed them.

For it is not inertia alone that is responsible for human relationships repeating themselves from case to case, indescribably monotonous and unrenewed; it is shyness before any sort of new, unforeseeable experience with which one does not think oneself able to cope.

...And if only we arrange our life according to that principle which counsels us that we must always hold to the difficult, then that which now still seems to us the most alien will become what we most trust and find most faithful.

- R. M. Rilke -

Contents

1 Introduction **12**

2 Literature Review on Phase Equilibria Prediction **16**

2.1 Introduction 16

2.2 Thermodynamics of Solvation 17

2.2.1 Intermolecular Forces 20

2.2.2 Activity Coefficient Models 22

2.3 Quantitative Structure-Property Relationships (QSPRs) 24

2.3.1 Correlations with Experimentally Determined Physicochemical Properties 25

2.3.2 Correlations based on Group Contributions 25

2.3.3 Correlations with Parameters from Molecular Structure 29

2.4 Computational Chemistry for Phase Equilibria 31

2.4.1 Monte Carlo and Molecular Dynamics Simulations 33

2.4.2 Deriving Parameters for QSPRs 35

2.4.3 Deriving Equation-of-state (EOS) parameters 37

2.4.4 Deriving Interaction Parameters for Activity Coefficient Models . . 38

2.5 Molecular Mechanics 42

2.5.1 Bond stretching 42

2.5.2 Angle Bending 43

2.5.3 Torsional terms 44

2.5.4 Out-of-plane angles 44

2.5.5 Electrostatic interactions 44

2.5.6 van der Waals interactions 46
2.5.7 Force field parameterisation 47
2.5.8 Description of available force fields 48
2.5.9 Applications of Molecular Modelling 52
2.6 Conclusions 52

3 Prediction of UNIQUAC Activity Coefficients 54
3.1 Introduction 54
3.2 General Principles 55
3.3 Molecular Pairs 56
3.3.1 Simple Alkane/Alkane and Alkane/Amine Systems 57
3.3.2 Paracetamol 60
3.4 Molecular Clusters 66
3.4.1 Solvent Accessible Surface Area 67
3.4.2 First Solvation Shell Modelling 68
3.5 Discussion of Methodology 74

4 Case Studies 77
4.1 Introduction 77
4.2 Infinite Dilution Activity Coefficients Predictions 77
4.2.1 Monofunctional Solutes 79
4.2.2 Bifunctional Solutes 86
4.3 Polymorph Solubility Prediction 93
4.4 Conclusions 96

5 Literature Review on Conformational Search of Molecular Systems 101
5.1 Introduction 101
5.2 Potential Energy Minimisation 102
5.3 Stochastic Search 105
5.4 Deterministic Global Search 107
5.5 Conclusions 112

6 Deterministic Global Optimisation of the Potential Energy of Molecular Systems 113

6.1 Introduction 113

6.2 Problem Definition 114

6.3 Dependent Variables 119

6.3.1 *Dependent* Angles Constraints 119

6.3.2 *Dependent* Torsions Constraints 121

6.3.3 Nonbonded Distances Constraints 123

6.4 Convexification of Potential Energy Functions 123

6.4.1 Convexification of Torsional Terms 124

6.4.2 Convexification of Nonbonded Interactions Term 129

6.5 Convexification of the Constraints 129

6.5.1 Symbolic Reformulations 130

6.5.2 Bounds on the *Dependent* Variables 134

6.5.3 Triangle and Tetrangle Inequalities 140

6.6 General Structure of Optimisation Framework 141

6.7 Case Studies 144

6.7.1 Ethylamine molecule 144

6.7.2 n-Butane/ethylamine pair 148

6.7.3 Ethylamine/ethylamine pair 156

6.8 Conclusions 158

7 Conclusions and Future Directions 159

7.1 Introduction 159

7.2 Conclusions 160

7.3 Future Directions 162

7.3.1 Space-filling model 162

7.3.2 Handling dissimilar molecules 163

7.4 Integration of property prediction methodology within a solvent design framework 163

A Modified D-UNIFAC Parameter Matrix 180

B Comparison of Force Fields **182**

C Boltzmann Jump Procedure utilising API routines from *Cerius*^{2} **186**

D Experimental Data for Bifunctional Benzene Derivatives **193**

E Once-Continuously Differentiable Underestimators for the Torsional Angle Terms **200**

F Analytical Proof of the Convexification of Nonbonded Interactions Term **204**

List of Figures

2.1 Ideal solubility where the curve shows the solubility predicted using the regular solution theory, o denotes solubilities below ideality and x denotes solubilities above ideality 19

2.2 Principal hydrogen bonding interaction sites 21

2.3 Possible relationships between different types of computational chemistry (Gubbins and Quirke, 1996) 31

2.4 Schematic representation of the four key contributions to a molecular mechanics force field: bond stretching, angle bending, torsional terms and non-bonded interactions 43

3.1 Global minimum potential energy of n-butane/n-pentane pair: -21.0588 kcal/mol where the grey atoms are carbon and the white atoms are hydrogen 58

3.2 Phase diagram for n-butane/n-pentane at 25^{o}C 59

3.3 Global minimum potential energy of n-butane/ethylamine pair: -28.0414 kcal/mol where the grey atoms are carbon, white atoms are hydrogen and the blue atom is nitrogen 59

3.4 Phase equilibria for n-butane/ethylamine at 25^{o}C 60

3.5 Paracetamol molecule where carbon atoms are shown in grey, oxygen in red, nitrogen in blue and hydrogen in white 61

3.6 Solubility of paracetamol showing experimental data and prediction by regular solution theory 62

3.7 Solubility of paracetamol in n-alcohols 63

3.8 Acetic acid dimer configuration 65

3.9 Top view: Surface area of chlorobenzene solute where a solvent molecule would be found after minimisation 66
3.10 Side view: Surface area of chlorobenzene solute where a solvent molecule would be found after minimisation 67
3.11 van der Waals surface area and solvent accessible surface area (SASA) . . 67
3.12 Initial configuration of chlorobenzene solute surrounded by 12 chlorobenzene molecules 68
3.13 Initial configuration of phenol solute surrounded by 12 phenol molecules . 69
3.14 Cluster of chlorobenzene solute surrounded by 12 chlorobenzene molecules (solute-solute interactions) - local energy minimisation from configuration in Fig. 3.12 70
3.15 Cluster of phenol solute surrounded by 12 phenol molecules (solute-solute interactions) - local energy minimisation from configuration in Fig. 3.13 . 70
3.16 Cluster of hexane solvent surrounded by 12 hexane molecules (solvent-solvent interactions) - local energy minimisation from starting configuration 71
3.17 Cluster of ethanol solvent surrounded by 12 ethanol molecules (solvent-solvent interactions) - local energy minimisation from starting configuration 71
3.18 Cluster of chlorohexane solute surrounded by 12 hexane molecules (solute-solvent interactions) - local energy minimisation from starting configuration 72
3.19 Cluster of phenol solute surrounded by 14 acetone molecules (solute-solvent interactions) - local energy minimisation from starting configuration . . . 73
3.20 Cluster of chlorohexane solute surrounded by 16 ethanol molecules (solute-solvent interactions) - local energy minimisation from starting configuration 74
3.21 Cluster of phenol solute surrounded by 22 water molecules (solute-solvent interactions) - local energy minimisation from starting configuration . . . 75
3.22 Unsuitable model of cluster of phenol solute surrounded by 22 water molecules (solute-solvent interactions) - local energy minimisation from starting configuration 75

4.1 Comparisons of infinite dilution activity coefficients of 6 monofunctional solutes in (i) *n*-hexane, (ii) acetone, (iii) ethanol and (iv) water at 298.15K 83

4.2 Graph showing correlation between infinite dilution activity coefficients, γ_1^∞, between experimental data with predicted data (some data points are out of bounds) for monofunctional solutes 86
4.3 Bifunctional benzene derivatives 87
4.4 Comparisons of infinite dilution activity coefficients of 7 bifunctional solutes in (i) *n*-hexane, (ii) acetone, (iii) ethanol and (iv) water at 298.15K 90
4.5 Graph showing correlation between infinite dilution activity coefficients, γ_1^∞, between experimental data with predicted data (some data points are out of bounds) for bifunctional solutes 93
4.6 Monoclinic form of paracetamol 94
4.7 Orthorhombic form of paracetamol 95
4.8 Transition temperature vs solvent solubility parameter 96
4.9 Solubility of the paracetamol polymorphs in diethylamine 97
4.10 Solubility of the paracetamol polymorphs in toluene 97
4.11 Solubility of the paracetamol polymorphs in acetone 98
4.12 Solubility of the paracetamol polymorphs in ethanol 98
4.13 Solubility of the paracetamol polymorphs in water 99

5.1 Branch and bound (Adjiman, 1998) 108
5.2 Convexification of the objective function and the feasible region (Adjiman, 1998) 110

6.1 *n*-butane molecule showing the atom numbers 114
6.2 Spherical triangle showing atoms 11, 12 and 13 119
6.3 Spherical triangle showing relationship between the angles 120
6.4 *Dependent* Torsional Angle 122
6.5 Spherical triangle showing atoms 3, 8, 10 with atom 4 at the origin 122
6.6 Convexification of the first torsional angle energy term - note that levels 1 and 2 have overlapping underestimates 126
6.7 Convexification of the second torsional angle energy term - note that the underestimators for levels 2 and 3 coincide 127
6.8 Convexification of the third torsional angle energy term 128

6.9 Global optimisation branch-and-bound framework for molecular systems . 143

6.10 Ethylamine molecule showing the atom numbers 144

6.11 *n*-Butane/ ethylamine molecule showing the atom numbers 144

6.12 Figure showing relationship for the *pseudo* torsion 152

6.13 Ethylamine/ ethylamine molecule showing the atom numbers 156

7.1 Traditional and new use of structure-property relationships 164

A.1 Currently available group interaction parameters for D-UNIFAC obtained from http://www.uni-oldenburg.de/tchemie/Consortium/Matrix_MoUniDo.html 181

Chapter 1

Introduction

The fine chemicals industry relies heavily on solvents for manufacturing. Therefore, the selection of solvents during the process development stage significantly affects the overall plant design and operation as well as the economics, safety and environmental impact of the process. Solvents may be used in numerous processing tasks such as extraction to separate the products of the synthetic reaction from the reagents, or as a reaction medium where the choice of solvent can affect the rate and yield of the reactions. Unfortunately, due to technical limitations, the choice of appropriate solvents has long had to be based on the combination of the chemists' expertise and expensive experimentation, often resulting in sub-optimal decisions. However, this method is time-consuming and expensive making the evaluation and comparison of a large number of alternative solvents difficult. For example, the cost of obtaining one vapour-liquid equilibria data point has been estimated to be about US$2600 (Gubbins and Quirke, 1996). Significant competitive advantage can be gained by companies that take a leading role in exploiting recent advances in physical and computational chemistry and in computer technology for the purpose of solvent selection.

Molecular Design is defined as the design of pure compounds or mixtures that exhibit certain desired properties (Odele and Macchietto, 1993). The key requirement in the development of a molecular design methodology is the availability of a reliable and low-cost prediction capability based on structure-property relationships for the phase equilibria of fine chemicals in different solvents and specially in the area of crystallisation. the iden-

tification of polymorph stability. The aim is to identify 'good' solvents to be used in processes even when data is not readily available and current models lack the necessary parameterisation.

The objectives of the project are then:

- Development of a reliable and low-cost approach to predicting solubility especially for pharmaceutical compounds
- Development of a prediction technique for polymorph stability as a function of the solvent and temperature
- Automation of the property prediction method using global optimisation

The ability to predict phase equilibria hinges on the determination of the relative strengths of the solute-solute, solvent-solvent interactions compared with the solute-solvent interactions. Solutes typical of the agrochemical and pharmaceutical industries are flexible molecules with 20 to 70 atoms, often containing aromatic rings and heteroatoms (O, N, P, S, F, etc.). The combination of the electronegative heteroatoms and aromatic π-electrons may lead to specific interactions of these molecules with small polar and aprotic solvents.

The extension of the phase equilibria prediction capabilities to the prediction of relative solubility for polymorphs of pharmaceutical compounds in various solvents and as a function of temperature is the second goal of this project. It is of critical importance for the pharmaceutical industry as it enables the identification of the transition temperature between the polymorphic forms and can thus provide greatly enhanced quality control.

The final goal of this project is to automate the property prediction technique using global optimisation techniques. This would enable reliable computations for new systems, facilitate technology transfer and ultimately allow the prediction technique to be used in a molecular design algorithm to design an optimal solvent for a particular solute.

This thesis is divided into two parts: the design of the phase equilibria prediction method-

ology and its use, and the mathematical development of the global optimisation algorithm. Chapter 2 provides an overview of property prediction techniques currently in use and gives a discussion on their respective strengths and weaknesses. New developments based on computational chemistry are also presented as these developments are promising for more cost-effective property prediction techniques. A summary of the theory behind molecular mechanics and available force fields are also presented.

Chapter 3 charts the development of a suitable prediction technique which provides a trade-off between computational cost and accuracy. A general methodology is first presented. Previous research has used pairs of molecules but although this works reasonably well for simple molecules, it fails for more complex solutes where the solute molecules are larger than the solvent molecules. The best approach appears to be the representation of the solvation shell around the solute molecule through a cluster. A suitable number of solvent molecules to be included in the solvation shell representation is also identified. Chapter 4 presents case studies comparing the methodology with D-UNIFAC, COSMO-RS and experimental results are shown for mono- and bifunctional benzene derivatives. An extension of the methodology to predict polymorph stability for various solvents at different temperatures is also shown in Chapter 4.

The subsequent two chapters highlight the work done in automating the property prediction methodology. In order to build the model, the strong interactions have to be identified. This is done using a deterministic global optimisation algorithm which aims to find the global minimum potential energy configuration and other low-lying energy minima for a molecular system which are usually where the strong interactions exist. In Chapter 5, a summary of available optimisation strategies to find the global minimum potential energy of a molecular system is presented. The global minimum potential energy is usually where the specific interactions are the strongest. These strategies are divided into stochastic and deterministic search methods. Chapter 6 begins with an explanation of the objective function detailing the potential energy function subject to various constraints and variable bounds. The potential energy is described using the OPLS force field (Jorgensen et al, 1984) and the minimisation problem is formulated using $3N$-6 *independent*

internal coordinates where N is the number of atoms. The rest of the internal coordinates are calculated from the *independent* variables. Tight convex underestimators are then described for the nonconvex terms and constraints are reformulated using symbolic reformation (Smith and Pantelides, 1998). Bounds for the *dependent* variables are determined using interval arithmetic. Case studies for a single molecule n-butane and for pairs of n-butane/ethylamine and ethylamine/ethylamine are presented.

Finally, Chapter 7 gives the conclusion of the research, and a discussion of future research directions. Extension and further development of the automation of the prediction technique is needed and research into the automated space-filling model of the solvation shell. Integration of the property prediction methodology into a *solvent design* framework is then discussed, with the objective of designing the *optimal* solvent for a particular solute subjected to process and product specifications.

Chapter 2

Literature Review on Phase Equilibria Prediction

2.1 Introduction

Different methodologies are currently available to predict properties. Zhao and Cabezas (1998) state that an ideal estimation technique for a physical property would provide reliable physical and thermodynamic properties for pure substances and for mixtures at any temperature, pressure, and composition, indicate the state (solid, liquid or gas), require a minimum of input data, choose the least-error route, indicate the probable error, and minimise computation time. Among the more popular property prediction methods currently available in the chemical and pharmaceutical industries are:

- thermodynamical models such as equations-of-state and activity coefficient models;
- quantitative structure-property relationships;
- computational chemistry.

These prediction methods should incorporate thermodynamic effects and intra- and intermolecular forces such as hydrogen bonding tendencies to better describe the phase behaviour for the systems.

This chapter will first describe the thermodynamics of solvation to illustrate further why

a solute dissolves in a solvent due to competing interactions between the solute and the solvent molecules. A thermodynamical method of predicting phase equilibria as a function of the interaction energies is then presented. Section 2.3 describes fitting techniques used to predict phase equilibria and Section 2.4 then presents computational chemistry techniques that may be used to model the systems. Finally, Section 2.5 provides a detailed description of molecular mechanics and the types of force fields available.

2.2 Thermodynamics of Solvation

The solvation of the solute in a solvent is considered to be composed of the following processes:

- Formation of a cavity in the solvent equal to the volume of the solute molecule,
- Reorganisation of the solvent molecules around the cavity and reorganisation of conformation of the solute molecule within the cavity,
- Formation of interactions between the solute molecule and solvent molecules in the solvation shell,
- Formation of hydrogen-bonds.

The solubility thus depends on the strength of the solute-solvent interactions compared with the solute-solute and solvent-solvent interactions. Lewis and Randall (1961) introduce the term fugacity, f, as a measure of escaping tendency. The thermodynamic criterion for multicomponent phase equilibria is that for every component i:

$$f_i^1 = f_i^2 \tag{2.1}$$

where 1 is the first phase and 2 is the second phase, and f_i refers to the fugacity of component i.

The fugacities of vapour and liquid are described in the following equations:

$$f_i^V = \phi_i y_i P \tag{2.2}$$

$$f_i^L = \gamma_i x_i f_i^o \tag{2.3}$$

where x_i, y_i are the liquid and vapour mole fractions of component i, ϕ_i is the fugacity coefficient and γ_i is the activity coefficient. The symbol o denotes the standard state. For an ideal liquid phase, $\gamma_i = 1$. The activity, a_i, of a compound i is defined as:

$$a_i = \frac{f_i}{f_i^o} \tag{2.4}$$

so for an ideal system, $x_i^{id} = a_i$. The theoretical definition of an ideal system is then a system where the interactions between unlike molecules i and j are identical to those between like molecules (i and i, j and j) in the same state. Therefore, any interchange between the molecules leaves the internal energy and volume of the system unchanged.

Regular solution theory was developed by Hildebrand and Scott (1950) and utilises knowledge of the solubility parameters of the solute, δ_i, and solvent, δ_j, to derive the activity coefficients of the system. δ_i and δ_j are defined as the square root of the interaction energies of the solute-solute, U_{ii}, and solvent-solvent, U_{jj}, systems respectively.

$$U_{ii} = \delta_i^2 \tag{2.5}$$

$$U_{jj} = \delta_j^2 \tag{2.6}$$

The activity coefficients are defined as a function of the square of the difference between the two solubility parameters so when $\delta_i = \delta_j$, $\gamma_i = 1$ and the mixture behaves ideally.

$$\ln \gamma_i = \frac{V_i \Phi_j^2 (\delta_j - \delta_i)^2}{RT} \tag{2.7}$$

where V_i is the molar volume of the solute i and Φ_j is the volume fraction of the solvent j. Therefore, the theory does not account for specific solute-solvent interactions and thus fails to predict negative deviations from ideality. In this case, the fact that the solubility

is greater than ideality (marked by x in Figure 2.1) is not predicted.

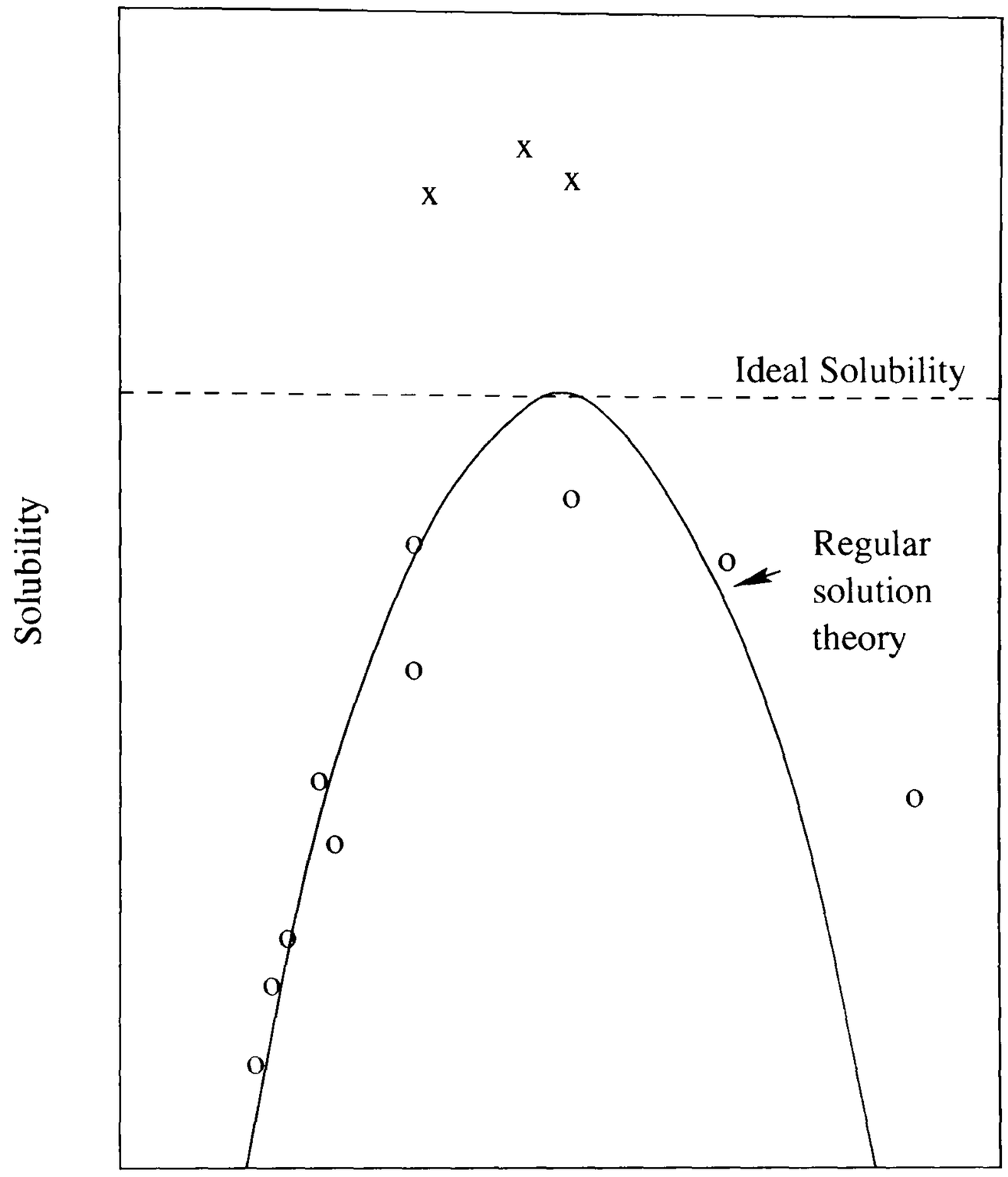


Figure 2.1: Ideal solubility where the curve shows the solubility predicted using the regular solution theory, o denotes solubilities below ideality and x denotes solubilities above ideality

For solid-liquid equilibria, the enthalpy of solution of a solid, $\triangle H_s$, is equivalent to the sum of the enthalpy of fusion at the absolute temperature T of the solution, $\triangle H_T^f$, and the enthalpy of liquid-liquid mixing, $\triangle H_{mix}$. At ideality, $\triangle H_{mix}$ is zero and the mole fraction of the solute in the liquid phase, x_i^{id}, may be calculated from:

$$x_i^{id} = \exp\left[\frac{\triangle H_m^f}{R}\left(\frac{1}{T_m} - \frac{1}{T}\right)\right] \tag{2.8}$$

where $\triangle H_m^f$ is the enthalpy of fusion at the melting point of component i, T_m, and R is the gas constant. For real systems, the equation becomes (Gmehling et al., 1978):

$$x_i\gamma_i = \exp\left[\frac{\triangle H_m^f}{R}(\frac{1}{T_m} - \frac{1}{T}) - \frac{\triangle C_p}{R}(\ln\frac{T_m}{T} - \frac{T_m}{T} + 1)\right] \tag{2.9}$$

where x_i is the mole fraction of the solute, γ_i is the activity coefficient, $\triangle C_p$ is the difference in heat capacity between the solid and liquid phases.

2.2.1 Intermolecular Forces

Solubility is highly dependent on the nature and relative importance of the intermolecular forces: solute-solute, solvent-solvent and solute-solvent. All molecules are subject to Coulombic and van der Waals interactions. The former describe the energy of interaction between two atoms depending on the atomic charges. The latter are the sum of the dipole-dipole and dispersion forces in the system. A more detailed description of these forces is given in Section 2.5.

Another very important intermolecular interaction is hydrogen bonding, an attractive specific interaction between a hydrogen bond donor, which is usually a hydrogen atom bonded to an electronegative atom, with a hydrogen bond acceptor, which may be any electron-rich atom. Hydrogen bonds are determined by their binding energy in the range 2-10 kcal/mol (Scheiner, 1997). Figure 2.2 shows the principal hydrogen bonding interaction sites (Laurence and Berthelot, 2000).

Solubility is then dependent on the competing nature of the solute-solvent interactions with the solute-solute and solvent-solvent interactions, both specific and non-specific. Specific interactions may be specified by (Grant and Higuchi, 1990):

- A high molar internal energy of interaction, U_{ij}, which is define as $U_{ij} >> RT$:
- A high steric requirement or an appreciable negative entropy change.

The first feature allows for the solute-solvent bond to persist through molecular vibrations, and the second prevents the molecules from spinning. Thus, hydrogen bonding and

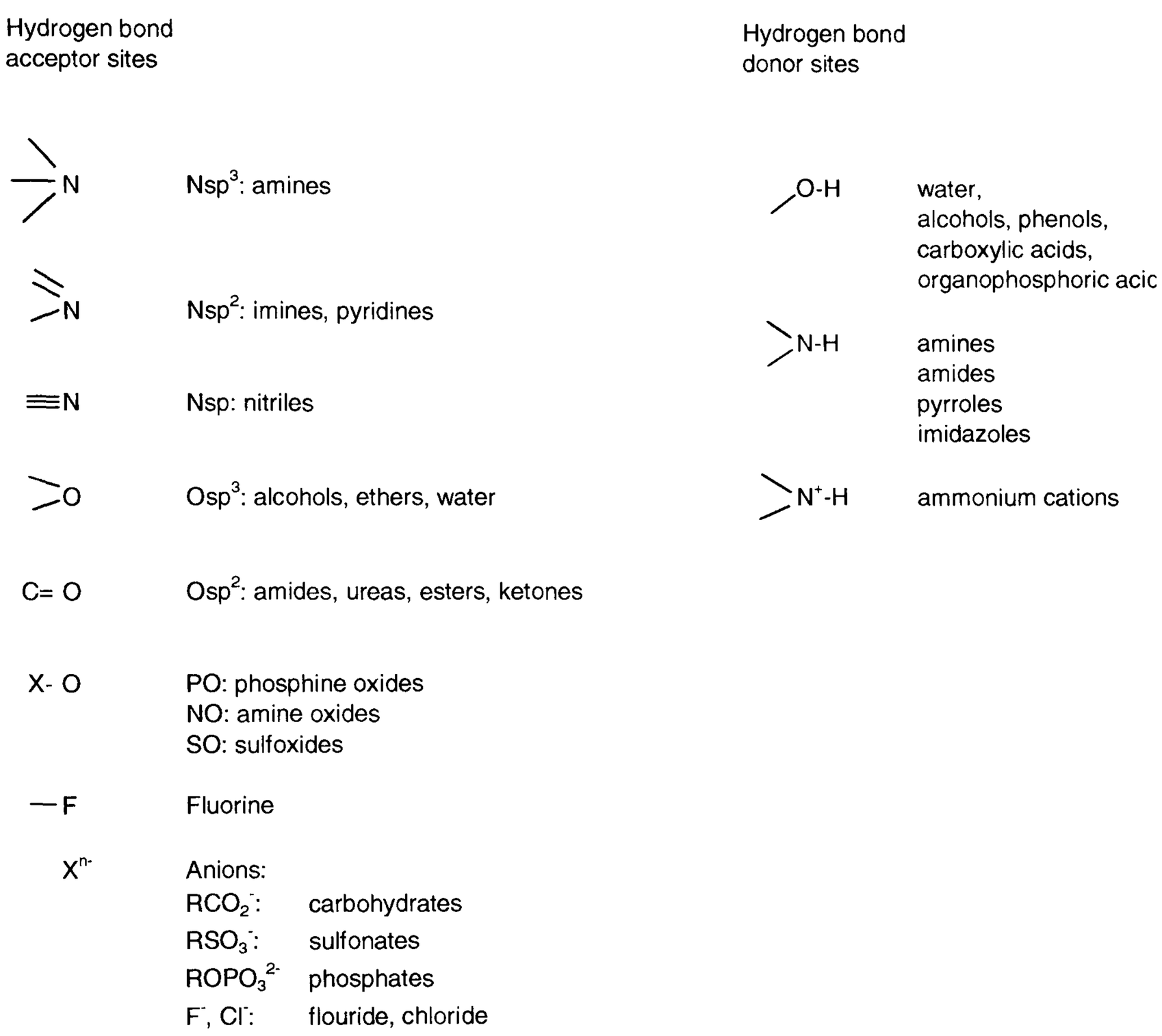


Figure 2.2: Principal hydrogen bonding interaction sites

certain charge transfer interactions are classified as specific interactions while van der Waals and Coulombic forces are nonspecific in nature.

The importance of suitably defining solvent effects was discussed by Vilaseca (1995) who studied solvents effects on a polar molecule using Monte Carlo simulations. This paper aimed to complete the study of the influence of polar and nonpolar solvents on the conformations of polar and nonpolar nonrigid molecules. For each simulation a solvent molecule was randomly selected and then translated and rotated while for a solute molecule, the internal rotational angle was also changed. Three different types of solute-solvent interactions were chosen: Lennard-Jones interaction sites without net charges, Lennard-Jones interaction sites with charges and the inclusion of polarisation effects due to the permanent C-Cl bond dipoles. The polarisation effects were found to predict accurately the favored conformation of the solute in the solvent.

The activity, a_i, describes the real, effective concentration of the species in the solution in a thermodynamically accurate sense. At infinite dilution, the solute i is infinitely dilute and the solvent j is unperturbed by any solute particles and each solute particle feels no forces other than from the solvent. Therefore, infinite dilution activity coefficients are a measure of the solute-solvent interactions with the solvent-solvent interactions (and thus the solubility) from:

$$x_i = \frac{1}{\gamma_i^{\infty}} \tag{2.10}$$

2.2.2 Activity Coefficient Models

The UNIQUAC activity coefficient model was developed by Abrams and Prausnitz (1975) following Guggenheim's (1952) quasi-chemical theory of liquid mixtures. Activity coefficients are related to the molar excess Gibbs energy g^E through

$$g^E = RT \sum_i x_i \ln \gamma_i \tag{2.11}$$

where R is the gas constant, T is the absolute temperature, and x_i is the mole fraction of component i. The UNIQUAC equation calculates the excess Gibbs free energy as a sum of a combinatorial part ($\ln \gamma_i^C$) to account for molecular size and shape, and a residual part ($\ln \gamma_i^R$) to account for molecular interactions. The model requires prior knowledge of the binary parameters of the solute-solvent systems, τ_{ij} and τ_{ji}, that are usually fitted to experimental data.

$$\ln\gamma_i = \ln\gamma_i^C + \ln\gamma_i^R \tag{2.12}$$

where

$$\ln \gamma_i^C = \ln \frac{\Phi_i}{x_i} + \frac{z}{2} q_i \ln \frac{\theta_i}{\Phi_i} + l_i - \frac{\Phi_i}{x_i} \sum_j x_j l_j \tag{2.13}$$

$$\ln \gamma_i^R = q_i - q_i \ln(\sum_j \theta_j \tau_{ji}) - q_i \sum_j \frac{\theta_j \tau_{ij}}{\sum_k \theta_k \tau_{kj}} \tag{2.14}$$

and

$$l_j = (\frac{z}{2})(r_j - q_j) - (r_j - 1) \tag{2.15}$$

$$\tau_{ji} = \exp\left(-(\frac{U_{ji} - U_{ii}}{RT})\right) \tag{2.16}$$

and where the value of z is usually 10 and the average area fraction θ and the average segment fraction Φ are defined by:

$$\theta_i = \frac{q_i x_i}{\sum_j q_j x_j} \tag{2.17}$$

$$\Phi_i = \frac{r_i x_i}{\sum_j r_j x_j} \tag{2.18}$$

where x is the mole fraction, r is a relative molecular volume parameter, q is a relative molecular surface area parameter and U_{ij} is the interaction energy between the binary system i and j. r, q and U are usually obtained by fitting experimental data.

Gupta et al. (1997) have used a modified activity coefficient model (UNIQUAC-HB)

that includes hydrogen-bonding interactions to predict the phase equilibrium behavior of antibiotics. The activity coefficient was expressed as a sum of the combinatorial, residual and hydrogen bonding contributions.

$$\ln\gamma_i = \ln\gamma_i^C + \ln\gamma_i^R + \ln\gamma_i^{HB} \tag{2.19}$$

A lattice-fluid hydrogen-bonding theory was chosen because it has well-defined parameters which may be measured spectroscopically. The hydrogen bonding contribution is defined as:

$$\ln\gamma_i^{HB} = r_i \sum_k \sum_l v_{kl} - \sum_k d_k^i \ln\frac{v_d^k}{v_{k0}} - \sum_l a_l^i \ln\frac{v_d^l}{v_{0l}} \tag{2.20}$$

where k is the type of proton-donor group, l is the type of proton-acceptor group, d_k^i is the number of k-type proton-donor groups on molecule i, a_l^i is the number of l-type proton-acceptor groups on molecule i, and $v_{kl} = N_{kl}/rN$, $v_{k0} = N_{k0}/rN$ and $v_d^k = N_d^k/rN$ etc. N_{kl} is the number of kl hydrogen bonds, N_{k0} is the number of free acceptor sites of l type, N is the total number of molecules, N_d^k is the total number of k-donor groups in the solution, and r is the average size parameter for the mixture defined as $r = \sum_i r_i x_i$. In this case, the aqueous/organic partitioning behaviour of an antibiotic, chloramphenicol, is measured. The calculations using this model can accurately represent the observed experimental behaviour.

2.3 Quantitative Structure-Property Relationships (QSPRs)

Quantitative Structure-Property Relationships (QSPRs) relate significant numerical descriptors of a compound to the physical property of interest. The use of QSPRs to predict solubility is logical because solubility is almost exclusively dependent on the intermolecular forces between the solute molecules and the solvent molecules. Therefore, the solute-solute, solute-solvent and solvent-solvent interactions that are largely responsible for the amount of solute dissolving in the the solvent should be numerically represented by the descriptors. Yalkowsky and Banerjee (1992) classified the prediction of solubility using QSPR techniques into three categories:

- correlations with experimentally determined physicochemical properties, such as partition coefficient, chromatographic retention time, melting point, boiling point, molar volume (derived from the liquid density and surface tension), or the parachor (derived from density and surface tension);

- correlations based on group contributions derived from measured aqueous solubilities;

- correlations with parameters calculated solely from the molecular structure based on topological, geometric and electronic descriptors.

2.3.1 Correlations with Experimentally Determined Physicochemical Properties

Yalkowsky and coworkers correlated the aqueous solubilities of polycyclic aromatic hydrocarbons (Yalkowsky and Valvani, 1979), halobenzenes (Yalkowsky et al., 1979) and nonelectrolytes (Yalkowsky and Valvani, 1980) using the melting point and either the molecular surface area or the octanol-water partition coefficient. The authors suggest that the melting point accounts for the relative energy loss of the lattice on dissolution and the octanol-water coefficient approximates the activity coefficient.

This method however cannot be applied to compounds that have not been synthesised as experimental values are needed to correlate with the solubilities.

2.3.2 Correlations based on Group Contributions

The concept behind the group contribution approach is that each molecule is considered as a combination of different functional groups (Joback and Reid, 1987). Although the number of molecules important to the chemical industries is huge, there exists a smaller number of atomic groups which constitute these molecules. Each such group contributes a certain value to a particular property based on the assumptions of transferability and additivity. The contributions from one group are assumed to be independent of that from another group. In actual fact, the contributions from a carbonyl group will vary slightly depending on whether it is interacting with a methyl group or a hydroxyl group, although

experience has shown that the values can be quite close. Nonetheless, for certain groups, it is difficult to obtain a group contribution, meaning that the environment (proximity effects) is significant. The properties of each group are usually fitted from experimental data. The group contribution method therefore strikes a balance between accuracy and computational ease.

The UNIFAC method (Universal Functional-group Activity Coefficients) (Fredenslund et al., 1975) is based on the UNIQUAC model and the group contribution concept. In the UNIFAC equation, the combinatorial part of the equation is the same as the UNIQUAC model but the parameters r_i and q_i are calculated as the sum of the group properties, R_k and Q_k respectively. The residual part of the equation is replaced by the solution of groups concept.

$$\ln\gamma_i^R = \sum_k v_k^{(i)}[\ln\Gamma_k - \ln\Gamma_k^{(i)}] \tag{2.21}$$

where Γ_k is the group residual activity coefficient, and $\Gamma_k^{(i)}$ is the residual activity coefficient of group k in a reference solution containing only molecules of type i. The group activity coefficient is given as:

$$\ln\Gamma_k = Q_k[1 - \ln(\sum_m \Theta_m\Psi_{mk}) - \sum_m(\frac{\Theta_m\Psi_{km}}{\sum_n \Theta_n\Psi_{nm}})] \tag{2.22}$$

where

$$\Theta_m = \frac{Q_m X_m}{\sum_n Q_n X_n} \tag{2.23}$$

$$\Psi_{mn} = \exp(-[\frac{U_{mn} - U_{nn}}{RT}]) \tag{2.24}$$

where U_{mn} is a measure of the interaction between groups m and n.

The original UNIFAC model has been modified to include the temperature dependence of the interaction parameters (Gmehling et al., 1993) called D-UNIFAC. At the moment, one major limitation of the model is the absence of interaction parameters for many of

the functional groups. Work is still continuing to complete the set as binary interaction parameters for certain groups, e.g., COOH with 18 other groups including CH_2NH_2 and CH_2NH, are not available. The present status of the publicly available modified UNIFAC parameter matrix is shown in the *Appendix A*.

The interaction parameters in the original UNIFAC correlation were fitted to vapour-liquid equilibria experimental values; as a result, they are not accurate for predictions of LLE. Magnussen et al. (1981) published a set of UNIFAC interaction parameters which are fitted to LLE data and thus give more accurate predictions for LLE. The interaction parameters were developed for the temperature range $25 - 40^oC$ and were temperature independent. Hooper et al. (1988) presented a UNIFAC model which focused on systems containing water and hydrocarbons over a wide range of temperature using temperature-dependent water-organic interaction parameters. The model's results give reasonable comparison with experimental values compared with Magnussen et al. (1981) which has not been parameterised for higher temperatures.

Gmehling et al. (1978) used UNIFAC to predict the SLE for several systems using Equation 2.8. The melting temperature and enthalpy of fusion have to be known beforehand. For binary systems, the largest difference between the predicted solubility with experimental values was 3.7 mol % for the naphthalene/*n*-hexane system, 0.82 mol % for the anthracene/chloroform system and 5.0 mol % for the phenanthrene/carbon disulfide system. The method can also be used to determine the eutectic temperature and composition where the largest deviation in terms of the eutectic composition was 8 mol % for the acetone/diethyl ether system and in terms of the eutectic temperature was 5^oC for the acetone/ethanol and benzene/chloroform systems.

The UNIFAC method is thus the most widely used prediction technique for phase equilibria. However, major limitations with the group contribution method are that it cannot differentiate between isomers and the errors increase as the molecule gets more complex. Certain group interaction parameters are also not available which limits the scope of the method.

Although not directly used to predict solubility, Constantinou and Gani (1994) have developed a new group contribution method that uses a set of second-order groups to predict properties of pure organic compounds. These properties could be used to predict solubility using QSPR equations with physicochemical properties as descriptors. Each property has a first-order group contribution and also a second-order group contribution that provides more structural information on the molecular structure of the compound and is able to distinguish between isomers. Properties are classified into primary and secondary properties. The former are estimated only from the molecular structure while the latter involve the primary properties as well as the molecular structure. The authors only considered primary properties in this paper: normal boiling point, normal melting point, critical properties, standard enthalpy of vapourisation at 298K, standard Gibbs energy and standard enthalpy of formation at 298K.

Marrero and Gani (2001) have expanded on the above method by using three 'levels' of estimation. The first level is used to describe simple and monofunctional compounds. The second level involves groups which allow for proximity effects and differentiation between isomers. Aromatic and aliphatic compounds with one ring and several substituents, and polyfunctional compounds may be accurately described at this level. The third level of description gives more structural information about molecular fragments which are insufficient in the previous two levels. Properties of large, complex heterocyclic and acyclic compounds are predicted at this level. Compared with other available group contribution methods used to predict pure component properties e.g. boiling and melting points, critical points, heats of fusion, heats of vapourisation etc, this method gives significant improvements in accuracy and applicability.

Marrero-Morejón and Pardillo-Fontdevila (1999) have proposed a new model for estimating critical constants and boiling points using the group interaction contribution (GIC) previously proposed by Pardillo-Fontdevila and González-Rubio (1997). This method considers the contributions of interactions between bonding groups and properties are defined as functions of structurally dependent parameters. This method has the ability

to distinguish between isomers. The GIC-based models perform well for most chemical classes except for alcohols, phenols and large heterocyclic compounds. Nonetheless, the predictions for these types of molecules have been found to be more accurate than the classic group contribution methods (Joback and Reid, 1987).

2.3.3 Correlations with Parameters from Molecular Structure

There are three types of descriptors available: topological, geometric, and electronic. The topological descriptors include structural information like atom counts, and molecular connectivity indices. The geometric descriptors include molecular volume and surface area. Finally, the electronic descriptors include partial atomic charges, number of electrons and extended Hückel calculations. Obviously the number of descriptors should be kept to a minimum while maintaining the integrity of the equations.

Nelson and Jurs (1994) derived regression models for each set of hydrocarbons, halohydrocarbons, alcohols and ethers, and a fourth model of the combined set. The nine descriptors used in the fourth model correlation were the dipole moment, difference of partial surface areas, fractional partial positive surface area, relative positive charge, relative negative charge, relative negative charged surface area, first moment of inertia (including hydrogens), first geometric moment/second geometric moment (including hydrogens) and valence corrected path-cluster-five molecular connectivity index. According to the authors, the dipole moment of the molecule is an important factor in aqueous solubility as it is a measure of the polarity of the molecule as is the difference between the total positive surface area and the total negative surface area which also accounts for the size of the molecule. The fractional partial positive surface area accounts for both the size and the charge of the compound as it is the total positive surface area times the total positive charge divided by the total surface area. The subsequent three descriptors differentiate between small, polar molecules that tend to be more soluble and large, nonpolar ones. The relative positive charge is the charge on the most positive atom divided by the total positive charge and similarly for the relative negative charge. The relative negative charged surface area is the relative negative charge times the surface area of the most negative atom. The last three descriptors describe the size and shape of the molec-

ular structure with the last descriptor encoding information on large, branched molecules.

A model for heteroatom-containing compounds was developed by Sutter and Jurs (1996) using nine descriptors. The descriptors consist of four topological descriptors: an oxygen atom count, a carbon atom count, a weighted path count based on the number of atoms adjacent to the atom in the path (molecular ID), and the molecular ID divided by the total number of atoms; one electronic descriptor: sum of the partial atomic charges; one geometric descriptor: molecular length divided by the thickness; and three combination ones to account for hydrogen bonding: the negative charged partial surface area divided by the total surface area, the sum of the surface area of hydrogen-bond acceptor atoms (SAAA1), and finally SAAA1 divided by the number of acceptor atoms. The dipole interactions were then accounted for with the number of oxygen atoms, the sum of partial charges and the fractional negative surface area, and the van der Waals forces were described using the number of carbon atoms, molecular ID and the one geometric descriptor.

The aqueous solubilities of compounds with a wide range of structural diversity were derived using six descriptors by Katritzky et al. (1998). The solute-solvent interactions and polarisability effects were accounted for using two descriptors: the most negative partial charge and the total number of electrons. Hydrogen-bond interactions were encoded in the fractional hydrogen donor surface area and the relative negative charge surface area. There is also a corrector descriptor to account for the deficiency of the hydrogen-bonding parameter for the nitrogen-containing compounds and finally the last descriptor contains information on the size, branching and atomic connectivity of the molecule.

McFarland (2001) used an approach that includes partial atomic charge and quantitative estimates of hydrogen-bonding strengths to predict aqueous solubilities of crystalline compounds. Raevsky et al. (1992) have derived hydrogen-bond factors that describe hydrogen-bonding donor and acceptor strengths for different groups and atoms. A total of 19 descriptors determined using a new software package HYBOTPLUS were used.

QSPRs methods are limited by the systems used in the data set as the equations can

only be used for the types of systems they have been fitted for. The structures correlated in the works summarised are currently limited to hydrocarbons, halogenated hydrocarbons and O and N containing compounds.

2.4 Computational Chemistry for Phase Equilibria

Chemistry is traditionally an experimental science. Recent developments however have enabled the study of structures, energies and reactions of chemicals to be done computationally, even for chemicals which may not exist under real conditions. The main approaches of computational chemistry are *ab initio* quantum mechanics which evaluates the interactions between the electrons explicitly, and molecular modelling which studies the molecular interactions using classical mechanics. A more detailed theoretical background on these different techniques is given in standard computational chemistry books (Leach, 1996).

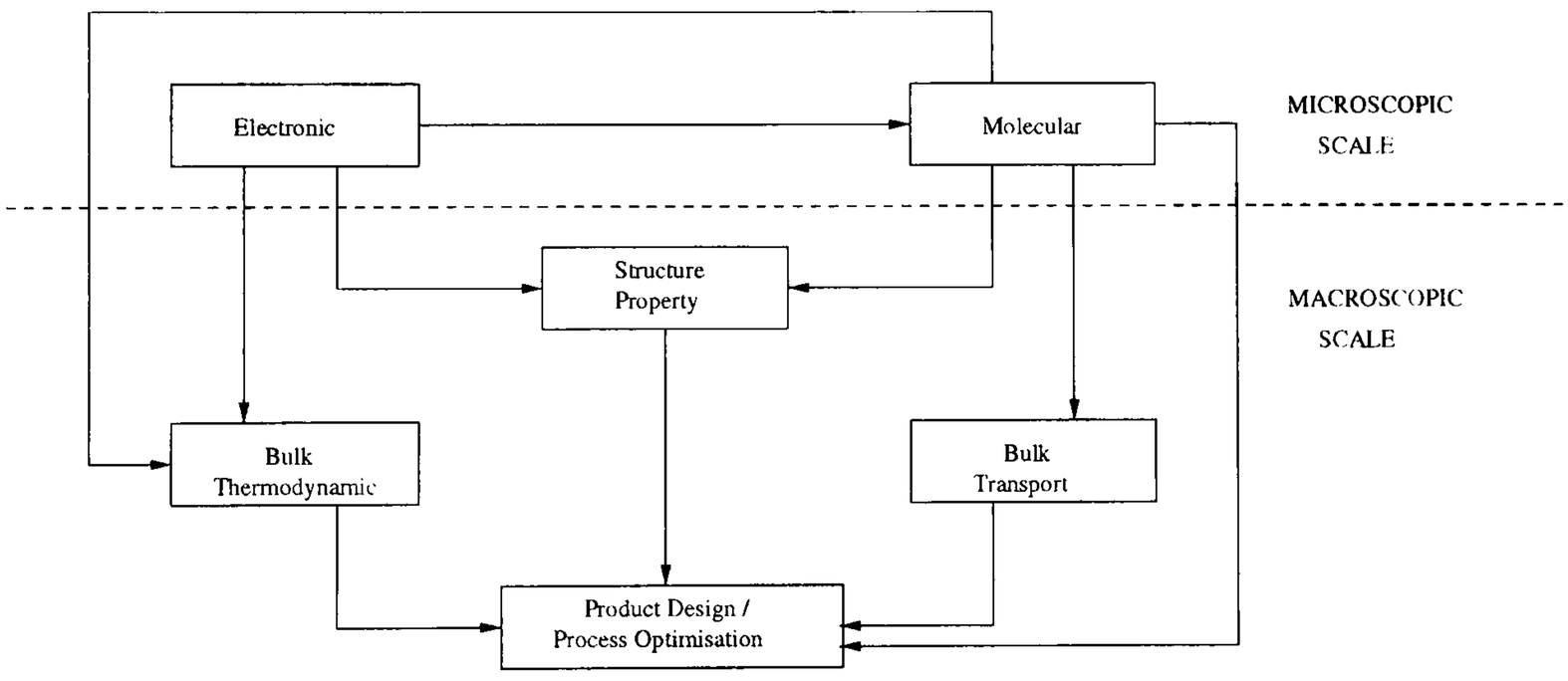


Figure 2.3: Possible relationships between different types of computational chemistry (Gubbins and Quirke, 1996)

The relationships between the different levels of computational chemistry may be described by Fig. 2.3 taken from Gubbins and Quirke (1996). Molecular simulations are an important tool in the understanding of macroscopic phenomena at the microscopic level. Any specific modelling activity may be a combination of various techniques depending on the accuracy and aims of the activity. More extensive information on the behaviour of a molecular system can be obtained by performing a simulation using an ensemble of

molecules. Microscopic level modeling in this case may be used either to predict static structural properties (e.g. interaction energies, binding energies) which are then either substituted into equations of state or to provide an initial configuration of Monte Carlo (MC) or Molecular Dynamics (MD) simulation of an ensemble of molecules.

MC techniques can only be used to predict equilibrium properties and attempt to find the equilibrium state by using a random number generator to displace the molecules randomly. For a fixed temperature T, volume V, and number of molecules N, the probability of finding the system in a particular configuration is proportional to $\exp(-U/k_bT)$, where U is the potential energy of the configuration and k_b is the Boltzmann constant. The simple Monte Carlo integration approach is usually performed in the canonical (N, V, T) ensemble and involves the following steps:

1. Obtain a configuration of the system by randomly generating $3N$ Cartesian coordinates for the particles.

2. Calculate the potential energy of a random particle in the configuration, $U_i(\mathbf{r}^N)$ and using the potential energy, determine the Boltzmann factor, $\exp(\frac{-U_i}{k_bT})$.

3. Calculate the instantaneous system properties (e.g. temperature, pressure) and add their contribution to the accumulating sum and return to step 1.

When the properties start fluctuating around a mean equilibrium value after a number of iterations, N_{trial}, the system has then reached equilibrium. The properties are then calculated using:

$$\langle U(\mathbf{r^N})\rangle = \frac{\sum_{i=1}^{N_{trial}} U_i(\mathbf{r^N}) \exp[-U_i(\mathbf{r^N})/k_bT]}{\sum_{i=1}^{N_{trial}} \exp[-U_i(\mathbf{r^N})/k_bT]} \tag{2.25}$$

MD generates successive configurations of the system by integrating Newton's laws of motions and thus may be used for the prediction of both equilibrium and non-equilibrium properties. Molecular dynamics simulations are usually performed in the microcanonical ensemble (N, V, E) where E is the total energy of the system and the steps are as follows:

1. Obtain a configuration of the system by randomly generating $3N$ Cartesian coordinates and velocities to the particles.

2. Calculate the potential energy for each particle of the configuration, $U(\underline{\mathrm{r}}^N)$.

3. Determine the force $\underline{\mathrm{f}}_i$ acting on each particle as the gradient of the potential energy field.

4. The subsequent new positions and velocities are obtained by solving the equations of motions defined by Newton's second law :

$$\underline{\mathrm{f}}_i = m\ddot{\mathbf{r}}_i, i = 1, ..., N.$$

5. The instantaneous system properties are then calculated and added to the accumulating sums. The system property, A, is defined as the average of the N_t state points v obtained along the system trajectory after equilibrium has been reached :

$$A = \frac{1}{N_t} \sum_{v=1}^{N_t} A_v \tag{2.26}$$

2.4.1 Monte Carlo and Molecular Dynamics Simulations

The first Monte Carlo (MC) simulation was carried out almost forty years ago by Metropolis et al. (1953) on a system of hard disks. Since then, molecular simulation techniques have rapidly evolved. The Gibbs ensemble Monte Carlo (Panagiotopoulos, 1987; Smit et al., 1989) is used widely to predict phase coexistence properties of complex fluids, while configurational-bias sampling (Siepmann and Frenkel, 1992; De Pablo et al., 1992) and thermodynamic scaling techniques (Kiyohara et al., 1996; Valleau, 1999) have broadened the range of systems which may be studied using these simulation techniques. These simulation methods however are currently limited by the reliability of the intermolecular potentials used.

Potoff et al. (1999) tested the predictive capabilities of several intermolecular potential models for mixtures of polar and non-polar components using Monte Carlo. These intermolecular potentials have be shown to accurately predict the phase behaviour of

pure components. The authors compared the Lorenz-Berthelot combining rules (Maitland et al., 1981) with the Kong rules (Kong, 1973) for the unlike pair potentials. The former usually results in more attractive unlike interactions compared with the latter. As a result, although the latter was better for nonpolar systems, the former was found to better predict phase behaviour for polar systems although often only qualitatively as the stronger attractive interactions accounted for the neglect of polarisation interactions between the compounds in the mixture.

Hloucha et al. (2000) state that as these intermolecular potentials are fitted to reproduce some experimental data, these models are not comprehensive. The authors address the suitability of *ab initio* pair interaction energy potentials for phase property prediction. The two-body intermolecular potentials were derived from the work by Bukowski et al. (1999) which used symmetry-adapted perturbation theory (SAPT) (Jeziorski et al., 1994) to compute pairwise interaction energies for various compounds. Hloucha et al. (2000) refitted the interaction energies to Lennard-Jones and modified Buckingham plus Coulomb site-site potentials for their Monte Carlo simulation of the pure component phase behaviour of acetonitrile and methanol. They found that the pair interaction potential was sufficient to describe the behaviour of acetonitrile but not for methanol. This was due to the slight inaccuracy of the potential energy surface due to level of the basis set used (Sum et al., 2002). Better prediction was achieved by scaling the interaction energies and incorporating multibody effects using a classical polarisable model. Sum et al. (2002) then extended the study to include the study of the phase behaviour of the acetonitrile and methanol mixture. This study was the first prediction of the phase behaviour of mixtures using *ab initio* potentials without the use of any combining rule.

Another limitation of using molecular simulations, aside from the difficulty in obtaining intermolecular potentials, is the high computing power required to for each run. Currently research has concentrated on aqueous solubility. However, current computing power is not sufficient to allow the comparison of the solubilities of a solute in different solvents in reasonable time.

2.4.2 Deriving Parameters for QSPRs

Molecular simulation can also be used to obtain fitting descriptors for QSPRs instead of using experimental data. Jorgensen and Duffy (2000) ran Monte Carlo simulations for 150 drug solutes in water to determine configurationally averaged results for five significant descriptors which are the Lennard-Jones interaction energy, number of hydrogen-bond acceptors, Coulomb tension which is a function of the Coulomb interaction energy (ESXC) and the solvent-accessible surface area (SASA), ESXC/SASA, the number of non-conjugated amine groups and the total number of nitro and carboxylic acid groups. The last two were corrections for the deficiencies in the charge distributions for these functional groups. The predictions obtained by this technique have a correlation coefficient of 0.9 and a rms error of 0.7.

Wang et al. (2001) developed a solvation model to predict solvation free energies based on weighted solvent-accessible surface area (WSAS).

$$\triangle G_{WSAS} = \sum_{i=0}^{m} \sum_{j=0}^{n_i} w_i S_j \tag{2.27}$$

where m is the number of the atom types for a given model, W_i is the solvation free energy weight of atom type i, n_i are the number of atoms with atom type i in a molecule, and S_j is the solvent accessible surface area (SASA) of atom j. The SASA calculation was carried out on the global minimum optimised conformation using the MMFF94 force field with MMFF charges (Halgen, 1996). Using a 401-molecule set to reproduce experimental solvation free energies in water, the rms error was 0.79 kcal/mol.

Instead of running simulations for large systems of molecules, Klamt et al. (1998) and Klamt and Eckert (2000) combine a dielectric continuum solvation model with a statistical thermodynamics of solute-solvent interactions (COSMO-RS). The model has been parameterised for the elements H, C, N, O, F, S, Cl, Br and I based on density functional theory. The parameters are a cavity radius and a dispersion constant for each element, and eight general parameters: an averaging radius for the screening charge density r_{av}, an effective contact area a_{eff}, the electrostatic interaction coefficient α', two hydrogen-

bonding parameters, a ring correction, a degeneracy difference between gas phase and liquid state, and a size dependence coefficient. The model allows for the calculation of the chemical potential differences to an accuracy of 0.4 (rms) kcal/mol.

The dielectric continuum solvation model uses the boundary condition for the total potential to calculate the ideal screening charges $q*$ for a cavity of solute X when embedded in a conductor. The ideal screening charges are then scaled to reflect the dielectric permeability ϵ using:

$$q = \frac{\epsilon - 1}{\epsilon + 0.5} q*$$

The surface of the molecular cavity is discretised into surface segments, each with the area of the parameterised effective contact area and with a corresponding local screening charge density, σ where $q_i = a_i \sigma_i$. The latter is a good descriptor of the local polarity of the molecular surface. The mean screening charge densities are obtained by averaging the original screening charge density over a region with the parameterised averaging radius. Each surface segment is then formed into pairs to reflect the interaction between two neighbouring segments. Molecular interactions are then considered as pairwise interactions of the molecular surface segments compared with UNIFAC where the interactions are represented by pairwise group parameters. The misfit energy $E_{misfit}(\sigma, \sigma')$ between the two interacting surfaces is:

$$E_{misfit}(\sigma, \sigma') = a_{eff} e_{misfit}(\sigma, \sigma') = a_{eff} \frac{\alpha'}{2} (\sigma + \sigma')^2 \quad (2.28)$$

where $e_{misfit}(\sigma, \sigma')$ is the misfit energy density and α' is the electrostatic interaction coefficient. The hydrogen bond energy between the interaction surfaces are parameterised additionally using:

$$E_{hb}(\sigma, \sigma') = a_{eff} e_{hb}(\sigma, \sigma') = a_{eff} c_{hb} \min(0, \min(0, \sigma_{don} + \sigma_{hb}) \max(0, \sigma_{acc} - \sigma_{hb})) \quad (2.29)$$

where $\sigma_{don} = \min(\sigma, \sigma')$ and $\sigma_{acc} = \max(\sigma, \sigma')$. The other two parameters, σ_{hb} and c_{hb}. are a threshold value for hydrogen bonding and a strength coefficient respectively, both of

which have to be parameterised. The van der Waals interactions are represented by the element specific dispersion constant. The chemical potential of a solute X in a solvent S. μ_S^X, is similarly represented as a residual part, $\mu_{S,res}^X$, and a combinatorial part, $\mu_{S,comb}^X$. The residual part represents the interactions of the surfaces of the liquid and is determined by the statistical averaging of the surface interactions.

$$\mu_S^X = \int \mu_S(\sigma) p^X(\sigma) d\sigma + \mu_{S,comb}^X \tag{2.30}$$

where $p^X(\sigma)$ describes the amount of surface of the pure compound X having a screening charge density between σ and $\sigma + \mathrm{d}\sigma$. The function $\mu_S(\sigma)$ is given by:

$$\mu_S(\sigma) = -RT \ln \left[\int d\sigma' p_S(\sigma') \exp\left(\frac{\mu_S(\sigma') - a_{eff} e(\sigma, \sigma')}{RT} \right) \right] \tag{2.31}$$

which has to be solved recursively. The combinatorial part arises from the difference in sizes of the solute and solvent molecules and is dependent on the size dependence coefficient. Currently the method has been used for vapour-liquid and liquid-liquid equilibria and has recently been extended for aqueous solid-liquid equilibria using COSMO-RSol (Klamt et al., 2002) . Here the solubility of a solid compound X in a solvent S, S_S^X is given by:

$$\log S_S^X = \log \left(\frac{MW^X \rho^S}{MW_S} \right) + \frac{\ln(10)}{kT} [-\triangle_S^X + \min(0, \triangle G_{fus}^X)] \tag{2.32}$$

where $\triangle_S^X = \mu_S^X$ - μ_X^X. The Gibbs free energy of fusion, $\triangle G_{fus}^X$, of a solute X is found using QSPR technique using three descriptors: the cavity volume which accounts for the molecular size, the number of ring atoms which describes the rigidity, and the chemical potential of the compound X in water, μ_W^X, which accounts for the polarity and hydrogen-bonding.

2.4.3 Deriving Equation-of-state (EOS) parameters

Fermeglia and Pricl (2001) presented two procedures of estimating EOS parameters using first, DMol^3/COSMO-RS calculations and secondly, molecular mechanics/dynamics simulations. The EOS used was the simplified PHSCT EOS (Song and Mason, 1991; Song

et al., 1996) perturbed hard-sphere chain. The parameters were A*, V*, and E* which are the molecular surface areas and volumes, and an energetic cohesive parameter respectively. For mixtures, an additional binary parameter k_{ij} is needed. For the first method using DMol^3/COSMO-RS, the parameters A* and V* were obtained directly. The third parameter E* was found by inserting the corresponding experimental vapour-pressure into the EOS expression for pure components together with the A* and V* parameters. The binary iteraction parameters were regressed from the activity coefficients obtained from the binary system and the EOS using the pure-component parameter sets. In the second method, again the first two parameters may be obtained directly. The energetic parameter can be calculated either as in the first method, or from the molecular dynamics simulation. The binary interaction parameters k_{ij} are found by performing NPT simulations of equimolar mixtures of each binary in the gas phase. Both methods were used to predict phase equilibria for 11 binary systems of which the DMol^3/COSMO-RS method showed satisfactory results except for two systems (however, this was attributed to the weakness of the EOS for polar/nonpolar and associating systems) and the second method gave slightly better results.

2.4.4 Deriving Interaction Parameters for Activity Coefficient Models

In the development of activity coefficient models as those described in Section 2.2.2, the adjustable binary parameters are interpreted as interaction energies between molecular pairs. If these models are theoretically sound then the energies could be calculated directly. The interaction parameter, A_{ji}, and UNIQUAC binary parameter, τ_{ji} are calculated as follows:

$$A_{ji} = \frac{\triangle U_{ji} - \triangle U_{ii}}{Rq_i} \tag{2.33}$$

$$\tau_{ji} = \exp\left(\frac{-A_{ji}}{T}\right) \tag{2.34}$$

where $\triangle U_{ji} = U_{ji} - (U_j + U_i)$, R is the gas constant, T is the temperature and q_i is the relative molecular surface area of molecule i.

In actuality, the parameters in the UNIQUAC equation may not necessarily represent

the difference in interaction energies between the molecules. The equations described in Section 2.2.2 are all based on the local composition concept. A physical model of a liquid is represented by a three-dimensional fully occupied lattice with a coordination number independent of concentration. However, Flemr (1976) has proved that the theory is inconsistent for any nonrandom mixture and the probability of finding a molecule i around a central molecule j is dependent on the mutual attraction u_{ij} and the interaction of molecule i with all its neighbours. McDermott and Ashton (1977) showed that the local composition should be consistent with the overall composition of the mixture. which was a point not considered by the authors of the above equations. On this basis, local composition equations should only be regarded as semi-empirical ones. In the original UNIQUAC formulation, the optimum binary parameters were obtained through a fitting technique using experimental results and are not derived from any equation nor do they represent any calculated value. Nevertheless, it is worthwhile investigating whether the predictions obtained by deriving UNIQUAC interaction parameters from interaction energies are sufficiently accurate or whether taking just the interaction energies into account would be sufficient in ranking the solubilities. The former will give quantitative results while the latter will only be valid qualitatively.

Meniai and Newsham (1992, 1995) have calculated the molecular interaction energies using a molecular graphics system, CHEM-X [1] to simulate liquid structures using Monte Carlo techniques for infinitely dilute systems. In the first paper (Meniai and Newsham, 1992), 64 solvent molecules surrounding one solute molecule are spaced to give the appropriate liquid density of the system. The numbers are chosen to keep the computational time within reasonable limits. A solvent molecule is then randomly displaced and the perturbation may be accepted or rejected depending on whether it passes a certain criterion. The authors investigate a variety of similar solvents interacting with heptane/toluene and hexane/benzene and find the results of the selectivity, S_{ij}^{∞}, useful in screening suitable solvents. Selectivity is defined as the ratio of the infinite dilution activity coefficients of components, i and j, defined as γ_i^{∞} and γ_j^{∞} respectively in the solvent-rich phase. However, for solutes and solvents differing in their chemical nature such as phenol/water

[1] Chemical Design Ltd., Oxford, England

separation, results are only useful qualitatively.

In the second paper (Meniai and Newsham, 1995), a faster method is proposed where only two molecules are used and the dimensions of a simple cube are calculated according to the experimental density of the system. The center of mass of one molecule is positioned at the center of the cube and the other at one of its corners. While one molecule is kept fixed, the other is rotated and shifted until the minimum energy configuration is reached. The difference in energies between rotating one molecule and both molecules is small. As a result, the former approach is used as it is less computationally intensive. Both molecular and group interaction parameters are calculated for phenol/water separation by different solvents. The agreement with experimental results reported in Magnussen et al. (1981) is reasonable, especially for mole fractions of the solute greater than 0.5. The average mean standard deviation for UNIQUAC parameters for various phenol/solvent systems is 0.097 while for group interaction parameters, it is 0.164.

Jónsdóttir et al. (1994) have calculated UNIQUAC interaction parameters for alkane/alkane and alkane/ketone systems using a Molecular Mechanics method, the Consistent Force Field (CFF) (Rasmussen et al., 1985; Rasmussen et al., 1993). The minimal potential energies for all possible combinations of conformers of the solvent-solvent, solute-solute and solvent-solute complexes are determined and used to calculate the average interaction parameters from the Boltzmann distribution. In order to compute the interaction energy between two molecules, they are located next to each other and their energy is computed with the CFF force-field and minimised using the Boltzmann Jump Procedure (BJP). At each iteration of the BJP, one of the molecules is rotated and translated, and the new configuration is accepted with a probability based on the Boltzmann factor. They obtained good agreement with experimental data from the Dortmund Data Bank for these few systems with a mean relative deviation of less than 5%. The only exception is the c-hexane/ c-hexanone system involving cyclic molecules which are more difficult to simulate. Jónsdóttir et al. (1996) extended the method to include polar systems with extensive hydrogen bonding, i.e. alkane/primary amines systems, using the Boltzmann Jump search procedure to investigate the conformational space. The average deviation of

the vapour pressure was within 5% of the experimental data except for the ethylamine/n-hexane system.

Sum and Sandler (1999) have used *ab initio* methods to predict phase equilibrium for binary systems of water with various organic compounds using activity coefficients models such as UNIQUAC and Wilson (Wilson, 1964). High-pressure VLE predictions were also conducted using the Peng-Robinson-Stryjek-Vera (PRSV) equation of state (Stryjek and Vera, 1986) and the Wong-Sandler mixing rule (Wong and Sandler, 1992). The energy of a cluster of eight molecules was first minimised using the PM3 semi-empirical method and then minimised again using the Hartree-Fock method. The interaction energies between the clustered molecular pairs for methanol/water, ethanol/water, acetic acid/water and formic acid/water mixtures were then determined from directly interacting molecular pairs. Some of the assumptions made in the calculations included neglecting the electron correlation, taking a cluster size of eight molecules to represent the system and calculating the interaction energies at 0K to predict phase equilibrium at other temperatures. The values predicted using the UNIQUAC method were in excellent agreement with experimental data and were better than with the Wilson equation.

In summary, these methods of estimating activity coefficients using interaction energies have yielded some promising results based on the assumption that the binary parameters are interaction energy parameters. Only simple molecules have been successfully modelled using this technique and our aim is to expand this method to include more complex molecules, leading up to pharmaceuticals. The force fields are less expensive computationally than quantum mechanical models and significant research is currently being done to improved their accuracy. It was thus decided that using force fields would be suitable for our requirements. Therefore, a summary of the theory and a description of available force-fields is provided in the next section.

2.5 Molecular Mechanics

Molecular mechanics ignores the electronic motions and calculates the energy of a system from the nuclei only using force fields. A force-field is a combination of the potential energy functions describing intra- and interatomic interactions together with the corresponding parameter set. The equilibrium molecular structures are determined through a minimisation of the total energy of the system.

Force fields are based on a classical representation of the energy of molecular systems. Neglecting electronic motion, the energy of the system may be calculated as a function of the nuclear coordinates only. Force fields also work from the assumption of transferability of force field parameters. This important assumption decreases the numbers of parameters needed significantly. Many force fields can only be used specifically on the molecules which have been used to parameterise them. It is therefore important to know the purpose which the force field was developed for before deciding to use it.

The potential energy of the system, E, is defined as a function of the positions, r, of N particles:

$$E(r^N) = \sum_{bonds} E_b + \sum_{angles} E_\theta + \sum_{torsions} E_\phi + \sum_{nonbonded} E_r \tag{2.35}$$

where E_b are bond interactions, E_θ are angle interactions, E_ϕ are torsional interactions and E_r are nonbonded interactions. The different terms are explained in Fig. 2.4. The ideal force-field will be able to predict accurately the conformational energies and the interaction energies of the molecules without incurring high computational costs.

2.5.1 Bond stretching

The typical form of equation used to express the bond stretch term as a function of bond length, l, is the Morse potential:

$$E_b = D_e(1 - \exp[-a(b - b_{eq})])^2 \tag{2.36}$$

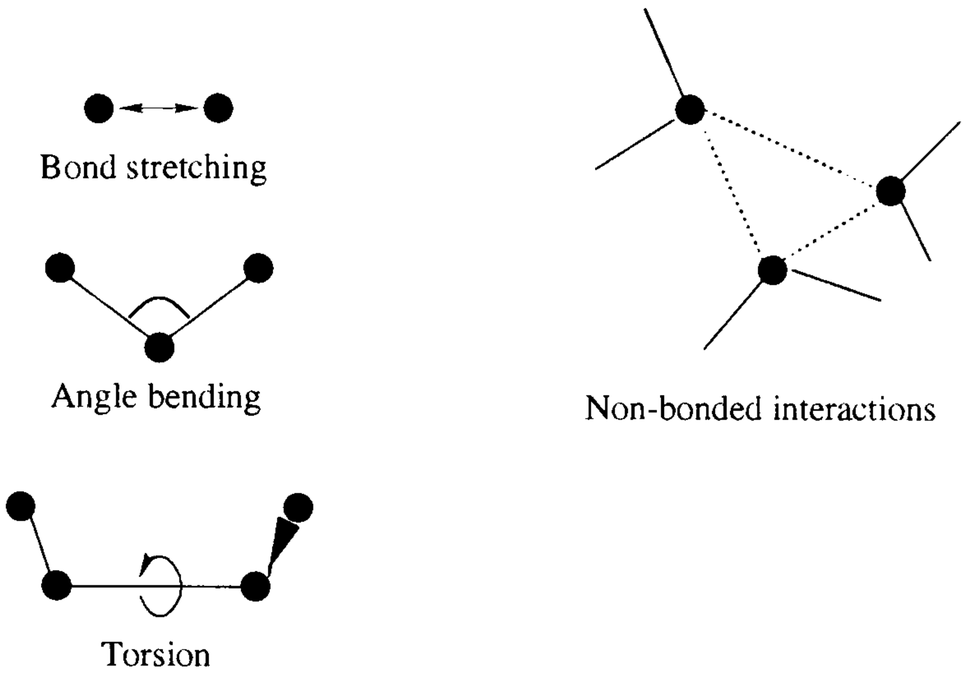


Figure 2.4: Schematic representation of the four key contributions to a molecular mechanics force field: bond stretching, angle bending, torsional terms and non-bonded interactions

D_e is the depth of the potential energy minimum and a is a parameter controlling the width of the potential well. b_{eq} is the reference length of the bond. The Morse potential however does require three parameter values and the evaluation of an exponential term is not computationally efficient. As a result, simpler forms of equations were developed. The simplest form is based on the Hooke's law formula:

$$E_b = K_b(b - b_{eq})^2 \tag{2.37}$$

As the Hooke's law formula is less accurate away from equilibrium, cubic and higher terms may be included to more accurately model the Morse curve:

$$E_b = K_b(b - b_{eq})^2[1 - 2K_b'(b - b_{eq}) - 2K_b''(b - b_{eq})^2 - 2K_b'''(b - b_{eq})^3 ...] \tag{2.38}$$

2.5.2 Angle Bending

Hooke's law is also frequently used to describe the deviation of bond angles, θ, from their reference values, θ_{eq}.

$$E_\theta = K_\theta(\theta - \theta_{eq})^2 \tag{2.39}$$

The accuracy of the force field may be improved by incorporating higher terms in the above equation as with the bond stretching terms.

2.5.3 Torsional terms

The torsional potential models how the energy changes as a bond rotates. There exist energy barriers towards bond rotation which are essential in the study of conformational analysis. Torsional potentials are used to reproduce the preference for staggered conformations about $sp^3 - sp^3$ bonds and the preference for eclipsed conformations about $sp^2 - sp^3$ bonds. Torsional potentials are usually expressed as a cosine series expansion.

$$E_{\phi 1} = \sum_{n=1}^{N} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \tag{2.40}$$

The parameter V_n is often referred to as 'barrier' height and it gives a qualitative indication of the relative barriers to rotation. γ is the phase factor which determines where the torsional angle, ϕ, passes through the minimum energy value.

2.5.4 Out-of-plane angles

The *out-of-plane* bending term is necessary to keep π-bonded atoms on the same plane as their neighbouring atoms as the π bond is maximised in the coplanar arrangement and the energy would be reduced if the bond was bent out of plane. One way of defining the *out-of-plane* bending term is to treat four connected atoms as an 'improper' torsion angle:

$$E_{\phi 2} = K_{\phi}(1 - \cos 2\phi) \tag{2.41}$$

The total torsional energy is then a sum of the 'proper' and the 'improper' torsional angles energy terms.

$$E_{\phi} = E_{\phi 1} + E_{\phi 2} \tag{2.42}$$

2.5.5 Electrostatic interactions

Molecules and atoms interact with each other through non-bonded forces which comprise van der Waals interactions and electrostatic interactions. The electrostatic charge plays an important part in the potential energy of a molecule and is one of the most difficult to define accurately and has been a topic of debate among specialists. The most commonly

used model is the simple Coulombic isotropic atom-atom potential. The electrostatic interaction between two molecules is calculated as a sum of interactions between pairs of point charges:

$$E_{r_{ij}} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.43}$$

where r_{ij} is the distance between atoms i and j, and ϵ_0 is the permittivity of vacuum. The atomic charges q_i are treated as fractional charges and usually each atom type is allocated a fixed charge. The charges can also be derived from the wavefunctions by computing electronic distributions quantum mechanically if experimental data are not available. Each atomic orbital is represented by a basis set composed of a series of basis functions. The electrostatic interactions can be represented by sets of atomic multipoles. The agreement between theory and experiment improves with the size of the basis set and whether electron correlations are included. The method of representing atomic charges must be able to include effects of solvent and other particles on the charges in the system. The most well-known method for assigning these charges is the Mulliken population analysis (Mulliken, 1955) but the results vary according to the quantum mechanical calculations level and the basis set chosen. A major limitation in the accuracy of predictions based on population analysis arises from the use of equal partitioning of the overlap charge between unequal atoms.

Gasteiger and Marsili (1980) presented a method to calculate atomic charges through an iterative partial equalisation of orbital electronegativity and considering only the connectivities of the atoms. The procedure effects a damping of the electrostatic field before total equalisation of electronegativity is reached. This approach provides rapid calculation for σ-bonded and nonconjugated π-systems.

Sometimes it is necessary to describe the electron distribution at all distances and in a detailed manner. The methods described above assume nuclear-centered charges. However this is not the case for lone pairs and π-electron systems. Stone and Alderton (1985) developed a distributed multipole analysis (DMA) to derive the charge distribution from

an *ab initio* wavefunction. This method may be used to facilitate the study of the effect of electron correlation on the density distribution and used to construct electrostatic models.

Many applications utilise point electrostatic charges for each nuclear position. However, atomic charges are usually determined at a single geometry but assumed to be valid over a range of conformations. Reynolds et al. (1992) developed two methods which generate atomic charges for each conformation by minimising the difference between quantum mechanical and classical molecular electrostatic potential (MEP) subject to the constraint that the sum of the atomic charges gives the total molecular charge. The first method determines the MEP for one conformation and constrains the charges to reproduce the dipole at an alternate geometry. The second method calculates the MEP for each conformation and weighs the MEP according to the Boltzmann factor. The latter method shows an improvement over the former and is useful especially in molecular dynamics simulations.

Price (1996) recommends representing the electronic distributions as accurately as possible by incorporating lone pairs and π bonds as it does make a significant difference in modeling molecules. The DMA method gives a better prediction of electrostatic distribution compared with other methods. However, the necessity of using anisotropic potentials will depend on the accuracy of the calculations needed as it results in an increase in computational expense.

2.5.6 van der Waals interactions

The most well-known van der Waals potential function is the Lennard-Jones 12-6 function:

$$E_{r_{ij}} = 4\epsilon[(\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}] \tag{2.44}$$

The above potential has two adjustable parameters. σ is the collision diameter for which the energy is zero and ϵ is the well depth.

Another frequently used van der Waals potential function is the Buckingham potential:

$$E_{r_{ij}} = \epsilon \left[\frac{6}{(\alpha - 6)} e^{-\alpha(\frac{r_{ij}}{r_m} - 1)} - \frac{\alpha}{(\alpha - 6)} (\frac{r_m}{r_{ij}})^6 \right] \tag{2.45}$$

r_m is the separation when the energy passes through a minimum and α is a positive scalar. When α is between approximately 14 and 15, the above energy potential models the Lennard-Jones 12-6 potential in the minimum energy region.

2.5.7 Force field parameterisation

The parameterisation of a force field is critical to its performance. The first step is to identify the data needed for the parameterisation process according to the use of the force field. The different conformational energies with their corresponding geometries and vibrational frequencies are usually included in the parameter set. Aside from experimental data, computer simulation techniques, such as *ab initio* calculations, have also been used to parameterise some force fields. Due to the lack of experimental data for particular classes of molecules, quantum mechanics calculations are increasingly used to provide the required data. A liability of using *ab initio* calculations is the lack of mean field effects which are naturally included in the experimental data.

Two ways of obtaining the parameters are 'parameterisation by trial and error' and by least-squares fitting. In the trial and error procedure, the method is basically manual and iterative in nature and requires gradual refinement of the parameters. The latter procedure enables the use of weighting factors for different experimental data which are then fitted and also allows the use of algorithms in minimising the sum of squares of the differences between the observed and the calculated values for the properties.

Transferability of the parameters also has significant consequences on the reliability of the force fields. If there are parameters missing for a particular molecule, those parameters may be 'guessed' or ignored. It is important to check that the results are accurate if this does happen and that the errors are balanced.

2.5.8 Description of available force fields

There are many levels of force fields available. Force fields differ in the choice of functions used for each term in the potential energy function and in the manner in which the parameters are determined. A detailed discussion on recent force field development is given by Rasmussen (1997). The first-generation, or classical, force fields were originally created for modeling proteins and peptides (Brooks et al., 1983; Weiner et al., 1984). The equations mainly used to describe the bond length and angle terms are quadratic, the torsion angle is a single-cosine function and the nonbonded terms are the Lennard-Jones and the Coulombic potential functions respectively. The second-generation force fields model the terms using more complex equations which more accurately describe the potential energy and include parameters for additional functional groups. A brief description of the best known force fields is found below.

The CHARMm force-field (Chemistry at Harvard Molecular Modelling) is a classical force field (Brooks et al., 1983). It was developed initially for biological molecules such as proteins, nucleic acids and carbohydrates. The force field includes alcohols, aldehydes, ketones, amides and phosphates in the parameter set and was parameterised on the basis of *ab initio* energies and geometries of small molecules. It performs well over a broad range of calculations and simulations and is designed to give good results for a wide variety of systems, from small isolated molecules to solvated complexes of large biological macromolecules. It is however not applicable to organometallic complexes. The electrostatic term can be scaled to mimic solvent effects and assigns fractional charges using the Gasteiger method of using iterative partial equalisation of orbital electronegativity which only considers the properties of the constituent atoms of a molecule and the network of bonds (Gasteiger and Marsili, 1980).

The AMBER force-field (Weiner et al., 1984) was also developed for nucleic acids and proteins. It is also a classical force-field. It was parameterised from experimental data against a limited number of organic fragments of proteins and nucleic acids. AMBER-95 uses an advanced potential-derived fractional charge model called RESP (restrained electrostatic potential fit) which evaluates the electrostatic potential over a range of inter-atomic dis-

tances using *ab initio* calculations followed by the fitting of the fractional charges to reproduce the *ab initio* results.

The MM2 (Allinger, 1977) and MM3 (Allinger et al., 1989) force-fields were developed by Allinger and co-workers for small organic molecules. The force-fields can calculate enthalpies of formation accurately and MM3 especially was developed to calculate gas phase structures to good accuracy and also vibrational frequencies of organic substances. MM3 uses the Buckingham potential and a bond dipole potential for treating the electrostatic potential which is quite different from most force fields as they tend to use Lennard Jones and Coulomb potentials respectively.

The Consistent Force-Field (CFF) was developed by Rasmussen et al. (1993). The parameters were optimised simultaneously using experimental data derived from x-ray, neutron and electron diffraction, as well as from vibrational spectra (IR, Raman and neutron scattering). The force field may be used for single molecules, molecular complexes and molecular and ionic crystals. It is however not suitable for larger biological structures. The parameter sets, also known as the PEF series, available currently are for alkanes, ketones, aldehydes, primary amines, ethers, alcohols, carbohydrates and glycolipids.

The CFF91, CFF95 and PCFF force fields (Hagler, 1988; Hagler, 1994) are *ab initio* force fields because their parameters were derived based on *ab initio* data. These force fields are second generation force fields. The choice of potential functions is similar to the MM3 force-field and introduces atomic dipoles and charge flux in order to perform better evaluations on fractional charges and dipole moments. CFF91 is applicable to hydrocarbons, proteins, protein-ligand interactions and is parameterised explicitly for acetals, acids, alcohols, alkanes, alkenes, amides, aromatics, ethers and esters. CFF95 has slightly more atom types than CFF91 and is parameterised for additional functional groups. It may be used for organic polymers such as polycarbonates and polysaccharides. PCFF is based on CFF91 with extended coverage of organic polymers, (inorganic) metals and zeolites.

BBL
LONDEL
UNIV.

Although force fields with parameters derived from *ab initio* calculations perform reasonably well in various aspects, it has been shown that these parameters are not suitable for molecular dynamics simulations at finite temperatures (Sun and Rigby, 1997). The reason behind this is that the parameters were developed based on static simulations corresponding to a classical state at 0K but the experimental data used to determine the parameters were measured at finite temperatures. Additionally, good agreement with experimental data can only be achieved when the experimental conditions at the time of measurement closely approximate the conditions under which the parameterisation data was collected.

The COMPASS force field (Condensed-phase Optimised Molecular Potentials for Atomistic Simulation Studies) (Sun, 1998) was developed to counter this problem, while preserving all the benefits of the CFF force fields, and it was extended to cover and consolidate most common organic and inorganic molecules. It is parameterised using condensed-phase properties in addition to *ab initio* and empirical data (to yield good agreement with experimental data) for molecules in isolation.

Jorgensen et al. (1984), Jorgensen and Tirado-Rives (1988) and Jorgensen et al. (1996) developed the OPLS force field (Optimised Potentials for Liquid Simulations) to reproduce experimental data on fluids. The optimised intermolecular potential functions are determined using Monte Carlo simulations. At the moment, the procedure is costly and requires a number of simulations but has good accuracy. The simulations give the conformer populations for each liquid along with thermodynamic and energy distributions. The parameterisation requires numerous Molecular Dynamics or Monte Carlo simulations using different sets of parameters to yield accurate thermodynamic and structural results for liquids. Torsional potential parameters are obtained by fitting to conformational energy profiles calculated using *ab initio* RHF/6-31G*. Intermolecular interactions are represented by interaction sites usually located on the nuclei. Bond stretching and angle bending parameters are derived from the AMBER all-atom force field.

Jónsdóttir et al. (1999) stressed the importance of choosing the correct force field for

a particular purpose. They calculated the UNIQUAC interaction parameters for several systems i.e., n-butane / ethylamine, acetone / cyclohexane, ethanol / water and methanol / water systems using CFF, CHARMm and AMBER. Only the PEF series gave good comparison with experimental data for the n-butane / ethylamine and acetone / cyclohexane systems. In the first case, this is due to the fact that the parameters for nitrogen in the CHARMm and AMBER force fields are based on amino acids rather than amines. In the second case however, as the ketone group was included in the CHARMm force field, the non-bonded interactions do not appear to be modeled to good accuracy. On the other hand, the CHARMm force field gave good results for interactions with water.

Halgren (1999) compared several force fields in their ability to predict conformational energies and intermolecular interactions. The force fields were MMFF94 and MMFF94s (Halgren, 1996); CFF95 and CVFF (accessed through Version 3.0 of Cerius^2 [2]); MSI CHARMm (accessed through Quanta 97); AMBER*, OPLS*, MM2* and MM3* force fields (accessed through Version 5.5 of BatchMin [3]). The MMFF94 and MMFF94s force fields performed well for both tests. CFF95 had large errors for conformational energies and the author was concerned about the force field's parameters being derived in a variety of ways from different types of data. MM3* performed well for conformational energies but is not suitable for condensed-phase simulations in high-dielectric media. MM2*, which also gives low interaction energies, and produces large errors for conformational energies. AMBER* was found to be unsuitable for conformational energies but gives good values for intermolecular interactions. CVFF and OPLS* performed poorly in both areas. MSI CHARMm was comparable to CFF95 in terms of calculating conformational energies but produces large errors for intermolecular interaction energies.

Two other types of force fields are the broadly applicable force fields which aim to cover the range of the periodic table, for example Dreiding (Mayo, 1990) and Universal (Rappe,

[2]Cerius^2, QUANTA 97 and DISCOVER are available from Molecular Simulations, Inc, 9685 Scranton Road, San Diego, CA 92121-2777

[3]BatchMin and MacroModel were developed in the laboratories of Professor W. Clark Still and are available from Schrodinger, Inc. (Portland, OR). Version 5.5 was released in 1996. The MM2*, MM3*, AMBER* and OPLS* force fields are modifications developed by the Stil group of the parent MM2, MM3, AMBER and OPLS force fields

1992) and the special purpose force fields, such as the Glass force field (Soules, 1979; Soules and Varshneya, 1981) and Zeolite force field (van Beest et al., 1990; Burchart, 1992). The former rely on atomic parameters coupled with theoretically and empirically derived rules for generating explicit force field parameters while the latter were parameterised for more specialised work e.g. glass and zeolites.

Force-fields yield accurate results only for those molecular classes for which they have been explicitly parameterised. For other functional groups not in the parameterisation list, the explicit parameters may be used for similar chemical structures based on the assumption of transferability. Comparisons of the various force fields showing the different types of equations used are shown in *Appendix B*.

2.5.9 Applications of Molecular Modelling

A molecular mechanics program will return an 'energy value' for any configuration of a system. Force-fields can be used for local energy minimisation usually using a derivative minimisation method such as the steepest descent and Newton's method (see Chapter 5). Molecular mechanics can be used to calculate heats of formations by adding the energy to form the bonds in the molecule to the steric energy.

The temperature at which the system is at also makes a significant contribution to the molecular energy. The kinetic energy may be estimated from a statistical thermodynamics analysis, through Molecular Dynamics or Monte Carlo simulation which is explained in Section 2.4.1.

2.6 Conclusions

Ruelle et al. (1997) determined the following thermodynamic terms to be important in calculating the solubility of drugs:

- ideal solubility,
- correction factor for entropy of mixing,

- changes in the non-specific cohesion forces,
- hydrophobic effect,
- hydrogen bonds between proton-donor solvents and the multiple proton-acceptor sites of the solute.

This is reflected in the choice of descriptors used in the QSPR equations. However, QSPRs are only suitable for the systems to which the equations have been fitted. Group contribution methods are based on the assumption that the interaction parameters are transferable and independent of neighbouring groups and that the contributions from a group are only a function of the mole fraction and temperature. Molecular simulations are currently very demanding computationally especially for medium to large solutes, aside from the difficulty in obtaining reliable unlike pair potentials.

The use of computational chemistry to obtain parameters previously only available from experimental data seems promising in the development of a property prediction model. Therefore, the use of activity coefficient models that require information on the solute-solvent interaction energies could be modelled using these tools. These methods however currently are only valid for simple systems and the model would need further development to enable its use for more complex systems.

Chapter 3

Prediction of UNIQUAC Activity Coefficients

3.1 Introduction

The aim of this chapter is to develop a reliable and qualitatively accurate solubility prediction methodology for pharmaceutical solutes. The selection of solvents during process development can significantly affect the overall economics, safety and environment impact of the process. In order to *optimally* select these solvents, the relevant physical properties of a large number of materials must be accessible at low cost. Due to the limitations of current solubility predictions techniques as discussed in the previous chapter, there is a need for a low-cost and reliable prediction method. As most pharmaceutical solutes often contain electronegative heteroatoms and aromatic rings which may lead to specific interactions of these molecules with small polar and aprotic solvents, the solubility may exhibit negative deviations from ideality. A reliable solubility prediction methodology must be able to identify these favourably interacting solvents.

Section 3.2 starts with explaining the general principles behind the methodology. First, pairwise interaction energies obtained from pairs of molecules are used to estimate the binary interaction parameters, and these results are compared with those obtained using clusters of molecules to represent the first solvation shell.

3.2 General Principles

As described in Section 2.4.4, new methods have been proposed to calculate UNIQUAC binary parameters using computational chemistry tools and taking into account the molecular structure and polarisation of the system. These parameters for any given binary system are calculated in the following manner. The interaction energies for three molecular pairs are determined: one pair of one solute molecule i and one solvent molecule j, and the other two pairs have two solute molecules of type i and two solvent molecules of type j respectively. The potential energy, U_{ij}, of each molecular pair is then minimised and the interaction energy for that particular configuration is calculated by:

$$\triangle U_{ij} = U_{ij} - (U_i + U_j) \tag{3.1}$$

where U_i and U_j are the potential energy of the isolated molecules i and j respectively. The interaction parameters are then a function of the differences in strength between the solute-solvent interaction energy, and the solute-solute and solvent-solvent interaction energies respectively. The parameterisation of the force field is thus important in accurately describing these interactions.

Jónsdóttir and co-workers (1994, 1996) only used the global minimum potential energy for each possible conformer pair to determine the interaction energy but Sum and Sandler (1999) found clusters to be a more accurate description. The aim is to appropriately represent the solvation shell either explicitly or implicitly to account for all possible molecular interactions and then take a suitable average. We have chosen to use the COMPASS force field as it is parameterised to give accurate results for most common organic and inorganic molecules under most practical conditions.

A Boltzmann Jump Procedure (BJP) stochastic search was used to sample the conformational space of the molecular pair as shown in *Appendix C* and called using the MSI *Cerius*^{2} software package. The algorithm is used in the QUANTA/CHARMm software package (QUANTA 4.0, Molecular Simulations, Cambridge). A molecule is randomly perturbed and a Metropolis criterion (Metropolis et al., 1953) is used to accept or reject

the new conformations. If the energy of the perturbed conformation is lower than the previous conformation, the move will be accepted. If the energy is higher, a stochastic decision is taken based on the Boltzmann factor $\exp(-\triangle U/RT)$ where $\triangle U = U_{new} - U_{old}$, the difference between the energy of the new and old conformation. The probability of accepting an upward move increases with higher temperatures (T) and a lower energy difference ($\triangle U$). On this basis, the Boltzmann Jump Procedure allows upward jumps on the potential energy surface. The molecules are perturbed until the torsional angles positions have changed sufficiently or a user-defined value for the root-mean-square (RMS) deviation is exceeded. The molecule is then minimised and a new cycle is initialised. The temperature used in the selection criterion is an important parameter and its value depends on the motivation of the search. If the temperature defined is too low, jumps across the energy barrier will not be allowed but if the temperature is too high, all upward moves will be accepted. Parameters controlling the change in torsional angles and intermolecular distances per step are also predefined. Ideally, all combinations of parameters should produce a minimisation step after 20-50 perturbations. For more complex molecules, a larger number of perturbation cycles is needed to cover the conformational space. However, one limitation of using a stochastic search is that there is no guarantee that all the minima have been found.

3.3 Molecular Pairs

In the determination of interaction parameters using pairs, two different strategies were used: the first strategy only used the global minimum potential energy of the pair while the second used all the minima found in the conformational search. The energies were then averaged using the Boltzmann distribution with each interaction energy $\triangle U_{ij}^m$ weighted by a probability p_m based on total energy:

$$p_m = \frac{e^{\frac{-U_{ij}^m}{kT}}}{\sum_n e^{\frac{-U_{ij}^n}{kT}}} \tag{3.2}$$

where k is the Boltzmann constant. The Boltzmann averaged interaction energy is then given by:

$$\triangle U_{ij}^{av} = \sum_m p_m \triangle U_{ij}^m \tag{3.3}$$

The use of the Boltzmann distribution as described here however only provides a partial sampling of the conformational space of the molecular pair and does not account for the width of the minima wells. The interaction parameters for the UNIQUAC equation are given by:

$$A_{ji} = \frac{\triangle U_{ji} - \triangle U_{ii}}{Rq_i} \tag{3.4}$$

$$A_{ij} = \frac{\triangle U_{ij} - \triangle U_{jj}}{Rq_j} \tag{3.5}$$

where R is the gas constant and q_i is the surface area fraction of molecule i. Note that $\triangle U_{ij} = \triangle U_{ji}$.

Simple systems i.e. an alkane/alkane pair and an alkane/amine pair were the initial case studies. The aim is to compare results with experimental data for nonpolar and then polar systems.

3.3.1 Simple Alkane/Alkane and Alkane/Amine Systems

A simple alkane/alkane system was the first chosen to investigate the validity of this method, n-butane/n-pentane. The former has two conformers while n-pentane has three. The parameters varied in this stochastic search are the number of iterations, number of perturbations between iterations, temperature and step size. As we are interested in determining all the minima of the system and the molecules are relatively simple, the number of iterations was 3000. The temperature, 1000K, and step size were high so that the chances of jumping the energy barrier to find high energy conformer pairs were increased. The number of perturbations was 50 so that the number of moves accepted for each iteration was sufficient enough to cover the entire conformational space.

Strategy	Global Minimum	Boltzmann Average
A_{12}	69.5K	71.6K
A_{21}	-49.1K	-48.2K

Table 3.1: Interaction parameters for n-butane/n-pentane system

The vapour-liquid equilibrium of the system was reproduced using the UNIQUAC interaction parameters calculated from Equation 3.4 using the two different strategies. Fig. 3.1 shows the global minimum potential energy configuration for the n-butane/n-pentane pair. Fig. 3.2 shows the predicted VLE curve using both the global minimum potential energies for the conformer pairs and all of the energy minima. As alkane/alkane systems are almost ideal, the residual contribution to the activity coefficient is almost negligible and the predicted VLE curve reproduces the experimental data from Dortmund Databank well. Table 3.1 shows the interaction parameters calculated using the two strategies.

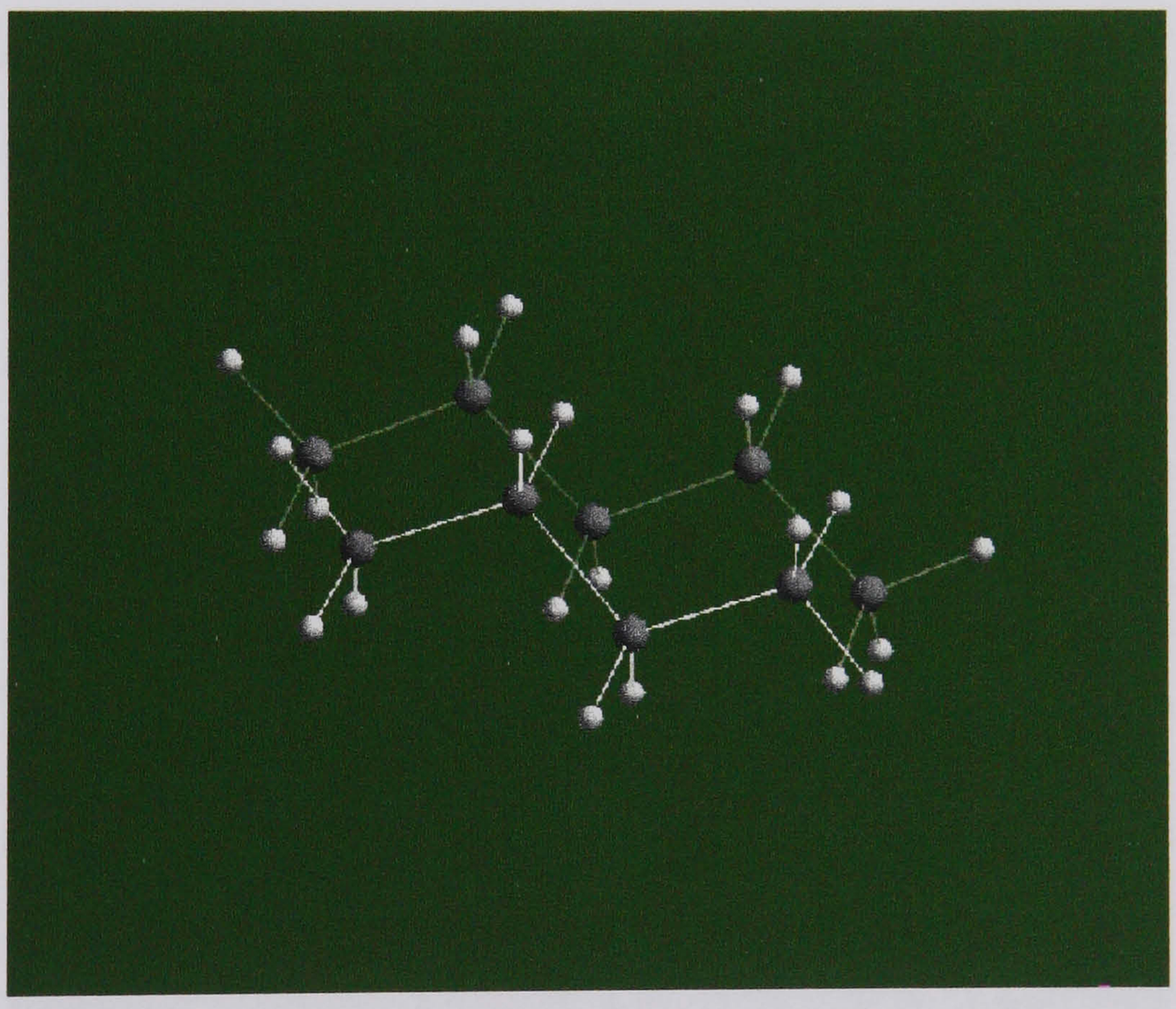


Figure 3.1: Global minimum potential energy of n-butane/n-pentane pair: -21.0588 kcal/mol where the grey atoms are carbon and the white atoms are hydrogen

The second system studied was a polar hydrogen-bonding alkane/amine pair, n-butane/ethylamine. Both molecules have two conformers each. Again the two different strategies

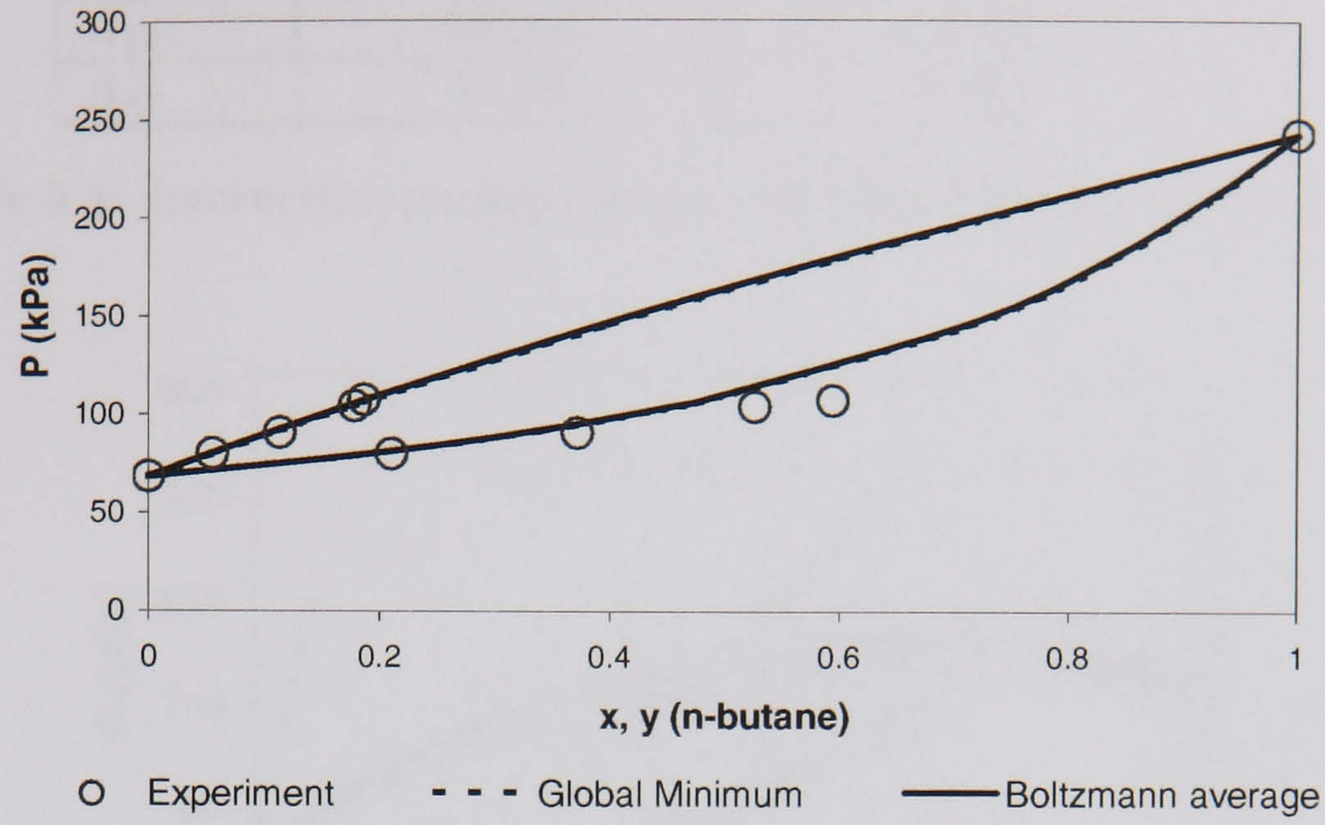


Figure 3.2: Phase diagram for n-butane/n-pentane at 25^{o}C

were compared with experimental data (Wolff et al., 1964). Fig. 3.3 shows the global minimum potential energy of an n-butane/ethylamine pair.

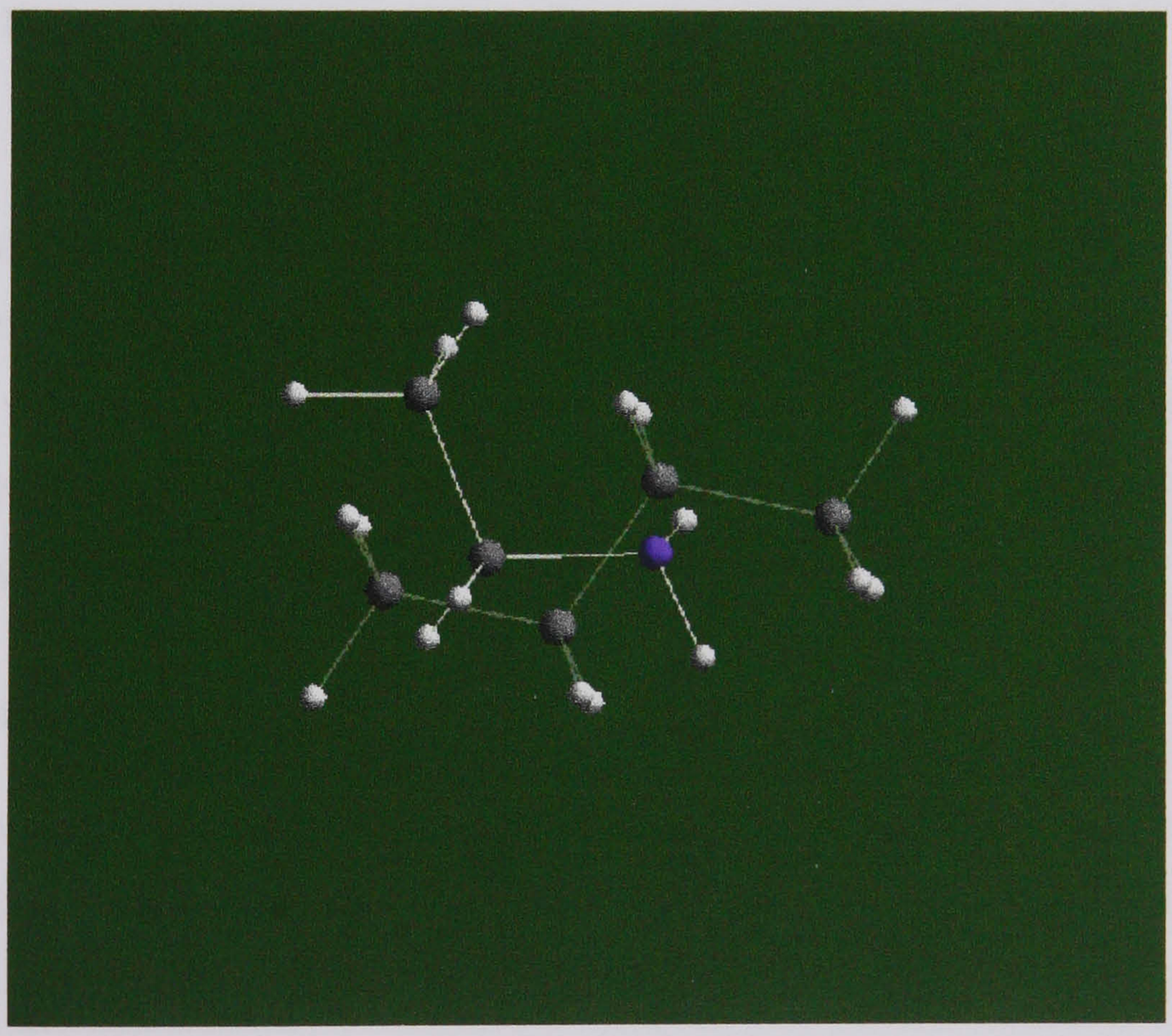


Figure 3.3: Global minimum potential energy of n-butane/ethylamine pair: -28.0414 kcal/mol where the grey atoms are carbon, white atoms are hydrogen and the blue atom is nitrogen

Strategy	Global Minimum	Boltzmann Average
A_{12}	498.4K	179.6K
A_{21}	48.1K	19.4K

Table 3.2: Interaction parameters for n-butane/ethylamine system

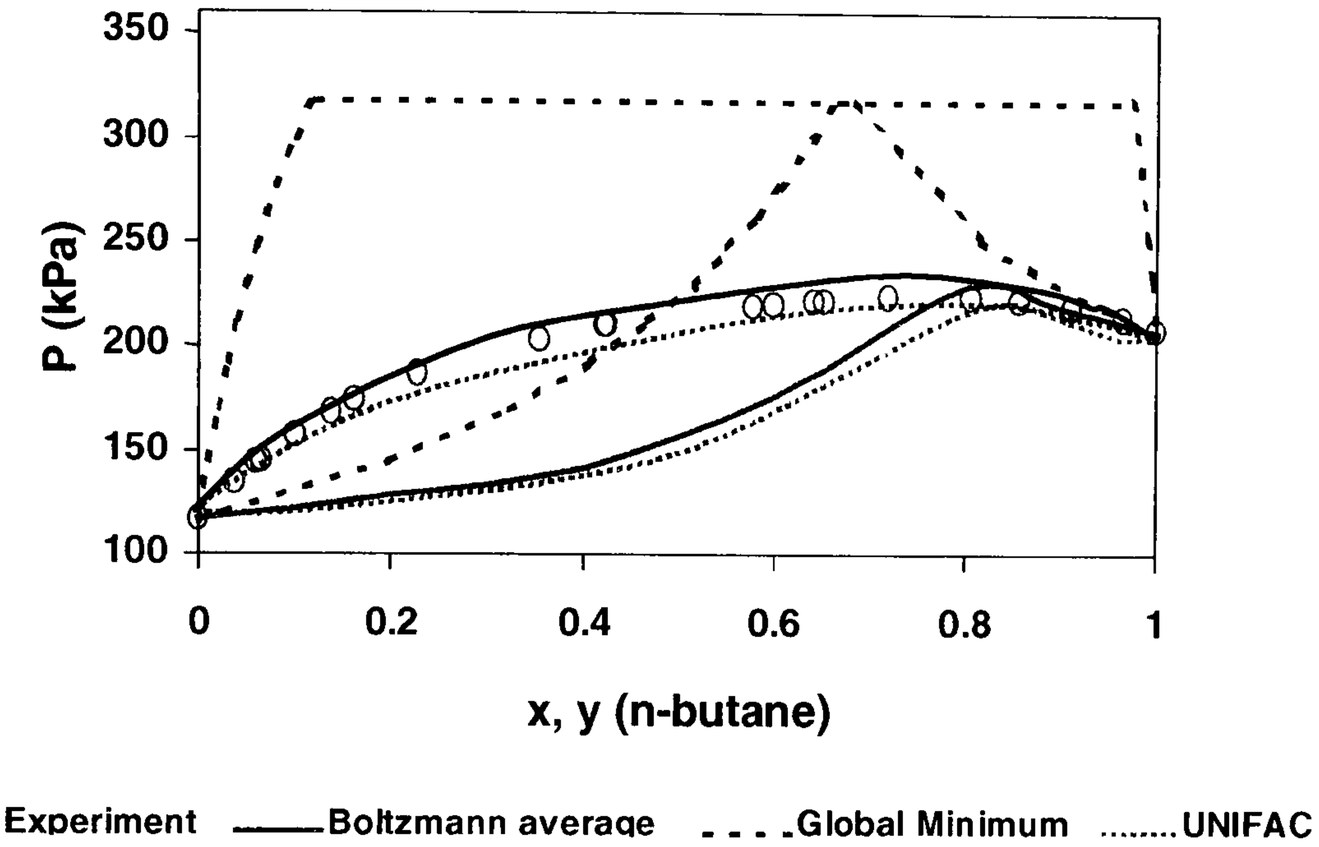


Figure 3.4: Phase equilibria for n-butane/ethylamine at 25^{o}C

Fig. 3.4 compares the predicted VLE data with experimental results. Both methods predict azeotropic behaviour but the global minimum strategy predicted two coexisting liquid phases which is in disagreement with experiments. The Boltzmann averaged interaction energies yielded a better representation. The UNIQUAC interaction parameters determined by the different strategies are shown in Table 3.2.

3.3.2 Paracetamol

The next step was to extend the work towards more complex molecules which is the aim of the solubility prediction method. Paracetamol (4-hydroxyacetanilide, $C_8H_9NO_2$) as shown in Fig. 3.5, a medium-sized molecule with an aromatic ring and a flexible tail, is an important analgesic and antipyretic agent. Granberg and Rasmuson (1999) reported its solubility in 26 solvents as shown in Fig. 3.6. It has a low solubility in nonpolar and chlorinated hydrocarbons and close to ideal solubility in solvents of medium polarity e.g. ketones, ester, alcohols. The solubility of paracetamol in n-alcohols and ketones decreases

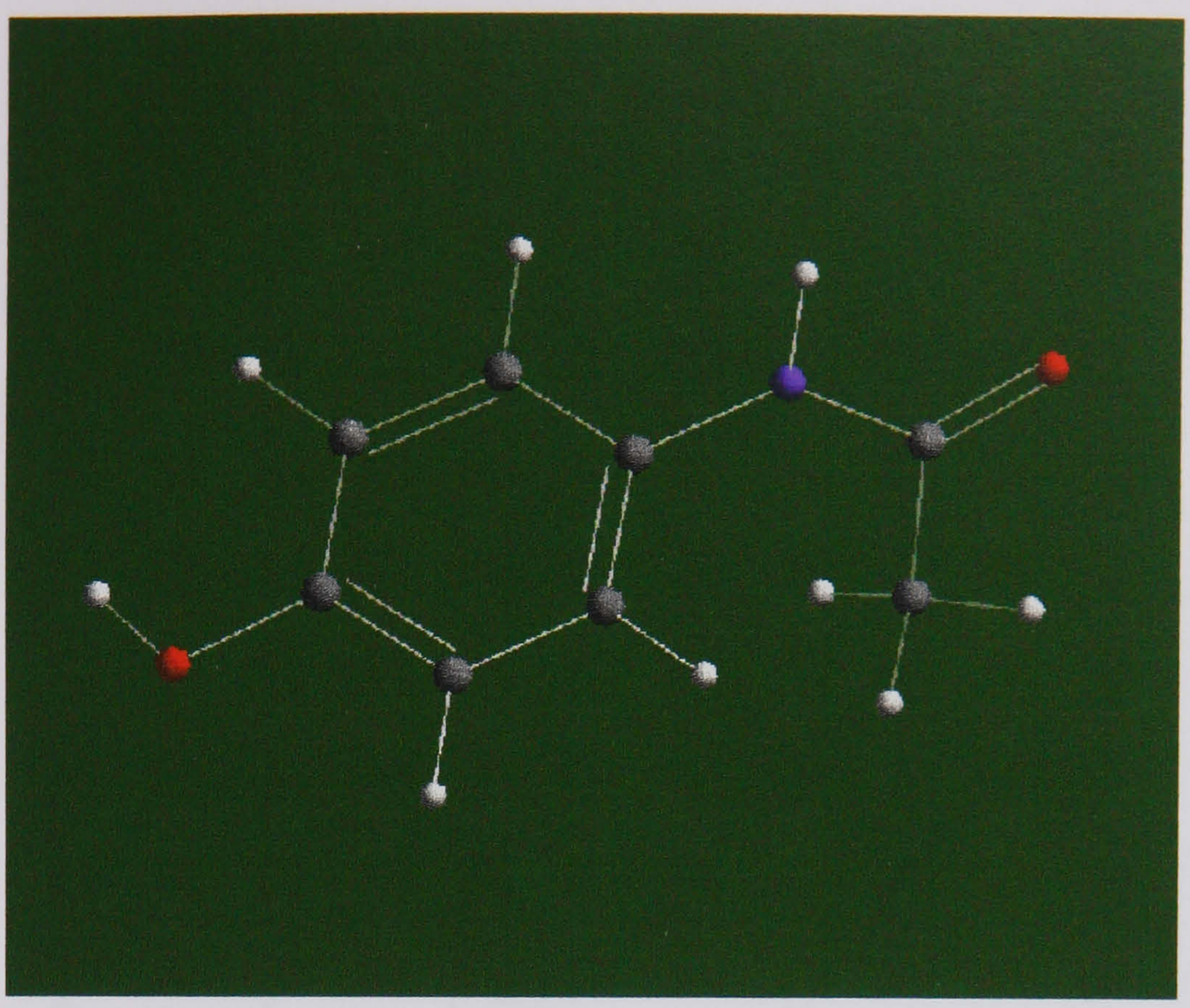


Figure 3.5: Paracetamol molecule where carbon atoms are shown in grey, oxygen in red, nitrogen in blue and hydrogen in white

with increasing length of the carbon chain as shown in Fig. 3.7. Paracetamol however shows strong negative deviation from ideality in aprotic solvents and thus has better than ideal solubility in diethylamine for example. The solubility of paracetamol is calculated using Equation 2.9.

Ten solvents were chosen to evaluate the two strategies: diethylamine (DEA), dimethyl formamide (DMF), dimethylsulfoxide (DMSO), methanol (MeOH), ethanol (EtOH), acetone, acetic acid (AcH), ethyl acetate (AcOEt), toluene and dichloromethane (DCM). These solvents vary in their polarity from methanol with a solubility parameter of 29.7 $(J/cm^3)^{1/2}$ to toluene which has a solubility parameter of 18.2 $(J/cm^3)^{1/2}$. The potential energy minima for each pair of molecules were found using the BJP with an iteration number of 5000. The solubility ranking is largely dependent on the difference between the solute-solvent and the solvent-solvent interaction energies, and the relative molecular volume and surface area of the solvent. The predicted solubilities using both strategies are shown in Table 3.3 and 3.4 where 1 refers to the solute and 2 refers to the solvent.

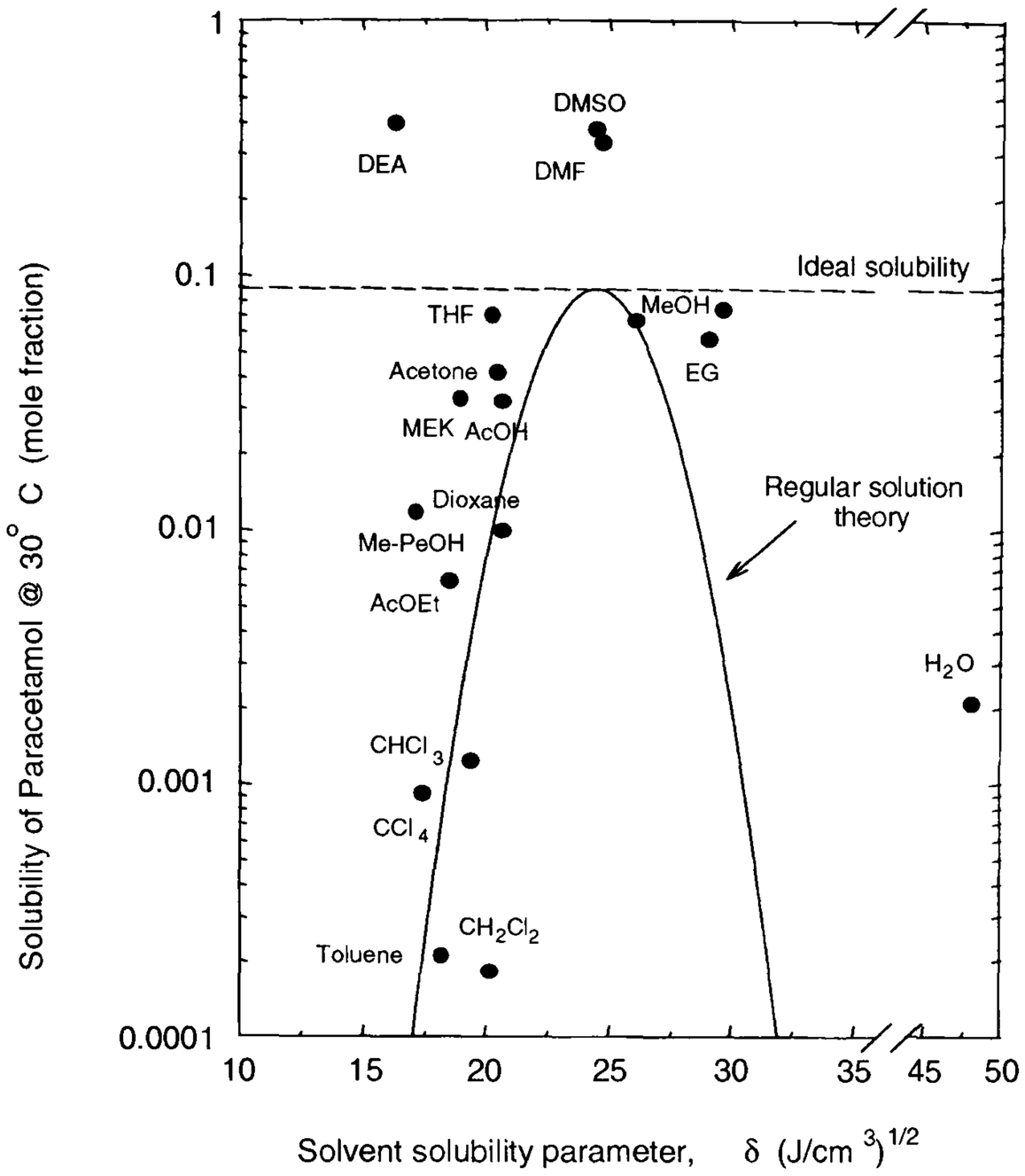


Figure 3.6: Solubility of paracetamol showing experimental data and prediction by regular solution theory

Solvent	$\triangle U_{12}$	$\triangle U_{22}$	x_1^{calc}	x_1^{expt}
DEA	-10.21	-5.80	0.361	0.389
DMF	-10.17	-6.11	0.343	0.369
DMSO	-7.61	-3.93	0.328	0.329
MeOH	-8.05	-6.60	0.242	0.073
EtOH	-8.52	-7.12	0.198	0.066
Acetone	-8.03	-4.82	0.313	0.041
AcH	-12.55	-13.50	1.1E-05	0.032
AcOEt	-8.27	-4.80	0.326	0.006
Toluene	-6.95	-3.32	0.334	0.0002
DCM	-5.57	-2.31	0.308	0.0002

Table 3.3: Interaction energies for solute-solvent, $\triangle U_{12}$, and solvent-solvent, $\triangle U_{22}$, calculated from global minimum potential energies of the different conformer pairs for paracetamol and solvent (kcal/mol), and the calculated solubilities, x_1^{calc}, compared with experimental solubilities, x_1^{expt}.

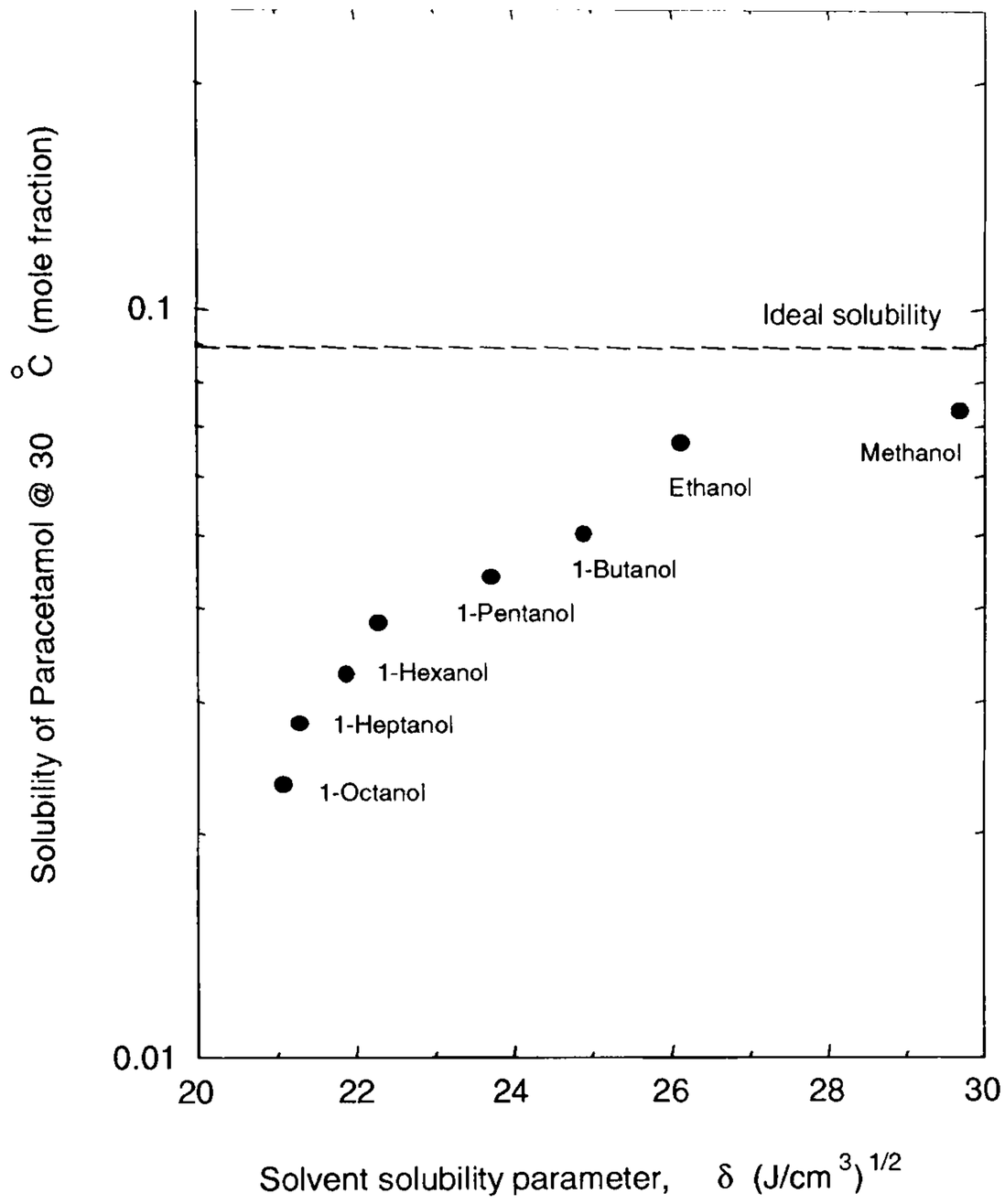


Figure 3.7: Solubility of paracetamol in n-alcohols

Solvent	$\triangle U_{12}$	$\triangle U_{22}$	x_1^{calc}	x_1^{expt}
DEA	-7.97	-3.66	0.364	0.389
DMF	-8.70	-5.65	0.322	0.369
DMSO	-5.73	-3.17	0.284	0.329
MeOH	-7.11	-6.28	0.171	0.073
EtOH	-7.15	-6.13	0.098	0.066
Acetone	-6.29	-3.23	0.309	0.041
AcH	-9.23	-11.00	2.3E-05	0.032
AcOEt	-6.10	-4.23	0.011	0.006
Toluene	-6.31	-3.00	0.314	0.0002
DCM	-4.52	-2.26	0.274	0.0002

Table 3.4: Interaction energies for solute-solvent, $\triangle U_{12}$, and solvent-solvent, $\triangle U_{22}$, calculated from the Boltzmann average of all potential energy minima for paracetamol and solvent (kcal/mol), and the calculated solubilities, x_1^{calc}, compared with experimental solubilities, x_1^{expt}.

Both strategies correctly identified the best solvent to be DEA. Acetic acid is rightly predicted to be a poor solvent but the quantitative prediction deviates by a large factor. Nonpolar molecules and acetone were predicted inaccurately to be good solvents. The other solubility predictions are quite comparable to experimental data if the solute-solvent interactions are similar to the solute-solute and solvent-solvent interactions. When using the Boltzmann averaged interaction energies however, the predicted solubilities improved when compared to experimental data especially for EtOH and AcOEt.

From the results shown in the previous section for the alkane/amine system, using the Boltzmann average of the interaction energies would give a good indication of the phase equilibria for similar types of molecules. However, if the molecules were of dissimilar size and different polarity, the method fails because stronger interactions would then dominate. In the case of acetic acid, low solubility for paracetamol was predicted because the two acetic acid molecules form dimer pairs as shown in Fig. 3.8 that strongly overemphasised the solvent-solvent interaction energies. The hydrogen bonds are shown as yellow dotted lines. Both the nonpolar molecules, toluene and DCM, were predicted to have high solubilities because the stronger interactions between the solvent molecules with the electronegative atoms dominated due to the lack of steric hindrance. In actuality, the solvation shell would consist of both strong and mostly weak interactions.

An alternative approach to ranking solvents is to consider only the solute-solvent interaction energies, U_{12}. In this case, there is a slight trend on the interactions in Table 3.4 as the solvents that exhibit good solubilities (DEA and DMF) have strong solute-solvent interactions while DCM has weak solute-solvent interactions. The interaction energy value however does not account for the solvent-solvent interaction which may either be weak as in the DEA case or strong in the case of MeOH and EtOH. Therefore, a cluster approach was used based on 10 solvent molecules surrounding the paracetamol solute and the solute-solvent interaction energies were then calculated as a Boltzmann average of the ten binary interactions as shown in Table 3.5. In this model, the solvent-solvent interactions are implicitly accounted for as their interactions with each other will affect their interactions with the solute molecule. Except for the case of DMSO, the solute-solvent

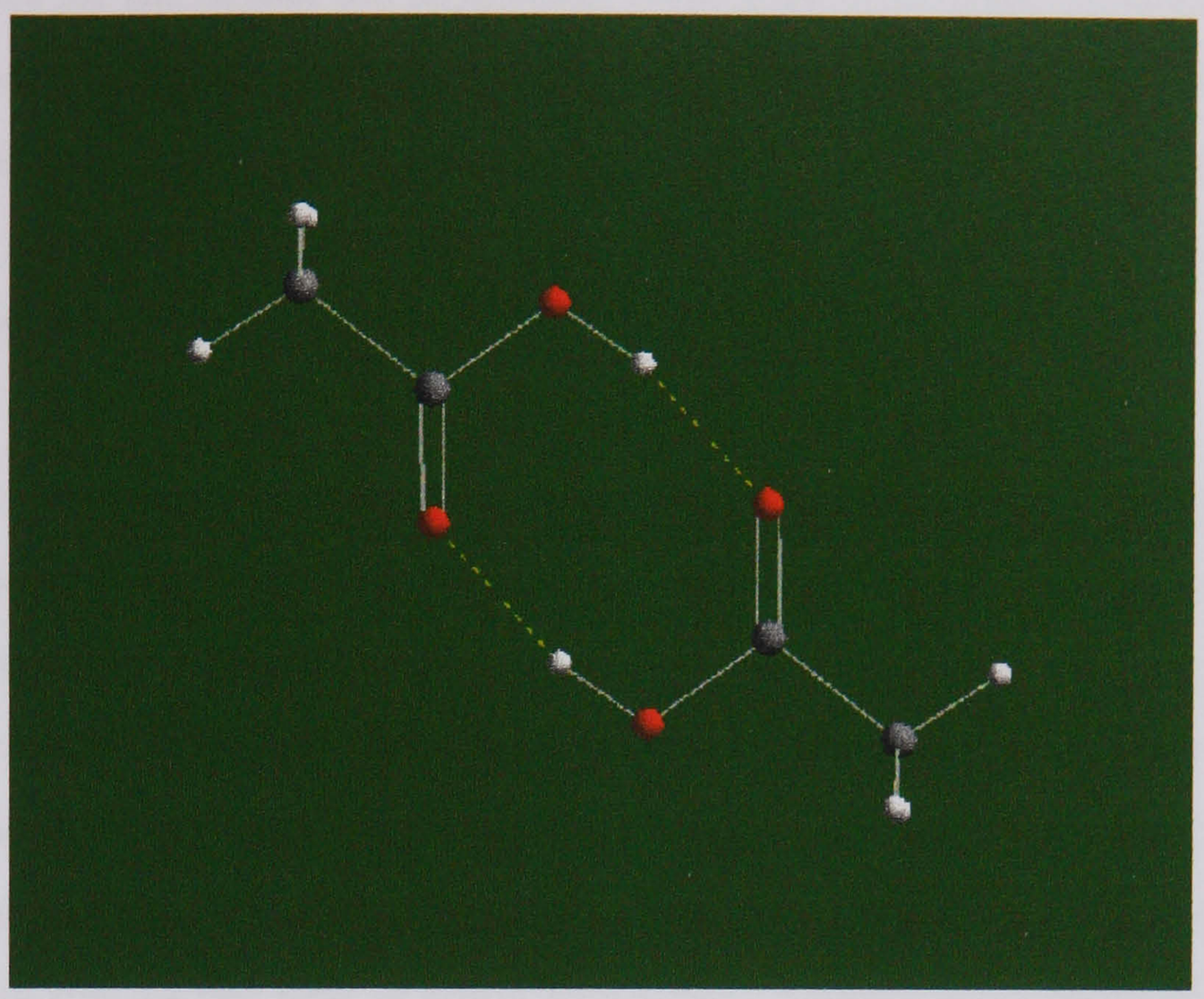


Figure 3.8: Acetic acid dimer configuration

Solvent	$\triangle U_{12}$	r_2	x_1^{expt}
DEA	-8.43	3.730	0.389
DMF	-7.25	3.104	0.369
DMSO	-5.66	2.768	0.329
MeOH	-6.12	1.468	0.073
EtOH	-6.09	2.268	0.066
Acetone	-5.62	2.540	0.041
AcH	-5.60	2.238	0.032
AcOEt	-5.44	3.672	0.006
Toluene	-4.98	3.930	0.0002
DCM	-4.52	2.187	0.0002

Table 3.5: Interaction energies for solute-solvent, $\triangle U_{12}$, using a cluster of 10 solvent molecules and the corresponding experimental solubilities, x_1^{expt}, with r_2 showing the solvent molecular volume.

interactions allow the solvents to be ranked in agreement with experimental data. Two possible reasons for the outlier are as follows. Firstly, the force field parameterisation may not be accurate for DMSO. Secondly, alcohols exhibit strong solute-solvent interactions but the solubilities are not high as these molecules also exhibit strong solvent-solvent interactions. Although solvent-solvent interactions are accounted for in this model, the number of solvent molecules might not be sufficient to fully account for it.

All possible solute-solvent interactions are not thoroughly investigated using pairs as only specific interactions are found especially if the solute molecule is much larger than the solvent molecule. An example of this is shown in Figs 3.9 and 3.10 which show the surface area of a chlorobenzene solute where a solvent molecule would be found after minimisation for pairs of molecules. For smaller molecules, the interaction energies are overestimated as the solvent molecule will be positioned in the few energetically favourable positions with the solute due to the lack of steric hindrance. Therefore a cluster-based approach may be a better model as it accounts both for specific interations as well as steric effects that may hinder such interactions.

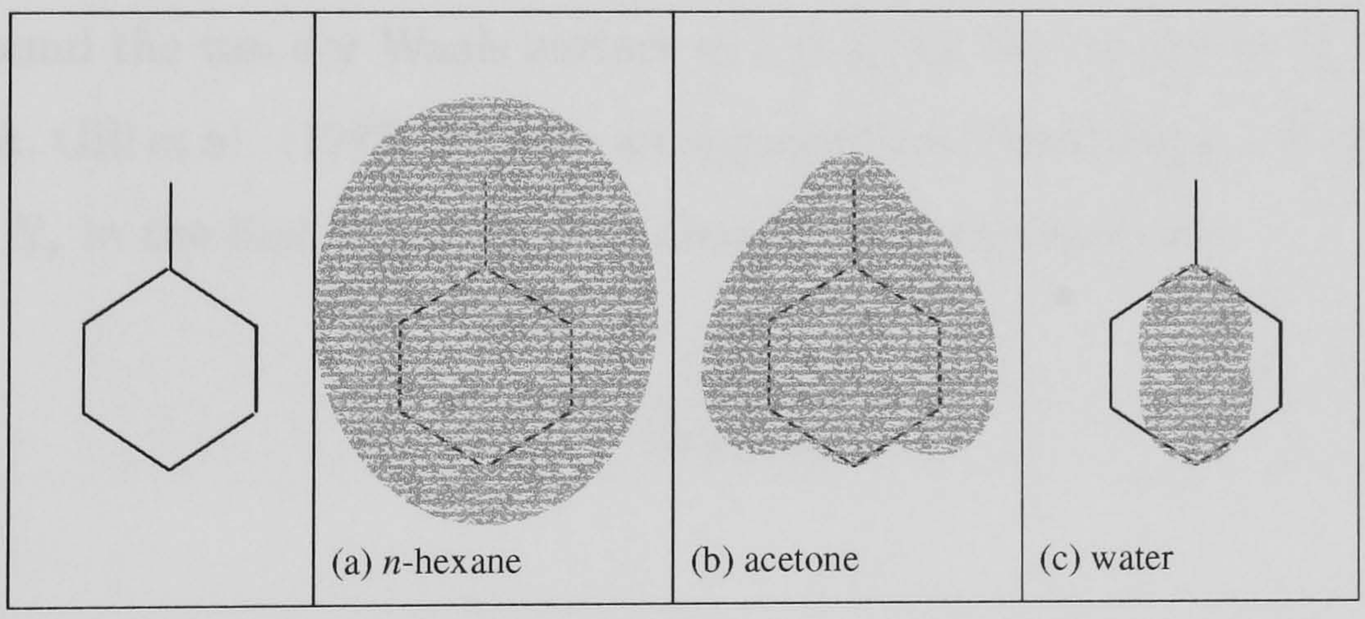


Figure 3.9: Top view: Surface area of chlorobenzene solute where a solvent molecule would be found after minimisation

3.4 Molecular Clusters

Two issues are considered in developing a cluster-based model: the number of solvent molecules and then solvation shell modelling. The basic idea behind using a cluster is to represent the first solvation shell.

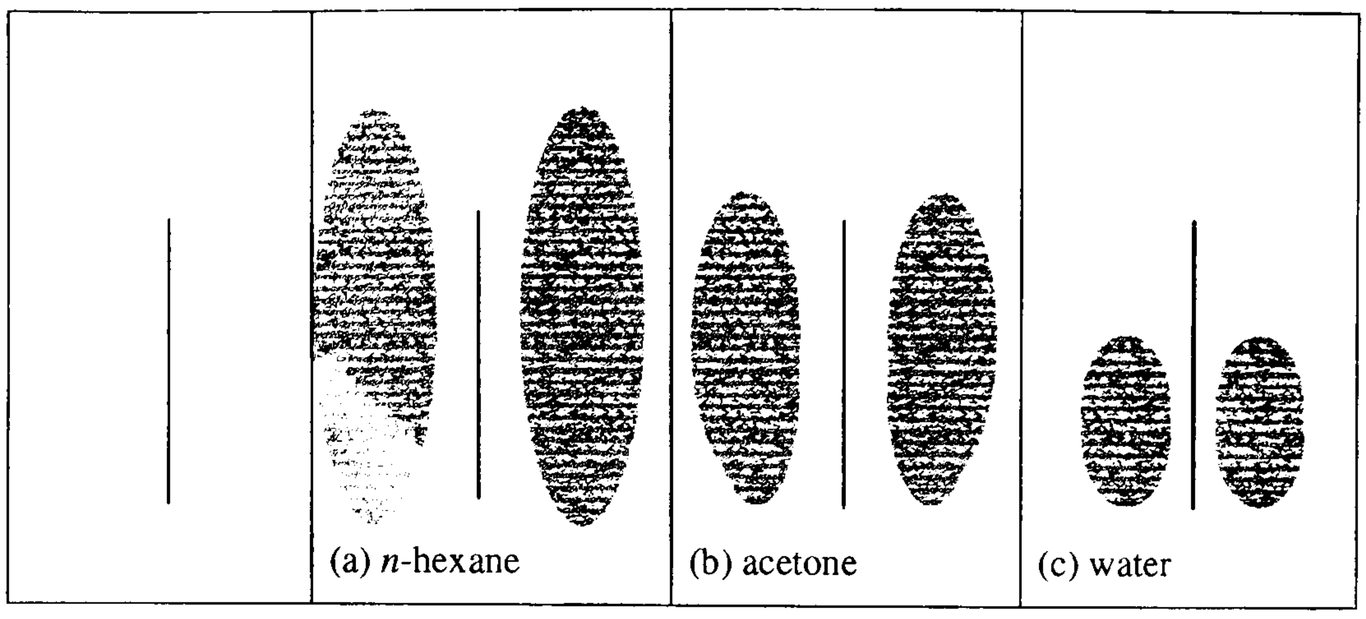


Figure 3.10: Side view: Surface area of chlorobenzene solute where a solvent molecule would be found after minimisation

3.4.1 Solvent Accessible Surface Area

The concept of solvent accessible surface area (SASA) was introduced by Richards (1977). The van der Waals surface area is defined as the external surface area resulting from placing a sphere of radius, $r_{vdW,i}$, centred on each atom i as shown in Fig. 3.11. The value of $r_{vdW,i}$ was estimated using the van der Waals volume found using MSI $Cerius^2$. The SASA is then the surface area generated by the centre of the sphere of radius $r_{vdW,solvent}$ moving around the van der Waals surface of the solute also as shown in Fig. 3.11. From this concept, Gill et al. (1985) propose an approach for determining the number of solvent molecules, N_s in the first solvation shell around a solute molecule:

$$N_s = \frac{SASA_{solute}}{(2r_{vdW,solvent})^2}$$

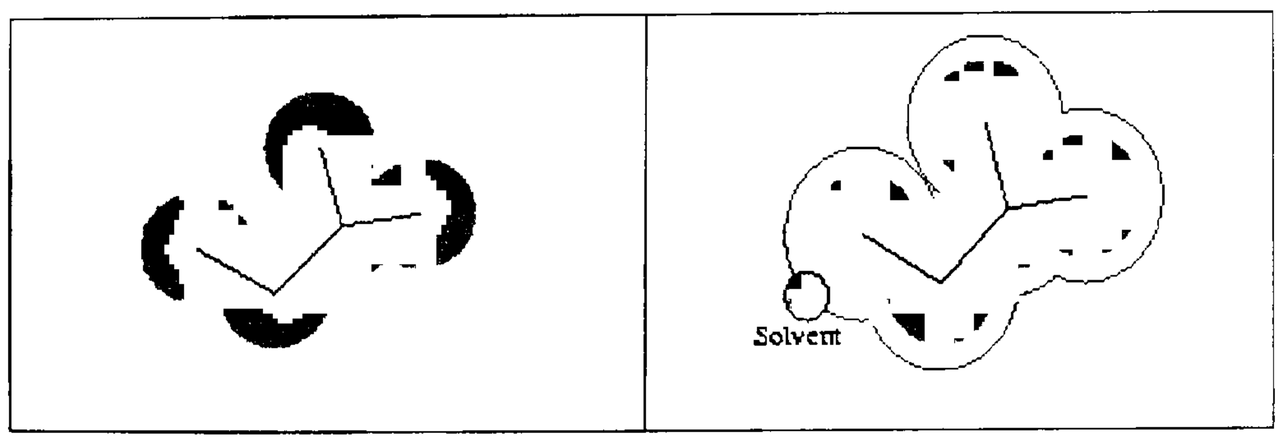


Figure 3.11: van der Waals surface area and solvent accessible surface area (SASA)

3.4.2 First Solvation Shell Modelling

The aim of the cluster is then to model the first solvation shell. As there are many ways of placing the solvent molecules surrounding the solvent which would produce different results, the objective is to have a systematic methodology of placing the solvent molecules before performing the local minimisation where specific interations would be emphasised. For most systems with similar molecules, the initial configuration is as shown in Figs. 3.12 and 3.13 for chlorobenzene and phenol. From the figures, the carbon atoms are shown in grey, oxygen in red, chlorine in magenta and hydrogen in white. The hydrogen bonds are shown in yellow dotted lines.

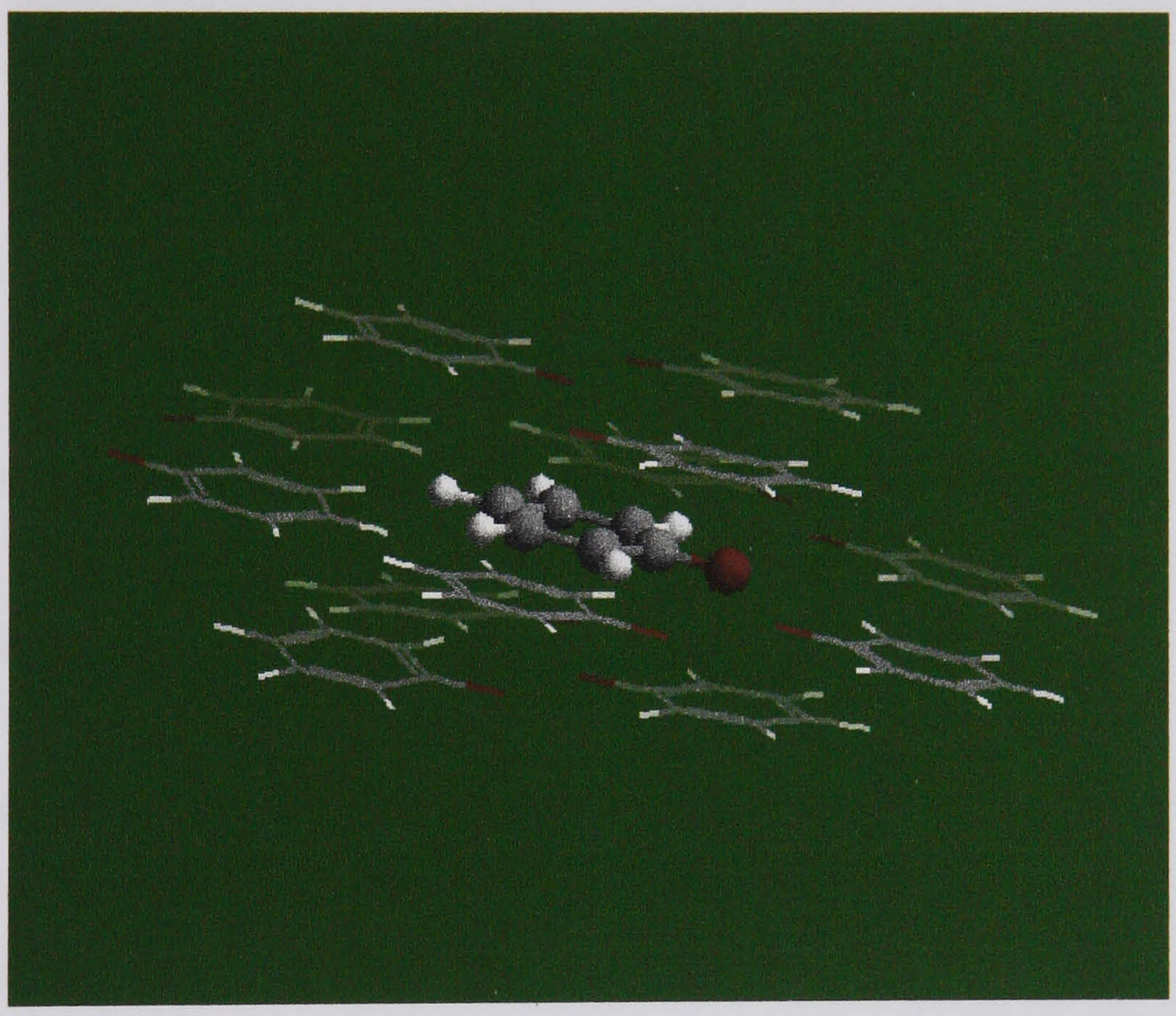


Figure 3.12: Initial configuration of chlorobenzene solute surrounded by 12 chlorobenzene molecules

Using the initial structure, Figs. 3.14 and 3.15 show the clusters of nonpolar solutes, chlorobenzene, and polar solutes, phenol, respectively after the minimisation step. The binary interaction energies for all pairs between the middle solute molecule and the other solute molecules are then determined. The solute-solute interaction energy for the system is calculated from the Boltzmann average of the binary interaction energies using Equa-

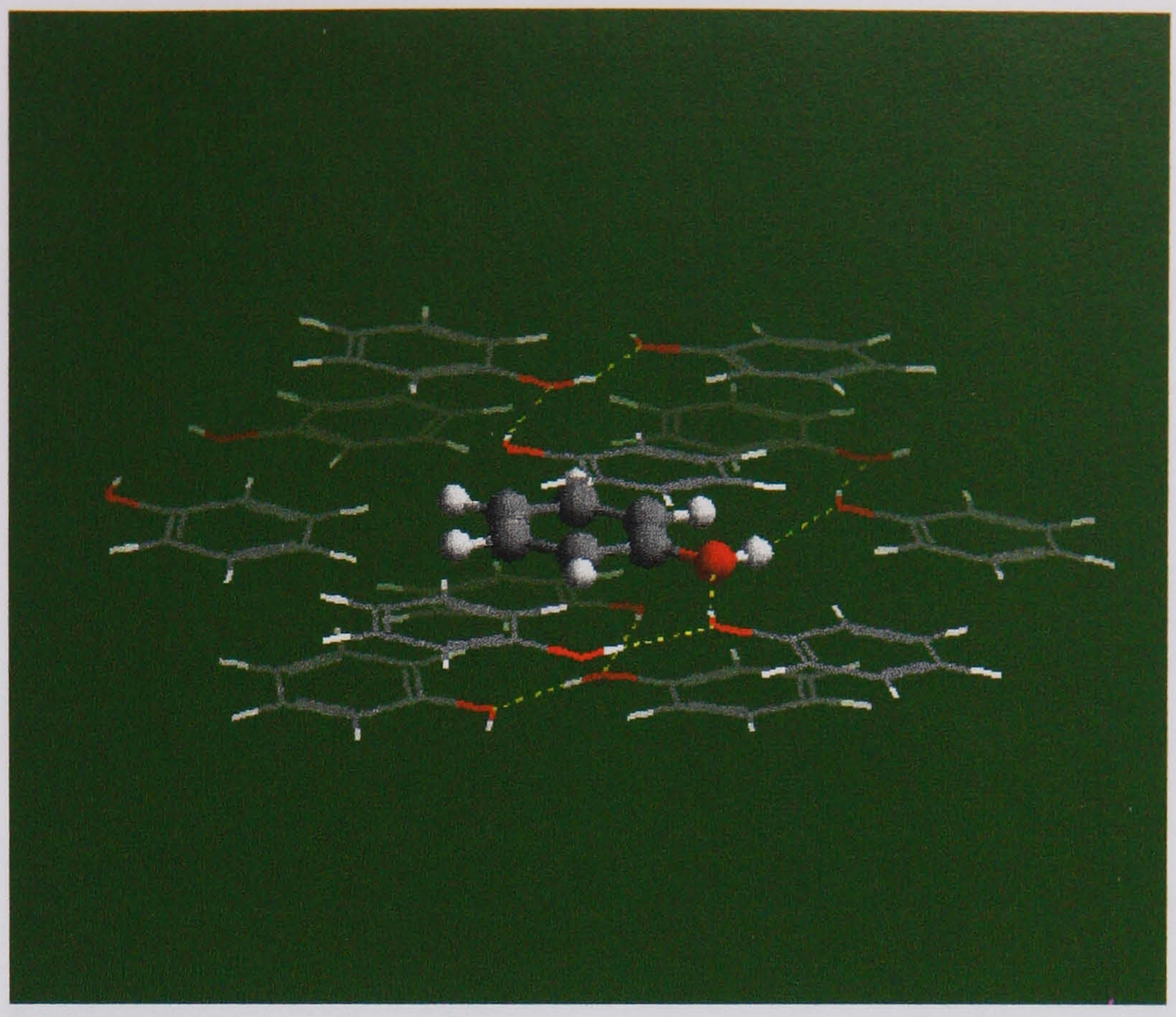


Figure 3.13: Initial configuration of phenol solute surrounded by 12 phenol molecules

tion 3.3. The solute-solute interaction energy for the chlorobenzene system is -1.9404 kcal/mol and for the phenol system is -3.4587 kcal/mol. The phenol solute molecule in the middle has two hydrogen bonds with the neighbouring molecules.

In order to compare solubilities, the prediction method should also be able to model both nonpolar and polar solvents. Therefore, we have chosen four different solvents of varying polarity, *n*-hexane, acetone, ethanol and water, for testing. Figs. 3.16 and 3.17 show the different configurations of nonpolar *n*-hexane and polar hydrogen-bonding ethanol after the minimisation step. The solvent-solvent interaction energy is -1.9663 kcal/mol for *n*-hexane and -4.0873 kcal/mol for ethanol. The similarity between the phenol and ethanol configurations are obvious. Again it must be emphasized that the real solvation shell will not be as structured. However since the UNIQUAC interaction parameters are based on the difference between the interaction energies, the difference is then what we are attempting to determine.

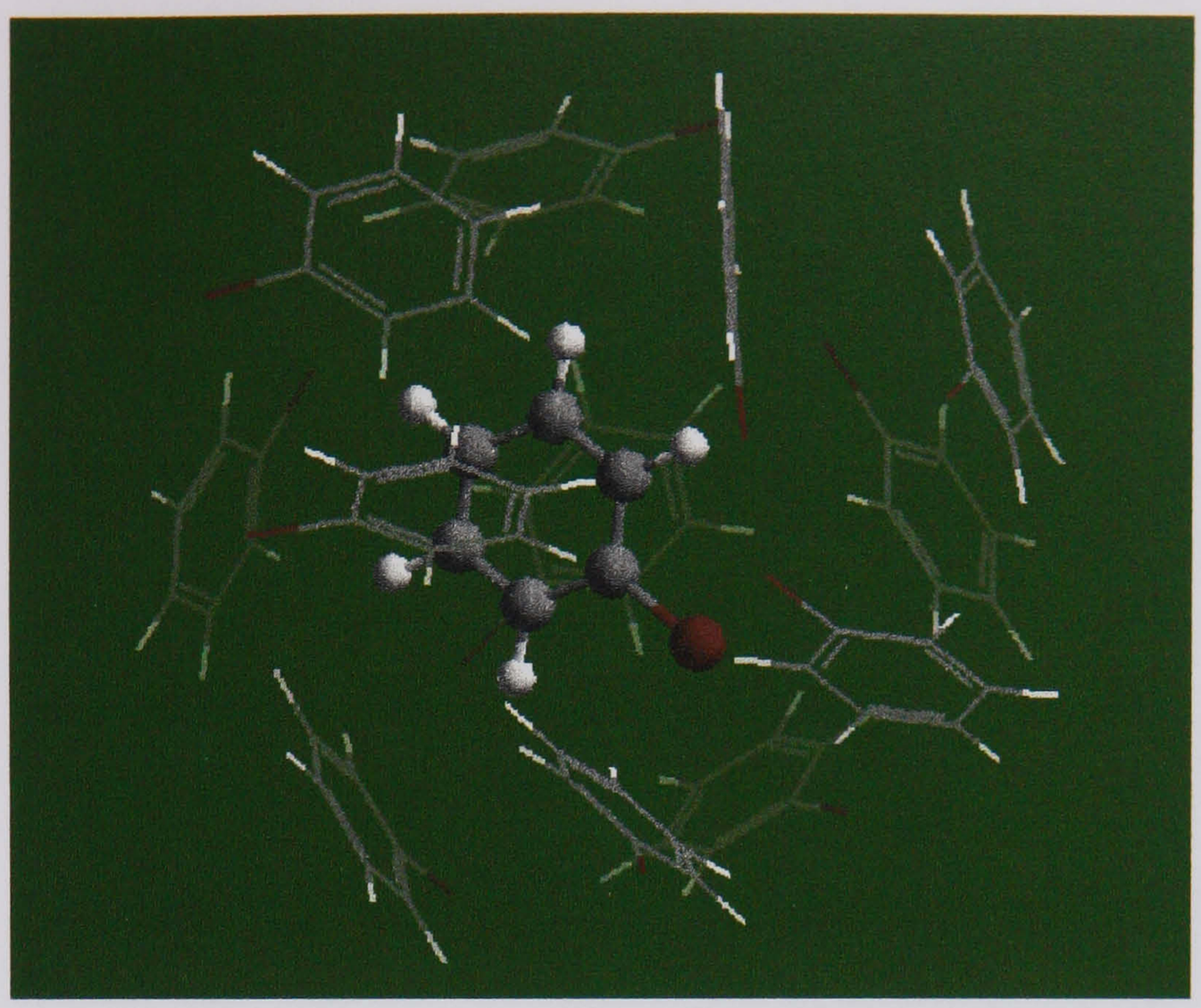


Figure 3.14: Cluster of chlorobenzene solute surrounded by 12 chlorobenzene molecules (solute-solute interactions) - local energy minimisation from configuration in Fig. 3.12

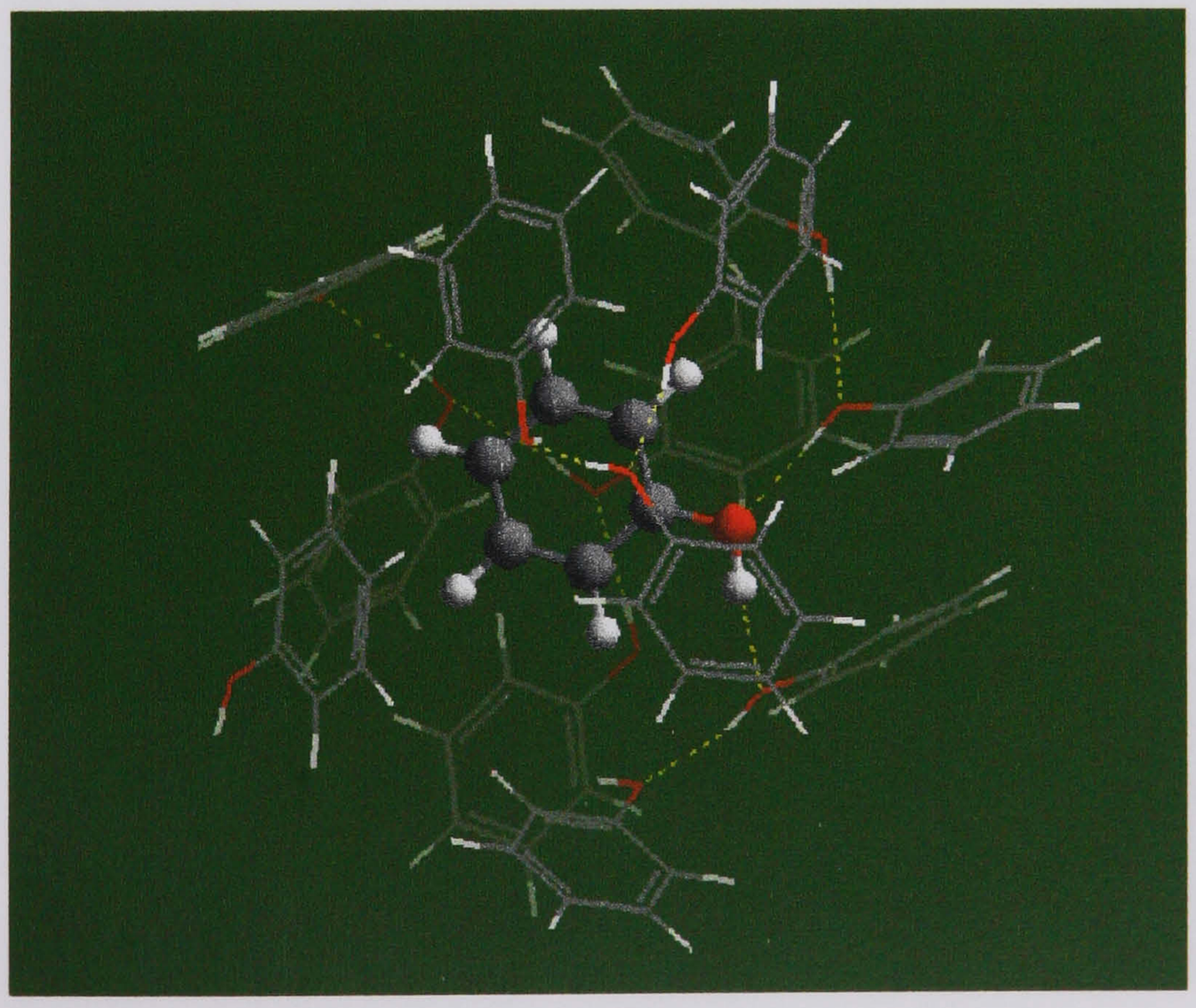


Figure 3.15: Cluster of phenol solute surrounded by 12 phenol molecules (solute-solute interactions) - local energy minimisation from configuration in Fig. 3.13

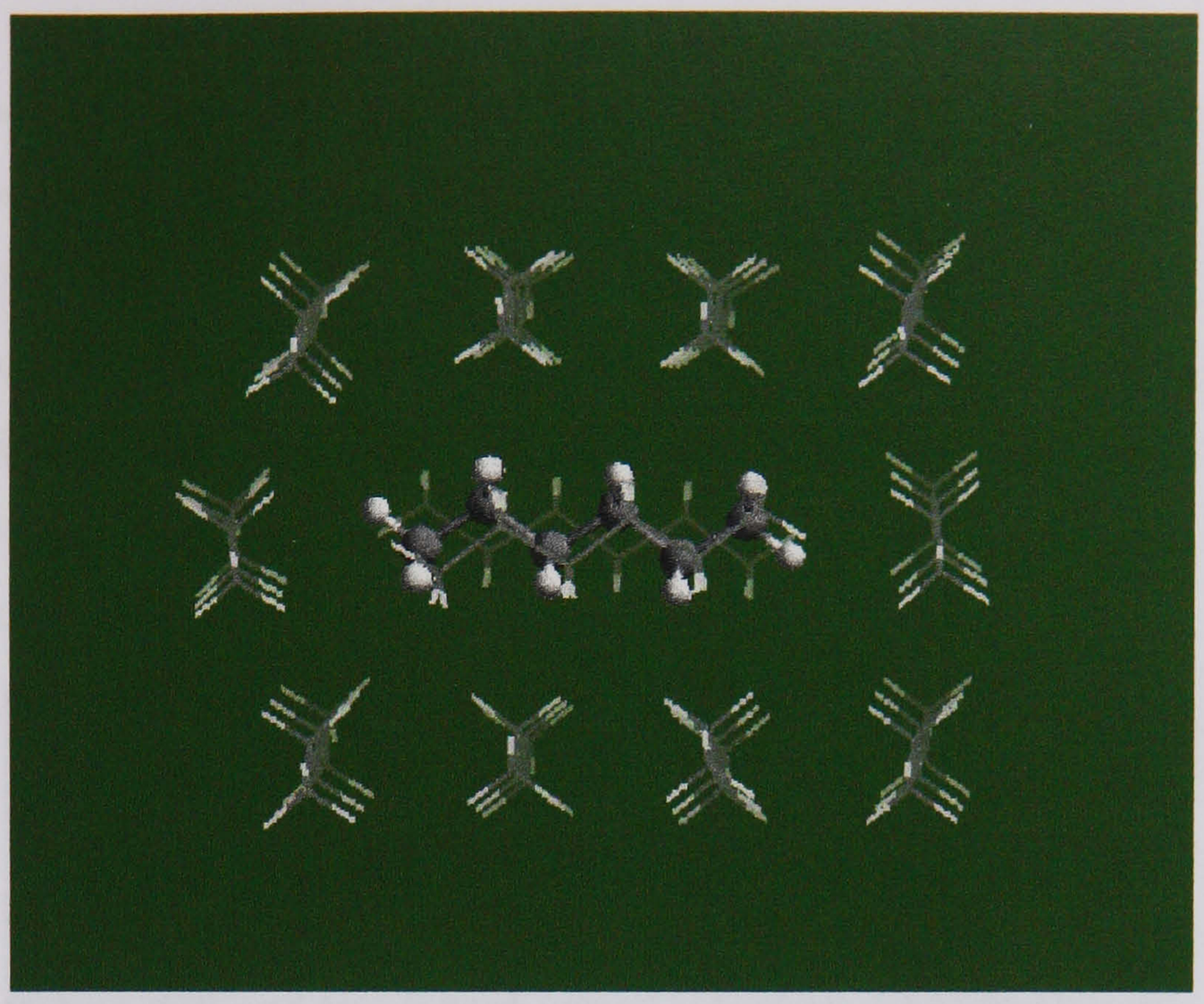


Figure 3.16: Cluster of hexane solvent surrounded by 12 hexane molecules (solvent-solvent interactions) - local energy minimisation from starting configuration

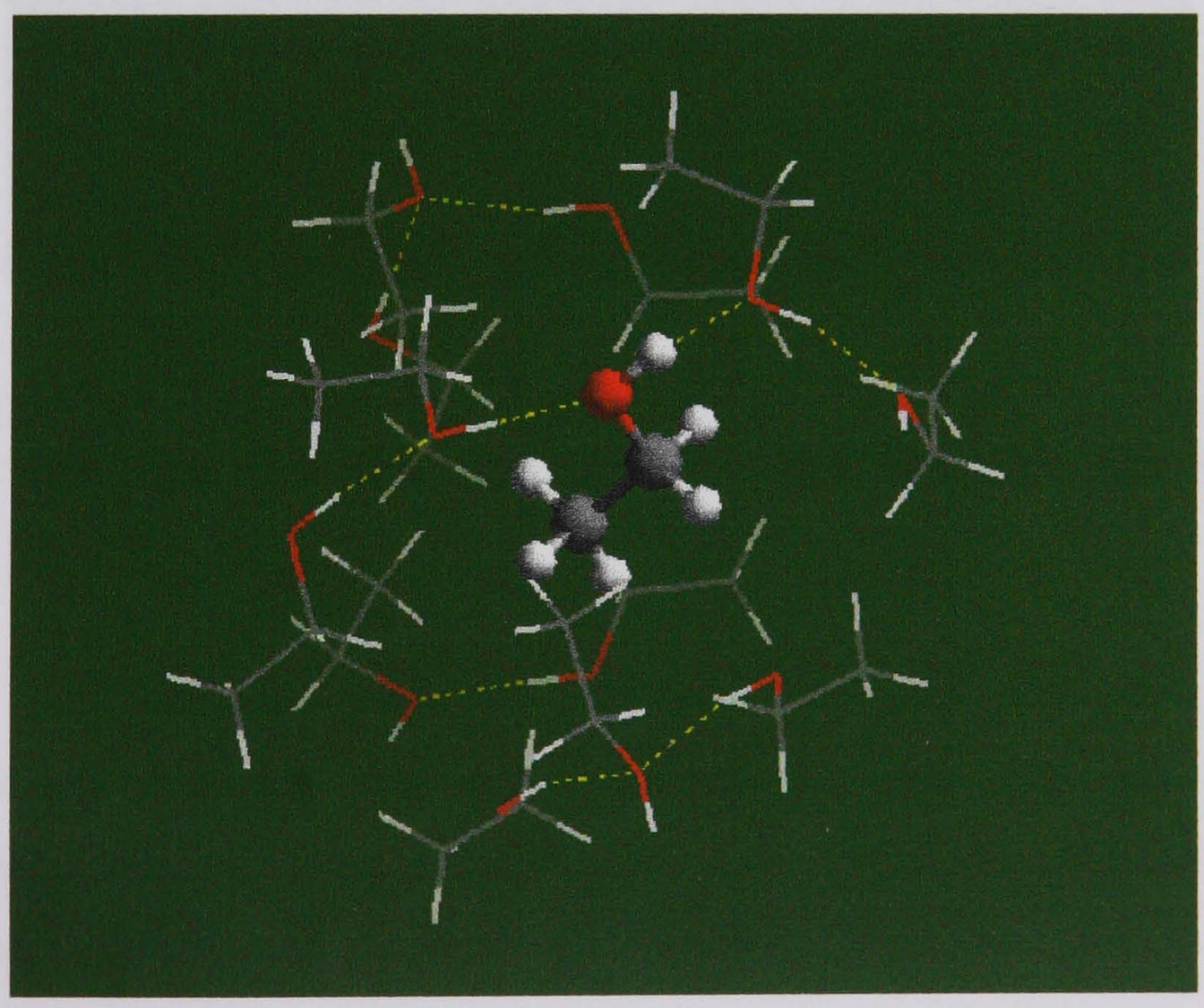


Figure 3.17: Cluster of ethanol solvent surrounded by 12 ethanol molecules (solvent-solvent interactions) - local energy minimisation from starting configuration

n-Hexane

Fig. 3.18 shows the final configuration of a chlorobenzene solute surrounded by 12 *n*-hexane solvent molecules. Nonpolar molecules are easy to model because due to lack of polarisability, the configuration maintains its initial structure. There are no strong specific interactions. The configuration gives a nonpolar solute-solvent interaction energy of -1.9093 kcal/mol. The UNIQUAC interaction parameters A_{12} and A_{21} are 7.4 and 5.5 which gives the infinite activity coefficients, γ_1^∞ and γ_2^∞ values of 1.29 and 1.39 compared with experimental data of 1.70 and 1.85.

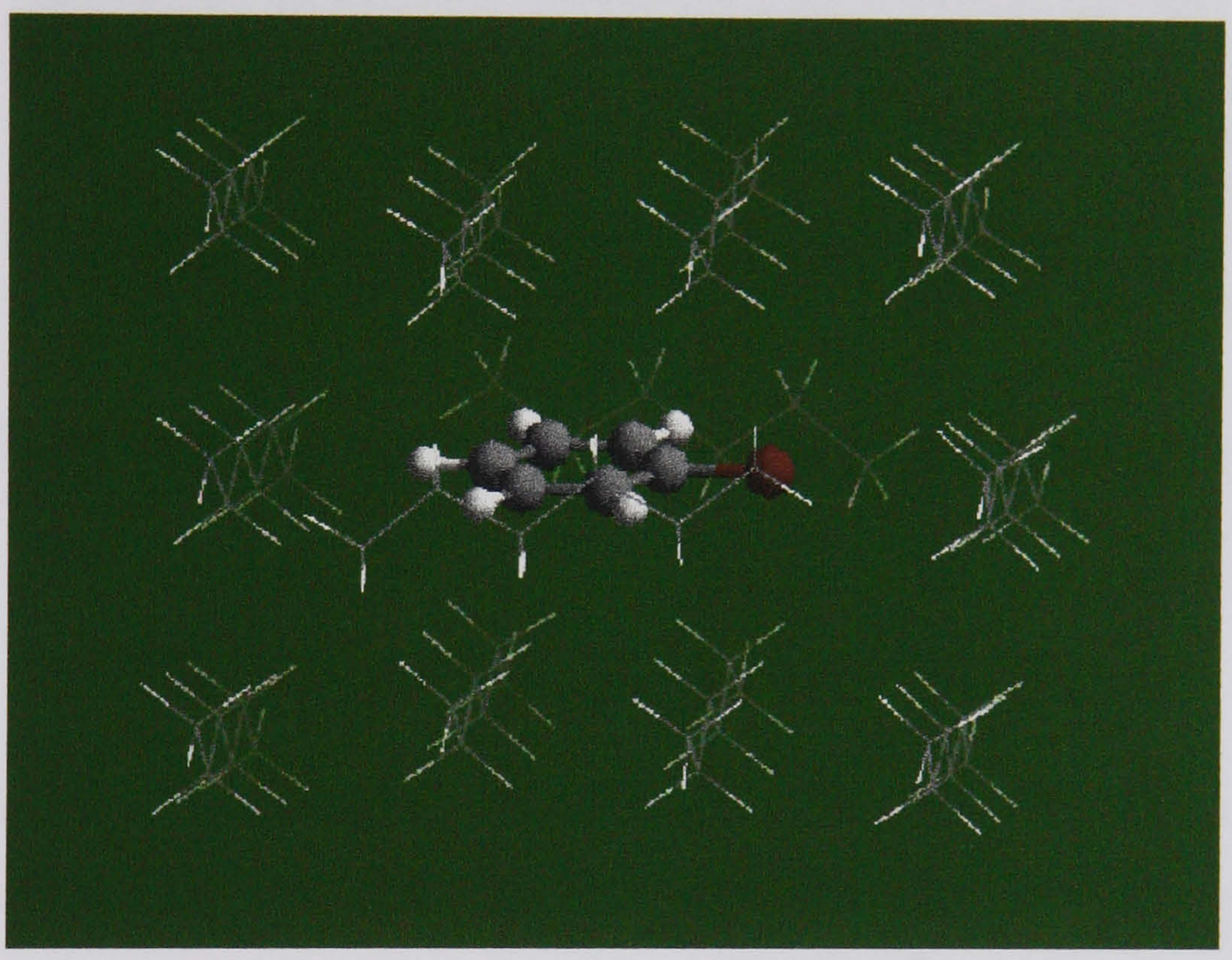


Figure 3.18: Cluster of chlorohexane solute surrounded by 12 hexane molecules (solute-solvent interactions) - local energy minimisation from starting configuration

Acetone

The acetone solvent consists of two methyl groups and a carbonyl group. The hydroxyl group in phenol acts as a hydrogen-bond donor to the double-bonded oxygen atom in acetone as shown in Fig. 3.19. The acetone solvent molecules do not form strong interactions with each other and the electronegative oxygen atoms tend to repel each other. The solute-solvent interaction energy is -3.3642 kcal/mol due to hydrogen-bonding, which gives the infinite activity coefficients, γ_1^∞ and γ_2^∞, values of 0.0173 and 0.1395 respectively

compared with the experimental values of 0.2 and 0.4 for γ_1^∞ and γ_2^∞ respectively.

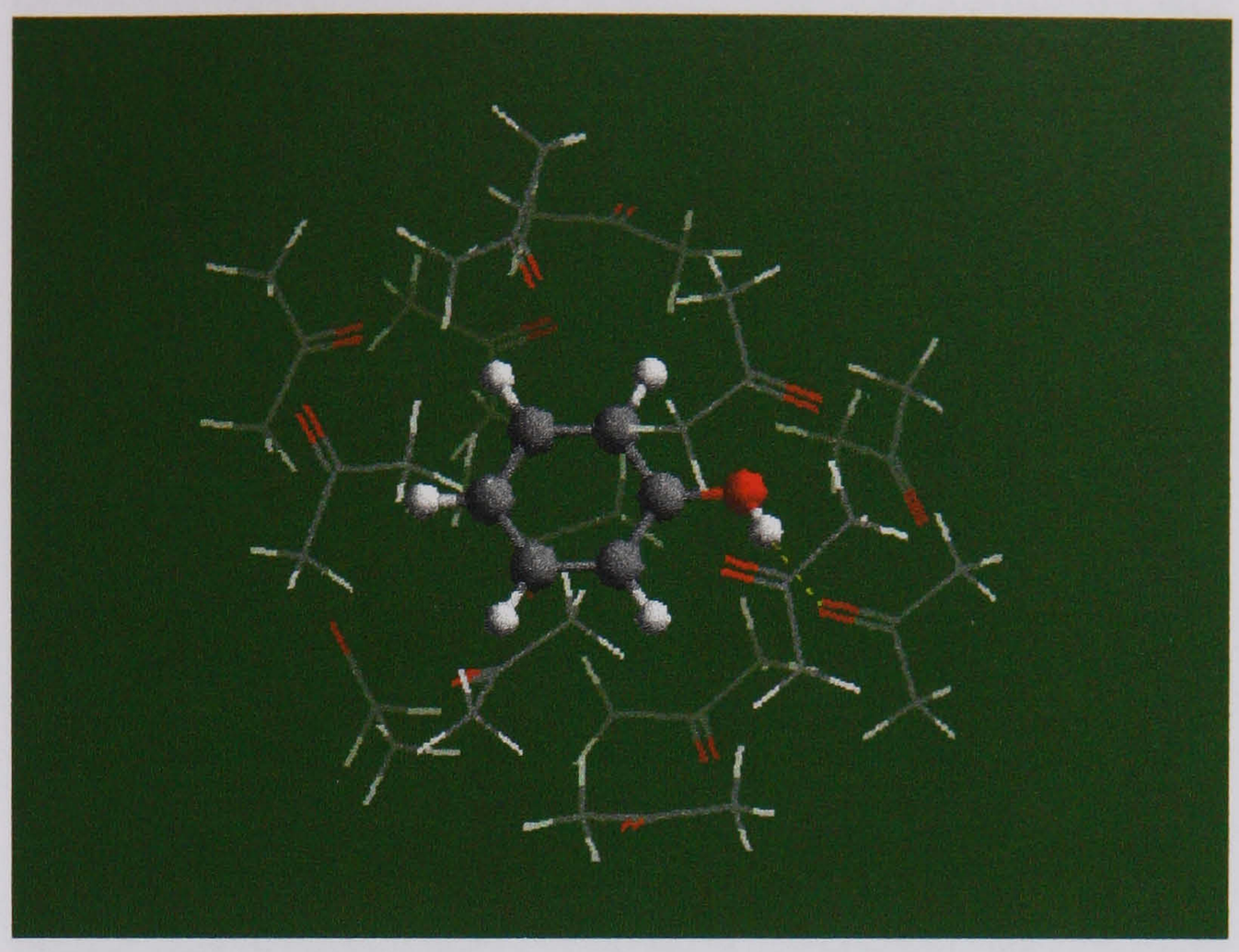


Figure 3.19: Cluster of phenol solute surrounded by 14 acetone molecules (solute-solvent interactions) - local energy minimisation from starting configuration

Ethanol

Ethanol has one ethyl group and one hydroxyl group. The ethanol solvent molecules are strongly attracted to each other and also the aromatic ring in chlorobenzene as shown in Fig. 3.20. The ethanol solvent however also has the nonpolar ethyl group so that would allow the ethanol solvent to spread surrounding the solute. The configuration gives a solute-solvent interaction energy of -2.1268 kcal/mol and from that, A_{12} and A_{21} values of 500.3 and -33.1 The γ_1^∞ and γ_2^∞ values are 9.181 and 24.02. There are no experimental data for this system but D-UNIFAC predictions (Gmehling et al., 1993) were 4.567 and 20.51 respectively.

Water

Water molecules are difficult to model as since these molecules are small and polar, they tend to form hydrogen bonds whenever possible and unlike ethanol, as water molecules are small, there is a lack of steric hindrance. Fig. 3.21 shows the configuration of water molecules that gives good result compared with Fig. 3.22. The configuration in Fig.

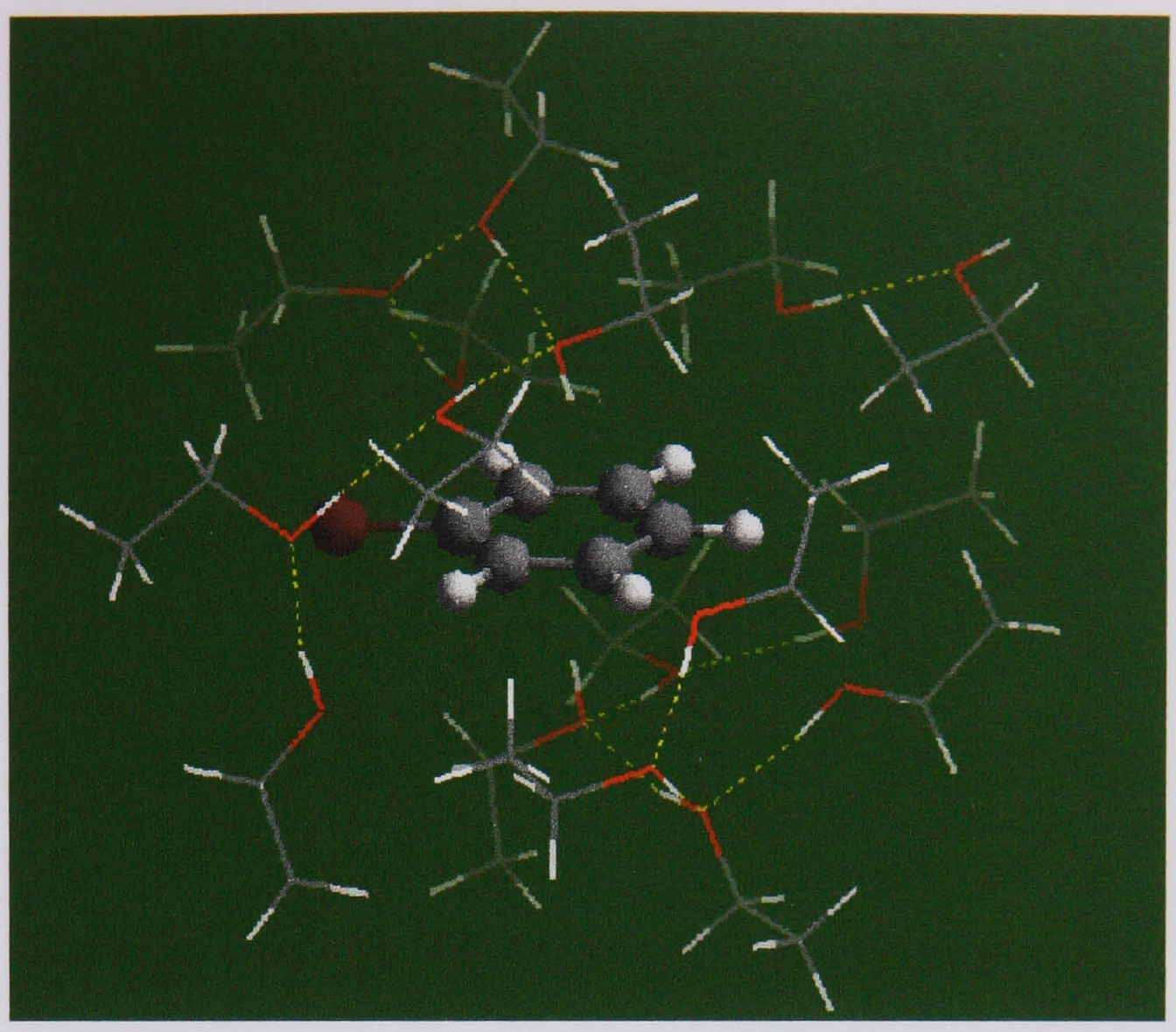


Figure 3.20: Cluster of chlorohexane solute surrounded by 16 ethanol molecules (solute-solvent interactions) - local energy minimisation from starting configuration

3.21 has the solute-solvent interaction energy of -2.9412 kcal/mol that gives γ_1^∞ and γ_2^∞ values of 39.86 and 3.36 respectively from UNIQUAC interaction parameters of 10.8 and 97.4. The experimental values of the infinite activity coefficients are 54 and 2. Using the other configuration, the solute-solvent interaction energy is -2.5553 kcal/mol with the γ_1^∞ and γ_2^∞ values of 197.9 and 7.98. The difference between Fig. 3.21 and Fig. 3.22 is that two of the water molecules that are hydrogen-bonded to the phenol solute have two hydrogen bonds each compared with three each in the latter configuration. As the phenol molecules hydrogen-bonded to the middle phenol solute also have two hydrogen bonds each as shown in Fig. 3.15, Fig. 3.21 more accurately describes the differences in strengths between the solute-solute and solute-solvent interactions and Fig. 3.22 underestimates the solute-solvent interactions.

3.5 Discussion of Methodology

The interaction energies calculated using Molecular Mechanics are necessarily biased because it is assumed that the system exists in a vacuum and thus the bulk solvent effects

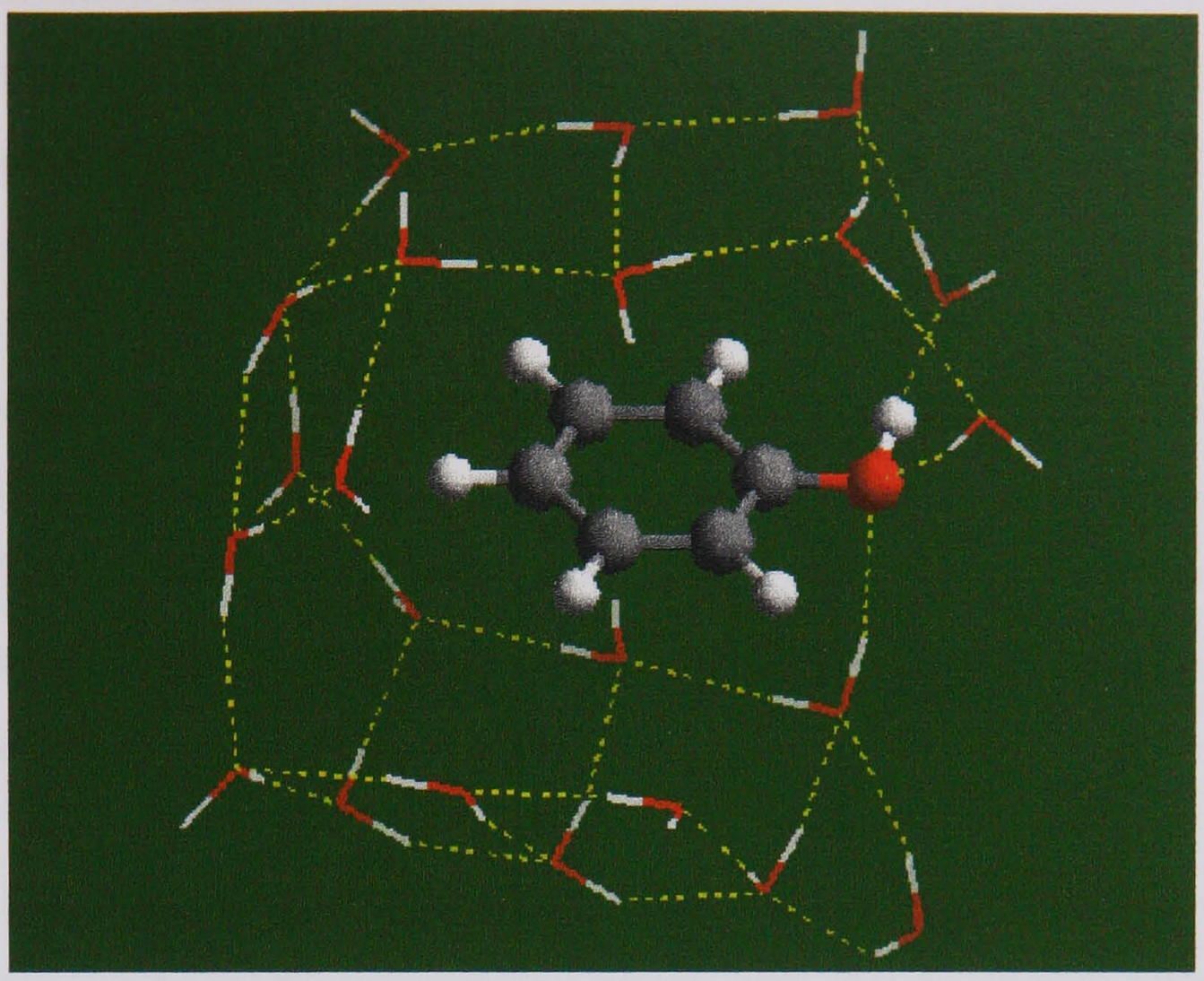


Figure 3.21: Cluster of phenol solute surrounded by 22 water molecules (solute-solvent interactions) - local energy minimisation from starting configuration

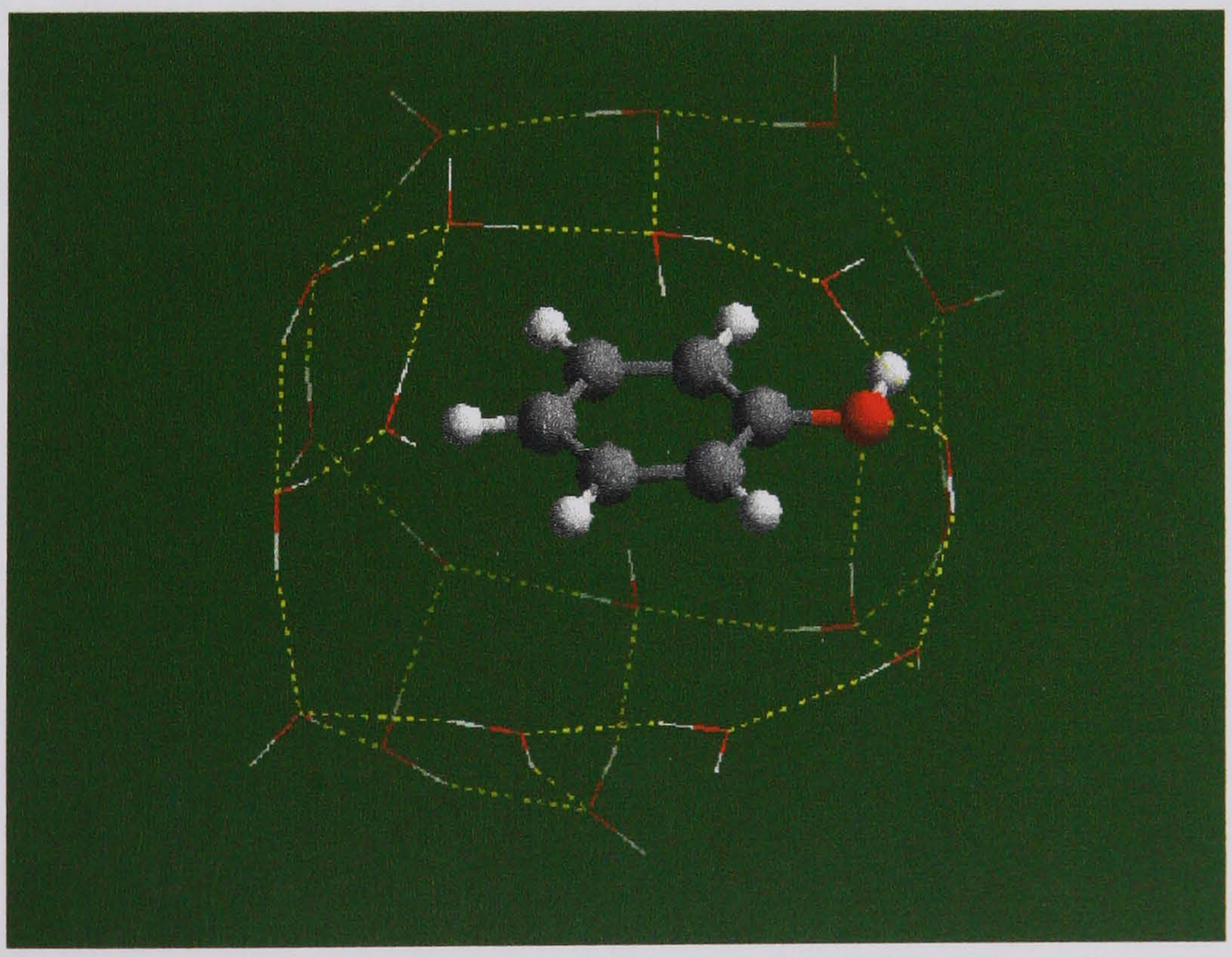


Figure 3.22: Unsuitable model of cluster of phenol solute surrounded by 22 water molecules (solute-solvent interactions) - local energy minimisation from starting configuration

on the first solvation shell are neglected. The interaction energies are therefore overestimated. Nonetheless, as the UNIQUAC model is based on the difference between the interaction energies, this approach should still be able to provide a reliable qualitative picture.

Small, polar solvents like water are difficult to model as due to lack of steric hindrance, one water molecule tends to hydrogen-bond with up to four of its neighbouring molecules in the first solvation shell. This causes the solvation shell to expand out to accommodate all the solvent molecules and decreases the strength of the solute-solvent interactions. The model thus has to be constructed carefully.

The specific interactions are also slightly overestimated as for example in Fig. 3.15, the middle solute molecule forms two hydrogen bonds with two other molecules. However in real systems, not all solute molecules will have two hydrogen bonds.

Nonetheless, the cluster-model of the solvation shell seems qualitatively reliable in predicting the infinite dilution activity coefficients for complex molecules compared with using pairs. This method also allows for the reliable and quick ranking of potential solvents. The interaction energy is calculated as the Boltzmann average of the pairwise interaction energies for the cluster. Case studies using mono- and bifunctional solutes are shown in the next chapter.

Chapter 4

Case Studies

4.1 Introduction

This chapter will present some case studies to illustrate the use of the prediction method. Firstly, infinite dilution activity coefficients for mono- and bifunctional solutes will be compared with experimental data. Secondly, as the structure of polymorphs is dependent on the interaction with the solvent molecules, the solubilities of two polymorphic forms of paracetamol in a variety of solvents are compared for a range of temperatures.

4.2 Infinite Dilution Activity Coefficients Predictions

Two solubility databases developed specifically with pharmaceutical solutes in mind (Kolar et al., 2002) are used to test the methodology developed. The solubility databases consist of mono- and bifunctional compounds based on core molecular fragments of common drugs and containing substituents of gradually increasing polarity. The solubilities of these compounds are then studied in various solvents of varying polarity and hydrogen-bonding tendency. The aim of the databases is to systematically investigate the solubility of pharmaceutical compounds and their intermediates in a wide range of solvents. As most pharmaceutical solutes have aromatic rings and heteroatoms, both solubility databases have a benzene ring as the core fragment. The first database consists of monofunctional benzene derivatives containing O, N and Group VII substituents, and the second consists of bifunctional derivatives also containing these substituents.

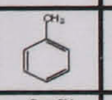
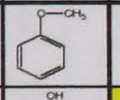
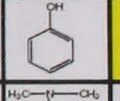
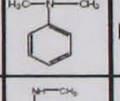
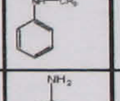
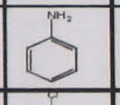
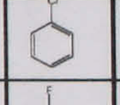
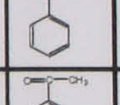
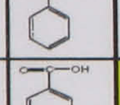
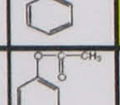
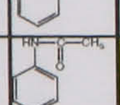
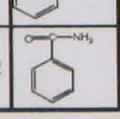
Substituent R_1		Structure	Solute (1)	DDB #	Ideal solubility @ 25 C	Gamma infinity	Solvent (2) DDB #	89	91	50	31	47	157	12	75	21	4	72	151	110	11	8	174	84	144	3
								Hexane	Heptane	Cyclo-hexane	Benzene	Chloro-form	CCl4	Di-Et ether	1,4-Dioxane	AcOEt	Acetone	DMF	DMSO	MeOH	EtOH	1,2-EG	H2O	Acetic acid	Pyridine	Aceto-nitrile
-CH3	-CH3		Toluene	161		(1) in (2)																				
						(2) in (1)																				
O	- O -		Anisole	18		(1) in (2)																				
						(2) in (1)																				
	- OH		Phenol	138	0.79409	(1) in (2)																				
						(2) in (1)																				
N	- N <		N,N-Dimethyl aniline	63		(1) in (2)																				
						(2) in (1)																				
	- NH -		N-Methyl aniline	325		(1) in (2)																				
						(2) in (1)																				
	- NH2		Aniline	17		(1) in (2)																				
						(2) in (1)																				
X	- Cl		Chlorobenzene	27		(1) in (2)																				
						(2) in (1)																				
	- F		Fluorobenzene	183		(1) in (2)																				
						(2) in (1)																				
CO	- CO -		Acetophenone	372		(1) in (2)																				
						(2) in (1)																				
	- COOH		Benzoic acid	425 M1122	0.16708	(1) in (2)																				
						(2) in (1)																				
CO-R	- CO-O-		Phenylacetate	1368 M7421		(1) in (2)																				
						(2) in (1)																				
	-CO-NH-		Acetanilide	833 M47	0.13356	(1) in (2)																				
						(2) in (1)																				
	-CO-NH2		Benzamide	2753, M1087	0.13415	(1) in (2)																				
						(2) in (1)																				
						δ $(MPa)^{1/2}$	14.9	15.1	16.8	18.74	19.42	17.49	15.6	20.3	18.6	20.5	24.8	26.4	29.7	26	29.1	48.13	20.7	21.7	24.77

Table 4.1: Solubility database for monofunctional benzene derivatives

4.2.1 Monofunctional Solutes

The monofunctional benzene derivatives database consists of 13 solutes in 17 solvents as shown in Table 4.1. Commonly used solutes and solvents are highlighted in yellow and combinations for phase equilibria data that are available in standard databases (Merck (Budavari et al., 2000), DDB [1], Beilstein [2]) are shaded in orange. From the database, the infinite activity coefficients of six solutes in four different solvents were predicted using the methodology described in Section 3.4 and compared with experimental data at 298.15K. In solvent selection, the infinite activity coefficients are usually used in the first approximation as a measure of the strength of interactions in the liquid phase. The six solutes are toluene with a nonpolar methyl group, phenol with a hydroxyl group, aniline with an amino group, chlorobenzene, benzoic acid with a carboxyl group and acetanilide with amino, carbonyl and methyl groups. The solvents are *n*-hexane, acetone, ethanol and water.

The experimental data are from three sources obtained from Petr Kolar from Mitsubishi Chemicals Corporation (MCC):

- Experimental measurements of the activity coefficients at infinite dilution (ACT),
- Regression of vapour-liquid (VLE) and liquid-liquid (LLE) equilibrium data,
- Interpolation of solid solubility data (SLE).

Complete VLE data at isothermal conditions where available were regressed to obtain the infinite activity coefficients data. Otherwise, incomplete data were fitted by the UNIQUAC equation and extrapolated to infinite dilution. SLE data were interpolated to 298.15K.

Our prediction methodology, the Molecular Mechanics method (MM), was also compared with two other prediction methodologies: D-UNIFAC (Gmehling et al., 1993) and quantum mechanical (QM)/ COSMO methods (Klamt and Eckert, 2000), both calculated by

[1] Dortmund Databank ver. 2000, DDBST GmbH, Oldenburg (http://www.ddbst.com)
[2] Beilstein Informationssysteme GmbH, Frankfurt am Main (http://www.beilstein.com)

Petr Kolar from MCC. Table 4.2 compares the infinite dilution activity coefficients predicted using the different techniques and the results are shown graphically in Fig. 4.1. Only the predictions that significantly deviate from experimental data are discussed.

In the case of nonpolar *n*-hexane, our method underpredicts the solubility of benzoic acid. As only the first solvation shell is accounted for, the interaction energies are slightly overestimated due to the fact that the effect of the surrounding molecules on the first solvation shell is neglected. However, the errors between the more polar molecules are larger than those of the nonpolar molecules. Due to the discrepancies between these errors, the solubilities for dissimilar systems like benzoic acid/ *n*-hexane deviate from experimental data. The D-UNIFAC method underpredicts the solubility of acetanilide in *n*-heptane. The acetanilide/*n*-heptane system was chosen as experimental data was not available for the acetanilide/*n*-hexane system.

As for using acetone as the solvent, both our method and COSMO-RS overpredict the solubilities of aniline, phenol and benzoic acid as the solute-solvent interactions were overestimated more than the solvent-solvent interactions. Again this is due to the strong specific interactions between the solute and solvent molecules and also the lack of specific interactions between the solvent molecules. The D-UNIFAC method underestimates the solubility of acetanilide.

In the case of ethanol, again the solubilities for aniline, phenol and benzoic acid for COSMO-RS were overpredicted. However, our method performed quite well as similar interactions exist between the solute-solvent and solvent-solvent systems. Therefore, the errors cancel out. D-UNIFAC slightly underestimated the solubility of benzoic acid.

As for water, our method performs poorly for the nonpolar solutes, toluene and chlorobenzene, as the solubilities were overestimated. COSMO-RS results were comparable with experimental data. D-UNIFAC underpredicted the solubility for the benzoic acid system.

In summary, our method seems to perform better for systems of similar polarity, com-

paring the results obtained for the polar systems in acetone and in ethanol. The binary parameters in the UNIQUAC equation are based on the difference between the solute-solvent interactions, and the solute-solute and solvent-solvent interactions respectively and for similar systems, any errors are minimised. The interaction energies calculated are shown in Table 4.3 and the UNIQUAC binary parameters calculated from these interaction energies are shown in Table 4.4. Water is hard to model due to its small size and tendency to hydrogen-bond with itself as explained in the Section 3.4.2. Fig. 4.2 however shows that the Molecular Mechanics prediction method proposed in this work is comparable to D-UNIFAC and COSMO-RS in predicting infinite activity coefficients for monofunctional solutes with rms relative error [3] of 25.38 for D-UNIFAC, 124.27 for COSMO-RS and 27.12 for our method. The high rms value for COSMO-RS was due to the results obtained for the toluene/ water and chlorobenzene/ water systems. Without these systems, the rms relative error for COSMO-RS is 11.54. For most systems, the activity coefficients show qualitative trends in accordance with experimental data as shown in Fig. 4.2. The next step is to extend the methodology to bifunctional compounds.

The method also allows for the qualitative ranking of solvents for the different solutes. In most cases, the predictions for all three methods are qualitatively accurate. Our method however predicts chlorobenzene to be more soluble in n-hexane than acetone and benzoic acid to be less soluble in n-hexane than water. D-UNIFAC and COSMO-RS predict toluene to be more soluble in acetone than n-hexane and D-UNIFAC also predicts that acetanilide is less soluble in acetone than ethanol.

---

[3] The root mean square relative error is calculated using the equation $\sum_{i=1}^{N} \sqrt{\frac{(x_i - x_i^{exp})^2}{N}}$ where x_i is the value of each point, x_i^{exp} is the experimental data point and N is the total number of data points

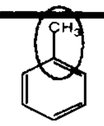
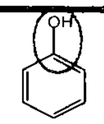
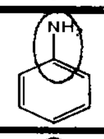
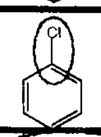
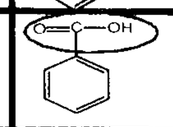
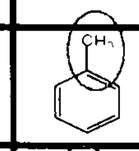
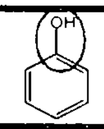
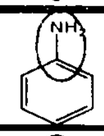
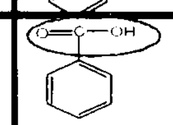
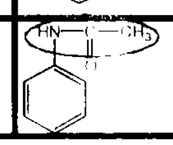
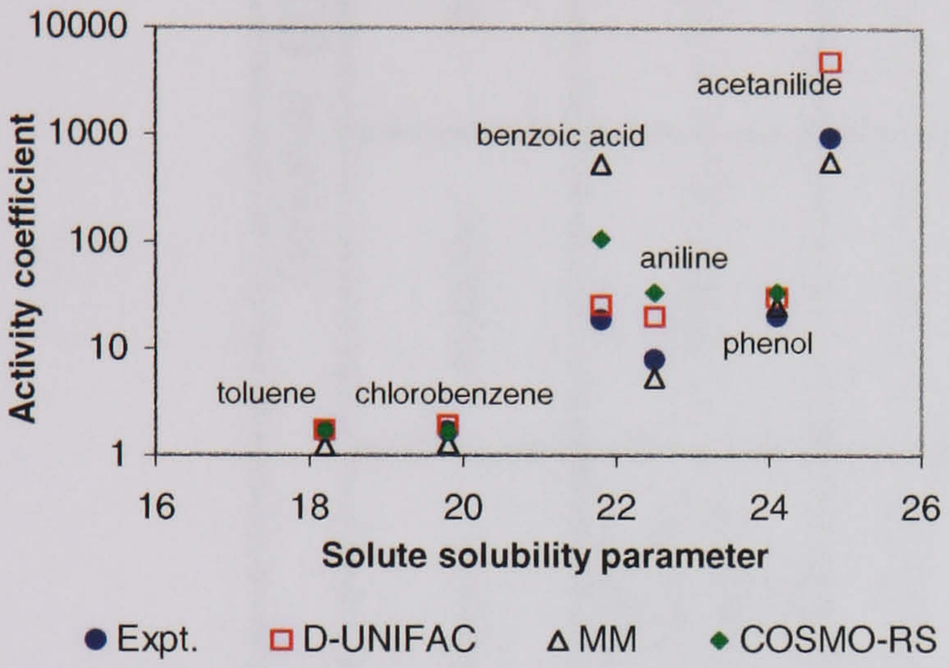
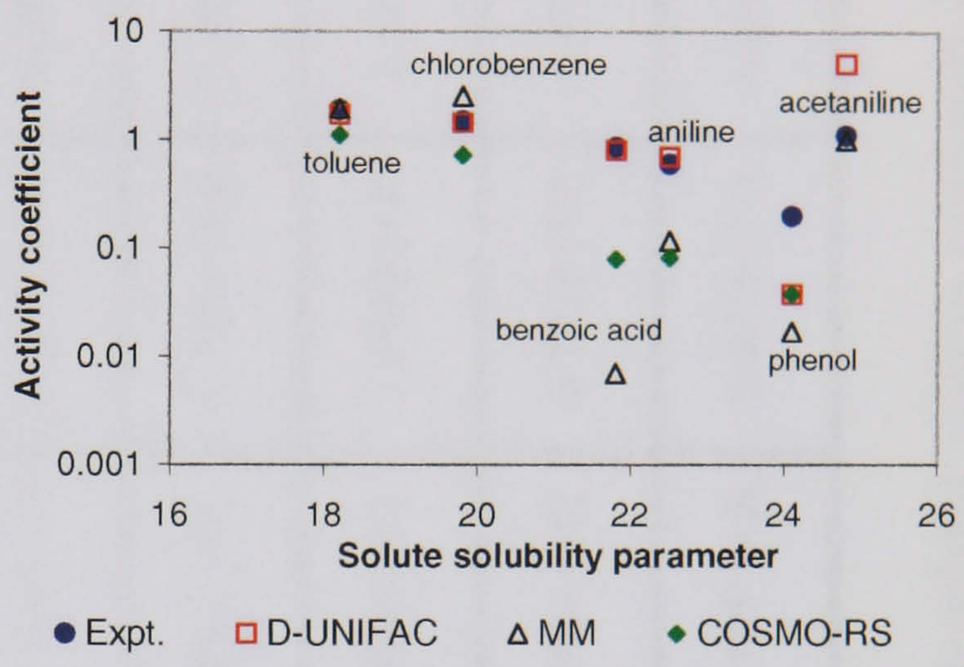
				Solvent (2)							
				Exp	D-UNIFAC	COSMO-RS	MM	Exp	D-UNIFAC	COSMO-RS	MM
Substituent	Structure	**Solute (1)**	**Gamma infinity**	**Hexane**	**Hexane**	**Hexane**	**Hexane**	**Acetone**	**Acetone**	**Acetone**	**Acetone**
-CH3		**Toluene**	(1) in (2)	1.640	1.726	1.707	1.215	1.89, 2.09	1.716	1.108	1.961
			(2) in (1)	1.870	1.717	1.641	1.281	1.98	1.752	1.449	1.722
- OH		**Phenol**	(1) in (2)	20.3	30.67	34.36	25.28	0.2	0.038	0.038	0.017
			(2) in (1)	15.2	14.53	8.1	20.92	0.4	0.070	0.016	0.140
- NH2		**Aniline**	(1) in (2)	8	20.24	33.69	5.376	0.6	0.701	0.082	0.117
			(2) in (1)	26.63	20.89	13.79	6.294	0.64	0.658	0.113	0.274
- Cl		**Chlorobenzene**	(1) in (2)	1.70	1.935	1.662	1.287	1.49	1.445	0.723	2.542
			(2) in (1)	1.85	2.704	1.616	1.389	1.55	1.496	1.005	2.131
- COOH		**Benzoic acid**	(1) in (2)	18.7	25.66	107.3	539.5	0.857	0.799	0.079	0.007
			(2) in (1)	9.93	12.60	5.907	17.83		0.850	0.1	0.246
-CO-NH-		**Acetanilide**	(1) in (2)	954*	4946*		570.2*	1.109	5.114		1.010
			(2) in (1)		24.34*		30.16*		2.053		0.760
				EtOH	**EtOH**	**EtOH**	**EtOH**	**H2O**	**H2O**	**H2O**	**H2O**
-CH3		**Toluene**	(1) in (2)	5.77	6.250	4.393	5.573	9165	7963.9	7922	1401
			(2) in (1)	15.4, 18.4	18.038	54.29	18.43	391.7	159.17	15650	131.7
- OH		**Phenol**	(1) in (2)	0.2	0.604	0.0655	0.437	54	42.08	52	39.86
			(2) in (1)	0.3 - 0.4	0.421	0.0684	0.445	2	2.807	1.518	3.36
- NH2		**Aniline**	(1) in (2)	2.84	3.443	0.514	4.393	143.6	137.4	161.1	264.6
			(2) in (1)	2.4	2.871	1.444	5.191	8.15, 6.65	6.327	29.82	12.94
- Cl		**Chlorobenzene**	(1) in (2)		4.568	3.597	9.181	12960	3325	7044	2703
			(2) in (1)		20.52	46.26	24.02	329.7	35.18	13049	261.6
- COOH		**Benzoic acid**	(1) in (2)	0.967	3.170	0.1548	0.870	136	646.9	194.2	352.7
			(2) in (1)		2.067	0.245	0.863	2.61	9.026	3.208	2.751
-CO-NH-		**Acetanilide**	(1) in (2)	1.97, 1.10	2.558	1.014	1.980	182.3	205.5		343.8
			(2) in (1)		1.519	1.532	1.518		2.469		2.228

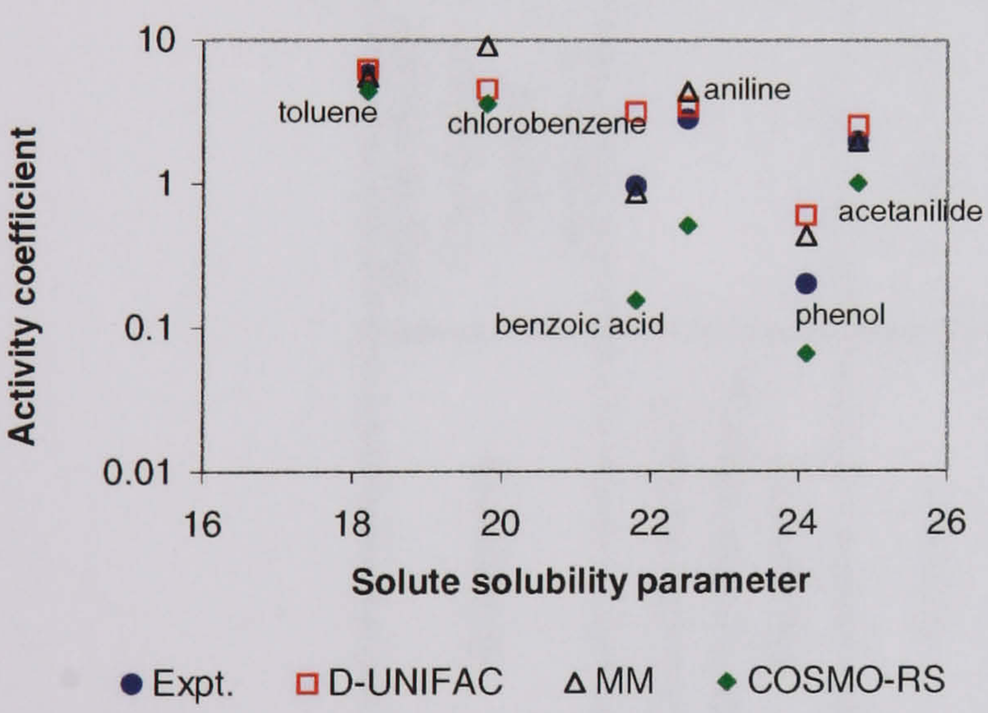
Table 4.2: Comparison of infinite dilution activity coefficients for monofunctionals. * denotes that n-heptane was the solvent instead of n-hexane



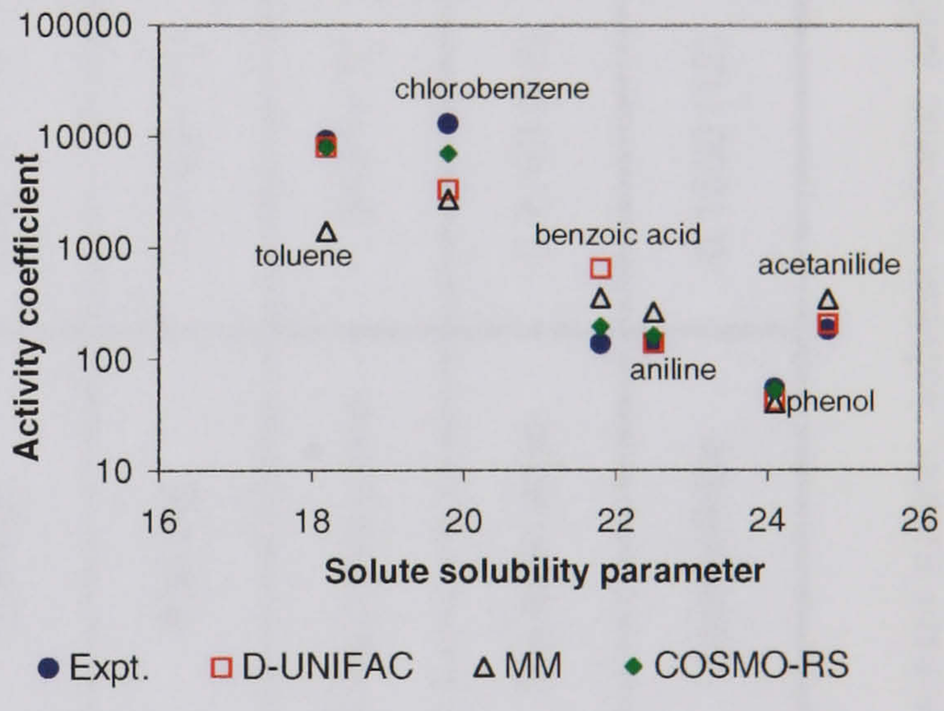
(i)



(ii)



(iii)



(iv)

Figure 4.1: Comparisons of infinite dilution activity coefficients of 6 monofunctional solutes in (i) n-hexane, (ii) acetone, (iii) ethanol and (iv) water at 298.15K

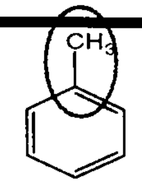
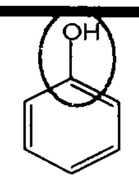
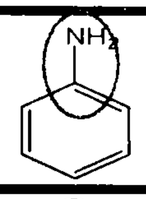
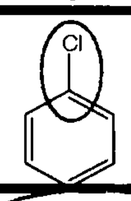
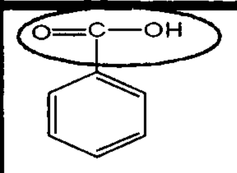
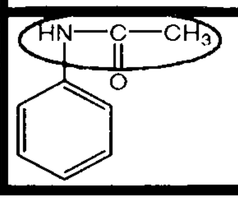
				Solvent (2)			
Substituent	**Structure**	**Solute (1)**	**Solute-solute interaction energy (kcal/mol)**	**Hexane**	**Acetone**	**EtOH**	**H2O**
		Solvent-solvent interaction energy (kcal/mol)		-1.9663 (-2.4243 for n-heptane)* (12)	-1.9782 (12)	-4.0873 (12)	-2.9700 (10)
-CH3		**Toluene**	-1.7591 (12)	-1.8192 (12)	-1.7332 (15)	-2.2730 (16)	-0.8431 (23)
- OH		**Phenol**	-3.4587 (12)	-1.7425 (11)	-3.3642 (14)	-4.0679 (15)	-2.9412 (22)
- NH2		**Aniline**	-2.7728 (12)	-1.8882 (11)	-2.7964 (14)	-3.0018 (15)	-2.1423 (22)
- Cl		**Chlorobenzene**	-1.9404 (12)	-1.9093 (11)	-1.7570 (14)	-2.1268 (15)	-0.5787 (22)
- COOH		**Benzoic acid**	-6.2789 (12)	-2.2952 (12)	-3.7212 (15)	-4.7720 (16)	-3.3296 (25)
-CO-NH-		**Acetanilide**	-6.1560 (12)	-2.4146* (12)	-3.5576 (17)	-4.5443 (19)	-3.4301 (29)

Table 4.3: Interaction energies for solute-solute, solvent-solvent and solute-solvent systems calculated using our methodology. * corresponds to calculations for n-heptane rather than n-hexane and the number in brackets corresponds to the number of molecules in the first solvation shell

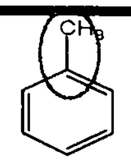
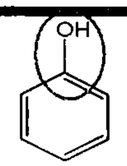
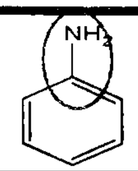
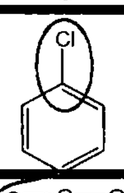
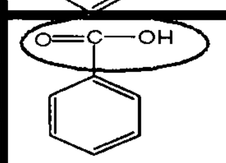
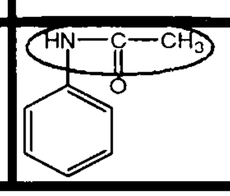
					Solvent (2)			
Substituent	**Structure**	**Solute (1)**	**UNIQUAC parameters r, q**	**UNIQUAC parameters**	**Hexane**	**Acetone**	**EtOH**	**H2O**
		UNIQUAC parameters r, q			4.500 (5.310)*	2.574	2.106	0.920
					3.856 (4.249)*	2.336	1.972	1.400
-CH3		**Toluene**	3.923	A_{12}	19.19	52.77	463.0	764.5
			2.968	A_{21}	-10.20	4.392	-87.14	155.3
- OH		**Phenol**	3.552	A_{12}	29.21	-298.6	4.962	10.35
			2.680	A_{21}	322.3	17.75	-114.4	97.18
- NH2		**Aniline**	3.717	A_{12}	10.20	-176.3	277.0	297.5
			2.816	A_{21}	158.1	-4.219	-40.92	112.7
- Cl		**Chlorobenzene**	3.813	A_{12}	7.439	47.65	500.3	859.6
			2.844	A_{21}	5.495	32.45	-33.05	240.9
- COOH		**Benzoic acid**	4.323	A_{12}	-42.92	-375.5	-174.7	-129.3
			3.344	A_{21}	599.5	384.9	226.8	443.8
-CO-NH-		**Acetanilide**	5.227	A_{12}	1.111*	-340.3	-116.6	-165.4
			4.004	A_{21}	470.2*	326.6	202.6	342.6

Table 4.4: UNIQUAC binary parameters calculated from the interaction energies shown in Table 4.3. * corresponds to calculations for n-heptane rather than n-hexane

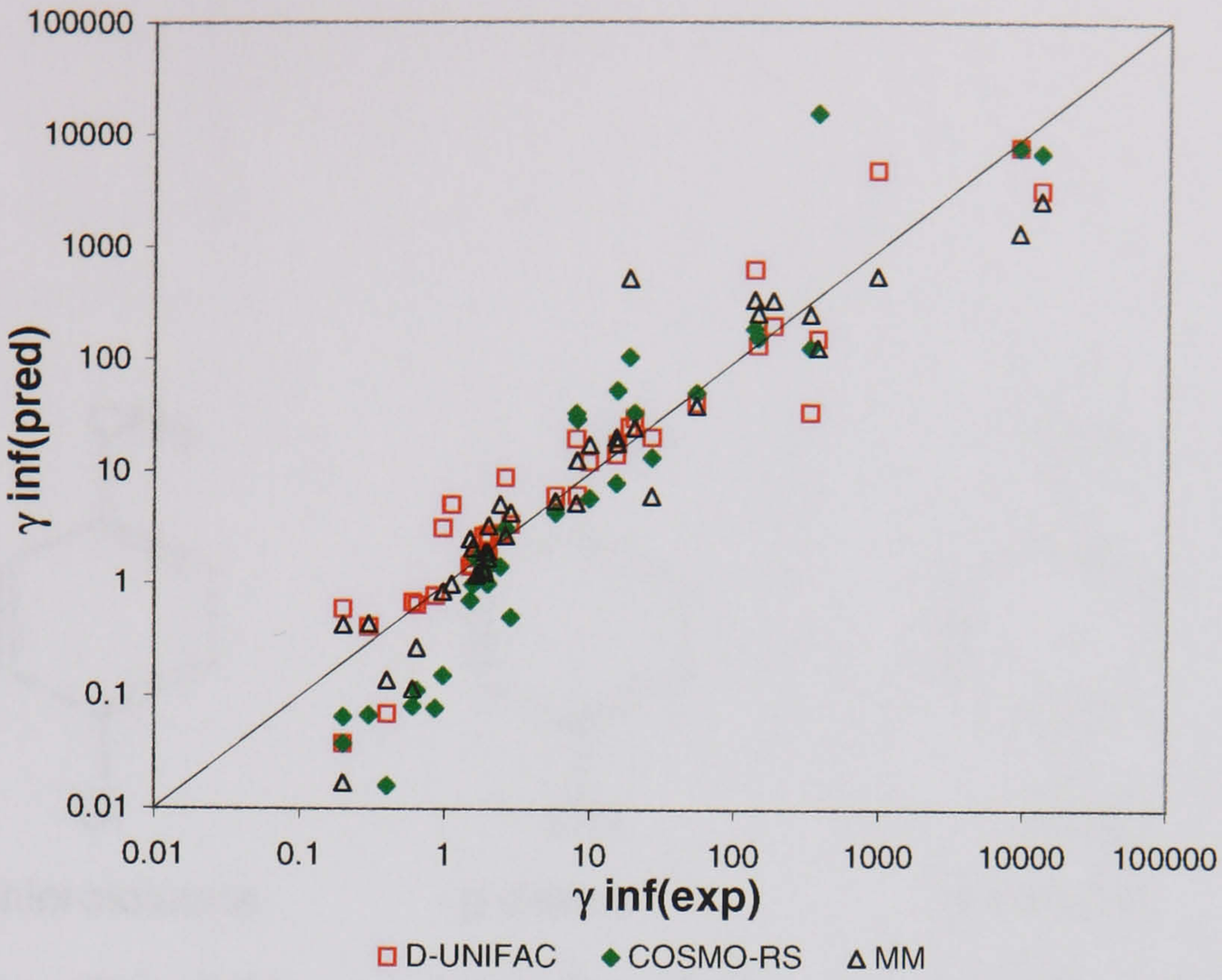


Figure 4.2: Graph showing correlation between infinite dilution activity coefficients, γ_1^∞, between experimental data with predicted data (some data points are out of bounds) for monofunctional solutes

4.2.2 Bifunctional Solutes

We have chosen to work with seven bifunctional benzene derivatives shown in Fig. 4.3: *p*-chlorotoluene, *p*-cresol, *p*-toluidine, *p*-chlorobenzoic acid, *p*-hydroxybenzoic acid, *p*-amino benzoic acid and paracetamol. Again the solubilities of these derivatives in the four solvents, *n*-hexane, acetone, ethanol and water, are compared with experimental data. Experimental infinite dilution activity coefficients are determined from the experimental results shown in *Appendix D* where only the experimental point with the temperature closest to 298.15K was used to calculate the infinite dilution activity coefficients. Therefore, these values only provide a rough estimate. Not all experimental data were available as shown in Table 4.5 and graphically in Fig. 4.4. Comparison is also made with D-UNIFAC except for *p*-aminobenzoic acid as D-UNIFAC lacked the necessary group interaction parameters. This is one of the main limitations of group contribution methods. In the case of *p*-chlorobenzoic acid and *p*-hydroxybenzoic acid, the solvent *n*-heptane was used instead of *n*-hexane as experimental data was available for these systems.

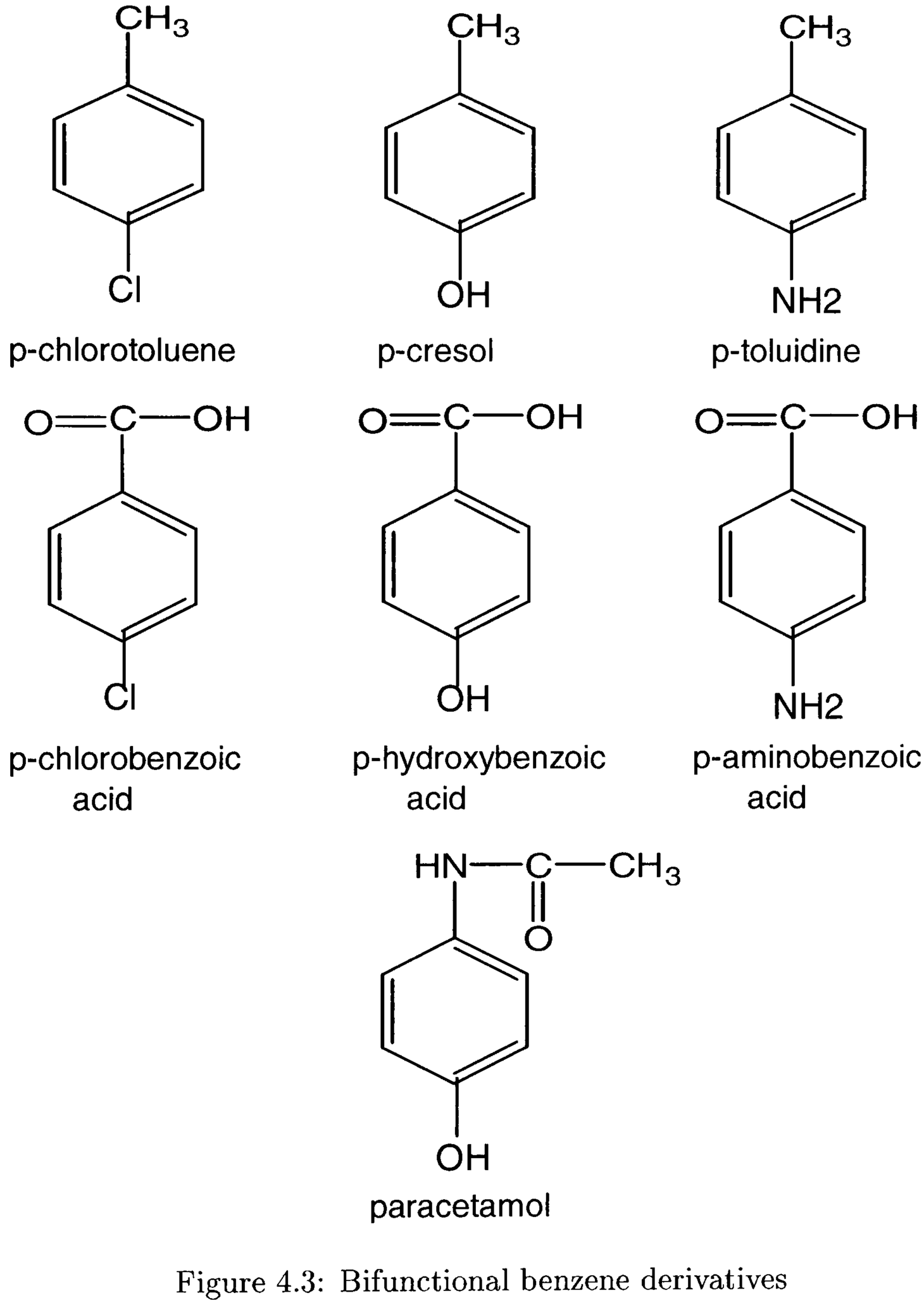


Figure 4.3: Bifunctional benzene derivatives

For the systems where n-hexane or n-heptane was the solvent, the solubility of paracetamol using D-UNIFAC was significantly underestimated. As for our methodology, p-chlorobenzoic acid in n-heptane was underestimated as solute-solute interaction energy is likely overestimated due to the fact that the chlorine atoms repel each other thus increasing the strength of the hydrogen-bond interaction between the carboxylic acid groups. On the other hand, the hydroxyl groups in p-hydroxybenzoic acid attract each other due to hydrogen-bonding as well as the carboxylic groups but as the benzene ring prevents the groups from coming close to the each other, the strength of the hydrogen-bonds are lessened. Therefore, although p-hydroxybenzoic acid has more hydrogen bonding sites than p-chlorobenzoic acid, the solute-solute interaction energy is weaker.

Our prediction method also overpredicts the solubility of the polar solutes in acetone as in the case of the monofunctional solutes. The solubility of p-hydroxybenzoic acid in ethanol is also overestimated. The interaction energies of the solute-solvent, solute-solute and solvent-solvent systems are shown in Table 4.6 and the UNIQUAC binary parameters calculated from these energies are shown in Table 4.7. Fig. 4.5 provides an overall comparison of the predicted data with experimental data. D-UNIFAC has a rms relative error of 18.66 and the Molecular Mechanics method has one of 12.03.

			Solvent (2)					
			Exp	D-UNIFAC	MM	Exp	D-UNIFAC	MM
Substituent	**Solute (1)**	**Gamma infinity**	**Hexane**	**Hexane**	**Hexane**	**Acetone**	**Acetone**	**Acetone**
-CH3, -Cl	**p-chlorotoluene**	(1) in (2)	0.5870	2.022	1.664		1.871	1.736
		(2) in (1)	5.880	2.101	1.685		1.774	1.380
-CH3, -OH	**p-cresol**	(1) in (2)		26.96	12.87		0.529	0.0032
		(2) in (1)		8.296	8.570		0.352	0.1141
-CH3, -NH2	**p-toluidine**	(1) in (2)		16.34	4.257		0.468	0.439
		(2) in (1)		9.68	3.647		0.322	0.646
- COOH, -Cl	**p-chlorobenzoic acid**	(1) in (2)	60.61*	36.47*	239.6*		1.192	0.0355
		(2) in (1)	9.75*	31.56*	24.24*		1.118	0.3700
- COOH, -OH	**p-hydroxybenzoic acid**	(1) in (2)	1515*	567.5*	258.8*		0.031	0.0013
		(2) in (1)	30.07*	116.4*	30.07*		0.116	0.1827
- COOH, -NH1	**p-aminobenzoic acid**	(1) in (2)			162.5			0.0011
		(2) in (1)			12.80			0.1770
-NH-CO-, -OH	**paracetamol**	(1) in (2)		2.075e+6	810.8	3.902	3.366	0.0056
		(2) in (1)		113.0	17.08	1.046	0.981	0.2706
			EtOH	**EtOH**	**EtOH**	**H2O**	**H2O**	**H2O**
-CH3, -Cl	**p-chlorotoluene**	(1) in (2)		5.950	16.57		12089	9438
		(2) in (1)		20.50	26.74		48.42	285.1
-CH3, -OH	**p-cresol**	(1) in (2)		0.576	0.6220	96.56	129.2	34.61
		(2) in (1)		0.342	0.7144	3.552	3.49	2.357
-CH3, -NH2	**p-toluidine**	(1) in (2)	4.823	3.440	10.01	81.00	387.5	506.6
		(2) in (1)	4.843	2.620	8.573	5.872	7.40	12.61
- COOH, -Cl	**p-chlorobenzoic acid**	(1) in (2)		4.152	4.006	173.1	1341	194.9
		(2) in (1)		2.429	1.469	2.174	6.35	2.227
- COOH, -OH	**p-hydroxybenzoic acid**	(1) in (2)	0.1468	0.466	0.0587	22.95	16.65	6.331
		(2) in (1)	0.4805	0.428	0.3634	1.525	1.60	1.205
- COOH, -NH2	**p-aminobenzoic acid**	(1) in (2)	5.815		9.380	74.10		15.03
		(2) in (1)	1.865		2.272	1.932		1.403
-NH-CO-, -OH	**paracetamol**	(1) in (2)	1.914	6.426	5.328	48.88	90.19	156.2
		(2) in (1)	1.017	1.939	1.383	1.448	2.302	1.758

Table 4.5: Comparison of infinite dilution activity coefficients for bifunctionals. * denotes that n-heptane was the solvent instead of n-hexane

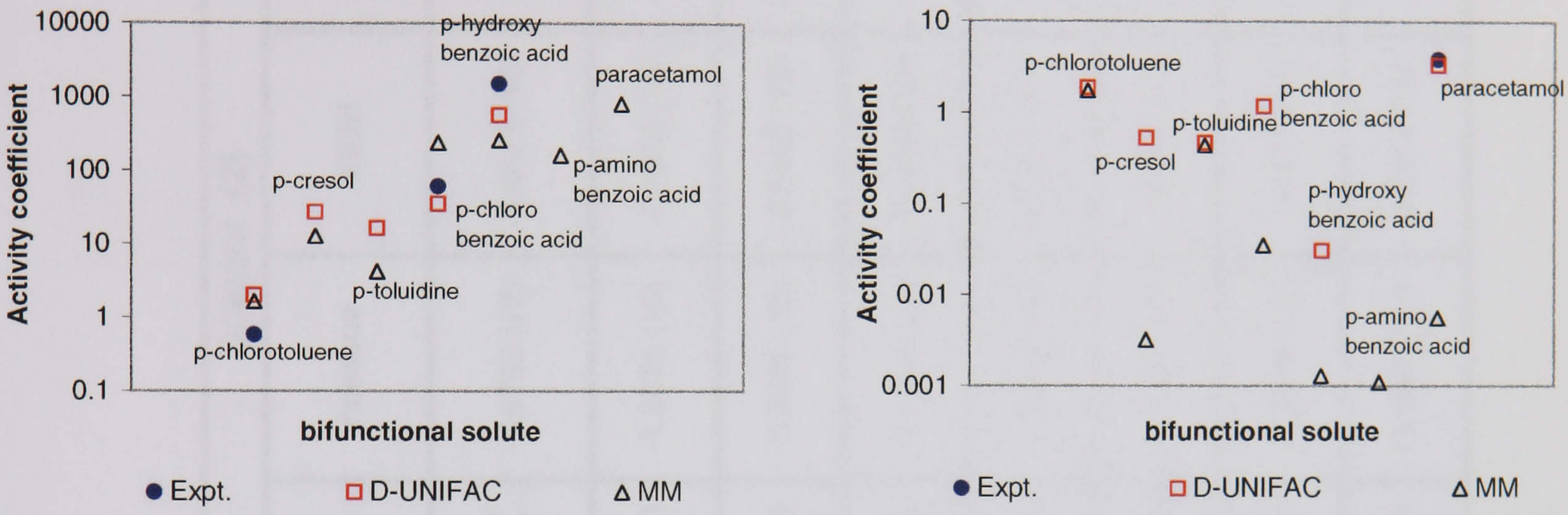


Figure 4.4: Comparisons of infinite dilution activity coefficients of 7 bifunctional solutes in (i) n-hexane, (ii) acetone, (iii) ethanol and (iv) water at 298.15K

			Solvent (2)			
Substituent	**Solute (1)**	**Solute-solute interaction energy (kcal/mol)**	**Hexane**	**Acetone**	**EtOH**	**H2O**
	Solvent-solvent interaction energy (kcal/mol)		-1.9663 (-2.4243 for n-heptane)* (12)	-1.9782 (12)	-4.0873 (12)	-2.9700 (10)
-CH3, -Cl	**p-chlorotoluene**	-2.1324 (12)	-1.9224 (12)	-1.9075 (15)	-2.1054 (16)	-0.3601 (25)
-CH3, -OH	**p-cresol**	-3.2038 (12)	-1.8398 (12)	-3.3326 (15)	-3.8108 (16)	-2.8608 (25)
-CH3, -NH2	**p-toluidine**	-2.8139 (12)	-1.9724 (12)	-2.4303 (15)	-2.7558 (16)	-1.9682 (25)
- COOH, -Cl	**p-chlorobenzoic acid**	-5.7091 (12)	(-2.4512)* (12)	-3.4009 (16)	-4.2915 (17)	-3.2974 (27)
- COOH, -OH	**p-hydroxybenzoic acid**	-5.5344 (12)	(-2.3259)* (12)	-3.7088 (16)	-4.8811 (17)	-3.5748 (27)
- COOH, -NH2	**p-aminobenzoic acid**	-5.3888 (12)	-2.1289 (12)	-3.6819 (16)	-4.0217 (17)	-3.4667 (27)
-CO-NH-, -OH	**paracetamol**	-6.2816 (12)	-2.1443 (12)	-3.5641 (17)	-4.3604 (19)	-3.3688 (30)

Table 4.6: Interaction energies for solute-solute, solvent-solvent and solute-solvent systems calculated using our methodology. * corresponds to calculations for n-heptane rather than n-hexane and the number in brackets correspond to the number of molecules in the first solvation shell

Substituent	Solute (1)	UNIQUAC parameters r, q	UNIQUAC parameters	Solvent (2) Hexane	Acetone	EtOH	H2O
				4.500 (5.310)*	2.574	2.106	0.920
				3.856 (4.249)*	2.336	1.972	1.400
-CH3, -Cl	**p-chlorotoluene**	4.632	A_{12}	5.729	15.23	505.8	938.2
		3.604	A_{21}	29.32	31.40	3.770	247.5
-CH3, -OH	**p-cresol**	4.420	A_{12}	16.51	-291.8	66.91	39.25
		3.478	A_{21}	197.4	-18.60	-87.83	49.63
-CH3, -NH2	**p-toluidine**	4.444	A_{12}	-0.7961	-97.40	339.8	360.1
		3.572	A_{21}	118.6	54.04	8.186	119.1
- COOH, -Cl	**p-chlorobenzoic acid**	5.119	A_{12}	-3.186*	-306.5	-52.11	-117.7
		3.905	A_{21}	419.8*	297.5	182.7	310.8
- COOH, -OH	**p-hydroxybenzoic acid**	4.771	A_{12}	11.65*	-372.8	-202.6	-217.4
		3.764	A_{21}	429.0*	244.1	87.35	262.0
- COOH, -NH2	**p-aminobenzoic acid**	4.858	A_{12}	-21.22	-367.0	16.74	-178.5
		3.829	A_{21}	428.4	224.3	179.7	252.6
-CO-NH-, -OH	**paracetamol**	5.545	A_{12}	-23.23	-341.7	-69.69	-143.4
		4.344	A_{21}	479.3	314.8	222.6	337.4

Table 4.7: UNIQUAC binary parameters calculated from the interaction energies shown in Table 4.6. * corresponds to calculations for n-heptane rather than n-hexane

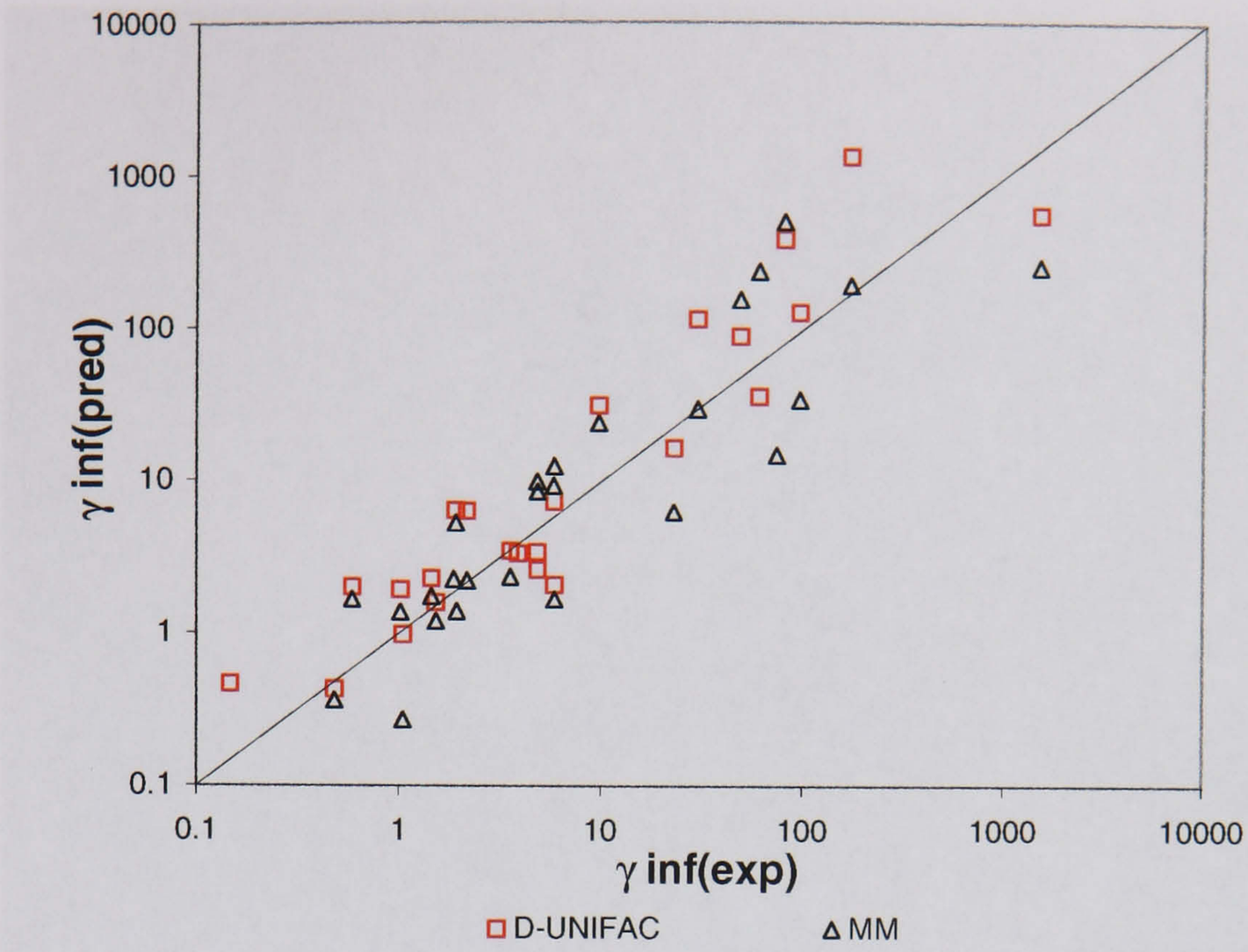


Figure 4.5: Graph showing correlation between infinite dilution activity coefficients, γ_1^∞, between experimental data with predicted data (some data points are out of bounds) for bifunctional solutes

4.3 Polymorph Solubility Prediction

As the methodology takes into account solute structure in determining the solute-solute interaction energies, we also evaluated the possibility of determining the stability of the polymorphs of paracetamol as a function of the solvent and temperature. Many molecular crystals can solidify in multiple crystalline structures. The crystalline form of an active pharmaceutical ingredient can have a large impact on its physical properties and thus its therapeutic performance. Compounds which have the same chemistry but different crystal structures are polymorphic. It is thus important to screen new active ingredients for polymorphic behaviour.

Paracetamol has two known polymorphs: the monoclinic form (Form I) and the orthorhombic form (Form II). Both polymorph crystalline structures differ by the arrangement of the molecules and therefore have different solute-solute interaction energies. The interaction energies are determined by orienting the molecules according to their crystalline structure as shown in Figs. 4.6 (Haisa et al., 1976) and 4.7 (Haisa et al., 1974).



Figure 4.6: Monoclinic form of paracetamol

After minimisation, the binary interaction energy is then calculated with Form I having -4.20 kcal/mol and Form II having -4.45 kcal/mol from Equation 3.3. In order to calculate the mole fraction of the solute dissolved in the solvent according to Equation 2.9, the standard heat of fusion, melting point and heat capacity of the two polymorph forms must be known and have been measured by Sacchetti (2000) as shown in Table 4.8. Although the heat capacity was assumed to be the same for both forms, Form I in actuality has the lower heat capacity but as the heat capacity equation as a function of temperature was not available, it was assumed to be the same and constant.

Polymorph	Monoclinic	Orthorhombic
T_m (K)	441.75	429.55
$\triangle H_f$ (kJ/kmol)	28100	27600
$\triangle C_p$ (kJ/kmol K)	99.8	99.8

Table 4.8: Properties of the two polymorph forms of paracetamol (Sacchetti, 2000)

Using the interaction energies calculated, the solubilities of the two forms at different

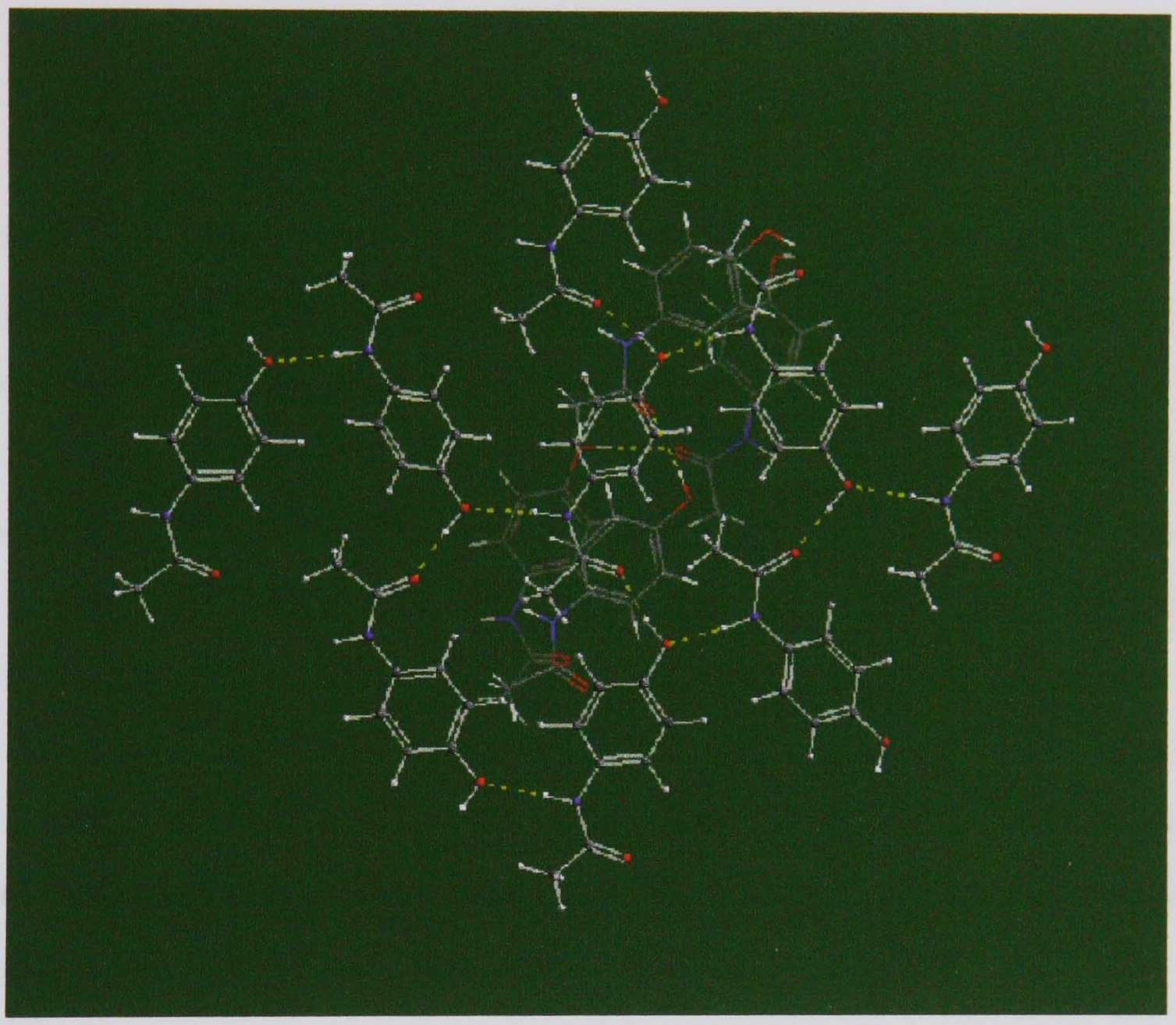


Figure 4.7: Orthorhombic form of paracetamol

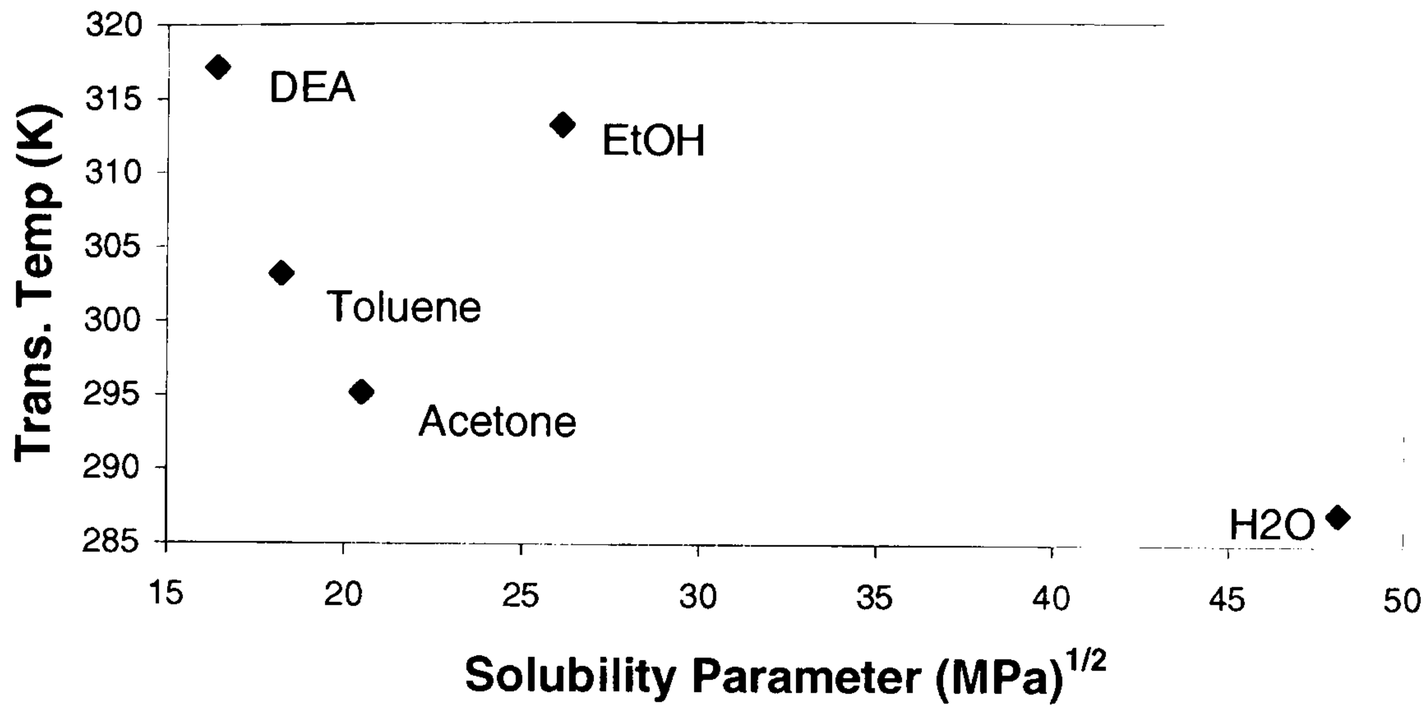


Figure 4.8: Transition temperature vs solvent solubility parameter

temperatures can be determined from Equation 2.9. The predictions show that, at lower temperatures, the monoclinic form is more soluble and hence more stable in all solvents. At a transition temperature that varies for different solvents as shown in Figure 4.8, the orthorhombic form becomes more soluble. Figs. 4.9 - 4.13 show the solubilities of the different polymorph forms of paracetamol in different solvents for a range of temperatures where the temperature is in Kelvin and m/f denotes the mole fraction. The polymorphic form which has the higher Gibbs free energy is the metastable form. Sacchetti (2000) found the free energy of Form II to be higher than that of Form I until -30^{o}C. The results are therefore qualitatively correct. Sohn (1990) however found that paracetamol polymorphs do not readily manifest their polymorphic behaviour when recrystallised in a variety of solvents and so, a quantitative conclusion cannot be made.

4.4 Conclusions

The methodology has been shown to be quite conclusive in its ability to predict satisfactory infinite dilution activity coefficients as compared with experimental data and other available prediction methods such as D-UNIFAC and COSMO-RS. This method has been tested extensively using mono- and bifunctional benzene derivatives since pharmaceutical solutes usually have aromatic rings. The method works better for similar systems as the slight overestimation of the interaction energies cancels out but it allows for a reliable solvent ranking for the solute. The methodology can also be used to compare the sta-

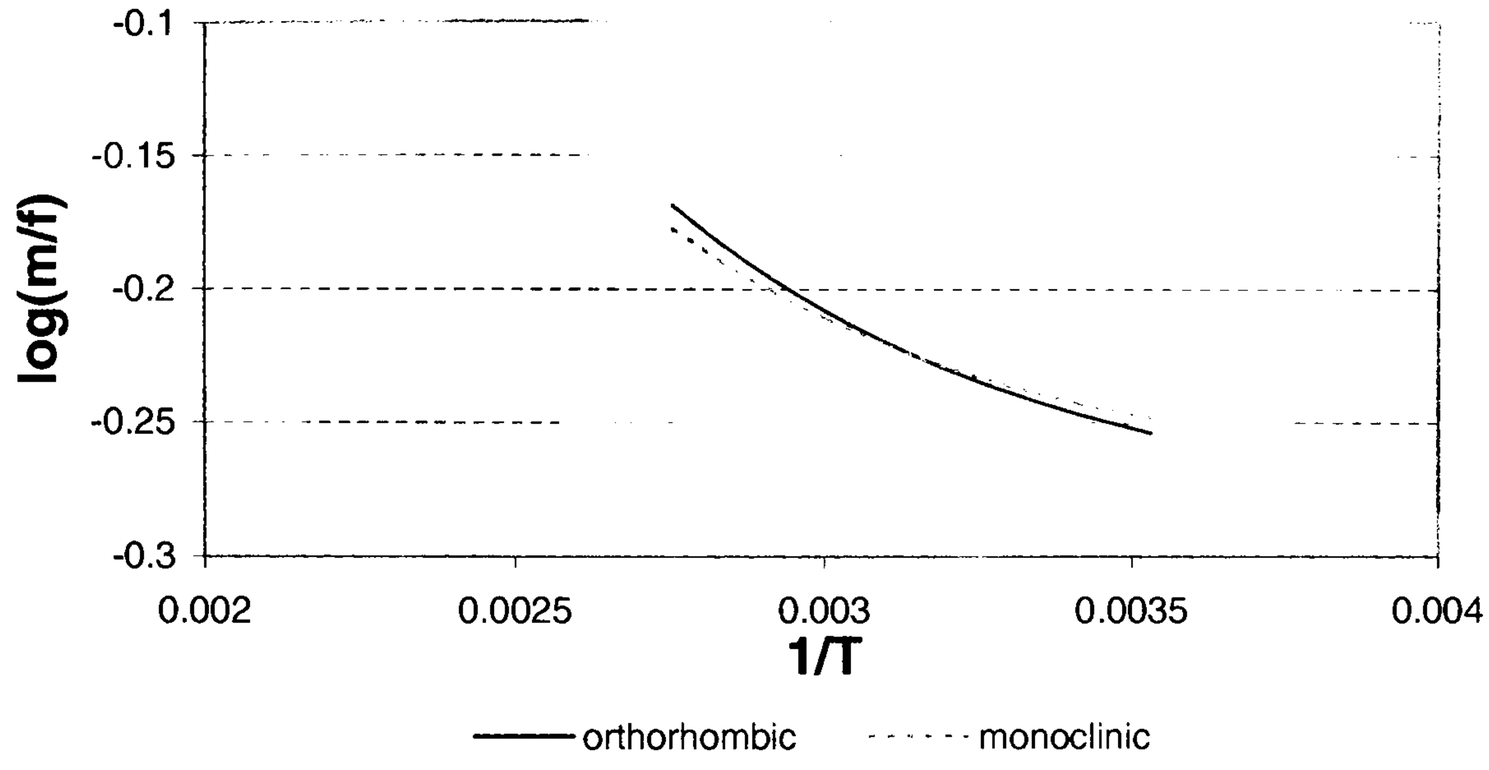


Figure 4.9: Solubility of the paracetamol polymorphs in diethylamine

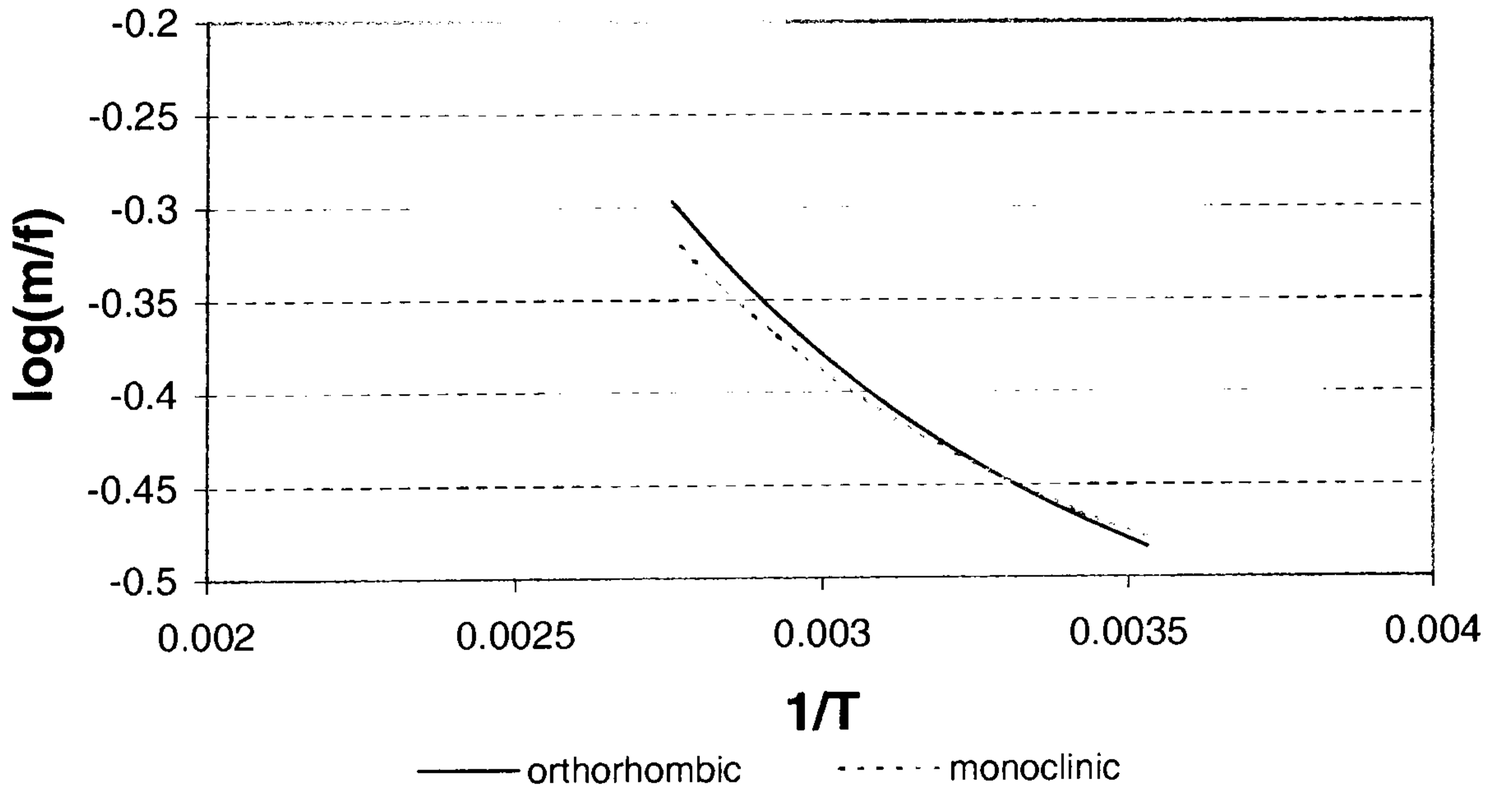


Figure 4.10: Solubility of the paracetamol polymorphs in toluene

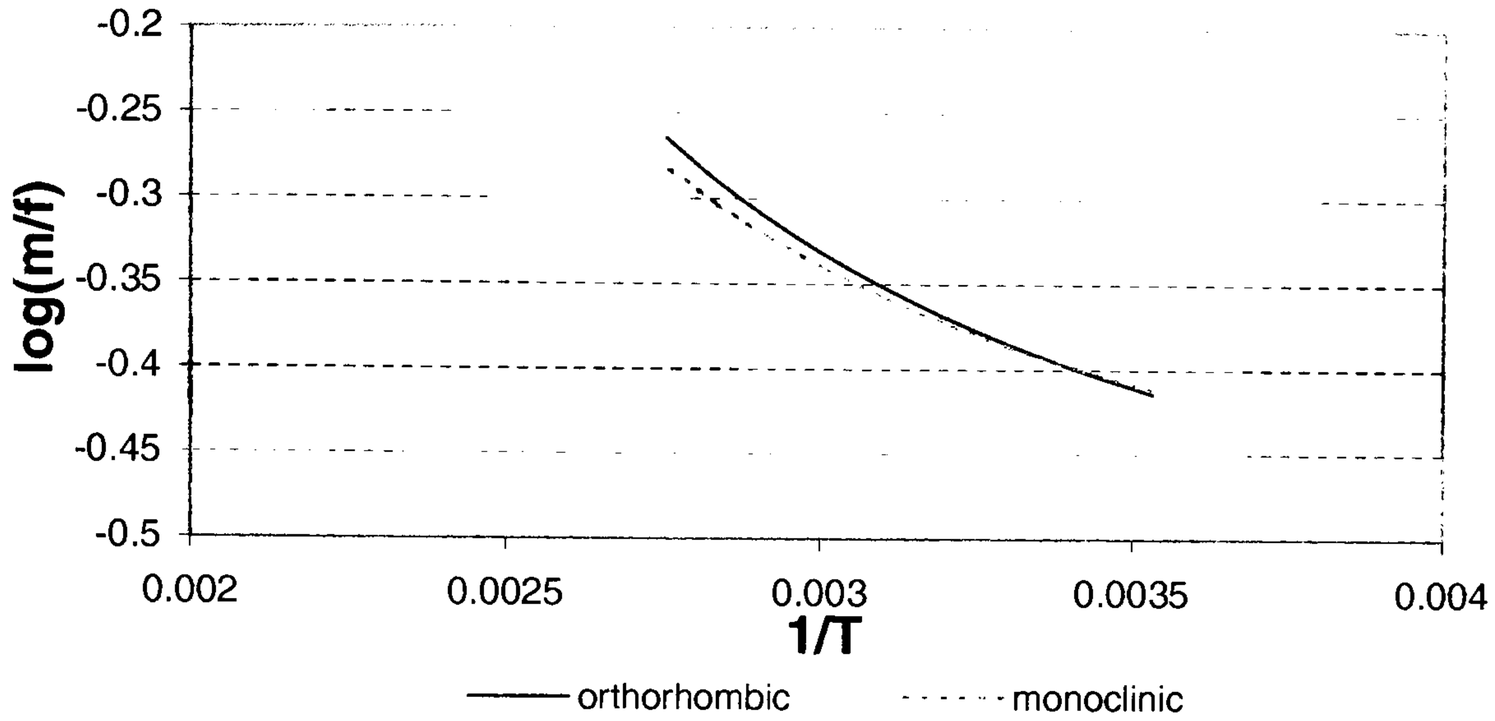


Figure 4.11: Solubility of the paracetamol polymorphs in acetone

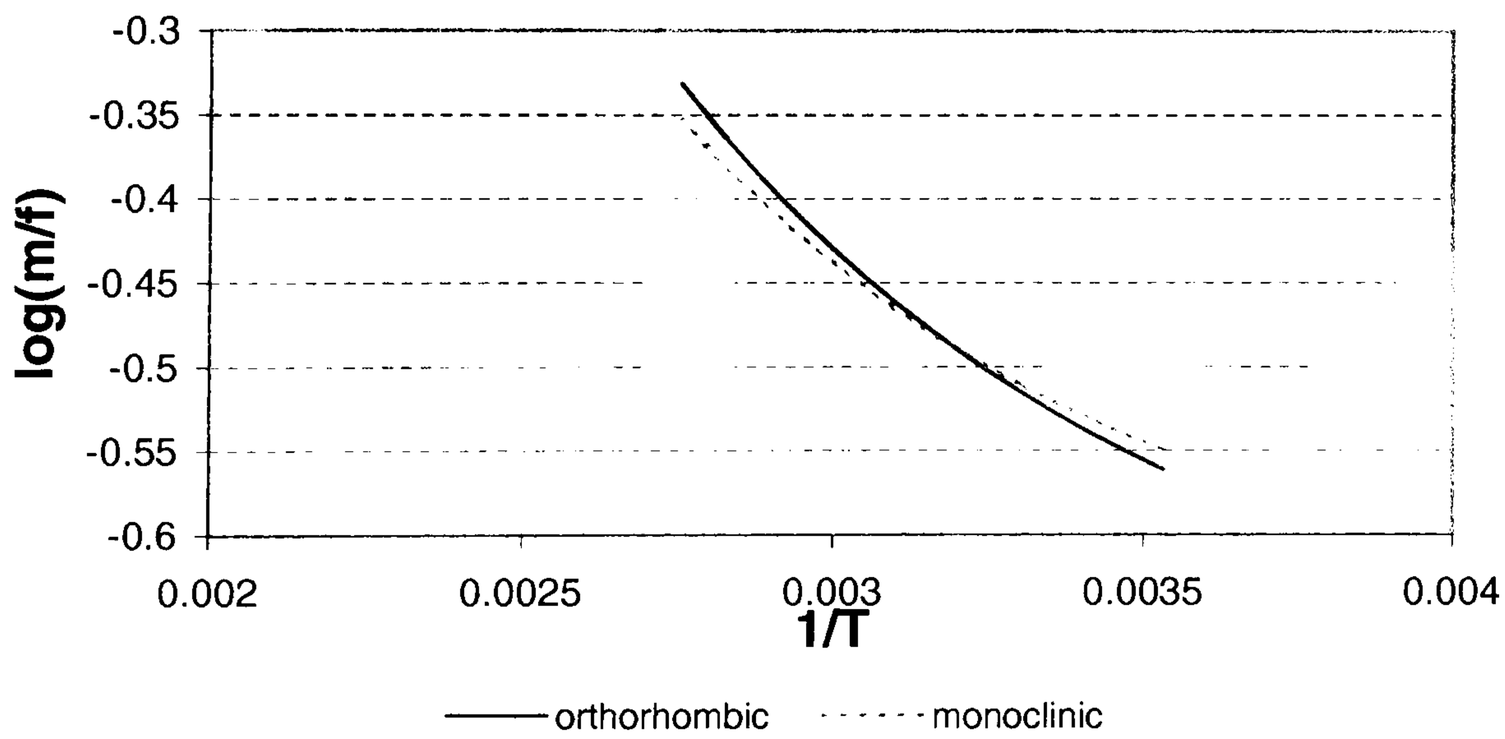


Figure 4.12: Solubility of the paracetamol polymorphs in ethanol

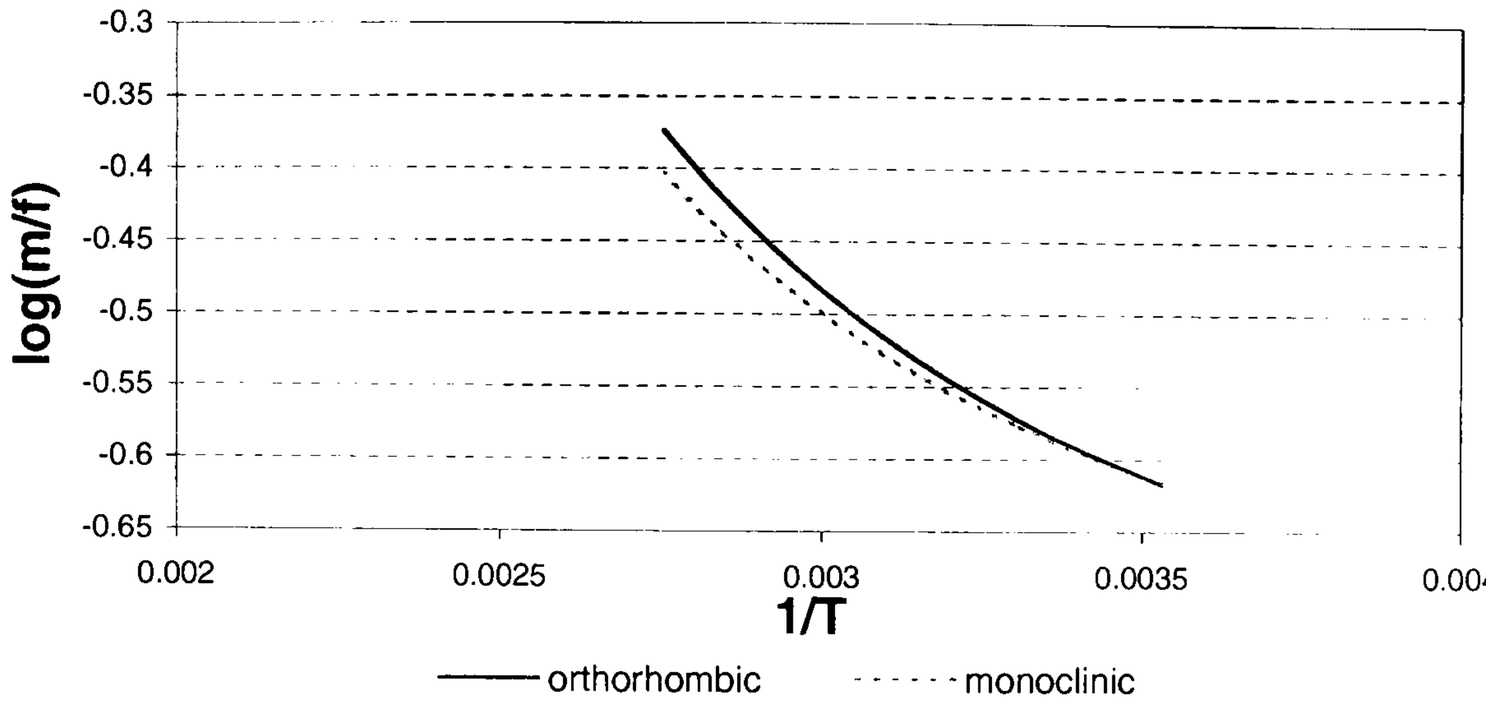


Figure 4.13: Solubility of the paracetamol polymorphs in water

bility of polymorphic forms for a range of temperature although more research has to be performed for more quantitatively accurate results.

The difficulty is in building a "good" model of the solvation shell especially for small, polar solvents. Therefore, an automated method would simplify the technique and allow the methodology to be used easily. The automated method has to identify specific interactions and then build the solvation shell model. The method can be summarised as follows:

1. Determine the number of solvent molecules from the solvent accessible surface area of the solute molecule

2. Identify solute-solvent specific iteractions

3. Place the solvent molecules systematically around the solute, emphasising the specific interactions

4. Perform local minimisation

5. Determine the binary interaction energies between all solute-solvent pairs

6. Calculate the Boltzmann average of the binary interaction energies using Equation 3.3.

Specific interactions for the solute-solvent, solvent-solvent and solute-solute pairs can ideally be found using a global optimisation algorithm which would list the minima of the pairs in increasing order. Such an algorithm has been devised. Details of the algorithm are explained in Chapter 6. After the specific interactions are identified, the first few solvent molecules are placed at these positions. The subsequent solvent molecules are placed systematically around the solute as in the previous chapter to form the first solvation shell.

Chapter 5

Literature Review on Conformational Search of Molecular Systems

5.1 Introduction

This chapter highlights the different techniques available to determine the configuration space for molecular systems. While it is not difficult to find a local minimum potential energy for a pair of molecules, it is a non-trivial problem to determine the configuration that will give the global minimum potential energy. As explained in Chapter 2, the energy of a molecule is determined by the Hamiltonian of the system which is determined by the set of atomic positions. The differences in energy of the various conformations are mostly due to nonbonded distances which act together to determine the spatial structure. For a molecule of N atoms, the number of local minima increases as $\exp(N^2)$ (Hoare, 1979). This exponential growth makes the location of the global minimum on the potential energy surface a non-trivial problem for large molecules.

The global minimum structure of a molecule (i.e. protein folding) or the global minimum energy configuration of microclusters provide important information on the physical, chemical and biological properties of these systems. For example, if the protein

folding problem could be solved, it would greatly simplify the tasks of drug design and the understanding of the mechanisms of biological function. Microclusters have important consequences on the areas of catalysis and crystallisation to astrophysics.

Two important criteria for the development of a successful conformational search algorithm are a fast, global optimisation search method and accurate potential energy functions to describe the molecular system. The search method used should be reliable and able to differentiate between different conformers i.e. new structures should be recognised. Most conformational search methods deal with the Molecular Mechanics minimisation of an initial geometry which is then compared with previously found conformers. The new conformation is usually differentiated from other structures by comparing the energies and dihedral angles. Existing global energy minimisation methods can be classified into stochastic and deterministic, with most methods falling in the former category. These search techniques usually make use of local energy minimisation techniques which are defined in Section 5.2 to help explore the conformational space. The performance of these methods can be assessed based on their ability to identify all minima, and their computational expense.

5.2 Potential Energy Minimisation

The potential energy of the molecular system is the sum of the various energy terms for a given conformation of the system using the defined set of potential energy functions and parameters for a particular force field. The structure of the system is optimised by 'energy minimisation' that finds a stable conformation by reducing the gradients to zero to find the minimum energy of the system. The potential energy surface of molecules is characterised by local minima and transition states. Minimisation algorithms are used to find the nearest local minimum from any starting configuration. For minimisation algorithms using derivative methods, a multidimensional function $f(\mathbf{x})$ may be expanded as a Taylor series round its minimum $\mathbf{x_k}$:

$$f(\mathbf{x}) = f(\mathbf{x_k}) + (\mathbf{x} - \mathbf{x_k})\nabla f(\mathbf{x_k}) + \frac{1}{2}(\mathbf{x} - \mathbf{x_k})^T \nabla^2 f(\mathbf{x_k}).(\mathbf{x} - \mathbf{x_k}) + \ldots$$

where $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ are the first and second derivatives of the function respectively.

The highest derivative used in the minimisation algorithm defines the *order* of the algorithm. If only the value of the function is used, the algorithm is of *order* 0. Algorithms that use the first derivatives which give slope information are of *order* 1 and those that use the second derivatives that define the curvature of the potential energy surface are of *order* 2.

An example of an *order* 0 minimisation algorithm is a grid search method. Here, a regular grid is placed on the potential energy surface and the value of the potential energy is measured at each node. The minimum is the grid point that has the lowest energy. The computational cost depends on the density of the grid as does the accuracy. There is therefore a trade-off between the two, however the method works well only for simple molecules.

Steepest Descent (Luenberger, 1984) and Conjugate Gradients techniques (Hestenes, 1980) are both *order* 1 minimisation algorithms as these methods use the first derivatives to direct the search towards the local minimum.

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k h(\mathbf{x}_k) \tag{5.1}$$

where λ is the step size, $h(\mathbf{x}_k)$ is the direction of the step and k is the iteration number. For Steepest Descent:

$$h(\mathbf{x}_k) = -g(\mathbf{x}_k)$$

where $g(\mathbf{x}_k)$ is the gradient at point $\mathbf{x}_k$. However, as there is no information on the curvature of the slope, a variable step size is used which is calculated using:

$$\frac{d}{d\lambda_k} f(\mathbf{x}_{k+1}) = \nabla f(\mathbf{x}_{k+1})^T . \frac{d}{d\lambda_k} \mathbf{x}_{k+1} = -\nabla f(\mathbf{x}_{k+1})^T . g(\mathbf{x}_k)$$

where $\nabla f(\mathbf{x}_{k+1})$ and $g(\mathbf{x}_k)$ are orthogonal. This method has slow convergence due to its zigzag descent method as the new gradient at the minimum point of any line minimisation is then perpendicular to the direction just taken. As a result, the technique is very inefficient for surfaces with long, narrow valleys as minima.

Conjugated Gradients techniques use information on the previous step sizes to determine the step size to use. The subsequent search direction, instead of being orthogonal to just the previous direction, is conjugate to the last gradient and to all the previous direction traversed. Two different vectors are conjugate to each other when they are orthogonal with respect to any symmetric positive definite matrix. The step length λ_k is determined using:

$$\min_{\lambda_k<0} f(\mathbf{x}_k + \lambda_k h(\mathbf{x}_k))$$

The new point $\mathbf{x}_{k+1}$ is calculated from Equation 5.1. The gradient at that point $g(\mathbf{x}_{k+1})$ is then determined to find the new search direction $h(\mathbf{x}_{k+1})$:

$$h(\mathbf{x}_{k+1}) = -g(\mathbf{x}_k) + \beta_k h(\mathbf{x}_k)$$

and

$$\beta_k = \frac{g(\mathbf{x}_{k+1})^T.(g(\mathbf{x}_{k+1}) - g(\mathbf{x}_k))}{g(\mathbf{x}_k)^T.g(\mathbf{x}_k)}$$

Order 2 minimisation techniques require a vector of $3N$ first derivatives and $(3N)^2$ second derivatives for the minimisation of a molecule with N atoms. The Newton-Raphson minimisation (Luenberger, 1984) is expressed as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\nabla f(\mathbf{x}_k)}{\nabla^2 f(\mathbf{x}_k)} \tag{5.2}$$

The Newton-Raphson method is quite accurate, however it is computationally expensive as the Hessian approximation has to be calculated at every iteration.

5.3 Stochastic Search

Stochastic search techniques aim to overcome the limitations of local gradient-based optimisation by allowing jumps over energy barriers. Stochastic searches tend to fall into several categories: random perturbations, genetic algorithms, simulated annealing and smoothing functions that aim to obtain improved 'generations' of minima.

The search method we implemented initially in Section 3.2, the Boltzmann Jump Procedure, falls into the first category. Li and Scheraga (1987) also discussed the use of this category of search method where a stable conformation is randomly perturbed and then minimised. The minimised structure is compared with existing structures and the search is terminated after the number of failed attempts at identifying a new conformation has reached a predefined number. This method however is not a global optimisation method in that it just provides a random sampling of the potential energy landscape and the search could run infinitely. However, it is easy to implement and with careful parameterisation of the perturbation method, could identify low-lying minima though not necessarily the global minimum especially for complex molecules e.g. proteins.

Barronó et al. (1999) used genetic algorithms to find the minimum Lennard-Jones configurations for clusters of N atoms. A set of configurations is used as the first generation. Then the configurational energy was used as a fitness criterion to choose a fraction of the first generation as 'parents'. 'Children' are obtained by crossover and mutation from the parents and the set of lower-energy configurations from the 'parents' and 'children' configuration is used as the next generation. The efficiency of the genetic algorithm depends on the selection of the crossing rules thus successful parameterisation of the algorithm needs an understanding of the particular problem.

Simulated annealing (Kirkpatrick et al., 1983) is based on the fact that slowly cooling a heated metal would bring it to a more uniformly crystalline state. In simulated annealing, the potential energy of the system is evaluated at a given temperature using either molecular dynamics or Monte Carlo simulation. The role of temperature is to allow the configurations to occupy high energy regions and to pass over those energy barriers. As

the temperature falls, lower energy states are more probable according to the Boltzmann distribution and at absolute zero, the system should be at the global minimum. Careful temperature control is important and convergence is slow. As global minimum cannot be guaranteed, successive runs are required to ensure a high probability that the global minimum has been reached.

The main idea behind the smoothing potential is to analytically transform the multi-minima potential energy surface by reducing the number of local minima and the heights of barriers. The reasons for the high maxima are the short-range repulsion between atoms and the absence of long-range attraction. The original potential energy function is transformed by successively applying a smoothing operator which may vary. For example, Pillardy and Piela (1997) used a multiple elliptic-Gaussian distance scaling method, and Pappu et al. (1998) included local searches in the smoothing diffusion equation method to potentially recognise errors due to earlier energy 'crossings'. The extent of smoothing is controlled by a parameter t. As the surface is smoothed, the high-lying minima merge into the regions with low-lying minima as the barriers between minima are lowered. Smoothing is carried out until all starting structures converge to the same structure at the smoothed potential energy surface. At that point, the reversed smoothing step is carried out and local minimisations conducted at regular intervals. The minimum points at each local minimisation step is used as a starting structure for the next iteration until the global minimum is found. Although the potential energy found could not be guaranteed to be the global minimum, proper implementation of these smoothing techniques give good local minima using a fraction of the function evaluations. However at present, smoothing the general objective function is very expensive computationally and these methods are only useful currently when the potential is a sum of univariate functions of distances between atomic coordinates (Neumaier, 1997).

Dill et al. (1997) developed a convex global underestimator (CGU) which underestimates all known minima by the least possible amount. A number of local potential energy conformations is generated, k, where $k \geq 2n + 1$ where n is the degrees of freedom. The locations of the minimum of the CGU and the lowest minimum of the known minima are

then used to limit the search domain. The steps are then repeated and the authors state that the global minimum can be found with reasonable assurance after several repetitions independent of the starting structure.

5.4 Deterministic Global Search

Deterministic global optimisation search methods implicitly cover the entire conformational space by identifying regions which cannot lead to low energy conformations. The absence of the global minimum energy from a given region of the energy surface can be ascertained without explicitly evaluating every configuration in that region. As a result, convergence of these algorithms to the global minimum can be achieved in finite time. Many deterministic techniques are based on the general branch-and-bound framework (Horst and Tuy, 1996). The advantanges of the branch-and-bound search method are that it focuses on the bottom parts of the energy landscape, does not get caught in kinetic traps and its speed is largely dependent on the size of the landscape rather than its shape. Branch-and-bound methods do not only guarantee the identification of the global minimum but also allow all local minima to be found.

The branch-and-bound framework is based on the generation of valid convex underestimators for the nonconvex functions. In the branch-and-bound framework, an upper bound E_U and a lower bound E_L on the global solution E^* are generated at each iteration such that $E^L \leq E^* \leq E^U$. Converging lower and upper bounds, E^{L*} and E^{U*}, bracket the global minimum solution and are succesively refined by successive partitioning of the initial domain. Each upper bound is obtained through a standard local minimisation of problem (P). Each lower bound is obtained by building a convex lower bounding problem denoted (LB). This problem is essentially an approximation of the original problem (P), such that the solution E^L of (LB) is less than the global solution of (P), E^*. Furthermore, the convexity of the lower bounding problem (LB) ensures that it has a unique solution, and that its solution E^L can therefore be found reliably with a standard local minimisation algorithm, without any risk of getting trapped at a local minimum. The construction of the lower bounding problem requires knowledge of the bounds on all

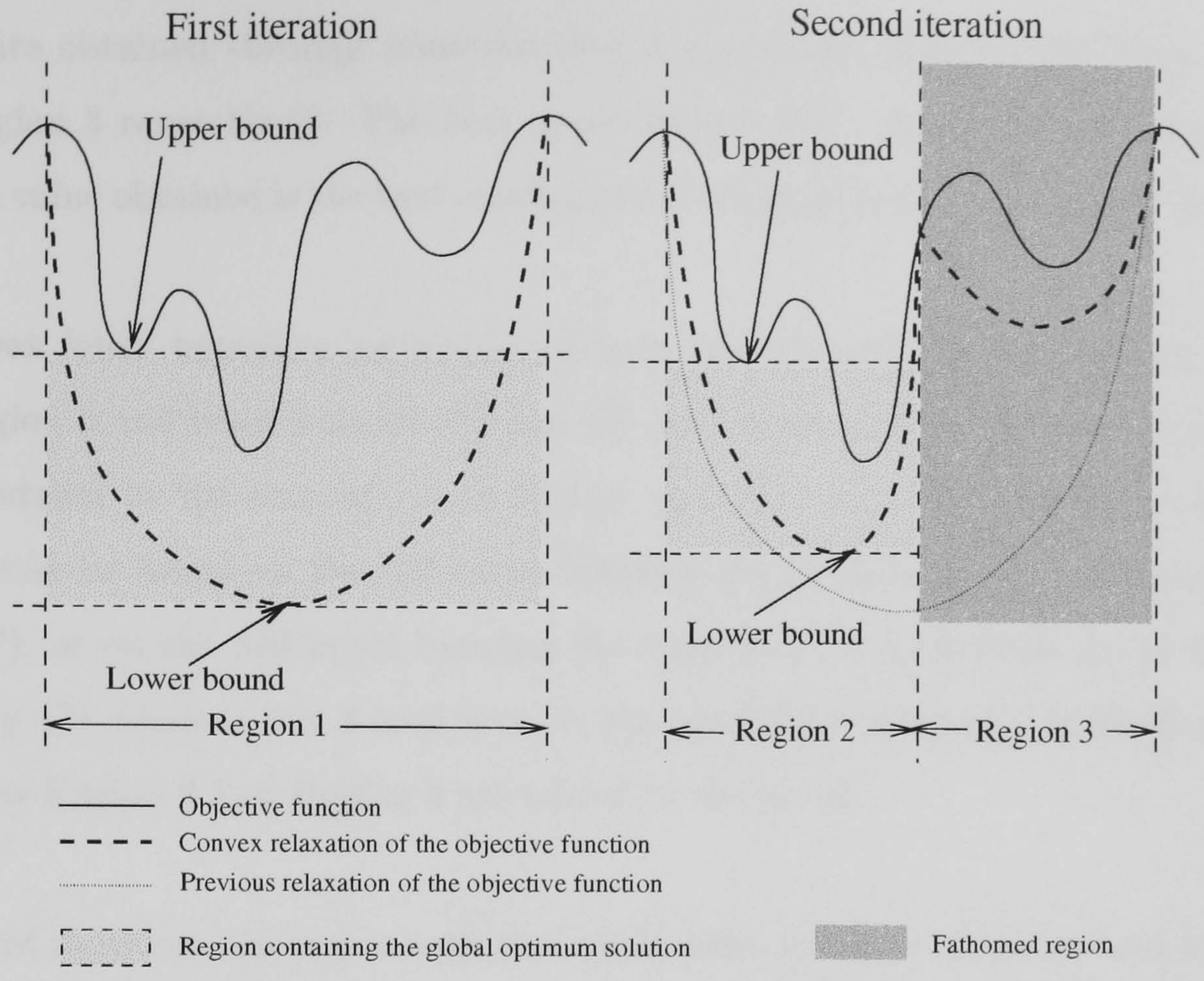


Figure 5.1: Branch and bound (Adjiman, 1998)

the variables of the problem. The convergence of the algorithm is achieved by constructing the lower bounding problem in such a manner that it is improving, that is, it becomes an increasingly better approximation of the original problem. This is made possible by partitioning the solution space and considering more narrow variable ranges.

The bounds on the independent variables define the entire configuration space denoted by Region 1 as shown in Fig. 5.1. At the first iteration, a local solution of (P), E_1^U, is found by selecting a starting point randomly within Region 1. (LB) is constructed based on the variable bounds and solved to give a lower bound E_1^L. In general, convergence is not achieved at the first iteration, and the best upper bound is set to $E^{U*} = E_1^U$. Region 1 and its corresponding lower bound E_1^L are added to a list of unexplored regions R. At the second iteration, Region 1 is removed from R and split into two subregions Region 2 and Region 3. This is done by *branching* on one of the independent variables. For example, if the torsional angle ϕ_1 is chosen for branching, its bounds in Region 2 are $0 \leq \phi_1 \leq \pi$ and in Region 3 are $\pi \leq \phi_1 \leq 2\pi$ from the original bounds of $0 \leq \phi_1 \leq 2\pi$. Upper bounds E_2^U

and E_3^U are obtained through minimisations from random starting positions in Region 2 and Region 3 respectively. The best upper bound, E^{U*}, is updated to min(E^{U*}, E_2^U, E_3^U). The value obtained is the best candidate for the global minimum energy found so far.

Two convex lower bounding problems are constructed based on the bounds in Region 2 and Region 3 and lower bounds E_2^L and E_3^L are obtained for each of these regions and E^{L*} is updated to the current lowest bound, min(E_2^L, E_3^L). A convergence test is applied: it may be based on the difference between the bounds on the solution (e.g. E^{U*} - $E^{L*} \leq \epsilon$?), or on the difference between the bounds on the variables in all the regions ($\nu^U - \nu^L \leq \epsilon$?) where $\nu = b$, θ or ϕ and ϵ is the predefined tolerance. If the test fails, the two regions Region 2 and Region 3 are added to the list R.

At the third iteration, the region with the lowest lower bound is removed from the list and partitioned as described above. This procedure is repeated until the convergence test is passed, indicating the global minimum energy $E^* = E^{U*}$ has been found. Any region R_i with a lower bound E_i^L greater than the global solution is guaranteed to remain in the list R until the global minimum is found and is thus implicitly eliminated from consideration.

The main difficulty in designing such an algorithm lies with the generation of a valid convex lower bounding problem as shown in Fig. 5.2. In the case of a constrained optimisation problem such as (P), a general convex underestimator for the objective function must be derived. This underestimator must be convex with respect to *all* the variables involved. Further, a convex relaxation of the feasible region must also be generated. This corresponds to a region that is convex and larger than the region defined by the constraints.

Several convexification strategies have been developed to build the lower bounding function: interval arithmetic (Hansen, 1980; Ratschek and Rokne, 1991), symbolic reformulation (Smith and Pantelides, 1999) and the αBB algorithm for general nonconvex twice-differentiable functions (Adjiman at al., 1998a, Adjiman at al., 1998b).

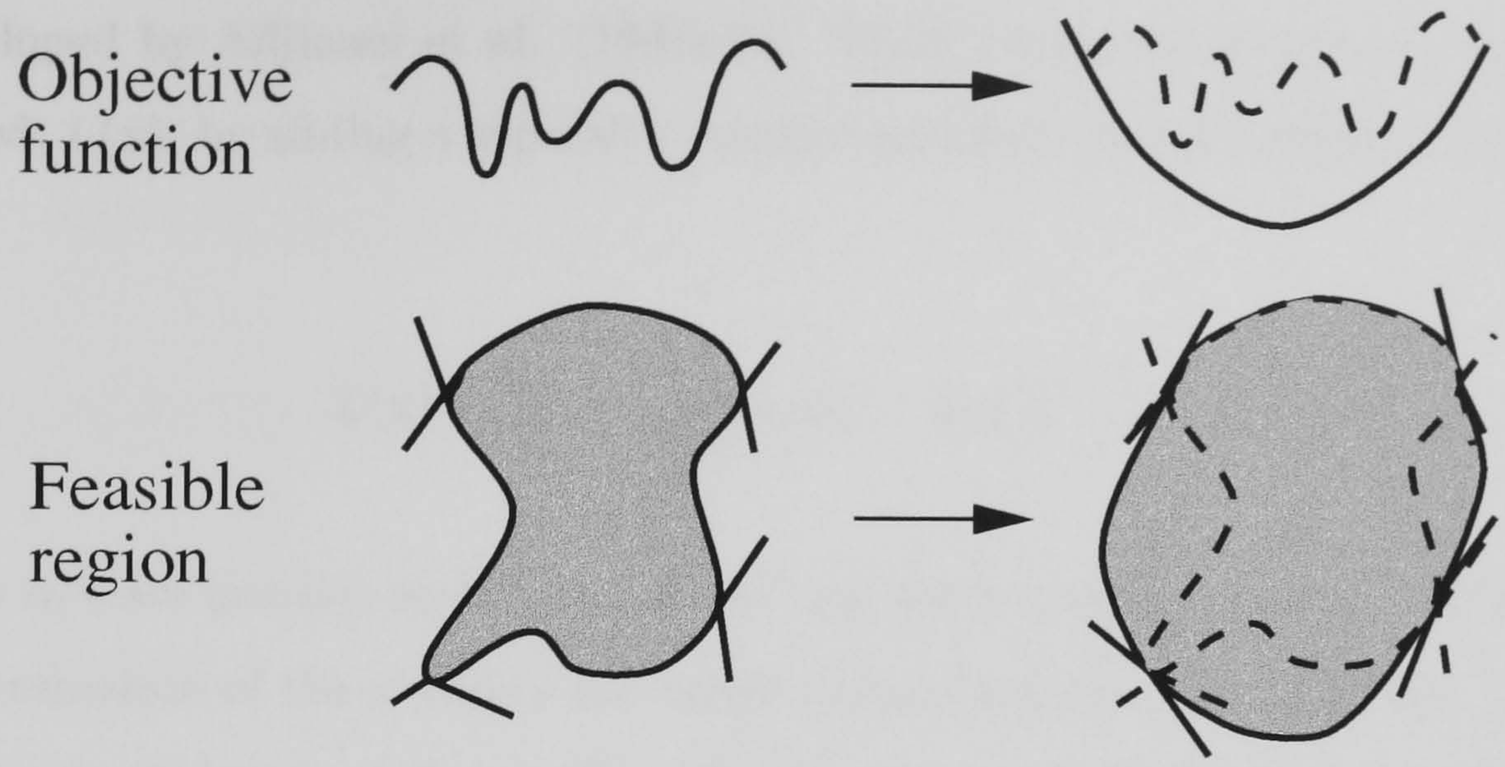


Figure 5.2: Convexification of the objective function and the feasible region (Adjiman, 1998)

The basis of interval arithmetic is deleting the portions of the optimisation domain that cannot contain the global solution. Vaidyanathan and El-Halwagi (1994) developed an algorithm that integrates the calculation of the lower and upper bounds with interval techniques to minimise the solution space. The algorithm also includes a 'distrust-region' method to eliminate infeasible regions without directly checking the constraints.

A symbolic reformulation/spatial branch-and-bound algorithm was developed by Smith and Pantelides (1999) to build tight underestimators for nonconvex problems. The algorithm is based on the fact that algebraic expressions are made out of binary operators corresponding to the five arithmetic operations and unary operators e.g. logarithms, exponentials etc. Tight convex underestimators have already been developed for bilinear and linear fractional functions (McCormick, 1976; Quesada and Grossmann, 1995), bilinear and trilinear terms (Maranas and Floudas, 1995), simple exponentiation and convex/concave univariate functions (Smith, 1996) and logarithmic mean temperature difference terms (Zamora and Grossmann, 1997). The nonconvex terms are reformulated using those convex relaxations by the introduction of new variables and constraints so the problem becomes a linear one.

The αBB algorithm which is applicable for general twice-differential functions was fur-

ther developed by Adjiman et al. (1998a,b). These nonconvex functions,$f(\mathbf{x})$, may be convexified, $L(\mathbf{x})$, by adding a separable positive quadratic terms (Maranas and Floudas. 1994):

$$L(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i (x_i^L - x_i)(x_i^U - x_i) \tag{5.3}$$

where the α_i's are positive scalars and x_i^L, x_i^U are lower and upper bounds of variable x_i. The determination of the α-values are based on the Hessian of the problem. The global minimum potential energy of a single molecule was found using this approach (Maranas and Floudas, 1994). The algorithm assumes that the covalent bond lengths and angles are at their equilibrium values. This assumption reduces the independent set of internal coordinates, $\mathbf{x}$, to a set of independent torsion angles. The values of the variables are bounded by upper and lower bounds. After each iteration, the feasible regions for each torsion are branched to give tighter underestimators and therefore an increasing lower bounding function. The algorithm guarantees that the non-linear problem is solved to global optimality with finite ϵ-convergence.

Klepeis et al. (1998) further refined the αBB optimisation algorithm to solve the protein folding and peptide docking problems. As the covalent bond lengths and bond angles are fixed at their equilibrium values, the conformation is only a function of the dihedral angles. A description of the energy of a polypeptide must include its interactions with the solvent and solvation contributions are included implicitly using empirical correlations with either the surface area or volume. The objective function of the problem is the sum of the potential energy function and the free energy of solvation. The algorithm has been used successfully to identify the global minimum conformation of met-enkeplalin (H-Tyr-Gly-Gly-Phe-Met-OH), a pentapeptide. The unconstrained optimisation problem was reformulated as a constrained problem by including nonlinear constraints based on distance restraints determined from NMR studies (Klepeis et al, 1999).

Maranas and Floudas (1992) have also proposed a global optimisation approach for finding the global minimum energy configuration of Lennard-Jones microclusters. The nonconvex

objective function is first transformed to the difference of two convex functions. A primal-relaxed dual global optimisation approach was then applied to find the global optimum in a finite number of iterations.

5.5 Conclusions

Based on a summary of these methods, a robust optimisation algorithm for flexible molecules has as yet not been developed. Our aim is to further build more solid models to further develop optimisation procedures for the global minimum optimisation of flexible molecular systems.

Chapter 6

Deterministic Global Optimisation of the Potential Energy of Molecular Systems

6.1 Introduction

In this chapter, the problem formulation of the global optimisation of the potential energy of molecules is explained. The objective of the algorithm is to identify the configurations of the global minimum and other low-lying energy minima for the potential energy. At these configurations, the interaction energies between the molecules in the system are usually stronger. Any specific interactions that may exist can then be identified. The entire conformational space is covered implicitly using a branch-and-bound framework (Horst and Tuy, 1996). A molecular system with N atoms can be described using $3N$-6 independent variables which can be either internal or Cartesian coordinates. We have chosen to work using internal coordinates for two main reasons. Firstly, the force fields that allow us to compute the energy as a function of system geometry are usually expressed in terms of internal coordinates. Secondly, the set of physically realistic values for some of the internal coordinates can be remarkably narrow. Bond lengths, for example, do not usually deviate by more than 0.03 Angstroms around equilibrium values (Jensen, 1999) and bond angles are similarly constrained. This in turn restricts the configuration

space in a way that cannot be achieved with Cartesian coordinates.

After defining the problem to be minimised in Sections 6.2 and 6.3, Section 6.4 explains the convexification strategies used to convexify the terms used in the objective function. Section 6.5 continues with convexification of the constraints used in the problem formulation. The general optimisation framework is discussed in Section 6.6 and case studies are presented in Section 6.7.

6.2 Problem Definition

The system to be minimised is expressed using a Z-matrix defining the *independent* bond lengths, bond angles and torsional angles of the system (Leach, 1996). The values of the remaining variables are *dependent* on the values of these *independent* variables and may be defined as functions of these variables.

A Z-matrix is used to define the connectivity between the atoms in a molecule. The parameters are the bond lengths, bond angles and torsional angles. Some of the torsional angles specified consist of atom chains that are not bonded together and these are known as *improper torsional angles*. For example in *n*-butane as shown in Figure 6.1, the Z-matrix is expressed as:

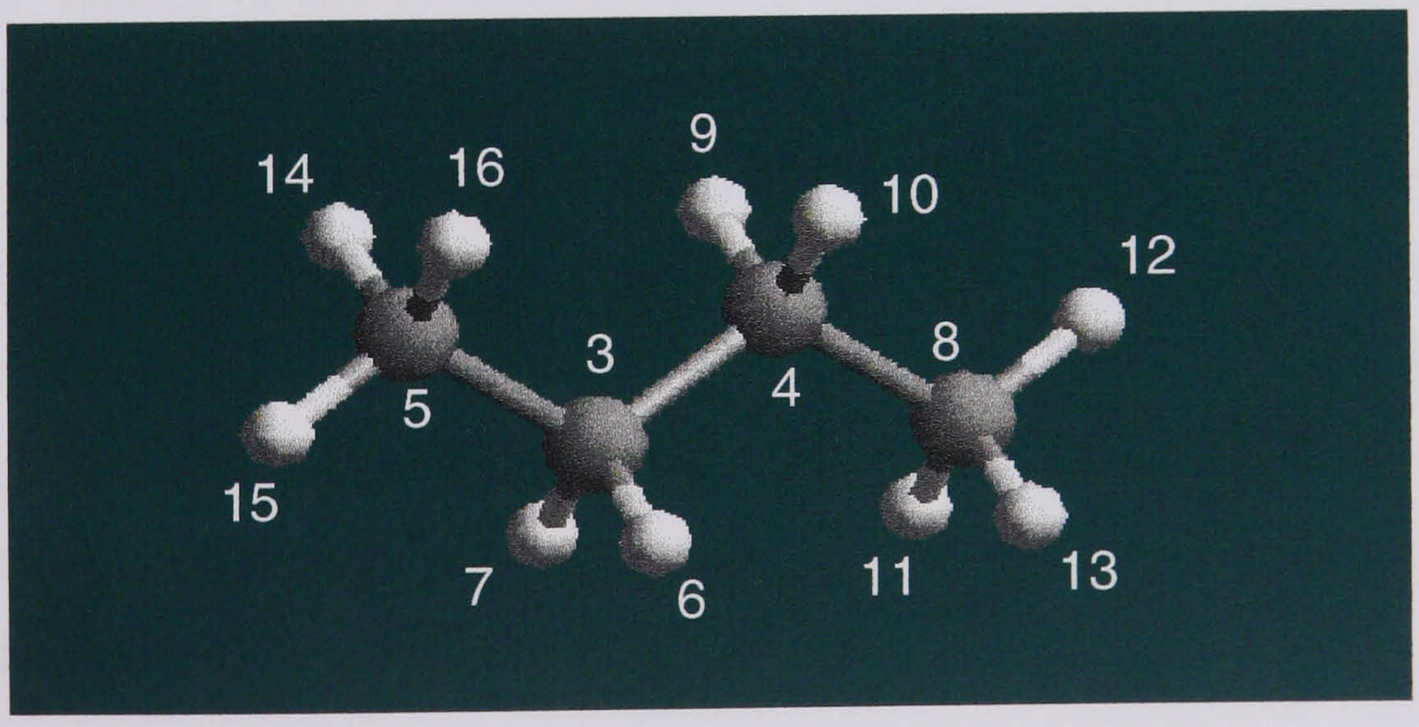


Figure 6.1: *n*-butane molecule showing the atom numbers

Atom Number	Atom Type	Bond Atom	Bond Length	Angle Atom	Bond Angle	Torsion Atom	Torsional Angle
1	X	0	0.000000	0	0.000000	0	0.000000
2	X	1	1.000000	0	0.000000	0	0.000000
3	C	2	1.000000	1	90.000000	0	0.000000
4	C	3	1.535393	2	90.000000	1	228.222667
5	C	3	1.532589	2	90.000000	4	247.520132
6	H	3	1.090504	4	109.955401	5	121.888845
7	H	3	1.090983	4	109.587650	6	116.437047
8	C	4	1.532835	3	112.541622	5	181.307267
9	H	4	1.090616	3	109.960386	8	122.132987
10	H	4	1.091019	3	109.544823	8	238.375932
11	H	8	1.090240	4	111.560538	3	59.958144
12	H	8	1.090526	4	110.839262	11	119.898634
13	H	8	1.090514	4	111.199223	11	239.550819
14	H	5	1.090269	3	111.339350	4	60.304044
15	H	5	1.090347	3	110.904731	14	119.844706
16	H	5	1.090470	3	111.240424	14	239.648020

where atoms 1 and 2 are dummy atoms denoted by X. Dummy atoms are sometimes used in the construction of the Z-matrix to define the geometry.

The atom number states the number of the atom in the molecule. The atom type determines the valence of the atom which is the number of covalent bonds that an atom participates in e.g. a carbon atom denoted by C has a valence of 4 and a nitrogen atom has 3. The covalent bonds have different bond orders depending on the number of electron pairs involved in the covalent bond. The bond length gives the length of the bond between the bond atom and the current atom. The bond angle defines the angle between the atoms given by the angle atom, bond atom and the current atom. The torsional angle is the torsion angle formed by the atoms given by the torsion atom, angle atom, bond atom and the current atom.

There are 13 bonds in the n-butane molecule and lengths of all the bonds are given in the Z-matrix so there are 13 *independent* bond lengths. There are also 24 angles in total and 27 torsions in the molecule. Without including the bond angles and torsional

angles in the Z-matrix involving the dummy atoms, the values of 11 of the angles, 9 improper torsional angles and 3 torsional angles are defined in the Z-matrix. These are defined as *independent* variables. The values of the *dependent* variables can be found subject to the values of these *independent* variables. As for nonbonded distances, there are 54. The values of the *dependent* variables may be found from the *independent* variables and these are specified as constraints in the optimisation algorithm.

In general, the minimisation of the potential energy function can be expressed as:

$$\min E_{system} = E_b(b) + E_\theta(\theta) + E_\phi(\phi) + E_r(r) \tag{6.1}$$

subject to

$$\begin{aligned}
\theta &= f_\theta(\theta^{ind}, \phi) \\
\phi &= f_\phi(\theta, \phi) \\
r &= f_r(b, \theta, \phi) \\
b^L &\leq b \leq b^U \\
\theta^L &\leq \theta^{ind} \leq \theta^U \\
\phi^L &\leq \phi^{ind} \leq \phi^U
\end{aligned}$$

where b, θ^{ind} and ϕ^{ind} are the *independent* internal coordinates and bounds on these variables are predefined beforehand. θ and ϕ are the sets of all the angles and torsional angles, both *independent* and *dependent*, and r is a vector of nonbonded distances. The superscripts U and L denote upper and lower bounds on the variables respectively.

The potential energy objective function is broken down into contributions from bond stretching (E_b), angle bending (E_θ), torsional angle (E_ϕ) and nonbonded interactions (E_r). A cross-term is sometimes included in some force fields but it is not considered in this work. In most studies, the minimisation problem above is expressed as an unconstrained problem but this implies that the nonbonded term is then expressed as a function of the independent internal coordinates rather than the nonbonded distances r. Since this function of the internal coordinates is complicated, we have decided to include the dependence of r on the internal coordinates explicitly as a set of constraints.

The OPLS force field (Jorgensen et al., 1996) was used as it provides a good representation of the energetics of gas phase and solvated systems (Duffy and Jorgensen, 2000) and was developed to reproduce experimental data on fluids. One important property of the OPLS force field is that it is separable and therefore we can use the additivity property of convex functions – the sum of convex functions is itself convex – in the design of a global optimisation algorithm. The force field may be separated into the following terms:

$$E_b = \sum_{i=1}^{n_{bonds}} K_{b_i} \left(b_i - b_{eq}\right)^2 \tag{6.2}$$

$$E_\theta = \sum_{i=1}^{n_{angles}} K_{\theta_i} \left(\theta_i - \theta_{eq}\right)^2 \tag{6.3}$$

$$E_\phi = \sum_{i=1}^{n_{torsions}} \sum_{n=1}^{3} \frac{V_{n_i}}{2} \left[1 + \cos\left(n\phi_i + \gamma_{n_i}\right)\right] \tag{6.4}$$

$$E_r = \sum_{i}^{n_{nonbonded}} \sum_{j}^{j>i} \left[\frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] \right] \tag{6.5}$$

The equations for bond lengths and bond angles reflect the fact that the variables do not vary much from equilibrium, b_{eq} and θ_{eq}, and K_b and K_θ define the width of the potential well. The torsional angle term presupposes torsional 'barriers' which may exclude certain values of torsional angles where V_n is the barrier 'height' and γ_n is the torsional angle shift in the minimum energy value. The nonbonded term is described by a Coulombic term where q defines the atomic charge and e is a function of the permittivity, and a Lennard-Jones term where σ_{ij} defines the collision diameter where energy is zero and ϵ_{ij} is the well depth.

The constraints on the *dependent* variables may be calculated in two ways. First, by

converting the internal coordinates into Cartesian coordinates and then directly calculating the bonds, angles and torsions from the Cartesian values. The nonbonded distance, r_{ij}, between atoms i and j can be calculated from:

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{6.6}$$

The angles are determined using the dot product equation for vectors:

$$\cos(\theta_{ijk}) = \frac{x_{ij}.x_{kj} + y_{ij}.y_{kj} + z_{ij}.z_{kj}}{\sqrt{b_{ij}b_{jk}}} \tag{6.7}$$

where $x_{ij} = x_i - x_j$.

A similar equation is used to calculate the torsional angle, ϕ_{ijkl}, as it is the angle between the planes ijk and jkl. The angle between the planes is equivalent to the angle between the normal vectors of the two planes. If the normal vector for plane ijk is $\mathbf{m}$ and for plane jkl is $\mathbf{n}$, the torsional angle is given by:

$$\cos(\phi_{ijkl}) = \frac{\mathbf{m}.\mathbf{n}}{|\mathbf{m}||\mathbf{n}|} \tag{6.8}$$

However, tight bounds on the *dependent* variables are difficult to obtain using these equations. Thus, rotation sequences relationships between the *independent* variables are used instead as further described in Section 6.3.

The main difficulty in designing a deterministic branch-and-bound algorithm as outlined in Section 6.6 lies with the generation of a valid convex lower bounding problem. In the case of a constrained optimisation problem, a general convex underestimator for the potential energy objective function must be derived. Furthermore, a convex relaxation of the feasible region must also be generated. This corresponds to a region which is convex and larger than the region defined by the constraints.

6.3 Dependent Variables

For any molecular system to be minimised, the system is defined using a Z-matrix. A list of all the bond lengths, bond angles, torsional angles and nonbonded distances of the system is then determined. If any of the bond angle and torsional angle are not explicitly stated in the Z-matrix, these variables are *dependent*. All the nonbonded distances are *dependent* variables. The constraints of the optimisation algorithm state the relationships between the *independent* and *dependent* variables and these constraints determine the values of the *dependent* angles, torsional angles and nonbonded distances.

6.3.1 *Dependent* Angles Constraints

The equation for the spherical triangle states the relationship between the *dependent* angle and the *independent* variables. The constraint $\theta = f_\theta(\theta^{ind}, \phi)$ is stated as:

$$\cos(\theta_{kjl}) = \cos(\theta_{ijk})\cos(\theta_{ijl}) + \sin(\theta_{ijk})\sin(\theta_{ijl})\cos(\phi_{kijl}) \tag{6.9}$$

As an example, one of the angles that is not explicitly stated in the Z-matrix is the angle 12-8-13. The calculation of the angle is independent of the actual bond lengths. Using spherical trigonometry (Todhunter and Leathem, 1901) for Fig. 6.1 with atom 8 as the origin, a spherical triangle can be constructed using the *independent* angles θ_{4-8-11}, θ_{4-8-12} and θ_{4-8-13} as shown in Fig. 6.2. At unit radius, the length of the arc between atoms 12 and 13 gives the value for the angle 12-8-13.

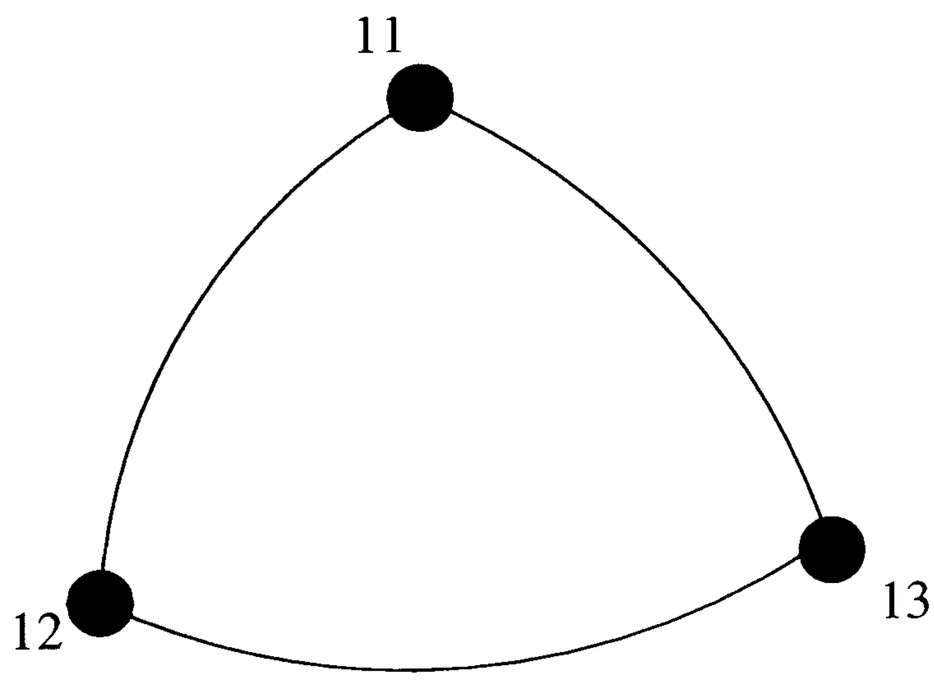


Figure 6.2: Spherical triangle showing atoms 11, 12 and 13

In order to construct a spherical coordinate system with origin based on atom 8. the direction from atom 8 to atom 4 is used to define the z-axis and the end atom of the *independent* torsion $\phi_{3-4-8-11}$ which is atom 11 is used to define the direction of the x-axis. The y-axis is found using the Right-Hand Rule. Fig. 6.3 shows the relationship

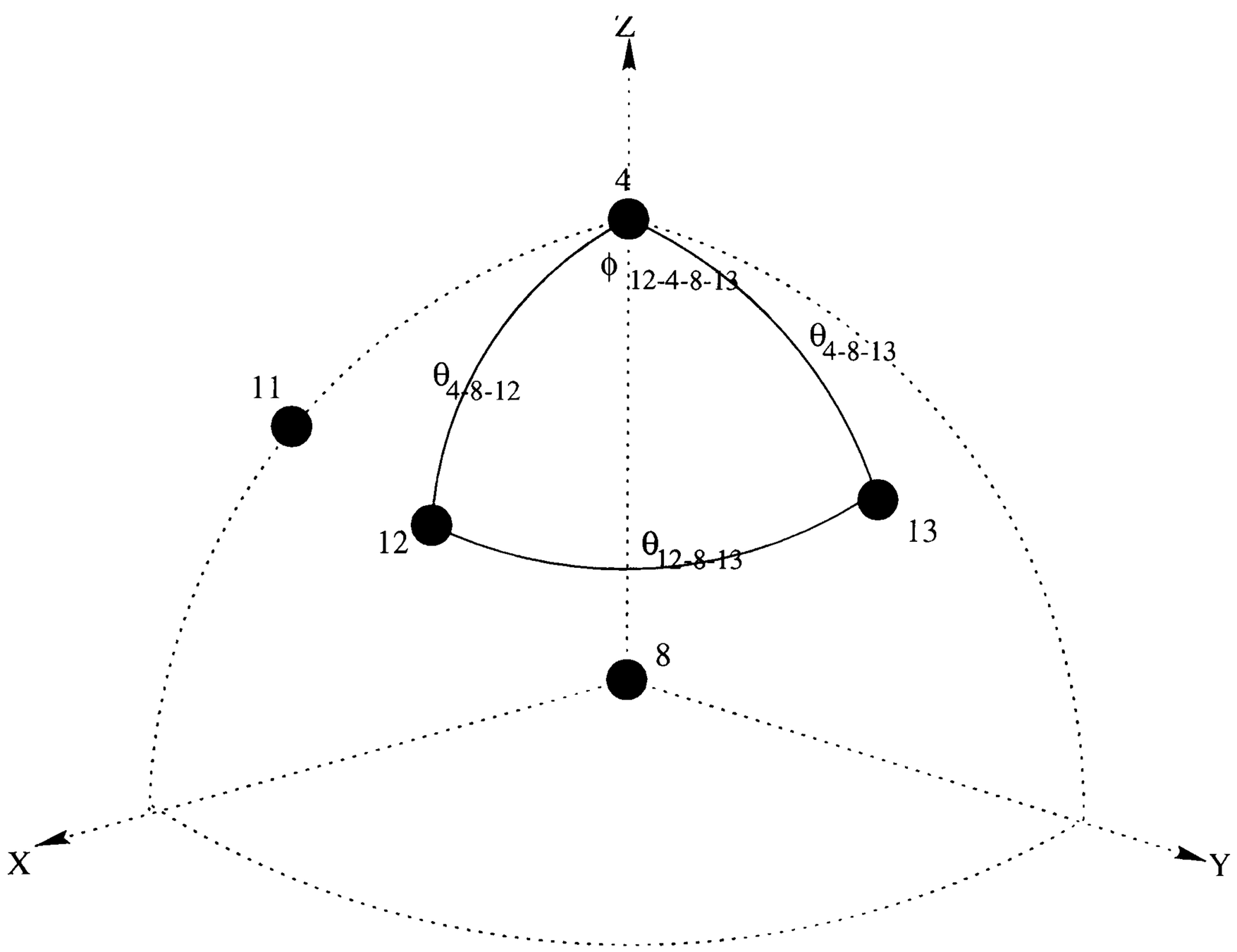


Figure 6.3: Spherical triangle showing relationship between the angles

between the *dependent* angle $\theta_{12-8-13}$ and the *independent* variables, the angles θ_{4-8-12}, θ_{4-8-13} and the improper torsional angle $\phi_{12-4-8-13}$ where the latter is itself a *dependent* variable as it is not stated in the Z-matrix. Equation 6.9 is used to calculate the values of the *dependent* angle:

$$\cos(\theta_{12-8-13}) = \cos(\theta_{4-8-12})\cos(\theta_{4-8-13}) + \sin(\theta_{4-8-12})\sin(\theta_{4-8-13})\cos(\phi_{12-4-8-13}) \tag{6.10}$$

As an example, it is given that angles θ_{4-8-11}, θ_{4-8-12} and θ_{4-8-13} have the values 111.561^o, 110.839^o and 111.199^o respectively. $\phi_{12-4-8-13}$ is a *dependent* variable whose value is 119.652^o. The method used to calculate it is presented in Section 6.3.2. Therefore,

substituting the values in:

$$\begin{aligned}\cos(\theta_{12-8-13}) &= \cos(110.839^o)\cos(111.199^o) + \sin(110.839^o)\sin(111.199^o)\cos(119.652^o)\\ &= -0.302438^o\\ \theta_{12-8-13} &= 107.604^o\end{aligned}$$

6.3.2 *Dependent* Torsions Constraints

The theoretical number of conformers for a molecule can be calculated by:

$$\text{Number of conformers} = (360^o/\text{angle increment})^{\text{number of rotatable bonds}}$$

where the angle increment is the angle by which the end atom of the torsion can be rotated. *n*-Butane has two conformers where the torsional backbone $\phi_{5-3-4-8}$ has the values 65.1^o and 180.0^o respectively. It can also have the value 294.9^o but due to symmetry, it is the same conformer as when the torsion has the value 65.1^o. The values of the *dependent* torsions are only dependent on the values of the *independent* torsion with the same middle atoms and the improper torsional angles between the two pairs of end atoms, $\phi = f_\phi(\theta, \phi)$. The *dependent* torsional angle constraint is:

$$\phi_{kijl} = \phi_{hijm} - \phi_{hijk} + \phi_{mijl} \tag{6.11}$$

The torsional angle $\phi_{12-4-8-13}$ was calculated from Equation 6.11. Given that $\phi_{11-4-8-12}$ in the Z-matrix is 119.899^o and $\phi_{11-4-8-13}$ is 239.551^o, then $\phi_{12-4-8-13}$ would have the value $\phi_{11-4-8-13} - \phi_{11-4-8-12} = 239.551^o - 119.899^o = 119.652^o$.

Another example, the torsion $\phi_{10-4-8-12}$ is *dependent* on the values of the torsion $\phi_{3-4-8-11}$ and the improper torsional angles $\phi_{3-4-8-10}$ and $\phi_{11-4-8-12}$ as shown in Fig. 6.4. The torsion $\phi_{10-4-8-12}$ is then obtained through a straightforward sum:

$$\phi_{10-4-8-12} = \phi_{3-4-8-11} - \phi_{3-4-8-10} + \phi_{11-4-8-12} \tag{6.12}$$

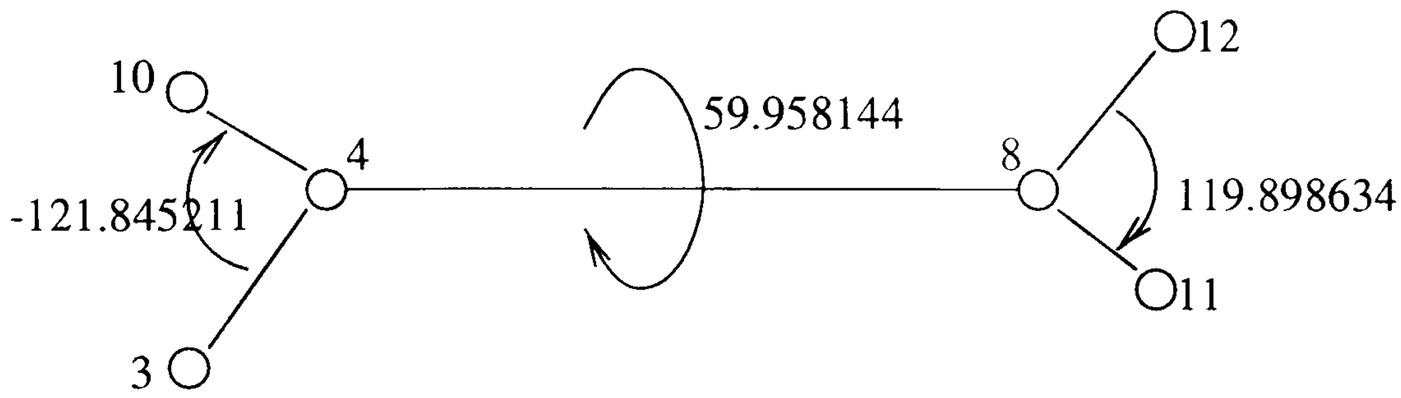


Figure 6.4: *Dependent* Torsional Angle

As it is given that $\phi_{11-4-8-12} = 119.899^o$, $\phi_{3-4-8-11} = 59.958^o$ and $\phi_{3-4-8-10}$ is itself a *dependent* variable. In order to calculate $\phi_{3-4-8-10}$ from Fig. 6.5, the angle θ_{8-4-10} first has to be calculated from θ_{3-4-8}, θ_{3-4-10} and $\phi_{8-3-4-10}$ using Equation 6.9 and was found to be 109.155^o. Also using Equation 6.9, the value of $\phi_{3-4-8-10}$ is then found to be 121.845^o or 238.155^o using the values of the variables θ_{3-4-8}, θ_{3-4-10} and θ_{8-4-10}. By analysing the tetrahedral with atom 4 as the middle atom using the value of $\phi_{8-3-4-10}$ which is 238.376^o, the value of $\phi_{3-4-8-10}$ is 238.155^o. Therefore, $\phi_{10-4-8-12} = 59.958^o - 238.155^o + 119.899^o = -58.012^o$. As the values have to be within the domain $[0^o, 360^o]$, the value is redefined as 301.702^o. As the domain for the torsional angles is $[0^o, 360^o]$, the

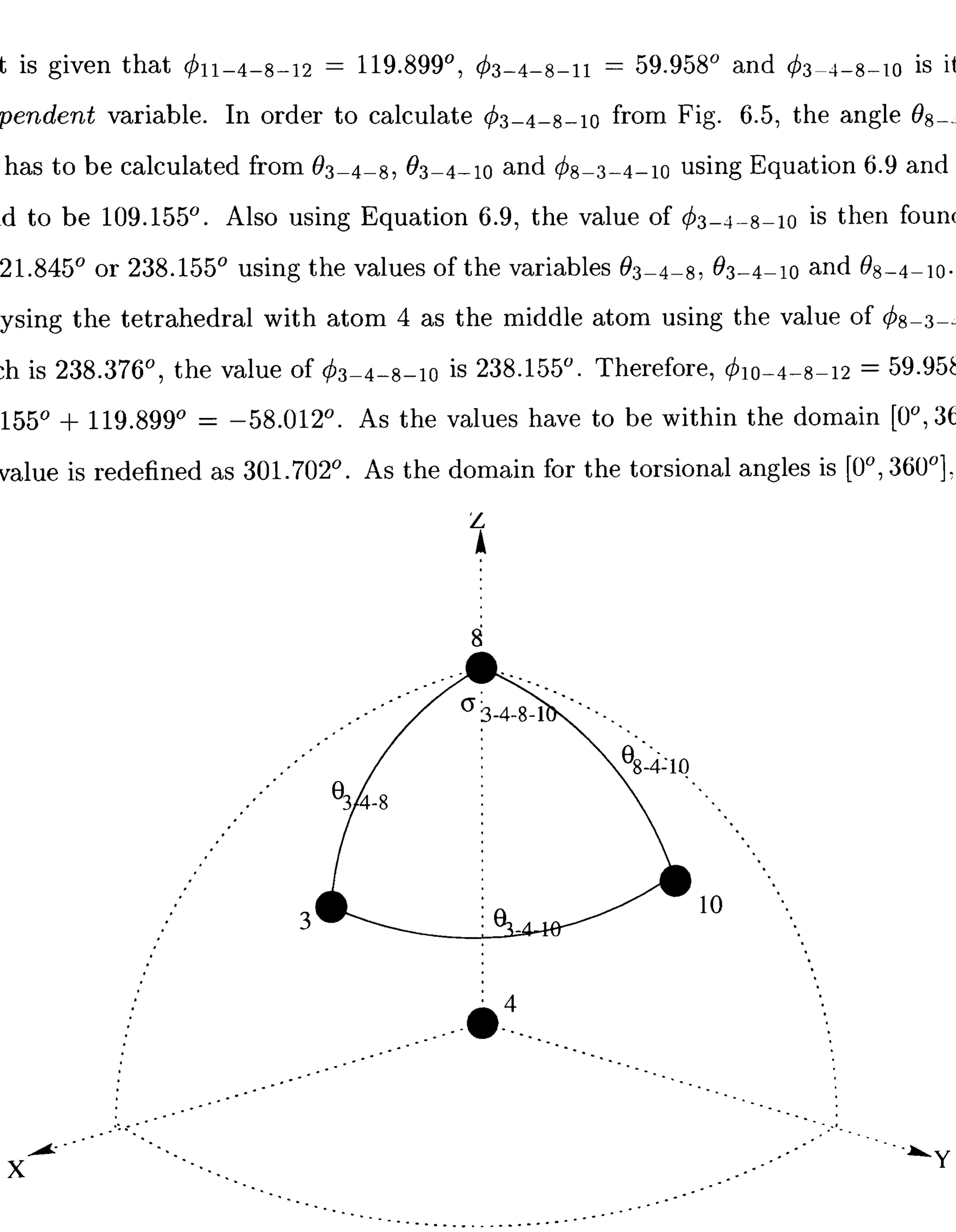


Figure 6.5: Spherical triangle showing atoms 3, 8, 10 with atom 4 at the origin

general rules for the redefinition of the torsional angle are:

1. If $(\phi < 0.0)$, $\phi = \phi + 360^o$

2. If $(\phi > 360^o)$, $\phi = \phi - 360^o$

6.3.3 Nonbonded Distances Constraints

The constraints which relate the nonbonded distance r to the internal coordinates, $r = f_r(b, \theta, \phi)$, were stated by Maranas and Floudas (1994) for the 3-atom chain to the 6-atom chain. For a 3-atom chain, the nonbonded distance r_{13} may be calculated from:

$$f_{r_{13}} = b_{12}^2 + b_{23}^2 - 2\cos(\theta_{123})b_{12}b_{23} \tag{6.13}$$

Similarly, for a 4-atom chain, the nonbonded distance r_{14} is:

$$\begin{aligned} f_{r_{14}} =& b_{12}^2 + b_{23}^2 + b_{34}^2 - 2\cos(\theta_{123})b_{12}b_{23} - 2\cos(\theta_{234})b_{23}b_{34} \\ & - 2(\sin(\theta_{123})\sin(\theta_{234})\cos(\phi_{1234}) - \cos(\theta_{123})\cos(\theta_{234}))b_{12}b_{34} \end{aligned} \tag{6.14}$$

6.4 Convexification of Potential Energy Functions

As discussed in Chapter 5, the nonconvex objective function has to be convexified in order to construct a convex lower bounding function for the branch-and-bound algorithm. This section shows the convexification of the potential energy terms in the OPLS force field. The proposed convex function has to be a rigorously valid underestimator which is once-continuously differentiable so that a gradient-based optimisation algorithm can be used (refer to *Appendix E*). Both the terms for E_b and E_θ are convex and thus only have one potential well. The other two terms E_ϕ and E_r are nonconvex and tight underestimators will need to be constructed for these terms.

6.4.1 Convexification of Torsional Terms

The torsional term is defined as:

$$E_\phi = \sum_{i=1}^{n_{torsions}} \frac{V_{1_i}}{2}\left[1+\cos\left(\phi_i+\gamma_{1_i}\right)\right] \tag{6.15}$$
$$+\frac{V_{2_i}}{2}\left[1-\cos\left(2\phi_i+\gamma_{2_i}\right)\right]+\frac{V_{3_i}}{2}\left[1+\cos\left(3\phi_i+\gamma_{3_i}\right)\right]$$

where each torsional angle i is a sum of three similar cosine terms $E_{\phi 1_i}$, $E_{\phi 2_i}$ and $E_{\phi 3_i}$ where

$$E_{\phi 1_i} = \frac{V_{1_i}}{2}\left[1+\cos\left(\phi_i+\gamma_{1_i}\right)\right]$$
$$E_{\phi 2_i} = \frac{V_{2_i}}{2}\left[1-\cos\left(2\phi_i+\gamma_{2_i}\right)\right]$$
$$E_{\phi 3_i} = \frac{V_{3_i}}{2}\left[1+\cos\left(3\phi_i+\gamma_{3_i}\right)\right]$$

As the sum of convex terms also gives a convex term, each cosine term is convexified separately.

For $V_{1_i} > 0$, $E_{\phi 1_i}$ has a maximum of V_{1_i} and a minimum of 0 at $\phi_{min} = \pi - \gamma_{1_i}$ while for $V_{1_i} < 0$, it has a maximum of 0 and a minimum of V_{1_i}. The two inflexion points are at $\phi^{*\prime} = \pi/2 - \gamma_{1_i}$ and $\phi^{*\prime\prime} = 3\pi/2 - \gamma_{1_i}$. A suitable underestimator for $E_{\phi 1_i}$ is a quadratic underestimator of the form:

$$L_{\phi 1_i}(\phi_i) = \frac{V_{1_i}}{C_{1_i}}(\phi_i - \phi_{min})^2 + K_{1_i} \tag{6.16}$$

where both C_{1_i} and K_{1_i} are constants given by

$$C_{1_i} = \frac{2((\phi_2-\phi_{min})^2-(\phi_1-\phi_{min})^2)}{\cos(\phi_2+\gamma_{1_i})-\cos(\phi_1+\gamma_{1_i})}$$

$$K_{1_i} = \frac{V_{1_i}}{2}(1+\cos(\phi_1+\gamma_{1_i})) - \frac{V_{1_i}}{C_{1_i}}(\phi_1-\phi_{min})^2$$

where $\frac{V_{1_i}}{C_{1_i}}$ defines the curvature of the quadratic equation and K_{1_i} is a constant which is added to the same equation. The values of ϕ_1 and ϕ_2 are chosen where the lower bounding

term, $L_{\phi 1_i}$, would meet the nonconvex cosine term, $E_{\phi 1_i}$. ϕ_1 and ϕ_2 can take the value of ϕ_i^L, ϕ_i^U or ϕ_{min} depending on the values of the bounds on ϕ_i which is shown in the pseudocode for the convexification of the first torsional term where $\phi_i^L \leq \phi_i \leq \phi_i^U$ and $0 \leq \phi_i^L \leq \phi_i^U \leq 2\pi$:

if $(\phi_i^L \leq \phi_{min} \leq \phi_i^U)$
  if ($\phi_i \leq \phi_{min}$)
    if ($\phi_i^L < \phi^{*\prime}$)
      $L_{\phi 1_i} = V_{1_i}/C_{1_i}(\phi_i - \phi_{min})^2 + K_{1_i}$ with $\phi_1 = \phi_i^L, \phi_2 = \phi_{min}$
    else
      $L_{\phi 1_i} = E_{\phi 1_i}$
  else
    if ($\phi_i^U > \phi^{*\prime\prime}$)
      $L_{\phi 1_i} = V_{1_i}/C_{1_i}(\phi_i - \phi_{min})^2 + K_{1_i}$ with $\phi_1 = \phi_{min}, \phi_2 = \phi_i^U$
    else
      $L_{\phi 1_i} = E_{\phi 1_i}$
else if ($\phi_{min} - \pi/2 \leq \phi_i^L < \phi_i^U \leq \phi_{min} + \pi/2$)
  $L_{\phi 1_i} = E_{\phi 1_i}$
else if ($\phi_i^L < \pi/2 - \gamma_1 < \phi_i^U$)
  $L_{\phi 1_i} = V_{1_i}/C_{1_i}(\phi_i - \phi_{min})^2 + K_{1_i}$ with $\phi_1 = \phi_i^L, \phi_2 = \phi_i^U$
else if ($\phi_i^L < 3\pi/2 - \gamma_1 < \phi_i^U$)
  $L_{\phi 1_i} = V_{1_i}/C_{1_i}(\phi_i - \phi_{min})^2 + K_{1_i}$ with $\phi_1 = \phi_i^L, \phi_2 = \phi_i^U$
else
  $L_{\phi 1_i} = [(E_{\phi 1_i}(\phi_i^U) - E_{\phi 1_i}(\phi_i^L))/(\phi_i^U - \phi_i^L)](\phi_i - \phi_i^L) + E_{\phi 1_i}(\phi_i^L)$

So for example in the first case where $\phi_i^L \leq \phi_{min} \leq \phi_i^U$, if $\phi_i \leq \phi_{min}$ and $\phi_i^L < \phi^{*\prime}$, $L_{\phi 1_i}$ would meet $E_{\phi 1_i}$ at ϕ_i^L and ϕ_{min}. If $\phi_i^L > \phi^{*\prime}$, the objective function is convex so the term does not need to be convexified. Therefore, if the bounds for ϕ_i lie only in the convex region of the potential term, the original objective function is used. If neither the minimum point nor an inflexion point are between the bounds, then the cosine term is concave over the domain of interest and a line between the lower and upper bounds can be used as an underestimator:

$$L_{\phi 1_i}(\phi_i) = \left(\frac{E_{\phi 1_i}(\phi_i^U) - E_{\phi 1_i}(\phi_i^L)}{\phi_i^U - \phi_i^L} \right) (\phi_i - \phi_i^L) + E_{\phi 1_i}(\phi_i^L) \tag{6.17}$$

Fig. 6.6 shows the underestimator for the first torsional term, $L_{\phi 1_i}(\phi)$, for different bounds on ϕ_i. The 'first level' corresponds to the domain $[0, 2\pi]$, the 'second level' to two domains $[0, \pi]$ and $[\pi, 2\pi]$ as obtained after branching on $\phi = \pi$, and the 'third level' to

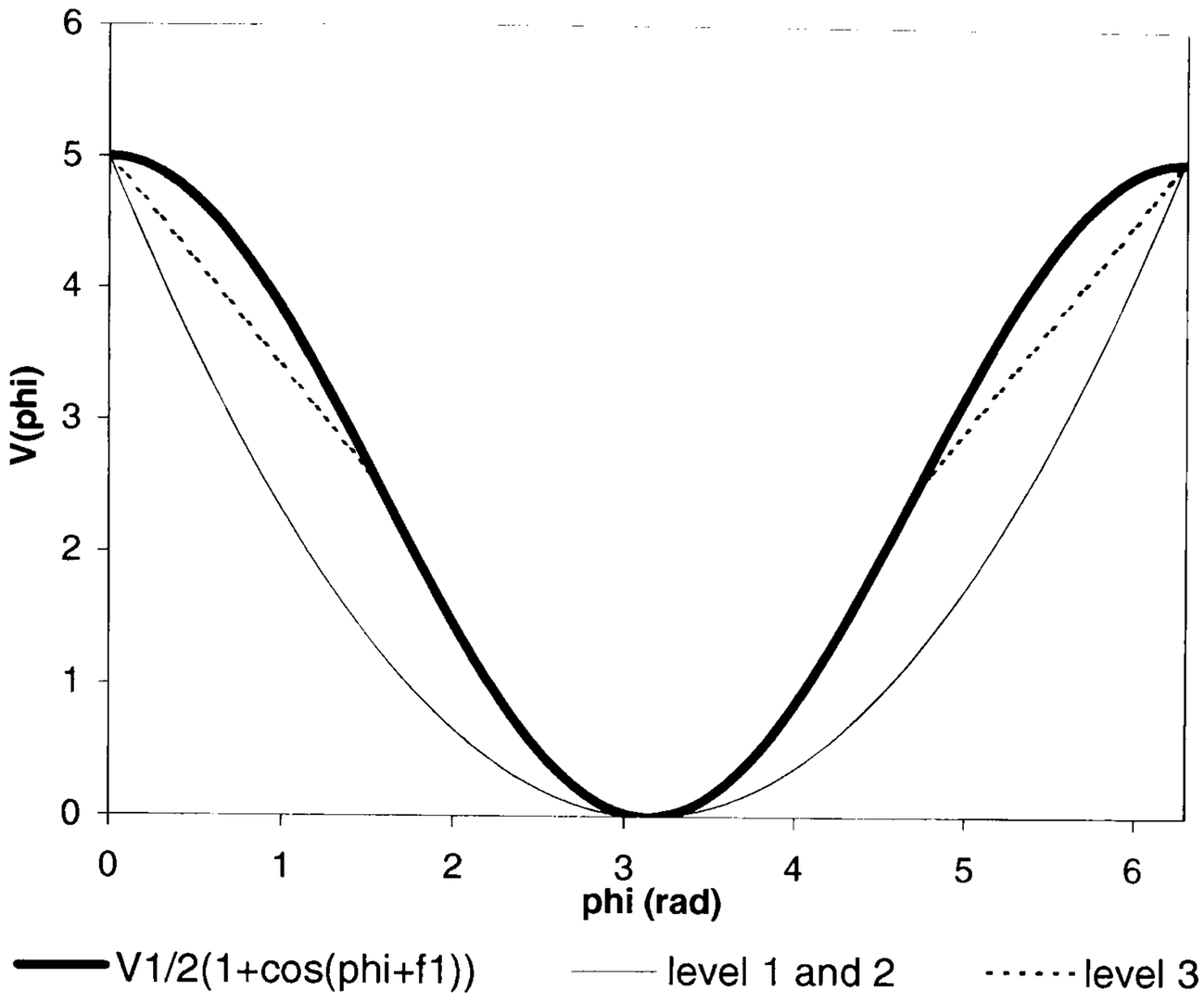


Figure 6.6: Convexification of the first torsional angle energy term - note that levels 1 and 2 have overlapping underestimates

four domains $[0, \pi/2]$, $[\pi/2, \pi]$, $[\pi, 3\pi/2]$ and $[3\pi/2, 2\pi]$. The values $V_{1_i} = 5.0$ and $\gamma_{1_i} = 0$ are used. After branching at $\phi = \pi$, the 'second level' still coincides with the 'first level'. If there is an inflexion point between the lower and upper bounds of the torsion angle ϕ_i, as in the 'first' and 'second level', then Equation 6.16 is used to underestimate the function. If the bounds for the torsion lie only in the convex region between the two inflexion points, then the cosine term is convex within the bounds and the term can be used as is. In the 'third level', since there is neither a minimum point nor an inflexion point for domains [0,$\pi/2$] and [$3\pi/2$,2π], the cosine term is concave in those regions.

A valid underestimator for the second term of the torsional angle, $E_{\phi 2_i}$, is:

$$L_{\phi 2_i}(\phi_i) = \frac{V_{2_i}}{C_{2_i}}(\phi_i - \phi_{avg_i})^2 + K_{2_i} \tag{6.18}$$

where ϕ_{avg_i} is ϕ_{min} if there is only one minimum point between the bounds or $(\phi_{min1} + \phi_{min2})/2$ if there are two minima points between the bounds, and both C_{2_i} and K_{2_i} are

constants given by

$$C_{2_i} = \frac{2((\phi_2 - \phi_{avg})^2 - (\phi_1 - \phi_{avg})^2)}{\cos(2\phi_1 + \gamma_{2_i}) - \cos(2\phi_2 + \gamma_{2_i})}$$

$$K_{2_i} = \frac{V_{2_i}}{2}(1 - \cos(2\phi_1 + \gamma_{2_i})) - \frac{V_{2_i}}{C_{2_i}}(\phi_1 - \phi_{avg})^2$$

Similar to the underestimator of the first torsional term, $\frac{V_{2_i}}{C_{2_i}}$ defines the curvature of the quadratic underestimator and K_{2_i} is added to it. ϕ_1 and ϕ_2 are where $L_{\phi 2_i}$ would meet the second torsional term and may take the values ϕ_i^L, ϕ_i^U, or ϕ_{avg_i}. In the case where ϕ_{avg_i} $= (\phi_{min1} + \phi_{min2})/2$, $K_{2_i} = 2V_{2_i}$. Convexification for the first three levels is shown in Fig. 6.7 where V_{2_i} is -0.05 and γ_{2_i} is 0.0. At the 'third level' after branching at $\phi = \pi/2$ and $\phi = 3\pi/2$, the underestimators are still the same underestimators as at the 'second level'.

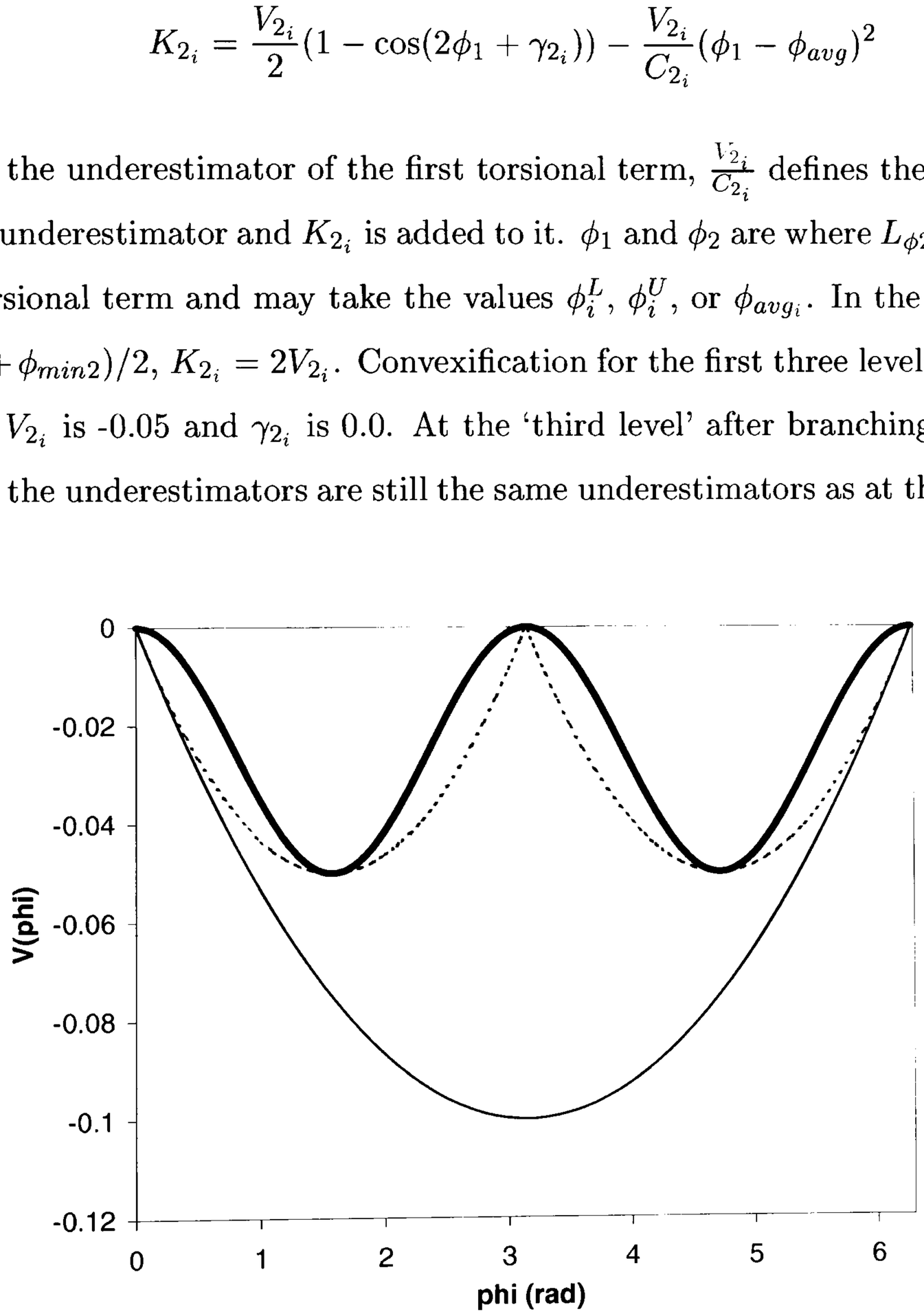


Figure 6.7: Convexification of the second torsional angle energy term - note that the underestimators for levels 2 and 3 coincide

Similarly, for the third torsional angle energy term, $E_{\phi 3}$, the underestimator is:

$$L_{\phi 3_i}(\phi_i) = \frac{V_{3_i}}{C_{3_i}}(\phi_i - \phi_{avg_i})^2 + K_{3_i} \tag{6.19}$$

where ϕ_{avg_i} is ϕ_{min} if there is only one minimum point between the bounds, $(\phi_{min1} + \phi_{min2})/2$ if there are two, and $(\phi_{min1} + \phi_{min2} + \phi_{min3})/3$ if there are three. Both C_{3_i} and K_{3_i} are constants given by

$$C_{3_i} = \frac{2((\phi_2 - \phi_{avg})^2 - (\phi_1 - \phi_{avg})^2)}{\cos(3\phi_2 + \gamma_{3_i}) - \cos(3\phi_1 + \gamma_{3_i})}$$

$$K_{3_i} = \frac{V_{3_i}}{2}(1 + \cos(3\phi_1 + \gamma_{3_i})) - \frac{V_{3_i}}{C_{3_i}}(\phi_1 - \phi_{avg})^2$$

where ϕ_1 and ϕ_2 can be either ϕ_i^L, ϕ_i^U, or ϕ_{avg_i}. If there are two minima within the bounds, K_{3_i} = -3/4V_{3_i} and if there are three, K_{3_i} = -2V_{3_i}. Fig. 6.8 shows the underestimator for the first three levels where V_{3_i} is 0.2 and γ_{3_i} is again 0.0.

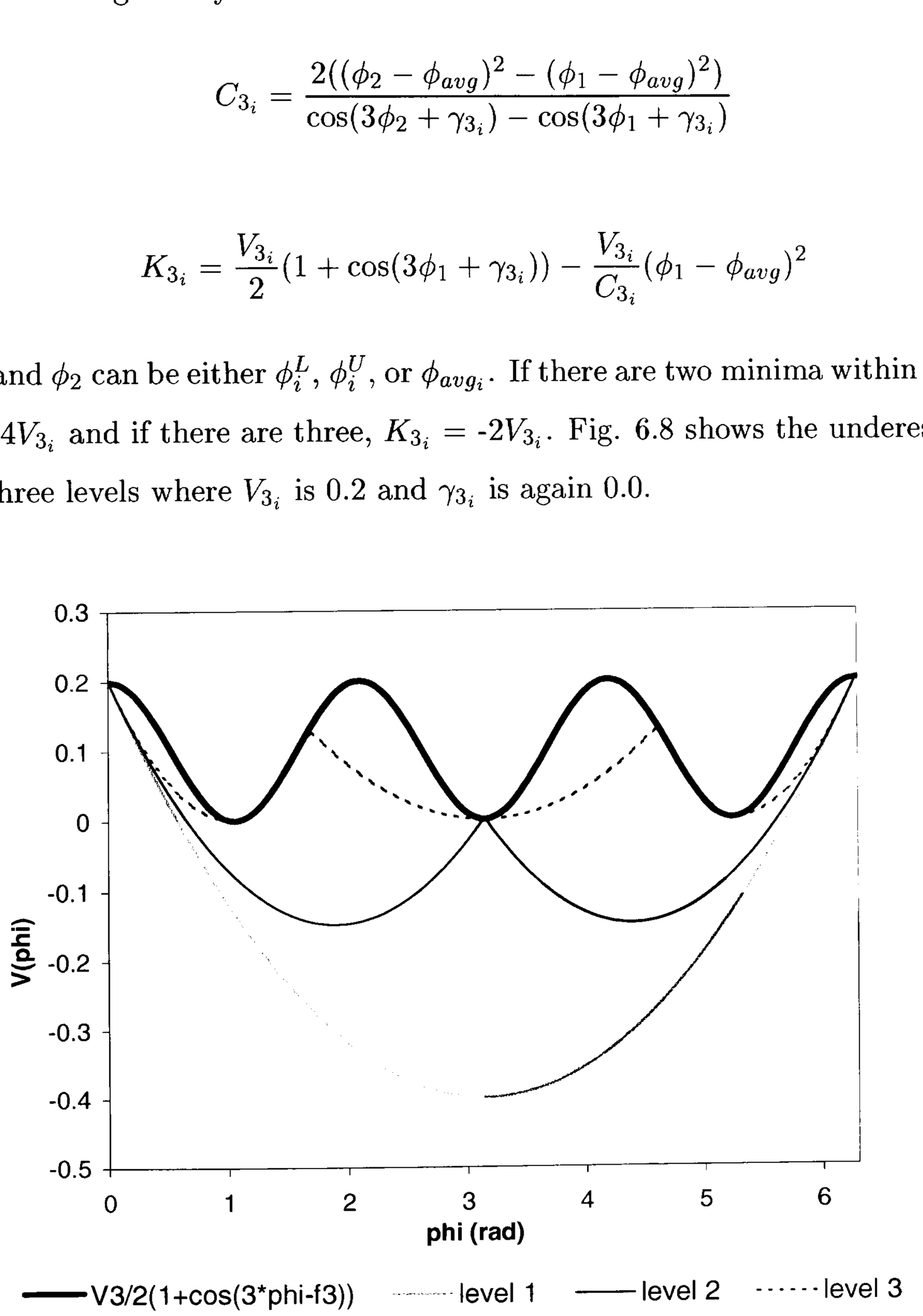


Figure 6.8: Convexification of the third torsional angle energy term

6.4.2 Convexification of Nonbonded Interactions Term

The nonbonded interactions term consists of a Coulombic term and a van der Waals term:

$$E_r = \sum_i \sum_j \left[\frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right] \tag{6.20}$$

The Coulombic term is convex if both the electronic charges q_i and q_j have the same sign. Otherwise, the term is concave. The van der Waals term may be divided into a convex part (power of 12) and a concave part (power of 6). The entire term may be convexified by augmenting it with a convex term to give a lower bound:

$$E_{r_{ij}} = \left[\frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right] + \alpha_{ij}(r_{ij}^U - r_{ij})(r_{ij}^L - r_{ij}) \tag{6.21}$$

where α_{ij} is defined as a non-negative parameter which is greater than or equal to the negative of one-half of the minimum value of the second order derivative $H_{r_{ij}}$, of the term, $E_{r_{ij}}$:

$$H_{r_{ij}} = \left[\frac{2q_i q_j e^2}{r_{ij}^3} + 4\epsilon_{ij} \left[\left(\frac{156\sigma_{ij}^{12}}{r_{ij}^{14}} \right) - \left(\frac{42\sigma_{ij}^{6}}{r_{ij}^{8}} \right) \right] \right] \tag{6.22}$$

The α_{ij} value is determined by performing a local minimisation on the second order derivative. If the term only has one minimum, this minimisation will always yield a rigorous value for α_{ij}. Graphical analysis indicates that this is the case but we will prove it analytically in *Appendix F*.

Convex terms are left unchanged while the secant underestimator is used to underestimate concave terms:

$$L_{r_{ij}}(r_{ij}) = \left(\frac{E_{r_{ij}}(r_{ij}^U) - E_{r_{ij}}(r_{ij}^L)}{r_{ij}^U - r_{ij}^L} \right) (r_{ij} - r_{ij}^L) + E_{r_{ij}}(r_{ij}^L) \tag{6.23}$$

6.5 Convexification of the Constraints

There are constraints in the original problem which must be convexified. These constraints state the relationships between the *dependent* variables and the *independent* variables. By

convexifying them, this relaxes the feasible region and removes some of the dependencies. However, the feasible domain is larger as a result and the solution found could be outside the original domain. This can be countered by adding further constraints and additional variables.

6.5.1 Symbolic Reformulations

Smith and Pantelides (1999) introduced a symbolic reformulation of the nonconvex equations into equivalent linear equations by introducing new variables. This technique will be used to convexify the nonconvex constraints.

Nonbonded Distances

The relationship between the nonbonded distance r with the internal coordinates in the atom chain, $r = f_r(b, \theta, \phi)$, is one of the constraints of the optimisation problem. The relationship is highly nonconvex so we have utilised a symbolic reformulation step to formulate a convex relaxation for the equation. For a 3-atom chain, we have:

$$r_{ik}^2 = b_{ij}^2 + b_{jk}^2 - 2\cos(\theta_{ijk}) b_{ij} b_{jk}$$

which can be reformulated as a linear constraint

$$w_1 = w_2 + w_3 - 2w_4 \tag{6.24}$$

with the additional nonlinear relationships

$$w_1 = r_{ik}^2 \tag{6.25}$$

$$w_2 = b_{ij}^2 \tag{6.26}$$

$$w_3 = b_{jk}^2 \tag{6.27}$$

$$w_4 = w_5 w_6 \tag{6.28}$$

$$w_5 = \cos(\theta_{ijk}) \tag{6.29}$$

$$w_6 = b_{ij} b_{jk} \tag{6.30}$$

where w_1, w_2 and w_3 are expressed as a function of quadratic terms, w_4 and w_6 are expressed as a function of bilinear terms and w_5 is expressed as a trigonometric term which may be convexified as a quadratic term if the term is nonconvex in that region or a linear term if the term is concave (similar to the torsional potential energy in Section 6.4).

The quadratic terms may be bounded using two convex inequalities:

$$r_{ik}^2 - w_1 \leq 0 \tag{6.31}$$

$$-\frac{(r_{ik}^U)^2 - (r_{ik}^L)^2}{r_{ik}^u - r_{ik}^L}(r_{ik} - r_{ik}^L) - (r_{ik}^L)^2 + w_1 \leq 0 \tag{6.32}$$

The bilinear term may be relaxed and included in the minimisation problem by the inclusion of four linear inequality constraints (Al-Khayyal and Falk, 1983):

$$w_4 \geq w_5^L w_6 + w_6^L w_5 - w_5^L w_6^L \tag{6.33}$$

$$w_4 \geq w_5^U w_6 + w_6^U w_5 - w_5^U w_6^U \tag{6.34}$$

$$w_4 \leq w_5^U w_6 + w_6^L w_5 - w_5^U w_6^L \tag{6.35}$$

$$w_4 \leq w_5^L w_6 + w_6^U w_5 - w_5^L w_6^U \tag{6.36}$$

The trigonometric term may be bounded by using two inequalities of opposite signs:

$$w_5 - \cos(\theta_{ijk}) \leq 0 \tag{6.37}$$

$$-w_5 + \cos(\theta_{ijk}) \leq 0 \tag{6.38}$$

The two inequalities are then convexified independently by converting the cosine term to a convex quadratic term. If θ^1 refers to $\max(\cos(\theta^L), \cos(\theta^U))$, and θ^2 refers to the angle which gives the lower bound of $\cos(\theta)$ from $\min(\cos(\theta^L), \cos(\theta^U), \cos(\theta_{min}))$ where θ_{min} is the minimum point of the cosine function in the domain and $\theta^L \leq \theta \leq \theta^U$, the constraint can be bounded by:

$$-\cos(\theta_{ijk}) = C'(\theta - \theta^1)^2 + K' \tag{6.39}$$

$$\cos(\theta_{ijk}) = C''(\theta - \theta^2)^2 + K'' \tag{6.40}$$

The constants are found by equating the lower bounds with the cosine terms at the two points where both functions meet as in Section 6.4. If $\theta^2 = \theta_{min}$, then for Equation 6.39. the two points are θ^L and θ^U and for Equation 6.40, the two points are θ^1 and θ_{min}. Otherwise, the two points are θ^L and θ^U for both equations.

As for the 4-atom chain, the constraint is defined as:

$$\begin{aligned} r_{il}^2 =& b_{ij}^2 + b_{jk}^2 + b_{kl}^2 - 2\cos(\theta_{ijk})b_{ij}b_{jk} - 2\cos(\theta_{jkl})b_{jk}b_{kl} \\ & - 2(\sin(\theta_{ijk})\sin(\theta_{jkl})\cos(\phi_{ijkl}) - \cos(\theta_{ijk})\cos(\theta_{jkl}))b_{ij}b_{kl} \end{aligned}$$

and the reformulation as a linear constraint is:

$$w_1 = w_2 + w_3 + w_4 - 2w_5 - 2w_8 - 2w_{17} + 2w_{19} \tag{6.41}$$

with the additional nonlinear relationships

$$w_1 = r_{il}^2$$

$$w_2 = b_{ij}^2$$

$$w_3 = b_{jk}^2$$

$$w_4 = b_{kl}^2$$

$$w_5 = w_6 w_7$$

$$w_6 = \cos(\theta_{ijk})$$

$$w_7 = b_{ij}b_{jk}$$

$$w_8 = w_9 w_{10}$$

$$w_9 = \cos(\theta_{jkl})$$

$$w_{10} = b_{jk}b_{kl}$$

$$w_{11} = \sin(\theta_{ijk})$$

$$w_{12} = \sin(\theta_{jkl})$$

$$w_{13} = \cos(\phi_{ijkl})$$

$$w_{14} = w_{11}w_{12}$$

$$w_{15} = w_{13}w_{14}$$

$$w_{16} = b_{12}b_{34}$$

$$w_{17} = w_{15}w_{16}$$

$$w_{18} = w_6 w_9$$

$$w_{19} = w_{18}w_{16}$$

These nonlinear relationships are bounded depending on whether they are quadratic, bilinear or trigonometric functions. The longer chains are convexified in a similar manner.

Dependent **Angles**

The constraints defining the relationships between the *dependent* angles with the independent variables are defined in Equation 6.9.

$$\cos(\theta_{jkl}) = \cos(\theta_{ikj})\cos(\theta_{ikl}) + \sin(\theta_{ikj})\sin(\theta_{ikl})\cos(\phi_{jikl})$$

where k is the atom bonded to atoms i, j and l and $\cos(\phi_{jikl}) = \cos(\phi_{rnikl} - \phi_{mikj})$.

The equation is reformulated as

$$w_1 = w_2 + w_3 \tag{6.42}$$

with the following additional constraints

$$w_1 = \cos(\theta_{jkl})$$

$$w_2 = w_4 w_5$$

$$w_3 = w_7 w_{10}$$

$$w_4 = \cos(\theta_{ikj})$$

$$w_5 = \cos(\theta_{ikl})$$

$$w_6 = \phi_{mikl} - \phi_{mikj}$$

$$w_7 = \cos(w_6)$$

$$w_8 = \sin(\theta_{ikj})$$

$$w_9 = \sin(\theta_{ikl})$$

$$w_{10} = w_8 w_9$$

Dependent Torsions

The torsional angle constraints based on Equation 6.11 state that:

$$\cos(\phi_{ijkl} - \phi_{ijkm}) = \cos(\phi_{hjkl} - \phi_{hjkm}) \tag{6.43}$$

$$\cos(\phi_{hjkm} - \phi_{ijkm}) = \cos(\phi_{hjkl} - \phi_{ijkl}) \tag{6.44}$$

The reformulation is then simply:

$$w_1 = w_2 \tag{6.45}$$

$$w_5 = w_6 \tag{6.46}$$

where

$$w_1 = \cos(w_3)$$

$$w_2 = \cos(w_4)$$

$$w_3 = \phi_{ijkl} - \phi_{ijkm}$$

$$w_4 = \phi_{hjkl} - \phi_{hjkm}$$

$$w_5 = \cos(w_7)$$

$$w_6 = \cos(w_8)$$

$$w_7 = \phi_{hjkm} - \phi_{ijkm}$$

$$w_8 = \phi_{hjkl} - \phi_{ijkl}$$

6.5.2 Bounds on the *Dependent* Variables

It is obvious from the form of the convex underestimators derived in Section 6.4 and 6.5.1 that bounds on the *dependent* variables must be known before an underestimator can be computed. Once these bounds are known, they can be used to generate a convex relaxation of the feasible region. As the bond lengths and angles do not vary much from equilibrium, these variables are already physically constrained. On the other hand, the torsional angles and nonbonded distances usually possess large differences in their lower and upper bounds. Therefore, arithmetic operations to determine the bounds have to be

conducted using interval arithmetic rules.

Rules for performing interval operations (Moore, 1966; Hansen, 1980) are:

Addition

$$[a,b]+[c,d]=[a+c,b+d] \tag{6.47}$$

Negation

$$[a,b]-[c,d]=[a-d,b-c] \tag{6.48}$$

Multiplication

$$[a,b]*[c,d]=[\min(ac,ad,bc,bd),\max(ac,ad,bc,bd)] \tag{6.49}$$

Division

$$[a,b]/[c,d]=[\min(a/c,a/d,b/c,b/d),\max(a/c,a/d,b/c,b/d)] \tag{6.50}$$

These interval arithmetic equations can be used to calculate bounds on the *dependent* variables. The bounds should be as tight as possible to achieve the optimal underestimator for the constraints. Whenever a variable is branched on, i.e. two subdomains are generated each corresponding to half of the original range of the variable, the bounds on the *dependent* variables are recalculated. Therefore, this ensures that the bounds get increasingly tighter as iterations proceed and so does the convex lower bounding approximation.

Dependent **Angles**

There are three steps to determining bounds for angles. First, for the *independent* angles, the bounds are already defined. Then for the *dependent* angles where the improper torsional angles for these angles are known, the bounds can be calculated from Equation 6.9. Finally, for the *dependent* angles where the improper torsional angles are not known,

the improper torsional angle will have to be calculated first as follows:

$$\phi^L_{ijkl} = \phi^L_{hjkl} - \phi^U_{hjki}$$
$$\phi^U_{ijkl} = \phi^U_{hjkl} - \phi^L_{hjki}$$

Once the bounds for ϕ_{ijkl} are known, the bounds for the *dependent* angle are calculated by applying interval arithmetic to Equation 6.51 which is similar to Equation 6.9.

$$\cos(\theta_{ikl}) = \sin(\theta_{jki})\sin(\theta_{jkl})\cos(\phi_{ijkl}) + \cos(\theta_{jki})\cos(\theta_{jkl}) \tag{6.51}$$

We use the same examples that are in Section 6.3, giving bounds on the angles rather than fixed values. The bounds for angles θ_{4-8-11}, θ_{4-8-12} and θ_{4-8-13} are given by [111.061^o, 112.061^o], [110.339^o, 112.339^o] and [110.699^o, 111.699^o], and those for $\phi_{11-4-8-12}$ and $\phi_{11-4-8-13}$ are [119.399^o, 120.399^o] and [239.051^o, 240.051^o]. The bounds for $\phi_{12-4-8-13}$ may be determined as $[\phi^L_{11-4-8-13} - \phi^U_{11-4-8-12}, \phi^U_{11-4-8-13} - \phi^L_{11-4-8-12}] = [239.051^o - 120.299^o, 240.051^o - 119.399^o] = [118.652^o, 120.652^o]$. The bounds for angle $\theta_{12-8-13}$ are then calculated using Equation 6.9 in this particular case as:

$$\begin{aligned}
(\cos(\theta_{12-8-13}))^L &= \sin(\theta^L_{4-8-12})\sin(\theta^L_{4-8-13})\cos(\phi^U_{12-4-8-13}) + \cos(\theta^L_{4-8-12})\cos(\theta^L_{4-8-13}) \\
&= \sin(110.339^o)\sin(110.699^o)\cos(120.652^o) + \cos(110.339^o)\cos(110.699^o) \\
&= -0.324324 \\
\theta^U_{12-8-13} &= \cos^{-1}(-0.324324) \\
&= 108.925^o
\end{aligned}$$

$$
\begin{aligned}
(\cos(\theta_{12-8-13}))^U &= \sin(\theta^U_{4-8-12})\sin(\theta^U_{4-8-13})\cos(\phi^L_{12-4-8-13}) + \cos(\theta^U_{4-8-12})\cos(\theta^U_{4-8-13}) \\
&= \sin(111.339^o)\sin(111.699^o)\cos(118.652^o) + \cos(111.339^o)\cos(111.699^o) \\
&= -0.280428 \\
\theta^L_{12-8-13} &= \cos^{-1}(-0.280428) \\
&= 106.285^o
\end{aligned}
$$

The lower bound should be rounded down and the upper bound rounded up to machine-representable numbers to ensure that the bounds are strictly correct. The values used in the interval arithmetic calculation depend on the bounds for the variables.

If however, the bounds for one of the angles θ_{4-8-13} are large i.e. $[30^o, 60^o]$, the calculated bounds for angle $\theta_{12-8-13}$ are also large.

$$
\begin{aligned}
(\cos(\theta_{12-8-13}))^L &= \sin(110.339^o)\sin(60^o)\cos(120.652^o) + \cos(111.339^o)\cos(30^o) \\
&= -0.729125 \\
\theta^U_{12-8-13} &= \cos^{-1}(-0.729125) \\
&= 136.813^o
\end{aligned}
$$

$$
\begin{aligned}
(\cos(\theta_{12-8-13}))^U &= \sin(111.339^o)\sin(60^o)\cos(118.652^o) + \cos(110.339^o)\cos(60^o) \\
&= -0.397095 \\
\theta^L_{12-8-13} &= \cos^{-1}(-0.397095) \\
&= 113.3967^o
\end{aligned}
$$

The bounds for the angle can also be found analytically with respect to the other variables. In order to find the min/ max values for the angle $\theta_{12-8-13}$ with respect to angle θ_{4-8-13} if the other variables are all given, the necessary condition is that the first order derivative

is equal to 0:

$$\begin{aligned} \frac{\partial \cos\theta_{12-8-13}(\theta_{4-8-12},\theta_{4-8-13},\phi_{12-4-8-13})}{\partial\theta_{4-8-13}} &= 0 \\ -\sin(\theta_{4-8-13})\cos(\theta_{4-8-12}) + \cos(\theta_{4-8-13})\sin(\theta_{4-8-12})\cos(\phi_{12-4-8-13}) &= 0 \end{aligned} \tag{6.52}$$

$$\tan(\theta_{4-8-13}) = \tan(\theta_{4-8-12})\cos(\phi_{12-4-8-13}) \tag{6.53}$$

To show if this point is minimum or maximum, we have to look at the second order derivative:

$$\begin{aligned} \frac{\partial^2 f(\theta,\phi)}{\partial\theta^2_{4-8-13}} &= -\cos(\theta_{4-8-13})\cos(\theta_{4-8-12}) - \sin(\theta_{4-8-13})\sin(\theta_{4-8-12})\cos(\phi_{12-4-8-13}) \\ \frac{\partial^2 f(\theta,\phi)}{\partial\theta^2_{4-8-13}} &= -\cos(\theta_{4-8-13})\cos(\theta_{4-8-12}) - \frac{\sin^2(\theta_{4-8-13})\cos(\theta_{4-8-12})}{\cos(\theta_{4-8-13})} \\ \frac{\partial^2 f(\theta,\phi)}{\partial\theta^2_{4-8-13}} &= -\frac{\cos(\theta_{4-8-12})}{\cos(\theta_{4-8-13})}(\cos^2(\theta_{4-8-13}) + \sin^2(\theta_{4-8-13})) \\ \frac{\partial^2 f(\theta,\phi)}{\partial\theta^2_{4-8-13}} &= -\frac{\cos(\theta_{4-8-12})}{\cos(\theta_{4-8-13})} \end{aligned}$$

$$\frac{\partial^2 f(\theta,\phi)}{\partial\theta^2_{4-8-13}} \geq 0 \text{ (min) if the cosines have different signs} \tag{6.54}$$

$$\frac{\partial^2 f(\theta,\phi)}{\partial\theta^2_{4-8-13}} \leq 0 \text{ (max) if the cosines have the same sign} \tag{6.55}$$

Applying interval arithmetic to Equation 6.53 gives bounds on the value of θ_{4-8-13} which corresponds to a stationary point in $\theta_{12-8-13}$:

$$\begin{aligned} \tan^L(\tilde{\theta}_{4-8-13}) &= \tan(111.339^o)\cos(118.652^o) \\ \tilde{\theta}^L_{4-8-13} &= 50.828^o \\ \tan^U(\tilde{\theta}_{4-8-13}) &= \tan(110.339^o)\cos(120.652^o) \\ \tilde{\theta}^U_{4-8-13} &= 53.979^o \end{aligned}$$

To determine whether Equation 6.53 corresponds a value of $\cos(\theta_{12-8-13})$ which is a maximum or a minimum in $\cos(\theta_{12-8-13})$, we consider the regions of $\cos(\tilde{\theta}_{4-8-13})$ and $\cos(\theta_{4-8-12})$ in the domain $\tilde{\theta}_{4-8-13} \in [50.828^o, 53.979^o]$ and $\theta_{4-8-12} \in [110.339^o, 111.339^o]$. In this case, $\cos(\tilde{\theta}_{4-8-13}) > 0$ and $\cos(\theta_{4-8-12}) < 0$, so Equation 6.54 holds and Equation 6.53 gives a minimum in $\cos(\theta_{12-8-13})$. Thus a maximum in $\theta_{12-8-13}$ occurs for a value of θ_{4-8-13} in the domain $[50.828^o, 53.979^o]$. This information can be used to calculate a more accurate value of $\theta^U_{12-8-13}$.

$$\begin{aligned}
\cos^L(\theta_{12-8-13}) =& \sin(53.979^o)\sin(110.339^o)\cos(120.652^o) + \cos(50.828^o)\cos(111.339^o) \\
=& -0.616491 \\
\theta^U_{12-8-13} =& 128.060^o
\end{aligned}$$

The above value gives a more accurate upper bound for $\theta_{12-8-13}$ than the straightforward application of interval arithmetic which gives 136.813^o.

In some cases, the bounds on the angles may be such that both Equations 6.54 and 6.55 are satisfied over the domain of interest. Thus if the situation exists where $\tan(\theta_{4-8-13})$ may be both negative and positive, then the bounds of $\tilde{\theta}_{4-8-13}$ using Equation 6.53 are found from the minimum positive value for $\tan(\tilde{\theta}_{4-8-13})$ using interval arithmetic for the lower bound of $\tilde{\theta}_{4-8-13}$ and maximum negative value of $\tan(\tilde{\theta}_{4-8-13})$ for the upper bound.

Dependent Torsions

The next step is to determine bounds for the *dependent* torsions using Equation 6.11.

$$\phi_{ijkl} = \phi_{hjkm} - \phi_{hjki} + \phi_{mjkl} \tag{6.56}$$

As the domain for the torsional angles is $[0.0^o, 360^o]$, if the bounds for the *dependent* torsions are calculated to be outside the domain, the bounds are redefined to be within the bounds as follows:

1. If $((\phi^L < 0.0^o$ AND $\phi^U > 0.0^o)$ OR $(\phi^L < 360^o$ AND $\phi^U > 360^o)$, the new bounds are $[0^o, 360^o]$

2. If ($\phi^L < 0.0^o$ AND $\phi^U < 0.0^o$), the new bounds are $[\phi^L + 360^o, \phi^U + 360^o]$

3. If ($\phi^L > 360^o$ AND $\phi^U > 360^o$), the new bounds are $[\phi^L - 360^o, \phi^U - 360^o]$

We used the same example to calculate the torsion $\phi_{10-4-8-12}$ given that the bounds for $\phi_{11-4-8-12}$ are $[119.399^o, 120.399^o]$ and for $\phi_{3-4-8-11}$ are $[58.958^o, 60.958^o]$. In order to determine the bounds for $\phi_{3-4-8-10}$, the bounds for angle θ_{8-4-10} were first calculated to be $[108.221^o, 110.087^o]$. As the bounds for angles θ_{3-4-8} and θ_{3-4-10} are $[112.042^o, 113.042^o]$ and $[109.045^o, 110.045^o]$ respectively, the bounds for $\phi_{3-4-8-10}$ are then calculated to be $[236.514^o, 239.741^o]$ as the value of the improper torsional angle $\phi_{8-3-4-10}$ is 238.376^o.

Therefore using interval arithmetic, the bounds for $\phi_{10-4-8-12}$ are calculated as:

$$\begin{aligned}\phi^U_{10-4-8-12} =& \phi^U_{3-4-8-11} - \phi^L_{3-4-8-10} + \phi^U_{11-4-8-12} \\ =& 60.958^o - (236.486^o) + 120.399^o \\ =& -55.129^o\end{aligned}$$

$$\begin{aligned}\phi^L_{10-4-8-12} =& \phi^L_{3-4-8-11} - \phi^U_{3-4-8-10} + \phi^L_{11-4-8-12} \\ =& 58.958^o - (239.741^o) + 119.398^o \\ =& -61.385^o\end{aligned}$$

The bounds are then $[298.615^o, 304.871^o]$ as the torsional angles are redefined to meet the bounds of $[0^o, 360^o]$.

6.5.3 Triangle and Tetrangle Inequalities

Bounds on the nonbonded distances can be tightened by using triangle inequalities. The triangle inequalities state that for any three variable sides of a triangle *abc*:

1. The upper bound of the side r_{ab} is not more than the sum of the upper bounds of sides r_{ac} and r_{bc}.

$$r^U_{ab} \leq r^U_{ac} + r^U_{bc} \tag{6.57}$$

2. The lower bound of the side r_{ab} is not less than the difference between the lower bound of r_{ac} and the upper bound of r_{bc}, or the lower bound of r_{bc} and the upper bound of r_{ac}.

$$r_{ab}^{L} \geq r_{ac}^{L} - r_{bc}^{U} \tag{6.58}$$

$$r_{ab}^{L} \geq r_{bc}^{L} - r_{ac}^{U} \tag{6.59}$$

Crippen and Havel (1988) presented a triangle inequality bound smoothing algorithm which is summarised as follows:

For any nonbonded distance (i,j), $\forall k$ where k is an atom not equal to i or j:

- If $r_{ij}^{U} > r_{ik}^{U} + r_{kj}^{U}$, then $r_{ij}^{U} = r_{ik}^{U} + r_{kj}^{U}$.
- If $r_{ij}^{L} < r_{ik}^{L} - r_{kj}^{U}$, then $r_{ij}^{L} = r_{ik}^{L} - r_{kj}^{U}$.
- Else if $r_{ij}^{L} < r_{jk}^{L} - r_{ki}^{U}$, then $r_{ij}^{L} = r_{jk}^{L} - r_{ki}^{U}$.

There exist also tetrangle inequality bounds but as an algebraic system of inequalities as Equations 6.57 and 6.58 which are self-consistent do not exist and there is no proof that the algorithm proposed by Crippen and Havel will converge in a finite number, they are not included here in this optimisation algorithm.

6.6 General Structure of Optimisation Framework

The molecular system to be optimised is presented in the form of the BOSS Z-matrix. The BOSS program is then called to determine all the bond lengths, bond angles, torsional angles and nonbonded distances that are in the system together with the parameters of the OPLS force field. The improper torsional angles are then removed from the list of torsional angles. In the case of the nonbonded distances, the length of the atom chain is calculated. Bounds on the *independent* variables are given. Bounds on the bonds and angles are tight as they do not vary much from equilibrium (Jensen, 1999). All the bond lengths are specified in the Z-matrix and the bounds given for them are $[b_i - 0.01, b_i + 0.01]$ in Angstroms as it has been observed that the lengths do not usually vary by more than

that in equilibrium systems. Similarly bounds for the *independent* angles are straightforward so the bounds are given as $[\theta_i - 0.02, \theta_i + 0.02]$ in radians. Torsional angles however are rotatable from [0.0, 2π] but the improper torsional angles are bounded using $[\phi_i - 0.02, \phi_i + 0.02]$ also in radians. The bounds are based on observation by comparing the different conformers of a molecule. Bounds on the *dependent* variables are calculated using interval arithmetic.

The nonbonded distances are tightened using triangle inequalities and constraints based on symbolic reformulation are added. α-values are then determined from a local minimisation of the Hessian which has been shown in Section 6.4 to give a valid lower bounding function to the nonbonded distance term.

In the optimisation branch-and-bound algorithm, the upper bound, E^U, is found from a local minimisation of the *independent* variables. If the upper bound found is lower than the best upper bound, E^{U*}, the value is substituted as the current best upper bound. The starting point of the lower bound minimisation is the optimised structure of the upper bound.

The lower bounding function is then minimised to get E^L subjected to the bounds for all the variables and the constraints. The results of the optimisation are added to the node of the tree. Each node contains the following information: the upper and lower bound values, the level of the node and pointers to the parent and child nodes. The lower bound E^L is added to the list of lower bounds where the lower bounds are listed in increasing order and contain the following information: lower bound, pointer to the next lower bound in the list and pointer to the node of the branch-and-bound tree which has that lower bound.

If the optimisation on the node is finished, the next best lower bound, E^{L*}, is obtained from the list of lower bounds. The choice of the branching variable is dependent on a comparison of the absolute ranges of all the independent variables and the variable with the largest range is chosen for branching. The entire process is continued until E^{L*} is within a tolerance ϵ of the best upper bound, E^{U*} as shown in Fig. 6.9.

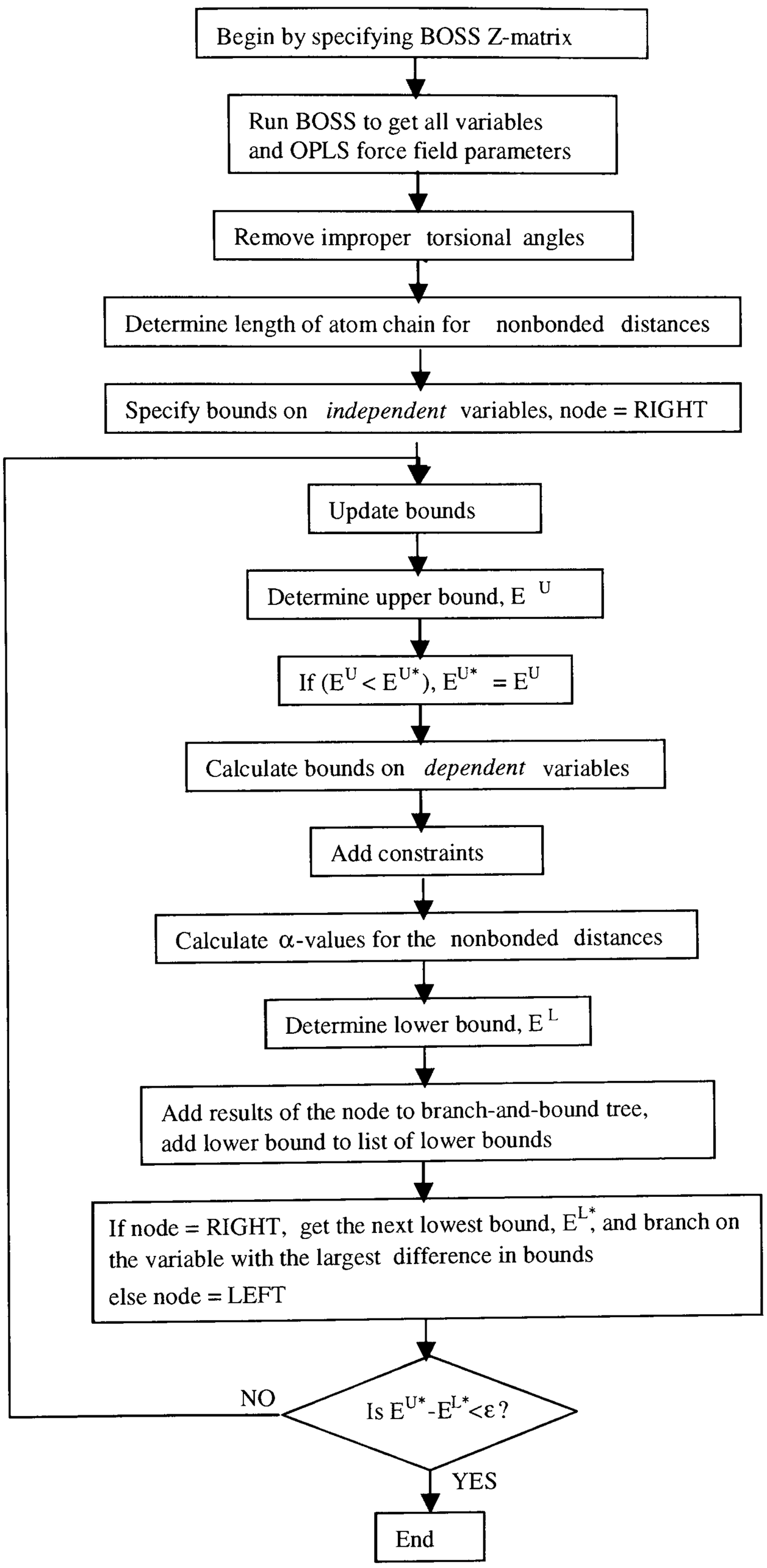


Figure 6.9: Global optimisation branch-and-bound framework for molecular systems

6.7 Case Studies

The optimisation algorithm is presented for three systems with several minima. The first is an ethylamine molecule, while the second and the third are an n-butane/ ethylamine pair and an ethylamine/ ethylamine pair respectively. The runs were conducted on a XP1000 667 MHz DEC Alpha. The convex lower bounding problem is solved using SNOPT (Stanford Business Software, Inc.), a sparse SQP algorithm with limited memory quasi-Newton approximations to the Hessian of the Lagrangian.

6.7.1 Ethylamine molecule

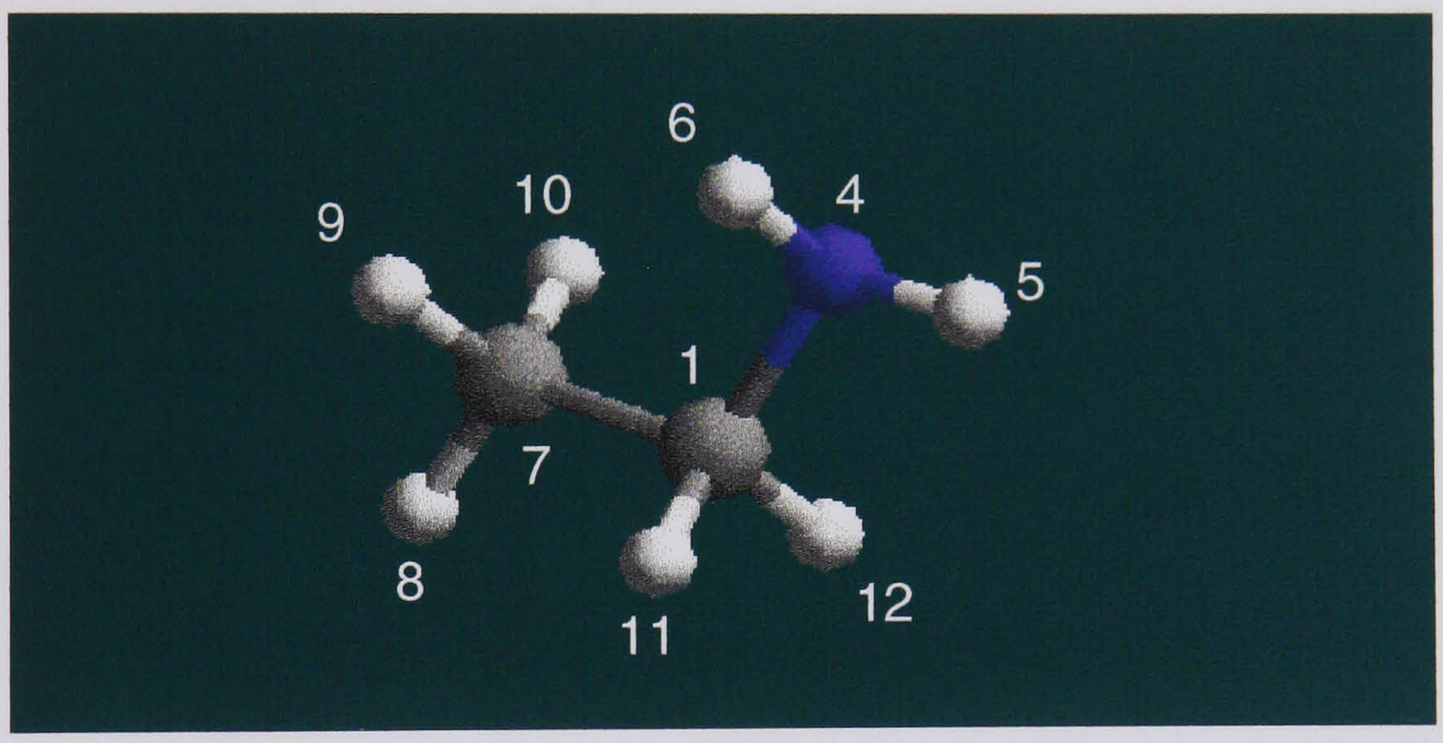


Figure 6.10: Ethylamine molecule showing the atom numbers

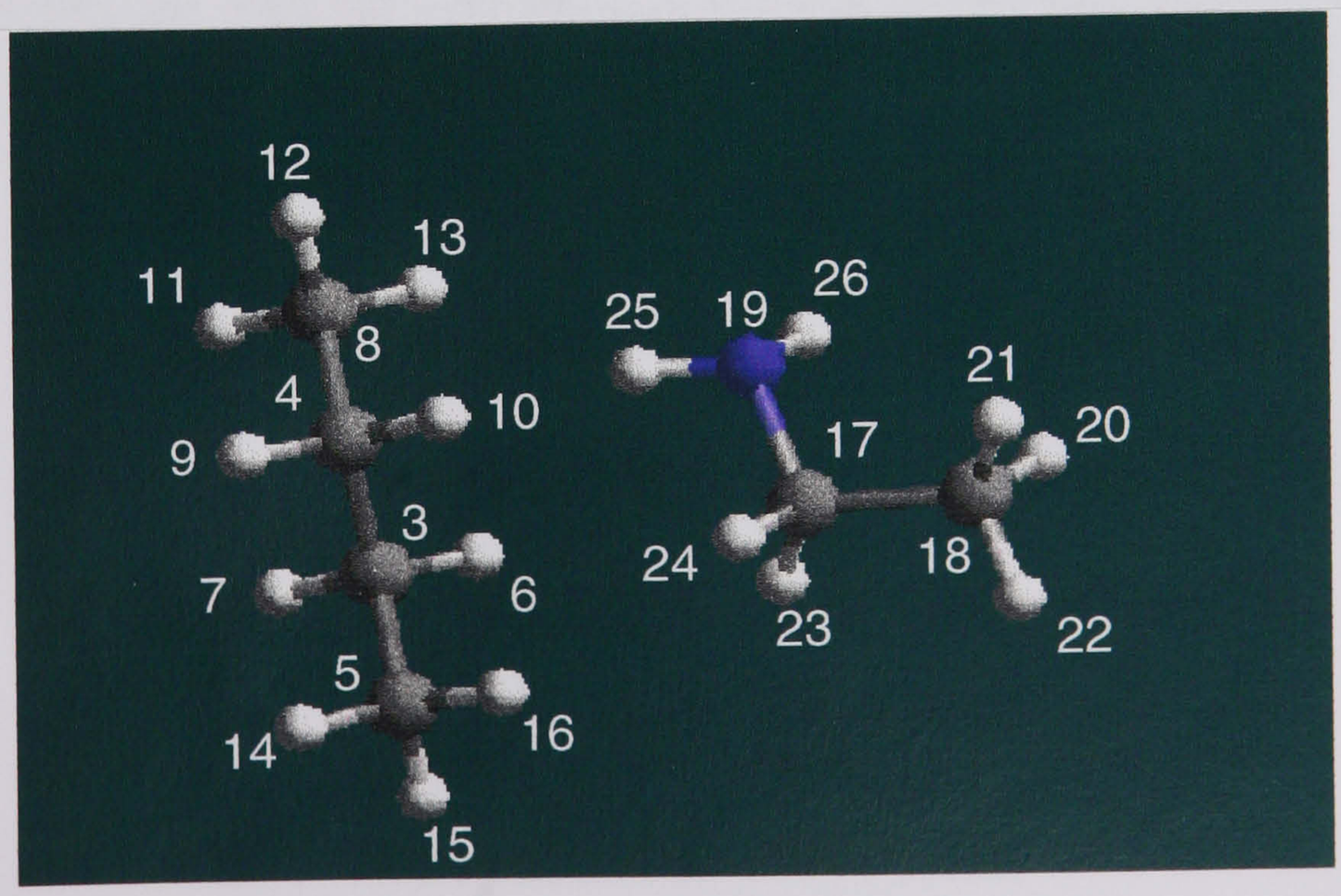


Figure 6.11: n-Butane/ ethylamine molecule showing the atom numbers

Using the Z-matrix shown below, the molecule is minimised as shown in Fig. 6.10.

Atom Number	Atom Type	Bond Atom	Bond Length	Angle Atom	Bond Angle	Torsion Atom	Torsional Angle
1	*C*	0	.000000	0	.000000	0	.000000
2	*X*	1	1.000000	0	.000000	0	.000000
3	*X*	2	1.000000	1	90.000000	0	.000000
4	*N*	1	1.450160	2	90.000000	3	270.000000
5	*H*	4	1.011243	1	110.328155	2	240.000000
6	*H*	4	1.010763	1	110.496898	5	242.773333
7	*C*	1	1.530398	4	109.336410	5	182.207408
8	*H*	7	1.090441	1	110.651715	4	179.100863
9	*H*	7	1.090405	1	111.733848	8	119.741254
10	*H*	7	1.090881	1	111.291973	8	240.679874
11	*H*	1	1.091310	4	109.984620	7	238.308621
12	*H*	1	1.091371	4	109.393610	7	120.616741

The nine bonds are all given in the Z-matrix. There are 15 angles of which 8 are *independent* variables, and 5 improper torsional angles are also defined. Then there are 15 torsions of which only 2 are defined in the Z-matrix and 21 nonbonded distances of which 15 have 4 atom chains and the other 6 have 5 atom chains.

To derive the lower bounding problem, the constraints for the torsional angles are added (Equation 6.43). For each two torsions ϕ_{ijkl} and ϕ_{hjkm}, there are 8 new variables and 4 new trigonometric terms and 6 linear terms.

$$w_{1,hijklm} = w_{2,hijklm}$$

$$w_{5,hijklm} = w_{6,hijklm}$$

where

$$w_{1,hijklm} = \cos(w_{3,hijklm})$$

$$w_{2,hijklm} = \cos(w_{4,hijklm})$$

$$w_{3,hijklm} = \phi_{ijkl} - \phi_{ijkm}$$

$$w_{4,hijklm} = \phi_{hjkl} - \phi_{hjkm}$$

$$w_{5,hijklm} = \cos(w_{7,hijklm})$$

$$w_{6,hijklm} = \cos(w_{8,hijklm})$$

$$w_{7,hijklm} = \phi_{hjkm} - \phi_{ijkm}$$

$$w_{8,hijklm} = \phi_{hjkl} - \phi_{ijkl}$$

The trigonometric terms are bounded by two nonlinear inequality constraints each. There are 24 combinations of ϕ_{ijkl} and ϕ_{hjkm} and that makes 192 additional variables, 192 nonlinear inequality constraints and 144 linear equality constraints.

As the constraints for the *dependent* angles and nonbonded distances sometimes share the same symbolic reformulation terms, these terms should be defined just once. Therefore for angles which are part of the nonbonded distance term for 4 atoms chains, the following terms have been reformulated:

$$\omega_{1,jkl} = \cos(\theta_{jkl})$$

$$\omega_{2,jkl} = \sin(\theta_{jkl})$$

$$\omega_{3,jkl} = b_{jk}b_{kl}$$

$$\omega_{4,jkl} = \omega_{1,jkl}\omega_{3,jkl}$$

That makes 40 additional variables, 40 nonlinear inequality constraints and 80 linear inequality constraints in addition to the torsional angle constraints.

For each *dependent* angle which does not participate in the nonbonded distance term, θ_{lkm}, the constraint is reformulated as Equation 6.42:

$$\varpi_{1,lkm} = \varpi_{2,lkm} + \varpi_{3,lkm}$$

where

$$\varpi_{1,lkm} = \cos(\theta_{lkm})$$

$$\varpi_{2,lkm} = \omega_{1,jkl}\omega_{1,jkm}$$

$$\varpi_{3,lkm} = w_{1,hijklm}\varpi_{4,lkm}$$

$$\varpi_{4,lkm} = \omega_{2,jkl}\omega_{2,jkm}$$

In this case, the number of additional variables is 20 with 10 nonlinear inequality constraints, 60 linear inequality constraints and 5 linear equality constraints.

The quadratic terms for the bonds lengths used in the nonbonded distances constraints are reformulated as:

$$\eta_{1,ij} = b_{ij}^2$$

As there are 9 bonds, that makes 9 new variables and 9 nonlinear inequality constraints and 9 linear inequality constraints.

As for the nonbonded distance having 4 atom chain, the constraint is reformulated as:

$$\vartheta_{1,ijkl} = \eta_{1,ij} + \eta_{1,jk} + \eta_{1,kl} - 2\omega_{4,ijk} - 2\omega_{4,jkl} - 2\vartheta_{9,ijkl} + 2\vartheta_{8,ijkl}$$

where

$$\eta_{1,jk} = b_{jk}^2$$

$$\eta_{1,kl} = b_{kl}^2$$

$$\omega_{4,ijk} = \omega_{1,ijk}\omega_{3,ijk}$$

$$\vartheta_{1,ijkl} = r_{il}$$

$$\vartheta_{2,ijkl} = b_{ij}b_{kl}$$

$$\vartheta_{3,ijkl} = \cos(\phi_{ijkl})$$

$$\vartheta_{4,ijkl} = \sin(\phi_{ijkl})$$

$$\vartheta_{5,ijkl} = \omega_{1,ijk}\omega_{1,jkl}$$

$$\vartheta_{6,ijkl} = \omega_{2,ijk}\omega_{2,jkl}$$

$$\vartheta_{7,ijkl} = \vartheta_{6,ijkl}\vartheta_{3,ijkl}$$

$$\vartheta_{8,ijkl} = \vartheta_{5,ijkl}\vartheta_{2,ijkl}$$

$$\vartheta_{9,ijkl} = \vartheta_{7,ijkl}\vartheta_{2,ijkl}$$

From 15 4-atom chain nonbonded distances, there are 135 new variables, 75 nonlinear inequality constraints, 375 linear equality constraints and 15 linear equality constraints. As for the 6 5-atom nonbonded distances, there are 153 new variables with 6 nonlinear inequality constraints, 390 linear inequality constraints and 6 linear constraints.

In total for the lower bounding optimisation problem, there are 565 variables with 1416 constraints. Two conformers were found, one with energy 0.5044 kcal/mol which has the torsion $\phi_{5-4-1-7}$ value of 1.036 rad and the global minimum potential energy conformer 0.4710 kcal/mol which has the torsion $\phi_{5-4-1-7}$ value of 3.180 rad.

6.7.2 *n*-Butane/ethylamine pair

The Z-matrix for the *n*-butane/ethylamine pair shown in Fig. 6.11 is as follows:

Atom Number	Atom Type	Bond Atom	Bond Length	Angle Atom	Bond Angle	Torsion Atom	Torsional Angle
1	*X*	0	0.000000	0	0.000000	0	0.000000
2	*X*	1	1.000000	0	0.000000	0	0.000000
3	*C*	2	1.000000	1	90.000000	0	0.000000
4	*C*	3	1.535393	2	90.000000	1	228.222667
5	*C*	3	1.532589	2	90.000000	4	247.520132
6	*H*	3	1.090504	4	109.955401	5	121.888845
7	*H*	3	1.090983	4	109.587650	6	116.437047
8	*C*	4	1.532835	3	112.541622	5	181.307267
9	*H*	4	1.090616	3	109.960386	8	122.132987
10	*H*	4	1.091019	3	109.544823	8	238.375932
11	*H*	8	1.090240	4	111.560538	3	59.958144
12	*H*	8	1.090526	4	110.839262	11	119.898634
13	*H*	8	1.090514	4	111.199223	11	239.550819
14	*H*	5	1.090269	3	111.339350	4	60.304044
15	*H*	5	1.090347	3	110.904731	14	119.844706
16	*H*	5	1.090470	3	111.240424	14	239.648020
17	*C*	3	3.816500	2	151.955467	1	324.784313
18	*C*	17	1.530337	3	172.982692	2	52.838495
19	*N*	17	1.449939	18	109.359699	3	181.361391
20	*H*	18	1.090072	17	111.823015	19	299.003647
21	*H*	18	1.090812	17	111.267904	20	120.922172
22	*H*	18	1.090366	17	110.636181	20	240.248747
23	*H*	17	1.091285	18	110.682262	19	121.301787
24	*H*	17	1.091303	18	110.043593	19	239.818638
25	*H*	19	1.011980	17	110.191857	18	181.507569
26	*H*	19	1.011047	17	110.242007	25	243.152111

The minimisation for pairs is simplified by the fact that the conformers are fixed. In

this case, the molecules are fixed at their global minimum potential energy conformers. The bond lengths, bond angles and torsional angles in each molecule have fixed values. The orientation of the molecules with respect to each other is dependent on *six* translational and rotational degrees of freedom i.e. the distance between atoms 3 and 17, the pseudo-angles θ_{2-3-17} and $\theta_{3-17-18}$, and the pseudo-torsions $\phi_{1-2-3-17}$, $\phi_{2-3-17-18}$ and $\phi_{3-18-17-19}$.

Pseudo Variables

For systems of molecular pairs, the orientation of the two molecules with respect to each other is determined from six variables: a nonbonded distance which defines the translational distance between the two central atoms in each molecule as defined in the Z-matrix, two angles which define the vertical rotation angles of the two molecules with each other and three torsions of which one defines the horizontal rotation angle and the other two specify the axial rotation of the molecules.

However in the determination of the nonbonded distances, in order to limit the number of atoms in the nonbonded distance chain to 4, *pseudo*-angles and torsions are used. In this case, the molecules are inflexible and the objective is to determine the global minimum potential energy orientation of the molecular pairs. A *pseudo*-angle is the angle between any two atoms in one molecule with an atom in the other molecule connected via the nonbonded distance specified in the Z-matrix. A *pseudo*-torsion is the torsion between two atoms in the first molecule with two atoms in the second also connected via the nonbonded distance specified in the Z-matrix. The aim is to simplify the convexification of the nonbonded distances constraints which will be explained in Section 6.5.

Pseudo-variables are used to determine bounds for the nonbonded distances. There are 140 nonbonded distances between the two molecules of which there is one between atoms 3 and 17. There are 13 between atom 17 and the other atoms in n-butane not including atom 3 and 9 between atom 3 and other atoms in the ethylamine molecule not including atom 17, which make 22 pseudo-angles. There are 117 pseudo-torsions which describe the torsional angle between an atom in n-butane 'connected' to atom 3 and an atom in ethy-

lamine 'connected' to atom 17 with respect to atoms 3-17. The objective is to simplify the calculation of the bounds for the nonbonded distances and also the derivation of the constraints.

Using Equation 6.9, the pseudo-angles may be calculated from bounds on the orientation variables. For example, in the case of the pseudo-angle θ_{8-3-17}, the bounds are dependent on the variables θ_{2-3-8}, θ_{2-3-17} and $\phi_{1-2-3-17}$. The angle θ_{2-3-8} is fixed while the bounds for angle θ_{2-3-17} are $[0, \pi]$ and for torsion $\phi_{1-2-3-17}$ are $[0, 2\pi]$.

As the range of values of bounds for angle θ_{2-3-17} are large, Equation 6.53 is used to determine if the bounds can be made tighter to calculate the bounds for pseudo-angle θ_{8-3-17}. The stable point is defined where $\tan(\theta_{2-3-17}) = \tan(\theta_{2-3-8})\cos(\phi_{8-2-3-17})$. As angle θ_{2-3-8} is fixed at 60^o, the maximum angle for θ_{8-3-17} is when the atom 17 is the antipodal of atom 8 when the difference in torsions between $\phi_{1-2-3-17}$ and $\phi_{1-2-3-8}$ is π. The minimum is when atom 17 is in the same direction as atom 8 with respect to atom 3. Therefore using interval arithmetic:

The maximum value of θ_{8-3-17} which satisfies Equation 6.54 is found when:

$$\begin{aligned}(\tan(\theta_{2-3-17}))^L &= \tan(60^o)\cos(180^o) \\ &= -1.732051 \\ \theta^U_{2-3-17} &= 120^o\end{aligned}$$

The minimum value of θ_{8-3-17} which satisfies Equation 6.55 is found when:

$$\begin{aligned}(\tan(\theta_{2-3-17}))^U &= \tan(60^o)\cos(0^o) \\ &= 1.732051 \\ \theta^L_{2-3-17} &= 60^o\end{aligned}$$

Therefore, the bounds for θ_{2-3-17} in the calculation of the bounds for pseudo-angle θ_{8-3-17} are $[60^o, 120^o]$. Using these new bounds to calculate bounds for the pseudo-angle

θ_{8-3-17}:

$$\begin{aligned}
\cos(\theta_{8-3-17}) &= \sin(\theta_{2-3-8})\sin(\theta_{2-3-17})\cos(\phi_{8-2-3-17}) + \cos(\theta_{2-3-8})\cos(\theta_{2-3-17}) \\
(\cos(\theta_{8-3-17}))^L &= \sin(\theta_{2-3-8})\sin(\theta^U_{2-3-17})\cos(\phi^U_{8-2-3-17}) + \cos(\theta_{2-3-8})\cos(\theta^U_{2-3-17}) \\
&= \sin(60^o)\sin(120^o)\cos(180^o) + \cos(60^o)\cos(120^o) \\
&= -1.00 \\
\theta^U_{8-3-17} &= 180^o \\
(\cos(\theta_{8-3-17}))^U &= \sin(\theta_{2-3-8})\sin(\theta^L_{2-3-17})\cos(\phi^L_{8-2-3-17}) + \cos(\theta_{2-3-8})\cos(\theta^L_{2-3-17}) \\
&= \sin(60^o)\sin(60^o)\cos(0^o) + \cos(60^o)\cos(60^o) \\
&= 1.00 \\
\theta^L_{8-3-17} &= 0^o
\end{aligned}$$

As in interval arithmetic, each variable is taken independently, therefore the value of the cosine of the angle may sometimes exceed the domain [-1,1]. If the value of the cosine is more than 1, it is taken as 1 and similarly if the value of the cosine is less than -1, it is taken as -1.

Equation 6.11 can also be used to determine bounds for pseudo-torsions. To calculate the pseudo-torsion $\phi_{8-3-17-24}$:

$$\phi_{8-3-17-24} = \phi_{2-3-17-18} - \phi_{2-3-17-8} + \phi_{18-3-17-24}$$

As the bounds for ϕ vary between $[0^o, 360^o]$, in order to distinguish between the location of atom 8 with respect to arc 2-3-17, an xyz coordinate using the orientation torsion $\phi_{1-2-3-17}$ has to be expressed with the atom 3 as the origin and the atoms $3-2$ axis showing the direction of the y-axis and atom 1 at the x-axis. The z-axis is then determined using the Right-Hand Rule as shown in Fig. 6.12.

The torsion $\phi_{1-2-3-8}$ is calculated and compared with the bounds for the orientation torsion $\phi_{1-2-3-17}$. The sphere is then divided into two hemispheres along the arc 2-3-8.

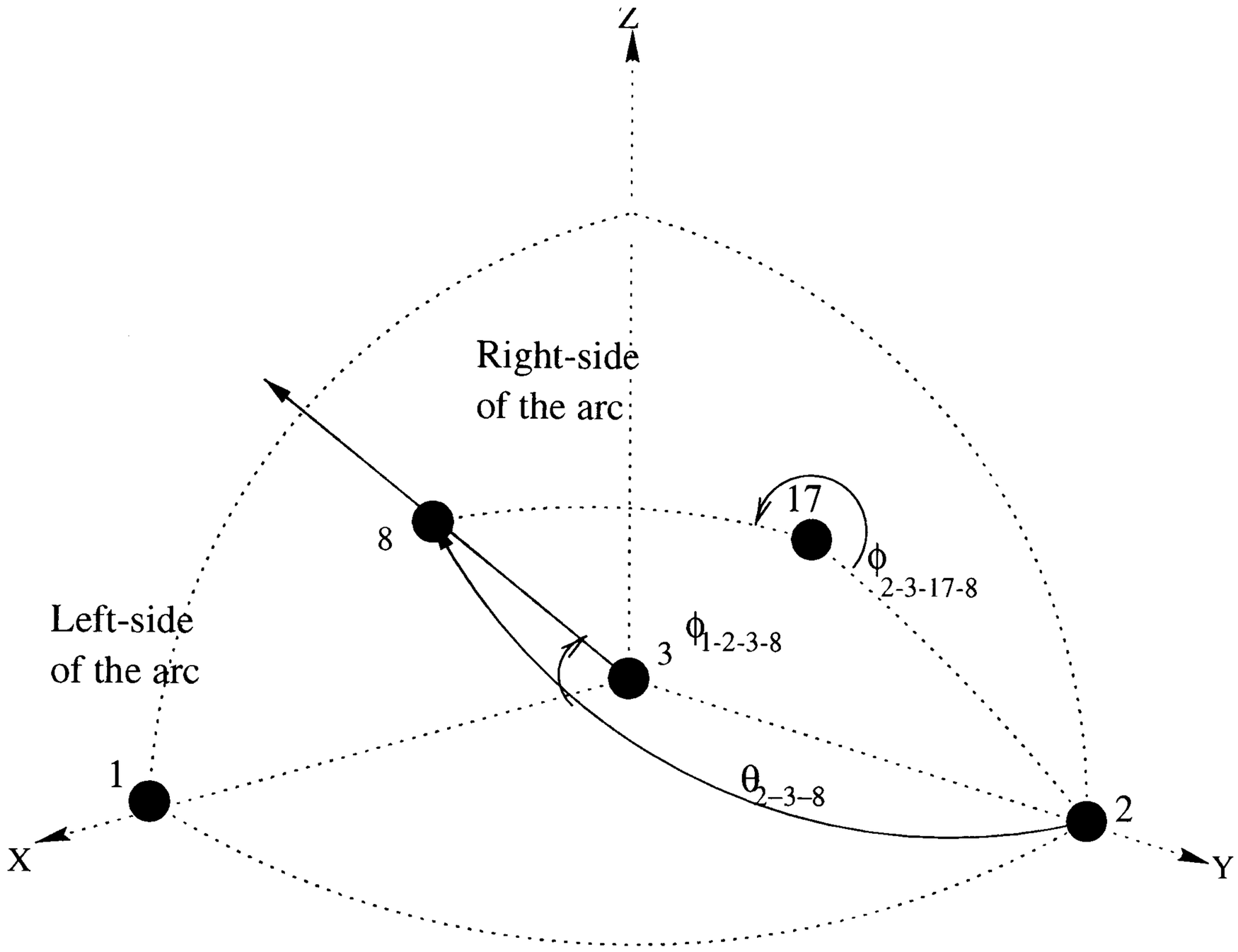


Figure 6.12: Figure showing relationship for the *pseudo* torsion

If the bounds of the orientation torsion lie only to the left of the angle arc 2-3-8, the bounds for $\phi_{2-3-17-8}$ are between $[0^o, 180^o]$. However if the bounds are to the right of the arc, the bounds for $\phi_{2-3-17-8}$ are $[180^o, 360^o]$.

The bounds for the orientation angle θ_{2-3-17} are then compared to the angle θ_{2-3-8}. If the torsion $\phi_{1-2-3-8}$ is in the bounds for the orientation torsion and the angle θ_{2-3-8} is less than the upper bound for θ_{2-3-17}, the bounds for $\phi_{2-3-17-8}$ is $[0^o, 360^o]$. Otherwise, if the upper bound for θ_{2-3-17} is less than θ_{2-3-8}, the upper bound for $\phi_{2-3-17-8}$ is found at the upper bound of the orientation angle and the furthest torsion on the right side. The lower bound is found at the upper bound of the orientation angle and the furthest torsion on the left side.

If the torsion bounds lie only to the left side with angle θ_{2-3-8} within the bounds for the orientation angle, the lower bound for $\phi_{2-3-17-8}$ is found at the upper bound of the orientation angle and the nearest torsion. The upper bound is found at the lower bound

of the orientation angle and the nearest torsion. Otherwise if the angle θ_{2-3-8} is less than the bounds, the upper bound is found at the lower bound of the orientation angle and the furthest torsion. Similarly if the angle θ_{2-3-8} is more than the bounds, the lower bound is found at the upper bound of the orientation torsion and the furthest torsion.

Conversely, if the torsion lies only to the right side of the arc, the upper bound for $\phi_{2-3-17-8}$ is found at the upper bound of the orientation angle and the nearest torsion. The lower bound is found at the lower bound of the orientation angle and the nearest torsion. If the angle θ_{2-3-8} is less than the bounds, the lower bound for $\phi_{2-3-17-8}$ is found at the lower bound of the orientation angle and the furthest torsion. Otherwise if the angle is more than the bounds, the upper bound is found at the upper bound of the orientation angle and the furthest torsion.

Therefore, if the angle θ_{2-3-8} is 60^o, the torsion $\phi_{1-2-3-8}$ is 280^o and the bounds for angle θ_{2-3-17} are $[90^o, 135^o]$ and for torsion $\phi_{1-2-3-17}$ are $[180^o, 270^o]$, the equation used to calculate $\phi_{2-3-17-8}$ are:

$$\cos(\phi_{2-3-17-8}) = \frac{\cos(\theta_{2-3-8}) - \cos(\theta_{8-3-17})\cos(\theta_{2-3-17})}{\sin(\theta_{8-3-17})\sin(\theta_{2-3-17})}$$

As the bounds for the torsion $\phi_{1-2-3-17}$ lie both to the right of $\phi_{1-2-3-8}$ and the bounds for the angle θ_{2-3-17} are larger than the angle θ_{2-3-8}, the lower bound for $\phi_{2-3-17-8}$ is at $\theta_{2-3-17} = 90^o$ and $\phi_{1-2-3-17} = 180^o$ and the upper bound for $\phi_{2-3-17-8}$ is at $\theta_{2-3-17} = 135^o$ and $\phi_{1-2-3-17} = 270^o$:

$$\cos(\phi^L_{2-3-17-8}) = \frac{\cos(60^o) - \cos(90^o)\cos(98.65^o)}{\sin(90^o)\sin(98.65^o)}$$

$$\phi^L_{2-3-17-8} = 59.62^o$$

$$\cos(\phi^L_{2-3-17-8}) = \frac{\cos(60^o) - \cos(135^o)\cos(75.55^o)}{\sin(135^o)\sin(75.55^o)}$$

$$\phi^L_{2-3-17-8} = 8.92^o$$

where the values for angle θ_{8-3-17} were calculated using Equation 6.9 and the different torsion angles as specified.

The bounds on the pseudo-angles and pseudo-torsions are then used to determine the bounds for the nonbonded distance using interval arithmetic and also in the constraints in the relationships of these distances with the pseudo-variables.

Constraints on the *Pseudo* Variables

The constraints on the pseudo-variables are similarly derived as in the ethylamine molecule example. There are 2640 additional variables from the pseudo-torsional angles constraints. Constraints for the pseudo-angles are slightly different to the previous case study as the bond lengths, bond angles and torsional angles in each molecule are fixed. The new variables for the pseudo-angles based on symbolic reformulation, where atom j is in the first molecule and atom k in the second molecule, are:

$$\omega_{1,ijk} = \cos(\theta_{ijk})$$
$$\omega_{2,ijk} = \sin(\theta_{ijk})$$
$$\omega_{3,ijk} = \omega_{1,ijk} b_{jk}$$

So for angle θ_{hjk}, the reformulation is:

$$\omega_{1,hjk} = \cos(\theta_{hjk})$$
$$\omega_{2,hjk} = \sin(\theta_{hjk})$$
$$\omega_{3,hjk} = \omega_{1,hjk} b_{jk}$$

That makes 66 additional variables from 22 pseudo-angles.

As the angles in each molecule are fixed, the constraints become:

$$\cos(\theta_{hji}) = \varpi_1 + \varpi_2$$

where

$$\varpi_{1,hji} = \omega_{1,hjk}\omega_{1,ijk}$$
$$\varpi_{2,hji} = w_{3,hjikl}\varpi_{3,hji}$$
$$\varpi_{3,hji} = \omega_{2,hjk}\omega_{5,ijk}$$

There are 110 of these linear equality constraints so that makes 330 additional variables.

The only quadratic term used for the bonds in the nonbonded distances constraints is:

$$\eta_{1,jk} = b_{jk}^2$$

where $j = 3$ and $k = 17$.

For each 3-atom chain, the only new variable is:

$$\vartheta_{1,ijk} = r_{ik}^2$$

and the linear equality constraint for the 3-atom chain is

$$\vartheta_{1,ijk} = b_{ij}^2 + \eta_{1,jk} - 2b_{ij}\omega_{3,ijk}$$

So that makes 22 new variables.

As for the nonbonded distance having 4 atom chain, r_{ijkl}, the constraint is reformulated as:

$$\vartheta_{1,ijkl} = \eta_{1,il} + b_{ij}^2 + b_{jk}^2 - 2b_{ij}\omega_{3,ijk} - 2b_{kl}\omega_{3,jkl} - 2b_{ij}b_{kl}\vartheta_{5,ijkl} + 2b_{ij}b_{kl}\vartheta_{3,ijkl}$$

where

$$\vartheta_{1,ijkl} = r_{il}^2$$
$$\vartheta_{2,ijkl} = \cos(\phi_{ijkl})$$
$$\vartheta_{3,ijkl} = \omega_{1,ijk}\omega_{1,jkl}$$
$$\vartheta_{4,ijkl} = \omega_{2,ijk}\omega_{2,jkl}$$
$$\vartheta_{5,ijkl} = \vartheta_{4,ijkl}\vartheta_{2,ijkl}$$

As there are 117 nonbonded distances with 4-atom chains, that makes 585 more variables. In total, there are 3923 variables in the optimisation algorithm and 8751 constraints.

The algorithm currently only offers a directed search and does not actually converge

as the lower bounding function only approaches the objective function when the bounds for the variables are small. However, after 1000 iterations the lowest potential energy minimum was found to be 0.8184 kcal/mol. The average time of each iteration is 443 CPU seconds. As the global minimum for the n-butane molecule was found to be 2.0724 kcal/mol, the interaction energy for the pair is 0.8184 - 2.0724 - 0.4710 = -1.7250 kcal/mol.

Although the algorithm was terminated after 1000 iterations, the global minimum was actually found after 32 iterations due to the directed search nature of the algorithm. The subsequent iterations were needed in order for the algorithm to converge i.e. the difference between the lower bound and the global minimum to be within ϵ tolerance. The CPU time is high as the local minimisation of the lower bound currently takes a long time due to the large number of constraints. Due to the dependencies between the molecular variables, some of the constraints are probably unnecessary. Further work on analysing the effect of the various constraints will make the algorithm more robust and efficient.

6.7.3 Ethylamine/ethylamine pair

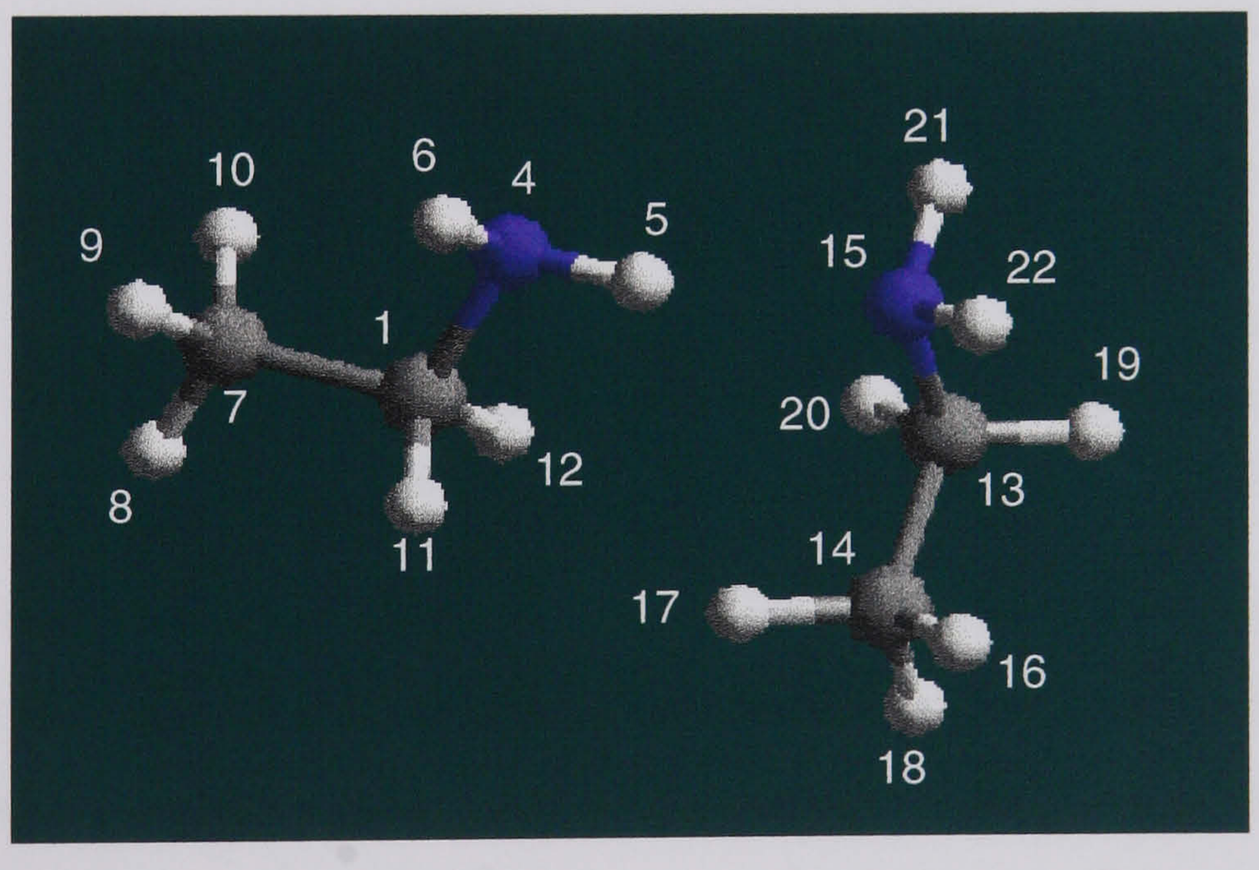


Figure 6.13: Ethylamine/ ethylamine molecule showing the atom numbers

The Z-matrix for the ethylamine/ethylamine pair shown in Fig. 6.13:

Atom Number	Atom Type	Bond Atom	Bond Length	Angle Atom	Bond Angle	Torsion Atom	Torsional Angle
1	*C*	0	0.000000	0	0.000000	0	0.000000
2	*X*	1	1.000000	0	0.000000	0	0.000000
3	*X*	2	1.000000	1	90.000000	0	0.000000
4	*N*	1	1.450767	2	90.000000	3	270.000000
5	*H*	4	1.012825	1	110.325606	2	240.000000
6	*H*	4	1.009375	1	110.495353	5	242.770316
7	*C*	1	1.530374	4	109.336800	5	182.220709
8	*H*	7	1.090448	1	110.648064	4	179.126874
9	*H*	7	1.090395	1	111.739479	8	119.738898
10	*H*	7	1.090878	1	111.294138	8	240.685122
11	*H*	1	1.091305	4	109.971507	7	238.327549
12	*H*	1	1.091357	4	109.404248	7	120.638338
13	*C*	1	3.833844	2	106.526288	3	206.624845
14	*C*	13	1.530422	1	80.983921	2	213.647624
15	*N*	13	1.450155	14	109.335145	1	75.852201
16	*H*	14	1.090398	13	111.872068	15	299.274266
17	*H*	14	1.090862	13	111.127528	16	120.803802
18	*H*	14	1.090422	13	110.678403	16	239.963993
19	*H*	13	1.091290	14	110.537750	15	121.074640
20	*H*	13	1.091363	14	110.198070	15	239.595996
21	*H*	15	1.010994	13	110.171040	14	181.544799
22	*H*	15	1.011039	13	110.418611	21	243.068570

As in the previous case study, the molecules are fixed at their global minimum conformer and only the positions of the molecules with respect to each other are optimised. The six degrees of freedom are the distance between atoms 1 and 13, the pseudo-angles θ_{2-1-13} and $\theta_{1-13-14}$ and the pseudo-torsions $\phi_{3-2-1-13}$, $\phi_{2-1-13-14}$ and $\phi_{1-14-13-15}$.

There are 100 nonbonded interactions between the two molecules, of which 18 have 3-atom chains and 81 have 4-atom chains. Therefore, that makes 18 pseudo-angles and 81 pseudo-torsions. Additional variables and constraints from symbolic reformulation of the *dependent* variables constraints are 2454 and 5917 respectively.

The global minimum found for the ethylamine/ ethylamine pair after 1000 iterations is -4.0148 kcal/mol. The global minimum was found after 14 iterations. The average time for each iteration is 373 CPU seconds. The interaction energy for the pair is -4.0148 -0.4710 $-0.4710 = -4.9568$ kcal/mol. The interaction is strong due to hydrogen bonding.

6.8 Conclusions

A global optimisation algorithm for molecular systems has been developed based on a branch-and-bound framework. The method offers a directed search technique towards the global minimum by constructing a tight lower bounding function that is consistently lower that the objective function.

Tight convex underestimators have been developed for the torsional angles terms and the nonbonded interactions terms. The feasible region defined by the bounds for the variables and the constraints between the *dependent* and *independent* variables have been convexified using symbolic reformulation and interval arithmetic.

This technique concentrates the search of low-lying potential energy minima and increases the chances for the global minimum to be found. These low energy minima allow for the identification of specific interactions between the molecular pairs which can then be used to build the phase equilibria model.

Chapter 7

Conclusions and Future Directions

7.1 Introduction

To conclude, the key objectives of this research have been:

- Development of a reliable and low-cost approach to predicting solubility especially for pharmaceutical compounds
- Development of a prediction technique for polymorph stability as a function of the solvent and temperature
- Automation of the property prediction method using global optimisation

In this chapter, the work of the previous chapters is highlighted in view of these objectives, and key contributions of the research are outlined. Further, future directions needed for further development of the research are identified. Finally, a discussion of how the automated property prediction methodology might be integrated within a solvent design framework is presented, where molecular groups are defined as building blocks in designing a solvent molecule. The *design* problem then is concerned with determining a set of structural blocks which combine to create the target molecule that best meets the desired properties specfications and process constraints.

7.2 Conclusions

A systematic methodology of predicting phase equilibria is introduced in Chapter 3. The methodology does not depend on the availability of experimental data for the mixture of interest and as a result, may be used for a wide variety of systems. It is based on the UNIQUAC activity coefficient model and uses molecular mechanics to determine the interaction parameters. These parameters are defined as a function of the differences between the solute-solvent interaction energies and the solvent-solvent and solute-solute interaction energies respectively. It has been shown that calculations based on using pairs of molecules is insufficient in determining the relative interaction energies. A better approach is to build a model of the first solvation shell around the solute molecule. Polar and nonpolar solutes are systematically positioned in a cluster before a local minimisation is performed. The binary interaction energies are then calculated from the configuration of the minimised structure.

Chapter 4 continues with the results obtained when applying the methodology to monofunctional solutes in solvents of varying polarity. Comparison is made between the predictions obtained using the methodology with experimental data, D-UNIFAC and COSMO-RS predictions. The latter two are quantitatively derived models based on obtaining parameters from experimental data. As only the first solvation shell is accounted for, the methodology naturally overemphasises slightly the specific interactions. Therefore, if the solute and the solvent molecules are similar, the method performs well as the errors cancel out. Otherwise, the solubility is slightly over- or underestimated. The methodology is quite comparable to D-UNIFAC or COSMO-RS. However the weakness of these methods is that both contain parameters that discount proximity effects. The chapter then follows on with results of the bifunctional solutes that again prove that the methodology may be extended to more complex molecules. The methodology can also be used to determine qualitatively the trends in polymorph stability for different solvents at a range of temperatures. However, as there were no experimental data available, no quantitative conclusion can be made.

The methodology until this stage is conducted manually. Therefore, in order to remove

the human aspect, it was decided to automate the step-by-step technique of building a solvation shell. First the number of solvent molecules making up the solvation shell has to be determined. Then specific interactions between the solute and the solvent molecules have to be identified. This is achieved using the global optimisation algorithm proposed in Chapter 6. Finally the subsequent solvent molecules should be filled in systematically using a space-filling technique.

Chapter 6 starts with the formulation of the optimisation problem to find the global minimum potential energy for a molecular system. The problem is defined using $3N$-6 independent variables which are given bounds where N is the number of atoms in the molecular system. The rest of the variables are defined as *dependent* variables. *Pseudo*-variables are also introduced to simplify the calculation of the bounds for the nonbonded distances. Bounds on these variables are determined using interval arithmetic. In the optimisation algorithm, nonconvex terms have to be convexified. Convex underestimators for the torsional and the nonbonded distances potential energy terms are constructed. Symbolic reformulation (Smith and Pantelides, 1998) is used to convexify the nonconvex constraints. Case studies were shown using an ethylamine molecule and also for pairs of n-butane/ ethylamine and ethylamine/ ethylamine. The algorithm is based on the branch-and-bound framework. This search method gives a directed search for the global minimum unlike stochastic methods which conduct the search randomly. The optimisation algorithm then gives a ranking of the potential energies of the molecular system beginning with the global minimum.

To summarise, a phase equilibria prediction methodology has been developed which derives interaction parameters from Molecular Mechanics force fields. It might also be used to predict polymorph stability in various solvents for a range of temperatures. A deterministic global optimisation algorithm is proposed to find the global minimum potential energy for molecular systems.

7.3 Future Directions

Currently, the automation of the phase equilibria prediction method is not finalised. Work on analysing the constraints of the algorithm has to be done to make it more robust and efficient. It has also not been tested on more complex molecules although it shows promising results for mono- and bifunctional benzene derivatives. Key directions for future research have been identified as analysing the algorithm, developing a space-filling model for automatic cluster generation, developing techniques for handling dissimilar solute and solvent molecules and testing the methodology for more complex molecules.

7.3.1 Space-filling model

Once the specific interactions have been identified using the global optimisation algorithm, the remaining solvent molecules still have to be systematically placed to model the first solvation shell. This space-filling task is important due to the many local minima available for molecular systems and will enable consistent results to be achieved. However, although easily done manually as the size and shape of the molecule can be visualised together with the locations of hydrogen-bonding interaction sites, it is not easily done computationally.

However, several guidelines for formulating the space-filling model utilising translation and rotation procedures:

- Determine the principal axis along which the torque of the molecule is nil
- Determine the possible hydrogen-bonding sites along the principal axis and the length of the site from the axis
- Rotate the neighbouring molecule 180^o placing hydrogen-bond donor sites opposite hydrogen-bond acceptor sites and vice versa if possible. If not, then place the neighbouring molecules where the Coulombic and van der Waals interactions are strongest by the side of the molecule without the molecules overlapping
- Build three layers of the solvent molecules around the solute placing the the middle molecule of the upper and lower layers where the Coulombic and van der Waals

interactions between these molecules are strongest with the solute molecule.

- Perform a local minimisation

The locations of the hydrogen-bonding sites are found using the global optimisation algorithm for potential energy of molecular systems.

7.3.2 Handling dissimilar molecules

Simple statistical tools can be utilised to determine the proportion by which the specific interactions are overestimated. This would involve performing Monte Carlo simulations for a number of systems both polar and nonpolar and then comparing the interaction energies with those obtained using the model of the first solvaton shell.

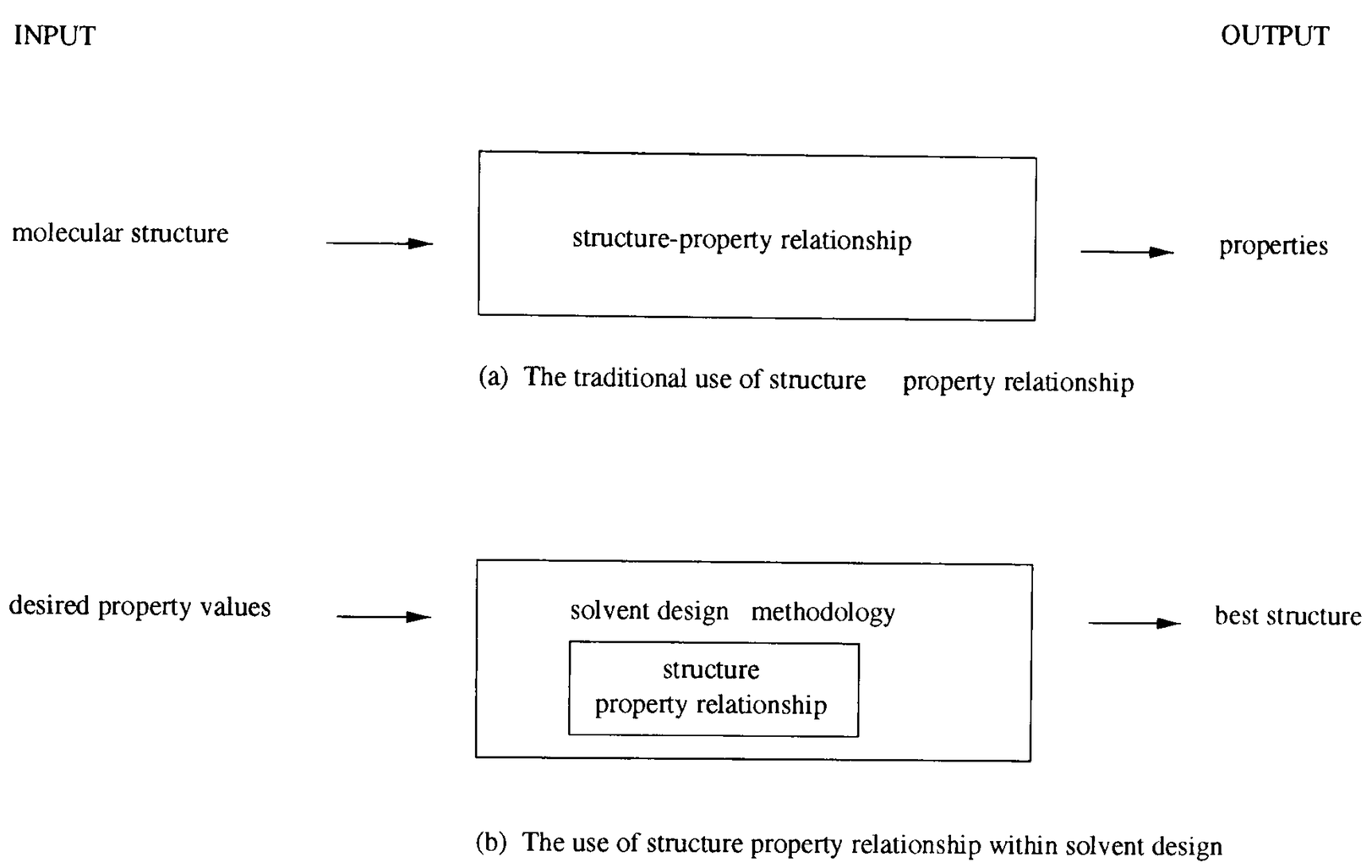
The BOSS simulation package using the OPLS force field runs NPT Monte Carlo simulations. Runs using n-hexane and water have been conducted. The best method of obtaining the interaction energies could be by giving a correction to the Boltzmann average equation.

7.4 Integration of property prediction methodology within a solvent design framework

Traditional molecular design methods involve synthesising and testing thousands of different types of compounds for their thermophysical and chemical properties. The search for new molecular products with optimal properties for a particular process is the driving force for many chemical companies and encompasses a wide range of products (e.g. refrigerants, solvents and extractants, lubricants, ceramics etc.).

In the past two decades, computer-aided molecular design (CAMD) has attracted widespread attention because of its ability to design compounds or mixtures of compounds with specific properties. The foundation of a good CAMD technique is a reliable method of predicting physico-chemical properties within small error bounds based on structure-property relationships. Therefore, rather than specifying a molecular structure and identifying the

corresponding properties, one aims to tackle the problem of finding the structure which best matches a given set of properties, or which minimises or maximises a given property (Maranas, 1996) as illustrated in Fig. 7.1. CAMD methodologies are divided into



(a) The traditional use of structure property relationship

(b) The use of structure property relationship within solvent design

Figure 7.1: Traditional and new use of structure-property relationships

generate-and-test and mathematical optimisation techniques. Both methods are similar in that a molecule is defined as a set of structural groups of which each has a contribution towards the target property. However, the former is conducted using heuristics while the latter uses mathematical programming tools.

Mathematical optimisation techniques are able to encompass all feasible molecular structures implicitly within the formulation (Knight and McRae, 1991; Odele and Macchietto, 1993; Duvedi and Achenie, 1996). Optimisation methods for molecular design require the formulation of a mixed-integer nonlinear problem (MINLP). The molecular design problem is formulated as an optimisation problem with the aim of designing the best molecular structure that would meet a pre-specified performance index subject to certain constraints relating to process and product specifications, environmental impact and safety. In order to facilitate the optimisation methodology, constraints are also specified so only chemically feasible compounds are generated. The molecular structures formed must satisfy the following criteria: no free attachments and meet the restrictions according to their

respective structural groups, i.e. restrictions on bonding between groups depending on the number of free attachments available and exhibit no proximity effects.

A property optimisation problem can be formulated as:

$$\min MP(p_j(\mathbf{n})) \tag{7.1}$$

subject to

$$p_j^L \leq p_j(\mathbf{n}) \leq p_j^U, \forall j \in P$$
$$n_i \in [n_i^L, n_i^U], \forall i \in G$$

where p_j is the jth property type with its upper and lower limits, p_j^U and p_j^L, and n_i is the number of times the ith molecular group participates in the structure. P is the set of properties and G is the set of molecular groups. The discrete variables, $\mathbf{n}$, are the numbers of atom groups in each molecule and the continuous variables consist of properties, mole fractions etc. Other constraints may include structural feasibility, process equations and design specifications. For example, the structural feasibility constraints are given by the octet rule (Odele and Macchietto, 1993):

$$\sum_j (2 - v_j) n_j = 2m \tag{7.2}$$

where n_j, v_j are the number and valency, respectively, of groups of type $\mathbf{j}$ and $\mathbf{m} = 1, 0$ or -1 for acyclic, monocyclic and bicyclic molecules respectively.

As our method needs a knowledge of molecular connectivity to determine the phase equilibria, it cannot be framed mathematically where each molecular group has a contribution towards the molecular property. However, full connectivity allows the differentiation between isomers, thus also taking into account steric hindrance and proximity effects. Therefore, the property optimisation problem could be formulated to minimise interaction energies for different groups. Once those strongly interaction groups are identified, these groups can be formed into solvent molecules based on the multi-level computer-

aided molecular design approach proposed by Harper et al. (1999). The first level defines the molecular fragments which Level 2 use to generate molecular structures. The molecular connectivity is then clearly defined in Level 3 and Level 4 obtains the 3-dimensional model of the molecules. Optimal solvents at this stage are then tested using our prediction methodology. The top few best solvents are verified experimentally.

The ability to design a solvent molecule with pre-specified physico-chemical properties reliably and accurately will enable better process design that will decrease costs and reduce environmental harm. Better understanding of interactions at the molecular level and how these interactions are manifested in macro-level properties will allow for better material and product design.

Bibliography

[1] D. S. Abrams and J. M. Prausnitz. Statistical Thermodynamics of Liquid Mixtures: A New Expression for the Excess Gibbs Energy of Partly or Completely Miscible Systems. *AIChE Journal*, 21:116–128, 1975.

[2] C. S. Adjiman, S. Dallwig, C. A. Floudas, and A. Neumaier. A Global Optimization Method, αBB, for General Twice-Differentiable Constrained NLPs – I. Theoretical Advances. *Computers and Chemical Engineering*, 22(9):1137–1158, 1998a.

[3] C. S. Adjiman, S. Dallwig, C. A. Floudas, and A. Neumaier. A Global Optimization Method, αBB, for General Twice-Differentiable Constrained NLPs – II. Implementation and Computational Results. *Computers and Chemical Engineering*, 22(9):1159–1179, 1998b.

[4] C. S. Adjiman, C. A. Schweiger, and C. A. Floudas. *Nonlinear and Mixed-Integer Optimization in Chemical Process Network Systems*, volume 40 of *DIMACs Series in Discrete Mathematics and Theoretical computer Science*, pages 429–452. American Mathematical Society, 1998.

[5] F. A. Al-Khayyal and J. E. Falk. Jointly Constrained Biconvex Programming. *Mathematics of Operations Research*, 8(2):273–286, 1983.

[6] N.L. Allinger. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *Journal of the American Chemical Society*. 99:8127–8134, 1977.

[7] N.L. Allinger, Y. H. Yuh, and J.-H. Lu. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. *Journal of the American Chemical Society*, 111:8551–8566, 1989.

[8] B.R. Brooks, R.E. Bruccoleri, B.D. Olafsson, D.J. States, S. Swaminathan, and M. Karplus. Charmm: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[9] S. Budavari, M. O'Neil, A. Smith, P. Heckelman, and J. Obenchain (Eds). *The Merck Index, 12th edition on CD-ROM, ver. 12:3.* CRC Press, Boca Raton, Florida, 2000.

[10] R. Bukowski, K. Szalewicz, and C. F. Chabalowski. *Ab Initio* Interaction Potentials for Simulations of Dimethylnitramine Solutions in Supercritical Carbon Dioxide with Cosolvents. *Journal of Physical Chemistry A*, 103(36):7322–7340, 1999.

[11] F. Cajori. *An Introduction to the Modern Theory of Equations.* MacMillan Company, New York, 1904.

[12] L. Constantinou and R. Gani. A New Group Contribution Method for Estimating Properties of Pure Components. *AIChE Journal*, 40:1697–1710, 1994.

[13] J. J. de Pablo, M. Laso, and U. W. Suter. Simulation of Polyethylene Above and Below the Melting-Point. *Journal of Chemical Physics*, 96(3):2395–2403, 1992.

[14] E. de Vos Burchart. *Studies on Zeolites: Molecular Mechanics, Framework Stability, and Crystal Growth.* Ph.D. Thesis, Technische Universiteit Delft, 1992.

[15] K. A. Dill, A. T. Phillips, and J. B. Rosen. *Molecular Structure Prediction by Global Optimisation*, pages 217–234. Developments in Global Optimisation. Kluwer Academic Publishers, Netherlands, 1997.

[16] E. M. Duffy and W. L. Jorgensen. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol and Water. *Journal of the American Chemical Society*, 122:2878–2888, 2000.

[17] A. P. Duvedi and L. E. K. Achenie. Designing Environmentally Safe Refrigerants using Mathematical Programming. *Chemical Engineering Science*, 51(15):3727 3739, 1996.

[18] M. Fermeglia and S. Pricl. Prediction of Phase Equilibria for Binary Mixtures by Molecular Modeling. *AIChE Journal*, 47(10):2371–2382, 2001.

[19] V. Flemr. A Note on Excess Gibbs Energy Equations based on Local Composition Concept. *Collection Czechoslov. Chem. Commun.*, 41:3347–3349, 1976.

[20] Aa. Fredenslund, R. L. Jones, and J. M. Prausnitz. Group Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE Journal*, 21:1086–1099, 1975.

[21] J. Gasteiger and M. Marsili. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron*, 36:3219–3288, 1980.

[22] J. Gmehling, J. Li, and M. Schiller. A Modified UNIFAC Model. 2. Present Parameter Matrix and Results for Different Thermodynamic Properties. *Industrial and Engineering Chemistry Research*, 32:178–193, 1993.

[23] J. G. Gmehling, T. F. Anderson, and J. M. Prausnitz. Solid-Liquid Equilibria using UNIFAC. *Industrial and Engineering Chemistry Fundamentals*, 17(4):269–273, 1978.

[24] R. A. Granberg and A. C. Rasmuson. Solubility of Paracetamol in Pure Solvents. *Journal of Chemical Engineering Data*, 44:1391–1395, 1999.

[25] K. E. Gubbins and N. E. Quirke. *Molecular Simulation and Industrial Applications: Methods, Examples and Prospects*, volume 1. Gordon and Breach, Amsterdam, 1996.

[26] E. A. Guggenheim. *Mixtures.* Clarendon Press, Oxford, 1952.

[27] R. B. Gupta, R. Kumar, and G. V. Betageri. Phase Behaviour of Mixtures Containing Antibiotics. Chloramphenicol Partitioning. *Industrial and Engineering Chemistry Research*, 36:3954–3959, 1997.

[28] M. Haisa, S. Kashino, R. Kawai, and H. Maeda. The Monoclinic Form of p-Hydroxyacetanilide. *Acta Crystallographica*, B32:1283–1285, 1976.

[29] M. Haisa, S. Kashino, and H. Maeda. The Orthorhombic Form of p-Hydroxyacetanilide. *Acta Crystallographica*, B30:2510–2512, 1974.

[30] T. A. Halgren. Merck Molecular Force Field. 1. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490 519, 1996.

[31] T. A. Halgren. MMFF VII. Characterization of MMFF94, MMFF94s, and Other Widely Available Force Fields for Conformational Energies and for Intermolecular Interaction Energies and Geometries. *Journal of Computational Chemistry*, 20(7):730–748, 1999.

[32] E. R. Hansen. Global Optimization using Interval Analysis - The Multidimensional Case. *Numerical Mathematics*, 34:247–270, 1980.

[33] P. M. Harper, R. Gani, P. Kolar, and T. Ishikawa. Computer-Aided Molecular Design with Combined Molecular Modeling and Group Contribution. *Fluid Phase Equilibria*, 158-160:337–347, 1999.

[34] M. R. Hestenes. *Conjugate Direction Methods in Optimization.* Springer-Verlag, New York, 1980.

[35] J. H. Hildebrand and R. L. Scott. *Solubility of Nonelectrolytes.* Reinhold, New York, 3rd edition, 1950.

[36] M. Hloucha, A. K. Sum, and S. I. Sandler. Computer Simulation of Acetonitrile and Methanol with *Ab Initio*-Based Pair Potentials. *Journal of Chemical Physics*, 113(13):5401–5406, 2000.

[37] M. R. Hoare. Structure and Dynamics of Simple Microclusters. *Advanced Chemical Physics*, 40:49–135, 1979.

[38] H.H. Hooper, S. Michel, and J.M. Prausnitz. Correlation of Liquid-Liquid Equilibria for Some Water-Organic Liquid Systems in the Region 20-250 ^oC. *Industrial and Engineering Chemistry Research*, 27:2182–2187, 1988.

[39] R. Horst and H. Tuy. *Global Optimization : Deterministic Approaches.* Springer Verlag, 3rd edition, 1996.

[40] F. Jensen. *Introduction to Computational Chemistry.* Wiley, 1999.

[41] B. Jeziorski, R. Moszyński, and K. Szalewicz. Perturbation-Theory Approach to Intermolecular Potential-Energy Surfaces of van der Waals Complexes. *Chemical Reviews*, 94(7):1887–1930, 1994.

[42] K.G. Joback and R.C. Reid. Estimation of Pure Component Properties from Group Contributions. *Chemical Engineering Communications*, 57:233–243, 1987.

[43] S. O. Jónsdóttir, R. A. Klein, and K. Rasmussen. UNIQUAC Interaction Parameters for Alkane/Amine Systems determined by Molecular Mechanics. *Fluid Phase Equilibria*, 115:59 72, 1996.

[44] S. O. Jónsdóttir, K. Rasmussen, and Aa. Fredenslund. UNIQUAC Parameters Determined by Molecular Mechanics. *Fluid Phase Equilibria*, 100:121–138, 1994.

[45] S. O. Jónsdóttir, W. J. Welsh, K. Rasmussen, and R. A. Klein. The Critical Role of Force-fields in Property Prediction. *New Journal of Chemistry*, 23(2):153–163, 1999.

[46] W. L. Jorgensen and E. M. Duffy. Prediction of Drug Solubility from Monte Carlo Simulations. *Bioorganic and Medicinal Chemistry Letters*, 10:1155–1158, 2000.

[47] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[48] W. L. Jorgensen and J. Tirado-Rives. The OPLS Potential Functions for Peptides. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[49] W.L. Jorgensen, J.D. Madura, and C.J. Swenson. Optimized Intermolecular Potential Functions for Liquid Hydrocarbons. *Journal of the American Chemical Society*, 106:6638–6646, 1984.

[50] A. R. Katritzky, Y. Wang, S. Sild, and T. Tamm. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water-Air Partition Coefficients. *Journal of Chemical Information and Computer Science*, 38:720–725, 1998.

[51] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220:671–680, 1983.

[52] K. Kiyohara, T. Spyriouni, K. E. Gubbins, and A. Z. Panagiotopoulos. Thermodynamic Scaling Gibbs Ensemble Monte Carlo: A New Method for Determination of Phase Coexistence Properties of Fluids. *Molecular Physics.*, 89(4):965–974, 1996.

[53] A. Klamt and F. Eckert. COSMO-RS: a Novel and Efficient Method for the A Priori Prediction of Thermophysical Data of Liquids. *Fluid Phase Equilibria*, 172:43–72, 2000.

[54] A. Klamt, F. Eckert, M. Hornig, M. E. Beck, and T. Bürger. Prediction of Aqueous Solubility of Drugs and Pesticides with COSMO-RS. *Journal of Computational Chemistry*, 23:275–281, 2002.

[55] A. Klamt, V. Jonas, T. Bürger, and J. C. W. Lohrenz. Refinement and Parametrization of COSMO-RS. *Journal of Physical Chemistry A*, 102:5074–5085, 1998.

[56] J. L. Klepeis, C. A. Floudas, D. Morikis, and J. D. Lambris. Predicting Peptide Structures using NMR Data and Deterministic Global Optimization. *Journal of Computational Chemisty*, 20:1354–1370, 1999.

[57] J. L. Klepeis, M. G. Ierapetritou, and C. A. Floudas. Protein Folding and Peptide Docking: A Molecular Modeling and Global Optimization Approach. *Computers and Chemical Engineering*, 22:S3–S10, 1998.

[58] J. P. Knight and G. J. MacRae. A Combinatorial Optimization Approach to Molecular Design. *Nanotechnology*, 2:142–148, 1991.

[59] P. Kolar, J. W. Shen, A. Tsuboi, and T. Ishikawa. Solvent Selection for Pharmaceuticals. *Fluid Phase Equilibria*, 192:771–782, 2002.

[60] C. L. Kong. Combining Rules for Intermolecular Potential Parameters. II. Rules for the Lennard-Jones (12-6) Potential and the Morse Potential. *Journal of Chemical Physics*, 59:2464–2467, 1973.

[61] C. Laurence and M. Berthelot. Observations on the Strength of Hydrogen Bonding. *Perspectives in Drug Discovery and Design*, 18:39–60, 2000.

[62] A. R. Leach. *Molecular Modelling: Principles and Applications.* Addison Wesley Longman Limited, Singapore, 1996.

[63] G. N. Lewis and M. Randall. *Thermodynamics.* McGraw-Hill, New York, 2nd edition, 1961. Revised by K. S. Pitzer and L. Brewer.

[64] Z. Li and H. A. Scheraga. Monte-Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding. *Proc. Natl. Aced. Sci. U.S.A*, 84:6611–6615, 1987.

[65] D. G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, Reading, Massachusetts, 2nd edition, 1984.

[66] T. Magnussen, P. Rasmussen, and A. Fredenslund. UNIFAC Parameter Table for Prediction of Liquid-Liquid Equilibria. *Industrial and Engineering Chemistry Process Design and Development*, 20:331–339, 1981.

[67] G. C. Maitland, M. Rigby, E. B. Smith, and W. A. Wakeham. *Intermolecular Forces.* Clarendon Press, Oxford, 1981.

[68] J.R. Maple, U. Dinur, and A.T. Hagler. Derivation of Force Fields for Molecular Mechanics and Dynamics from Ab Initio Energy Surfaces. *Proc. natl. Acad. Sci*, 85:5350–5354, 1988.

[69] C. D. Maranas and C. A. Floudas. A Global Optimization Approach for Lennard-Jones Microclusters. *Journal of Chemical Physics*, 97:7667–7678, 1992.

[70] C. D. Maranas and C. A. Floudas. Finding All Solutions of Nonlinearly Constrained Systems of Equations. *Journal of Global Optimization*, 7:135–170, 1995.

[71] C.D. Maranas and C.A. Floudas. Global Minimum Potential Energy Conformations of Small Molecules. *Journal of Global Optimization*, 4:135–170, 1994.

[72] J. Marrero and R. Gani. Group-Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilibria*, 183-184:183–208, 2001.

[73] J. Marrero-Morejón and E. Pardillo-Fontdevila. Estimation of Pure Compound Properties using Group-Interaction contributions. *Thermodynamics*, 45(3):615–621, 1999.

[74] S.L. Mayo, B.D. Olafson, and W.A. Goddard III. Dreiding: A Generic Force Field. *Journal of Physical Chemistry*, 94:8897–8909, 1990.

[75] G. P. McCormick. Computability of Global Solutions to Factorable Nonconvex Programs: Part I - Convex Underestimating Problems. *Mathematical Programming*, 10:146–175, 1976.

[76] C. McDermott and N. Ashton. Note on the Definition of Local Composition. *Fluid Phase Equilibria*, 1:33–35, 1977.

[77] J. W. McFarland, A. Avdeef, C. M. Berger, and O. A. Raevsky. Estimating the Water Solubilities of Crystalline Compounds from their Chemical Structures Alone. *Journal of Chemical Information and Computer Science*, 41:1355–1359, 2001.

[78] A.-H. Meniai and D. M. T. Newsham. The Selection of Solvents for Liquid-liquid Extraction. *Trans IChemE*, 70:78–87, 1992.

[79] A. H. Meniai and D. M. T. Newsham. Computer-Aided Method for Interaction Parameter Calculations. *Trans IChemE*, 73:842–848, 1995.

[80] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of State Calculations by Fast-Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[81] R. E. Moore. *Interval Analysis.* Prentice-Hall, Englewood Cliffs, NJ, 1966.

[82] R.S. Mulliken. Electronic Population Analysis of LCAO-MO Molecular Wave Functions I. *The Journal of Chemical Physics*, 23:1833–1846, 1955.

[83] T. M. Nelson and P. C. Jurs. Prediction of Aqueous Solubility of Organic Compounds. *Journal of Chemical Information and Computer Science*, 34:601–609, 1994.

[84] A. Neumaier. Molecular Modeling of Proteins and Mathematical Prediction of Protein Structure. *SIAM Review*, 39(3):407–460, 1997.

[85] O. Odele and S. Macchietto. Computer Aided Molecular Design: A Novel Method for Optimal Solvent Selection. *Fluid Phase Equilibria*, 82:47–54, 1993.

[86] A. Z. Panagiotopoulos. Direct Determination of Phase Coexistence Properties of Fluids by Monte Carlo Simulation in a New Ensemble. *Molecular Physics*, 61:813–826, 1987.

[87] R. V. Pappu, R. K. Hart, and J. W. Ponder. Analysis and Application of Potential Energy Smoothing and Search Methods for Global Optimization. *Journal of Physical Chemistry B*, 102(48):9725–9742, 1999.

[88] E. Pardillo-Fontdevila and R. González-Rubio. A Group-Interaction Contribution Approach. A New Strategy for the Estimation of Physico-Chemical Properties of Branched Isomers. *Chemical Engineering Communications*, 163:245–254, 1997.

[89] J. Pillardy and L. Piela. Smoothing Techniques of Global Optimization: Distance Scaling Method in Searches for Most Stable Lennard-Jones Atomic Clusters. *Journal of Computational Chemistry*, 18(16):2040–2049, 1997.

[90] J. J. Potoff, J. R. Errington, and A. Z. Panagiotopoulos. Molecular Simulation of Phase Equilibria for Mixtures of Polar and Non-Polar Components. *Molecular Physics*, 97(10):1073–1083, 1999.

[91] S. L. Price. Applications of Realistic Electrostatic Modelling to Molecules in Complexes, Solids and Proteins. *Journal of Chemical Society, Faraday Transactions*, 92(17):2997–3008, 1996.

[92] I. Quesada and I. E. Grossmann. A Global Optimization Algorithm for Linear Fractional and Bilinear Programs. *Journal of Global Optimization*, 6:39–76, 1995.

[93] O. A. Raevsky, V. Y. Grigor'ev, D. B. Kireev, and N. S. Zefirov. Complete Thermodynamic Description of H-Bonding in the Framework of Multiplicative Approach. *Quantitative Structure-Activity Relationships*, 11:49–63, 1992.

[94] A.K. Ráppe, C.J. Casewit, K.S. Colwell, W.A. Goddard III, and W.M. Skiff. UFF: A Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *Journal of the American Chemical Society*, 114:10024–10035, 1992.

[95] K. Rasmussen. How to Develop Force Fields: An Account of the Emergence of Potential Energy Functions for Saccharides. *J. Mol. Struct.(Theochem)*, 395:91–106, 1997.

[96] Kj. Rasmussen. *Potential Energy Functions in Conformational Analysis*, volume 37 of *Lecture Notes in Chemistry*. Springer-Verlag, Berlin Heidelberg New York Tokyo, 1985.

[97] Kj. Rasmussen, S.B. Engelsen, J. Fabricius, and B. Rasmussen. *The Consistent Force Field: Development of Potential Energy Functions for Conformational Analysis*, volume 406 of *Recent Experimental and Computational Advances in Molecular Spectroscopy, NATO ASI Series C*, pages 381–419. Kluwer Academic Publishers, Dordrecht, 1993.

[98] H. Ratschek and J. Rokne. Interval Tools for Global Optimization. *Computers and Mathematical Applications*, 21:41–50, 1991.

[99] C.A. Reynolds, J.W. Essex, and W.G. Richards. Atomic Charges for Variable Molecular Conformations. *Journal of the American Chemical Society*, 114:9075–9079, 1992.

[100] M. Sacchetti. Thermodynamic Analysis of DSC Data for Acetaminophen Polymorphs. *Journal of Thermal Analysis and Calorimetry*, 63(2):345–350, 2000.

[101] S. Scheiner. *Hydrogen Bonding, A Theoretical Perspective*. Oxford University Press, New York, 1997.

[102] J. I. Siepmann and D. Frenkel. Configurational Bias Monte Carlo - A New Sampling Scheme for Flexible Chains. *Molecular Physics*, 75(1):59–70, 1992.

[103] B. Smit, P. Desmedt, and D. Frenkel. Computer-Simulations in the Gibbs Ensemble. *Molecular Physics*, 68(4):931–950, 1989.

[104] E. M. B. Smith and C. C. Pantelides. *Global Optimisation of General Process Models*, pages 355–386. Global Optimisation in Engineering Design. Kluwer Academic Publishers, 1996.

[105] Y. T. Sohn. *Journal of Korean Pharmaceutical Science*, 20(2):97–104, 1990.

[106] Y. Song, T. Hino, S. M. Lambert, and J. M. Prausnitz. Liquid-Liquid Equilibria for Polymer Solutions and Blends, Including Copolymers. *Fluid Phase Equilibria*, 117:69–76, 1996.

[107] Y. Song and E. A. Mason. Statistical-Mechanical Theory of a New Analytical Equation of State. *Journal of Chemical Physics*, 12:7840–7853, 1991.

[108] T. F. Soules. A Molecular Dynamic Calculation of the Structure of Sodium Silicate Glasses. *J. Chem. Phys.*, 71:4570–4578, 1979.

[109] T. F. Soules and A. K. Varshneya. Molecular Dynamic Calculations of a Sodium Borosilicate Glass Structure. *J. Amer. Ceram. Soc.*, 64:145–150, 1981.

[110] A.J. Stone and M. Alderton. Distributed Multipole Analysis: Methods and Applications. *Molecular Physics*, 56(5):1047–1064, 1985.

[111] R. Stryjek and H. Vera. An Improved Peng-Robinson Equation of State for Accurate Vapor-Liquid Equilibrium Calculations. *Canadian Journal of Chemical Engineering*, 64:323–333, 1986.

[112] A. K. Sum and S. I. Sandler. A Novel Approach to Phase Equilibria predictions Using *Ab Initio* Methods. *Industrial and Engineering Chemistry Research*, 38:2849–2855, 1999.

[113] A. K. Sum, S. I. Sandler, R. Bukowski, and K. Szalewicz. Prediction of the Phase Behavior of Acetonitrile and Methanol with *Ab Initio* Pair Potentials. I. Pure Components. *Journal of Chemical Physics*, 116(17):7627–7636, 2002.

[114] A. K. Sum, S. I. Sandler, R. Bukowski, and K. Szalewicz. Prediction of the Phase Behavior of Acetonitrile and Methanol with *Ab Initio* Pair Potentials. II. The Mixture. *Journal of Chemical Physics*, 116(17):7637–7644, 2002.

[115] H. Sun. COMPASS: An Ab Initio Force Field Optimized for Condensed-Phase Applications - Overview with Details on Alkane and Benzene Compounds. *Journal of Physical Chemistry*, 102:7338–7364, 1998.

[116] H. Sun, S.J. Mumby, J.R. Maple, and A.T. Hagler. An Ab Initio CFF93 All-atom Force Field for Polycarbonates. *Journal of the American Chemical Society*, 116:2978–2987, 1994.

[117] H. Sun and D. Rigby. Polysiloxanes: Ab initio Force Field and Structural, Conformational and Thermophysical Properties. *Spectrochimica Acta (A)*, 53:1301–1323, 1997.

[118] J. M. Sutter and P. C. Jurs. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds using a Quantitative Structure-Property Relationship. *Journal of Chemical Information and Computer Science*, 36:100–107, 1996.

[119] I. Todhunter and J. G. Leathem. *Spherical Trigonometry.* Macmillan, London, 1901.

[120] J. P. Valleau. *Advances in Chemical Physics*, volume 105. John Wiley and Sons, New York, 1999.

[121] B. W. H. van Beest, G. J. Kramer, and R. A. van Santen. Force-Fields for Silicas and Aluminophosphates Based on Ab Initio Calculations. *Phys. Rev.*, 159:1955–1958, 1990.

[122] E. Vilaseca. Solvent Effect on Conformational Equilibrium: A Monte Carlo Study of 1,2-dichloroethane in Carbon Tetrachloride. *Journal of Chemical Physics*, 104:4243-4257, 1995.

[123] J. Wang, W. Wang, S. Huo, M. Lee, and P. A. Kollman. Solvation Model Based on Weighted Solvent Accessible Surface Area. *Journal of Physical Chemistry B*, 105:5055–5067, 2001.

[124] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta Jr., and P. Weiner. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *Journal of the American Chemical Society*, 106:765 784, 1984.

[125] G. M. Wilson. Vapor-liquid Equilibrium XI. A New Expression for the Excess Free Energy of Mixing. *AIChE Journal*, 86:168–176, 1964.

[126] H. Wolff, A. Hoepfner, and H. M. Hoepfner. *Ber. Bunsenges. Phys. Chem.*, 68:410, 1964.

[127] D. S. H. Wong and S. I. Sandler. A Theoretically Correct Mixing Rule for Cubic Equations of State. *AIChE Journal*, 38:671–680, 1992.

[128] S. H. Yalkowsky and S. Banerjee. *Aqueous Solubility. Methods of Estimation for Organic Compounds.* Dekker, New York, 1992.

[129] S. H. Yalkowsky, R. Orr, and S. C. Valvani. Solubilities and Partitioning 2. The Solubility of Halobenzenes in Water. *Industrial and Engineering Chemical Fundamentals*, 18:351–353, 1979.

[130] S. H. Yalkowsky and S. C. Valvani. Solubilities and Partitioning 2. Relationships between Octanol-Water Partition Coefficient and Aqueous Solubility. *Journal of Chemical Engineering Data*, 24:127–129, 1979.

[131] S. H. Yalkowsky and S. C. Valvani. Solubilities and partitioning 1. Solubility of Nonelectrolytes in Water. *Journal of Pharmaceutical Science*, 69:912–922, 1980.

[132] J. M. Zamora and I. E. Grossmann. A Comprehensive Global Optimization Approach for the Synthesis of Heat Exchanger Networks with No Stream Splits. *Computers and Chemical Engineering*, 21:S65–S70, 1997.

[133] R. Zhao and H. Cabezas. Molecular Thermodynamics in the Design of Substitute Solvents. *I EC Research*, 40(8):3268–3280, 1998.

Appendix A

Modified D-UNIFAC Parameter Matrix

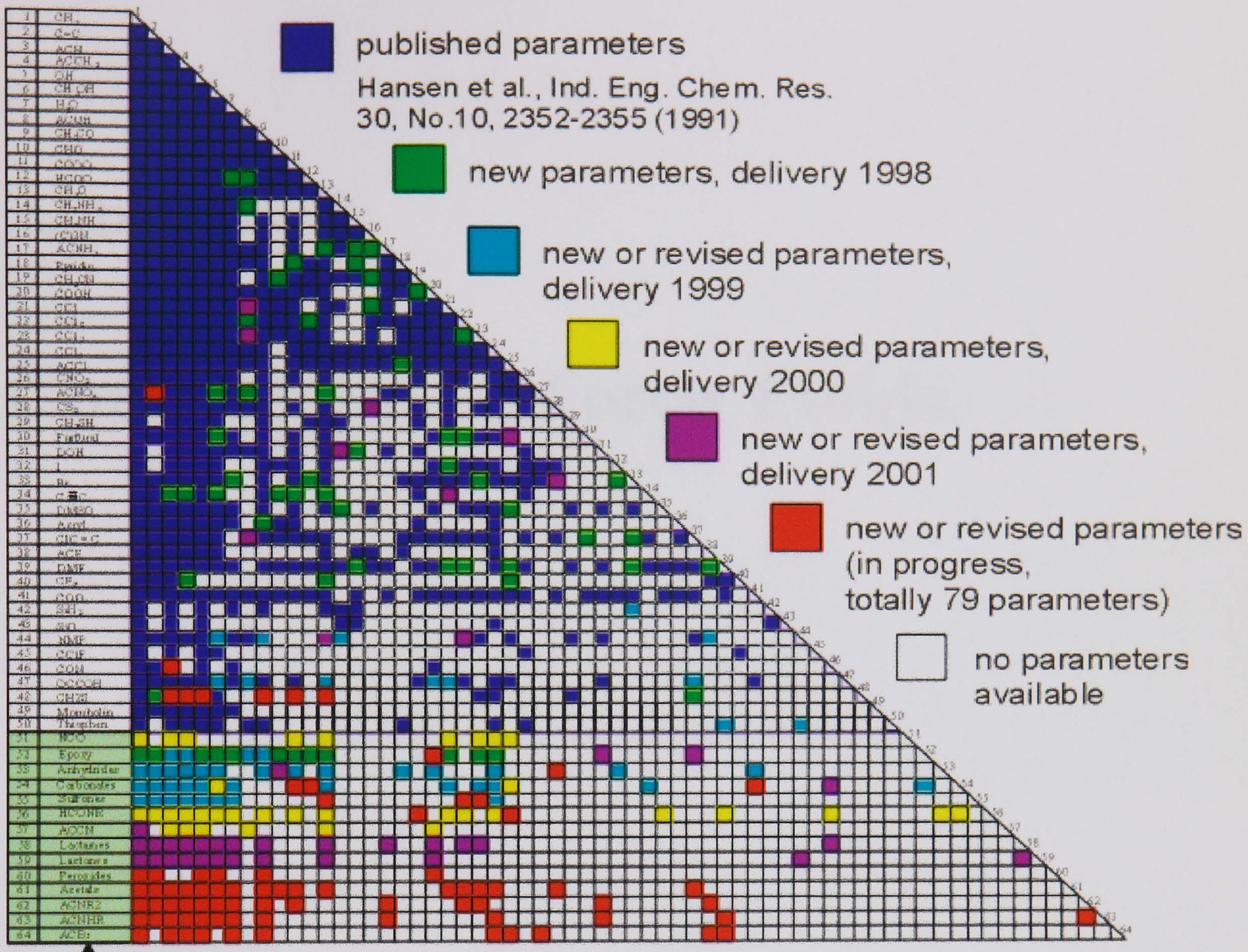


Figure A.1: Currently available group interaction parameters for D-UNIFAC obtained from http://www.uni-oldenburg.de/tchemie/Consortium/Matrix_MoUniDo.html

Appendix B

Comparison of Force Fields

	Bond length			Bond angle			Torsion angle			
Force-field	Quadratic bond-stretching	Quartic bond-stretching	Morse bond-stretching	Quadratic angle-bending	Quartic angle-bending	Cosine angle-bending	Single-cosine torsion	Three-term cosine torsion	Cosine-Fourier torsion	Sin-cos torsion
COMPASS		√			√			√		
CFF		√			√			√		
OPLS	√			√				√		
AMBER	√			√			√			
CHARMm	√			√			√			
CVFF			√	√			√			
UFF	√					√			√	
ESFF			√			√				√
Dreiding	√			√					√	
MM2		√ (cubic)			√ (2^{nd} & 6^{th} power)			√		
MM3		√			√ (2^{nd} – 6^{th} power)			√		

	Out-of-plane				Van der Waals				Electrostatic	
Force-field	Improper cosine out-of-plane	Improper quadratic out-of-plane	Improper square out-of-plane	Wilson out-of-plane	6-9 van der Waals	6-12 van der Waals	Exp-6	Hydrogen bonds	Atomic monopoles & screened Coulombic terms	Coulombic interaction
COMPASS			√		√					√
CFF				√	√					√
OPLS						√				√
AMBER	√					√		√ [10-12] (opt)		√
CHARMm		√				√		√ [6-4cos] (opt)		√
CVFF	√		√			√				√
UFF	√			√	√				√	
ESFF				√	√					√
Dreiding				√		√		√ [10-12]	√	
MM2							√			√
MM3							√			√ (bond-dipole)

Note: OPLS, AMBER, CHARMm, ESFF, Dreiding and UFF contain no cross-terms.

Bond length

Quadratic bond-stretching $k (r - r_0)^2$

Quartic bond-stretching $k_2 (r - r_0)^2 + k_3 (r - r_0)^3 + k_4 (r - r_0)^4$

Morse bond-stretching $k[1 - e^{(-\alpha(r_b - r_b^0))}]^2$

Bond angle

Quadratic angle-bending $k (\theta - \theta_0)^2$

Quartic angle-bending $k_2 (\theta - \theta_0)^2 + k_3 (\theta - \theta_0)^3 + k_4 (\theta - \theta_0)^4$

Cosine angle-bending various

Torsion angle

Single-cosine torsion $k (1 + cos(n\phi - \phi_0))$

Three-term cosine torsion $k_1 [1 - cos(\phi - \phi_{01})] + k_2 [1 - cos(2\phi - \phi_{02})] + k_3 [1 - cos(3\phi - \phi_{03})]$

Cosine-Fourier torsion $k (1 \pm cos\, n\phi)$

Sin-cos torsion $k_\phi (\frac{\sin^2 \theta_1 \sin^2 \theta_2}{\sin^2 \theta_1^0 \sin^2 \theta_2^0} + sign \frac{\sin^n \theta_1 \sin^2 \theta_2}{\sin^n \theta_1^0 \sin^n \theta_2^0} \cos[n\phi]$

Out-of-plane

Improper cosine out-of-plane $k [1 + cos(n\chi - \chi_0)]$

Improper quadratic out-of-plane $k (\chi - \chi_0)^2$

Improper square out-of-plane $k\chi^2$

Wilson out-of plane $k\chi^2$

$k (cos\, \chi - cos\, \chi_0)^2$

Force-field	Range
COMPASS	• Condensed-phase Optimised Molecular Potentials for Atomistic Simulation Studies • *Ab initio* force-field • First high quality force field to incorporate parameters for organic and inorganic materials • Optimised empirically to yield good agreement with experimental data • Energy terms include diagonal and off-diagonal cross coupling terms • Uses partial atomic charges
CFF	• 2nd generation force-field • Parameterised against a wide range of experimental observables for organic compounds containing H, C, N, O, S, P, halogen atoms and ions, alkali metal cations, and several biochemically important divalent metal cations • All the CFF force-fields (CFF91, CFF95, PCFF) have the same functional form but differ in the range of functional groups to which they were parameterised • CFF91 is applicable to hydrocarbons, proteins, protein-ligand interactions and is parameterised explicitly for acetals, acids, alcohols, alkanes, alkenes, amides, amines, aromatics, ethers and esters • CFF95 has slightly more atom types than CFF91 and is parameterised for additional functional groups. Recommended for all life sciences applications and for organic polymers e.g. polycarbonates and polysaccharides • PCFF is based on CFF91, extended to have a broader coverage of organic polymers, (inorganic) metals and zeolites. Parameterised for the functional groups listed in CFF91 and also He, Ne, Kr, Xe. • PEF series divides charges obtained from Mulliken population analysis
OPLS	• Optimised Potentials for Liquid Simulations developed to reproduce experimental data on fluids • Parameterised from Molecular Dynamics or Monte Carlo simulations using different sets of parameters to yield accurate thermodynamic and structural results for liquids. • Torsional potential parameters are obtained by fitting to conformational energy profiles calculated using *ab initio* RHF/6-13G*. • Intermolecular interactions are represented by interaction sites located on the nuclei. • Bond stretching and angle bending parameters are derived from the AMBER force field.
AMBER	• Classical force-field • Parameterised from experimental data against a limited number of organic fragments of proteins and nucleic acids. Widely used for proteins, DNA and other classes of molecules but the standard version is useful mainly for proteins and nucleic acids. • AMBER-95 uses an advanced potential-derived fractional charge model called RESP (restrained electrostatic potential fit). PD model evaluates the electrostatic potential over a range of inter-atomic distances using *ab initio* calculations, followed by fitting of the fractional charges based on the resulting curve.
CHARMm	• Classical force-field • Parameterised on the basis of ab initio energies and geometries of small organic molecules • Performs well over a broad range of calculations and simulations and is designed to give good results for a wide variety of systems, from small isolated molecules to solvated complexes of large biological macromolecules but it is not applicable to organometallic complexes • Assigns fractional charges using the method developed by Gasteiger using iterative partial equalization of orbital electronegativity which only considers the properties of the constituent atoms of a molecule and the network of bonds
CVFF	• Classical force-field • Fit to small organic crystals and gas phase structures and handles peptides, proteins and a wide range of organic systems. • Has anharmonic and cross-terms enhancements
UFF	• Broadly applicable force-field (a purely diagonal, harmonic force-field) • Parameters are generated from a set of rules based on element, hybridization and connectivity for the full periodic table. • Validated for main group compounds, organic molecules and metal complexes and other systems for which other forcefields do not have parameters • Moderately accurate for predicting geometries and conformational energy differences of organic molecules, main group inorganics and metal complexes • Uses the charge equilibration method
ESFF	• Broadly applicable force-field • A rule-based force field to provide a wide coverage of the periodic table to reproduce the structures of isolated molecules and crystals. • Its scope does not extend to highly accurate vibrational frequencies or other properties such as conformational energies • Used for structure prediction of organic, inorganic, and organometallic systems in gas or condensed phases
Dreiding	• Broadly applicable force-field • General force constants and geometry parameters are based on simple hybridization rules rather than on specific combinations of atoms • Allows reasonable predictions for a very much larger number of structures and has good coverage for organic, biological and main-group inorganic molecules • Moderately accurate for geometries, conformational energies, intermolecular binding energies and crystal packing • Uses the charge equilibration method
MM2 / MM3	• Developed for performing accurate calculations on gas phase structures and for determining vibrational frequencies of organic substances • Can calculated enthalpies of formation with good accuracy • MM3 force-field is intended for relatively small organic molecules not for large biological structures

Note: Some classical force-fields were originally created for modeling proteins and peptides, others for DNA and RNA. Some have been extended to handle more general systems having similar functional groups.

Rule-based force-fields have extended the range of force-field simulations to a wider range of elements.

2nd generation force-fields include parameters for all functional groups appropriate for protein simulations.

Broadly applicable force-fields rely on atomic parameters coupled with theoretically and empirically derived rules for generating explicit force-field parameters.

Appendix C

Boltzmann Jump Procedure utilising API routines from *Cerius*^{2}

```
void run_bjp ( double tem, int iter, int *istatus )
{
#define WAVELENGTH 36.00
        energy_terms[0] = msF_Total;

        *istatus = 0;

/* ... Get the object id of the current model */

        msD_MOD_Current(&model, &status);

/* ... Find out how many atoms are in this model */

        msD_MOD_NumAtoms(model, msD_SYM_ALL, &lenbuf, &status);
        if (lenbuf < 1)
        {
                msU_MSG_Display(msU_MSG_TEXTPORT, "boltzmann_ERR_NoAtoms", &status);
                return;
        }

/* ... Check how many atoms are currently selected in this model */
```

```
msD_SEL_NumSelected(&model, 1, msD_ATOM_ID, &numsel, &status);
if (numsel < 1)
{
        msU_MSG_Display(msU_MSG_TEXTPORT, "boltzmann_ERR_NoAtomsSel", &status);
        return;
}

/* ... Get the object id of the selected atoms */

        atsel = (int *) malloc (numsel *sizeof(int));

        msD_SEL_FindSelected(model, msD_ATOM_ID, atsel, numsel, &n_sel, &status);

        if (status != SUCCESS)
        {
                msU_MSG_ArgS("msD_SEL_FindSelected");
                msU_MSG_Display(msU_MSG_TEXTPORT, "boltzmann_ERR_API", &status);
                *istatus = 1;
                return;
        }

/* ... use this to get the Cartesian coordinates of the selected atoms */

        selxyz = (dTriplet *) malloc (n_sel * sizeof(dTriplet));

        msD_XYZ(atsel, n_sel, selxyz, &status);

        if (status != SUCCESS)
        {
                msU_MSG_ArgS("msD_XYZ");
                msU_MSG_Display(msU_MSG_TEXTPORT, "boltzmann_ERR_API", &status);
                *istatus = 1;
                return;
        }

        selmass    = (float *) malloc(n_sel * sizeof(float));

        msD_ATM_Mass(atsel, n_sel, selmass, &status);
```

```
if (status != SUCCESS)
{
        msU_MSG_ArgS("msD_ATM_Mass");
        msU_MSG_Display(msU_MSG_TEXTPORT, "boltzmann_ERR_API", &status);
        *istatus = 1;
        goto error_line;
        }


/* ... save the xyz coordinates in temporary arrays */

        tempselxyz  = (dTriplet *) malloc(n_sel * sizeof(dTriplet));

        if (tempselxyz == NULL)
        {
                *istatus = 1;
                goto error_line;
        }

/* ... start perturbation of molecules */

        for (counter = 0; counter <= 40; counter++) {

        printf("The iteration number is %d\n", counter);

/* ... now randomly perturb the molecules. If the RMS of the torsional angle
   and the Cartesian space is less than the parameters AND the rigid body
   translation window and the rigid body angular window also meet the parameters
   set, accept the move or reinitialize and perturb again */

        save_xyz(tempselxyz, selxyz);

/* ... calculate centre of mass */

        calculate_com();

/* ... calculate energies */
```

```
msF_EEX_MSetup(model, msF_Automate, &status);
printf("Status msF_EEX_MSetup: %d\n", status);

msF_EEX_MValidity(model, &valid, &status);
printf("Status msF_EEX_MValidity: %d %d\n", valid, status);

msF_EEX_Evaluate(msF_EVAL_E, &status);
printf("Status msF_EEX_Evaluate: %d\n", status);

msF_EEX_Energies(old_energy, energy_terms, 1, &status);
printf("The energy of the system is %lf\n", old_energy[0]);

perturb();

msD_SetXYZ(atsel, n_sel, tempselxyz, &status);

if (status != SUCCESS)
{
        msU_MSG_ArgS("msD_SetXYZ");
        msU_MSG_Display(msU_MSG_TEXTPORT, "boltzmann_ERR_API", &status);
        *istatus = 1;
        return;
}

msG_RedisplayModels ( RD_OLDSCALE, RD_COMPLETE);

/* ... compare energies */

msF_EEX_MSetup(model, msF_Automate, &status);
printf("Status msF_EEX_MSetup: %d\n", status);

msF_EEX_MValidity(model, &valid, &status);
printf("Status msF_EEX_MValidity: %d %d\n", valid, status);

msF_EEX_Evaluate(msF_EVAL_E, &status);
printf("Status msF_EEX_Evaluate: %d\n", status);
```

```
msF_EEX_Energies(new_energy, energy_terms, 1, &status);
printf("The energy of the system is %lf\n", new_energy[0]);

if (new_energy[0] <= old_energy[0]) {
        save_xyz(selxyz, tempselxyz);
}
        else {
        ran_num = ran2(&idum);
        x = (double) exp(-(new_energy[0]-old_energy[0])/(0.001987*2000));
        printf("x = %lf\n", x);
        if (x  < ran_num) {
                save_xyz(selxyz, tempselxyz);
        }
}

        msD_SetXYZ(atsel, n_sel, selxyz, &status);
        if (status != SUCCESS)
        {
                msU_MSG_ArgS("msD_SetXYZ");
                msU_MSG_Display(msU_MSG_TEXTPORT, "boltzmann_ERR_API", &status);
                *istatus = 1;
                return;
        }

        msG_RedisplayModels ( RD_OLDSCALE, RD_COMPLETE);
}

error_line:
        free (tempselxyz);
        free (selxyz);
        free (atsel);
        free (selmass);

return;
}
```

```
void perturb(void)
{
        int i;
        double ra1, ra2, ra3, ra4, ran;

        ra1 = ran2(&idum);
        ra2 = ran2(&idum);
        ra3 = ran2(&idum);
        ra4 = ran2(&idum);
        ran = ran2(&idum);

        for (i=0; i < n_sel; i++)
        {
                tempselxyz[i][0] = com[0]+(tempselxyz[i][0]-com[0])*cos(ra1*(2*PI/WAVELENGTH))
                                 - (tempselxyz[i][1]-com[1])*sin(ra1*(2*PI/WAVELENGTH));
                tempselxyz[i][0] = com[0]+(tempselxyz[i][2]-com[2])*sin(ra1*(2*PI/WAVELENGTH))
                                 + (tempselxyz[i][0]-com[0])*cos(ra1*(2*PI/WAVELENGTH));
                tempselxyz[i][1] = com[1]+(tempselxyz[i][0]-com[0])*sin(ra1*(2*PI/WAVELENGTH))
                                 + (tempselxyz[i][1]-com[1])*cos(ra1*(2*PI/WAVELENGTH));
                tempselxyz[i][1] = com[1]+(tempselxyz[i][1]-com[1])*cos(ra1*(2*PI/WAVELENGTH))
                                 - (tempselxyz[i][2]-com[2])*sin(ra1*(2*PI/WAVELENGTH));
                tempselxyz[i][2] = com[2]+(tempselxyz[i][1]-com[1])*sin(ra1*(2*PI/WAVELENGTH))
                                 + (tempselxyz[i][2]-com[2])*cos(ra1*(2*PI/WAVELENGTH));
                tempselxyz[i][2] = com[2]+(tempselxyz[i][2]-com[2])*cos(ra1*(2*PI/WAVELENGTH))
                                 - (tempselxyz[i][0]-com[0])*sin(ra1*(2*PI/WAVELENGTH));
        }

        if (ran < 0.5) {
           for (i=0; i < n_sel; i++) {
                tempselxyz[i][0] = tempselxyz[i][0] - ra2*0.5;
                tempselxyz[i][1] = tempselxyz[i][1] - ra3*0.5;
                tempselxyz[i][2] = tempselxyz[i][2] - ra4*0.5;
        }}
        else {
           for (i=0; i < n_sel; i++) {
                tempselxyz[i][0] = tempselxyz[i][0] + ra2*0.5;
```

```
                tempselxyz[i][1] = tempselxyz[i][1] + ra3*0.5;
                tempselxyz[i][2] = tempselxyz[i][2] + ra4*0.5;
        }}

return;
}
```

Appendix D

Experimental Data for Bifunctional Benzene Derivatives

1. *p*-chlorotoluene (2) / *n*-hexane (1)

```
                ANTOINE CONSTANTS          RANGE

(1)    6.87776 1171.530   224.366   -25-   92 C

(2)    7.77830 2107.520   267.777    31- 162 C

TEMPERATURE=   120.05 DEGREE C    ( 393.20 K )

LIT: OSCARSON J.L.,LUNDELL S.O.,CUNNINGHAM J.R.,
     AICHE SYMP.SER. 0,1(1987).

    experimental data
 P [mm Hg]        x1       y1

   478.54      0.06740 0.50900
   1111.6      0.28700 0.82800
   1860.9      0.56100 0.91900
   2430.2      0.77300 0.96300
   2533.7      0.82400 0.97200
   2952.2      0.98600 0.99800
```

2. *p*-cresol (2) / water (1)

```
SLE type                    : unknown type
number of  components       :      2
```

number of data points : 9
pressure : not given
reference number : 254

ref.: SIDGWICK N.V.,SPURRELL W.J.,DAVIES T.E.,
J.CHEM.SOC.LONDON 0,1202(1915).

experimental data

T [K]	xl1
279.85	0.5242000
282.35	0.5476000
283.95	0.6024000
290.25	0.7478000
293.45	0.8004000
297.15	0.8559000
300.65	0.9331000
303.05	0.9461000
306.95	1.0000000

3. *p*-toluidine (2) / ethanol (1)

SLE type : unknown type
number of components : 2
number of data points : 1
pressure : not given
reference number : 828

ref.: PUCHER G.,DEHN W.M.,
J.AM.CHEM.SOC. 43,1753(1921).

experimental data

T [K]	xl1
298.15	0.3210722

SLE type : unknown type
number of components : 2
number of data points : 3
pressure : not given
reference number : 1002

ref.: SPEYERS C.L.,
AM.J.SCI. 4,293(1902).

experimental data

T [K]	xl1
295.25	0.4910000
284.85	0.6612000
273.15	0.7928000

4. *p*-toluidine (2) / water (1)

SLE type : unknown type
number of components : 2
number of data points : 1
pressure : not given
reference number : 854

ref.: DEHN W.M.,
J.AM.CHEM.SOC. 39,1399(1917).

experimental data

T [K]	xl1
298.15	0.9891221

5. *p*-chlorobenzoic acid (2) *n*-heptane (1)

SLE type : unknown type
number of components : 2
number of data points : 6
pressure : not given
reference number : 312

ref.: SIDGWICK N.V.,EWBANK E.K.,
J.CHEM.SOC. 979(1921).

experimental data

T [K]	xl1
409.25	0.9891000
438.45	0.9677000
454.05	0.9330000

480.35 0.7748000

491.45 0.5973000

500.75 0.3199000

6. *p*-chlorobenzoic acid (2) water (1)

SLE type : unknown type

number of components : 2

number of data points : 10

pressure : not given

reference number : 472

ref.: FLASCHNER O.,RANKIN I.G.,
Z.PHYS.CHEM. 23(1909).

experimental data

T [K]	xl1
435.15	0.9964530
453.15	0.9873733
456.95	0.9723705
458.75	0.9529929
459.15	0.9287388
460.15	0.8967857
462.15	0.8553676
465.15	0.7867038
477.15	0.6402347
513.15	0.0000000

7. *p*-hydroxybenzoic acid (2) /*n*-heptane (1)

SLE type : unknown type

number of components : 2

number of data points : 2

pressure : not given

reference number : 312

ref.: SIDGWICK N.V.,EWBANK E.K.,
J.CHEM.SOC. 979(1921).

experimental data

T [K]	xl1

```
470.15 0.9923000
481.65 0.9869000
```

8. *p*-hydroxybenzoic acid (2) / ethanol (1)

```
SLE type                  : unknown type
number of  components     :     2
number of data points     :     4
pressure                  :  not given
reference number          :   312

ref.: SIDGWICK N.V.,EWBANK E.K.,
      J.CHEM.SOC. 979(1921).

 experimental data
    T [K]     x11
  340.15 0.8257000
  409.65 0.6581000
  457.15 0.3821000
  486.15 0.0000000
```

9. *p*-hydroxybenzoic acid (2) / water (1)

```
SLE type                  : unknown type
number of  components     :     2
number of data points     :    15
pressure                  :  not given
reference number          :   312

ref.: SIDGWICK N.V.,EWBANK E.K.,
      J.CHEM.SOC. 979(1921).

 experimental data
    T [K]     x11
  324.05 0.9968000
  329.15 0.9957000
  333.35 0.9943000
  336.15 0.9935000
  338.25 0.9929000
  341.35 0.9916000
  348.85 0.9868000
```

353.25 0.9828000

362.75 0.9691000

370.15 0.9483000

382.65 0.8978000

387.15 0.8764000

410.65 0.7552000

440.15 0.5711000

486.15 0.0000000

SLE type : unknown type

number of components : 2

number of data points : 10

pressure : not given

reference number : 472

ref.: FLASCHNER O.,RANKIN I.G.,

Z.PHYS.CHEM. 23(1909).

experimental data

T [K] xl1

350.15 0.9857112

363.15 0.9687958

370.55 0.9482347

377.55 0.9196765

384.95 0.8845927

393.15 0.8385980

407.15 0.7733234

427.55 0.6570927

453.75 0.4487255

486.15 0.0000000

10. *p*-aminobenzoic acid (2) / ethanol (1)

SLE type : unknown type

number of components : 2

number of data points : 9

pressure : not given

reference number : 1022

ref.: LAZZELL C.L.,JOHNSTON J.,
J.PHYS.CHEM. 32,1331(1928).

experimental data

T [K]	xl1
298.15	0.9503000
348.95	0.8644000
354.45	0.8471000
361.85	0.7996000
372.15	0.7796000
388.15	0.7000000
413.25	0.5199000
428.35	0.3448000
460.15	0.0000000

11. *p*-aminobenzoic acid (2) / water (1)

SLE type : unknown type
number of components : 2
number of data points : 12
pressure : not given
reference number : 472

ref.: FLASCHNER O.,RANKIN I.G.,
Z.PHYS.CHEM. 23(1909).

experimental data

T [K]	xl1
355.35	0.9931317
367.15	0.9856100
376.75	0.9680020
382.15	0.9466878
385.35	0.9209881
388.35	0.8830367
392.35	0.8347776
396.55	0.7801478
405.15	0.6554791
417.95	0.5092679
431.15	0.3069653
459.15	0.0000000

Appendix E

Once-Continuously Differentiable Underestimators for the Torsional Angle Terms

Taking the example from the pseudocode for the first torsional term, $E_{\phi 1_i}$:

if $(\phi_i^L \leq \phi_{min} \leq \phi_i^U)$

  if ($\phi_i \leq \phi_{min}$)

    if ($\phi_i^L < \phi^{*\prime}$)

      $L_{\phi 1_i} = V_{1_i}/C_{1_i}(\phi_i - \phi_{min})^2 + K_{1_i}$ with $\phi_1 = \phi_i^L, \phi_2 = \phi_{min}$

    else

      $L_{\phi 1_i} = E_{\phi 1_i}$

  else

    if ($\phi_i^U > \phi^{*\prime\prime}$)

      $L_{\phi 1_i} = V_{1_i}/C_{1_i}(\phi_i - \phi_{min})^2 + K_{1_i}$ with $\phi_1 = \phi_{min}, \phi_2 = \phi_i^U$

    else

      $L_{\phi 1_i} = E_{\phi 1_i}$

else if ($\phi_{min} - \pi/2 \leq \phi_i^L < \phi_i^U \leq \phi_{min} + \pi/2$)

  $L_{\phi 1_i} = E_{\phi 1_i}$

else if ($\phi_i^L < \pi/2 - \gamma_1 < \phi_i^U$)

  $L_{\phi 1_i} = V_{1_i}/C_{1_i}(\phi_i - \phi_{min})^2 + K_{1_i}$ with $\phi_1 = \phi_i^L, \phi_2 = \phi_i^U$

else if ($\phi_i^L < 3\pi/2 - \gamma_1 < \phi_i^U$)

  $L_{\phi 1_i} = V_{1_i}/C_{1_i}(\phi_i - \phi_{min})^2 + K_{1_i}$ with $\phi_1 = \phi_i^L, \phi_2 = \phi_i^U$

else

  $L_{\phi 1_i} = [(E_{\phi 1_i}(\phi_i^U) - E_{\phi 1_i}(\phi_i^L))/(\phi_i^U - \phi_i^L)](\phi_i - \phi_i^L) + E_{\phi 1_i}(\phi_i^L)$

In order to show that the underestimators are continuous and first-order continuously differentiable, the different functions and first-order derivatives will be shown to meet at the boundaries.

Case 1: Are underestimators continuous?

The first torsional term, $E_{\phi 1_i}$, states that:

$$E_{\phi 1_i} = \frac{V_{1_i}}{2}(1 + \cos(\phi_i + \gamma_{1_i}))$$

At $\phi_i = \phi_{min}$, $E_{\phi 1_i} = 0$.

The underestimators for the case where $(\phi_i^L \leq \phi_{min} \leq \phi_i^U)$:

If $(\phi_i \leq \phi_{min}$ AND $\phi_i^L < \phi^{*\prime})$,

$$L_{\phi 1_i} = \frac{V_{1_i}(\cos(\phi_{min}) - \cos(\phi_i^L + \gamma_{1_i}))}{2((\phi_{min} - \phi_{min})^2 - (\phi_i^L - \phi_{min})^2)}(\phi_i - \phi_{min})^2$$
$$+ \frac{V_{1_i}}{2}(1 + \cos(\phi_i^L + \gamma_{1_i})) - \frac{V_{1_i}(\cos(\phi_{min}) - \cos(\phi_i^L + \gamma_{1_i}))}{2((\phi_{min} - \phi_{min})^2 - (\phi_i^L - \phi_{min})^2)}(\phi_i^L - \phi_{min})^2$$

At $\phi_i = \phi_{min}$,

$$\begin{aligned} L_{\phi 1_i} &= 0 + \frac{V_{1_i}}{2}(1 + \cos(\phi_i^L + \gamma_{1_i})) - \frac{V_{1_i}(-1.0 - \cos(\phi_i^L + \gamma_{1_i}))}{2(0.0 - (\phi_i^L - \phi_{min})^2)}(\phi_i^L - \phi_{min})^2 \\ &= 0 + \frac{V_{1_i}}{2}(1 + \cos(\phi_i^L + \gamma_{1_i})) - \frac{V_{1_i}}{2}(1 + \cos(\phi_i^L + \gamma_{1_i})) \\ &= 0 \end{aligned}$$

If $(\phi_i \geq \phi_{min}$ AND $\phi_i^U < \phi^{*\prime\prime})$,

$$L_{\phi 1_i} = \frac{V_{1_i}(\cos(\phi_i^U + \gamma_{1_i}) - \cos(\phi min))}{2((\phi_i^U - \phi_{min})^2 - (\phi_{min} - \phi_{min})^2)}(\phi_i - \phi_{min})^2$$
$$+ \frac{V_{1_i}}{2}(1 + \cos(\phi_{min})) - \frac{V_{1_i}(\cos(\phi_i^U + \gamma_{1_i}) - \cos(\phi_{min}))}{2((\phi_i^U - \phi_{min})^2 - (\phi_{min} - \phi_{min})^2)}(\phi_{min} - \phi_{min})^2$$

At $\phi_i = \phi_{min}$,

$$L_{\phi 1_i} = 0 + \frac{V_{1_i}}{2}(1 + \cos(\phi_{min})) - 0$$
$$= 0 + \frac{V_{1_i}}{2}(1 + (-1)) - 0$$
$$= 0$$

As all the functions that contain ϕ_{min} have the value 0 at that point, these functions are all continuous.

Case 2: Are First order derivatives of underestimators continuous?

The first-order derivative of the first torsional term, $\frac{\partial E_{\phi 1_i}}{\partial \phi 1_i}$, states that:

$$\frac{\partial E_{\phi 1_i}}{\partial \phi 1_i} = \frac{V_{1_i}}{2}(-\sin(\phi_i + \gamma_{1_i}))$$

At ϕ_{min}, $\frac{\partial E_{\phi 1_i}}{\partial \phi 1_i} = 0$.

The first order derivatives for the case where $(\phi_i^L \leq \phi_{min} \leq \phi_i^U)$:

If $(\phi_i \leq \phi_{min}$ AND $\phi_i^L < \phi^{*\prime})$,

$$\frac{\partial L_{\phi 1_i}}{\partial \phi 1_i} = 2\frac{V_{1_i}(\cos(\phi_{min}) - \cos(\phi_i^L + \gamma_{1_i}))}{2((\phi_{min} - \phi_{min})^2 - (\phi_i^L - \phi_{min})^2)}(\phi_i - \phi_{min})$$

At $\phi_i = \phi_{min}$,

$$\frac{\partial L_{\phi 1_i}}{\partial \phi 1_i} = 2\frac{V_{1_i}(\cos(\phi_{min}) - \cos(\phi_i^L + \gamma_{1_i}))}{2((\phi_{min} - \phi_{min})^2 - (\phi_i^L - \phi_{min})^2)}(\phi_{min} - \phi_{min})$$
$$= 0$$

If $(\phi_i \geq \phi_{min}$ AND $\phi_i^U < \phi^{*\prime\prime})$,

$$\frac{\partial L_{\phi 1_i}}{\partial \phi 1_i} = 2\frac{V_{1_i}(\cos(\phi_i^U + \gamma_{1_i}) - \cos(\phi min))}{2((\phi_i^U - \phi_{min})^2 - (\phi_{min} - \phi_{min})^2)}(\phi_i - \phi_{min})$$

At $\phi_i = \phi_{min}$,

$$\frac{\partial L_{\phi 1_i}}{\partial \phi 1_i} = 2\frac{V_{1_i}(\cos(\phi_i^U + \gamma_{1_i}) - \cos(\phi min))}{2((\phi_i^U - \phi_{min})^2 - (\phi_{min} - \phi_{min})^2)}(\phi_{min} - \phi_{min})$$

$$= 0$$

As all the functions that contain ϕ_{min} have zero gradient at that point, these functions are all first order continuously differentiable.

Appendix F

Analytical Proof of the Convexification of Nonbonded Interactions Term

We begin by writing the necessary optimality conditions for a minimum in $H_{r_{ij}}$. At the minimum, the value of the derivative of $H_{r_{ij}}$ is 0. This condition can be expressed as an 11^{th} order polynomial in r_{ij} as shown in Equation F.1.

$$P_{r_{ij}} = \left[-6q_iq_je^2r_{ij}^{11} + 4\epsilon_{ij}\left[\left(-2184\sigma_{ij}^{12}\right) + \left(336\sigma_{ij}^{6}r_{ij}^{6}\right)\right]\right] = 0 \tag{F.1}$$

Using Descartes' Rule (Cajori, 1904), the number of times the sign changes in the polynomial coefficients indicates the maximum number of roots. Depending on the signs of the charges on the atoms, the polynomial P may be described using two cases.

For the case where $q_iq_j < 0$, $P_{r_{ij}}$ has at most one root since there is only one sign change. In the case where $P_{r_{ij}}(r_{ij}^L)P_{r_{ij}}(r_{ij}^U) < 0$, the number of real roots within the interval $[r_{ij}^L, r_{ij}^U]$ must be odd and therefore the number of roots within the bounds is then only 1 (*Type* 1). On the other hand, if $P_{r_{ij}}(r_{ij}^L)P_{r_{ij}}(r_{ij}^U) > 0$, then the number of real roots is 0 (*Type* 2).

As for the case where $q_iq_j > 0$, $P_{r_{ij}}$ has at most two roots. Similarly when $P_{r_{ij}}(r_{ij}^L)P_{r_{ij}}(r_{ij}^U) < 0$, the number of real roots within the interval $[r_{ij}^L, r_{ij}^U]$ must be odd and so the number

of roots is 1 (*Type* 3). Conversely, when $P_{r_{ij}}(r_{ij}^L)P_{r_{ij}}(r_{ij}^U) > 0$. the number of roots may either be 0 or 2 depending on the coefficients of the equation (*Type* 4). However for the coefficients used in the force field, the number of roots would be 0 as the number of roots for $E_{r_{ij}} \leq 1$ as the term reflects the attraction and repulsion between two nonbonded atoms. Table F.1 shows an analysis of the different atom pairs found in a *n*-butane/ethylamine system. For pairs of *Type* 1 using the $\alpha_{ij}-$value as a result of the direct minimisation of

Pairs	Type
C, RCH_3 - C, RCH_3	*Type* 4 (0 root)
C, RCH_3 - C, R_2CH_2	*Type* 4 (0 root)
C, RCH_3 - H, RH	*Type* 1
C, RCH_3 - N, RNH_2	*Type* 4 (0 root)
C, RCH_3 - H, RNH_2	concave
C, R_2CH_2 - C, R_2CH_2	*Type* 4 (0 root)
C, R_2CH_2 - H, RH	*Type* 1
C, R_2CH_2 - N, RNH_2	*Type* 4 (0 root)
C, R_2CH_2 - H, RNH_2	concave
H, RH - H, RH	*Type* 4 (0 root)
H, RH - N, RNH_2	*Type* 1
H, RH - H, RNH_2	convex
N, RNH_2 - N, RNH_2	*Type* 4 (0 root)
N, RNH_2 - H, RNH_2	concave
H, RNH_2 - H, RNH_2	convex

Table F.1: Equation type for $P_{r_{ij}}$ for *n*-butane/ethylamine system

the Hessian term would give a tight underestimator for the non-convex term. For pairs of *Type* 4, the minimum value is given by $\min(H_{r_{ij}}(r_{ij}^L), H_{r_{ij}}(r_{ij}^U))$.

BIBL.
LONDIN.
UNIV.