Microscopic Views of Drug Solubility

Laban Bondesson

Theoretical Chemistry
Royal Institute of Technology
Stockholm 2006
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 accurate as desired. In this thesis different aspects that are known to have a large impact on the aqueous solubility of a molecule have been studied in detail using various theoretical methods with intension to provide microscopic view on drug solubility. The first aspect studied is the hydrogen bond energies. Eight drug molecules have been calculated using density functional theory and the validity of additive model that has often been used in solubility models is examined. 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 was shown that the calculated volume differed significantly dependent on which method that had been used, especially for larger molecules.

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 seen that the concentration close to the interface was almost three times higher than in the bulk. The increase in concentration close to the surface depends on the larger gap between the interface energy and the gas phase energy than between the bulk energy and the gas phase energy.
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 IV** Calculations of the cavitation volumes and partial molar volumes of drugs in water, L. Bondesson and H. W. Hugosson, in preparation.
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.
Acknowledgments

This Licentiate 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 and Dr. Luca Frediani, who explained different theories and brought good ideas to the projects.

Thanks to Elias Rudberg 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.
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 used (log(S)). Since the molecule may protonate or deprotonate dependent on the pH value in water, the solubility is measured in its least soluble environment which is when the molecule is in its neutral form. This is usually referred to as the intrinsic solubility. In fact what is measured is the amount of the liquid or crystal
A solute that is dissolved in water, as shown in the schematic Figure 1, which depends on thermodynamic properties of the investigated solutes in the 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. Chapter four will highlight some of the key effects that determine the solvation energy of a molecule. The fifth 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

As 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 aqueous solubility 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 and the standard deviation was 0.19 log units. For the tabulated experimental data in general the measurement procedure is 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 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.

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. The octanol/water partition coefficient log(P) is slightly 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), 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:

$$
\log(S) = 0.5 - \log(P) - 0.01(t_m - 0.25).
$$

(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 had 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 also many models to predict the log(P) value with different accuracy.\textsuperscript{7-10} The melting temperature is much 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.\textsuperscript{1,3,11,12} Delaney\textsuperscript{11} has recently compared present models. He divides the models dependent on regression/modeling method and descriptors. The same classification of models dependent on regression models has been used in the review of Jorgensen and Duffy.\textsuperscript{3} Four regression methods, namely the Artificial Neural Network (ANN), the Linear Regression(LR), the Group contribution and the Artificial Intelligence (AI), are 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 $n_i$ of this fragment is performed. The solubility can then be calculated using the equation

$$
\log S = \sum_i a_i n_i + a_0.
$$

(2.2)

The model is optimized with respect to the different fragments’ contribution to the model, which is done with regression analysis. The only descriptor in this method is the structural fragment. Examples of this model is given by K"{u}ne et. al\textsuperscript{13} which is built on 694 organic nonelectrolyte 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 their model less suitable for drug molecules. Another more recent group contribution model is developed by Klopman et al.\textsuperscript{14} who uses a data set of 1168 organic molecules. The standard
deviation (rms error) for their study was 0.79 log units. It is noted that in that study most of the investigated molecules were classic organic molecules with one functional group.

It should also be noted that for predicting log(P) there exists a quite accurate group contribution model proposed by Leo and Hanch. 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. 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

$$\text{log}S = \sum_i a_i c_i + a_0$$

(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 calculated from different kinds of methods like fragment counts, molecular or quantum mechanical calculated properties, Monte Carlo or molecular dynamics (MD) simulated properties.

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 contribution of the original variables. The number of components that should be used in the model is determined by leave-one-out cross-validation. 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 of studies predicting the water solubility using MLR methods.\textsuperscript{3,16–18} Jurs et al.\textsuperscript{19–21} have developed both MLR and Neural Network (NN) models. They use a set of 200 topological, geometric and electronic descriptors obtained from calculations using the AM1 and PM3 semi-empirical quantum chemistry methods. 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\textsuperscript{17} 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 a 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\textsuperscript{22} 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).\textsuperscript{7,3,23} 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 a 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 proprieties 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 refereed to as COSMO-RS.\textsuperscript{16} This model differs from those methods that 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 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\textsuperscript{11} 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 common 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 over training 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. 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. 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 has many similarities with the model developed by Huskonen. The difference is that this model uses fewer descriptors and has three times fewer hidden neurons. According to the authors the model should therefore be more robust. 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. 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 present today, one can say that the NN models achieve a better rms than the MLR models. Although the drawback by using NN models is that there is a risk that these models are overtrained and that it is difficult or even impossible to modify the molecule dependent on the results of the models due to the lack of connection to real physical properties. It should 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 performed by Bergström et al. who produced a data set built upon 85 molecules, all measured using the same conditions. However, to develop accurate QSAR models a much larger data set is required.

2.4 Models built on combinations of experimental and calculated data

There also exist models that combine experimental and computed properties. For instance, the method of Thompson et al. partly partly experimental vapor pressure data to predict the solubility. Since their model is built upon thermodynamic methods, the investigated data set consists mostly of small organic molecules and therefore it is not comparable to the computational models described above. However, since this model deals with the aqueous solubility problem in a different way by using statistical thermodynamic approaches it is worth to note. Meylan and Howard have developed a method using calculated values of log(P) and experimental melting temperature. They, like Yalkowski, found that the melting temperatures is an important parameter when predicting log(S) for solids. The models that use combination of experimental and calculated descriptors are of course not suitable for in silico screening of new compounds. However these methods can give deeper understanding of what is missing in the pure computational models and therefore be valuable for the development of new computational models. This could, for example, be done by developing new descriptors that are related to experimental properties that are important for the solubility.
Chapter 3

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 system $s$ is defined as

$$\mu_s = \left( \frac{\partial G}{\partial N_s} \right)_{P,T,N'_s} = \left( \frac{\partial A}{\partial N_s} \right)_{T,V,N'_s},$$  \hspace{1cm} (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 $N'_s$ the number of molecules in the system except the $N_s$ molecules. For a two component system the chemical potential is defined as

$$\mu_A = G(T, P, N_A + 1, N_B) - G(T, P, N_A, N_B),$$  \hspace{1cm} (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 is taking one molecule $s$ from phase $\alpha$ into phase $\beta$. This process is carried out under constant pressure $P$ and temperature $T$. The Gibbs energy of solvation of $s$ from phase $\alpha$ into phase $\beta$ then becomes

$$\Delta G_s = \mu_s^\beta - \mu_s^\alpha.$$  \hspace{1cm} (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:

$$A(g) \leftrightarrow A(l),$$  \hspace{1cm} (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

$$\Delta G_{l\rightarrow g} = RT\ln \frac{P_A}{P_{M_A}^l},$$  \hspace{1cm} (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:

\[
A(l) \leftrightarrow A(aq).
\]  

(3.6)

The free energy for this process is defined as:

\[
\Delta G_{l\rightarrow aq} = -RT\ln \frac{M_A^{aq}}{M_A^l},
\]

(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

\[
A(g) \leftrightarrow A(aq),
\]

(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

\[
\Delta G_{g\rightarrow aq} = RT\ln \frac{P_A}{P M_A^{aq}}.
\]

(3.9)

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

\[
A(g) \leftrightarrow A(s)
\]

(3.10)

and

\[
A(s) \leftrightarrow A(aq)
\]

(3.11)

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

\[
\Delta G_{s\rightarrow aq} = RT\ln \frac{P_A}{P M_A^s}
\]

(3.12)

and

\[
\Delta G_{s\rightarrow aq} = -RT\ln \frac{M_A^{aq}}{M_A^s}.
\]

(3.13)

These relations are the same as for liquids except that the \( P_A \) 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

\[
S = M_A^{aq} = M_A^s e^{\left(\frac{\Delta G_{s\rightarrow aq}}{RT}\right)}.
\]

(3.14)
If one considers the difference between free energy of hydration and free energy of vaporization $\Delta\Delta G$ it is evident that the 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 which is dependent on Helmholtz energy, defined as

$$A = -k_B T \ln Z,$$

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

$$\Delta G_{free} = kT \ln \left( \frac{\rho^a}{\rho^g} \right).$$
where $\rho^\alpha$ and $\rho^\beta$ 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:

$$\Delta G = \Delta H - T\Delta S,$$

(3.17)

where $T$ is the temperature.

This relation has been widely used in solubility described 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 weight of the entropy and enthalpy terms depends on the molecule. There are several properties of the molecule that determine its solubility. The size of the molecule is an important factor and widely used in solubility models. For models described in Chapter 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 two it was claimed that the octanol/water partition log(P) had 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

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

(3.18)

$\Delta G_{\text{elec}}$ is the electrostatic contribution which is important for polar solutes due to the
polarization of the solvent. $\Delta G_{vdw}$ is the van der Waals interaction and is usually split into a repulsive $\Delta G_{rep}$ and an attractive dispersion term $\Delta G_{disp}$. $\Delta G_{cav}$ is the free energy to form the cavity for the solute in the solvent. The last term $\Delta G_{hb}$ is sometimes used for systems where there exist hydrogen bonds between the solute and the solvent. In this thesis some of the above components will be examined and studied in detail.
Chapter 4

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 dependent 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:

\[ \Delta G_{sol} = \Delta G_{el} + \Delta G_{cav} + \Delta G_{disp} + \Delta G_{rep} \]  

(4.1)

Where the \( \Delta G_{el} \) 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:

\[ G_{el} = \langle \Psi | H(\Psi) - 1/2V(\Psi) | \Psi \rangle, \]  

(4.2)

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

\[ H(\Psi) = H^0 + V(\Psi). \]  

(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

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

(4.4)

where \( \rho_M \) and \( \sigma \) are two electrostatic potentials and \( \Sigma \) 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:

\[ G_{cav} = \sum_{i}^{\text{spheres}} \frac{A_i}{4\pi R_i^2} G_i^{HS}, \]  

(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:

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

(4.6)

where \( P_M(0K| r) \) and \( \omega_{0K}^M \) are, respectively, the transition densities and energies for solute \( M \) (for transition from ground state \( 0 \) to excited state \( K \)) and \( \sigma_S \) 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:

\[ G_{rep} = \alpha \int_{r \in C} drP(r), \]  

(4.7)

where \( \alpha \) is a constant defined by properties of the solvent, \( P(r) \) 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 field.) This method has been used in this thesis and will therefore be described in more detail.

To deal with both structural and dynamic properties Molecular Dynamic simulations (MD) 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

\[
\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i},
\]

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

\[
F_i(t) = -\frac{\partial V}{\partial r_i}.
\]

Since it is only computationally possible to use a finite system, periodic boundary conditions are usually introduced when performing MD simulations. By introducing periodic boundary conditions a virtual infinite system is created and the investigated cell is repeated in all directions. 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 use a correction for the long range inter-atomic interactions that may affect the behavior of other atoms beyond the size of the cell. Methods for dealing with this kind of corrections are Ewald summation, group based truncation and atom based force shift.\textsuperscript{29,30}

When performing 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 the ensemble is called NVT. In the case when the number of particles (N), the pressure (P), and the temperature (T) are kept constant the ensemble is called NTP. In this thesis only the NPT ensemble has been used.

Force fields play the essential role in MD simulations. In this thesis the so-called *general amber force field parameters*\(^{31}\) (GAFF) and TIP3P\(^{32}\) water force field parameters have been used. The potential function used in the AMBER program\(^{33}\) are defined as follows.

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

where \(r\) is the bond length, \(\theta\) the bond angle, \(\phi\) the dihedral angle, \(n\) the multiplicity and \(\delta\) the phase. \(K_r\) and \(K_\theta\) the force constants for bond and angle respectively. \(A\) and \(B\) are atom dependent Lennard Jones parameters, \(q\) the atomic charges and \(\epsilon\) the dielectric constant.

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 by 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: a soft and a hard part. 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)\textsuperscript{34,35} and scaled particle theory.\textsuperscript{36} In this thesis Molecular Dynamics simulations have been used to study the molecular volume using two different methods and the results have been compared with computed volumes obtained by the GEPOL\textsuperscript{37,38} algorithm, which is an algorithm for calculating the molecule surface area and volume, and experimental partial molar volumes. One of the methods is MD simulations where the impact of solute molecule on the total volume of the system is investigated. The following procedure is used in the calculations. 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 water in the simulated box. This procedure will be referred to as $V_{\text{tot}}^{\text{MD}}$ below. Another method is the Voronoi volume of the solute obtained 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 check if a solute or a solvent atom is the closest one to the cell. If the solute atom is the closest one to the investigated cell, the cell volume is added to the total Voronoi volume, otherwise the cell volume is discarded. This model will be labeled as $V_{\text{Vor}}^{\text{MD}}$ below.

The calculated volumes from these two methods presented in this thesis and from the GEPOL\textsuperscript{37,38} algorithm using PCM are shown in Figure 5.1 for comparison. The volumes of
the different models are very similar for volumes below 350 Å³. When looking at volumes larger than 350 Å³ one can observe that the $V_{tot}^{MD}$ gives a larger volume than the ones obtained with GEPOL algorithm. In the comparison between $V_{tot}^{MD}$ 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 $V_{Vor}^{MD}$ is similar to those obtained by the GEPOL algorithm. That the two models described above give different volumes is not something 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 volume calculated by the two MD methods and the GEPOL algorithm has been shown for a set of five small organic molecules in Table 5.1.

It can be seen that the $V_{tot}^{MD}$ model has the best agreement to 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. When performing a PCM calculation in the Gaussian program one obtains a cavity energy term from the cavity used in the calculation. An almost linear relation between the GEPOL cavity volume and the cavity energy is observed from these calculations. A least square fitting
Table 5.1: Calculated and experimental partial molar volumes for four alcohols and benzene molecules.

<table>
<thead>
<tr>
<th>Molecule</th>
<th>$V_{TOT}^{MD}$</th>
<th>$V_{TOT}^{MD}$</th>
<th>$V_{Exp}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methanol</td>
<td>73.051</td>
<td>66.0636</td>
<td>62.865</td>
</tr>
<tr>
<td>2-propanol</td>
<td>124.673</td>
<td>125.349</td>
<td>110.872</td>
</tr>
<tr>
<td>Benzene</td>
<td>136.232</td>
<td>136.705</td>
<td>139.067</td>
</tr>
<tr>
<td>2-methyl-2-propanol</td>
<td>140.902</td>
<td>153.352</td>
<td>137.578</td>
</tr>
<tr>
<td>3-hexanol</td>
<td>194.092</td>
<td>213.246</td>
<td>205.803</td>
</tr>
</tbody>
</table>

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. This is only 10-15 percent relative to 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 log(S) units and energy units in KCAL/mol is roughly 1.
Chapter 6

Hydrogen bonding energy

The solubility is strongly dependent on how hydrophilic the solute is. Generally one can say that a molecule that can form many hydrogen bonds to the water solvent becomes more soluble and vice versa. When using continuum models the electrostatic interactions are well covered by these models. However, the hydrogen bonding energy is not completely covered by the electrostatic interactions. Today many aqueous solubility models based on QSPR use the number of hydrogen bonds for certain groups present in the investigated molecule as the main descriptor. This descriptor is then multiplied by a weight in the regression step during the development of the model. The hydrogen bond energy and the impact of the hydrogen bondings for molecules in solution have therefore been investigated in this thesis.

Hydrogen bonds have been studied extensively by a number of authors. Our focus is to examine if there is an additive behavior of some of the hydrogen bonding groups present in drug molecules. The hydrogen bonding energies to water have been calculated for a set of small organic molecules which have different specific hydrogen bonding groups. A different number of hydrogen bonded waters was added to the different groups dependent on how many hydrogen bonds they formed; for example two water molecules to alcohol groups and three water molecules to carboxyl acid group. The starting point for the water molecules was selected visually and then a geometry optimization was performed. The hydrogen bonding energy was calculated using the following procedure: First the 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 hydrogen bonding energy for each group was then stored for use in an additive model.

The validity of the additive approach was investigated for a set of eight common drug molecules. The hydrogen bonding energy for each drug molecule was first calculated using
the same computational scheme as described above. Each molecule was then visually in-
vestigated and the occurrence of each hydrogen bonding group was counted. The additive
hydrogen bonding energy was calculated by multiplying the number of occurrences of each
group by the additive hydrogen bonding value. 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 two molecular properties was investigated for a set
of three drug molecules dissolved in water. The investigated properties were their 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 nm. IR radiation therefore has not enough energy to cause electronic transitions
in the molecule but may induce vibrational and rotational excitations of covalent bonds.
In order to measure the IR spectrum of a molecule, a beam of monochromatic light passes
through the sample and the amount of energy absorbed by the molecule is recorded. IR
spectroscopy works because different chemical bonds have specific frequencies at which they
vibrate, corresponding to 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) falls on the molecule and interacts with the electric dipole moment. The
photon excites one of the electrons into a virtual state since 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 is measured in Raman spectroscopy and gives a fingerprint for a specific molecule.

The computational procedure for the investigation in this thesis involves several steps. A fre-
quency 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 un-
derstand the effects of long-range and short-range interactions on the vibrational frequencies
and 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 using explicit hydrogen bonds. The investigated frequencies have been divided into three regions for the caffeine molecule. In the high frequency region 3000-4000 cm$^{-1}$ there is only small IR absorbance in the frequency region 3250 cm$^{-1}$ and lower, except for peaks related to the presence of OH bonds in 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 relating to C-H stretching motions, which were lowered a bit by using dielectric continuum models but which were almost unaffected by 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 gas phase calculated value for this peak was 1752 cm$^{-1}$. 

Figure 6.1: Caffeine molecule with hydrogen bonded waters
When using PCM the frequency was decreased by 17 cm\(^{-1}\), and for the calculations with explicit 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\(^{40}\) which was 1720 cm\(^{-1}\). The frequency with the highest absorbance in the caffeine molecule corresponds to C=O bending and pyrimidine ring bending. The gas phase calculated 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\(^{40}\) 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 including oxygen atoms that tend to form hydrogen bonds. The study has also shown 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.
Chapter 7

Ion concentration at the interface

The behavior of ions at the gas/liquid interface is different than for the ions in 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 it has been studied using a dielectric continuum model for the azide ion. This method is, however, different from the PCM method developed for bulk calculations. The electrostatic contribution for the interface calculation differs in how the Green’s function is formulated. The cavity surface is discretized in points $x_i$, $y_i$. The expression for the $G(x,y)$ then becomes

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

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 of the electrostatic term can be found in.\textsuperscript{41} 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

$$G_{rep} = \alpha \int_{r \in C} dr P(r), \quad (7.2)$$

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

$$h_{rep} = \frac{\delta G_{rep}}{\delta P} = \alpha \left[ S - S^{(in)} \right], \quad (7.3)$$
where $S$ is the overlap matrix and $S^{(\text{in})}$ 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 the paper 4 of the thesis. The repulsion free energy term then becomes

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

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

$$h_{\text{rep}} = \frac{\delta G}{\delta P} = \sum_i \alpha' \rho_M(s_i)f(s_i)\delta_i \quad (7.5)$$

The major difference between the formulation for interfaces and for bulk calculations is that the density has to be modeled at each point in the surface for the interface formulation, whereas for the bulk formulation the $\alpha$ 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 dependent on the position in the interface for the azide ion. The interface between the liquid and gas phase has a sigmoidal shape where the negative values refer to the solvent and positive values to gas phase. Since the molecule 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 $\theta = 0^\circ$, $\theta = 45^\circ$ and $\theta = 90^\circ$ corresponding to the angle to the normal of the interface. All calculations have been performed with both a pure electrostatic model and an electrostatic and repulsive model. 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 observed minimum was also similar for all three orientations and was 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 repulsive contribution. The magnitude of the maximum differed corresponding to the orientation. The trace of the polarizabilities was also investigated. One could see a maximum for the model including repulsive contribution, which was not present for the electrostatic model, observed around -2.5 Å for all orientations. A concentration profile, see figure 7.1, based on the Boltzmann
distribution was also calculated from the expression:

\[
c(z) = c_0 \exp \left( \frac{G_0 - G(z)}{RT} \right),
\]

(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 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


