Microscopic Interpretations of Drug Solubility

Laban Bondesson

Theoretical Chemistry
Royal Institute of Technology
Stockholm 2007

Microscopic Interpretations of Drug Solubility
Doctoral thesis

ISBN 978-91-7178-691-3
ISSN 1654-2312
TRITA-BIO-Report 2007:7
Printed by Universitetsservice US AB,
Stockholm, Sweden, 2007
Typeset in LaTeX by the author.

Abstract

The development of computational models for predicting drug solubility has increased drastically during the last decades. Nevertheless these models still have difficulties to estimate the aqueous solubility as accurately as desired. Different aspects that are known to have a large impact on the aqueous solubility of a molecule have been studied in detail in this thesis using various theoretical methods with intension to provide a microscopic view on drug solubility. The first aspect studied is the hydrogen bond energies. The validity of the additive model, often used in the field of solubility models has been tested using density functional theory by examining eight drug molecules. The impact of hydrogen bonds in Infrared and Raman spectra of three commonly used drug molecules has also been demonstrated. The calculated spectra are found to be in good agreement with the experimental data. Another aspect that is important in solubility models is the volume that a molecule occupies when it is dissolved in water. The volume term and its impact on the solvation energy has therefore also been calculated using three different methods. It is shown that the calculated volumes is strongly dependent on the computational methods employed, especially for larger molecules. The interaction energy between a molecule and the surrounding solute can be estimated in different ways. In this thesis a new computational scheme has been developed for calculating solute-solvent interaction energies. It applies molecular dynamics simulations to generate structures of solute-solvent complexes and linear scaling quantum chemical methods to calculate the electronic structures of the selected complexes and the interaction energies. Some technical details, such as the convergence of solute-solvent interaction energies with respect to the number of solute molecules included, the use of mixed basis sets and the basis set superposition error, have also been provided.

Most of the solubility models assume the solute molecule to be in the bulk of the solvent. The molecular behavior at the water/gas interface has been investigated to see how it differs from bulk. It was found that the concentration close to the interface was almost three times

higher than in the bulk. This results from the fact that the energy gap between the interface and the gas phase is larger than that between the bulk and the gas phase.

Preface

The work presented in this thesis has been carried out at the Department of Theoretical Chemistry, Royal Institute of Technology, Stockholm, Sweden.

List of papers included in the thesis

Paper I Density functional theory calculations of hydrogen bonding energies of drug molecules, **L. Bondesson**, K. V. Mikkelsen, Y. Luo, P. Garberg and H. Ågren, J. Mol. Struct. (THEOCHEM) 81, 776 (2006).

Paper II Hydrogen bonding effects on infrared and Raman spectra of drug molecules, **L. Bondesson**, K. V. Mikkelsen, Y. Luo, P. Garberg and H. Ågren, Spectrochimica Acta A: Mol. Bio. Spectro. 66, 213 (2007).

Paper III Solvation of N_3^- at the water surface: the Polarizable Continuum Model approach, **L. Bondesson**, L. Frediani, H. Ågren and B. Menucci, J. Phys. Chem. B. 110, 11361 (2006).

Paper IV Calculations of the cavitation volumes and partial molar volumes of drugs in water, **L. Bondesson** and H. W. Hugosson, in preparation.

Paper V A linear scaling study of solvent-solute interaction energy of drug molecules in aqua solution, **L. Bondesson**, E. Rudberg, Y. Luo and P Salek, submitted, 2007

Paper VI Basis set dependence of solvent-solute interaction energy of benzene in water: A linear scaling ab initio study, **L. Bondesson**, E. Rudberg, Y. Luo and P Salek, submitted, 2007.

Comments on my contribution to the papers included

- I was responsible for calculations and for writing of Paper I.
- I was responsible for calculations and for writing of Paper II.
- I was responsible for calculations and part of writing of Paper III.
- I was responsible for calculations and part of writing of the first draft for Paper IV.
- I was responsible for calculations and for writing of Paper V.
- I was responsible for calculations and for writing of Paper VI.

Acknowledgments

This Doctoral thesis would have been very difficult if not impossible to produce without the help of many people, whom I would like to thank:

I would like to thank my supervisor Prof. Hans Ågren for giving me the opportunity to study at the Department of Theoretical Chemistry.

I wish to thank Dr. Per Garberg at Biovitrum who introduced me to the subject and helped me financially for the first year of my studies.

And, special thanks to Prof. Yi Luo who is always optimistic and helpful.

I would also like to thank my collaborators Dr. Håkan Hugosson, Prof. Kurt V. Mikkelsen, Dr. Luca Frediani, Elias Rudberg and Dr. Pawel Salek who explained different theories and brought good ideas to the projects.

Thanks to Elias Rudberg and Emanuel Rubensson who helped me writing computer programs.

I would like to thank all my colleagues of the Theoretical Chemistry group in Stockholm, Biovitrum, and the Department of Chemistry in Copenhagen.

Finally, my special thanks go to my kids Theo and Douglas and my wife Jenny for their love and support during these years.

Contents

Chapter 1

Introduction

In order to pass through biological membranes a molecule must be soluble in water. If the solubility of the drug is too low, drug administration via the oral route becomes impossible and the medical intake will be less convenient for patients, whereas highly soluble molecules are quickly distributed. The solubility of compounds therefore represents a significant problem in drug discovery research today. When large compound collections are screened, many compounds with a low solubility tend to be proposed as candidates for new drugs. Unfortunately these low solubility compounds are unsuitable as medicine and can therefore never be developed to drugs. If these compounds could be eliminated at an early stage, either prior to activity screening or early in the hit-to-lead phase when hundreds of "hits" need to be evaluated, resources could be saved and the lead finding process improved. Throughout the last decades a lot of effort has been spent to develop fast experimental and computational methods to predict the solubility of these candidate drugs. In an economic and humanitarian perspective an accurate computational method to predict the aqueous solubility could lead to less expensive medicine in the future. The use of an accurate model that could reject proposed structures that are not in the desired solubility range is also favorable from an environmental point of view since fewer molecules need to be synthesized, which leads to less pollutants from the pharmacological industry.

The term aqueous solubility (S) that will be used throughout this thesis is defined as the amount (mol) of the investigated molecule that can be dissolved in one liter of water. The range of the aqueous solubility is large between different compounds and therefore the logarithm of the amount dissolved is normally used (log(S)). Since the molecule may protonate or deprotonate depending on the pH value in water, the solubility is measured in its least soluble environment, i.e. the neutral form. This is usually referred to as the intrinsic solubility. In fact what is measured is the amount of the liquid or crystal solute that is dissolved in water, as shown in the schematic Figure 1, which depends on thermodynamic proper-

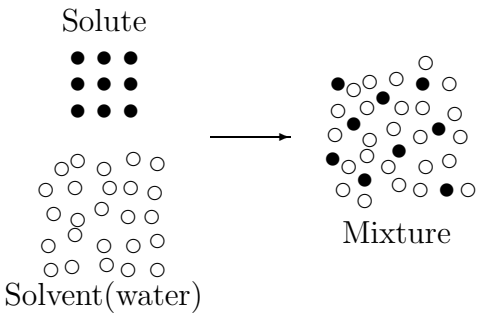

Figure 1.1: A schematic picture of solubility.

ties of the investigated solutes in different environments. These thermodynamic properties are determined by the structure of the molecule where the size of the molecule, number of hydrogen donors and acceptors, and hydrophilicity of the molecule are considered to be important factors.

In this thesis we have focused on parameters and properties that have impact in solubility models from an atomistic point of view. The next chapter will give a short background of the development of solubility models and discuss some of the present ones. In chapter three the theories behind solubility will be introduced. The fourth chapter is dedicated to some of the computational tools that are used to determine properties that are important for the solubility. In the final chapters a survey of obtained results is presented.

Chapter 2

Background: Solubility models

Since the aqueous solubility is an important property, an increasing number of methods to predict solubility have been developed during the last decades. Such development is illustrated in terms of the number of publications in the field in Figure 2.1. These models vary in how the solubility is calculated/predicted. Some of the models are empirical and efficient, others use discrete calculations of the investigated systems, and are therefore slower and not suitable for screening large compound collections. The target log(S) range is -1 to -5 log units in the development of new drugs. An accurate model to predict log(S) is therefore desired since the uncertainty of the modeled log(S) value must be small to be sure that the investigated compound is useful as a drug.

2.1 Experimental accuracy

Since the accuracy of the solubility models is always evaluated by comparing calculated values to experimental data, a short discussion of the accuracy of the experimental data is relevant. The accuracy of measured aqueous solubilities was investigated by Kishi and Hashimoto[2] who focussed on the aqueous solubility of antracene measured by 17 different laboratories using the same protocol. The largest variation in the measured solubility was 0.85 log units with the standard deviation of 0.19 log units. For the tabulated experimental data the measurement procedure is in general not the same and a larger variation in measured log(S) is therefore expected. When looking at molecules that tend to ionize, the pH dependence of the solubility must also be taken into account since most of the computational models deal with intrinsic solubility. Jorgensen and Duffy[3] discussed the accuracy of experimental data and concluded that Quantitative Structure-Property Relationship (QSPR) models can not be more accurate than the experimental uncertainty of 0.6 log units. It

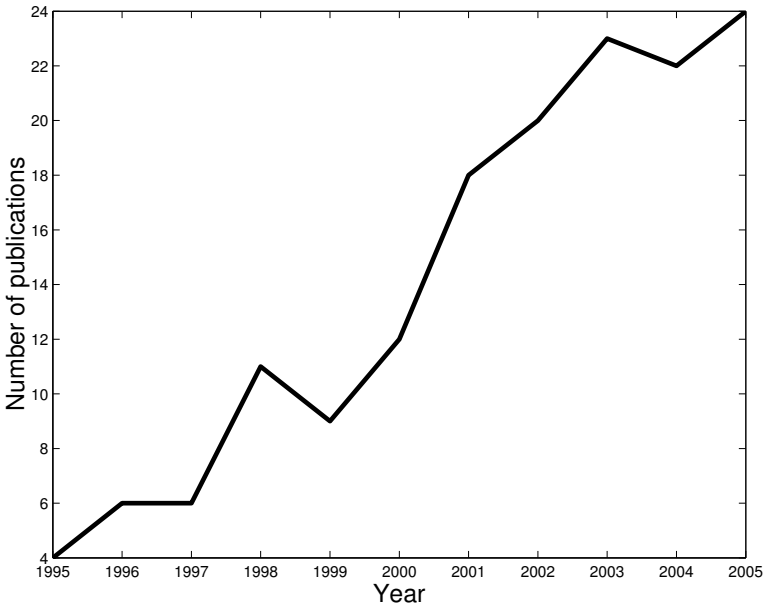

Figure 2.1: Number of methodological publications per year concerning aqueous solubility predictions between 1995-2005. Taken from Konstantin et. al.[1]

should also be mentioned that the traditional approach to measure solubility is time consuming (tens of hours per compound) and requires at least 1-2 mg of sample. The expensive and time consuming synthetical procedure is thus the main drawback of the experimental determination of the solubility for new drugs. Accurate computational models would be an advantage for the modification of existing structures that are known to be potential drugs but have low solubility. Coates et al.[4] have performed very long solubility experiments on sparingly soluble organic liquids and have found that the shake flask method can overestimate the solubility. For such compounds a measurement for at least 100 days is required to achieve a steady state solubility. The error between short and long experiments could be in the order of 10-100 times larger than the "true" value.

2.2 Empirical models

Hanch and coworkers showed in the late sixties that there is a linear relation between octanol/water partition and water solubility of liquids.[5] The octanol/water partition coefficient log(P) is easier to determine experimentally than the solubility and the relation still

has importance for the development of new models.

One experimental model of Yalkowski, called as the General Solubility Equation (GSE),[6,7] uses the octanol/water partition coefficient $log(P)$ to describe the difference between liquid phase and water solution. The energy required to go from solid to liquid phase is related to the melting temperature t_m. The resulting equation is as follows:

$$log(S) = 0.5 - log(P) - 0.01(t_m - 25). \tag{2.1}$$

However, the GSE is not valid for compounds with very high melting point nor for compounds with too high or too low log(P) value. The accuracy of GSE for drug like compounds has also been questioned. Even if the GSE would have been an accurate method one would still need experimental data which take almost the same amount of work as measuring the solubility directly. There are many models to predict the log(P) value with different accuracy.[8–11] The melting temperature is even more difficult to predict and there are also fewer existing models that can predict this property. Even though the empirical models described above are not suitable for new compounds they have been important for the understanding of the solubility.

2.3 Computational models

Today there are several different methods that have been developed to predict the solubility, and which are summarized in different review articles.[1,3,12,13] Delaney[12] has recently compared present models. He classifies the models based on the choice of regression/modeling method and descriptors. A similar classification has also been used in the review of Jorgensen and Duffy.[3] Four regression methods, namely the Artificial Neural Network (ANN), the Linear Regression(LR), the Group contribution and the Artificial Intelligence (AI), are presently available.

The group contribution method is a fast and conceptually simple method. In this model, the molecular structure in 2-D or 3-D format is screened for fragment a_i and a count of the occurrence of this fragment (n_i) is performed. The solubility can then be calculated using equation 2.2

$$logS = \sum_i a_i n_i + a_0, \tag{2.2}$$

The model is optimized with respect to contributions of different fragments to the model using regression analysis. The only descriptor in this method is the structural fragment.

Examples of this model are given by Küne et. al[14] which is built on 694 organic nonelelectrolyte solids and liquids. The absolute average error for their data set was 0.4-0.5 log units. However, the lack of polyfunctional molecules in their data set has made this model less suitable for drug molecules. Another more recent group contribution model is developed by Klopman et al.[15] who uses a data set of 1168 organic molecules. The standard deviation (rms error) for their model was 0.79 log units. It is noted that in this study most of the investigated molecules were classic organic molecules with only one functional group.

It should also be mentioned that for predicting log(P) there exists a quite accurate group contribution model proposed by Leo and Hanch.[8] This model is also computationally fast. However, a huge set of experimental values is required to build such a model. For the octanol-water systems there exist about 20 000 experimental data entries which are generally considered to be more accurate than aqueous solubility data.[16] The model developed by Leo and Hanch is considered to be one of the most accurate models for predicting log(P).

Another commonly used approach to predict the aqueous solubility is to use Multiple Linear Regression (MLR). In this QSPR method the equation used is similar to the group contribution method and is formulated as

$$logS = \sum_i a_i c_i + a_0 \tag{2.3}$$

where index i refers to descriptor, c a value for the investigated structure and a a coefficient that will be determined by the regression method. The descriptors are calculated from the investigated structures and differ from model to model. Examples of descriptors that are used for these models are: molecular weight, solvent-accessible surface area (SASA), molecular volume, counts of functional groups, hydrogen bonding acceptors and donors (HBAC,HBDN). The descriptors can be obtained from different methods, such as fragment counts, molecular or quantum mechanical calculations, and Monte Carlo or molecular dynamics (MD) simulations.

The calculated descriptors are usually dependent on which program that has been used to generate them. An example is the SASA whose value depends on the choice of the probe radius of the solvent. The 3-D structure that is used for calculating descriptors also affects the values of the generated descriptors. To make things even more complicated, after the generation of the descriptors one has to perform some kind of regression analysis to fit coefficients. There are different regression methods that can be used for this fitting procedure. Here the Partial Least Square (PLS) method will be described briefly. The PLS method calculates one component at time as long as the new components makes a difference to the model. The difference is controlled by a cross validation after each added component. Consider a variable matrix X consisting of predictors and a matrix Y consisting of response

variables. The variables in the X matrix are projected to the line that best describes the matrix. The same procedure is carried out for the Y matrix, however, the predictions of those lines are also dependent on the correlation between the lines. When calculating the new components the original variables are projected to new vectors and stored as new variables called scores. The scores are linear combinations of the original variables. The outcome of this is that the PLS method finds the coefficients for the contributions of the original variables. The number of components that should be used in the model is determined by leave-one-out cross-validation approach. In the cross validation each observation is removed at one time and the rest of the observations are used to create the model. The removed observation is then predicted by the model. If the sum of the difference between the predicted values and the given values is smaller than that for the previous component, the component is used and a new component is then calculated. Otherwise no further components will be calculated.

There have been several studies predicting the water solubility using MLR methods.[3,17–19] Jurs et. al.[20–22] have developed both MLR and Neural Network (NN) models. They use a set of 200 topological, geometric and electronic descriptors obtained from the AM1 and PM3 semi-empirical quantum chemistry calculations. For the MLR model the rms was 0.72 log units and for the test set 0.80 log units. Their data set contains mostly classical organic molecules. Huskonen[18] has also developed MLR and NN models. He used a huge data set of 1297 molecules of which 413 have randomly been removed to make a test set. The final model used 30 descriptors. The MLR model of Huskonen has given an rms of 0.67 log units for the training set and 0.71 log units for the test set. A comparison with a "benchmark" data[23] set (consisting of 21 common compounds tested in many models) gave an rms of 0.88 log units for the two training sets he used.

Jorgensen and Duffy have developed models to predict log(P) and log(S).[3,25] Two kinds of log(S) models were proposed. The first is a Monte Carlo method for solutes dissolved in water. Eleven descriptors were averaged from the MC simulation and used in the MLR. The resulting model has an rms error of 0.72 log units and uses 5 descriptors. It should be noted that the employed MC simulations are computationally time consuming and it makes the model less suitable for screening of large compound libraries. Their second model is a QSPR model based on the results of the MC simulations. In the QSPR model the MC properties have been replaced by similar properties that are computationally faster to generate. The final model has an rms of 0.90 for the used test set.

Another model, developed by Klamt et al., is referred to as COSMO-RS.[17] This model differs from the methods where many quickly computed descriptors have been generated. The COSMO-RS approach calculates the chemical potential in the solute and solvent. This is done by embedding the solute in conductors describing the solvent and the solute and

integrating the polarization charge densities over the surfaces. However, this method also includes descriptors for volume, chemical potential in water, and number of ring atoms to describe the free energy of fusion. According to the author of the COSMO-RS model the error should be 0.66 log units, although a comparison by Delaney[12] indicated that the error for common compounds (13 of 21 compounds in the "benchmark" data set) was 0.91 log units. The advantage of this model is the possibility of using different pH, salt concentration and different solvents. A drawback is that the model is considerably slower than models built on quickly generated descriptors. However, it should also be mentioned that the obscure descriptors in the MLR models are often unfamiliar to medicinal chemists.

Another commonly used method to predict the solubility is the Neural Network (NN) method. The major difference between MLR and NN is that NN introduces non-linear terms for the descriptors. NN is often used in combination with MLR to reduce the number of descriptors needed. The introduction of non-linear terms seems to have a good impact when using a large data set. There is a risk of so-called overtraining when using NN. Some of the developed models seem to obtain better accuracy than the experiment. The physical explanation of the non-linear terms is also often missing in the published models. There are several examples of prediction models for aqueous solubility which use NN. We will briefly describe three of them.

Huskonen has constructed an NN model to his MLR model.[18] The NN model uses the same set of structural parameters as for the MLR model. For the NN model the rms was 0.47 log units for the training set and 0.60 log units for the test set. This model was also compared to the "benchmark" data set and the rms then became 0.63 log units. Jurs et al. have developed two NN models.[21] The first model uses the final 10 MLR descriptors and obtains an rms of 0.88 and 0.50 log units for the training set and the test set, respectively. Another NN model of them uses all 100 initial descriptors and leads to a rms of 0.88 log units for the training set and 0.51 log units for the test set. The NN method of Tetko et al[26] has many similarities with the model developed by Huskonen. The difference is that this model uses fewer descriptors and three times fewer hidden neurons. According to the authors the model should therefore be more robust. Indeed, the rms was 0.62, 0.60 and 0.64 log units for the training set, the test set and the "benchmark" data set, respectively.

There has also been at least one study where Artificial Intelligence (AI) has been used to predict the aqueous solubility.[27] The result of that study is comparable with the MLR and NN studies. The rms for the "benchmark" data set was 0.82 log units. However, only 11 of the 21 molecules in the "benchmark" test set were present.

To summarize the computational models available today, one can say that the NN models achieve a better rms than that of the MLR models. The drawback of the NN models is the

risk of overtraining and it is difficult or even impossible to modify the molecule depended results due to the lack of connection to real physical properties. One should also be aware of that the accuracy of a QSAR model can not exceed the experimental accuracy. It would therefore be an advantage if a more accurate data set with a diverse set of drug molecules could be developed. An attempt of doing this has been carried out by Bergström et al. who produced a data set of 85 molecules[19]and made all measurements under the same conditions. However, to develop accurate QSAR models a much larger data set is required.

Solubility Theories

The solvation is a thermodynamic process and can therefore be formulated with thermodynamics and statistical mechanics. One of the key thermodynamic quantities for solvation is the chemical potential. The chemical potential of a component in a system s is defined as

$$\mu_s = \left(\frac{\partial G}{\partial N_s}\right)_{P,T,N'} = \left(\frac{\partial A}{\partial N_s}\right)_{T,V,N'}, \tag{3.1}$$

where G is the Gibbs energy of the system, A the Helmholtz energy, P the pressure, T the absolute temperature, V the volume, and N' the number of molecules in the system except the N_s molecules. For a two components system the chemical potential is defined as

$$\mu_A = G(T, P, N_A + 1, N_B) - G(T, P, N_A, N_B), \tag{3.2}$$

which means the change in Gibbs energy of adding one molecule of A to the system while the temperature, pressure and number of components of B are constant.

The solvation process is here defined as if one molecule s is taken from phase α into phase β. This process is carried out under constant pressure P and temperature T. The Gibbs energy of solvation of s from phase α into phase β then becomes

$$\Delta G_s = \mu_s^\beta - \mu_s^\alpha. \tag{3.3}$$

To relate the aqueous solubility to free energy terms there are different relations that can be considered. The first relation is when a liquid compound is in equilibrium with its vapor:

$$A(g) \leftrightarrow A(l), \tag{3.4}$$

where the g and l indicates the gas phase and the liquid state, respectively. The free energy of vaporization can then be calculated from

$$\Delta G_{l\to g} = RTln\frac{P_A}{PM_A^l}, \tag{3.5}$$

where P_A is the vapor pressure for species A over its pure liquid and P is the reference pressure, M_A^l the equilibrium molarity of liquid A.

The second relation is the equilibrium between sample A in liquid and dissolved in aqueous solution:

$$A(l) \leftrightarrow A(aq). \tag{3.6}$$

The free energy for this process is defined as:

$$\Delta G_{l\rightarrow aq} = -RTln\frac{M_A^{aq}}{M_A^l}, \tag{3.7}$$

where the M_A^{aq} is the equilibrium aqueous molarity of solute A, i.e the solubility S. Combining equation 3.4 and 3.6 gives

$$A(g) \leftrightarrow A(aq), \tag{3.8}$$

which is usually called *free energy of hydration* and is the equilibrium between aqueous solution and the vapor of sample A. The free energy of hydration can be defined by combining equations 3.5 and 3.7, which leads to

$$\Delta G_{g\rightarrow aq} = RTln\frac{P_A}{PM_A^{aq}}. \tag{3.9}$$

When the solute starts from the solid phase equation 3.4 and 3.6 become

$$A(g) \leftrightarrow A(s) \tag{3.10}$$

and

$$A(s) \leftrightarrow A(aq) \tag{3.11}$$

respectively. The free energy relations for equations 3.4 and 3.6 then become

$$\Delta G_{s\rightarrow aq} = RTln\frac{P_A}{PM_A^s} \tag{3.12}$$

and

$$\Delta G_{s\rightarrow aq} = -RTln\frac{M_A^{aq}}{M_A^s}. \tag{3.13}$$

These relations are the same as for liquids except that P_A corresponds to pure substance vapor pressure of solid A and M_A^s is the molarity of solid A. From equation 3.13 the solubility can be written as

$$S = M^{aq} = M_A^s e^{(\frac{\Delta G(s\rightarrow aq)}{RT})}. \tag{3.14}$$

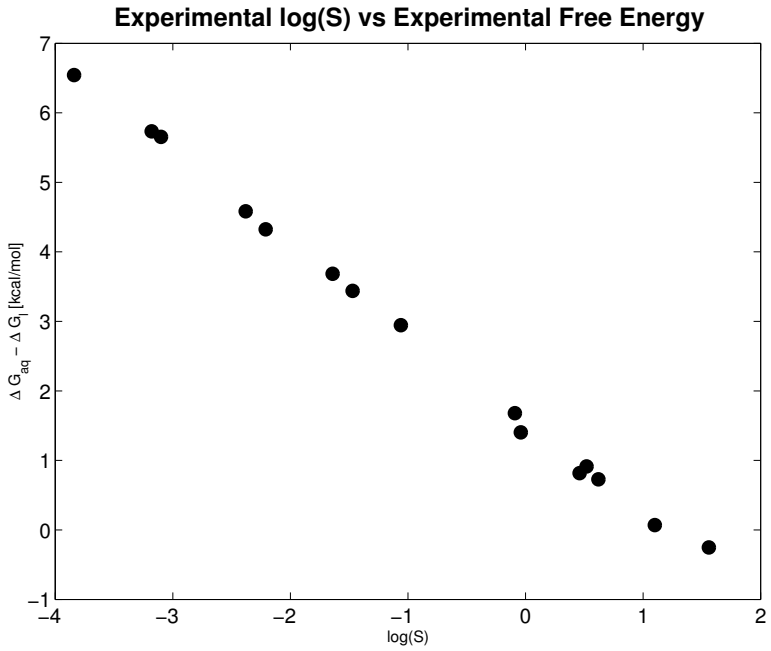

Figure 3.1: Difference in free energy of hydration and vaporization as function of the aqueous solubility log(S).

If one considers the difference between free energy of hydration and free energy of vaporization $\Delta\Delta G$ it is evident that its relation to solubility is linear as clearly demonstrated in Figure 3.1 where experimental results for 15 small organic molecules have been plotted.

However, to be able to solve equation 3.14 the chemical potential must be determined and this is dependent on Helmholtz energy, defined as

$$A = -k_B T ln Z, \tag{3.15}$$

where Z is the partition function, which is far from a trivial task to calculate for systems relevant in this thesis. There are computational approaches to deal with this problem such as Monte Carlo simulations. However, such approaches require simplifications in the way of treating the partition functions which may affect the result.

Instead of looking at the derivative of the Helmholtz energy one may look at the derivative of the Gibbs energy. This relation is more easy to relate to experimental results since it is easier to control constant pressure than constant volume. The free energy of hydration $\Delta G_{g\rightarrow aq}$ and free energy of vaporization $\Delta G_{l\rightarrow g}$ can experimentally be obtained from

$$\Delta G_{free} = kTln\left(\frac{\rho^{\alpha}}{\rho^{\beta}}\right). \tag{3.16}$$

where ρ^{α} and ρ^{β} are the densities in the different phases.

Since the chemical potential is difficult to calculate the free energy is usually related to enthalpy H and entropy S as:

$$\Delta G = \Delta H - T\Delta S, \tag{3.17}$$

where T is the temperature.

This relation has been widely used in solubility describing models. The enthalpy contribution to the solubility comes from several solvent-solvent, solute-solute, and solute-solvent interactions, such as breaking of solute-solute bonds (positive enthalpy), breaking of solvent-solvent bonds in the formation of cavity solvent (positive enthalpy), restructuring of water (negative enthalpy), and formation of solvent-solute bonds (negative enthalpy).

When looking at the entropy term there are three components that contribute, namely the construction of a cavity in the solvent (negative entropy), the "iceberg formation" of the solvent around the solute (negative entropy), and the mixing of the two substances (positive entropy).

The weights of the entropy and enthalpy terms depend on the molecule. There are several properties of the molecule that determine its solubility. The size of the molecule is an important factor and is widely used in solubility models. For models described in Chapter

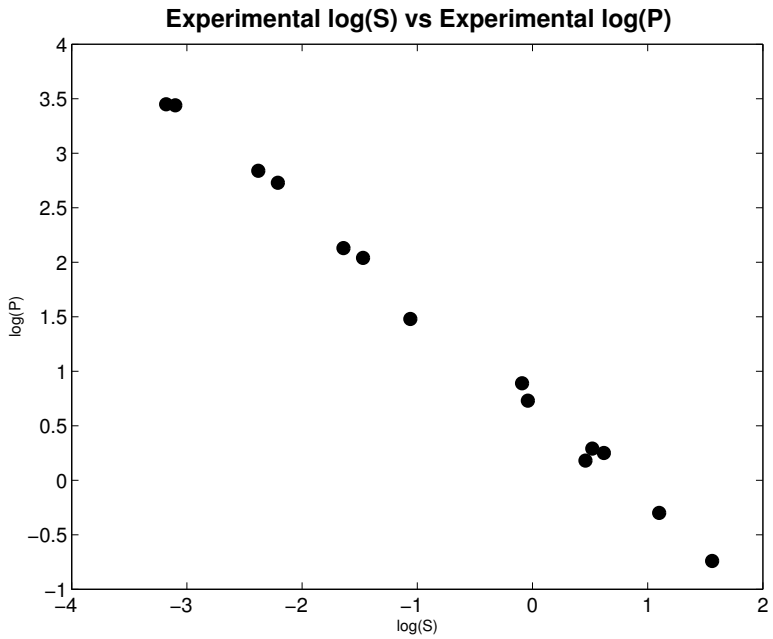

Figure 3.2: Octanol/water log(P) partition v.s. solubility log(S).

2 the molecular weight or the volume of the molecule has been included in most of the models. It is also important to know if the compound is polar or not. A molecule that can form many hydrogen bonds to the environment is significantly more soluble than molecules that cannot. The aromaticity of the molecule has also an impact on the solubility. For instance, a molecule with many double bonds is more soluble than a molecule with fewer double bonds.

In Chapter 2 it was claimed that the octanol/water partition log(P) has a linear relation to log(S) for liquids. To verify this statement the same compounds that have been used in Figure 3.1 are here plotted to relate their log(S) and log(P).

As seen in Figure 3.2 the relation is as linear as for the relation between free energy and log(S). This relation justifies the log(P) term in the General Solubility Equation and other models using the log(P) term to determine the energy change of going from liquid state to aqueous solution.

Solvation models used for calculating the free energy of solvation usually split the free energy in different components as for example

$$\Delta G_{free} = \Delta G_{elec} + \Delta G_{vdw} + \Delta G_{cav} + \Delta G_{hb}. \tag{3.18}$$

ΔG_{elec} is here the electrostatic contribution which is important for polar solutes due to the

polarization of the solvent. ΔG_{vdw} is the van der Waals interaction and is usually split into a repulsive ΔG_{rep} and an attractive dispersion term ΔG_{disp}. ΔG_{cav} is the free energy to form the cavity for the solute in the solvent. The last term ΔG_{hb} is sometimes used for systems where there exist hydrogen bonds between the solute and the solvent. More discussion about the different energy terms will be found in Chapter 6. In this thesis some of the above components will be examined and studied in detail.

Computational methods

The calculation of molecular properties in the condensed phase is a big challenge since the number of involved molecules is huge. There are two fundamentally different ways to deal with the solvent effects in theoretical modelling, using either a continuum or explicit molecules to represent the solvent.

4.1 Dielectric continuum models

In the continuum models the solvent is described by a homogeneous dielectric continuum which is dependent on its dielectric constant. There are different kinds of continuum models depending on the choice of the cavity shape. The well-known Onsager reaction field model uses a spherical or ellipsoidal cavity whereas the Polarizable Continuum Model (PCM) uses a molecular shaped cavity. The PCM has been used in this thesis and will therefore be described in more detail.

With the polarizable continuum model, like other apparent surface charge models, it is possible to investigate the molecular properties in solution. It is also possible to calculate the free energy difference between a molecule in gas phase and in a liquid solvent. In PCM the molecule is surrounded by a cavity with a molecular shape. The free energy of solvation for the PCM solvated molecule is defined as:

$$\Delta G_{sol} = \Delta G_{el} + \Delta G_{cav} + \Delta G_{disp} + \Delta G_{rep} \tag{4.1}$$

Where the ΔG_{el} is the electrostatic contribution, G_{cav} the work needed to form the cavity, G_{disp} the short range solute-solvent interactions and G_{rep} the short range solute-solvent repulsive forces. The electrostatic term collects the electrostatic effects and is defined as:

$$G_{el} = \langle \Psi | H(\Psi) - 1/2V(\Psi) | \Psi \rangle \,, \tag{4.2}$$

where the solute wavefunction is determined by the Schrödinger equation $H(\Psi)\Psi = E\Psi$. The Hamiltonian $H(\Psi)$ is

$$H(\Psi) = H^0 + V(\Psi). \tag{4.3}$$

Where H^0 is the Hamiltonian for an isolated molecule and $V(\Psi)$ the mean solute-solvent interaction potential which is defined as

$$V(x) = V_M(x) + V_\sigma(x) = \int_{\mathbb{R}^3} \frac{\rho_M(y)}{|x-y|} dy + \int_\Sigma \frac{\sigma(s)}{|x-s|} ds, \tag{4.4}$$

where ρ_M and σ are two electrostatic potentials and Σ the interface. When the equation above has been defined the problem consists of screening apparent surface charge density $\sigma(s)$. The cavity contribution to equation 4.1 is only dependent on the shape of the cavity and is defined as:

$$G_{cav} = \sum_i^{spheres} \frac{A_i}{4\pi R_i^2} G_i^{HS}, \tag{4.5}$$

where R_i is the radius of the sphere, $G_i^{HS}$is the cavitation energy for a sphere of radius R_i, and A_i is the area of the portion of the sphere i. In equation 4.1 there is also a dispersion contribution which is calculated as:

$$G_{dis} = \frac{1}{\pi} \int_0^\infty d\omega \sum_{K\neq 0} \frac{\omega_{0K}^M}{(\omega_{0K}^M)^2 + \omega^2} \int dr_1$$
$$\times \int_\Sigma \frac{dr_2}{r_{12}} P_M(0K|r_1) \sigma_S[\epsilon(i\omega), P_M(0K|r); r_2], \tag{4.6}$$

where $P_M(0K|r)$ and ω_{0K}^M are, respectively, the transition densities and energies for solute M (for transition from ground state (0) to excited state K) and σ_S is the surface charge density induced in the solvent by the electric field of the charge distribution $P_M(0K|r)$. $\epsilon(i\omega)$ is the calculated dielectric constant at imaginary frequencies. The repulsion term in equation 4.1 is defined as:

$$G_{rep} = \alpha \int_{\mathbf{r}\notin\mathbf{C}} dr P(\mathbf{r}), \tag{4.7}$$

where α is a constant defined by properties of the solvent, $P(\mathbf{r})$ is the solute electronic charge distribution, and C the cavity domain.

4.2 Discrete models

In the discrete models all the solvent molecules are described explicitly. Examples of discrete models are:

- Monte Carlo: The probability of a certain configuration is determined by the Boltzmann factor. This method is suitable for generation of configurations for investigations of properties that are not dependent on time.
- Quantum Mechanics/Molecular Mechanics (QM/MM): In this method quantum mechanics is used for the central part of the system and the surroundings are treated with classical molecular mechanics.
- First-Principles Molecular Dynamics: All the molecules are treated quantum mechanically. The gradient and energy are calculated in each step to determine the dynamics of the investigated system.
- Molecular Dynamics: (Here meaning using classical force fields.) This method has been used in this thesis and will therefore be described in more detail later.
- Supermolecular Approach: The molecule under investigation and a few other molecules that are assumed to interact with it are included in the system and calculated quantum mechanically.

4.2.1 Molecular dynamics

To deal with both structural and dynamic properties, Molecular Dynamic (MD) simulations is a commonly used approach since it is possible to do time-dependent calculations within a decent time. MD simulations solve Newton's equations of motion for a system of N atoms

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i}, \tag{4.8}$$

where i = 1...N and the forces are the negative derivative of the potential function V.

$$F_i(t) = -\frac{\partial V}{\partial r_i}. \tag{4.9}$$

Since it is only computationally possible to calculate a finite system, periodic boundary conditions are usually introduced when performing MD simulations to create a virtual infinite system. When a molecule in the periodic system leaves a box at one side it will enter the box at the opposite side. The use of a finite system makes it necessary to include a correction for the long range inter-molecular interactions that may affect the behavior of other molecules beyond the size of the unit cell. There are few methods can be used to obtain these corrections, such as Ewald summation, group based truncation and atom based force shift.[30,31]

In MD simulations different kinds of ensembles can be used. If the number of particles(N), the volume of the system(V), and the energy of the system(E) are kept constant the ensemble is called NVE. When the number of particles(N), the volume(V), and the temperature(T) are kept constant a NVT ensemble is formed. In the case when the number of particles(N), the pressure(P), and the temperature(T) are kept constant the ensemble is named as NPT. In this thesis only the NPT ensemble has been employed.

Force fields play the essential role in MD simulations. In this thesis the so-called *general amber force field parameters*[32] (GAFF) and TIP3P[33] water force field parameters have been used. The potential function defined in the AMBER program[34] has the form

$$\begin{aligned} U(R) &= \sum_{bonds} K_r \left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_{eq}\right)^2 + \\ &\sum_{dihedrals} \frac{V_n}{2} \left(1 + cos[n\phi - \gamma]\right) + \sum_{i<j}^{atoms} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \sum_{i<j}^{atoms} \frac{q_i q_j}{\epsilon R_{ij}} \end{aligned} \tag{4.10}$$

where r is the bond length, θ the bond angle, ϕ the dihedral angle, n the multiplicity and δ the phase. K_r and K_θ the force constants for bond and angle respectively. A and B are atom dependent Lennard Jones parameters, q the atomic charges and ϵ the dielectric constant.

4.2.2 Linear scaling Quantum Chemistry

Due to the fast growth of computer power in the last decades it has become possible to treat larger and larger systems quantum mechanically. However, as the system becomes larger the computer programs meet new challenges and need to be modified to treat such large systems efficiently. Here we present a short description of the main features of a linear scaling program that has been used in this thesis. The basic concepts of modern quantum chemical methods can be found in textbooks, for instance the one entitled *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory.*[35] Three key computational techniques that are important for this thesis, namely the construction of the Fock matrix, sparse matrix storage and solving the generalized eigenvalue problem, will be discussed below.

The construction of the Fock matrix is the most time consuming part in a quantum chemical calculation where the most time is spent on calculating the two electron integrals. The Fock matrix is defined as

$$\mathbf{F} = \mathbf{H_{core}} + \mathbf{G} \tag{4.11}$$

where the $\mathbf{H_{core}}$ is the sum of a kinetic term and an electron-nuclear attraction term. The

$\mathbf{H_{core}}$ is quite fast to calculate even for large systems and it is only needed to calculate it once for a Hartree-Fock (HF) or density functional theory (DFT) calculation. The second term in the Fock matrix is much more important to improve since it takes most of computational time. This term is defined as

$$\mathbf{G} = \mathbf{J} + \mathbf{K} \tag{4.12}$$

where $\mathbf{J}$ is the coulomb matrix and $\mathbf{K}$ is the exchange matrix. The elements in these matrices are computed as follows:

$$J_{pq} = \sum_{rs} D_{rs}(pq|rs) \tag{4.13}$$

$$K_{pq} = -\frac{1}{2}\sum_{rs} D_{rs}(pr|sq) \tag{4.14}$$

The two electron integral part, $(ab|cd)$, in these equations is defined as:

$$(pq|rs) = \int \frac{b_p(\mathbf{r_1})b_q(\mathbf{r_1})b_r(\mathbf{r_2})b_s(\mathbf{r_2})}{|\mathbf{r_1} - \mathbf{r_2}|} d\mathbf{r_1} d\mathbf{r_2} \tag{4.15}$$

The number of these two-electron integrals is proportional to n^4 where n is the number of basis functions. If one takes into account the symmetry the number of integrals can be reduced to $\frac{1}{8}n^4$. For large systems some of the integrals will also be neglected as they have a very small contribution. This gives an n^2 scaling factor which is usually the case for the conventional quantum chemical codes. Unfortunately these improvements are not sufficient to allow a computer program to treat large system in a descent amount of time, therefore more improvements have to be made.

In Figure 4.1 a schematic draw of interaction regions for Coulomb and Exchange integrals is shown. It can be seen that the Coulomb interaction spans over the whole molecule whereas the exchange interaction is more local. For the long range Coulomb interaction the integrals can be substituted by the Fast Multipole Method (FMM). FMM was originally developed for computing Coulomb interactions among classical point charges, and has been adopted to quantum chemistry programs. The idea of FMM is that charges close to each other can be expressed as multipole expansions. The calculated multipole moments are then used instead of calculating all integrals for long ranges. When it comes to the exchange matrix the same approximations are difficult to be adopted although in this matrix the interaction range that has an impact on the energy is smaller than that for the Coulomb matrix. Therefore it is bearable to calculate this matrix only using truncations. A more thorough explanation can be found in Rudbergs licentiate thesis.[44]

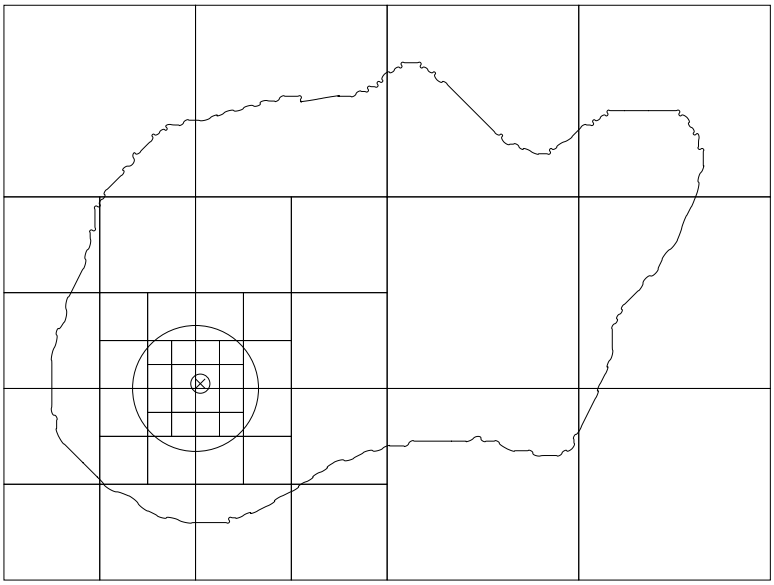

Figure 4.1: Schematic draw of FMM and exchange region in Linear scaling QM programs. FMM hierarchy and exchange interaction region are marked with boxes and large circles.

When performing QM calculations on small systems the matrices where the data is stored and manipulated mostly consist of non-negligible elements. However, when the system becomes larger an increasing number of negligible elements can be found. It is thus possible to reduce memory usage and computational time if only non-negligible elements are stored and manipulated. The traditional way to implement sparse matrix storage is to have one vector for the values in the matrix, one for the number of entries in each row and another for the position in each row. This kind of sparse matrix implementation is not the best for quantum chemical codes since the occurring matrices are not sparse enough to use a data structure with that much addressing overhead. A better approach for quantum chemistry codes is to use block-sparse matrices. Block sparse matrix storage is preferable since the highly optimized Linear Algebra packages can be used for the small full block matrices and the remaining parts of the matrix with negligible elements do not need to be calculated.

The time to solve the generalized eigenvalue problem in the Roothan Hall equation grows cubically with the system size. It is therefore important to use another scheme to improve this for large systems. The common way to get the density matrix D is to solve the generalized eigenvalue problem.

$$\mathbf{FC} = \mathbf{SC}\epsilon \tag{4.16}$$

which gives the eigenvectors $\mathbf{C}$ and eigenvalues ϵ. The density matrix can then be con-

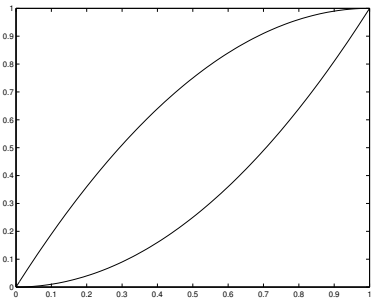

Figure 4.2: Trace purification functions x^2 and $x - x^2$

structed from the following equation

$$D_{pq} = 2 \sum_{a=1}^{\frac{N}{2}} C_{pa} C_{qa} \tag{4.17}$$

An alternative approach is to not calculate the eigenvectors. This can be done because the eigenvectors of $\mathbf{D}$ and $\mathbf{F}$ are the same. In this scheme the $\mathbf{F}$ matrix is transformed into $\mathbf{F_{ort}}$ by multiplying the inverse Cholesky decomposition of $\mathbf{S}$. The purification method is then applied to the $\mathbf{F_{ort}}$ matrix where one of the functions x^2 and $x - x^2$ (see Figure) is applied depending on if the trace of D_{ort} is larger or smaller than the number of occupied orbitals. When the trace is equal to the number of occupied orbitals the D_{ort} matrix is multiplied by an inverse Cholesky decomposition of $\mathbf{S}$ to form the density matrix D. A deeper description of the handling and storage of matrices can be found in the licentiate thesis of Rubensson.[45]

All these improvements have been implemented in the linear scaling quantum chemistry code *ergo*. All quantum chemical calculations in papers V and VI have been carried out using this program.

Before ending this Chapter it should be mentioned that one can also combine the discrete models with the continuum models. In this case, one can describe the solvent molecules close to the solute molecule with a discrete model and the surrounding with a continuum model.

Chapter 5

Cavitation energy

When investigating aqueous solubility, the work of introducing the solute into the aqueous solvent is usually split into two different parts: soft and hard. The soft part is referred to as the introduction of the solute into the solvent cavity while the hard part is associated with the formation of the cavity in the solvent. The cavity formation energy have previously been investigated by many authors mostly using free energy of perturbation (FEP)[36,37] and scaled particle theory.[38] In this thesis molecular dynamics simulations have been used to study the molecular volume and the results have been compared with that obtained from the GEPOL[39,40] algorithm, which is an algorithm for calculating the molecular surface area and volume, and experimental partial molar volumes. The following procedure is used in the MD simulations. The solute is dissolved in a water box containing around 3500 water molecules. The box is equilibrated for 200 ps and simulated during 400 ps. The box volume is recorded during the simulation and the average value is stored. The volume change of the system caused by the solute is then calculated by subtracting an average volume of a single water times the number of waters in the simulated box. This volume will be referred as V_{tot}^{MD}. We have also obtained the Voronoi volume of the solute from the same MD simulations. The Voronoi volume is calculated by making a 3-dimensional grid of the box and then from the center of each cell in the grid one checks if a solute or a solvent atom is the closest one to the cell. If the solute atom is the closest one, the cell volume is added to the total Voronoi volume, otherwise the cell volume remains unchanged. This volume is labeled as V_{Vor}^{MD}.

The calculated volumes from these two methods mentioned above and from the GEPOL[39,40] algorithm with PCM are shown in Figure 5.1 for comparison. The volumes obtained from different models are very similar for volumes below 350 Å^3. When looking at volumes larger than 350 Å^3 it is found the V_{tot}^{MD} gives a larger volume than the ones obtained with GEPOL algorithm. For V_{tot}^{MD} and V_{Vor}^{MD} one can see that for half of the investigated compounds the two models gives similar volumes. However, for the rest of the molecules V_{Vor}^{MD} becomes

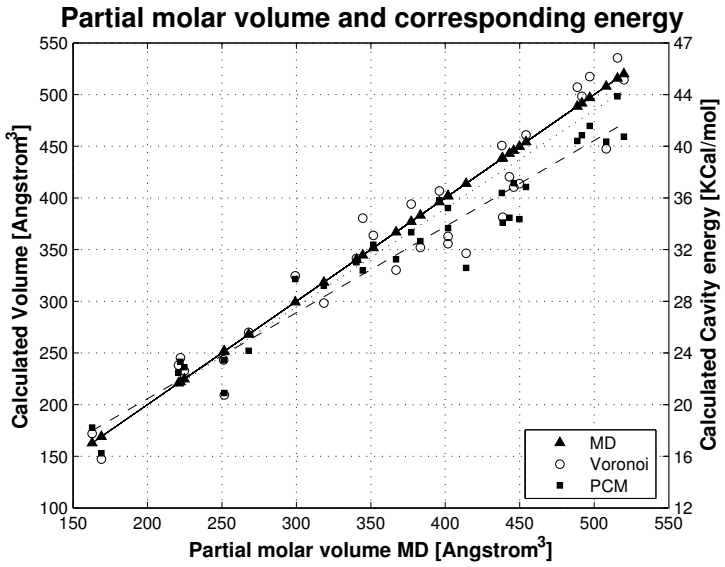

Figure 5.1: Calculated cavity and partial molar volumes and the corresponding formation energy.

similar to those obtained by the GEPOL algorithm. Such observations are not entirely unexpected since V_{tot}^{MD} includes the volume changes in the water layers around the molecule i.e. the iceberg effect, whereas the V_{Vor}^{MD} model does not. Since the volumes obtained by the V_{tot}^{MD} model include the whole volume change of the system these values should have best agreement with the experimental partial molar volume. A comparison between experimental values of partial molar volumes and the calculated volumes is given for a set of five small organic molecules in Table 5.1.

Molecule	V_{TOT}^{MD}	V_{Vor}^{MD}	V^{PCM}	V^{Exp}
Methanol	73.051	66.0636	62.865	63.351
2-propanol	124.673	125.349	110.872	119.230
Benzene	136.232	136.705	139.067	137.164
2-methyl-2-propanol	140.902	153.352	137.578	145.799
3-hexanol	194.092	213.246	205.803	194.454

Table 5.1: Calculated and experimental partial molar volumes for four alcohols and one benzene molecule.

It can be seen that the V_{tot}^{MD} model has the best agreement with the experimental values as expected since the iceberg effect is only included in that model. The relation between

the volume that the molecule occupies and the cavity energy has also been investigated. By performing a PCM calculation with the Gaussian program[41] a cavity energy can be obtained. An almost linear relation between the GEPOL cavity volume and the cavity energy is observed from these calculations. A least square fitting procedure has therefore been performed to relate the occupied volume to an energy. The related energy values obtained from different methods for 32 molecules have been included in Figure 5.1. One can see that the calculated energy is dependent on the computational model used. The energy difference for a molecule from different methods is in the worst cases around 5 Kcal/mol, which is only 10-15 percent of the total energy. However, such a small difference may have large impact on a solubility model where the total solvation energy is much lower than the cavitation energy. The relation between one log(S) unit and an energy unit in KCal/mol is roughly 1.

Chapter 6

Interaction energy

In Chapter 3 it was shown that the interaction energies between a solute and the solvent determines its solubility. Within this thesis the interaction energies have been studied by several different methods. In this Chapter some results of these investigations will be discussed. As mentioned in Chapter 3 the interaction energy between a drug molecule and its surrounding can be calculated direct or as a sum of different energy contributions. Both approaches have been employed here. At first we discuss the approach that sums up contributions of different important components,

- Hydrogen bonding energy: Many drug molecules tend to form hydrogen bonds with surroundings which are often the key property in their biological binding. Typical hydrogen bonding groups are -OH and -NH_2.
- Ionic attraction: Ions tend to have strong interactions to other molecules and molecules in ionic form usually are more soluble. However in this thesis the solubility of drug molecules are limited to the neutral form.
- Dipole dipole interaction: Substances with molecules possessing dipole moments have higher melting point or boiling point than those with similar molecular mass but no dipole moments. This is simply due to the stronger interaction between dipolar molecules.
- van der Waals/Dispersion energy: This is an energy term that arises for any substance. It is heavily dependent on the size of the molecule.
- Covalent bond: This is not really an intermolecular interaction, but rather an intramolecular interaction. It is mentioned here because some solids like diamond, silicon etc. form covalent bonds. This interaction is very strong.

There are also combinations of these interactions such as dipole induced dipole, ion - HB etc.

The different contributions described above are usually divided into two classes namely electrostatic and vdW interactions. The electrostatic interactions are sum of dipole-dipole hydrogen bonding and dipole-induced dipolar interactions, whereas the vdW interaction operates between induced dipoles. In MD methods the electrostatic and vdW terms are given in the non bonding terms, see equation 4.10. There, the first term corresponds to the vdW interactions and the second term to the electrostatic interactions. When looking at the Polarized Continuum model the electrostatic interaction is calculated separately and the vdW term is considered as a sum of dispersion and repulsive contributions. There are other continuum models that have an explicit term for hydrogen bonding groups[17] to compensate for the poor HB interaction energy given by the electrostatic contribution. In the super-molecular approach where the whole system is treated quantum mechanically, the electrostatic interaction energy can be calculated using different levels of theory. However, when dealing with the van der Waals interactions, a high level theoretical approach, such as the second order perturbation method (MP2), should be employed to obtain accurate van der Waals energy. In this thesis a variety of methods have been tested for calculating interaction energies.

6.1 Hydrogen bonding energy

Hydrogen bonds form strong intermolecular interactions. It is therefore an important property in solubility models. Today many aqueous solubility models based on QSPR use the number of hydrogen bonds for certain groups present in the drug molecule as the main descriptor. This descriptor is then multiplied by a weight in the regression step during the development of the model. We have studied the hydrogen bond energy and the impact of the hydrogen bondings on molecules in solution. One of our focuses is to examine the additive behavior of hydrogen bonding energy in drug molecules.

A set of small organic molecules with specific hydrogen bonding groups has been chosen (see paper I for details). A number of hydrogen bonded waters was added to the different polar groups in the molecule, for example two water molecules bonded to alcohol groups and three water molecules to the carboxyl acid group. The initial positions of the water molecules were selected visually and their equilibrium structures were obtained through geometry optimization. The hydrogen bonding energy was calculated using the following procedure: the total energy of the water-molecule complex was calculated, then the energies of the waters and the molecule were calculated separately. The hydrogen bonding energy was then

obtained by subtracting the water energy and the molecule energy from the water-molecule complex energy.

The validity of the additive approach was investigated for eight commonly used drug molecules. The additive hydrogen bonding energy was obtained by summing up the hydrogen bonding energy of each polar group in the drug molecule. It is shown that the validity of the additive model is strongly conditional, and to some extent predictable: in cases where the hydrogen bonding group is isolated the addition model can have relevance, while in cases where the hydrogen bonding groups are interconnected through π-conjugation rings or chains of the drug molecules it in general introduces substantial errors. It is found that in general the strong cooperative effects of hydrogen bonds should be taken into account for evaluation of the hydrogen bonding energies of drug molecules.

The impact of the hydrogen bonding on molecular structure and properties was investigated in paper II for three drug molecules dissolved in water by means of infrared and Raman spectra. Infrared (IR) spectroscopy is a powerful tool to detect the molecular structure in organic chemistry. The photons in infrared light have a wavelength between 0.78 and 1000 mm. IR radiation therefore lacks sufficient energy to cause electronic transitions in the molecule but may induce vibrational and rotational excitations of covalent bonds. IR spectroscopy works because different chemical bonds have specific frequencies at which they vibrate, corresponding to different energy levels. The frequencies are determined by the shape of potential energy surface of the molecule, the masses of the atoms and by associated vibronic coupling. Raman spectroscopy is also a valuable tool in the characterization of molecules and is a good compliment to IR spectroscopy. The Raman effect occurs when monochromatic light (usually laser) acts on the molecule and interacts with its electronic dipole moment. The photon excites one of the electrons into a virtual state even when the energy is not large enough to excite the electron into a full quantum state. Then almost immediately another photon is released and the molecule falls back into its lower state. However, when the electron relaxes it may fall back to a higher vibrational state. The outcome of this is that the excitation photon has a higher energy than the photon coming out of the molecule. This phenomenon is called a red shift or Stokes shift. Anti-Stokes shift is also possible but unusual. The energy shift measured in Raman spectroscopy gives a fingerprint for a specific molecule.

The computational procedure employed in this thesis involves several steps. A frequency calculation was performed for the molecules in gas phase and water solutions. The solvent effect has been described by different approaches: a dielectric continuum model, molecule-water complexes through hydrogen bonds, and complexes in dielectric continuum models. The infrared and Raman spectra of the molecule were in each case analyzed to understand the effects of long-range and short-range interactions on the vibrational frequencies and

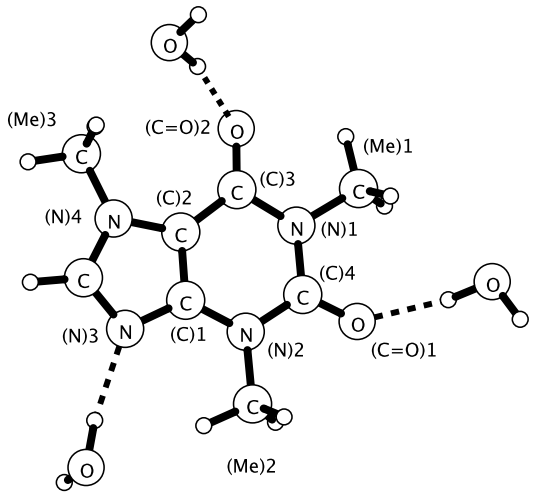

Figure 6.1: Caffeine molecule with hydrogen bonded waters

spectral intensities. In Figure 6.1, the caffeine molecule and water complex is presented. For this particular molecule three waters have been added in the calculations through explicit hydrogen bonds. The calculated frequencies have been divided into three regions for the caffeine molecule. In the high frequency region 3000-4000 cm^{-1} there is only a small IR absorbance in the frequency region 3250 cm^{-1} and lower, except for peaks related to the OH bonds of water that could be found when explicit water was used. There are several non-water related peaks for the Raman spectra in this region. The frequencies in this region are related to C-H stretching motions, whose intensities were slightly lowered by using dielectric continuum models but almost unaffected with the inclusion of explicit hydrogen bonds. Most of the observable peaks for the IR spectra were found in the middle region i.e. at frequencies between 900-1800 cm^{-1}. The IR spectra of this region are shown in Figure 6.2.

The intensities for the Raman spectra in this region were smaller than for the high frequency region. The highest frequency in this region corresponds to a C=O stretching and pyrimidine ring bending and the calculated gas phase value for this peak was 1752 cm^{-1}. When using PCM the frequency was decreased by 17 cm^{-1}, and for the complex with explicit

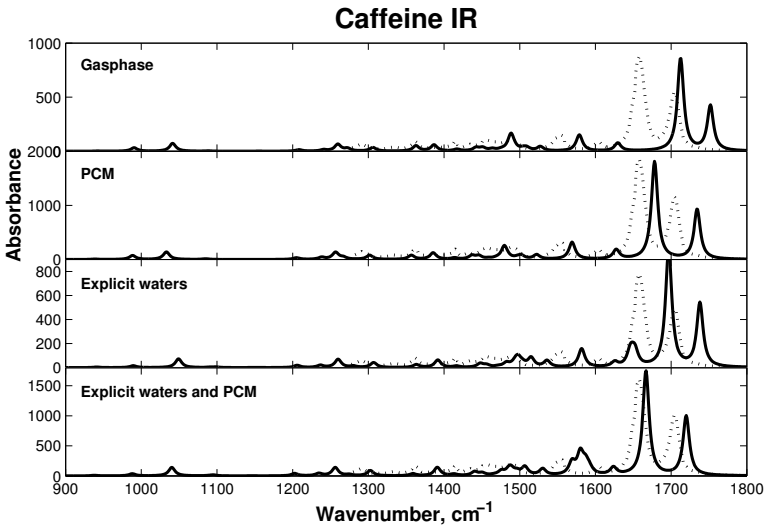

Figure 6.2: IR spectra of caffeine in middle frequency region. Solid line: calculations. Dotted line: experimental results from Ohnsmann et al.[42]

hydrogen bonding the frequency was decreased by 14 and 32 cm^{-1} without and with PCM, respectively. The calculated frequency, with both explicit hydrogen bonds and PCM, has the best agreement with the experimental value obtained by Ohnsmann et al[42] which was 1720cm^{-1}. The frequency with the highest absorbance in the caffeine molecule corresponds to C=O bending and pyrimidine ring bending. The calculated gas phase frequency for this peak was 1713 cm^{-1} and is decreased by 34 cm^{-1} when PCM is added. The calculations including hydrogen bonds without and with PCM decreases the energy by 16 and 46 cm^{-1}. The highest peak in the study by Ohnsmann et al[42] has a frequency of 1659 cm^{-1} which is also in good agreement with the value of 1667 cm^{-1} calculated using explicit hydrogen bondings and PCM. In the low frequency region, below 900 cm^{-1}, there are no strong spectral peaks except those related solely to water molecules.

It is evident that the solvation effect has an impact on the spectra, in particular on vibrational modes associated with oxygen atoms that tend to form hydrogen bonds. Our study has also showed that the use of PCM changes atom pair motions for non hydrogen bonding atoms and provides good agreement with experiment for non hydrogen bonded systems. The model with explicit hydrogen bonds is shown to improve the vibrational frequencies for the pairs including hydrogen bonded atoms.

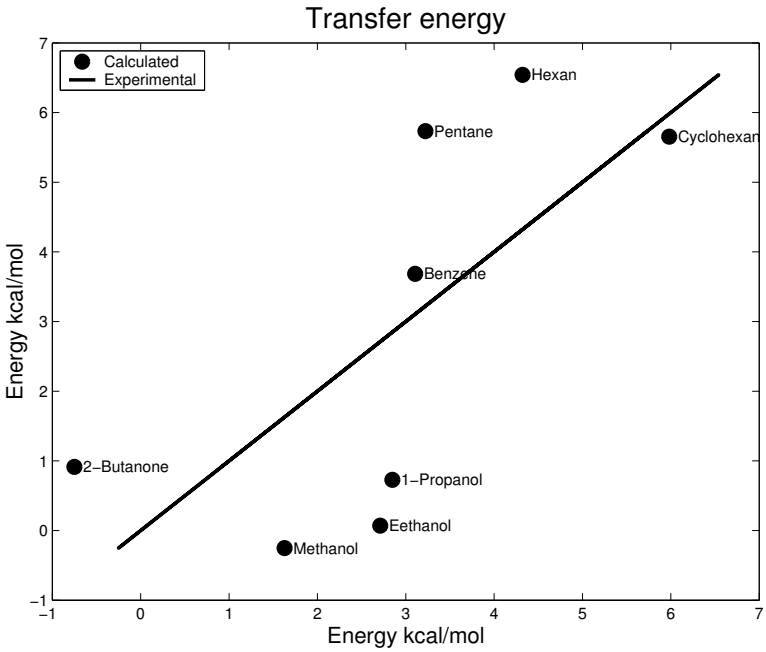

Figure 6.3: Free energy calculation using LIE.

6.2 Interaction energy calculations

In this thesis several different methods have been used to calculate the total interaction energy between molecule and solvent. To show the difficulties to calculate interaction energies, some of the investigated methods will be examplified. The first of these is the linear interaction energy (LIE)[47–49] approach. A set of 8 small organic molecules(liquids) have been investigated using LIE. The computational procedure was first to do MD simulations with the investigated molecule dissolved in water and dissolved in its own environment and apply the following equation:

$$\Delta G_{m\rightarrow aq} = \beta\left(\langle E^{el}_{m-surr}\rangle^{aq} - \langle E^{el}_{m-surr}\rangle^{m}\right) + \alpha\left(\langle E^{LJ}_{m-surr}\rangle^{aq} - \langle E^{LJ}_{m-surr}\rangle^{m}\right) \tag{6.1}$$

The results of the test molecules is plotted in Figure 6.3. The force field that has been used is OPLS.[54] If more work were spent on this method using other force fields or changing the parameters manually for the molecules under investigation, a better agreement with experimental data may have been obtained. Compared to the traditional QSAR methods the use of the LIE method gives a poor correlation. Another approach of calculating interaction energy that has been used extensively in the work of this thesis is the polarizable continuum

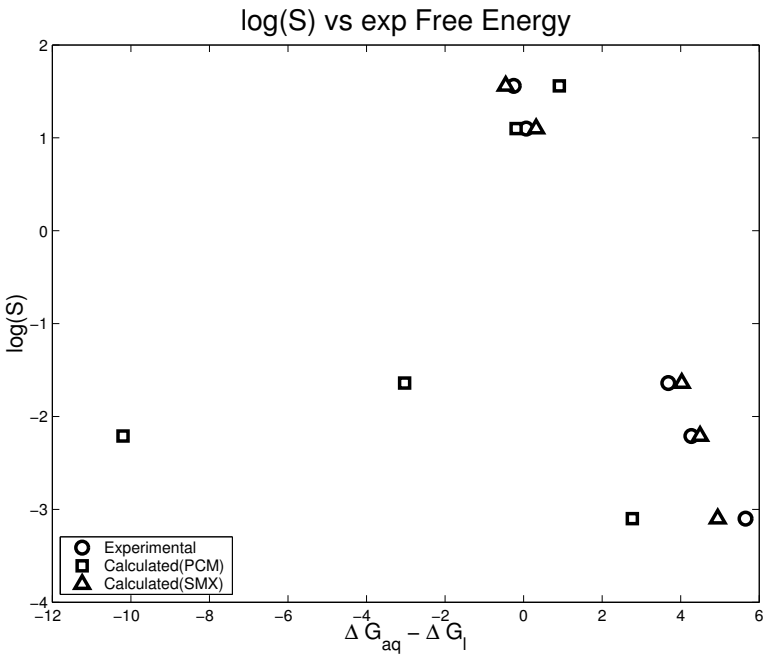

Figure 6.4: Free energies calculated by polarizable continuum models.

models, such as Tomasi' PCM[50–53] and SMxGauss approaches.[55] The main difference between the two models is that the SMX model has more empirical parameters than the PCM model. Calculated free energies for a test set of five small organic molecules using default parameters in both PCM and SMX methods are illustrated in Figure 6.4 together with experimental results. As seen, the PCM results give a poor correlation to the experimental values whereas the results of SMX method agree well with the experimental data. A larger set of molecules has also been tested with the SMX approach, however, when the tested molecules become less common the good agreement with experiment disappears and the results become worse than those obtained from QSAR models. A systematic optimization of various parameters has been carried out but only small improvement has been achieved.

The quantum chemical linear scaling (QCLS) approach has also been used to calculate interaction energies by means of the super-molecule model. In combination with the QCLS calculations, the geometries of super-molecules that consist of the solute and a large number of solvent molecules have been taken out from a MD simulation. The converged size of the supermolecule is obtained when adding additional solvents has very small impact on the interaction energy. In the work of this thesis, solvent molecules within a radius of 10 Å from the center of the solute are included. The interaction energy between the solute and the

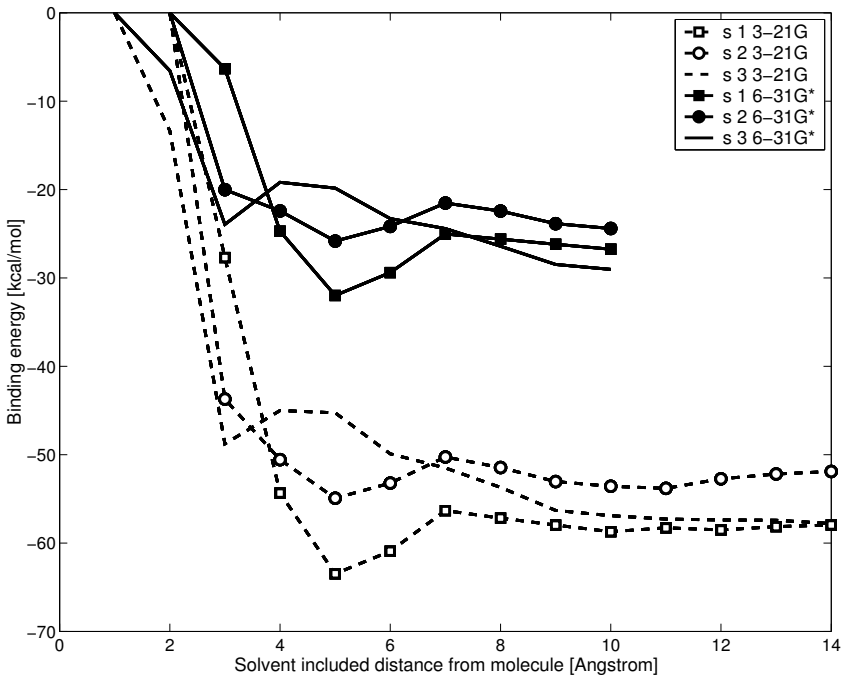

Figure 6.5: Calculated quantum mechanical interaction energies for Aspirin in liquid water.

solvents has then been calculated using the following equation

$$E_{\text{interaction}} = E_{\text{complex}} - E_{\text{solvent}} - E_{\text{solute}} \tag{6.2}$$

As an example, the interaction energies of the drug molecule Aspirine in water calculated with different basis sets are shown in Figure 6.5. It can be seen that it is required to include many layers of water solvents before the energy changes are negligibly small. This behavior has also been seen for other molecules. It is found that the basis set superposition error (BSSE) can have extremely strong impact on the absolute energy values, for which the counterpoise correction is adopted. In Figure 6.6 the basis set superposition error corrected and uncorrected interaction energies for three drug molecules Aspirin, Caffeine and Ibuprofen in water are demonstrated. Since the sizes of the systems are very large, the calculations are quite demanding even though efficient code for large systems has been used. An improved and more efficient scheme has been tested for the benzene molecule. In this scheme the solute and its closed solvents are calculated with larger basis set while the rest of the system with a smaller basis set. The performance of this scheme is shown in Figure 6.7. It is noted that if molecules within the radius of 4 Å from the solute are calculated with large basis sets and the rest with small basis sets, the obtained energy is very close to

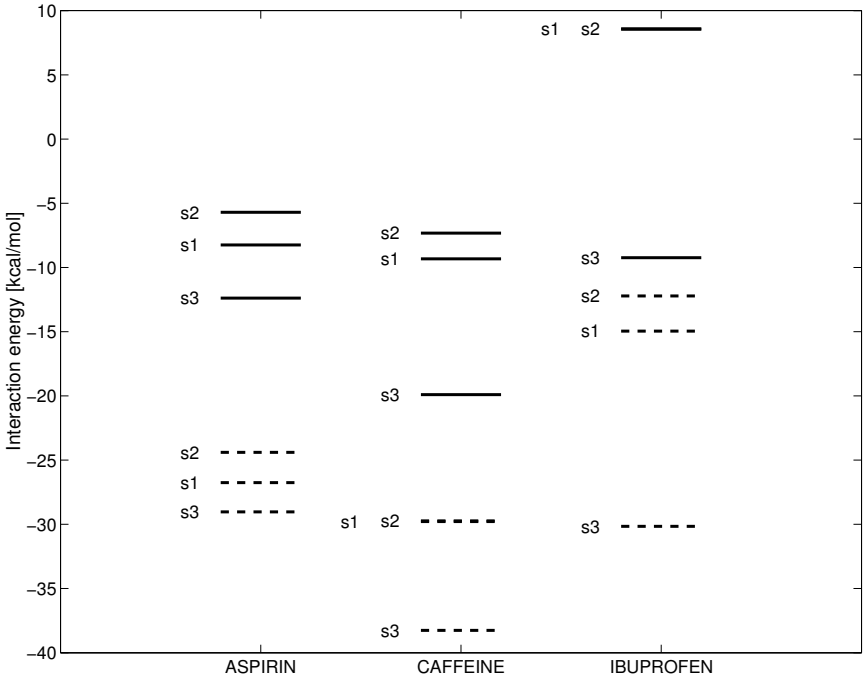

Figure 6.6: Calculated interaction energies with and without BSSE for Aspirin, Caffeine and Ibuprofen in water. s1, s2 and s3 correspond to the different snapshots of MD simulations. The solid and the dashed lines are the BSSE corrected and uncorrected interaction energies, respectively.

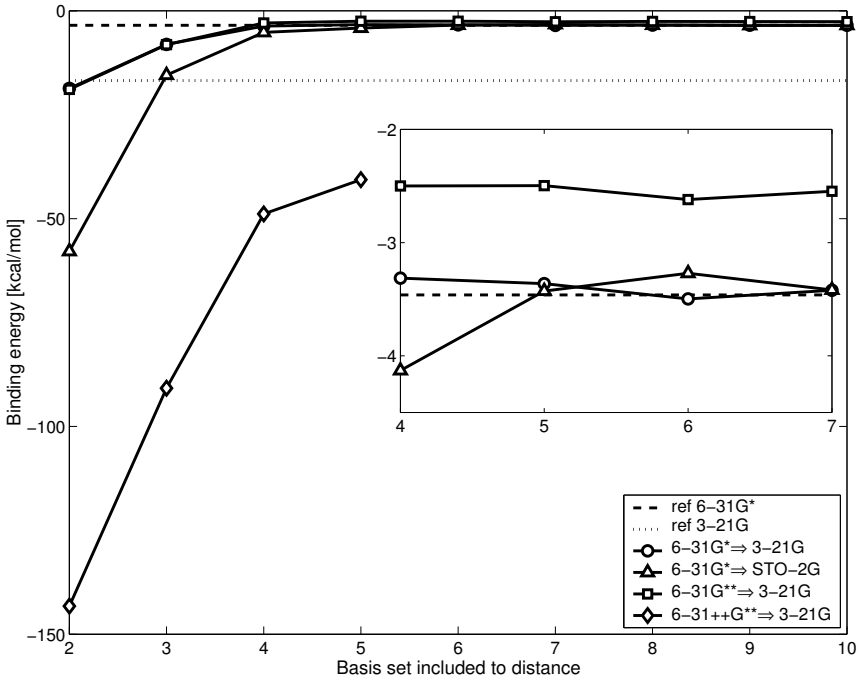

Figure 6.7: Interaction energies calculated with mixture of basis sets for benzene in water.

the calculated value using large basis sets for the whole system. The mixture of basis sets scheme can substantially reduce the computational costs. It is also seen that the inclusion of many solvent molecules have larger effects on the interaction energy than using a model with few solvent molecules with bigger large basis sets. The proposed scheme with mixture of basis sets could easily be applied to drug molecules dissolved in water and in their crystal form.

Ion concentration at the interface

The behavior of ions at gas/liquid interfaces is different from the bulk. The most common way to theoretically simulate gas/liquid interfaces is to use discrete models such as molecular dynamics or Monte Carlo methods. In this thesis an alternative approach, namely the dielectric continuum model, has been employed for the azide ion at a water surface. The computational scheme is, however, different from the PCM method developed for bulk calculations. The cavity surface is discretized in points x_i y_i. The expression for the electrostatic term, G(x,y), then becomes

$$G_E(x,y) = \frac{1}{D(\epsilon(z)|x-y|)} + G_E^{img}(x,y), \tag{7.1}$$

where the first term represents a Coulomb-like interaction to a homogeneous environment having a dielectric permittivity D. This permittivity depends on the permittivity profile at the interface. The second term represents the image-charge interaction. A full derivation for electrostatic contributions can be found in Ref.[43] The repulsion term for this interface model also differs from the bulk model since in the bulk it is assumed that the density is equal around the investigated molecule as shown in equation 7.2

$$G_{rep} = \alpha \int_{\mathbf{r}\notin\mathbf{C}} drP(\mathbf{r}), \tag{7.2}$$

where C is the cavity domain, P(r) is the solute molecule density and $\alpha = 0.063\rho_B\frac{n_{val}^B}{M_B}$, where n_{val}^B is the number of valence electrons and M_B the solvent molecular weight. The above expression for the repulsion free energy leads to the following expression for the interaction Hamiltonian

$$\mathbf{h_{rep}} = \frac{\delta \mathbf{G_{rep}}}{\delta \mathbf{P}} = \alpha \left[\mathbf{S} - \mathbf{S^{(in)}}\right], \tag{7.3}$$

where $\mathbf{S}$ is the overlap matrix and $\mathbf{S^{(in)}}$ involves a sum of electric field integrals on the surface. However, such an expression is not suitable for interfaces since it assumes that the

density is equal around the cavity. Therefore a new expression for the repulsion term was derived. The full derivation can be found in Paper 4 of the thesis. The repulsion free energy term has the form

$$G_{rep} = \sum_i \alpha' \rho_B(s_i) P(s_i) f(s_i), \tag{7.4}$$

where the s_i is the surface element, $\rho_B(r)$ the solvent density, $P(s_i)$ the solute molecule density, and $f(s_i)$ is the weight which now depends on the position and an $\alpha' = 0.063 \frac{n^B_{val}}{M_B}$. The repulsion operator can then be written as follows

$$\mathbf{h_{rep}} = \frac{\delta \mathbf{G}}{\delta \mathbf{P}} = \sum_{\mathbf{i}} \alpha' \rho_{\mathbf{M}}(\mathbf{s_i}) \mathbf{f}(\mathbf{s_i}) \delta_{\mathbf{i}} \tag{7.5}$$

The major difference between the formulation for interface and for bulk calculations is that the density has to be modeled at each point in the surface for the case of the interface, whereas for the bulk formulation the α term including the density is a constant and is not included in the integration.

The model above was used to calculate the energy profile and some properties depending on the position in the interface for the azide ion. The interface between the liquid and gas phase has a sigmodial shape where the negative values refer to the solvent and positive values to the gas phase. Since the ion is linear, three different orientations have been considered, namely: i) perpendicular to the interface, ii) an angle of 45 degrees to the surface, and iii) parallel to the surface. These orientations will be refereed to as θ 0 , θ 45 and θ 90, corresponding to the angle to the normal of the interface. All calculations have been performed with both a pure electrostatic model and a combination of electrostatic and repulsive models. For all orientations there was a small energy minimum observed close to the interface for the repulsive model, whereas no minimum was observed for the electrostatic model. The depth of the observed minimum was also similar for all three orientations, around -0.60 Kcal/mol with respect to the corresponding bulk value. The dipole moment was also investigated for the three orientations. For the dipole moment oriented along the normal to the interface one could see that the three orientations had a maximum around 3 Å for both models and a small minimum around -2.5 Å for the model including the repulsive contribution. The trace of the polarizability was also investigated. One could see a maximum for the model including the repulsive contribution, around -2.5 Å for all orientations, which was not presented in the results of electrostatic model. A concentration profile, see Figure 7.1, based on the Boltzmann distribution was also calculated from the expression:

$$c(z) = c_0 e^{\frac{G_0 - G(z)}{RT}}, \tag{7.6}$$

where G_0 is the bulk energy, z the position at the interface, T the temperature (here assumed to be 300K) and c_0 the concentration (here set to 1). As seen in Figure 7.1 it is necessary

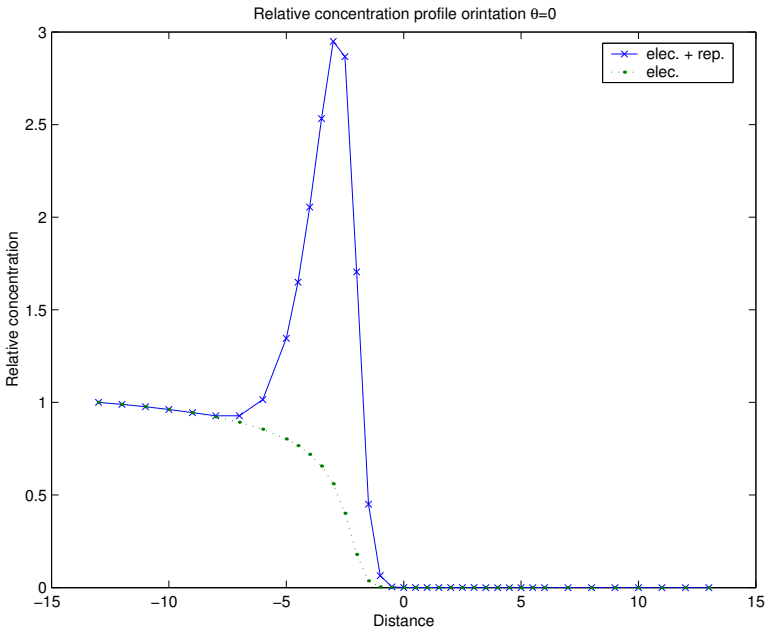

Figure 7.1: Calculated concentration profile of the azide ion

to use a model that includes the repulsion contribution and is able to detect the minimum that gives an increased concentration close to the interface. Our obtained results for the gained increase in concentration close to the surface are in good agreement with other computational studies. It has been found that the peak concentration is about 3-4 times higher than that in the bulk.

Bibliography

[1] V. B. Konstantin, P. S. Nikolay and I. G. Tetko, Curr. Medicinal Chemistry, **13**, 223, 2006.

[2] H. Kishi and Y. Hashimoto, Chemosphere, **18**, 1749, 1989.

[3] L. W. Jorgensen and E. M. Duffy, Adv. Drug Delivery Reviews, **54**, 355, 2002.

[4] M. Coates, D. W. Conell and D. M Barron, Envir. Sci. Technology, **19**, 628, 1985.

[5] C. Hansch, and J. E. Quinlan, and G. L. Lawrence,J. Org. chem. 33, 347, 1968.

[6] S. H. Yalkowsky and S. C. Valvani, J. Pharm. Sci., **69**, 912, 1980.epsps

[7] N. Jain and S. H. Yalkowsky, J. Pharm. Sci., **90**, 234, 2001.

[8] C. Hansch and A. J. Leo, Substituent parameters for Correlation analysis in Chemistry and Biology, Wiley, New York, 1979.

[9] W. M. Meyland and P. H. Howard, J. Pharm. Sci. **84**,83,1995.

[10] A. K. Ghose, A. Pritchett and G. M. Crippen. J. Comput. Chem. **9**, 80, 1988.

[11] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome and Y. Matsushita, Chem. Pharm. Bull. **40**, 127, 1992.

[12] J. S. Delaney , Drug Disc. Today., 4, 289, 2005.

[13] D. Erös, G. Keri, I. Kövesdi, C. Szantai-Kis, G. Meszaros and L. Örfi, Mini Reviews Medicinal Chemistry, **4**, 167, 2004.

[14] R. Kühne, R.-U. Ebert, F. Kleint, G. Schmidt, G. Schüürmann, Chemosphere. **30**, 2061, 1995.

[15] G. Klopman and H. Zhu,- Chem. Inf. Comput. Sci. **41**, 439, 2001.

[16] A. Klamt, F. Eckert and M. Hornig, Journal of Computer-Aided Molecular Design, **15**, 355, 2001.

[17] A. Klamt, F. Eckert, M. Hornig, M. E. Beck and T. Bürger, J. Comp. Chem. **23**, 275, 2002.

[18] J. Huuskonen **40**, 773, 2000.

[19] C. A. Bergström, C. M. Wassvik, U. Norinder, K. Luthman and P Artursson, J. Chem. Inf. Comput. Sci. **44**, 1477, 2004.

[20] J. M. Sutter and P. C. Jurs, J. Chem. Inf. Comput. Sci. **36**, 100, 1996.

[21] N. R. McElroy and P. C. Jurs, J. Chem. Inf. Comput. Sci. **41**, 1237, 2001.

[22] B. E. Mitchell and P. C. Jurs, J. Chem. Inf. Comput. Sci. **38**, 489, 1998.

[23] G. Klopman, S. Wang and D. M. Balthasar, J. Chem. Inf. Comput. Sci. **32**, 474, 1992.

[24] W.L Jorgensen and E. M. Duffy, Bioorg. Med. Chem. Lett., **10**,1155, 2000.

[25] E. M. Duffy and W.L Jorgensen, J. Am. Chem. Soc., **122**,2878, 2000.

[26] I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva and A. E. P. Villa, J. Chem. Inf. Comput. Sci. **41**, 1488, 2001.

[27] D. Butina and J. M. R. Gola, J. Chem. Inf. Comput. Sci. **43**, 837, 2003.

[28] J. D. Thompson, C. J. Cramer and D. G. Truhlar, J. Chem. Phys. **119**, 1661, 2003.

[29] W. M. Meylan and P. H. Howard, Perspect. Drug Discov. Des. **19** 67 2000.

[30] P. J. Steinbach and B. R. Brooks, J. Comp. Chem. **15**, 667, 1994.

[31] P. Ewald, Ann. Phys. **64**, 253, 1921.

[32] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case., J. Comput. Chem., **25**,1157, 2004.

[33] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klien, J. Chem. Phys. **79**, 726, 1983.

[34] D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, R.V. Stanton, A.I. Cheng, and P.A. Kollman(2004), AMBER 8, University of California, San Francisco.

[35] A. Szabo, N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory.* Dover Publications, Inc., Mineola, New York, 1996.

[36] S. Höfinger and F. Zerbetto, Chem. Eur. J. **9**, 566, 2003.

[37] S. Höfinger and F. Zerbetto, Theor. Chem. Acc. **112**,240, 2004.

[38] R. A. Pierotti, Chem. Rev.,**76**,719, 1976.

[39] J. L Pascual-Ahuir and E. Silla, J. Comp. Chem., **11**, 1047, 1990.

[40] E. Silla and Inaki Tunon and J. L Pascual-Ahuir, J. Comp. Chem., **12**, 1077, 1991.

[41] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[42] J. Ohnsmann, G. Quints, S. Garrigues, and M. Guardia. *Anal. Bioanal. Chem.*, 374:561, 2002.

[43] L. Frediani and R. Cammi and S. Corni and J. Tomasi, J. Chem. Phys. **120**, 3893, 2004.

[44] E. Rudberg, *Fock Matrix Construction for Large Systems*, Licentiate thesis, Theoretical Chemistry, KTH, Stockholm (2006)

[45] E. H. Rubensson, *Sparce Matrices in Self-Consistent Field Methods*, Licentiate thesis, Theoretical Chemistry, KTH, Stockholm (2006)

[46] E. Rubensson, E. Rudberg and P. Sałek. ergo version 1.1 a quantum chemistry program for large-scale self consistent field calculations, 2006.

[47] J. Åqvist, C. Medina and J.E Samuelsson. Prot. Eng. **7**,385,1994.

[48] D.K. Jones-Hertzog and W. L. Jorgensen. J. Med. Chem. **40**,1539,1997.

[49] T. Hansson , J. Marelius and J. Åqvist. J. Comput.-Aided Mol. Des. **7**,385,1998.

[50] S. Miertus, E.Scrocco and J Tomasi. Chem. Phys. **55**, 117, 1981.

[51] S. Miertus and J. Tomasi. Chem. Phys. **65**, 239, 1982.

[52] J. L. Pascual-Ahuir, E. Sila. J. Tomasi and R. J. Bonaccorsi. J Comput. Chem. **8**, 778, 1987.

[53] M. A. Aquilar, F. J. Olivares del Valle and J. Tomasi. J. Chem. Phys. **98**, 778, 1993.

[54] W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives. J. Am Chem. Soc., **118**, 11225, 1996.

[55] G. D. Hawkins, C. J. Cramer and D. Thrular. J. Phys. Chem. B **102**,3257,1998.