# AQUEOUS SOLUBILITY OF DRUG-LIKE COMPOUNDS

A Dissertation submitted to the
Rheinische Friedrich-Wilhelms-University Bonn
for the degree of
Doctor of Natural Sciences

Presented by

## LEI DU-CUNY

Dipl. Ing.
University Darmstadt (Germany)
born June 11, 1977
Shanghai, China

Bonn 2006

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

Dekan: Prof. Dr. A. B. Cremers, Institut für Informatik der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Prof. Dr. M. Wiese, Abteilung Pharmazeutische Chemie, Pharmazeutisches Institut, Rheinischen Friedrich-Wilhelms-Universität Bonn

2. Referent: PD Dr. J. Huwyler, Abteilung Toxikologie und Klinische Pharmakologie, Pharmazentrum, Universität Basel und F. Hoffmann-La Roche Ltd, Basel

Tag der Disputation: 25. April 2006

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert

Erscheinungsjahr: 2006

# Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbst und ohne jede unzulässige Hilfe angefertigt habe. Aus fremden Quelle entnommene Gedanken und Daten sind als solche kenntlich gemacht. Diese oder eine ähnliche Arbeit sind von mir noch an keiner anderen Stelle einer Prüfungsbehörde vorgelegt worden. Ich habe vormals noch keinen Promotionsversuch unternommen. Die Ergebnisse dieser Dissertation sind an der nachstehend aufgeführten Stelle auszugsweise veröffentlicht worden.

Du-Cuny, L.; Fischer, H.; Huwyler, J.; Kansy, M. Method for crystallization of a weakly acidic and/or weakly basic compound. EP patent Appln. No.05018750.9 filed August 30, 2005.

Bonn, den 30. 01. 2006

# Acknowledgements

The studies in this thesis were carried out at the F. Hoffmann La Roche Ltd, Basel, Switzerland and supported by the department of pharmacy, faculty of pharmacy, Rheinische Friedrich-Wilhelms-University Bonn.

I wish to express my sincere gratitude to

My supervisor, **Prof. Dr. Michael Wiese**, without whose recommendation and acceptation this thesis would not be what it is today and to **PD Dr. Jörg Huwyler** who has acted as the second supervisor, and whose wide experience in the area of PK/PD studies was an invaluable gain of knowledge for me in the area of ADMET studies. Special thanks to **Dr. Manfred Kansy**, whose novel ideas and extensive know-how in the area of compound profiling accompanied me, during the whole Ph-D study, to the remarkable success. His energetic and attentive enthusiasm, and steady willingness for discussion was a great motivation for me and the described projects. Special thanks also to **Prof. Klaus Müller** who taught me medicinal chemistry during my later working phase and whose brilliant scientific suggestions lead me to the breakthrough in the in silico prediction described herein.

My working group: **Dr. Stefanie Bendels** for assisting my computer-based activities and the unforgettable time spent together in the room 5.88; **Dr. Holger Fischer**, for many valuable solubility- related discussions and suggestions; **Mr. Björn Wagner** for performing the best p$K_a$ measurements in the world; **Mrs. Virginie Micallef** and **Mrs. Isabelle Parrilla** for providing precise solubility data; **Mr. Frank Senner** for allowing my curiosity in the PAMPA and Caco2 studies.

My colleagues in the molecular modeling group: **Dr. Martin Stahl** for many promising topics discussed in the CAMM meetings; **Dr. Wolfgang Guba** for introducing the Free-Wilson analysis to me; **Dr. Paul Gerber** for enabling me to calculate the binding strength of hydrogen bonds; **Dr. Bernd Kuhn** for giving me a longed-for CSD course; **Dr. Harald Mauser** and **Dr. Olivier Roche** for solving any software-related problems; **Dr. Nicole Kratochwil** for kind conversations in diverse areas.

My “crystallography group”: **Mr. André Alker** for exciting discussions about the crystal breeding and the identification of crystal 3D structures; **Mrs. Martina Stihle** for her magic treatment of crystals under the microscope.

My chemist: **Dr. Synèse Jolidon** for guaranteeing my function as ordinary synthesis chemist in the laboratory.

My project supporters: **Dr. Günter Gross**, **Dr. Ulrich Widmer** and **Dr. Primin Hidber** for providing the data for the DDPD (Drug Development Profile DB) database; **Mr. André Thiele** for keeping my WINDOWS system running smoothly; **Mrs. Regina Mehlin** for administering the conferences and meetings.

Finally, my greatest debt is due to my family. My mother **Prof. Dr. Du Minqiong** and my aunt **Mrs. Du Huaqiong** for offering me the best opportunity to learn the worldly wisdom in the overseas and for encouraging and supporting me to overcome difficulties in my life bravely and successfully; my husband **Mr. Roland Cuny** for helping me to open the door to European society; my guest family, **Dr. Wolfgang Schaub** and **Mrs. Anje Schaub** for taking care of me as my German foster parents. Without all of them this thesis would not be……..

# Summary

New effective experimental techniques in medicinal chemistry and pharmacology have resulted in a vast increase in the number of pharmacologically interesting compounds. However, the possibility of producing drug candidates with optimal biopharmaceutical and pharmacokinetic properties is still improvable. A large fraction of typical drug candidates is poorly soluble in water, which results in low drug concentrations in gastrointestinal fluids and related acceptable low drug absorption. Therefore, gaining knowledge to improve the solubility of compounds is an indispensable requirement for developing compounds with drug-like properties.

The main objective of this thesis was to investigate whether computer-based models derived from calculated molecular descriptors and structural fragments can be used to predict aqueous solubility for drug-like compounds with similar structures. For this purpose, both experimental and computational studies were performed. In the experimental work, a novel crystallization method for weak acids and bases was developed and applied for European patent. The obtained crystalline materials could be used for solubility measurements. A novel recognition method was developed to evaluate the tendency of compounds to form amorphous forms. This method could be used to ensure that only solubilities of crystalline materials were collected for the development of solubility prediction. In the development of improved in silico solubility models, lipophilicity was confirmed as the major driving factor and crystal information related descriptors as the second important factor for solubility. Reasons for the limited precision of commercial solubility prediction tools were identified. A general solubility model of high accuracy was obtained for drug-like compounds in congeneric series when lipophilicity was used as descriptor in combination with the structural fragments. Rules were derived from the prediction models of solubility which could be used by chemists or interested scientists as a rough guideline on the contribution of structural fragments on solubility: Aliphatic and polar fragments with high dipole moments are always considered as solubility enhancing. Strong acids and bases usually have lower intrinsic solubility than neutral ones. In summary, an improved solubility prediction method for congeneric series was developed using

high quality solubility results of drugs and drug precursors as input parameter. The derived model tried to overcome difficulties of commercially available prediction tools for solubility by focusing on structurally related series and showed higher predictive power for drug-like compounds in comparison to commercially available tools. Parts of the results of this work were protected by a patent application[1], which was filed by F. Hoffmann-La Roche Ltd on August 30, 2005.

# Contents

**1 INTRODUCTION 13**

**2 CRYSTALLIZATION 17**

***2.1 Introduction 17***

***2.2 Materials and methods 18***

2.2.1 Materials 18
2.2.2 Methods 19

***2.3 Results and discussion 29***

2.3.1 Crystallization of known drugs 29
2.3.2 Crystallization of internal development compounds 30
2.3.3 Crystallization of external polymorphs 32

**3 EVALUATION OF THE TENDENCY TO FORM AMORPHOUS MATERIAL 39**

***3.1 Introduction 39***

***3.2 Materials and methods 40***

3.2.1 Materials 40
3.2.2 Methods 42

***3.3 Results and discussion 47***

3.3.1 Bosentan 47
3.3.2 Trazodone 50
3.3.3 Glibenclamide and Iopanoic acid 51

***3.4 Conclusion 52***

**4 AQUEOUS SOLUBILITY PREDICTION OF DRUG-LIKE COMPOUNDS 55**

***4.1 Introduction 55***

4.1.1 Solubility prediction tools 55

4.1.2 Objectives 67

*4.2 Materials and methods* ***67***

4.2.1 Materials 67
4.2.2 Methods 90

*4.3 Results and discussions* ***91***

4.3.1 Melting point prediction 91
4.3.2 Solubility prediction considering crystal structure information 96
4.3.3 Solubility prediction using melting point as a parameter 102
4.3.4 Solubility prediction for drugs in congeneric series 104
4.3.5 The impact of solid state on solubility 124

**5 CONCLUSIONS AND OUTLOOK 139**

**6 ABBREVIATIONS 141**

**7 REFERENCES 143**

**8 APPENDIX 153**

# 1 INTRODUCTION

The therapeutic effect of a drug is based on the interaction between the drug and its specific receptor. Its strength and duration depend on the concentration of the drug near the receptor and the stability of the drug binding to the receptor. In order to reach the necessary concentration at the receptor site, the drug must be dissolved in the gastrointestinal (GI) tract at first and traverse several membrane barriers. In other words, the drug must be sensibly absorbed at first. Therefore, a good absorption is one precondition for high drug concentrations in the biophase. (Figure 1)

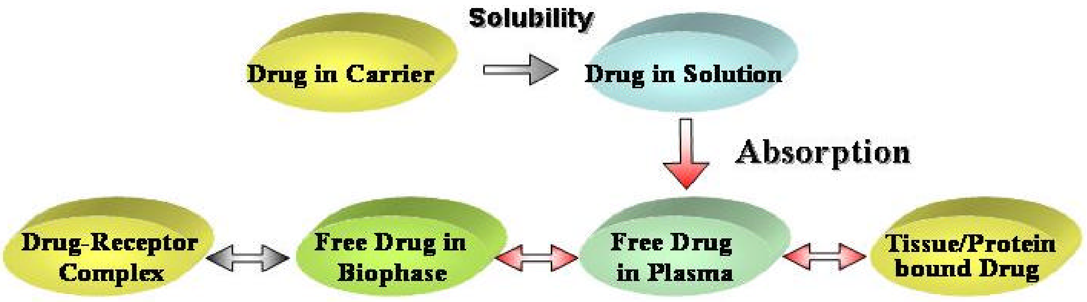


Figure 1: The influence of the absorption on the drug therapeutic effect[2].

Gastrointestinal absorption is dominated by passive uptake in the jejunum and ileum because of their high surface area. The majority of orally administered drugs are absorbed via the passive transcellular route[3], so that in most situations, the intestinal absorption can be simplified as a passive diffusion process of a solute through the membrane. Such a simplified transport model can be described with Fick's first law, in which the flux equation reduces to a product of permeability and solubility, when certain assumptions are made. In case of an ionizable molecule, the permeation by passive diffusion can only be very efficient, when the molecule is in its uncharged form at the membrane surface[4]. The amount of the uncharged form at a given pH depends on several important factors, such as pH, binding to indigenous carriers (proteins and bile acids), self-binding (aggregate or micelle formation), and solubility (a solid-state form of self-binding)[4]. Thus, low aqueous solubility is usually related to low drug concentrations in gastrointestinal fluids, which can lead to impaired drug absorption. Therefore, gaining knowledge in improvement of solubility is a key prerequisite for successful development of drugs.

Computational models for the prediction of aqueous solubility from electrotopology, molecular surface areas, lipophilicity, and hydrophilic measures have been devised, and several of these show impressive statistics[5-10]. However, all tools either commercially available or published by academia are usually restricted to deal with non drug-like molecules due to the limited number of published solubility data of drug-like compounds. Thus, developing a structurally based solubility prediction tool with high predictive power is an absolute necessity to give medicinal chemists constructive feedback on how to design better drug-like compounds with improved solubility.

The main objective of this thesis was to investigate whether computer-based models derived from calculated molecular descriptors and structural fragments can be used to predict aqueous solubility for drug-like compounds with similar structures. For this purpose, both experimental and computational studies were performed. One objective in the experimental work was directed to the development of a novel crystallization method for weak acids and bases. Thus, crystalline material could be obtained for solubility measurements. Another objective in the experimental work was to develop a novel method to evaluate the tendency of compounds to form amorphous materials. This method could be used to ensure that only solubilities of crystalline materials were collected for the development of solubility prediction. The goal in the computational work was to find suitable descriptors for solubility prediction of drug-like compounds in congeneric series. The influence of crystal lattice on solubility was evaluated using compounds with information related to solid state, e.g. with known crystal structure or melting point. One of the obtained models was modified and improved using an extended dataset to predict the solubility of drug-like compounds in congeneric series. As a result of the improved prediction tool, structurally based solubility rules were derived, which can be the basis for the guidance of decision processes in the synthesis of more soluble drug-like candidates.

# Experimental Part

# 2 CRYSTALLIZATION

## 2.1 Introduction

Crystallization is an important purification and separation technique in a variety of commercial processes, as for example biotechnology, mineral processing, waste treatment, energy storage, production of new materials and electronic chemicals[11]. Crystallization can occur in solution, from vapor or from melt. Most processes in the chemical industries use crystallization from solution. The starting point for crystallization is the creation of a saturated solution. However, formation of a saturated solution is often a time-consuming process. Usually, it takes days until the equilibrium between the compound's soluble and insoluble forms has been reached. Hence, the most currently known methods use a supersaturated solution, instead of a saturated one, as the starting point for the crystallization. In such cases, it is important to know the level of supersaturation, since supersaturation appropriate for crystallization varies from compound to compound. In general, with decreasing level of supersaturation, the crystal growth becomes slower and the crystal quality improves[12].

Crystallization using pH variation is a well-known method for proteins[13-18], but rarely for drug molecules. Among the few publications found to use pH variation for drug crystallization, nicotinic acid was an example. Wang[19] tried to obtain highly supersaturated solution of nicotinic acid by adding hydrochloride acid to an aqueous sodium nicotinate solution, which was then used as the starting point for the crystallization of nicotinic acid. Controlled batch crystallization by pH variation was another example developed by Zhu[20]. According to Zhu[20], crystallization was initialized using a short pulse of supersaturation. pH was modified, during the whole crystallization process, in order to maintain a constant level of supersaturation. Furthermore, Zhu[20] tried to raise the level of supersaturation to the highest concentration, in order to shorten the operation time. However, supersaturation could also be a risk for the formation of amorphous materials and the occurrence of crystal defects[21]. An alternative method is using a saturated solution. However, up to now, there is no scientific-based method available to identify the condition of formation of saturated solution. Thus crystallization via saturated solutions is considered as impracticable for commercial purposes.

This study closes the aforementioned gap in knowledge and application of saturated solution. For the first time, the invented new crystallization method successfully enables the generation of saturated solutions using pH-variations. Fine granular pH variations are applied to obtain the saturated solution. Crystallization process can be smoothly initialized and controlled, a key prerequisite for further optimization and application in a productive commercial environment. The breakthrough advantages of the new method are summarized as following:

- Avoidance of buffer systems in the crystallization of compounds.
- Improved control of crystal growth due to the use of the saturated solution.
- Reduction of the possibility to obtain non-crystalline (amorphous) materials.

## 2.2 Materials and methods

### 2.2.1 Materials

Diclofenac, famotidine, flurbiprofen, furosemide, hydrochlorothiazide, ketoprofen, propranolol, quinine were commercial compounds used for crystallization.

Cyclopenthiazide and codeine are compounds with known polymorphic forms. Additionally, an internal compound with known polymorphs was included in the study. Their solubilities were determined via a potentiometric method. Crystalline materials of all compounds were successfully obtained using the invented new crystallization method.

The pSol[22] instrument usually foreseen for the potentiometric solubility measurements was used here to study crystallization processes. The pH-solubility profile obtained via pSol[22] delivered a plot of pH against solubility, which was the key procedure for planning the crystallization experiments.

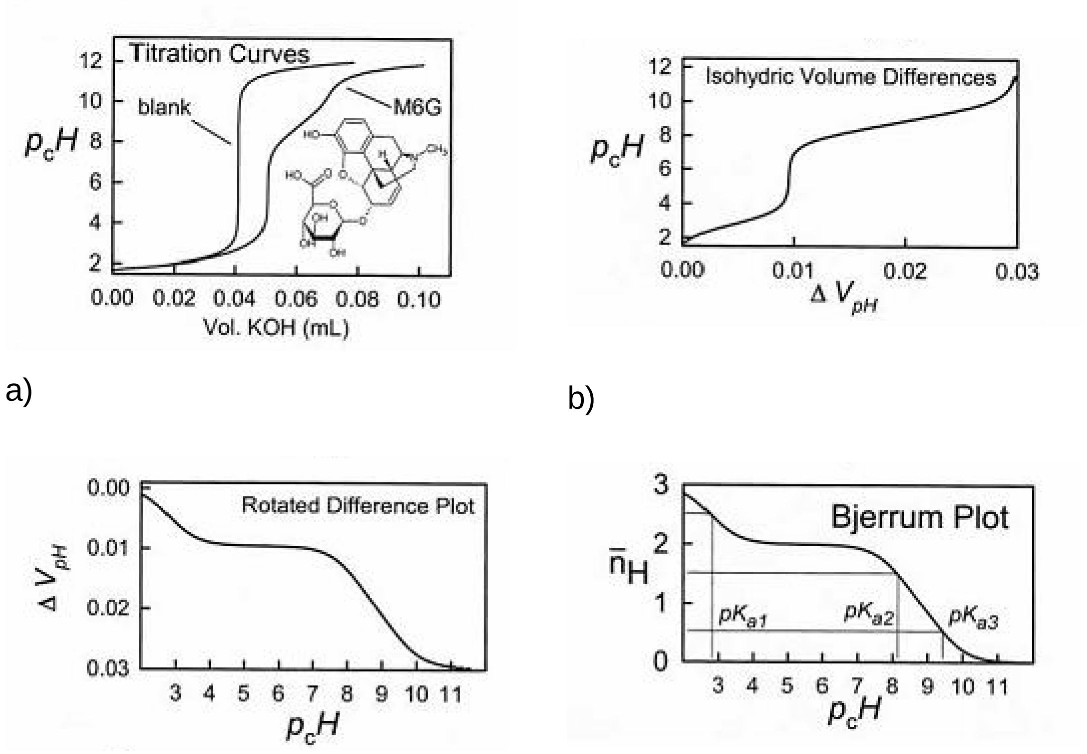
## 2.2.2 Methods

### 2.2.2.1 p$K_a$ assays applied

A potentiometric titration method was used for the p$K_a$ determination of UV inactive compounds via the GLpKa[23] equipment and a photometric method for UV active compounds via the Profiler SGA[24] equipment. The methods are described in detail, in order to explain potential restrictions.

#### 2.2.2.1.1 Potentiometric Determination

Usually, a blank titration is performed at the beginning of the measurement to calibrate the electrode. Afterwards, precisely known volumes of a standardized strong acid or base are added to a vigorously-stirred solution of a protogenic substance, while the pH is continuously measured with a pH-electrode. The results of an experiment deliver two potentiometric titration curves, one with and one without sample as shown in Figure 2a[4].



a)

b)

c)

d)

Figure 2: Four step construction of the Bjerrum difference plot for a molecule with three $pK_a$ values, whose constants are observed in the simple titration curve[4].

The potentiometric titration curve depicts the measured pH against titrant volume added. The shape can give information on the amount of substance present and its characteristic acid-base ionization properties. To reveal overlapping $pK_a$s, it is necessary to transform the titration curves into Bjerrum plots. Such a plot can be obtained by subtracting a titration curve containing no sample, “blank” titration, (left curve in Figure 2a), from a titration curve with sample, (right curve in Figure 2a), at fixed pH values. The difference between the total and the free concentrations is equal to the concentration of the bound hydrogen ions. The latter concentration divided by that of the sample gives the average number of bound hydrogen atoms per molecule of substances, $\bar{n}_H$. The Bjerrum curve is a plot of $\bar{n}_H$ vs. $p_cH$. It reveals all the $pK_a$s as $p_cH$ values at half-integral $\bar{n}_H$.

#### 2.2.2.1.2 Spectrophotometric determination of ionization constants

The spectrophotometric method is based on the multiwavelength spectrophotometric approach from Tam and coworkers[25]. A UV light source, a fiber optic dip probe and a diode array detector are used to monitor the spectral changes that arise in the course of pH-metric titration of an ionizable compound. At the end of the measurement, TFA (target factor analysis)[25] is used to calculate $pK_a$ values from the multi-wavelength spectrophotometric absorption titration data.

### 2.2.2.2 Solubility assay applied

#### 2.2.2.2.1 Equilibrium solubility measurement by shake flask

The measurement requires different diluted DMSO stock solutions for the calibration and a saturated buffer solution. A saturated buffer solution is obtained by adding a compound to a standard buffer solution until saturation occurs, indicated by undissolved excess of the compound. The thermostated saturated solution is shaken until equilibration between the solution and the solid phase is established. After micro-filtration or centrifugation, the concentration of the substance in the supernatant solution is determined using HPLC, usually via UV detection. (Figure 3)

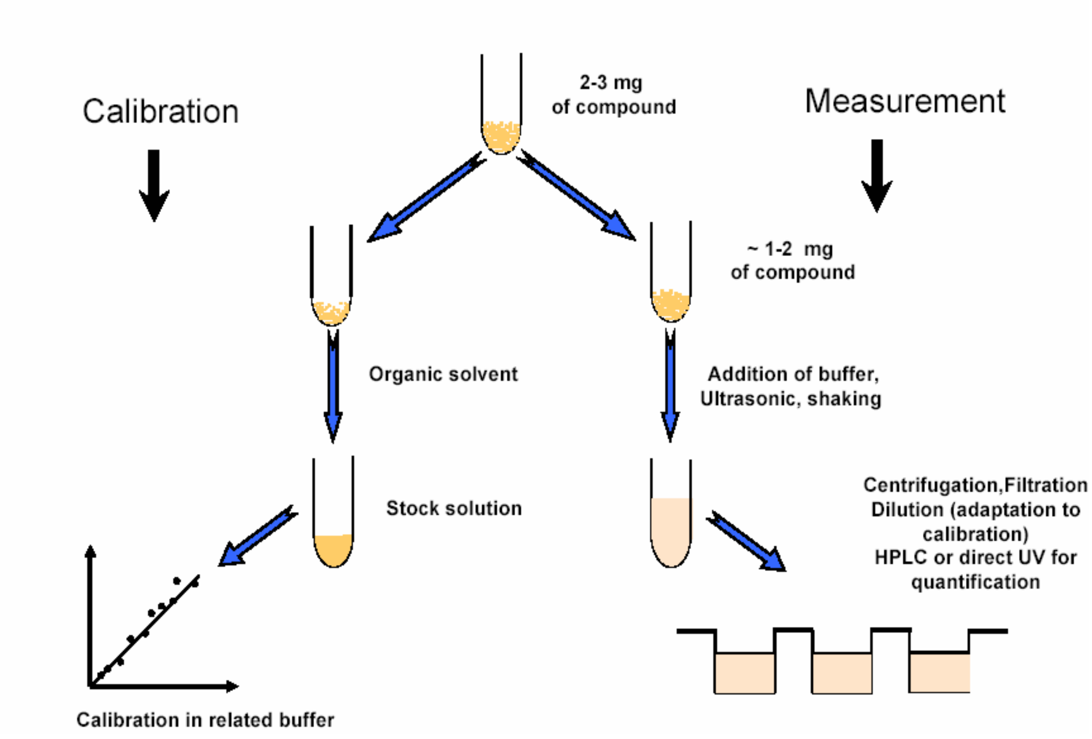


Figure 3: The principal steps of an equilibrium solubility measurement[26].

In order to evaluate the experimental error of the Shake-Flask measurements, solubility of 6 drugs were measured five times. (Table 1).

| Name | ApK$_a$1 | BpK$_a$1 | BpK$_a$2 | MW | Exp1 | | | Exp2 | | | Exp3 | | | Exp4 | | | log1/S average | SE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | S (µg/mL) | log1/S | pH | S (µg/mL) | log1/S | pH | S (µg/mL) | log1/S | pH | S (µg/mL) | log1/S | pH | | |
| Mefenamic acid | 4.2 | 0 | 0 | 241 | 11.42 | 4.32 | 6.5 | 2.78 | 4.94 | 6.5 | 3.95 | 4.79 | 6.5 | 5.2 | 4.67 | 6.5 | 4.68 | 0.19 |
| Flurbiprofen | 4.03 | 0 | 0 | 244 | 2050.25 | 2.08 | 6.3 | 1128.56 | 2.33 | 6.3 | 1387.95 | 2.25 | 6.3 | 1442.54 | 2.23 | 6.3 | 2.22 | 0.07 |
| Astemizole | 0 | 9.93 | 8.87 | 458 | 3.19 | 5.16 | 6.5 | 11.28 | 4.61 | 6.5 | 19.90 | 4.36 | 6.5 | 12.44 | 4.57 | 6.5 | 4.67 | 0.24 |
| Terfenadine | 0 | 9.53 | 0 | 471 | 2.48 | 5.28 | 6.7 | 5.41 | 4.94 | 6.6 | 11.06 | 4.63 | 6.5 | 7.81 | 4.78 | 6.6 | 4.91 | 0.20 |
| Warfarin | 0 | 0 | 0 | 308 | 157.91 | 3.29 | 6.5 | | | 6.5 | 129.24 | 3.38 | 6.4 | 123.79 | 3.40 | 6.5 | 3.35 | 0.04 |
| Iopanoic acid | 4.5 | 0 | 0 | 570 | 38.02 | 4.18 | 6.5 | 15.63 | 4.56 | 6.5 | 23.39 | 4.39 | 6.5 | 19.50 | 4.47 | 6.5 | 4.40 | 0.12 |

Table 1: Equilibrium solubility measurements of 6 drugs, which are acids, bases and neutral compounds in the solubility range from low to high.

The compounds included acids, bases and neutral ones with a solubility range from low to high. The experimental standard error was calculated with Eq. 1 for each compound.

$$SE = \frac{\sum_{i=0}^{n} ABS\left|\log 1/S_n - \log 1/S_{avg}\right|}{n} \qquad \text{Eq. 1}$$

The average standard error for the performed measurements was ±0.143 for the analyzed data set.

### 2.2.2.2.2 Potentiometric solubility assay

The potentiometric solubility assay was first described by Avdeef[22]. The measurement via this method requires an ionizable compound as reactant and a strong acid or a strong base as titrant. A blank titration is performed at the beginning of the measurement, similar to the procedure described for the determination of ionization constants. Afterwards, a certain amount of compound is placed in a reaction beaker and dissolved in a given volume of solvent. A titration is then performed in the direction of complete dissolution. During the measurement, the pH value is continually determined via a pH electrode. Similar to the potentiometric $pK_a$ assay, two potentiometric titration curves are obtained, and the corresponding Bjerrum plot is derived. Thus, the value of apparent $pK_a$, ($pK_a^{App}$), can be determined at the half-integral $\bar{n}_H$ positions of Bjerrum plot. In case of weak acid, the apparent ionization constant, $K_a^{App}$, is defined as Eq. 2[22].

$$\begin{aligned} K_a^{APP} &= \frac{[A^-][H^+]}{([HA]+[HA]_s)} \\ &= K_a \frac{[HA]}{([HA]+[HA]_s)} \end{aligned} \qquad \text{Eq. 2}$$

[HA] is the concentration of the molecule HA in the solution. $[HA]_{(s)}$ is the moles of the molecule HA, which precipitated per liter of aqueous solution.

At the half-integral $\bar{n}_H$ positions of Bjerrum plot, half of the total amount of the substance is protonated, thus, the concentration of the free acid, HA, equals that of the conjugate base, $A^-$. (Eq. 3)

$$[HA]+[HA]_s=[A^-]=\frac{C}{2} \qquad \text{Eq. 3}$$

$[A^-]$ is the concentration of the conjugate base, $A^-$ in the solution. C is the total amount of substance in the solution and solid phase.

Combining Eq. 2 and Eq. 3 together, the value of intrinsic solubility can then be deduced from the Eq. 4 using the experimentally determined $pK_a$ value and the sample concentration, C, as input parameters.

$$\log S_0=\log[HA]=\log\frac{C}{2}-pK_a^{App}+pK_a \qquad \text{Eq. 4}$$

### 2.2.2.3 Description of the crystallization assay

Crystallization is considered as a kinetic process and illustrated with the help of the pH-solubility profile using a weak base as an example. (Figure 4)

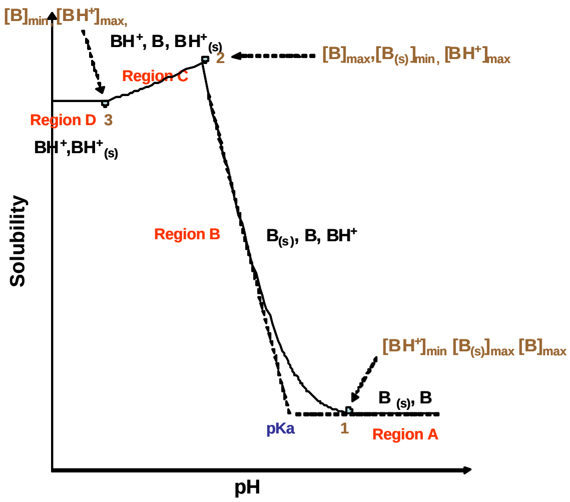


Figure 4: Solubility-pH profile of a weak base. B is the soluble form of the weak base. $B_{(s)}$ is the solid form of the weak base. $BH^+$ is the charged form of the weak base.

In region A, the compound is in equilibrium and solubility stays constant. Eq. 5 describes the equilibrium in region A.

$$BH^+ \rightleftharpoons B \rightleftharpoons B_{(s)} \qquad \text{Eq. 5}$$

In region B, solubility rises with the increasing amount of $BH^+$, when pH changes from high to low. This means, when a basic compound is titrated from its insoluble to its soluble form, an increasing amount of uncharged precipitate $B_{(s)}$ will go into solution with increasing hydrogen concentration $[H^+]$. This will continue, until point 2 is reached. At point 2, a "perfect" buffer system[4] exists. The simultaneous presence of solid free base and its solid conjugate acid force the pH and solubility to be constant, as long as the two interconverting solids are present. This special pH point has been designated as the Gibbs' $pK_a$ ($pK_a^{GIBBS}$)[4]. The equilibrium equation associated with this phenomenon is Eq. 6.

$$BH^+_{(s)} \rightleftharpoons B_{(s)} + H^+ \qquad \text{Eq. 6}$$

$$K_a^{GIBBS} = \frac{\{H^+\}\{B_{(s)}\}}{\{BH^+_{(s)}\}} \qquad \text{Eq. 7}$$

The solubility at point 2 is $S = S_0 + S_i$. The constants $S_0$ and $S_i$ are intrinsic and salt solubility[4].

From point 2 on, (in region C), $BH^+_{(s)}$ will be only won in credit of B in the solution and solubility decreases with $B + H^+ \rightarrow BH^+_{(s)}$, until region D is reached. In region D, no more B will be changed into $BH^+_{(s)}$ and the minimum of [B] is achieved. The equilibrium existing there is described by Eq. 8:

$$BH^+_{(s)} \rightleftharpoons BH^+ \rightleftharpoons B + H^+ \qquad \text{Eq. 8}$$

However, during the potentiometric titration, one is not very frequently able to observe the phenomenon of the “perfect” buffer system, because in order to get a good titration, it is always recommended to use a small amount of compound. And this leads to the situation, that the whole amount of compound is dissolved before the maximal concentration of salt in solution ($[BH^+]_{max}$ at point 2) is reached. The point where the whole amount of compound dissolved in the solution is signified as 2’ in Figure 5; at this point, the compound reaches its total solubility. The total solubility does not change with pH and is signified in Figure 5 by the blue dashed line.

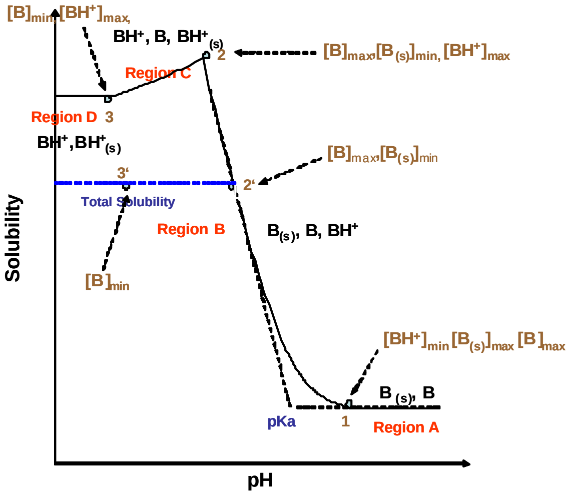


Figure 5: Solubility-pH profile of a weak base. B is the soluble form of weak base. $B_{(s)}$ is the solid form of weak base. $BH^+$ is the charged form of weak base.

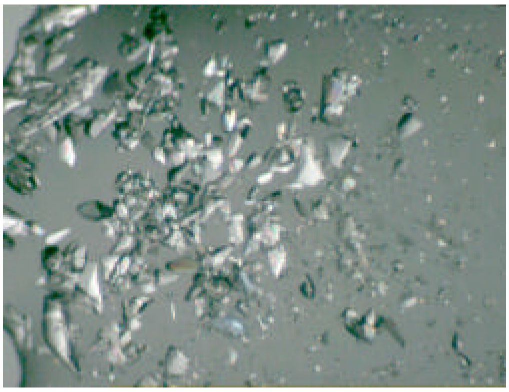
According to the new crystallization method, crystals can be easily obtained, when the direction of titration described above is reversed. In the case of a weakly basic compound B, one starts with an unsaturated solution of the compound at a low initial pH-value as illustrated in Figure 5 by means of point 3', which can be varied by the amount of the compound used. Subsequently, the pH value is gradually increased by adding a strong basic titrant to the solution. This leads to an increasing deprotonation of $BH^+$ to B, but initially, there is no precipitation of solid phase. By reaching point 2', the titration is stopped. At this target point 2', the concentration of the uncharged form has reached its maximal value $[B]_{max}$, which is equal to the intrinsic solubility $S_0$. Therefore, a saturated solution of the compound of interest has been reached that may serve to carry out a crystallization under substantially saturated conditions. Hence, at the point 2', the probability for the formation of the neutral form is at its maximum.

The point of saturation can be precisely identified via the pH-solubility profile and can be easily reached using pH-titration. Therefore, generation of saturated solution is no more a time-consuming process since the advent of the invented new crystallization method. Furthermore, the newly developed method uses saturated solution instead of highly supersaturated solution as the starting point for the crystallization. Hence, the shortcoming of the currently known method can be avoided. The control of the crystal growth can be improved and the possibility to obtain non-crystalline form can be reduced using the new crystallization method. However, due to practical limitations, it may be difficult to reach the target point 2' very precisely. If too much base is added, the pH-value goes beyond the targeted pH-value corresponding to point 2' and a supersaturated solution is formed. Therefore, the titration is usually stopped at a point very close to the solubility-pH profile that corresponds to a slightly unsaturated solution. By keeping the solution at defined conditions allowing controlled slow solvent evaporation, the concentration of the solution will slowly increase so that the saturated state is reached. In order to obtain good crystallization results and reduce the risk of forming amorphous solid materials, an improved system might be of advantage for monitoring the concentration of the uncharged form and regulating the pH-value so that the concentration of the uncharged form is kept within a predefined tolerance range above the intrinsic solubility. Alternatively, the improved system may monitor the total concentration of the compound and regulate the pH-value so that the total concentration is kept within a predefined tolerance range above the predetermined total solubility profile.

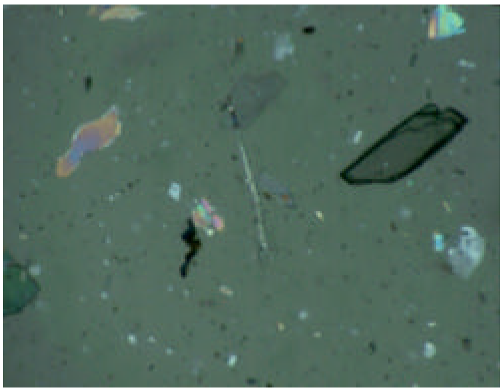
# 2.3 Results and discussion

## 2.3.1 Crystallization of known drugs

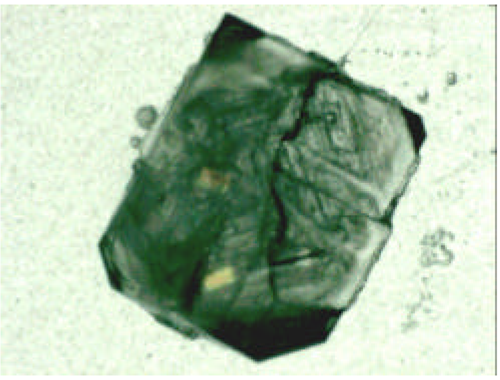
The known drugs famotidine, diclofenac, flurbiprofen, furosemide, hydrochlorothiazide, ketoprofen, propranolol and quinine were used to verify the readiness and applicability of the new crystallization method. Crystals with high quality were obtained and their microscopic pictures are depicted in Figure 6.



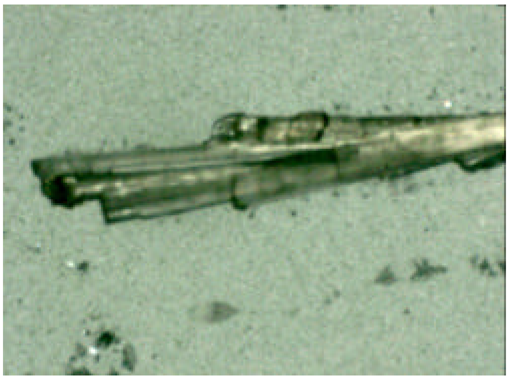
Famotidine



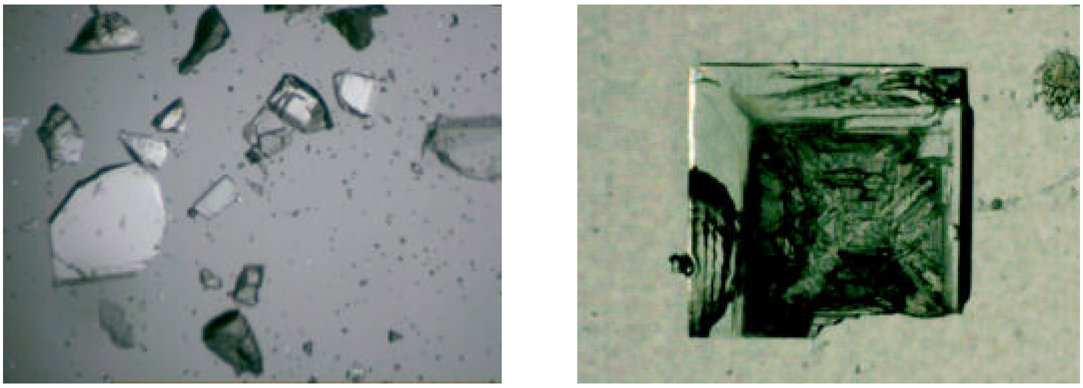
Diclofenac



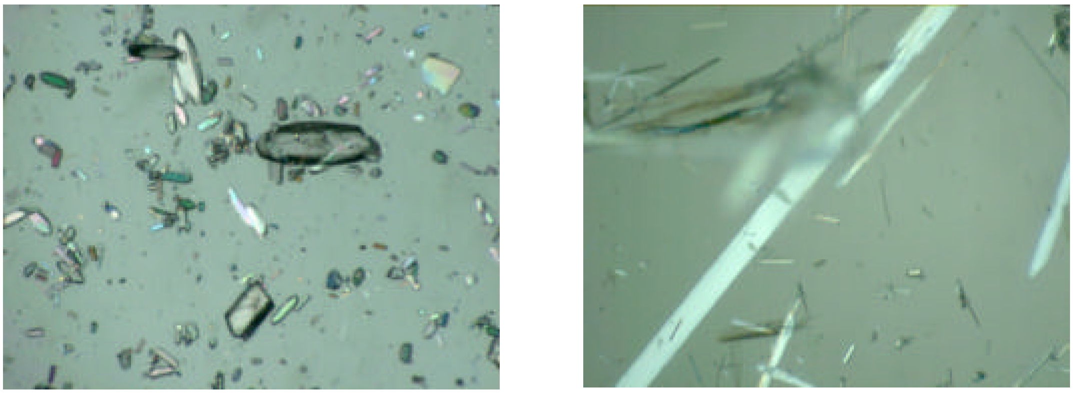
Flurbiprofen



Furosemide



Hydrochlorothiazide Ketoprofen



Propranolol Quinine

Figure 6: Crystals of diclofenac, famotidine, flurbiprofen, furosemide, hydrochlorothiazide, ketoprofen, propranolol, and quinine obtained using the new crystallization method.

### 2.3.2 Crystallization of internal development compounds

The application of the newly developed crystallization method was further extended to several development compounds. Compound 1 is an internal compound with ability of forming amorphous and polymorphic forms. Its $pK_a$, equilibrium and potentiometric results are summarized in Table 2.

| | | | | Equilibrium solubility | | | | pSol |
|---|---|---|---|---|---|---|---|---|
| Name | Polymorphic forms | $ApK_a1$ | $BpK_a1$ | S (µg/mL) | pH | Solution | $S_0$ (µg/mL) | $S_0$ (µg/mL) |
| compound 1 | Mod C | 7.62 | 4.01 | 43 | 9.6 | Boric acid/KCl-NaOH pH=10 | 0.5 | 0.135 |

Table 2: Solubility measured using equilibrium and potentiometric methods.

Three polymorphic forms are known for compound 1. Among them, modification A is known to be the most stable form, which is formed via transition from modification B; modification B is an anhydrate and modification C is hydrate. The crystals obtained via the crystallization method had the form of yellow needles. (Figure 7) Through the comparison with the reference data, the obtained crystals were characterized by powder diffraction as modification C. (Figure 8 and Figure 9)

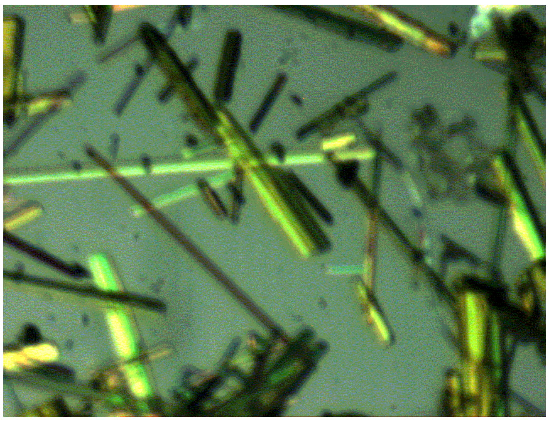


Figure 7: Crystals of compound 1.

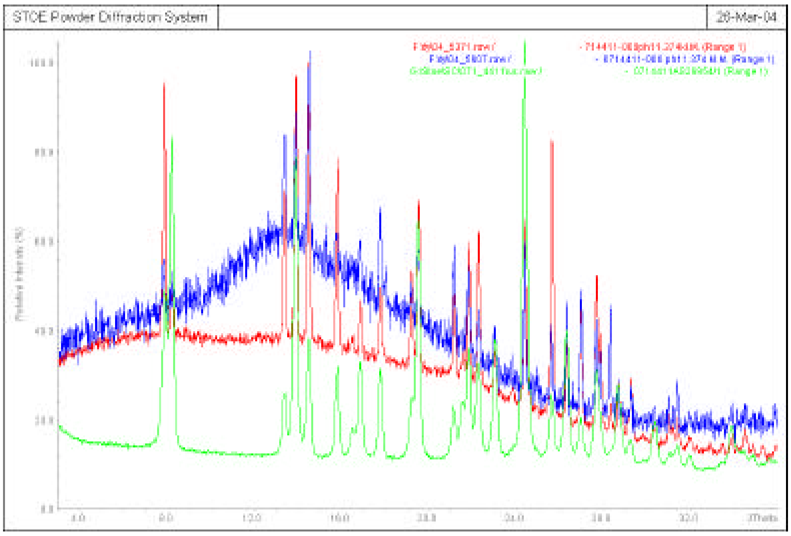


Figure 8: Powder diffraction diagram of crystal forms (red and blue) of compound 1 obtained via crystallization method. The obtained crystals show the same diffraction pattern as crystals in modification C (green) in reference diagram.

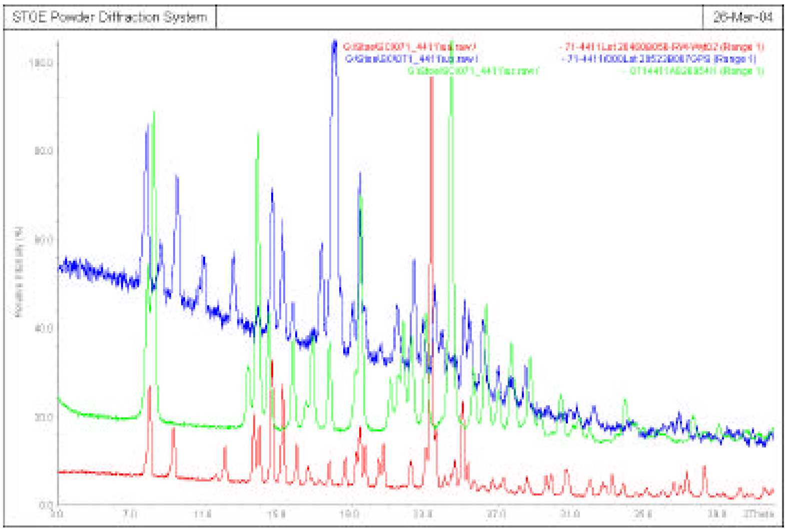


Figure 9: Powder diffraction diagram of three different crystal forms of compound 1. They are used as reference diagrams in order to identify the crystal form obtained via the new crystallization method. In the reference diagram the form A is colored in red, form B in blue and form C in green.

The result of the crystallization of modification C confirmed the readiness of the new crystallization method. Since the method is based on titration in aqueous solution, hydrates are usually obtained.

## 2.3.3 Crystallization of external polymorphs

Pudipeddi[27] has shown that in a data set of 72 compounds with different polymorphic forms, usually small differences in their solubilities were determined. The described differences in the solubilities were often in the range of the experimental error of the solubility measurements. Extensive literature searches were performed in order to identify compounds showing large differences in the measured solubilities of their polymorphic subtypes. Several interesting drugs could be found with much larger solubility differences. One of those examples is premafloxacin.[28] There is a 30 fold solubility difference described between polymorphic form I and III of premafloxacin. The other examples are codeine[29] and cyclopenthiazide[30] with a 13 fold difference for codeine between hydrate and other crystal forms. A 4 fold difference was described between the polymorphic form II and III of cyclopenthiazide.

Except for premafloxacin, cyclopenthiazide and codeine were available for further characterization and application of the new crystallization method.

The physicochemical properties of the three cyclopenthiazide polymorphic forms, according to Gerber[30], are summarized in Table 3:

| Polymorph | Melting point (°C) | Solubility in water (µg/mL) |
|---|---|---|
| I | 239.33 | 34.7 |
| II | 223.03 | 61.8 |
| III | 187.87 and 233.48 | 17.15 |

Table 3: Physicochemical properties of diverse cyclopenthiazide polymorphic forms.

After the crystallization, white needles were obtained (Figure 10) with a melting point of 233°C. No powder diffraction diagram was described for cyclopenthiazide by Gerber[30]. Therefore a direct comparison between obtained crystals and those described in the literature was not possible. Based on the similarity in melting points, it can be assumed that the obtained crystals belonged to the polymorphic form III, the most stable one with the lowest solubility.

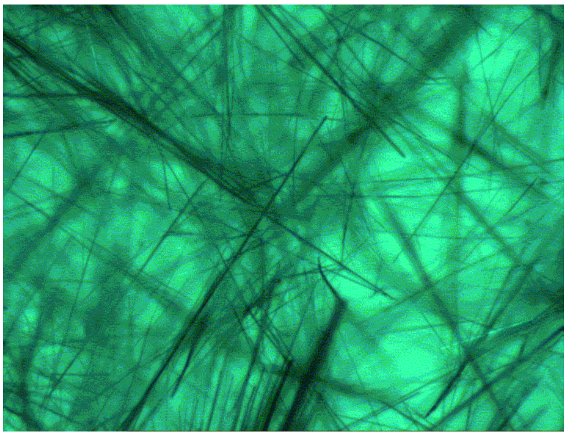


Figure 10: Crystals of cyclopenthiazide.

El-Gindy[29,31] described the solubility of three polymorphic forms of codeine, which are summarized in the Table 4.

| Polymorph | Solubility in water (g/mL) |
|---|---|
| I | 8.103 |
| II | 11.123 |
| III | 80.431 |

Table 4: Solubility of diverse polymorphic forms of codeine.

White needles were obtained by applying the new crystallization method. (Figure 11b) Its 3D structure was solved by single crystal X-ray analysis. (Figure 11a and c)

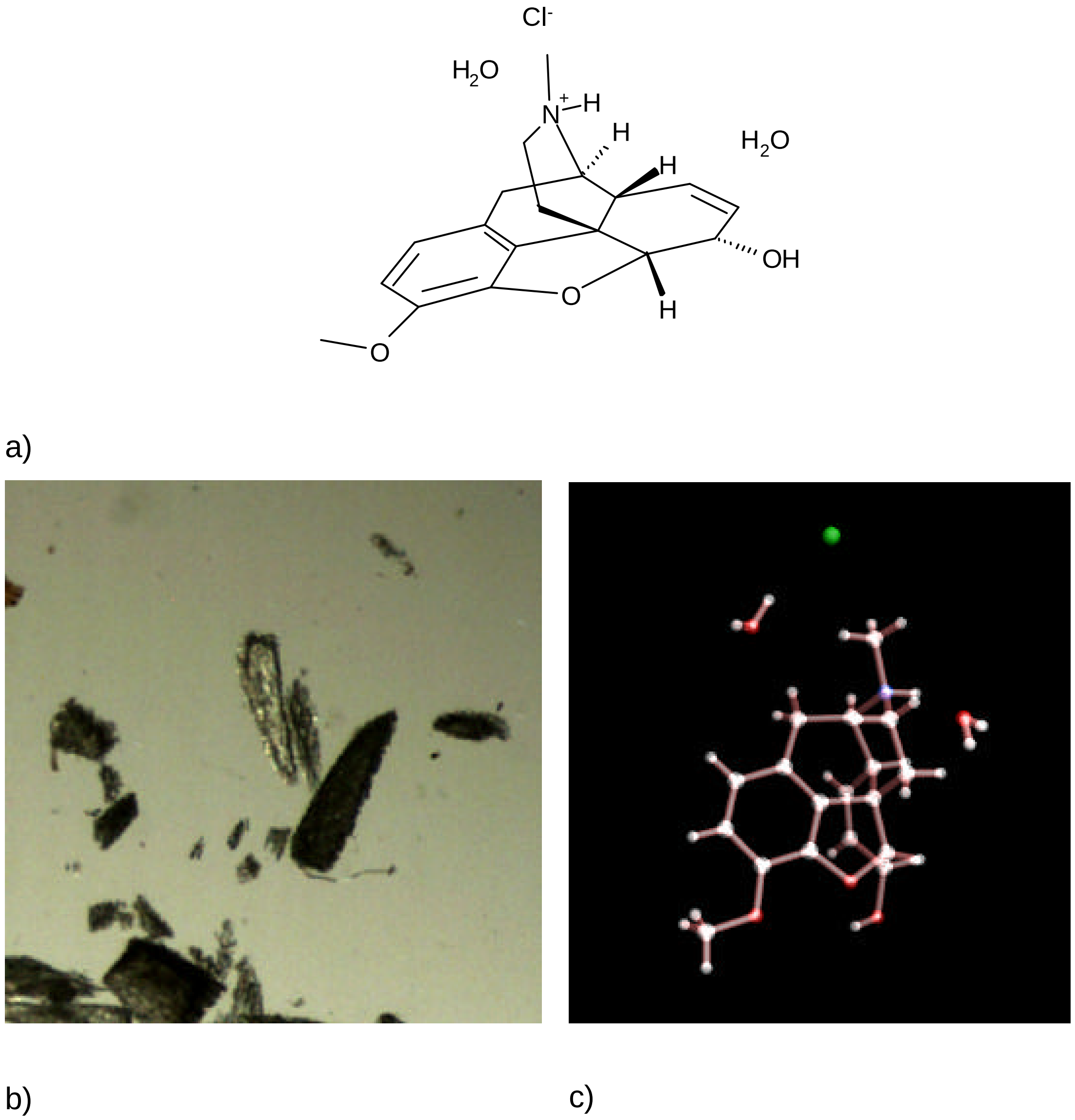


Figure 11: Codeine crystals obtained by the new method. a) 2D structure; b) microscopic photo; c) the 3D structure identified by the single crystal X-ray analysis.

Figure 11 shows, that instead of the free basic form, a chloride salt form of codeine with two water molecules in crystal packing was obtained.

In this case, the formation of codeine salt solids shows the practical limitations of the newly developed crystallization method. As in chapter 2.2.2.3 described, the titration is stopped at a point very close to the pH-solubility profile which corresponds to a slightly unsaturated solution. During the slight evaporation process, the concentration of the solution increases, until the saturated state, resp. a point 2' in the Figure 12 is reached.

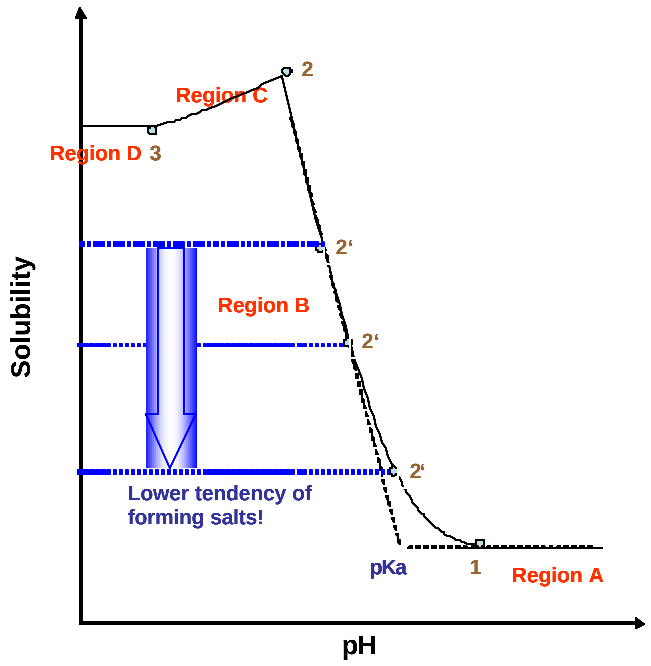


Figure 12: Solubility-pH profile of a weak base. Reduction in concentration of a compound leads to lower tendency of forming salts.

The point 2’ is considered as the starting point for the crystallization. Dependent on the concentration of the compound and the relative orientation of the point 2’, the concentration of counter ions can be high, which can influence the crystallization process and the obtained crystal form. When crystallization starts at point 2, then the crystallization process is a competition between charged and uncharged form. The charged salt solid has generally a stronger crystal lattice than the uncharged solid form, because of the strong ionic interactions between the cations and anions. Thus, the closer 2’ moves to the point 2, the higher the tendency of obtaining salt solid than the formation of uncharged materials. Therefore, in order to enhance the possibility of obtaining an uncharged form, a reduced concentration of the compound is recommended to be utilized for the crystallization, i.e. 2’ should be sufficiently kept away from 2.

In conclusion, a new crystallization method has been developed for weak acidic and basic compounds. According to this method, one can rapidly proceed to a situation in which the solution is in a substantially saturated state, by gradually changing the pH-value of the solution in a direction that leads to a decrease of said compound’s solubility. In particular, one can avoid the drawbacks associated with crystallization from a supersaturated state, because crystallization is then carried out under the most desirable conditions, by maintaining the solution in a substantially saturated state. According to the described results, there is a high probability to get hydrates with low solubilities by the described method. However the example of codeine shows that further optimization of the method can probably improve the results. In all the eleven analyzed cases, crystals could be obtained easily via the pH-solubility profile using the sample concentration and experimental $pK_a$ as input parameters.

# 3 EVALUATION OF THE TENDENCY TO FORM AMORPHOUS MATERIAL

## 3.1 Introduction

Amorphous solids, or glasses, are phase intermediates between solids and liquids. The atoms in an amorphous solid are aligned in a rigid disordered structure, instead of a regular lattice like ordinary ("crystalline") solid. Various degrees of disorder in the solid form result in inconsistent properties of amorphous solids in comparison to their cystalline counterparts. Additionally, the instability of amorphous solids may lead to crystallization after long time of storage. Therefore, it is often a significant risk for pharmaceutical industry to produce amorphous instead of crystalline solids for medicines. The following chapter investigates the tendency of compounds to form amorphous materials. The achieved results help to improve the design and production of pure and stable pharmaceuticals.

The formation of amorphous solids is firstly dependent on the condition of crystal growth. For example, a crystal system can be driven by a high degree of supersaturation to an order-disorder transition, resulting in an amorphous solid[32-35]. Secondly, the formation of amorphous solids is compound-specific[21]. For example, relatively large molecules and molecules with a certain degree of rotational flexibility tend to form a disordered state even at mild crystallization conditions[21]. Therefore, being able to identify the degree of supersaturation, is helpful in reducing the possibility of obtaining amorphous materials and crystal defects. Hence, the first goal of this work was directed toward the evaluation of potential rules for the formation of amorphous materials.

High Throughput (HT) solubility assay uses freeze-drying procedure to eliminate DMSO from the stock solution. Prepared solids can be used as basis for the solubility determination. Usually, the solubility results of equilibrium and HT-solubility assays are similar, but can be different in specific cases if the characteristics of the solid forms change during the evaporation process in HT-solubility measurements.

Therefore, the second direction of this work is researching potential differences in the results of equilibrium and HT-solubility measurements.

## 3.2 Materials and methods

### 3.2.1 Materials

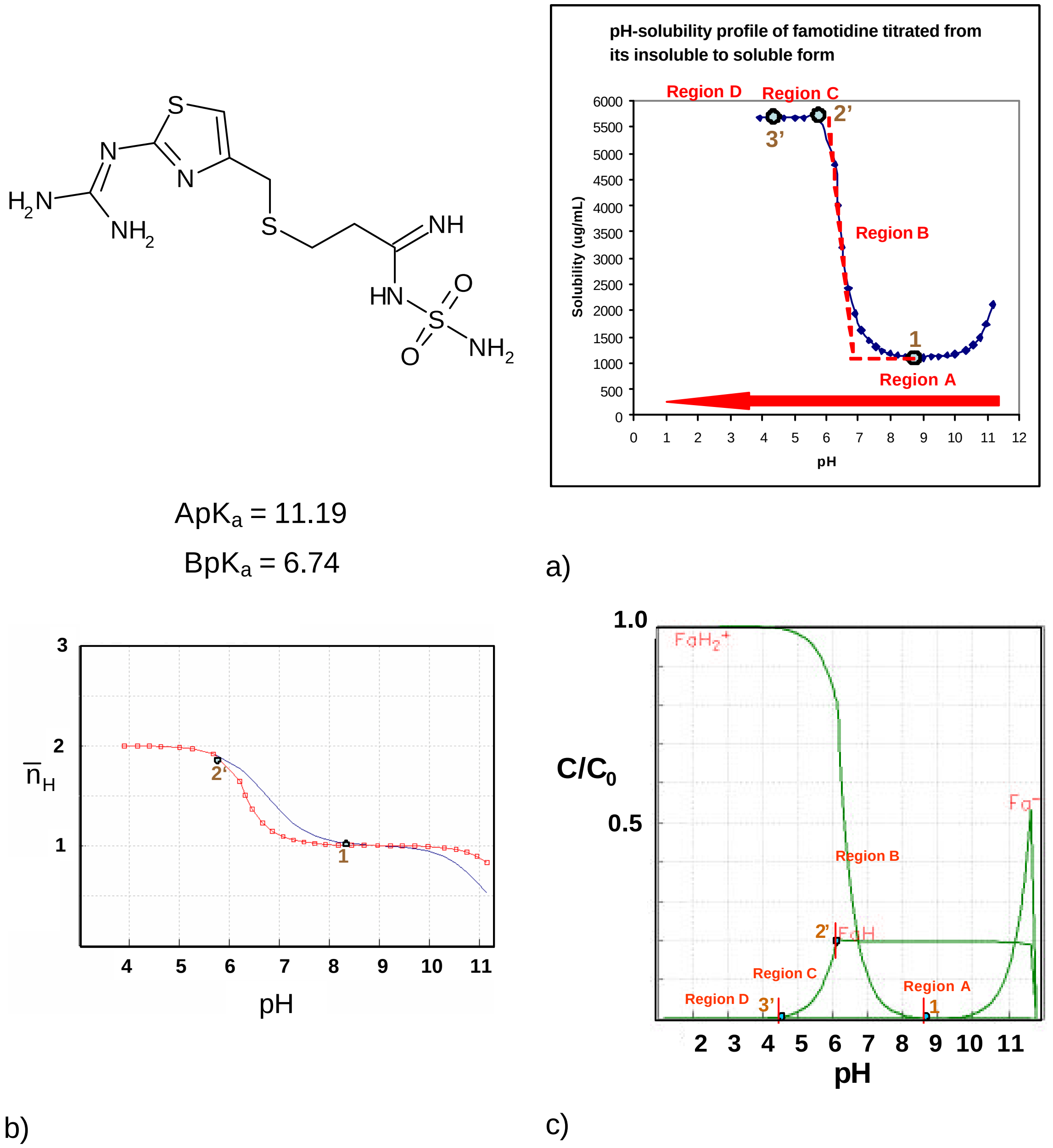
Bosentan, trazodone, glibenclamide, iodopanoic acid are commercial compounds used to test the working principle of the new evaluation method.

The pSol[22] equipment for the potentiometric solubility measurement was used here to evaluate the tendency for the formation of amorphous materials. The characteristics of pSol[22] have already been described in chapter 2.

pH-Solubility profile, Bjerrum plot and speciation profile are obtained by potentiometric solubility measurements.

- The pH-solubility profile describes the plot of pH against solubility.
- The Bjerrum plot depicts pH against $\bar{n}_H$. $\bar{n}_H$ is the average number of the bound hydrogen atoms per molecule of substance. Therefore, the Bjerrum plot reveals all $pK_a$s as pH values at half-integral $\bar{n}_H$ positions.
- The normalized speciation profile depicts pH against the normalized concentration of all compound species.

The function of these three plots and the relationship between them is demonstrated using famotidine as an example.



ApK$_a$ = 11.19

BpK$_a$ = 6.74

a)

b)

c)

Figure 13: Famotidine titrated from its insoluble to soluble form. a) pH-solubility profile; b) Bjerrum plot; c) Speciation profile. The blue colored curve describes the titration in the absence of precipitate FaH. The blue colored curve describes the titration in the presence of precipitate FaH. The red points represent the collected experimental data.

The pH-solubility profile of famotidine (Figure 13a) can be divided into four regions by three defined points, resp. point 1, 2’ and 3’, which can be retrieved via the speciation profile. (Figure 13c) In the speciation profile, famotidine exists in region A dominantly in its precipitated form, FaH$_{(s)}$, and the soluble form FaH; in region B, FaH$_{(s)}$, FaH and its ionic form $FaH_2^+$; in region C, FaH and $FaH_2^+$ coexist; in region D, $FaH_2^+$ is the major component.

In Figure 13b, the Bjerrum plot of famotidine is depicted. The blue colored curve is the reference curve and describes the titration in the absence of precipitate FaH. With the help of the reference curve, points 1 and 2' in the solubility profile can be defined in Bjerrum plot, as well. When famotidine is titrated from its insoluble to soluble form, data are collected and represented as red points in Figure 13b. In case of missing supersaturation, experimental data collection runs along the red curve, which stays for the titration in the presence of precipitate. The red curve meets the reference curve at two different points. The first one is the same as point 1 in the solubility profile. It indicates the status of the minimal concentration of the charged form in solution. Continuing the titration to lower pH, the total concentration in solution increases. More and more uncharged precipitate $FaH_{(s)}$ is transformed to the charged form, $FaH^+$, while the concentration of the uncharged form in the solution stays constant. Finally, the intersection of the titration and reference curve is reached. This intersection point is equal to point 2' in the solubility profile. At this point, the uncharged precipitate $FaH_{(s)}$ has reached its minimum concentration and the whole amount of the compound is dissolved in the solution. From point 2' to 3', the charged form $BH^+$ will only go into the solution at the expense of the dissolved uncharged form B, therefore, the solubility does not change with pH.

## 3.2.2 Methods

### 3.2.2.1 High Throughput solubility assay

High Throughput (HT) solubility assays use the solvent evaporation process and in principle are modified Shake-Flask assays. Saturated buffer solutions are prepared by adding buffer to the solid materials which are obtained through freeze-drying to eliminate the DMSO from the stock solution. (Figure 14)

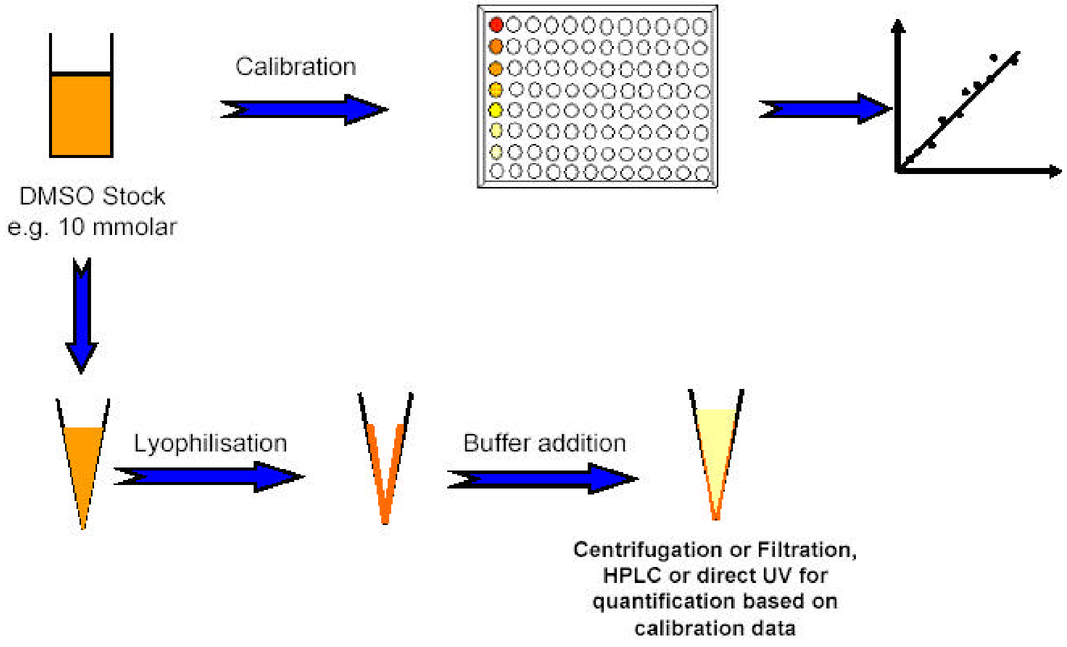
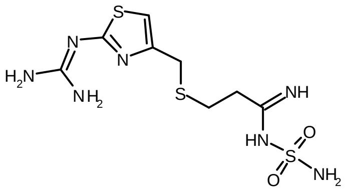


Figure 14: The principal steps of a High Throughput solubility measurement[26].

### 3.2.2.2 Assay description for the evaluation of the tendency to form amorphous material

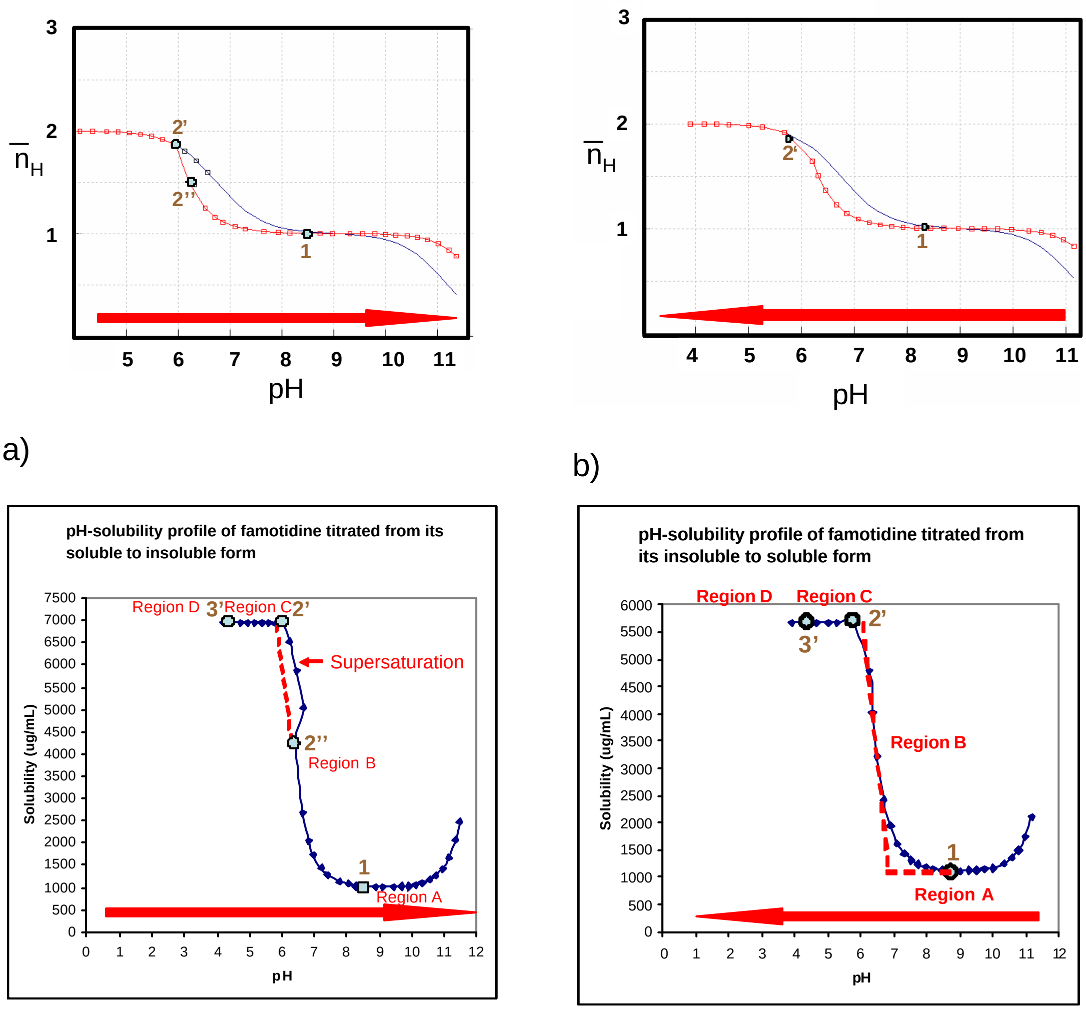
The tendency to form amorphous forms is compound-specific[21] and can be related to the occurrence of high degree of supersaturation, observed in crystallization experiments. The crystallization method described in chapter 2 is based on the titration of a compound from its soluble to its insoluble form. The reverse titration procedure described by Avdeef[22] is used in the determination of intrinsic solubility. Hence, a compound can be titrated in both directions, resp. from soluble to insoluble or reverse. The obtained curves should usually be identical, but can be different in cases when compounds have the tendency to form highly supersaturated solutions. Therefore, comparing the curves obtained by reverse titration experiments allows to detect supersaturation effects and can give some insight into tendencies of compounds to form amorphous materials.

Famotidine has two $pK_a$ values, about 6.74 and 11.19. Two titrations were performed for famotidine. One was from its soluble to insoluble form and the other was from its insoluble to soluble form as depicted in Figure 15 by focusing on the basic $pK_a$.



$ApK_a = 11.19$

$BpK_a = 6.74$



a)

b)

c)

d)

Figure 15: Bjerrum plot and pH-solubility profile of famotidine. The direction of titration is given by red arrow. The red curve in the Bjerrum plot is the calculated titration curve in the presence of precipitate, the blue curve is the calculated titration curve in the absence of precipitate and the unfilled circles are the experimental points registered during the titration. a) Bjerrum plot titrated from the soluble to insoluble form; b) Bjerrum plot titrated from the insoluble to soluble form; c) pH-solubility profile titrated from the soluble to insoluble form; d) pH-solubility profile titrated from the insoluble to soluble form.

When famotidine was titrated from its soluble to insoluble form, the curve determined by the potentiometric method was not identical with the calculated Bjerrum curve from pH 6 to 6.5. The experimental data were following preferably the blue curve to some extend and jumped back to the red curve, when the precipitation started. Therefore, in this pH range, more famotidine was dissolved than expected and this phenomenon indicated the occurrence of supersaturation[36]. Hence, in this case, precipitation began not at point 2', but at point 2''. Furthermore, at point 2'', where the precipitate appeared, the precipitation rate could be large which was one precondition for the generation of amorphous materials. Figure 15c shows the corresponding pH-solubility profile. The unusual curve form in region B was another indicator for the occurrence of supersaturation.

Figure 15b and d show the Bjerrum plot and pH-solubility profile, when famotidine was titrated from its insoluble to soluble form. In this titration direction, the experimental data moved along the red curve and supersaturation did not occur. This was the appropriate direction to measure the intrinsic solubility.

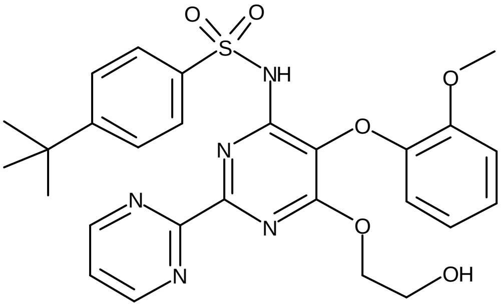
For the first time, this study demonstrated that potentiometric titration from reverse directions can be used to detect the occurrence of supersaturation and to evaluate the level of the supersaturation. Due to its easy and comfortable performance, this new method can be utilized in the early drug discovery phase as a quick recognition procedure to identify the tendency of compounds to form amorphous materials. The results of the new method may explain complicated, difficult-to-understand biological processes, e.g. high absorption caused by formation of supersaturation in the in vivo test. Furthermore, this finding can streamline formulation activities in the later drug development where information on supersaturation and relative probability for the formation of amorphous forms is mandatory.

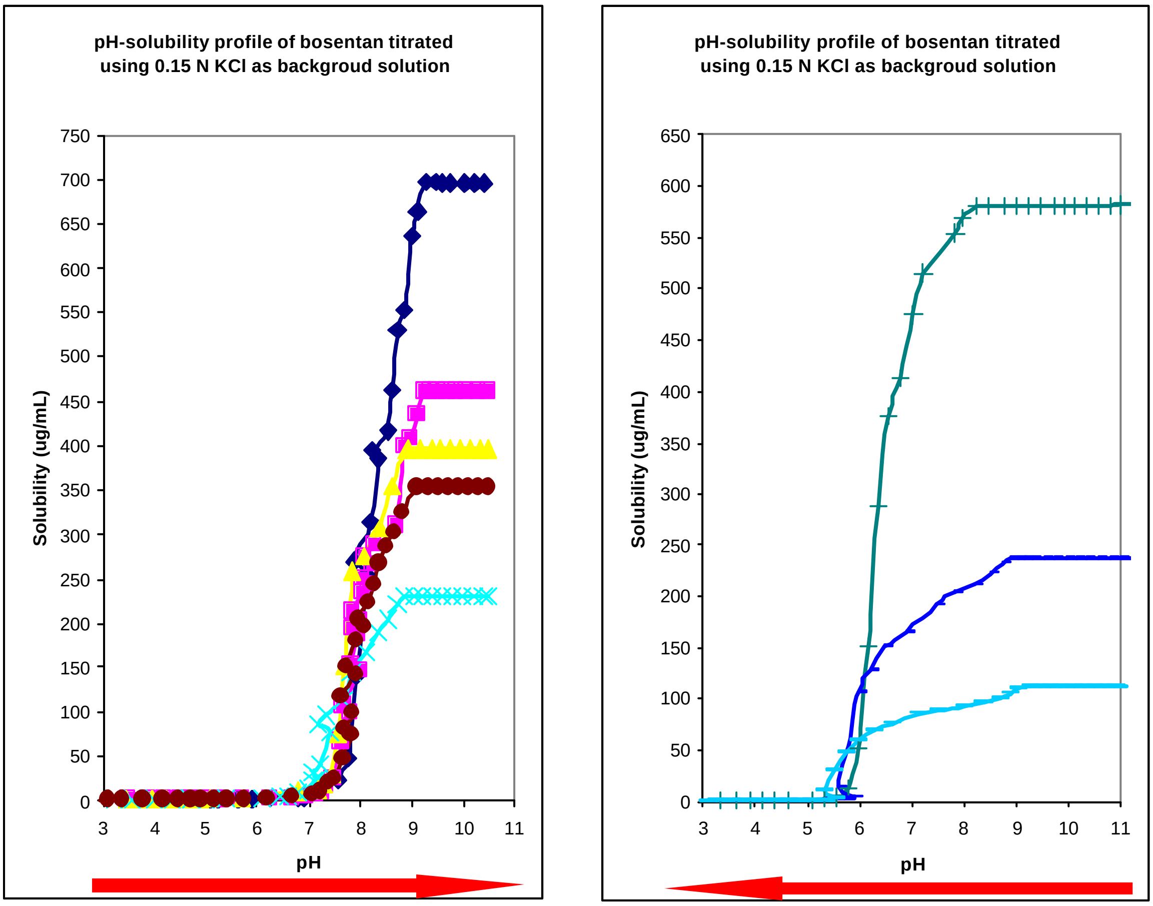
## 3.3 Results and discussion

Four drugs, bosentan, trazodone, glibenclamide and iodopanoic acid showed large differences in their equilibrium and HT solubility values. They were taken as examples to evaluate the working principle of the newly developed method. Bosentan has an acid $pK_a$ of 5.46, trazodone a basic $pK_a$ of 6.6, glibenclamide an acid $pK_a$ of 4.5 and iopanoic acid an acid $pK_a$ of 4.5.

### 3.3.1 Bosentan

Bosentan is a drug identified with large differences in solubility results obtained using equilibrium and HT-solubility assays. The titration behavior of bosentan was investigated by bidirectional titration experiments as described.





a)

b)

Figure 16: pH-solubility profile of bosentan. a) Profiles obtained by titration of bosentan from its insoluble to soluble form; b) profiles obtained by titration of bosentan from its soluble to insoluble form.

The profiles obtained from titrations should be independent on the titration directions, in case when the compound does not form supersaturated solution. However, for bosentan, this was not the case. In comparison with the profiles obtained by titration from the insoluble to soluble form, the profiles obtained by reverse direction were not reproducible and were partly dependent on degrees of supersaturation. Furthermore, at the same pH value, the solubility value of bosentan obtained by titration from its insoluble to soluble form was much lower than from the reverse direction, which was a strong indication for supersaturation.

The reverse titration results are helpful to explain the differences obtained by equilibrium and HT-solubility assays. (Table 5)

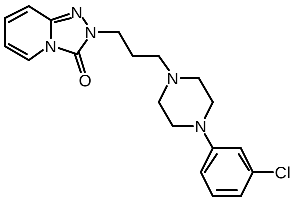
| Name | Method | $pK_a$ | S (µg/mL) | pH | $S_0$ (µg/mL) | Back ground solution | Titration direction |
|---|---|---|---|---|---|---|---|
| Bosentan | Equilibrium solubility | 5.46 | 14 | 6.6 | 0.946 | 50mM Phosphate pH=6.5 | |
| Bosentan | HT-solubility | 5.46 | 1260 | 6.5 | 105.3 | 50mM Phosphate pH=6.5 | |
| Bosentan | pSol | 5.46 | | | 0.942 | 0.15 N KCl | Insoluble to soluble |
| Bosentan | pSol | 5.46 | | | 48.5 – 116.6 | 0.15 N KCl | Soluble to insoluble |

Table 5: Equilibrium, High Throughput and potentiometric solubility measurements of bosentan.

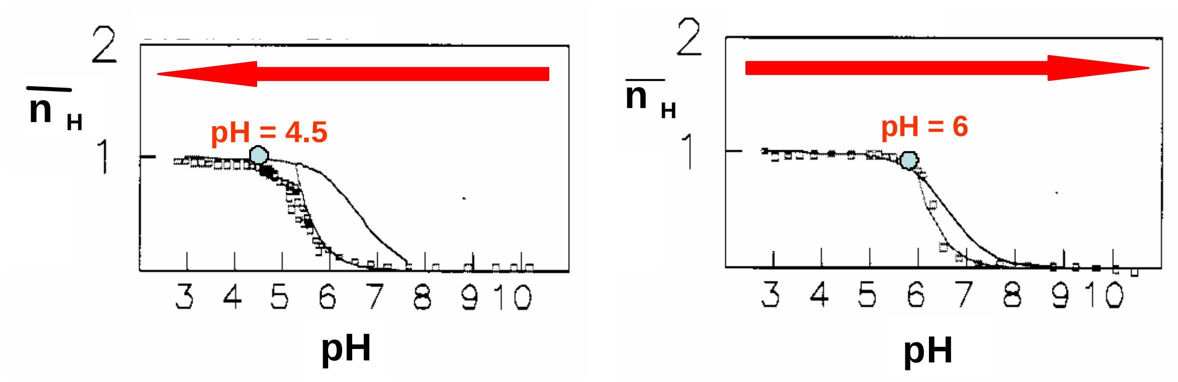
Table 5 shows that for bosentan, higher HT-solubility value in comparison to equilibrium solubility was obtained. In case of the HT-solubility assay, freeze-drying was used to eliminate DMSO from the stock solution. Thus, the amorphous form of bosentan could possibly be obtained because of its preference to form supersaturated solutions. This assumption was confirmed by the potentiometric results as well. Titrating bosentan from its soluble to insoluble form, higher solubility values were obtained than in reverse direction. Furthermore, the higher potentiometric result ($S_0$ = 116.6) agreed with the higher HT-solubility value ($S_{0\ (HT\text{-}solubility)}$ = 105.3) and the lower potentiometric result ($S_0$ = 0.942) with the lower equilibrium solubility value ($S_{0\ (equilibrium\ solubility)}$ = 0.946). Therefore, differences in equilibrium and HT-solubility can be explained by analysis of different solid forms. Reverse potentiometric titrations can be used to estimate the tendency of compounds to form supersaturated solutions, which is one precondition for the formation of amorphous materials.

## 3.3.2 Trazodone

Trazodone is another example showing a large deviation between its equilibrium and HT-solubility. Because the purchasable basic form of trazodone was only available as a methanol solution, the salt solid form, trazodone hydrochloride was used to study the behavior of trazodone in titration experiments.



BpK$_a$ = 6.6



a) b)

Figure 17: Bjerrum plot of trazodone. a) Titrated from insoluble to soluble form; b) titrated from soluble to insoluble form.

| Name | Method | $pK_a$ | S (µg/mL) | pH | $S_0$ (µg/mL) | Background solution | Titration direction |
|---|---|---|---|---|---|---|---|
| Trazodone | Equilibrium solubility | 6.6 | 68 | 13 | 68 | 500mM KOH | |
| Trazodone | HT-solubility | 6.6 | > 454 | 6.5 | > 201 | 50mM Phosphate pH=6.5 | |
| Trazodone | pSol | 6.6 | | | 105.1 | 0.15 N KCl | Insoluble to soluble |
| Trazodone | pSol | 6.6 | | | 453.7 | 0.15 N KCl | Soluble to insoluble |

Table 6: Equilibrium, High Throughput and potentiometric solubility measurement of trazodone.

Similar amounts of trazodone HCl were taken for the titration in opposite directions. From soluble to insoluble form, the precipitation occurred at pH = 6, much later than in the reverse direction (pH = 4.5). Therefore, high supersaturation was assumed to occur in this direction, which was confirmed by the agreement between the solubility result obtained in this direction ($S_0$ = 453.7 ug/mL) and the high HT-solubility value ($S_{0\ (HT\text{-}solubility)}$ > 201 ug/mL).

## 3.3.3 Glibenclamide and Iopanoic acid

Glibenclamide and iopanoic acid are two compounds described by Hancock[37] to show high solubility differences between amorphous and crystalline materials. (Table 7)

| Compound | Forms | Solubility ratio | Comments |
|---|---|---|---|
| Glibenclamide | Amorphous/crystal | 14 | 23 °C, buffer (aq.) |
| Iopanoic acid | Amorphous/I-crystal | 3.7 | 37°C, phosphate buffer (aq.) |

Table 7: Experimental solubility ratios for glibenclamide and iopanoic acid[37].

Equilibrium, High Throughput and potentiometric solubility results of glibenclamide and iopanoic acid are summarized in Table 8.

| Name | $ApK_a$ | Equilibrium solubility | | | HT-solubility | | | Potentiometric method | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | From insoluble to soluble | From soluble to insoluble |
| | | S (µg/mL) | pH | $S_0$ (µg/mL) | S (µg/mL) | pH | $S_0$ (µg/mL) | $S_0$ (µg/mL) | $S_0$ (µg/mL) |
| Glibenclamide | 4.5 | 0.33 | 6.5 | 0.003 | 52 | 6.5 | 0.515 | 0.009 | 0.046 |
| Iopanoic acid | 4.5 | 38.02 | 6.5 | 0.377 | 213 | 6.5 | 2.11 | 0.405 | 6.2 |

Table 8: Equilibrium, High Throughput and potentiometric solubility of glibenclamide and iopanoic acid.

Table 8 shows the amorphous formation tendencies of glibenclamide and iopanoic acid were confirmed, firstly by the different equilibrium and HT-solubility results; secondly by the different potentiometric results obtained using reverse titration experiments.

## 3.4 Conclusion

In conclusion, the crystallization procedure described in chapter 2 is based on the titration of a compound from its soluble to insoluble form. The often observed occurrence of high supersaturation in this direction can be assumed to be one precondition for the formation of amorphous materials. Thus, comparing the titration behavior of compounds from opposite directions is a new and easy procedure to evaluate the tendency for the formation of amorphous solids and one possibility to explain differences in the results of solubility experiments. Therefore, this work discovered a new procedure with the following advantages:

- Quickly identifies the tendency of compounds to form amorphous materials.
- Helps to improve the design and production of pure and stable pharmaceuticals.
- Streamlines formulation activities in the later drug development where information on supersaturation and relative probability for the formation of amorphous forms is mandatory.
- May explain complicated, difficult-to-understand biological processes, e.g. high absorption caused by formation of supersaturation in the in vivo test.

# Prediction Part

# 4 AQUEOUS SOLUBILITY PREDICTION OF DRUG-LIKE COMPOUNDS

## 4.1 Introduction

The majority of prediction tools for solubility are generic tools. There is only a small number of tools dealing with congeneric series. Usually, all tools either commercially available or published by academia are restricted to deal with non drug-like molecules due to the limited number of published solubility data of drug-like compounds. The following short summary will give an overview on the data sets, descriptors and methods which have been used in the development of prediction tools for solubility in the past. The predictive power of commercially available tools was evaluated using a newly collected data set of drug-like compounds. Finally, a comparison of published and newly collected data sets was performed and the difference in the related data sets is described.

### 4.1.1 Solubility prediction tools

#### 4.1.1.1 Data sets used in solubility prediction

Usually, data sets from PHYSPROP database[38-43], Huuskonen[44], AQUASOL[5,39,40,45] are used in the parameterization of solubility prediction tools. A few groups[38-42,46,47] collected data sets from different literature sources[9,48], with focus on solubility measured under identical experimental conditions. High quality measurement data have been used in the prediction tools developed by McFarland[10], Klamt[49] and Bergström[50].

#### 4.1.1.2 Methods and descriptors

The principal computational approaches for solubility prediction can be grouped into two classes:

- Multiple linear regression (MLR) based
- Neural networks (NN) based

#### 4.1.1.2.1 Multiple linear regression

In multiple linear regression based approaches, the correlation between solubility (S) and its relevant descriptors is computed, according to the Eq. 9.

$$\log S = \sum_i a_i c_i + a_0 \qquad \text{Eq. 9}$$

where $c_i$ are values of different molecular descriptors i for the given molecule and $a_i$ are the corresponding coefficients determined by regression analysis, in order to maximize the correlation coefficient $r^2$ between the measured and computed solubility results.

When structural fragments are used as descriptors, the multiple linear regression method can be defined as a Group Contribution (GC) approach and $a_i$ are increments assigned to the number of occurrences $c_i$ of a structural fragment i in the molecule of interest.

When structural properties are used as descriptors, the multiple linear regression method is named Property Contribution approach and $c_i$ are values of different molecular properties for the given structure. The properties used in multiple linear regressions can be divided into two classes: experimental and calculated. The experimental properties can be, for example, melting points, boiling points, and lipophilicity. Calculated properties usually used are molecular weight, solvent-accessible surface area (SASA), counts of potential donor and acceptor hydrogen bonds (HBDN, HBAC), counts of specific functional groups and rotatable bonds, electrostatic potential data from quantum mechanical calculations, and a wide-range of topological and electronic indices such as those developed by Hall and Kier[51,52].

Standard statistical packages are usually applied for the descriptor selection in multiple linear regression based approaches[5-7,53-55]. Beside those, some novel methods for the descriptor selection have been described recently, e.g. Jorgensen[7] used Monte Carlo Simulation to select descriptors for the solute and water interaction, Wegner[56] and Sahura[39] used entropy-based descriptor selection.

#### 4.1.1.2.2 Neural networks

In comparison to multiple linear regression, the principal advantage of neural networks (NN) is related to the introduction of non-linear terms into the solubility equations. Furthermore, neural networks can consider descriptors in specific range of the measurement space. The disadvantage is that the internal processing of data in the NN approach is hidden. Usually, NN systems are treated as black boxes and often difficult to provide further insights in the nature of the major properties or features governing solubility. Therefore, the application of neural networks can be considered as promising in the treatment of large data sets with high content of non-linearity.

### 4.1.1.3 Available solubility prediction tools

#### 4.1.1.3.1 Available tools based on multiple linear regression

Numerous approaches based on multiple linear regression have been published for the prediction of aqueous solubility. Most studies include a large collection of various, relatively complex descriptors[42,54,57-60]. The probably most successful studies based on multiple linear regression approaches are from Abraham and Le[5], Meylan and Howard[6], Jorgensen and Duffy[7].

Abraham and Le[5] used experimentally determined descriptors for developing the logS prediction and ended up with a six-descriptor model with $r^2 = 0.92$ and rms = 0.56 for 594 molecules. (Eq. 10)

$$\begin{aligned}\log S &= 0.510 - 1.02 * R_2 + 0.813 * p_2^H + 2.124 * \sum a_2^H \\ &+ 4.187 * \sum b_2^H - 3.337 * \sum a_2^H * \sum b_2^H - 3.986 * V_x\end{aligned} \quad \text{Eq. 10}$$

where $R_2$ is the molar refractivity, $p_2^H$ is the dipolarity, $\sum a_2^H$ is the hydrogen-bond acidity, $\sum b_2^H$ is hydrogen-bond basicity and $V_x$ is volume.

Meylan and Howard[6] used experimental $\log P_{o/w}$ and molecular weight ($M_W$) as descriptors along with 15 correction factors ($f_i$) to predict solubility. An $r^2 = 0.84$ and rms = 0.90 was obtained for a data set of 3000 compounds. (Eq. 11)

$$\log S = 0.796 - 0.854 \log P_{o/w} - 0.00728 M_w + \sum_i f_i \qquad \text{Eq. 11}$$

$f_i$ describes various sub-rules accounting for the presence of specific functional groups. 12 compound classes are identified: aliphatic alcohol, aliphatic acid, aliphatic amine, aromatic acid, phenol, alkyl pyridine, azo, nitrile, hydrocarbon, nitro, $SO_2$, fluoroalkane, polycyclic aromatic hydrocarbons (PAH), multi-amino acid. Each class of these has a corresponding $f_i$ value. $\sum f_i$ is the sum of all correction factors applicable to a given compound. Each factor applies to a compound, if the related substructural fragment is present, but each factor is counted only once no matter how many times the functional group appears in a molecule.

Jorgensen and Duffy[7] selected their descriptors via a Monte Carlo (MC) simulation for different solutes in water. Five terms were used in the final regression equation and yielded $r^2 = 0.88$, $q^2 = 0.87$, and rms = 0.72 for 230 compounds.

$$\log S = 0.32\,\mathrm{ESXL} + 0.65\,\mathrm{HBAC} + 2.19\,\#\mathrm{amine} - 1.76\,\#\mathrm{nitro} - 162\,(\mathrm{HBAC} * \mathrm{HBDN})^{1/2} / \mathrm{SASA} + 1.18 \qquad \text{Eq. 12}$$

where ESXL is solute-water Lennard-Jones interaction energy. It is highly correlated with molecular size, which can be represented alternatively by SASA or volume. HBAC is the number of hydrogen acceptors, HBDN is the number of hydrogen donors, #amine is the number of non-conjugated amine groups and #nitro is the total number of nitro groups.

In 2002, Jorgensen and Duffy[61] developed three diverse QSPR equations for alkane, PAHs and remaining molecules.

| | | |
|---|---|---|
| For alkanes | $\log S = 1.302 - 0.0104\,\mathrm{VOL}$ | Eq. 13 |
| For PAH class | $\log S = 4.182 - 0.0155\,\mathrm{VOL} + 0.670\,\#\mathrm{rotor}$ | Eq. 14 |

For remaining molecules

$$\log S = 3.886 - 0.0194\ \text{SASA} + 0.514\ \text{HBAC} + 0.578\ \text{HBDN} + 1.343\ \#\text{amine} + 1.224\ \#\text{amide} - 116\ (\text{HBAC} * \text{HBDN})^{1/2} / \text{SASA} + 0.182\ \#\text{rotor} - 0.00405\ \text{WPSA} \qquad \text{Eq. 15}$$

where #rotor is the number of rotable bonds. #amide is the number of amides. The WPSA (weakly polar components of SASA) term is the surface area for all halogens, sulfur, and phosphorous atoms.

Yalkowsky and Valvani[53] used melting points to consider the impact of crystal state on solubility. A regression was achieved for 155 compounds with $r^2 = 0.979$ and SD = 0.308. (Eq. 16)

$$\log S_w = -1.05\ \log P_{oct} - 0.012(mp - 25) + 0.87 \qquad \text{Eq. 16}$$

where logP is octanol-water partition coefficient. It approximates the activity coefficient of the un-ionized solute in water in equilibrium with the un-ionized molecular species in octanol. Mp is melting point in °C, an approximation for the relative energy it takes to break the crystal lattice of the solute.

According to Eq. 16, the melting point is a valuable descriptor for describing the influence of solid state on solubility. However, the disadvantage of Yalkowsky's method relates to the fact that there are currently no reliable models to predict melting points. Usually, experimental values have to be used, which are not suitable in the early drug discovery phase, because ranking schemes are necessary before synthesis.

Beside the works of Meylan and Howard[6], Jorgensen and Duffy[61], several structural series orientated studies were preformed. In order to prove the molecular similarities, Chen[54] divided a data set of 321 structurally diverse drugs or related compounds into three groups, according to the Euclidean distance calculated using 8 molecular descriptors of the compounds. His QSAR model could predict the properties of unknown compounds that were structurally similar to those used to build the model. Delgado[62] made a solubility study for chlorinated hydrocarbons, McElroy[55] focused on heteroatom-containing organic compounds, Nikolic[63] on aliphatic alcohols and Yin[64] on sulfur-containing aromatic esters.

#### 4.1.1.3.2 Available tools based on neural networks

Tetko and Tanchuk[9,65] used multiple linear regression for identifying subsets of significant descriptors in the application of NN. They started with three different types of 55 topological indices introduced by Kier and Hall[51,52]. These indices were analyzed via multiple linear regression. The resulting final equation contained 33 significant parameters. The selected parameters were 24 E-state and six other topological indices including indicator variables for aliphatic hydrocarbons and aromaticity. Artificial Neural Networks were then applied to analyze the set of 33 selected descriptors and a model was provided with $r^2$ = 0.91 and RMS = 0.62 to estimate the aqueous solubility for a diverse set of 1291 organic compounds with 33-4-1 neurons.

#### 4.1.1.3.3 Other available tools

Klamt[10] combined the COSMO-RS method, based on quantum chemical calculations, with a QSPR approach in order to predict the aqueous solubility of a wide range of typical neutral drugs and pesticides. The COSMO-RS, originally developed for the prediction of liquid-liquid and liquid-vapor equilibrium constants, was extended to solid compounds by the addition of an expression for the Gibbs free energy of fusion $\Delta G_{\mathrm{fus}}^{\mathrm{X}}$, which was related to the free energetic difference between the compound in its solid and liquid state. Klamt[10] first identified a small set of descriptors of potential significance for $\Delta G_{\mathrm{fus}}^{\mathrm{X}}$. The selected descriptors were the molecular size, rigidity, polarity, and number of hydrogen bonds. He tried to describe $\Delta G_{\mathrm{fus}}^{\mathrm{X}}$ via a QSPR approach. It finally turned out that the descriptor combination of cavity volume $V^X$, the number of ring atoms $N_{ringatom}^{X}$, and the chemical potential of a compound X in water ($m_W^X$) was the best suited for the description of $\Delta G_{\mathrm{fus}}^{\mathrm{X}}$. On a data set of 150 neutral drug-like compounds, the COSMO-RS model achieved a rms deviation of 0.66 log-units. One possible advantage of this prediction method is that COSMO-RS is able to predict solubility in almost arbitrary solvents and solvent mixtures due to the capability of COSMO-RS to estimate the chemical potential of a compound in arbitrary liquids.

#### 4.1.1.4 Performance of commercially available tools on drug-like compounds

In a recent evaluation[66], commercially available solubility prediction tools were tested on a set of 384 neutral drug-like compounds.

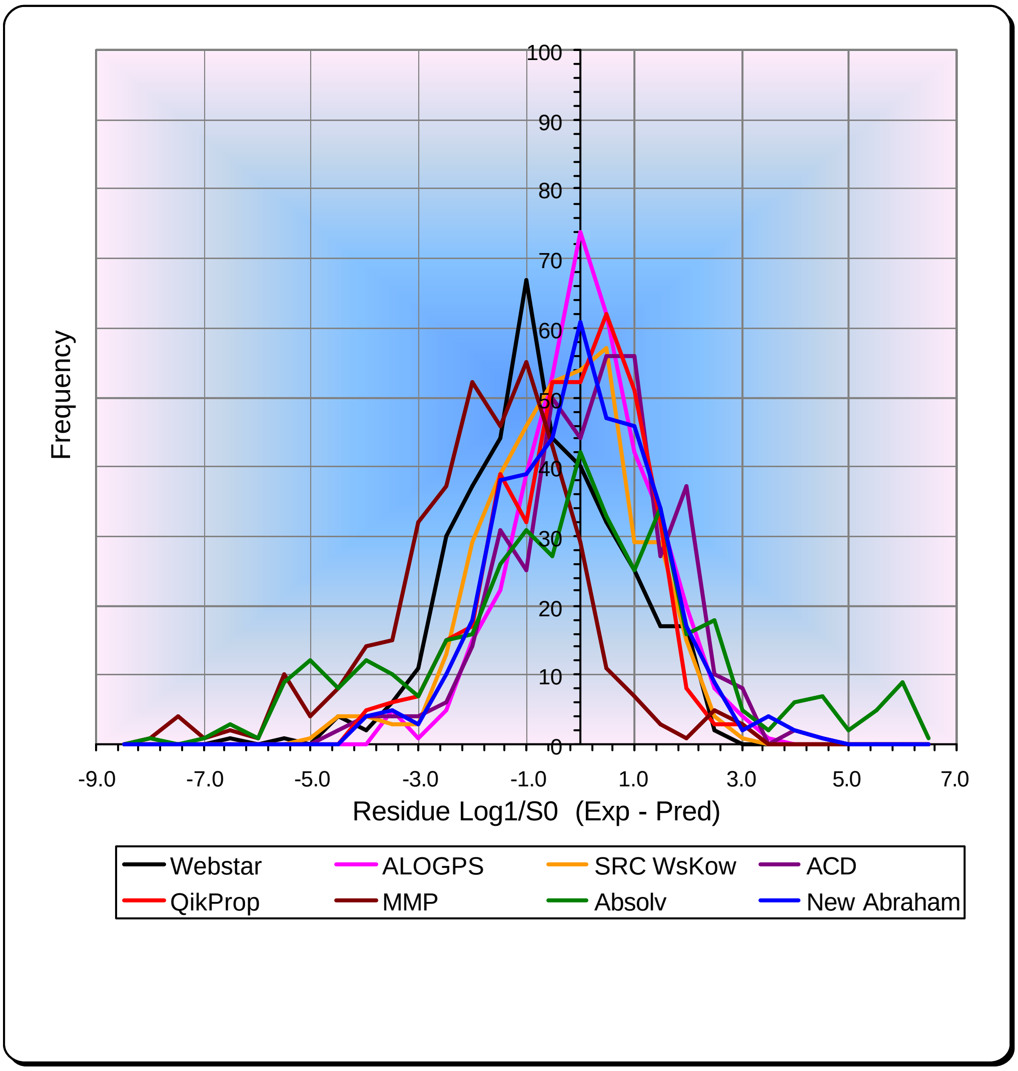
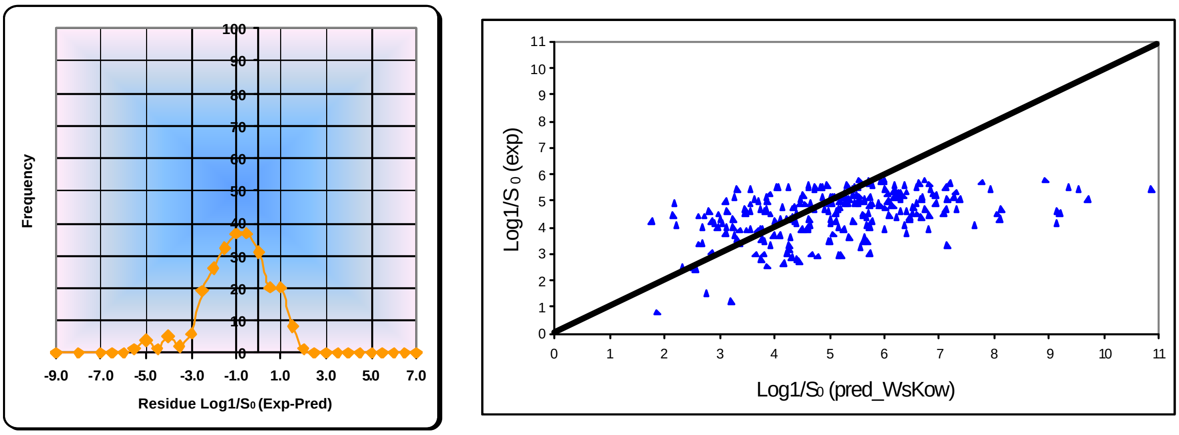


Figure 18: Experimental minus predicted Log1/$S_0$ versus frequency of 384 neutral compounds in each residual range. $S_0$ is the molarity of the unionized molecular species. (graph from Le[66])

According to the results shown in Figure 18, AlogPS was the best available "off-the-shelf package" and predicted 49.2% of the compounds within an error of 0.5 log units of the experimental intrinsic solubility. No single residual was above 3.5 log units for the 384 compounds in the data set.

The program SRC WsKow[6] was the second best solubility prediction tool after AlogPS. (Figure 18) Because AlogPS was not available, SRC WsKow[6] was re-evaluated with 253 more precisely characterized compounds taken from the data set selected by Le[66]. (Figure 19)



a)

b)

Figure 19: The solubility of 253 neutral drugs predicted with WsKow[6]. a) Experimental minus predicted Log1/$S_0$ versus frequency of compounds in each residual range; b) experimental versus predicted Log1/$S_0$ . $S_0$ is the molarity of the unionized molecular species.

Although most residues of the 253 neutral drugs lay within 2 log units (Figure 19a), there was no correlation between experimental and predicted solubility. (Figure 19b) This finding could be confirmed by a larger data set of 2473 drug-like compounds, which was used for the development of an improved solubility prediction tool as described in the chapter 4.2.1.1.3. (Figure 20)

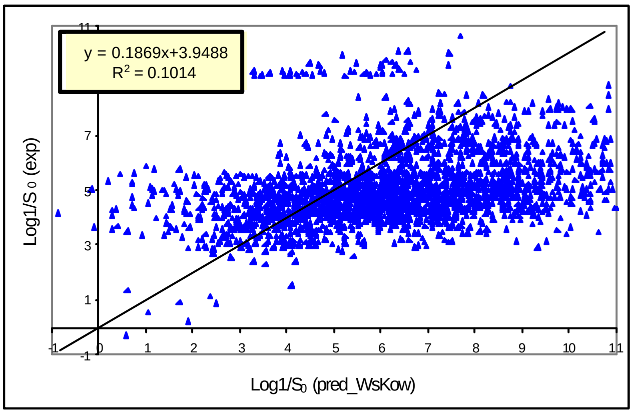


Figure 20: Solubility prediction for 2473 drugs-like compounds using WsKow[6]. $S_0$ is the molarity of the unionized molecular species.

### 4.1.1.5 Fundamental differences between aqueous solubilities of compounds from the AQUASOL database and drug-like compounds

In order to understand the reason for the poor performance in the solubility prediction of drug-like compounds by commercial tools, solubility values of 1770 organic compounds were extracted from the AQUASOL database[67]. Calculated properties of these 1770 organic compounds from AQUASOL database[67] and 2473 drug-like compounds were examined.

Evaluated descriptors were calculated with programs Msrfvl[68] and CallistoGen[69]. Principal component analysis[70] (PCA) was applied to reveal groupings in the observations. PCA summarized the information contained in the original variables by calculation of four new latent variables. The first three components described 75.9% of the X-space.

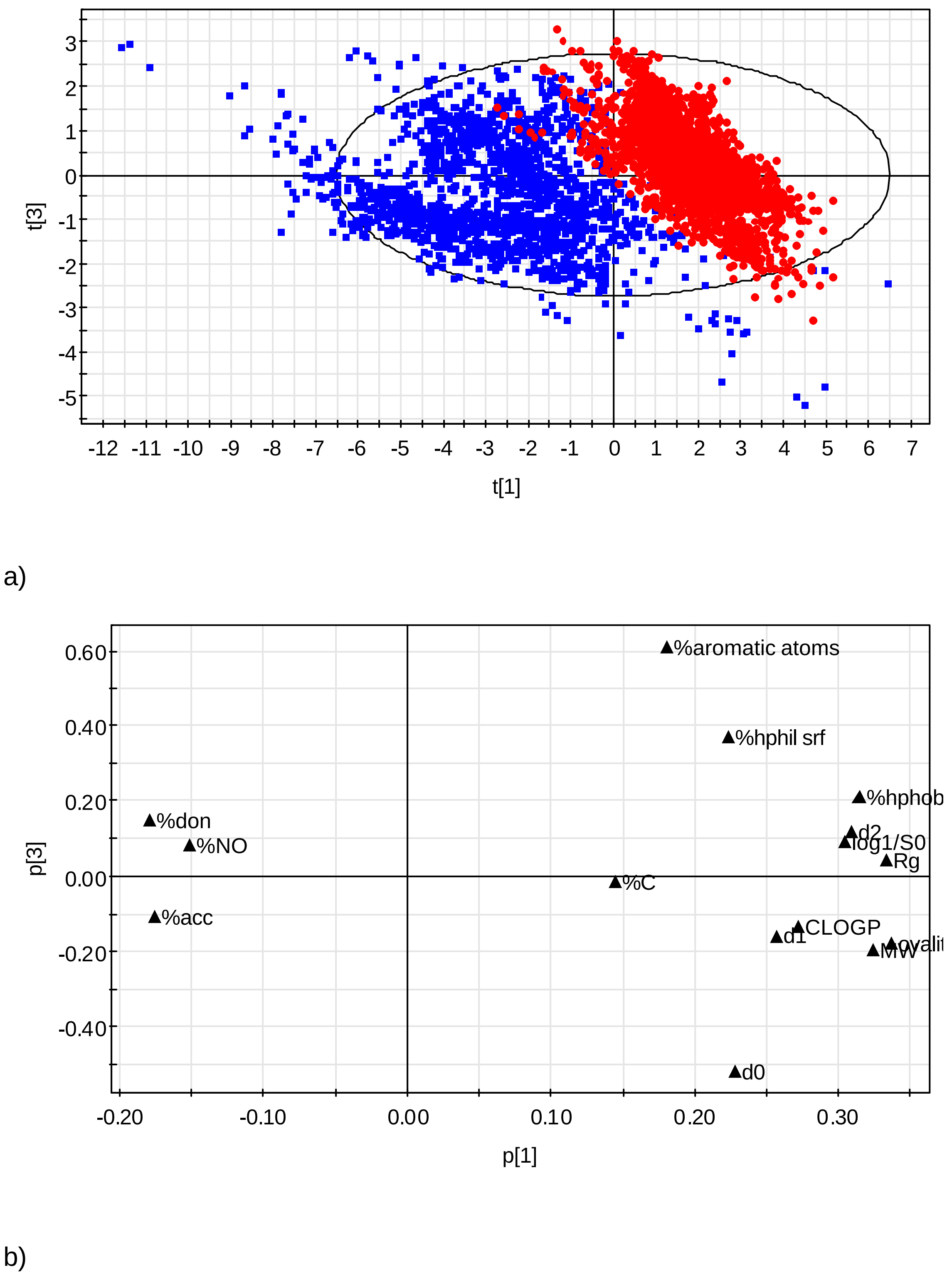


Figure 21: PCA analysis for 1770 organic compounds from the AQUASOL database[67] (blue) and 2473 internal drug-like molecules (red). a) PCA score plot[70]; b) PCA loading plot[70].

A clear separation between 1770 organic compounds and 2473 drugs was observed as depicted in Figure 21a. The descriptors responsible for this separation were molecular weight, %aromatic atoms and solubility (Figure 21b). Histograms were used in Table 9 to compare the important properties of 1770 organic and 2473 drug-like compounds directly.

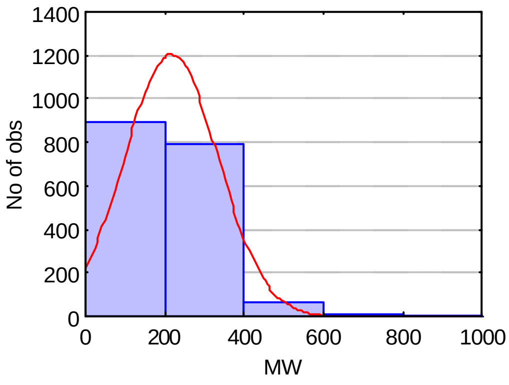
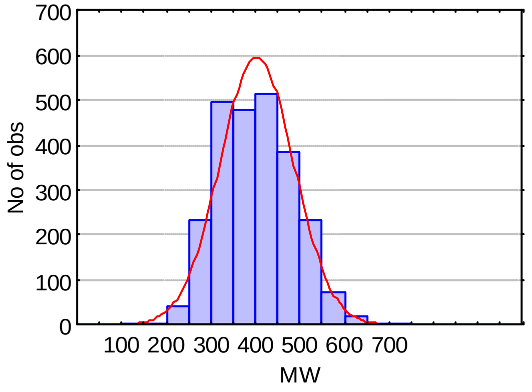
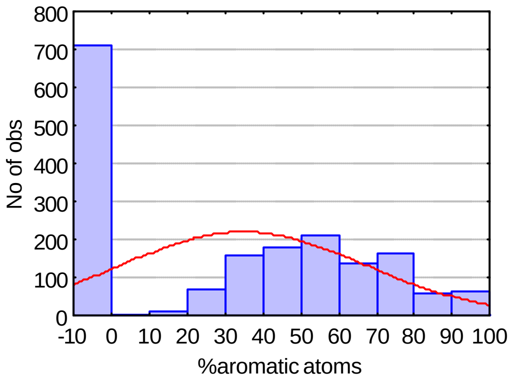
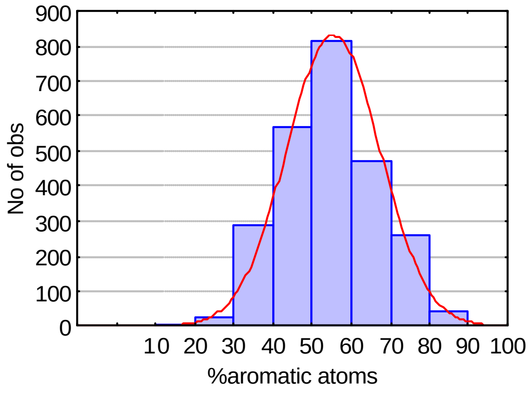
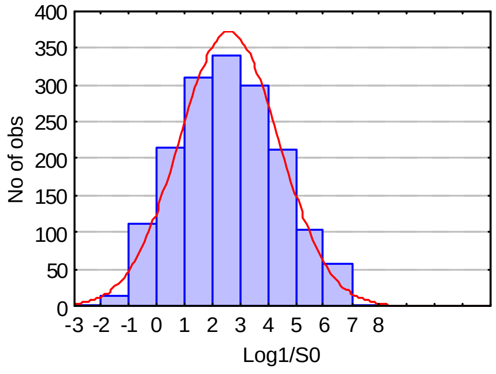
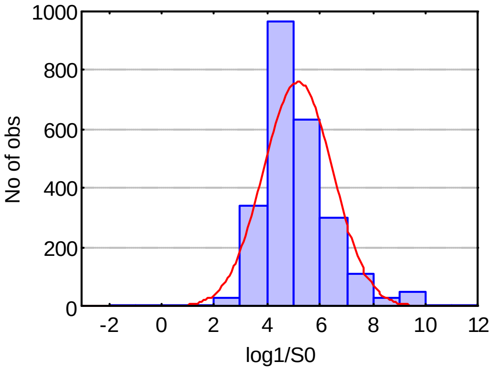
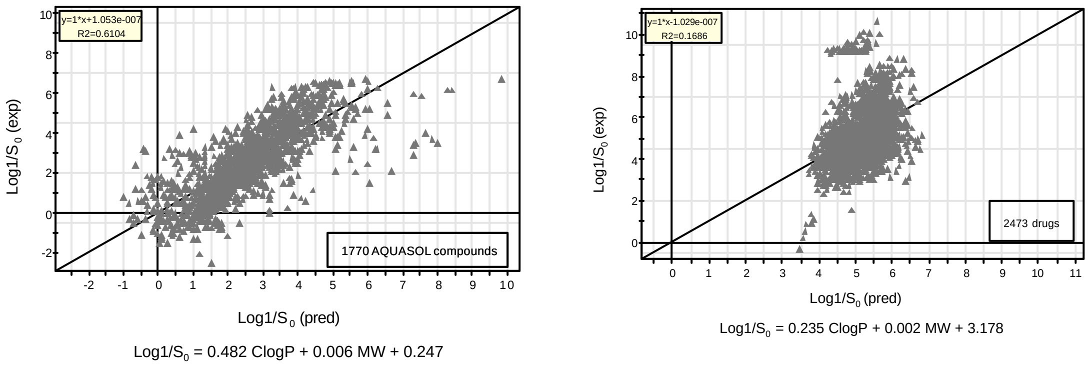
| Compounds | 1770 organic compounds | 2473 drugs |
|---|---|---|
| MW |  |  |
| %aromatic atoms |  |  |
| Log1/$S_0$ |  |  |

Table 9: Comparison of molecular properties of 1770 organic and 2473 drug-like compounds. $S_0$ is the molarity of the unionized molecular species.

Table 9 shows that in comparison to the organic compounds from the AQUASOL database[67], drug-like compounds have higher molecular weight, and usually include a larger fraction of aromatic atoms and occupy poor solubility. Additionally, 1770 compounds in the AQUASOL database[67] were inspected, according to their drug-likeness. 206 compounds were found as drug-like, resp. 11.6%. Therefore, the compounds in AQUASOL database[67] do not represent the properties of drug-like molecules. Due to this drawback, the commercial tools based on data extracted from the AQUASOL database[67] can not predict the solubility of drug-like compounds well.

The relationship between solubility, lipophilicity and molecular weight for 1770 organic compounds in AQUASOL database[67] and 2473 drug-like compounds are shown in Figure 22.



a) b)

Figure 22: Solubility prediction using lipophilicity and molecular weight as descriptors. a) For 1770 organic compounds in the AQUASOL[67] database; b) For 2473 drug-like compounds.

Figure 22a shows that the solubility of 1770 organic compounds in the AQUASOL database[67] can be predicted using lipophilicity and molecular weight as descriptors. However, the same does not work in the solubility prediction of drug-like molecules. Solubility prediction of drug-like molecules seems to be more complicated in comparison to simple organic molecules. Thus, commercially available tools can not be expected to work well for the solubility prediction of drug-like compounds, because they are calibrated with data of simple organic compounds.

### 4.1.2 Objectives

Solvation process and solid state related factors are governing solubility. In a large number of publications[5,7,53,61], lipophilicity is found to describe the liquid-liquid interaction important for the solvation process. Melting point, hydrogen bond donor and acceptor counts are usually used for the interpretation of the cohesive energy in the crystal packing. However, the relationship between crystal packing, melting point and intermolecular hydrogen bonding has not been completely understood. Furthermore, few publications can be found dealing with the extent of the influence of solid state on solubility. Therefore, the first goal of the prediction part was to collect compounds with measured 3D structures and melting points in order to explain the differences in solubility caused by diverse crystal packings. The second goal was related to polymorphism for evaluating the impact of solid state on solubility.

Several tools[6,54,55,61-64,71-73] are available for the solubility prediction of compounds with certain structural or property similarities. However, most of them fail to predict the aqueous solubility of drug-like compounds in congeneric series, because the descriptors used in such tools are calibrated with small subsets of organic compounds. Therefore, the third goal of this part focused on the identification of suitable descriptors for predicting solubility of an extended large data set comprising of congeneric series of drug-like compounds.

## 4.2 Materials and methods

### 4.2.1 Materials

#### 4.2.1.1 Data sets

Three data sets were collected. The first data set contained 74 compounds with known 3D crystal structures. The second data set contained 51 compounds with melting points, which were well characterized during the late development phase. The third data set contained 2473 compounds in 81 congeneric series. Among these 2473 compounds, 983 were uncharged, 166 had measured $pK_a$ values and for 1324 compounds, their $pK_a$ values were assigned according to structural similarity comparisons. Additionally, the 2473 compounds were classified using the clustering package of Daylight[74] and singletons were eliminated.

#### 4.2.1.1.1 Compounds with known crystal information: First data set

74 drug-like compounds were found to have their 3D crystal structures registered in the internal and external Cambridge Structure Database (CSD)[75]. Among them, 34 compounds had similar structures and belonged to four different series, resp. deoxyuridine-, diazepam-, sulfonamide- and sulfanilyurea-derivates. A large number of these 74 compounds showed polymorphism or pseudo polymorphism. Although polymorphic forms usually do not differ in their solubility to a large extent[27], solubility prediction can be complicated due to different conformations of diverse polymorphic forms, when the value of solubility is dependent on 3D structures.

Following rules for the selection of crystal structures were applied. If compounds occurred in both polymorphic and pseudo polymorphic forms, the polymorphic form was preferably taken. Usually polymorphs are energetically more stable, in comparison to the pseudo polymorphs. In case compounds occurred in several polymorphic or pseudo polymorphic forms, the conformations were compared with each other, using the superposition function of MOLOC[68]. In those cases, different possibilities were observed and had to be considered as described in the following:

Case 1. Different polymorphic forms with similar conformations but different packing schemes. These were diazepam, progesterone, sulfamethoxazole, hydrocortisone, trifluorothymine, deoxyriboside, sulfameter.

Case 2. Different polymorphic forms with different conformations and different packing schemes. These were bosentan, restosterone, carbamazepine, furosemide, diclofenac, sulfamerazine, 5-fluoro-2-deoxyuridine.

Case 3. An asymmetric crystal unit contains two molecules with identical constitution but different conformations. These were testosterone, furosemide, sulfamethoxypyridazine, sulfamerazine, 5-fluoro-2-deoxyuridine, sulfameter, medazepam, 2-methyl-4-methoxy-6-sulfanilamidopyrimidine, 3-azido-3-deoxythymidine, prazepam, sulfabenzamide.

According to the obtained analytical results for the different conformations, 59 compounds with one conformation were used for the model development and 15 compounds with more than one conformation to test the model.

#### 4.2.1.1.2 Compounds with known melting points: Second data set

51 compounds with predefined solubility and melting points during the late development phase were found. Out of this data set, 11 compounds had measured equilibrium solubility values. In order to evaluate the quality of the solubility data determined during the late development phase, the equilibrium solubility of these 11 compounds were listed together with the predetermined solubility in Table 10.

| Predetermined solubility | | | | | | | Equilibrium solubility | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ID | Ap$K_a$1 | Bp$K_a$1 | S (µg/mL) | log1/$S_0$ | Buffer | T (°C) | S (µg/mL) | log1/$S_0$ | Buffer | pH |
| compound 2 | 5.7 | 8 | 7700 | 1.69 | Buffer pH=7.5 | 37 | 6330 | 1.75 | Phosphate 0.05M pH 6.5 | 6.8 |
| compound 3 | 0 | 0 | <10 | <4.70 | 0.1N HCl | 25 | 22 | 4.35 | Phosphate 0.05M pH 6.5 | 6.5 |
| compound 4 | 0 | 8.9 | >2500 | >3.32 | Buffer pH=7.5 | 37 | 7740 | 3.71 | Phosphate 0.05M pH 6.5 | 6.6 |
| compound 5 | 0 | 6.13 | 320 | 3.38 | Buffer pH=6.8 | 37 | 602 | 3.15 | Phosphate 0.05M pH 6.5 | 6.6 |
| bosentan | 5.46 | 0 | 430 | 5.15 | Buffer pH=7.5 | 37 | 14 | 5.77 | Phosphate 0.05M pH 6.5 | 6.6 |
| compound 6 | 0 | 10.29 | 110 | 8.63 | Buffer pH=5 | 25 | 17 | 8.33 | Phosphate 0.05M pH 6.5 | 6.3 |
| compound 7 | 0 | 4.07 | 13 | 7.71 | Buffer pH=1 | 25 | 7 | 6.01 | Phosphate 0.05M pH 6.5 | 3 |
| compound 8 | 7.62 | 4.01 | 0.8 | 5.58 | Buffer pH=7 | 25 | 1 | 5.89 | Phosphate 0.05M pH 6.5 | 6.6 |
| compound 9 | 0 | 0 | <0.1 | >6.8 | Buffer pH=7 | 25 | 1 | 5.80 | Phosphate 0.05M pH 6.5 | 6.6 |
| compound 10 | 6.66 | 0 | 0.02 | 8.19 | Buffer pH=7.5 | 25 | 1 | 6.83 | Phosphate 0.05M pH 6.5 | 7.8 |
| compound 11 | 8.07 | 0 | <0.02 | >7.42 | Buffer pH=7 | 25 | 1 | 5.68 | Phosphate 0.05M pH 6.5 | 6.5 |

Table 10: Overview on solubilities of 11 selected compounds measured at certain pH value and temperature.

Table 10 shows that the predetermined solubilities agree with data determined via the equilibrium solubility method, except for compound 10. Thus, the predetermined solubility data of all these 51 compounds were considered as well characterized and used in the further development of a new solubility model.

#### 4.2.1.1.3 Compounds belonging to congeneric series: Third data set

2473 compounds with measured equilibrium or HT-solubilities were collected. Before the solubility data of these 2473 compounds were combined and used in the development of the prediction tool, the available solubility data obtained by both methods were compared.

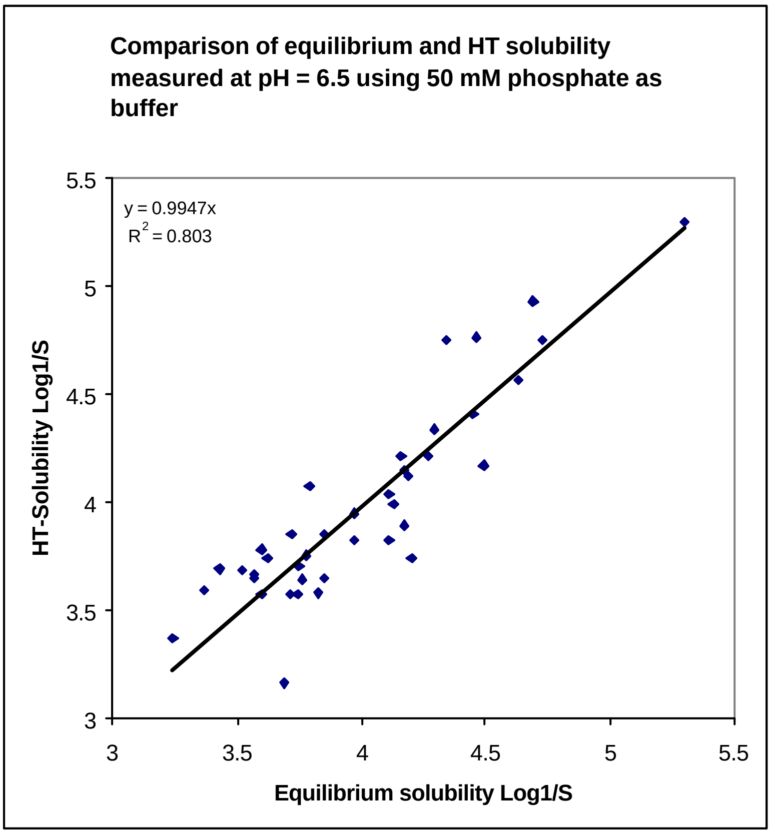


Figure 23: Comparison of solubility data obtained using equilibrium and HT solubility measurements. S is the molarity of molecular species.

Figure 23 shows the correlation of both solubility measuring methods. Usually, data generated by these methods do correlate well. Due to the differences in the solid state properties, sometimes, differences can occur in solubilities when lower crystallinity is obtained after lyophilisation or compounds with low molecular weight are lost during the evaporation process. Therefore, HT solubility results were left out, when compounds showed high tendency to form amorphous materials.

#### 4.2.1.1.4 Criteria for selection of high quality solubility data

Solubility data were collected for neutral and ionizable compounds. In order to overcome difficulties due to the ionization, the intrinsic solubility was calculated according to the Henderson-Hasselbalch[76] equation. The aqueous solubility used for prediction was expressed as log1/$S_0$. $S_0$ is the molarity of the unionized molecular species.

In order to ensure high quality data used for the development of prediction tools, all three data sets mentioned above were selected based on the following criteria:

1. Availability of aqueous solubility data determined by potentiometric titrations with 0.15 N KCl as background solution or by equilibrium solubility and HT-solubility methods with 50 mM phosphate as buffer. In case of equilibrium and HT solubility measurements, pH value of the saturated solution and measurement temperature were registered.
2. The solubility of compounds available as salts were considered, when no large pH shift was observed for saturated solutions or the intrinsic solubility of the neutral form was determined via the potentiometric method.
3. HT-solubility data were only taken, when compounds did not show high tendency to form amorphous materials.
4. Ionizable compounds were used in the data set only if their $pK_a$ values were known or could be derived from structural similar compounds.
5. For the first data set, crystal structures were collected from the Roche X-ray or Cambridge structure database (CSD)[75].
6. For the second data set, the experimental melting points were mandatory.

### 4.2.1.2 Descriptors

#### 4.2.1.2.1 Property based descriptors

35 descriptors assumed to influence both the crystal energy and solute water interactions were considered to model solubility. Except melting points which were experimentally determined, the other 34 descriptors were calculated. These 34 descriptors were used to express the molecular size, polarity, flexibility, rigidity, electronic properties, formation of hydrogen bonds, hydrophilicity and lipophilicity of the molecules.

A detailed overview on used 2D and 3D descriptors is given in Table 11, together with the information on the applied software packages.

| Dimension | Descriptors | Tools |
|---|---|---|
| 2D | molecular weight<br>the number of aromatic rings<br>the number of non aromatic rings<br>the number of possible internal hydrogen bonds<br>the number of rotable bonds<br>the number of nitrogen and oxygen atoms<br>the number of aromatic atoms<br>the number of aliphatic carbons | CALLISTOGEN[69] |
| | lipophilicity | KOWWIN[6]<br>ClogP[77] |
| | $pK_a$ | ACD[78] |
| 3D | molecular volume and surface<br>hydrophilic volume and surface<br>hydrophobic volume and surface<br>the number of hydrogen donors and acceptors<br>the maximum, minimum and mean value of the hydrogen donor and acceptor strength<br>rotational volume<br>ovality<br>Rg<br>d0, d1, d2 | MOLOC[68] |
| | $E_{min1-3}$<br>$HL_{1-2}$<br>A<br>CP | Volsurf[79] |
| | dipole moment<br>polarizability<br>HOMO LUMO gap | VAMP[80] |

Table 11: 2D and 3D descriptors listed together with the applied softwares. For explanation of d0, d1, d2, $E_{min1-3}$, $HL_{1-2}$, A, CP see the chapter of Abbreviation.

Before the calculation of 3D descriptors for the first data set, the crystal structures obtained from CSD[75] were inspected for the correct adjustment of hydrogen atoms using MOLOC[68]. 3D descriptors were then generated by keeping the crystal 3D structure fixed. 3D descriptors for the second data set were derived after the conversion of 2D to 3D-structures using CORINA[81].

In order to evaluate the quality of the calculated parameters used in the development of improved solubility prediction tools, measured and calculated descriptors were compared.

### 4.2.1.2.2 Evaluation of the property based descriptors

#### *4.2.1.2.2.1 Dipole Moments*

Dipole moments were calculated using VAMP[80]. VAMP[80] is a AM1 based method which uses the natural atomic orbital/point charge (NAO-PC) model to calculate the molecular electrostatic potentials. Calculated VAMP[80] dipole moments were compared with experimentally measured ones to estimate their quality. (Table 12, Figure 24)

| Molecule | Dipole moment(vamp[80]) [Debye] | Dipole moment(exp) [Debye] |
|---|---|---|
| $C_2H_5OH$ | 1.57 | 1.69 |
| $C_6H_5CH_3$ | 0.47 | 0.36 |
| $CH_2Cl_2$ | 1.51 | 1.57 |
| $CH_3Cl$ | 1.68 | 1.87 |
| $CH_3OH$ | 1.70 | 1.71 |
| $CHCl_3$ | 0.99 | 1.01 |
| $H_2O$ | 1.87 | 1.85 |
| $NH_3$ | 1.92 | 1.47 |
| $C_6H_4(CH_3)_2$ | 0.78 | 0.62 |

Table 12: 9 compounds with experimental and calculated dipole moments.

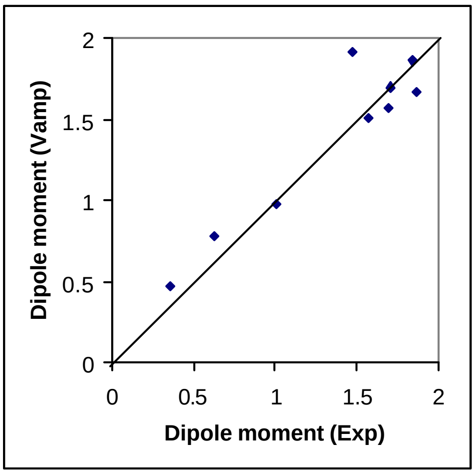
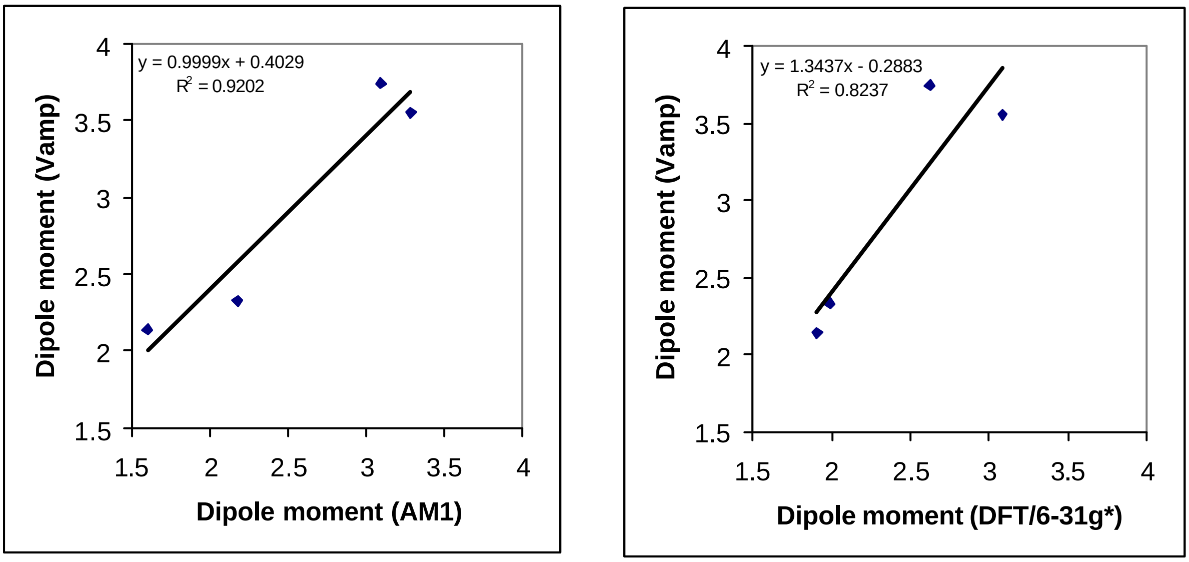


Figure 24: Comparison of experimental and calculated dipole moments using VAMP[80].

In Table 12, 9 small organic compounds are listed and their predicted dipole moments meet experimental values well. (Figure 24)

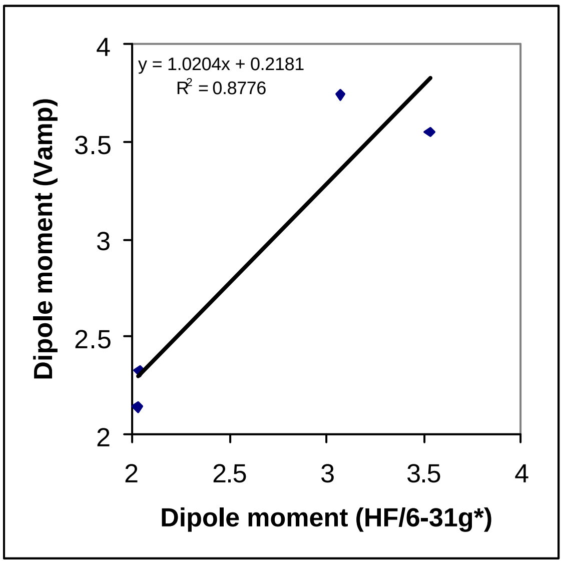
| Molecule | Dipole moment (GAUSSIAN[82]) [Debye] | | | Dipole moment (VAMP[80]) [Debye] |
|---|---|---|---|---|
| | AM1 | DFT/6-31g* | HF/6-31g* | AM1/NAO |
| compound 12 | 3.09 | 2.63 | 3.07 | 3.74 |
| Diazepam | 3.28 | 3.08 | 3.53 | 3.55 |
| Dimethomorph | 2.18 | 1.99 | 2.04 | 2.33 |
| compound 13 | 1.6 | 1.91 | 2.03 | 2.14 |

Table 13: 4 drug-like compounds are listed together with their dipole moments calculated with GAUSSIAN[82] and VAMP[80].



a)

b)



c)

Figure 25: Graphical comparison of dipole moments calculated with VAMP[80] and GAUSSIAN[82].

In the case of unknown dipole moments, dipole moments were calculated with Gaussian[82] using AM1, DFT/6-31g* and HF/6-31g* methods and the results were compared with the dipole moments calculated via VAMP[80]. (Table 13, Figure 25) It is well known, that 6-31g* ab initio methods give the most similar electronic property results to those observed in the X-ray structure[83-85]. However, for the four drug-like molecules, it took two hours to do electropotential calculations with DFT and two days with HF. Therefore, the AM1 method of VAMP[80] was the preferred choice for the calculation of dipole moments. It was much faster and the dipole moments calculated with VAMP AM1 were close to those obtained with ab initio methods. (Table 13, Figure 25)

#### *4.2.1.2.2.2 Lipophilicity*

Lipophilicity is an important descriptor for the prediction of solubility. 664 compounds with experimentally measured lipophilicity values were selected out of the third database to evaluate the error of the lipophilicity calculation program, e.g. ClogP[77]. These 664 compounds were classified in 51 congeneric series using the clustering package of Daylight[74]. 611 of them were uncharged and 53 had measured $pK_a$ values. The LogP values of charged compounds were calculated, according to the

Henderson-Hasselbalch[76] equation, using measured LogD and p$K_a$ values as input parameters.

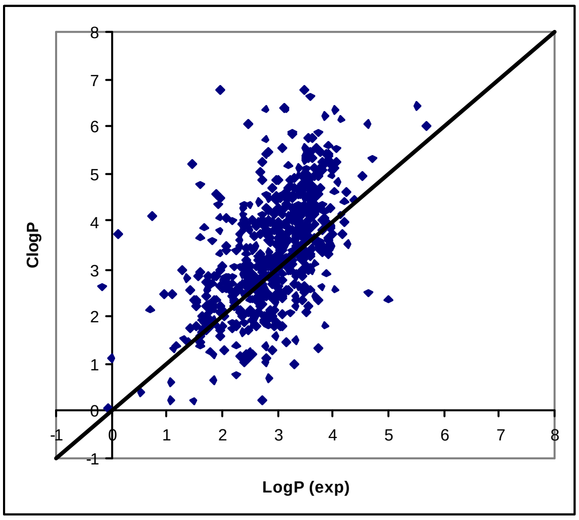


Figure 26: Experimental lipophilicity of 664 compounds is plotted against ClogP[77].

In Figure 26, experimental lipophilicities are plotted against the values calculated with ClogP[77]. It is obvious that the values calculated with ClogP[77] do not correspond to the experimental lipophilicities. The average standard error between ClogP[77] and experimental value was ±0.849 log units.

Nevertheless, the correlation between experimental and calculated lipophilicity could be improved via Eq. 17, which took the index of congeneric series as additional indicator variables into consideration.

$$LogP = a * C\log P + \sum_{i=1}^{n} c_i * f_{series,i} + b \qquad \text{Eq. 17}$$

where a is the coefficient of ClogP[77]. $c_i$ is the constant for the congeneric series $f_{series,i,}$. b describes the constant term in Eq. 17.

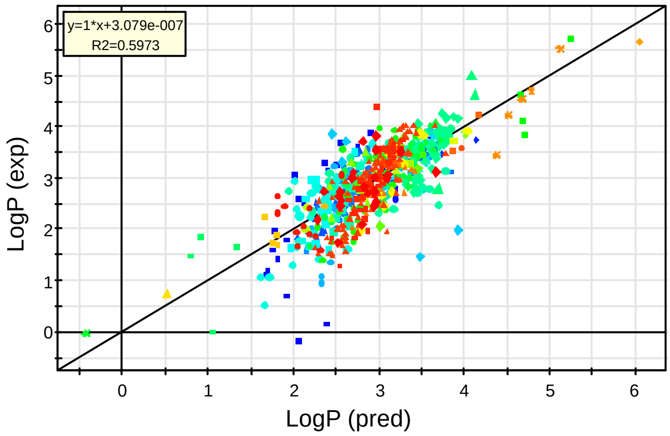


Figure 27: The lipophilicity of 664 compounds in 51 congeneric series is calculated with Eq. 17 and plotted against experimental logP. Different colors and shapes are used to identify these 51 congeneric series.

$R^2 = 0.597$, $Q^2 = 0.482$ and rmse = 0.481 were calculated using SIMCA[70]. The correlation between experimental and calculated lipophilicity was improved. The average standard error between calculated and experimental lipophilicity was ±0.371 log units. Hence, the congeneric series index can be used to correct the error in the lipophilicity calculation for diverse scaffolds and is helpful in the improvement of the solubility prediction for the compounds in the third data set.

#### *4.2.1.2.2.3 Melting Points*

The melting point is an useful factor for studying the solid cohesive energy of crystal packing. 51 compounds of the second data set were used to testify the predictive power of the program MPBPVP[86,87], which is the only commercially available tool for this task. (Figure 28)

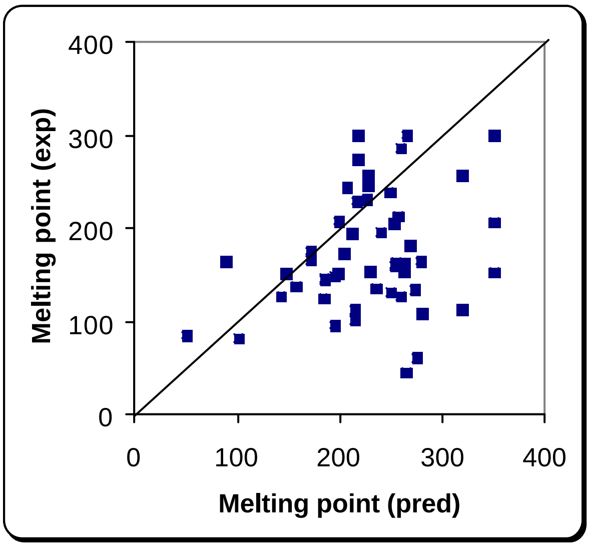


Figure 28: Comparison of experimental and calculated melting points using MPBPVP[86,87].

Figure 28 shows that program MPBPVP[86,87] fails to sufficiently predict the melting points of the second data set. Therefore, a new model had to be developed, in order to allow the consideration of melting points in the solubility prediction.

##### *4.2.1.2.2.4 pK$_a$*

pK$_a$ values were considered in the calculation of the solubility shift caused by the ionization. In case when no experimental pK$_a$ values were available, the program ACD[78] could be used to predict the ionization constants. In order to evaluate the predictive power of ACD[78], 23 structural similar compounds were selected and are listed together in Table 14. The first 13 compounds in Table 14 had their pK$_a$ values measured and the remaining 10 compounds had no measured pK$_a$ values.

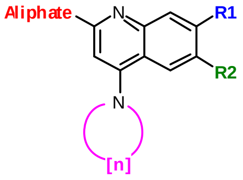
| Nr. | Structure | BpK$_a$ (exp) | BpK$_a$ (ACD) | S (µg/mL) | pH |
|---|---|---|---|---|---|
| 1 |  | 9.85 | 4.09 | - | - |
| 2 | | 9.81 | 4.26 | - | - |
| 3 | | 9.92 | 3.99 | - | - |
| 4 | | 9.99 | 4.04 | - | - |
| 5 | | 10 | 4.17 | - | - |
| 6 | | 9.67 | 3.61 | - | - |
| 7 | | 9.97 | 4.19 | - | - |
| 8 | | 9.55 | 4.05 | - | - |
| 9 | | 9.46 | 3.83 | - | - |
| 10 | | 9.86 | 4.49 | - | - |
| 11 | | 9.54 | 4.24 | - | - |
| 12 | | 9.11 | 3.15 | - | - |
| 13 | | 9.23 | 3.94 | - | - |
| 14 | | - | 4.26 | 5 | 6.5 |
| 15 | | - | 3.92 | 303 | 6.5 |
| 16 | | - | 4.27 | 229 | 6.5 |
| 17 | | - | 4.11 | 36 | 6.5 |
| 18 | | - | 4.27 | 6 | 6.5 |
| 19 | | - | 3.67 | 14 | 6.5 |
| 20 | | - | 3.67 | 27 | 6.5 |
| 21 | | - | 4.26 | 390 | 6.5 |
| 22 | | - | 3.93 | 84 | 6.5 |
| 23 | | - | 3.97 | 6 | 6.5 |

Table 14: Comparison of experimental and calculated pK$_a$ values. The calculation was performed with ACD[78] for 23 structural similar compounds.

Experimental and calculated pK$_a$ values of the first 13 compounds were compared, in order to decide whether pK$_a$ calculated by ACD[78] could be taken for the compounds without experimental pK$_a$ values. In comparison with otho- and meta-substituted quinoline derivates, the para-substituted ones are known to have a high base pK$_a$ value. An average BpK$_a$ could be calculated to 9.68 for the first 13 compounds in Table 14 using the experimentally measured data. The standard error of pK$_a$ shift caused by different substituent patterns was ±0.24 log units. However, ACD[78] treated the para-substituted quinoline derivates as compounds containing isolated quinoline moieties. Much lower base pK$_a$ values were calculated using ACD[78] than the experimentally determined. In order to overcome the limitations of the calculation tool for drug-like compounds, pK$_a$ values were adjusted considering information on structural similar compounds where several measured values of pK$_a$ existed. For example, for the last 10 compounds listed in Table 14 with known solubility but unknown $pK_a$ values, the formerly calculated average $pK_a = 9.68$ with a standard error of ±0.24 log units were used in the prediction tool development for the correction of solubility shift caused by ionization. Such manual $pK_a$ adjustment was performed for 2473 compounds collected in the third data set for the development of an superior solubility prediction tool, as well. Among them, 983 were uncharged at pH = 6.5, 166 had their $pK_a$ values measured and for the remaining 1324 charged compounds, $pK_a$ values were assigned as described.

#### 4.2.1.2.3 Fragment based descriptors

The structural fragmentation scheme of ClogP[77] was found to be the easiest way to obtain molecular fragments. In ClogP[77], the molecules are dissected according to the rule of “Isolating Carbon”. An “Isolating Carbon” atom (IC) is a carbon which is not double- or triple-bonded to a hetero atom[77]. Isolating carbons can, however, be multiply bonded to one another, such as those in $CH_3CH{=}CH_2$. An IC is an atomic fragment that, for calculation purposes at least, is always hydrophobic. Any hydrogen atom attached to an isolating carbon (ICH) is also a hydrophobic atomic fragment. All atoms or groups of covalently bonded atoms that remain after removal of ICs and ICHs are polar fragments. Thus a polar fragment contains no ICs but each has one or more bonds to ICs. These bonds are used to label the environments of a polar fragment, and are usually designated as A for aliphatic, Z for benzyl, V for vinyl, Y for styryl and a for aromatic.

Smarts[77] is a language for the specification of substructures using rules that are straightforward extensions of Smiles[77]. In order to enable flexible and efficient fragment search, Smarts[77] notations were used to reproduce the five connection environments defined in ClogP[77]. (Table 15)

| Type | Symbol | Smarts |
|---|---|---|
| Alkyl | A | [C; !$(*=,#[!#6]); !$(C(-*)a; !$(*=C)] |
| Benzyl | Z | [C; !$(C=*); $(C(-*)a)] |
| Vinyl | V | [C; $(*=C); !$(*=Ca); !$(*(=C)a] |
| Styryl | Y | [C; $(*=Ca); $(*(=C)a)] |
| Aromatic | a | [c; !$(*=, #[!#6])] |

Table 15: Smarts[77] notations for the five connection environments of the “Isolating Carbon”.

The difference in the application of ClogP[77] connection environments and the newly defined ones is shown for benzyl and styryl substituents on the first position of hetero aromatic ring system. (Table 16)

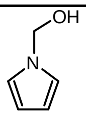
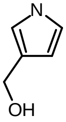
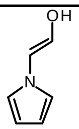
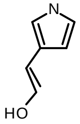
| Nr. | Hydroxyl group | Newly defined Fragment | ClogP fragment |
|---|---|---|---|
| 1 |  | **(Z)[OH]** | **(A)[OH]**<br>**not correct** |
| 2 |  | (Z)[OH] | (Z)[OH] |
| 3 |  | **(Y)[OH]** | **(V)[OH]**<br>**not correct** |
| 4 |  | (Y)[OH] | (Y)[OH] |

Table 16: Comparing the definition of newly defined connection environments with ClogP[77]. A and Z as defined in Table 15.

Table 16 shows that ClogP[77] treats the nitrogen atom in pyrrole rings as an aliphatic atom, which is not chemically right defined. In contrast, the newly-defined connection environments overcome this problem. The same nitrogen atom is correctly handled as an aromatic atom, which meets the chemical definition of aromatic atom well. Thus, the hydroxyl group in compound 1 and 3 is correctly recognized as benzyl and styryl bounded substituent using the newly-defined connection environments.

Additionally, the ClogP[77] fragments are so defined that each heavy atom in the molecule belongs only to one certain fragment. Thus, the presence of fragments can be easily checked. Eq. 18 shows, the total number of heavy atoms in a molecule should be equal to the sum of the number of the heavy atoms in the fragments. The precondition for this equation is the availability of all fragments of this molecule in the newly developed database.

$$\mathrm{Nr}_{\text{heavy atoms of a molecule}} = \sum_{i=1}^{n} \mathrm{Nr}_{\text{heavy atoms of a fragment}} \qquad \text{Eq. 18}$$

where n is the number of the fragments in a molecule.

Due to this simplified test method for missing fragments, ClogP[77] fragments were preferably used to Kowwin LogP[6] fragments as descriptors for the development of solubility prediction tool.

In addition to the 170 structural fragments defined in the chapter Appendix, four fragments were used as correction factors to improve the predictive power of the new solubility tool. (Table 17)

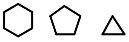
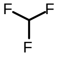
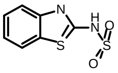
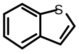
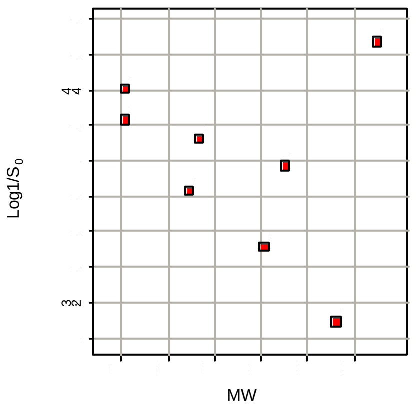
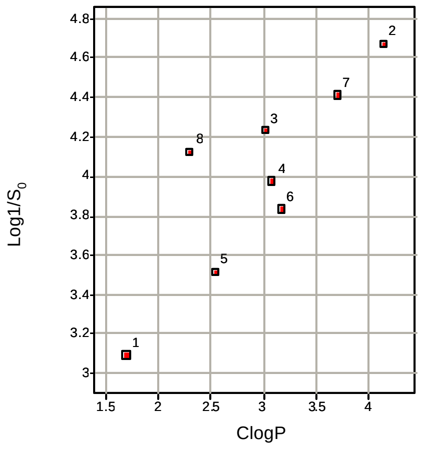
| Fragments as correction factors | Structures |
|---|---|
| Aliphatic ring |  |
| Trifluoromethyl C(F)(F)F |  |
| aS(=O)(=O)[NH]c1sc2ccccc2n1 |  |
| s1ccc2ccccc12 |  |

Table 17: Four correction factors for the solubility prediction.

#### *4.2.1.2.3.1 Evaluation of the fragment based descriptors*

Fragment based descriptors were used to predict the solubility of drug-like compounds in congeneric series. Figure 29 takes the derivates of diazepam as example to demonstrate the usage of fragments. Usually, increasing the lipophilicity and the molecular weight results in reduced solubility[6]. However, diazepam and temazepam have higher solubility values than nordiazepam and oxazepam, although their lipophilicity and molecular weight is higher. (Figure 29)





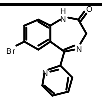
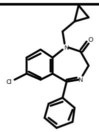
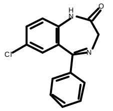
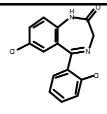
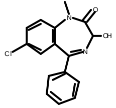
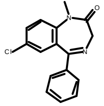
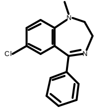
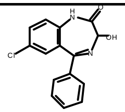
| Nr | Compound | Structure | MW | Melting point (°C) | ClogP | Log1/S₀ |
|---|---|---|---|---|---|---|
| 1 | Bromazepam |  | 316.16 |  | 1.703 | 3.09 |
| 2 | Prazepam |  | 324.81 |  | 4.143 | 4.67 |
| 3 | Nordiazepam |  | 270.72 | 216 | 3.021 | 4.23 |
| 4 | 7-Chloro-5-(o-chlorophenyl)-1,3-dihydro-2H-1,4-benzodiazepin-2-one |  | 305.16 |  | 3.084 | 3.97 |
| 5 | Temazepam |  | 300.74 | 119 | 2.549 | 3.51 |
| 6 | Diazepam |  | 284.75 | 132 | 3.17 | 3.83 |
| 7 | Medazepam |  | 270.78 |  | 3.71 | 4.41 |
| 8 | Oxazepam |  | 286.72 | 197 | 2.305 | 4.12 |

Figure 29: The correlation between solubility, lipophilicity and molecular weight for diazepam derivates.

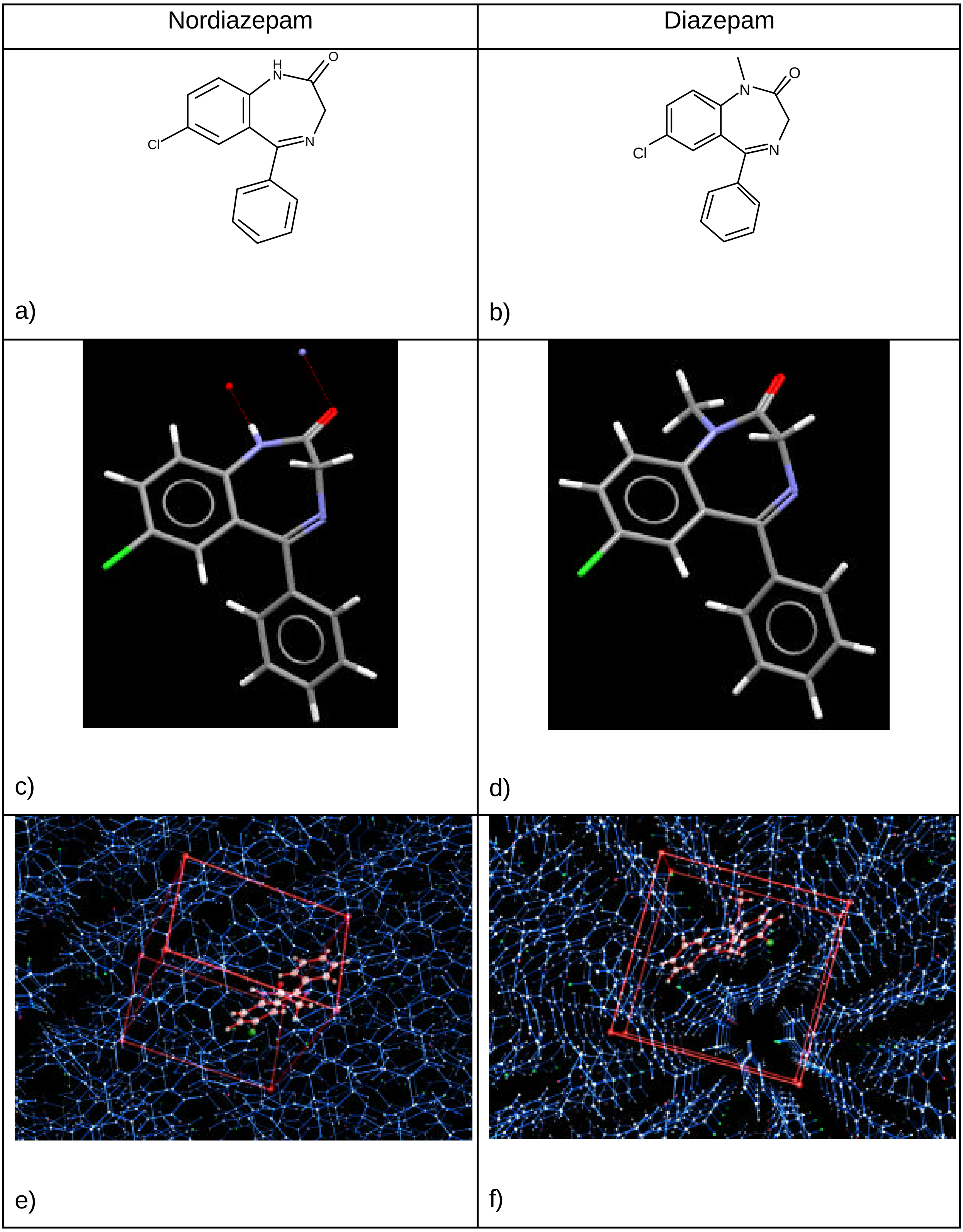


Figure 30: a) 2D structure of nordiazepam; b) 2D structure of diazepam; c) 3D crystal structure of nordiazepam; d) 3D crystal structure of diazepam; e) crystal packing of nordiazepam; f) crystal packing of diazepam.

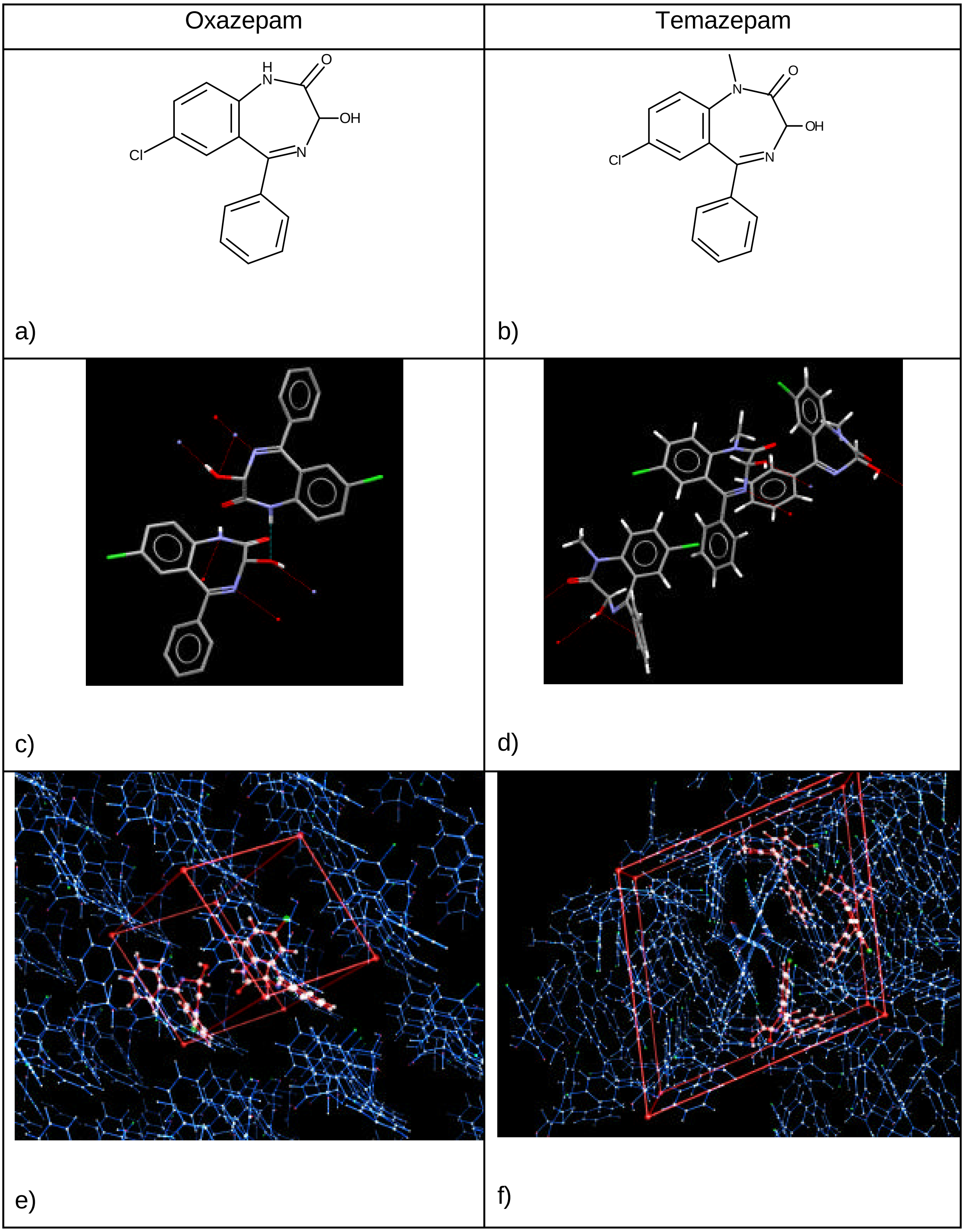


Figure 31: a) 2D structure of oxazepam; b) 2D structure of temazepam; c) 3D crystal structure of oxazepam; d) 3D crystal structure of temazepam; e) crystal packing of oxazepam; f) crystal packing of temazepam.

The abnormal solubility phenomenon observed for the benzodiazepines (Figure 29) can be explained by comparing the compounds' crystal structures and melting points. The N-alkylation of the amide group in the temazepam replaces the amido hydrogen atom in oxazepam, which is responsible for strong hydrogen and dipolar bonding within the crystal lattice. The melting point of temazepam is lower than that of oxazepam, which illustrates the remarkable impact of eliminating the amido hydrogen atom of the oxazepam molecule. The crystal structures shown in Figure 30 and Figure 31 reflect the effect of amido hydrogen atom, as well. The flat layer of oxazepam in the crystalline state is the result of its strong hydrogen bonding, which makes the process of dissolution much more difficult than for temazepam. An analogous example has been described by Goosen[88] on a series of thalidomide and its N-alkyl analogues. Therefore, two fragments, resp. (*)N(*)C(*)=O and (*)[NH]C(*)=O can be used as descriptors to consider the influence of the crystal lattice on the solubility and distinguish the methyl group as present in temazepam from the general aliphatic chains. Due to hydrogen bonding, diazepam and temazepam have higher solubilities than nordiazepam and oxazepam, despite of its higher molecular weight. Thus, in the series of diazepam derivates, the negative proportionality of molecular weight to $\log 1/S_0$ shows, that the molecular weight is not always a suitable descriptor for solubility prediction. The prediction of solubility can be improved when structural based fragments are used as descriptors, instead of molecular weight.

Inspecting the calculated lipophilicity of the benzodiazepines, several anomalies were detected. The increment of lipophilicity caused by methyl group is usually about 0.5 log unit. However, the ClogP[77] values of diazepam and nordiazepam differ only by 0.15 log units, although diazepam occupies a methyl group more than nordiazepam. Therefore, experimental values were collected in order to analyze the effect of small structural differences on lipophilicity.

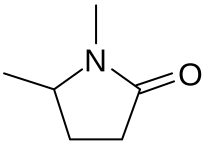
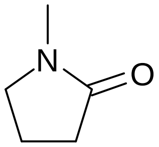
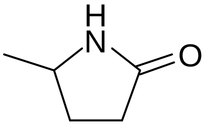
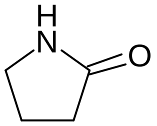
| | $\Delta$LogP$_{1-2}$ = 0.5 | | |
|---|---|---|---|
| $\Delta$LogP$_{1-3}$ = 0.03 | <br>LogP = -0.07<br>**1** | <br>LogP = -0.54<br>**2** | $\Delta$LogP$_{2-4}$ = 0.3 |
| | <br>LogP = -0.1<br>**3** | <br>LogP = -0.85<br>**4** | |
| | $\Delta$logP$_{3-4}$ = 0.75 | | |

Table 18: The experimental lipophilicity[89] of four similar cyclic amides.

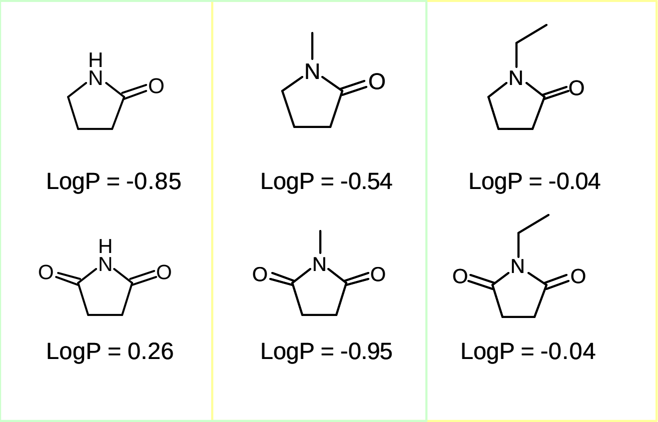


Figure 32: The experimental lipophilicity[89] of six similar cyclic amides.

Table 18 and Figure 32 show that the change in lipophilicity by the addition of an amido methyl group is normally $\Delta$LogP = 0.3, much smaller than a normal methyl group ($\Delta$LogP = 0.5).

According to these findings, a fragmental constant for the amido methyl group was introduced and used in the model development for the third data set.

## 4.2.2 Methods

### 4.2.2.1 Data analysis

Multivariate data analysis was performed using the program SIMCA[70]. Variable preprocessing was performed. Thus, all the descriptors were mean-centered and scaled to unit variance (UV). Descriptors with a higher skewness than 1.5 were log-transformed. Principal component analysis (PCA) was performed to get an overview on the data sets. The information contained in original variables was summarized by calculation of new latent variables. The compounds, which could not be well explained with the latent variables were classified as outliers in PCA. Outliers conforming to the overall correlation structure, but occupying extreme characteristics were strong outliers and were identified using the 95% tolerance interval signified as ellipse in the PCA loading plot[70]. Outliers found by inspecting residuals for each observation were moderate outliers and were identified by the “distance to the model in X space” (DModX) plot[70]. Furthermore, PCA loading plots were used to detect reason for the outliers in PCA and were sometimes helpful in explanation of PLS results. Projections to latent structure (PLS) was performed to predict solubility. The goodness of fit of a PLS model was given by a regression coefficient $R^2$. The goodness of prediction was evaluated by a cross-validated $R^2$, designated as $Q^2$. The $Q^2$ value was the main criterion for assessing the quality of a model. In general, a model with a $Q^2$ of 0.3 or higher is statistically meaningful, while a $Q^2$ greater than 0.5 is regarded as a good model and 0.9 or above is excellent[70]. Variable Influence on Projection (VIP) estimated the influence of every original variable on the matrix Y. Variables with larger VIPs were the most relevant for explaining Y, and those with VIPs less than 0.8 were of lesser importance[70].

The PLS models were refined through stepwise selection of the variables and exclusion of the outliers. The excluded variables were those which showed colinearity with other variables or had low importance on solubility prediction. A variable was excluded, if a more predictive model (higher $Q^2$) was obtained after exclusion. There were two criteria for identifying real outliers. First, the experimental value was wrong. Second, the compound showed great standard deviation in PLS Y-residue and its extreme characteristics caused heterogeneity in X-matrix. If by removing an outlier, the model was greatly improved, that outlier was dropped from the data set permanently. This refinement procedure was repeated until no further improvement of the model was achieved.

Once a model was chosen, it was validated by a permutation test using scrambled Y values to ensure that the model was not obtained by chance. The result of the response permutation test was summarized in the validation plot in SIMCA[70]. The $R^2$- and $Q^2$ intercept in the validate plot are interpretable as measures of the significance of the model's predictive power[70]. A model with $R^2$Y-intercept below 0.3-0.4 and the $Q^2$ intercept below 0.05 can be assumed not to be overfitted[70].

In case of large data sets (N >100), the data set was divided into a training data set and a test data set. A PLS model usually was built by only using the training data set and obtained model was tested with an independent test data set. When additional observations were available, they were also used to test the predictive power of the model.

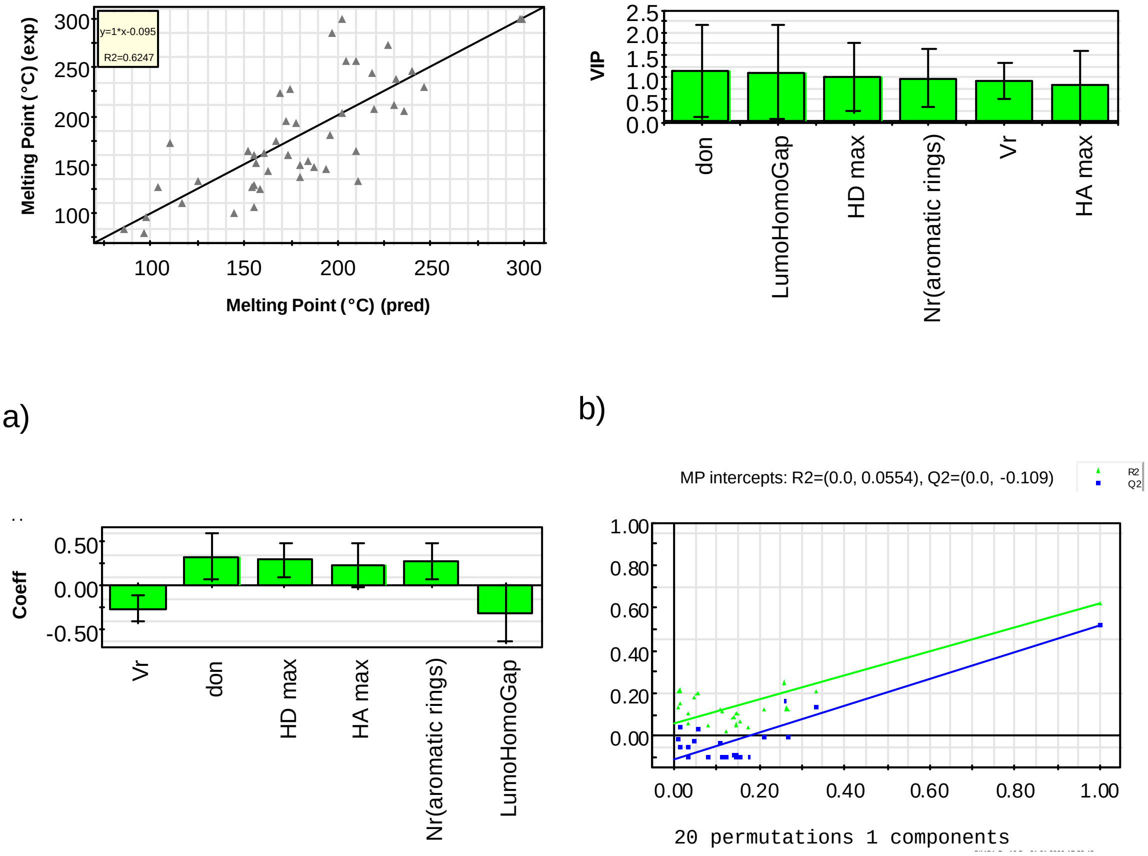
# 4.3 Results and discussions

## 4.3.1 Melting point prediction

51 compounds of the second data set were used to develop an improved model for the prediction of melting points.

After the PCA analysis and descriptor selection, three outliers were detected and left out of the model. The first one had a melting point of 44°C, while the other compounds had melting points in the range of 80 to 300°C (Figure 33a) and therefore were excluded. The second one was identified to have a higher melting polymorphic form and therefore its current registered melting point was not reliable. The third one showed the highest residue value in the resulting PLS model and exclusion of this compound enhanced the predictive power of the model dramatically. After outlier detection, a much more improved model (Eq. 19), in comparison to the program Mpbpvp[86,87], was obtained with $R^2 = 0.625$, $Q^2 = 0.518$ and rmse = 36.601 for 48 compounds. (Figure 33)

MP (°C) = 13.3671 * the number of hydrogen donors – 12.7269 * LumoHomo gap + 12.19 *the maximum value of hydrogen donor strength + 70.4612 * the maximum value of hydrogen acceptor strength + 15.0007 * the number of aromatic rings – 85.8744 * rotational volume + 213.05 Eq. 19



a)

b)

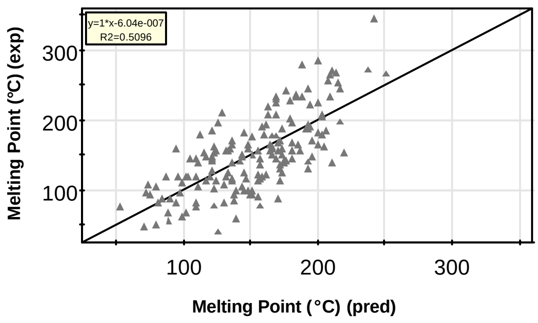
c)

d)

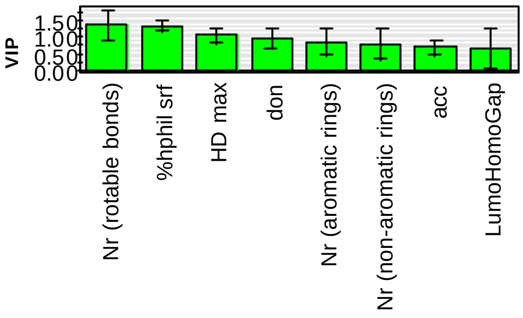
Figure 33: The final melting point model. a) Correlation between experimental and predicted melting point; b) PLS VIP plot; c) PLS coefficient plot; d) PLS permutation test.

Descriptors used in Eq. 19 are listed in the VIP plot (Figure 33b), according to their importance for explanation of melting points. The number of hydrogen donors (don) was found as the most important descriptor followed by LUMO HOMO Gap, the maximum value of hydrogen donor strength (HD max), the number of aromatic rings, the rotable volume and the maximum of hydrogen acceptor strength (HA max). The coefficient plot (Figure 33c) shows that the most descriptors responsible for hydrogen bonding were positively correlated to melting point (don, HD max, HA max) except LUMO HOMO Gap, which could be used to describe the hydrogen bonding strength. The higher the gap between LUMO and HOMO, the more energy is necessary to bring the electron from HOMO to LUMO orbital and more difficult is the formation of hydrogen bonds. Therefore, the LUMO HOMO Gap in Figure 33c showed a negative proportionality to the melting point. Furthermore, ring structures were found to increase the melting point, whereas a large degree of molecular flexibility resulted in a lowered melting point. The permutation test shows that the PLS model was well validated.

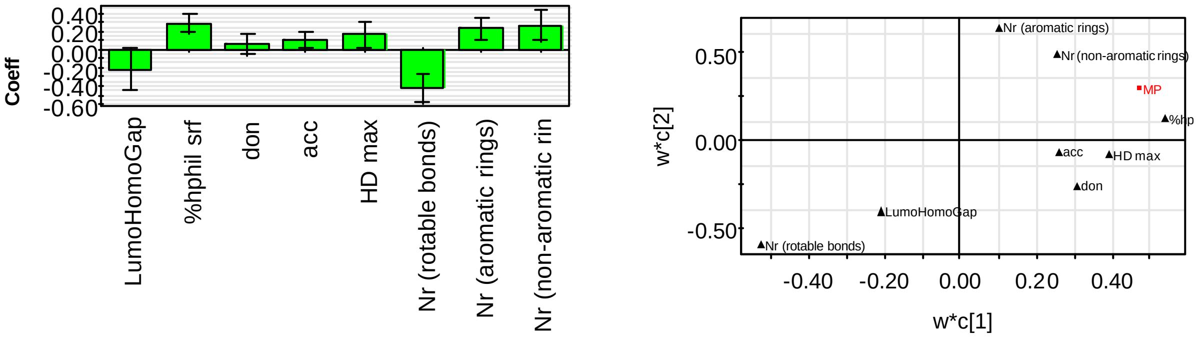
The important variables detected in the above mentioned model express similar molecular properties as those used by Bergström[90] in her melting point study of drug-like compounds. In order to reproduce the prediction results obtained by Bergström[90], her training data set with well characterized melting points was taken and used for the model development.



a)



b)

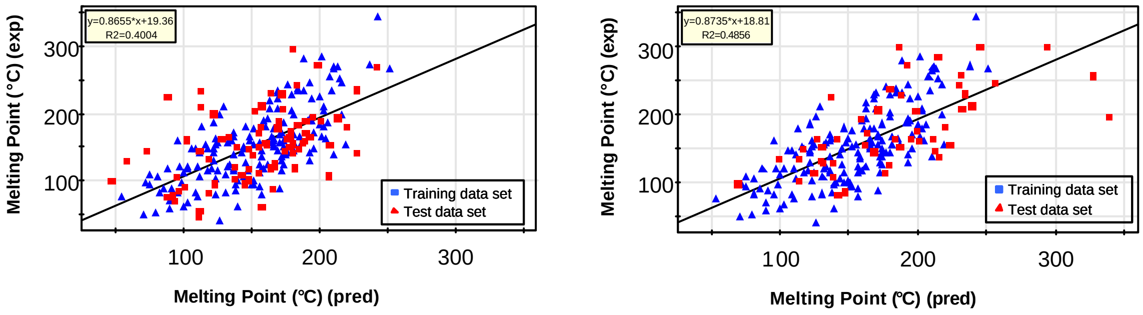


c) d)

Figure 34: Melting point model developed using the training data set of Bergström[90]. a) Correlation between experimental and predicted melting points; b) PLS VIP plot; c) PLS coefficient plot d) PLS loading plot.

A melting point model was obtained with $R^2$ = 0.51, $Q^2$ = 0.463 and rmse = 38.6761. The VIP plot shows the most important descriptors detected in this melting point model were responsible for the molecular flexibility, rigidity, polar surface and the formation of intermolecular hydrogen bonds, which agreed with the descriptors identified by the former melting point model and the original published model of Bergström[90]. The coefficient plot shows that melting points increased with the formation of hydrogen bonds, polar surface, the molecular rigidity and decreased with the molecular flexibility. The loading plot shows the contribution of the descriptors to the melting point was similar weighted as those in the model of Bergström[90].

The obtained model was validated using the test data set of Bergström[90] and 48 compounds of the second data set. (Figure 35)



a) b)

Figure 35: Melting point model validated with a) the test data set of Bergström[90]; b) 48 compounds of the second data set.

Figure 35 shows that the developed model could be used to predict the melting points of the test data set of Bergström[90] and the 48 compounds of the second data set.

According to the frequency plot showed for the Bergström[90] data set together with 48 compounds of the second data set, melting points can be divided into three categories.

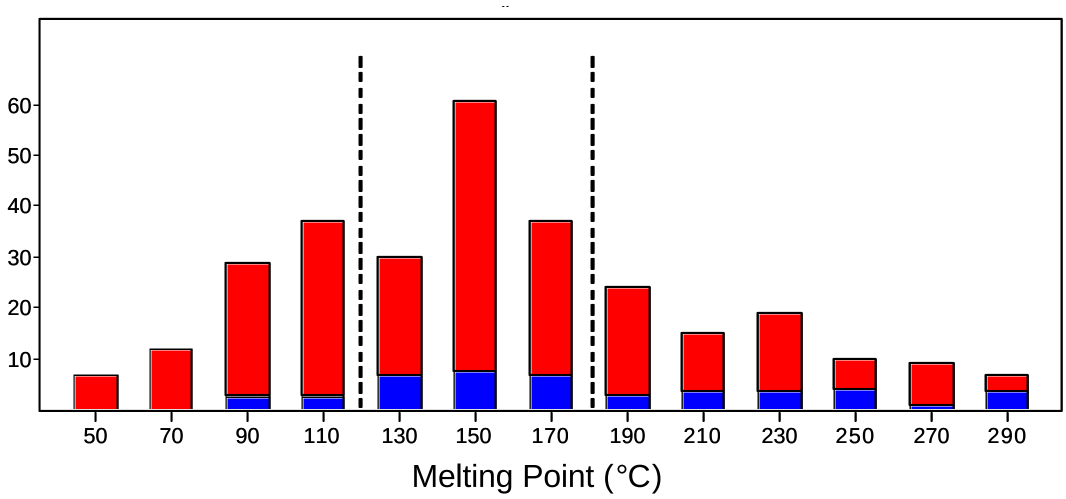


Figure 36: Frequency plot of the Bergström[90] data set (red) and 48 compounds of the second data set (blue). The bars present the bin centers ± 10 °C. The dashed lines show the cutoff between low, intermediate and high melting point values.

Figure 36 shows the majority of compounds displayed melting points between 120 and 180°C. Thus, 120 and 180 °C were used as thresholds to define low, intermediate and high melting point values. The classification results are listed in Table 19 and more than 50% of drug-like compounds with melting points from medium to high were correctly classified.

| Melting point (°C) | % correctness |
|---|---|
| 40 - 120 | 43.6 |
| 120 - 180 | 61.2 |
| 180 - 300 | 64.3 |

Table 19: The correctness of melting point classification.

### 4.3.2 Solubility prediction considering crystal structure information

74 compounds of the first data set were divided into two data sets, according to the diversity in the conformations of the polymorphic forms. As already described in chapter 4.2.1.1.1, 59 compounds of the first data set had only one conformation in the internal and external CSD[75] database. Calculated 3D descriptors for the remaining 15 polymorphic compounds with more than one conformation were quite similar. Therefore, for the first data set, 3D descriptors were considered independent on the conformation. Hence, all 74 compounds were used for the development of a solubility prediction model. Lipophilicities of some compounds were found to be falsely calculated. Their experimental and calculated lipophilicities are listed together in the following tables. (Table 20 and Table 21)

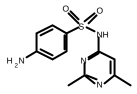
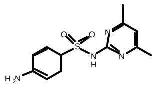
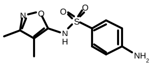
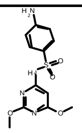
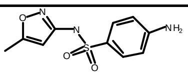
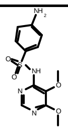
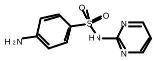
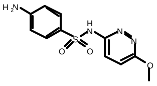
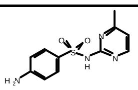
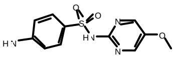
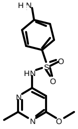
| Name | Structure | KowlogP | ClogP | logP (exp) |
|---|---|---|---|---|
| sulfisomidine |  | 0.757 | 1.097 | -0.3 |
| sulfamethazine |  | 0.757 | 1.097 | 0.89 |
| sulfisoxaz ole |  | 1.031 | 0.222 | 1.15 |
| sulfadimethoxine |  | 1.174 | 1.981 | 1.56 |
| sulfamethoxazole |  | 0.484 | 0.563 | 1.75 |
| sulfadoxine |  | -0.238 | 1.231 | 1.06 |
| sulfadiazine |  | -0.338 | 0.1 | -0.13 |
| sulfamethoxypyridazine |  | 0.198 | 0.41 | 0.4 |
| sulfamerazine |  | 0.21 | 0.599 | 0.13 |
| sulfameter |  | -0.257 | 0.648 | 0.46 |
| 2-methyl-4-methoxy-6-sulfanilamidopyrimidine |  | 0.745 | 1.547 | 0.61 |

Table 20: Comparison of the experimental and calculated lipophilicities for sulfonamide derivates.

Table 20 shows the lipophilicity of sulfonamide derivates were not correctly calculated, therefore the experimental values collected from MedChem[89] database were used instead of the calculated ones. The same was true for L-phenylalanine and an additional internal compound of the data set.

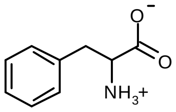
| Name | Structure | KowlogP | ClogP | logP (exp) |
|---|---|---|---|---|
| L-phenylalanine |  | -1.283 | -1.556 | 1.114 |
| compound 14 | | 4.212 | 4.38 | 5.3 |

Table 21: Comparison of the experimental and calculated lipophilicity for L-phenylalanine and an internal compound.

After the PCA analysis, descriptor selection and outlier detection, a PLS model was generated for 70 compounds with $R^2 = 0.827$, $Q^2 = 0.79$ and rmse = 0.576 (Eq. 20), where four outliers were omitted.

$$\text{Log}1/S_0 = 0.628723 * \text{KowlogP} + 0.0088498 * \text{MW} + 0.239609 * \text{the number of hydrogen donors} - 0.814466 * \text{the number of possible intramolecular hydrogen bonds} – 0.649414 \qquad \text{Eq. 20}$$

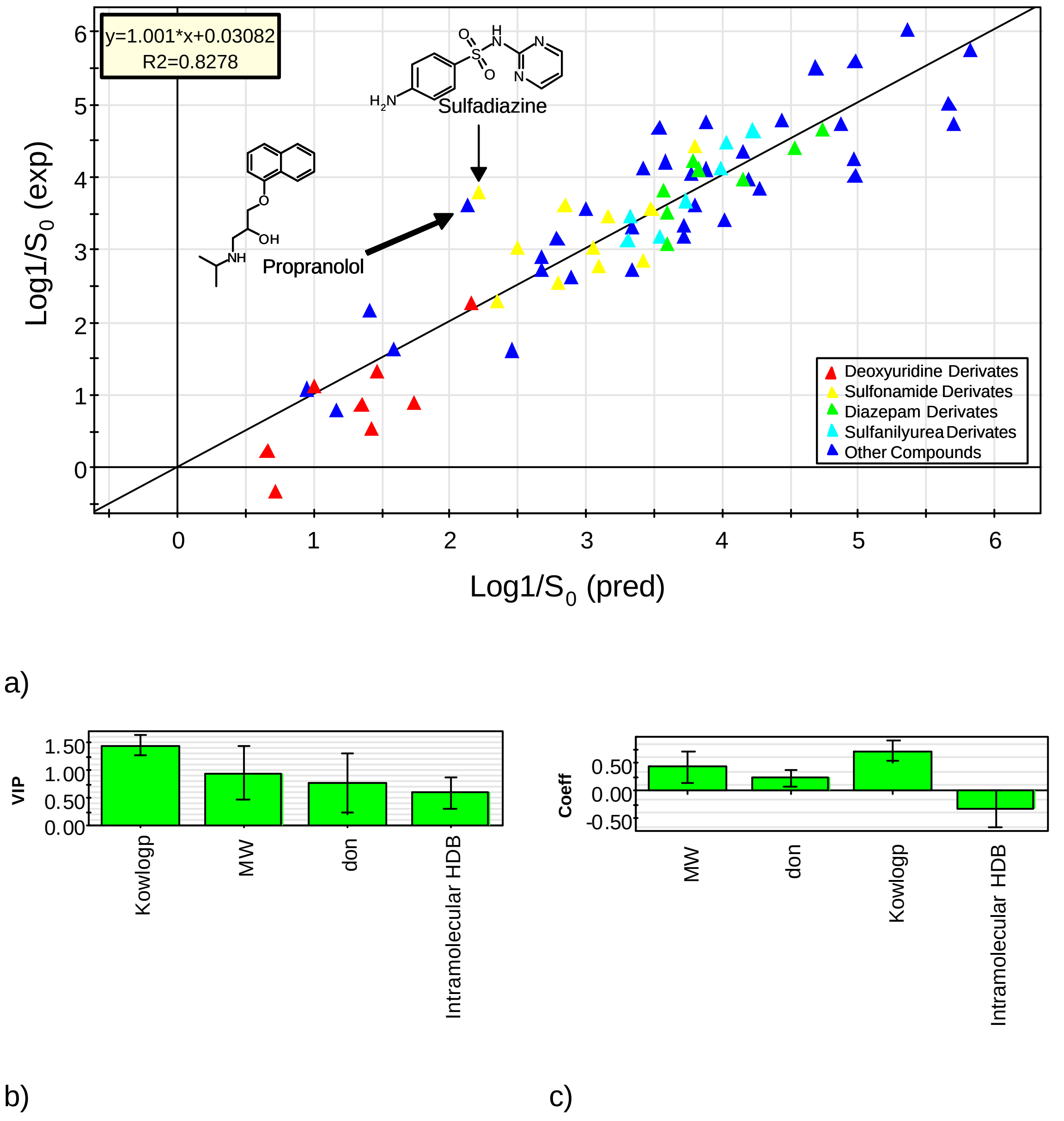
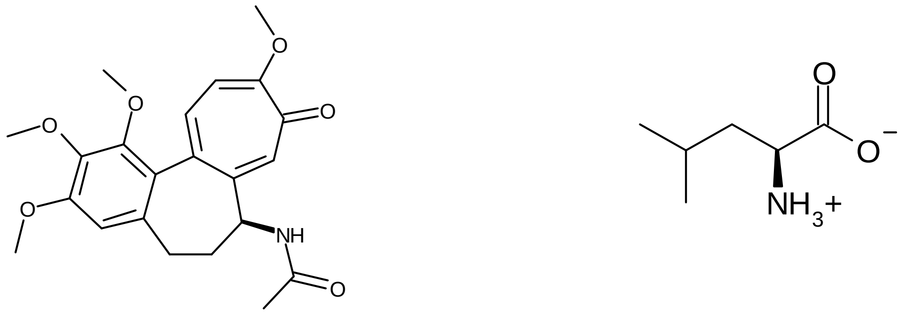


Figure 37: The final solubility model. a) Correlation between experimental and predicted solubility values; b) VIP plot; c) coefficient plot.

Descriptors used in Eq. 20 are shown in the VIP plot (Figure 37b), according to their importance for explanation of solubility in this model. Lipophilicity was found to be the most important descriptor followed by molecular weight, the number of hydrogen donors and the number of possible intramolecular hydrogen bonds. The coefficient plot shows (Figure 37c) that possible formation of intramolecular hydrogen bonds increased the value of solubility. Additionally, the higher the lipophilicity, molecular weight and the number of hydrogen donors, the lower the solubility. Furthermore, the importance of hydrogen donors, used in Eq. 20 confirmed the relationship between solubility and cohesive energy in solid state.



a) b)

Figure 38: Two Outliers. a) Colchicine; b) L-leucine.

Four outliers were detected and not included in the model generation. The first and second one had a molecular weight higher than 600 Da, while the others had a molecular weight in the range between 100 and 450 Da. The third one was colchicine, whose relative high solubility value (log1/$S_0$ = 1.24) could not be correctly predicted using its molecular weight (MW = 399.4 Da) as descriptor in this model. Exclusion of colchicine improved the predictive power of the model dramatically. The fourth one was L-leucine, the only compound in this data set without aromatic ring in its structure.

In Figure 37a, propranolol and sulfadiazine show the largest deviation between experimental and predicted solubility values. The prediction error of propranolol could be a result of its falsely calculated values of descriptors. According to the crystal structure registered in the CSD[75] database, no intramolecular hydrogen bond was observed for propranolol, although two intramolecular hydrogen bonds were calculated which led to a reduced predicted solubility value.

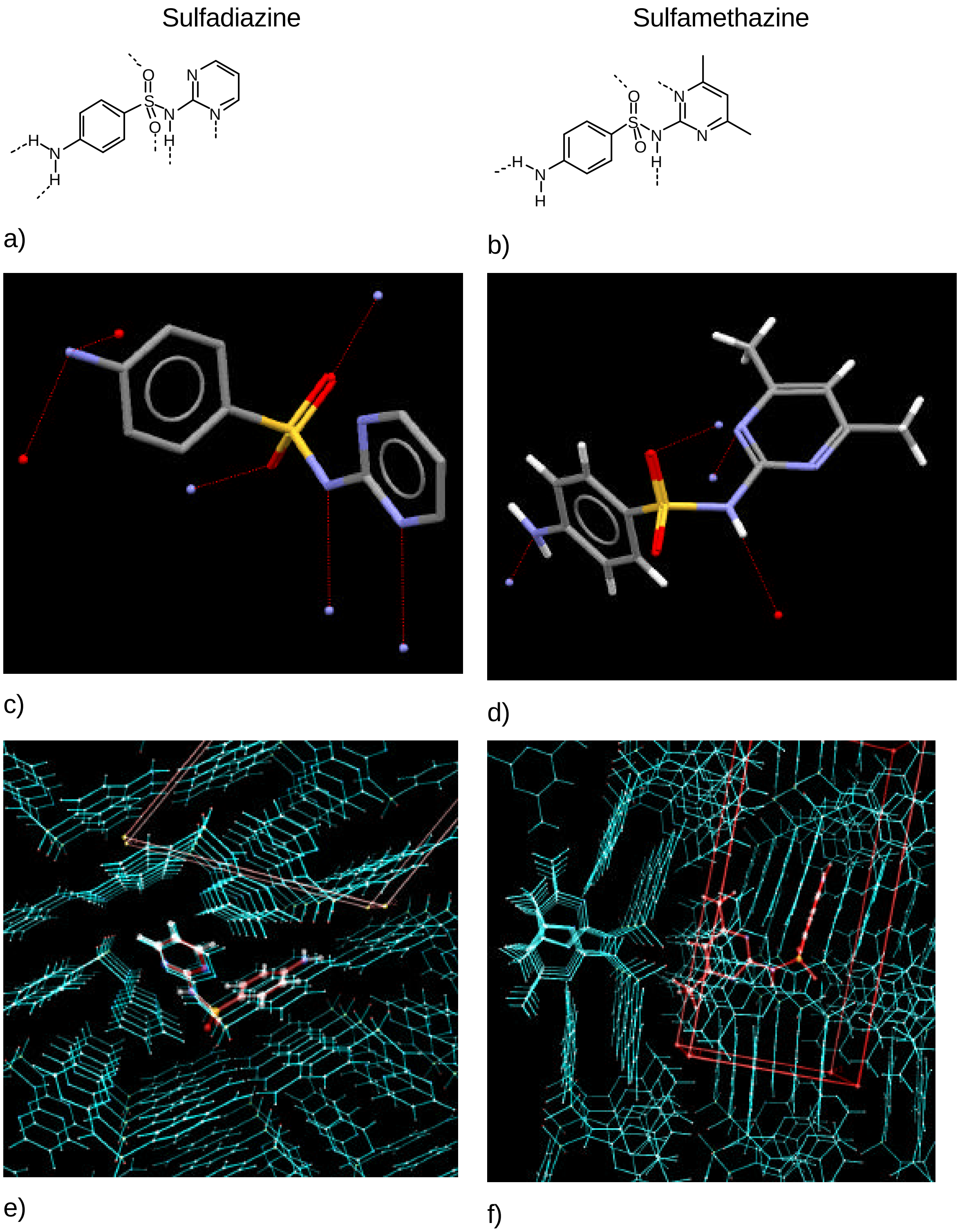


Figure 39: a) 2D structure of sulfadiazine; b) 2D structure of sulfamethazine; c) 3D crystal structure of sulfadiazine; d) 3D crystal structure of sulfamethazine; e) crystal packing of sulfadiazine; f) crystal packing of sulfamethazine.

Figure 39 shows that sulfadiazine and sulfamethazine have similar 2D structures, differing only by two methyl groups. However, by losing two methyl substituents, the molecular moiety containing the pyrimidine ring is in case of sulfadiazine flatter than sulfamethazine. Therefore, in contrast to sulfamethazine, the pyrimidine rings of sulfadiazine can be superimposed directly on top of each other and molecules build 6 intermolecular hydrogen bonds, which would lead to a higher density in the crystal packing, a higher energetic cost for crystal lattice degradation and therefore a poorer solubility.

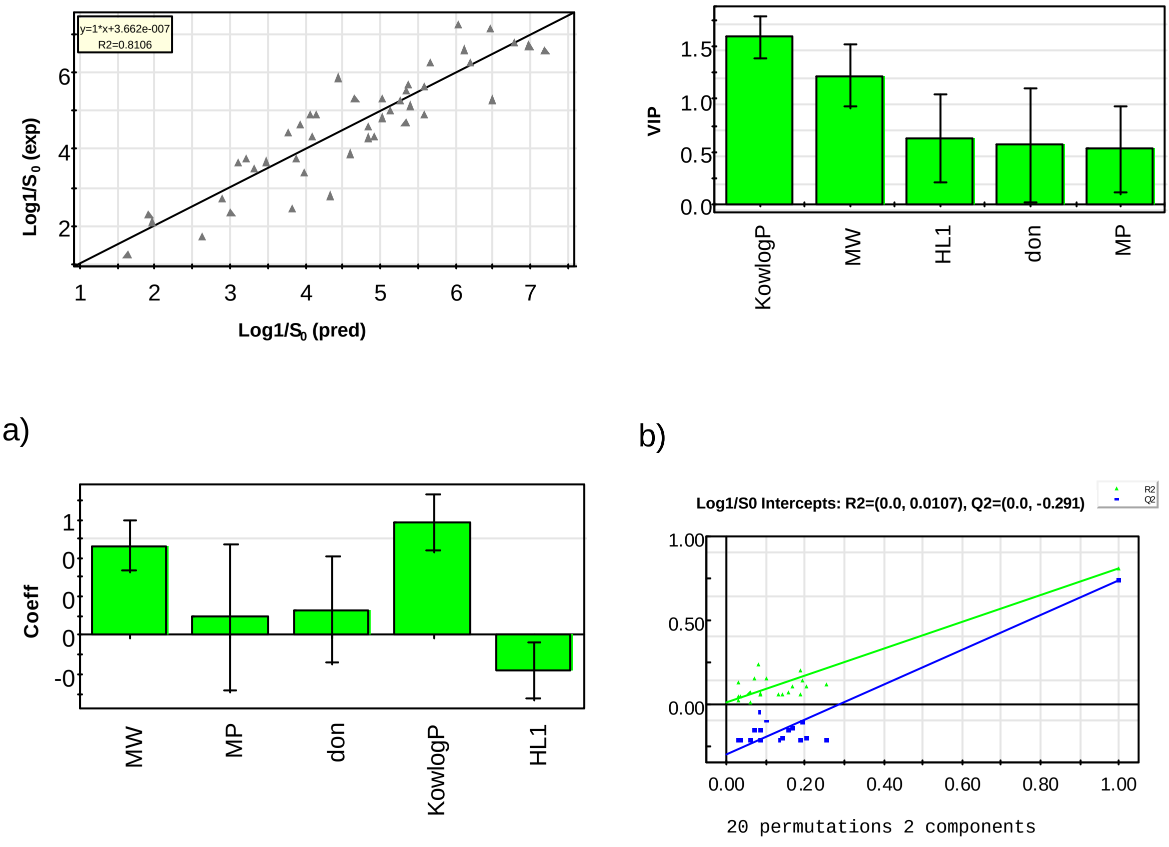
| Name | KowlogP | logP (exp) | ApK$_a$ | S (µg/mL) | pH | log1/S$_0$ (exp) | log1/S$_0$ (pred) | MP (°C) |
|---|---|---|---|---|---|---|---|---|
| Sulfadiazine | -0.338 | -0.13 | 7.45 | 113 | 6.5 | 3.8 | 2 | 255-256 |
| Sulfamethazine | 0.484 | 1.75 | 6.5 | 525 | 6.5 | 2.87 | 2.91 | 178-179 |

Table 22: The calculated and experimental lipophilicity, melting points and solubility of sulfadiazine and sulfamethazine.

Table 22 indicates that a higher melting point as a result of more intense crystal packing of sulfadiazine leads to lower solubility, although the lipophilicity of sulfadiazine is lower than sulfamethazine, which would indicate a trend in the other direction. Therefore, the significant prediction error for sulfadiazine can be assumed as a result of insufficient consideration of solid state properties.

### 4.3.3 Solubility prediction using melting point as a parameter

51 compounds of the second data set were used to develop a solubility prediction model, by considering melting point information. After PCA analysis, descriptor selection and outlier detection, a PLS model (Eq. 21) with $R^2 = 0.811$, $Q^2 = 0.746$ and rmse = 0.677 was obtained for 44 compounds, while seven outliers were identified and omitted.



a)

b)

c)

d)

Figure 40: The final solubility model. a) Correlation between experimental and predicted solubility; b) PLS VIP plot; c) PLS coefficient plot; d) PLS permutation test.

$$\mathrm{Log1/S_0} = 0.344659 * \mathrm{KowlogP} + 0.0076349 * \mathrm{MW} + 0.169565 * \text{the number of hydrogen donors} + 0.00251848 * \mathrm{MP} - 0.216864 * \mathrm{HL1} + 0.159355 \qquad \text{Eq. 21}$$

Descriptors used in Eq. 21 are listed in Figure 40b. According to its importance for explanation of solubility, lipophilicity was found as the most important descriptor followed by molecular weight, the hydrophilic-lipophilic balance, the number of hydrogen donors and melting point. The coefficient plot shows (Figure 40c), the higher the lipophilicity, molecular weight, the number of hydrogen donors, melting point, the lower the solubility. Additionally solubility increased with higher values of hydrophilic-lipophilic balance in molecule. The permutation test (Figure 40d) shows that the obtained PLS model was not overfitted.

Seven outliers were detected and left out, mainly due to exceptionally low or high descriptor values lying out of the covered descriptor range or possibly imprecise solubility values.

### 4.3.4 Solubility prediction for drugs in congeneric series

The third data set containing 2473 compounds in 81 congeneric series was used to develop an improved model for solubility prediction. Lipophilicity, 170 structural fragments plus 4 fragmental based correction factors and 81 congeneric series indices were used in the model generation. A model (Eq. 22) with $R^2 = 0.844$, $Q^2 = 0.79$ and rmse = 0.510 was obtained for 1515 compounds in the training data set. The quality of the model was tested with 958 compounds in the test data set and $R^2 = 0.813$ was obtained.

$$Log1/S_0 = 0.131493 * C\log P + \sum_{i=1}^{n=174} b_i * frag_i + \sum_{i=1}^{n=81} c_i * f_{series,i} + 3.7551 \qquad \text{Eq. 22}$$

Eq. 22 uses ClogP[77] to describe the liquid-liquid interaction in the solvation process and the fitted coefficients $b_i$ to study the cohesive energy caused by each fragment in the solid state. Thus, the solubility value of a fragment could be calculated, which was the sum of the fragmental contribution to lipophilicity and to crystal packing. The solubility values of fragments are listed in the appendix and used later for the solubility prediction of external data described in the literature. (chapter 4.3.4.1)

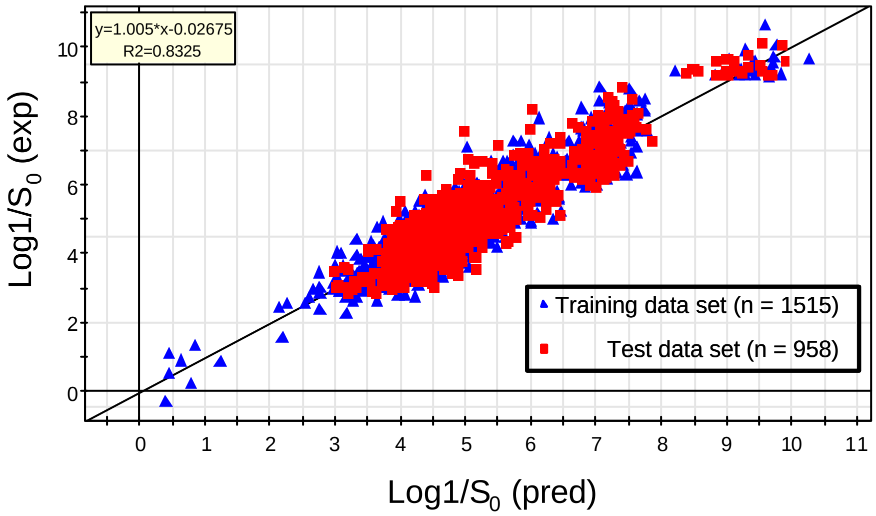


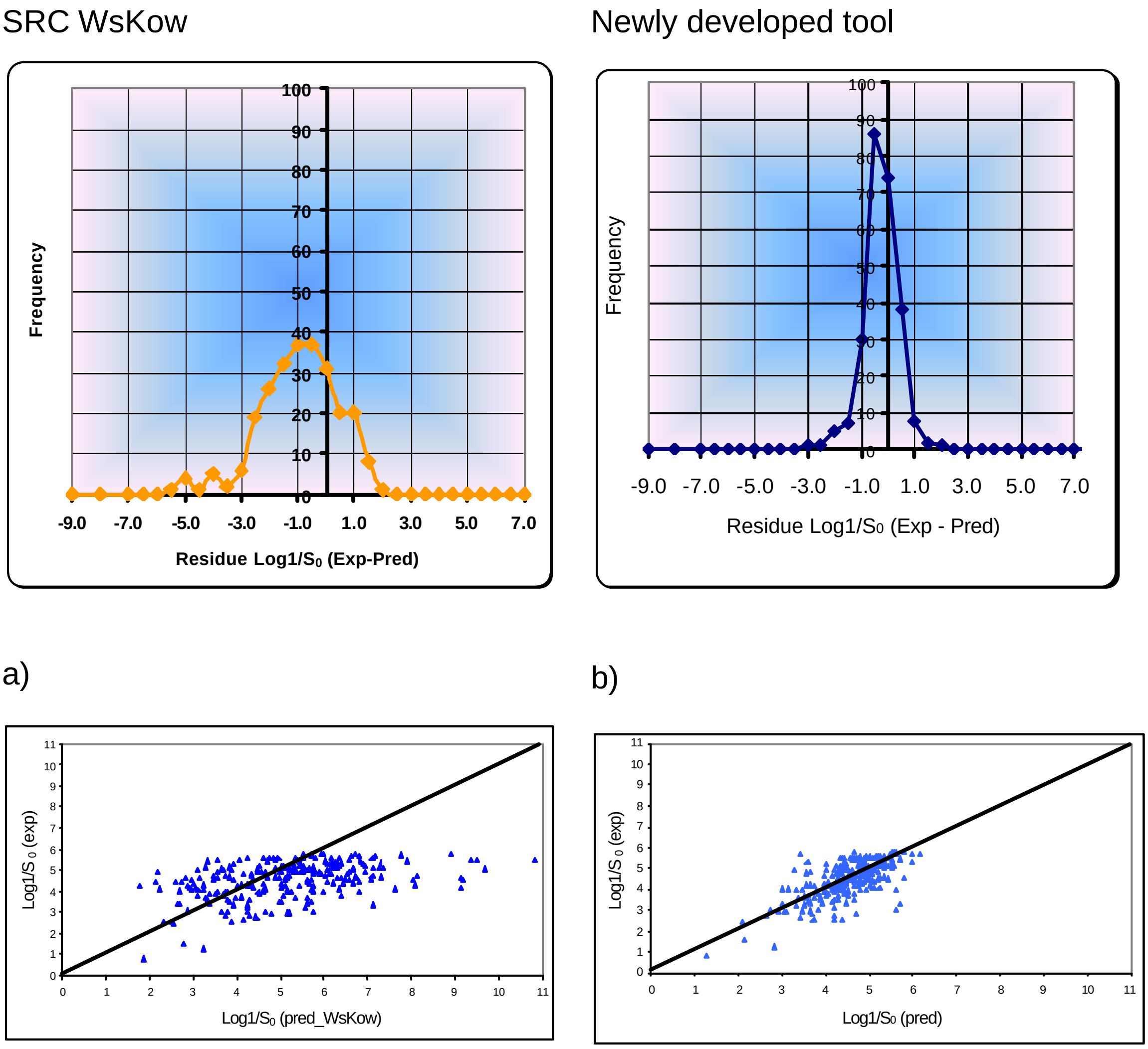
Figure 41: The final solubility model generated with 1515 compounds (blue) in the training data set and tested with 958 (red) in the test data set.

Figure 41 shows the solubility of most compounds is predicted within an error of one log unit. The standard error of the predicted solubilities is 0.42 log units. The correctness of the solubility classification are listed in Table 23. More than 50% compounds in each solubility range were correctly classified.

| Solubility | S (µg/mL) | $Nr_{compounds}$ | %correct classification |
|---|---|---|---|
| low | S <= 10 | 944 | 66% |
| medium | 10 < S <= 100 | 992 | 69% |
| high | S > 100 | 537 | 51% |

Table 23: The correctness of solubility classification for 2473 compounds with the model described in Eq. 22.

253 neutral drugs used in the chapter 4.1.1.4 to test the program SRC WsKow[91] were also used to test the newly developed prediction tool.



a)

b)

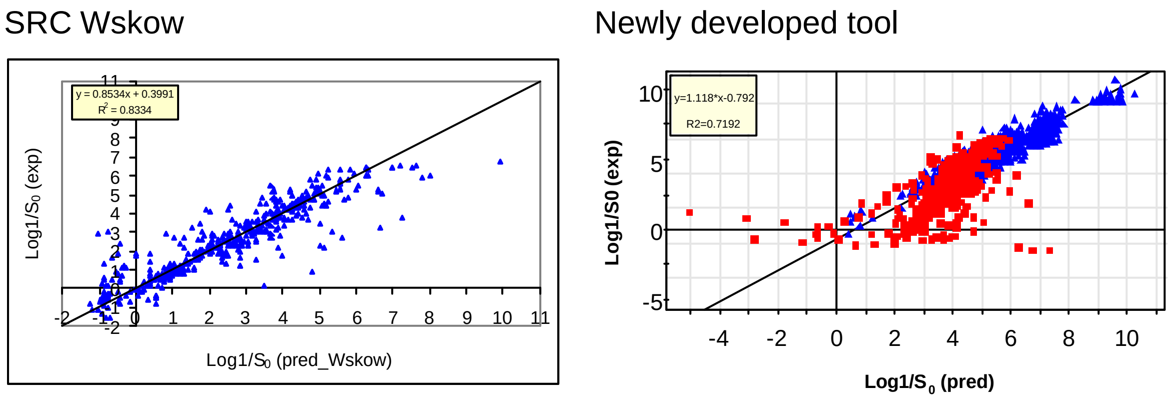
c)

d)

Figure 42: Solubility prediction of 253 neutral drugs. a) Residue diagram for prediction with SRC WsKow[91]; b) residue diagram for prediction with the newly developed tool; c) experimental versus predicted solubility with SRC WsKow[91]; d) experimental versus predicted solubility with the newly developed tool.

In comparison to the program SRC WsKow[91], the residues of most compounds predicted with the newly developed tool were much lower. Most lay within 1 log unit and no single residual was higher than 2 log units. (Figure 42a and b) Additionally, the correlation between predicted and experimental solubility was much better when using the newly developed tool compares to the program SRC WsKow[91]. (Figure 42c and d)

467 organic compounds, for which the fragments were present in the newly developed fragmental database were selected from AQUASOL[67] database and used as test data set for the solubility prediction.

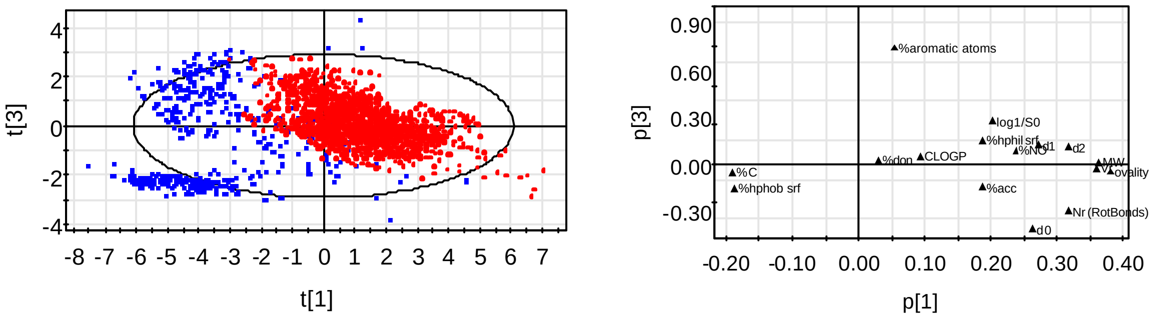


a) b)

Figure 43: Solubility prediction of 467 organic compounds in AQUASOL database. a) Solubility prediction with SRC WsKow[91]; b) solubility prediction with the newly developed tool. The blue colored compounds are 1515 drugs used as training data set for the prediction tool. The red colored are 467 organic compounds used as test data set.

In comparison to the program SRC WsKow[91] (Figure 43a), the solubility of these 467 compounds were not well predicted with the newly developed prediction tool, resp. two separate data sets could be observed in the Figure 43b.

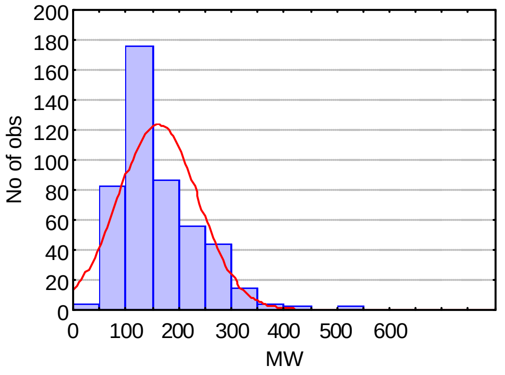
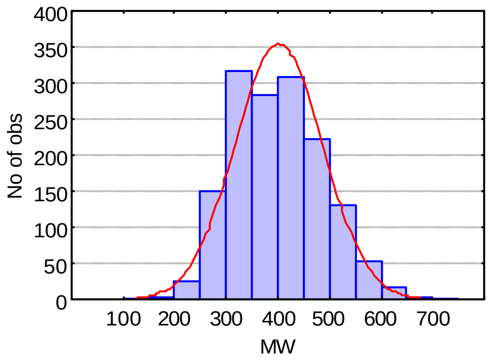
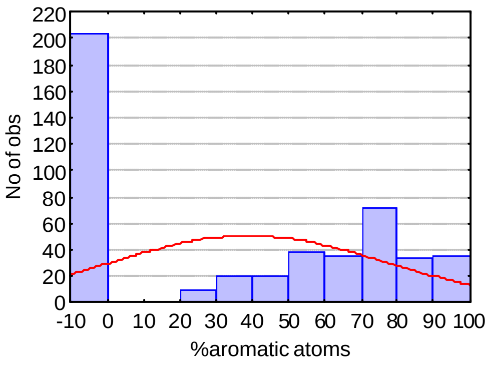
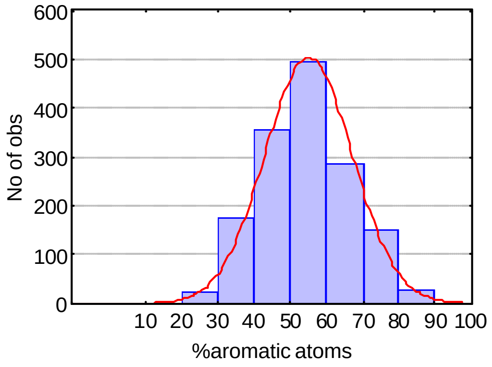
In order to explain the differences between drug-like compounds and 467 organic compounds from the AQUASOL[67] database, PCA analysis was performed for the related data sets. Five components were calculated for PCA and its first three component (t1-t3) described 78.1% of the x-space.



a) b)

Figure 44: PCA analysis for 467 organic compounds in the AQUASOL[67] database (blue) and 1515 drug-like compounds (red). a) PCA score plot; b) PCA loading plot.

The 467 organic compounds in AQUASOL[67] database and 1515 drug-like compounds were detected in PCA analysis as two separate clusters. (Figure 44a) The dominating descriptors for this separation were %aromatic atoms, %C and molecular weight. (Figure 44b) A comparison of the descriptors is shown in Figure 45.

| | 467 organic compounds | 1515 drug-like compounds |
|---|---|---|
| MW |  |  |
| %aromatic atoms |  |  |

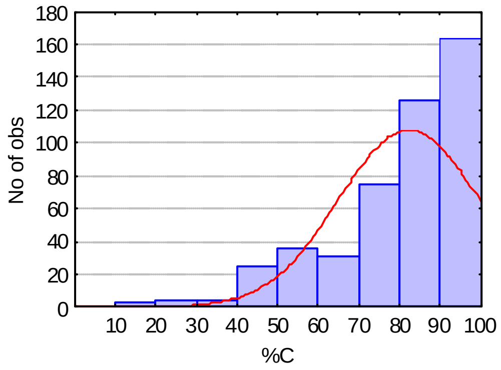
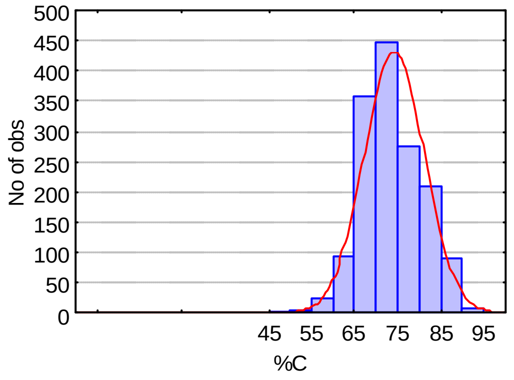
| | | |
|---|---|---|
| %C |  |  |

Figure 45: Comparing the properties of 467 organic and 1515 drug-like compounds.

Figure 45 shows drug-like compounds contain mostly aromatic atoms and have higher molecular weight. Additionally, a more compact crystal packing can always be observed for drug-like compounds, because of the intermolecular hydrogen bonding. In contrast to the drug-like compounds, 35% of 467 organic compounds used here contain only aliphatic carbons and such molecules are held together in the crystal state through the van der Waals interactions. Therefore, the contribution of an aliphatic carbon atom to the solubility is different for simple organic compounds, in comparison to drug-like molecules. Hence, the application of the newly generated solubility prediction model is directed to the prediction of solubility of drug-like compounds.

As mentioned in chapter 4.1.1.5, among the 1770 compounds in the AQUASOL[67] database, there are 206 drug-like, resp. 11.6%, which are not part of the third data set. The model in Figure 46 was generated for 206 drug-like compounds in the AQUASOL[67] database together with 2473 reference compounds by including the definition of 58 new fragmental constants for the 206 compounds from the AQUASOL[67] database.

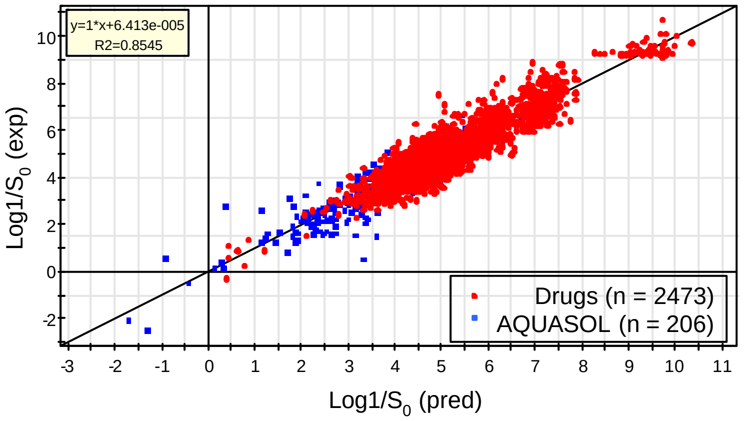


Figure 46: Solubility prediction for 206 drug-like compounds in AQUASOL[67] database and 2473 reference compounds.

Figure 46 shows that the solubilities of 206 drug-like compounds from the AQUASOL[67] database are much higher than the reference compounds. Nevertheless, the correlation between the predicted and experimental value (Figure 46) shows the arrangement of these 206 AQUASOL[67] drug-like compounds is in line with the red colored reference compounds.

### 4.3.4.1 Solubility prediction for external data described in the literature

Several literature studies[88,92-95] on physicochemical characterization of drug-like congeneric series were found. The described solubility values were used as external validation data sets to test the predictive power of the newly developed solubility prediction tool.

The derived fragment related coefficients were applied to predict the solubility of external congeneric series. Two methods were evaluated to derive solubilities for compounds with similar structures:

1. The experimental solubility of a compound in the congeneric series was taken as a starting point. The scaffold solubility value of this compound was calculated by subtracting the solubility values of substituents from the compound's experimental solubility value. The required solubility prediction

value of any other compound was a result of the calculated value of the scaffold and the solubility values of substituents derived from the generated solubility model.

2. The scaffold solubility value would be first calculated for each compound in the congeneric series, by subtracting the solubility values of substituents from the compound's experimental solubility value. Finally, the mean value of the scaffold was assigned to all compounds. The predicted solubility value of each compound resulted then from the mean value of the scaffold and its substituents' fragmental solubility values.

The predictive power of the newly developed solubility prediction tool was directly compared with the commercially available tool, WaterFrag[96] developed by Meylan.

#### 4.3.4.1.1 Validation with the external data set 1

Goosen[88] measured the solubility of thalidomide and its N-alkyl analogues in water at pH = 6.4 and 25°C. Because the fragmental value of AC(=O)[NH]C(=O)A in thalidomide was not available, N-methyl-thalidomide was taken as the starting point for the solubility calculation following method 1.

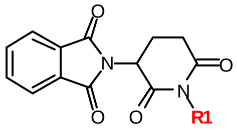


Figure 47: The scaffold of thalidomide derivates.

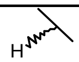
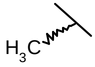
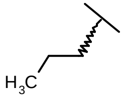
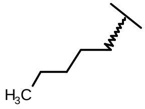
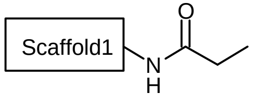
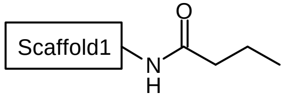
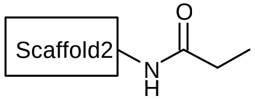
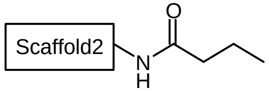
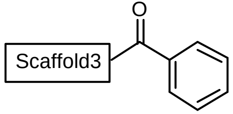
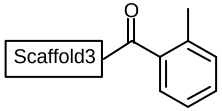
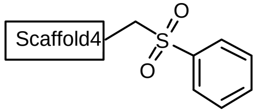
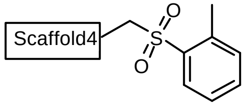
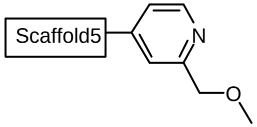
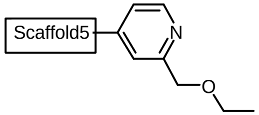
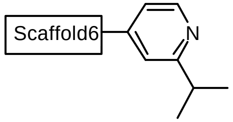
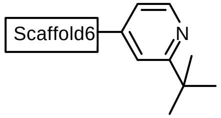
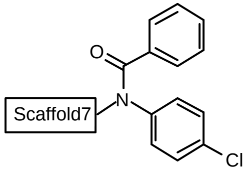
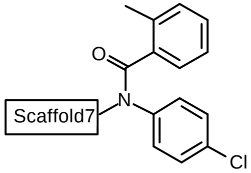
| Name | Substituent R1 | $S_{exp}$ (µg/mL) | MW | Log1/$S_{0(exp)}$ | Log1/$S_{0(pred)}$ | Log1/$S_{0(pred\ WaterFrag)}$ |
|---|---|---|---|---|---|---|
| Thalidomide |  | 52.1 | 258 | 3.69 |  | 1.34 |
| N-Methyl thalidomide* |  | 275.9 | 272 | 2.99 | 2.99 | 1.83 |
| N-Propyl thalidomide |  | 57.3 | 300 | 3.72 | 3.08 | 2.9 |
| N-Pentyl thalidomide |  | 6.54 | 328 | 4.7 | 3.16 | 4 |

Table 24: The solubility data of thalidomide and its N-alkyl analogues[88]. *: Compound was taken as starting point for the prediction with the newly developed solubility tool.

Table 24 shows the solubility decrement caused by lengthening the chain length is correctly predicted using both programs, resp. the newly developed tool and the program WaterFrag[96] developed by Meylan. However, the decrement caused by one aliphatic carbon was correctly predicted by WaterFrag[96] ($\Delta$log1/$S_{0(CH3)}$ = 0.5 log units), but not by the newly developed tool ($\Delta$log1/$S_{0(CH3)}$ = 0.05 log units). Databases containing drug-like compounds were scanned to find similar examples. Such examples should be compounds with experimental solubilities and structures containing the fragment *C(=O)N(*)C(=O)* with corresponding modification on the N-alkyl chains. However, the glutarimide ring with long N-alkyl chain is considered as instable. Therefore, it is difficult to evaluate the quality of the prediction by only using these four compounds of Goosen[88]. Additionally, it is known, that the program WaterFrag[96] is based on the same fragmental database as the program Kowwin[97] and the increment of lipophilicity caused by a methylene group has a value of 0.5 log units. Thus, it can be assumed that the solubility coefficient of the methylene group in the WaterFrag[96] is derived from its lipophilicity value. Furthermore, the shift of solubility caused by one aliphatic carbon is not always about 0.5 log units. It can vary between 0.01 and 1.09 log units, which depends on the structural environment of the methylene group, as collected in Table 25.

| Name | Structure | Log1/$S_0$ | Name | Structure | Log1/$S_0$ | Shift of log1/$S_0$ caused by one methylene group |
|---|---|---|---|---|---|---|
| compound 15 |  | 4.56 | compound 16 |  | 4.68 | 0.12 |
| compound 17 |  | 4.34 | compound 18 |  | 4.48 | 0.14 |
| compound 19 |  | 5.55 | compound 20 |  | 5.09 | -0.46 |
| compound 21 |  | 5.64 | compound 22 |  | 4.95 | -0.69 |
| compound 23 |  | 4.66 | compound 24 |  | 4.93 | 0.27 |
| compound 25 |  | 4.77 | compound 26 |  | 5.03 | 0.26 |
| compound 27 |  | 5.53 | compound 28 |  | 5.68 | 0.15 |

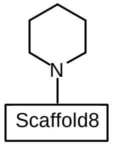
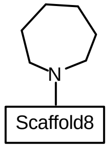
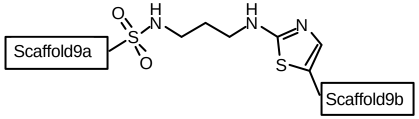
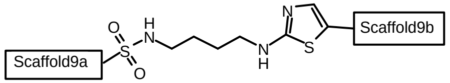
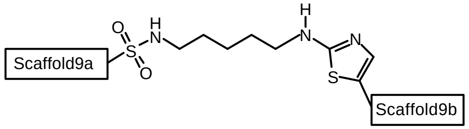
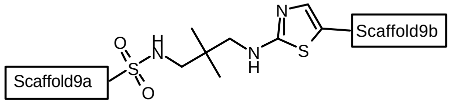
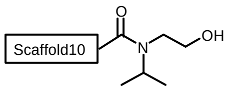
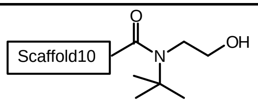
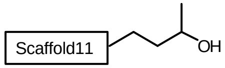
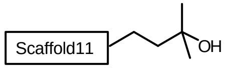
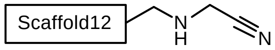
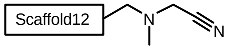
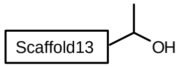
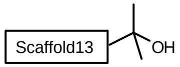
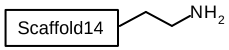
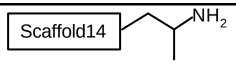
| | | | | | | |
|---|---|---|---|---|---|---|
| compound 29 |  | 7.87 | compound 30 |  | 7.61 | -0.26 |
| compound 31 |  | 4.56 | compound 32 |  | 5.65 | 1.09 |
| compound 33 |  | 5.36 | compound 34 |  | 5.66 | 0.3 |
| compound 35 |  | 3.44 | compound 36 |  | 3.53 | 0.09 |
| compound 37 |  | 4.51 | compound 38 |  | 4.16 | -0.35 |
| compound 39 |  | 3.90 | compound 40 |  | 4.78 | 0.88 |
| compound 41 |  | 4.42 | compound 42 |  | 4.78 | 0.36 |
| compound 43 |  | 7.04 | compound 44 |  | 6.31 | -0.37 |

Table 25: Example for the solubility shift caused by adding a methylene group to compounds with the same scaffold. The compounds listed in the same raw have the same scaffold, which is signified with the number of the scaffold.

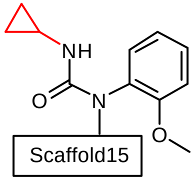
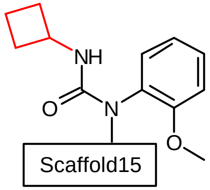
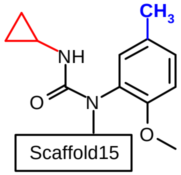
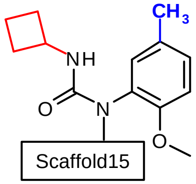
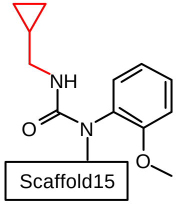
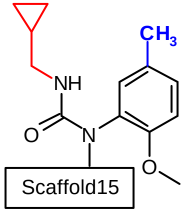
| Name | Structure | Log1/$S_0$ | Name | Structure | Log1/$S_0$ |
|---|---|---|---|---|---|
| compound 45 |  | 4.88 | compound 46 |  | 5.29 |
| compound 47 |  | 5.11 | compound 48 |  | 5.12 |
| compound 49 |  | 5.28 | | | |
| compound 50 |  | 4.76 | | | |

Table 26: Collection of six compounds with the same scaffold and comparison of changes in their solubility caused by adding a methylene group on two different substituent positions.

Table 25 shows, that adding a methylene group can increase the solubility in some cases. Such phenomenon happens frequently, when the crystal packing is changed. (Chapter 4.3.2) Additionally, Table 26 shows that it is not easy to find a general rule for describing the influence of a methylene group on the solubility, because the solubility shift can be different, even when the addition of the methylene group occurs at the same position.

#### 4.3.4.1.2 Validation with the external data set 2

Bavetsias[92] measured the solubility of CB30865 analogues at pH = 7.4 in 10 mM potassium dihydrogen phosphate containing 150 mM sodium chloride.

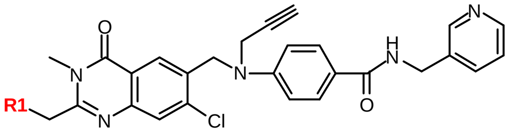


Figure 48: The scaffold of CB30865 analogues.

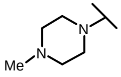
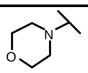
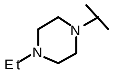
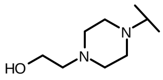
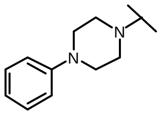
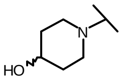
| Nr | Substituent | $pK_a$ | | MW | S (µM) | Log1/$S_0$ | Log1/$S_{0\ (pred)}$ | Log1/$S_{0\ (WaterFrag)}$ |
|---|---|---|---|---|---|---|---|---|
| | | $pK_a1$ | $pK_a2$ | | | | | |
| 1* |  | 4.65 | 7.86 | 584 | 146 | 4.49 | 4.49 | 2.99 |
| 2 |  | 4.65 | | 571 | 2 | 5.79 | 4.27 | 3.36 |
| 3 |  | 4.65 | 7.16 | 598 | 286 | 3.77 | 4.53 | 3.53 |
| 4 |  | 4.65 | 7.16 | 614 | 75 | 4.36 | 3.95 | 1.76 |
| 5 |  | 4.65 | | 646 | 5 | 5.39 | 5.10 | 6.85 |
| 6 |  | 4.65 | | 585 | 0.5 | 6.39 | 4.10 | 3.18 |

Table 27: The solubility of CB30865 analogues[92]. *: The compound was taken as starting point for the prediction with the newly developed solubility tool.

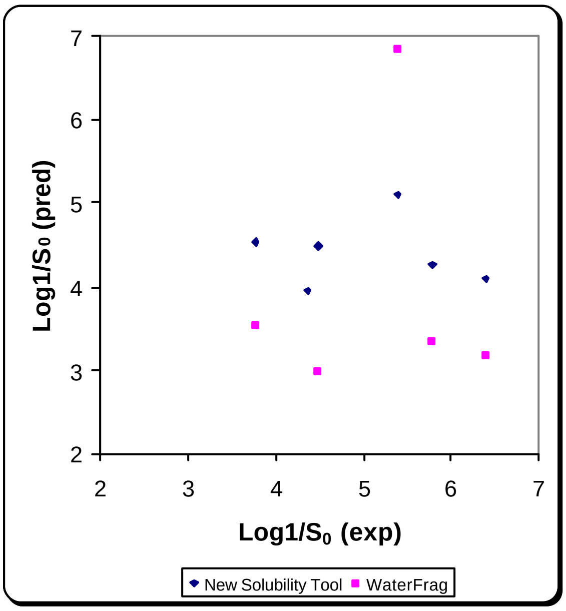


Figure 49: Comparison of solubility of CB30865 analogues[92] predicted with the new solubility tool and WaterFrag.

$pK_a$ calculated by ACD[78] was used to consider the pH dependence of solubility. For the 6 compounds in Table 27, the solubility prediction results achieved by the newly developed tool were much better in comparison to WaterFrag[96].

#### 4.3.4.1.3 Validation with the external data set 3

Edwards[94] measured the solubility of pyridopyrimidine trifluoromethyl ketones in 0.01 M sodium phosphate buffer at pH = 7.4.

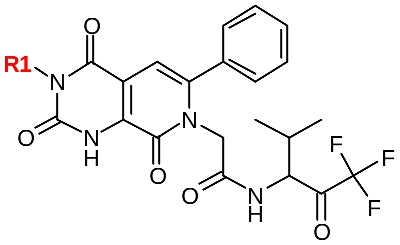


Figure 50: The scaffold of pyridopyrimidine trifluoromethyl ketones[94].

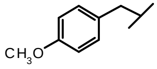
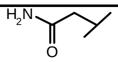
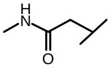
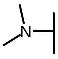
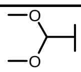
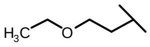
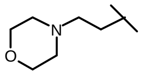
| Nr | Substituent | $pK_a$ | MW | S (mg/mL) | $Log1/S_0$ | $Log1/S_{0\ (pred)}$ | $Log1/S_{0\ (pred\ WaterFrag)}$ |
|---|---|---|---|---|---|---|---|
| 1 | H | | 464 | 0.22 | 3.32 | | 2.82 |
| 2* | $CH_3$ | | 478 | 0.044 | 4.03 | 4.03 | 3.31 |
| 3 |  | | 584 | 0.23 | 3.40 | 4.35 | 6.08 |
| 4 |  | | 521 | 0.13 | 3.60 | 3.65 | 1.81 |
| 5 |  | | 535 | 0.1 | 3.73 | 3.34 | 1.96 |
| 6 |  | | 507 | 0.32 | 3.20 | | 1.67 |
| 7 |  | | 538 | 0.42 | 3.11 | 4.27 | 2.36 |
| 8 |  | | 536 | 0.008 | 4.83 | 3.80 | 3.65 |
| 9 |  | 7.04 | 577 | 0.30 | 3.44 | 3.74 | 2.15 |

Table 28: The solubility of pyridopyrimidine trifluoromethyl ketones[94]. *: The compound was taken as starting point for the prediction with the newly developed solubility tool.

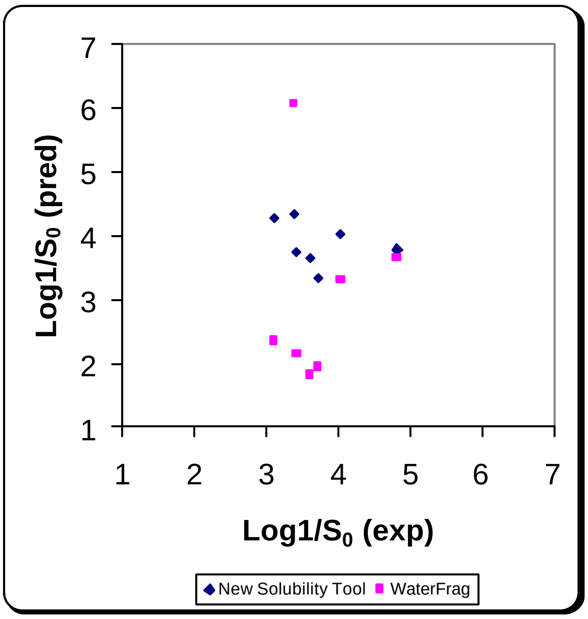


Figure 51: Comparison of solubility of pyridopyrimidine trifluoromethyl ketones[94] predicted with the new solubility tool and WaterFrag.

Because of the lack of the fragments, the solubility of compounds 1 and 6 could not be predicted. The solubility prediction results for the remaining 7 compounds using the newly developed tool was much better in comparison to WaterFrag[96].

#### 4.3.4.1.4 Validation with the external data set 4

Bernstein[95] measured the solubility of 3-amino-6-phenylpyridin-2-one trifluroromethyl ketones in 0.01 M sodium phosphate buffer at pH = 7.4.

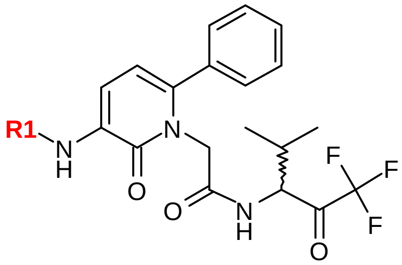
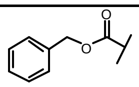
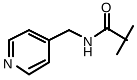
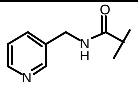
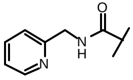
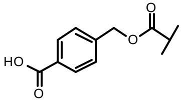
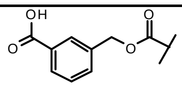
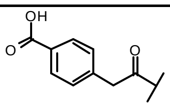
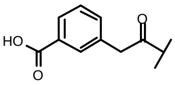


Figure 52: The scaffold of 3-amino-6-phenylpyridin-2-one trifluoromethyl ketones[95].

| Nr | Substituent | pK$_a$ | MW | LogP[95] (exp) | S (µg/mL) | Log1/S$_0$ | Log1/S$_0$ (pred | Log1/S$_0$ (pred WaterFrag) |
|---|---|---|---|---|---|---|---|---|
| 1* |  |  | 529 |  | 1.8 | 5.47 | 5.47 | 4.22 |
| 2 |  | 5.46 | 529 | 2.16 | 140 | 3.58 | 4.53 | 1.34 |
| 3 |  | 5.03 | 529 | 2.84 | 430 | 3.09 | 4.53 | 1.34 |
| 4 |  | 4.88 | 529 | 2.15 | 300 | 3.25 | 4.53 | 1.34 |
| 5 |  | 4.1 | 573 | 0.84 | 2600 | 5.64 | 6.53 | 4.37 |
| 6 |  | 4.09 | 573 |  | 1570 | 5.87 | 6.53 | 4.37 |
| 7 |  | 4.17 | 557 | 0.35 | 2500 | 5.58 | 5.30 | 3.53 |
| 8 |  | 4.13 | 557 |  | 900 | 6.06 | 5.30 | 3.53 |

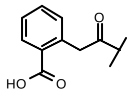
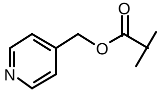
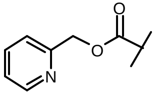
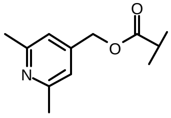
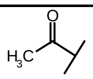
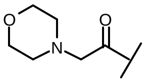
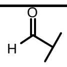
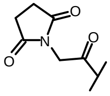
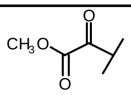
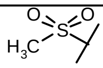
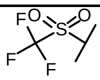
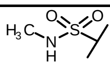
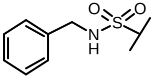
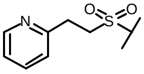
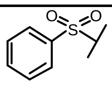
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 |  | 3.85 | 557 | 1.14 | 2500 | 5.90 | 5.30 | 2.87 |
| 10 |  | 4.65 | 530 | 2.41 | 16 | 4.52 | 5.30 | 1.96 |
| 11 |  | 4.65 | 530 | 1.91 | 2.1 | 5.42 | 5.30 | 1.96 |
| 12 |  | 4.65 | 558 | 2.38 | 8.5 | 4.82 | 5.39 | 3.01 |
| 13 |  | | 437 | 1.78 | 92 | 3.68 | 3.94 | 0.95 |
| 14 |  | 5.43 | 522 | 2.00 | 23 | 4.36 | 3.60 | -0.28 |
| 15 |  | | 423 | 1.84 | 220 | 3.28 | 4.18 | 0.38 |
| 16 |  | | 534 | 0.96 | 3100 | 2.24 | | 0.34 |
| 17 |  | | 481 | | 840 | 2.76 | 4.29 | 0.37 |
| 18 | H | 3.49 | 395 | 1.74 | 490 | 2.91 | 3.78 | 0.55 |
| 19 | $CH_3CH_2$ | 3.86 | 423 | 2.37 | 280 | 3.18 | 4.04 | 1.24 |
| 20 |  | 8 | 473 | 1.77 | 4100 | 2.16 | 1.72 | 0.67 |
| 21 |  | 8 | 527 | 0.94 | 940 | 2.85 | 1.46 | 1.87 |
| 22 |  | 8 | 488 | 1.79 | 1730 | 2.45 | | 0.62 |
| 23 |  | 8 | 564 | 2.57 | 21 | 4.42 | | 3.06 |
| 24 |  | 5 | 564 | | 100 | 3.85 | 1.95 | 1.38 |
| 25 |  | 8 | 535 | 2.53 | 180 | 3.57 | 4.20 | 3.80 |

Table 29: The solubility of 3-amino-6-phenylpyridin-2-one trifluoromethyl ketones[95]. *: The compound was taken as starting point for the prediction with the newly developed solubility tool.

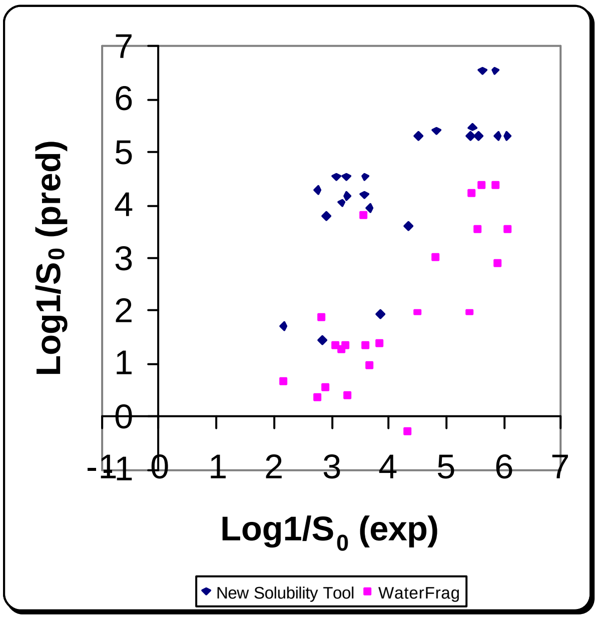


Figure 53: Comparison of solubility of 3-amino-6-phenylpyridin-2-one trifluoromethyl ketones[95] predicted with the new solubility tool and WaterFrag.

Because of the lack of the fragments, the solubility of compounds 16, 22 and 23 could not be predicted. The solubility prediction results for the rest of 22 compounds were much better with the newly developed tool than with WaterFrag[96]. Table 29 shows the solubility difference caused by diverse substituent positions on the aromatic ring, resp. ortho, meta and para, can be 0.5 log units. Unfortunately, such position caused solubility difference can not be correctly predicted in the current form of the newly developed tool or program WaterFrag[96].

### 4.3.4.2 Structure based solubility rules

The contribution of fragments to solubility were derived as a result of the weighting of the structural fragments used in the new solubility prediction model. Thus, 460 structure based solubility rules were derived and listed in the appendix. Furthermore, molecular properties important for the solubility enhancement can be identified by inspection of the structure based solubility rules. A small section from the appendix is taken here as example to visualize the influence of small structural changes on intrinsic solubility as shown in Figure 54.

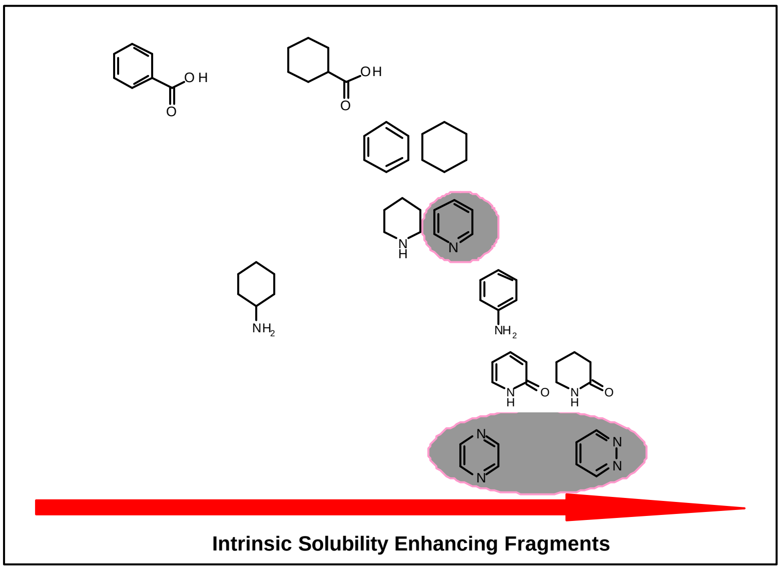
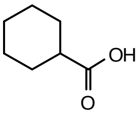
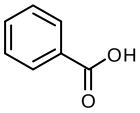
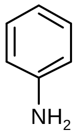
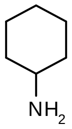


Figure 54: Examples for intrinsic solubility enhancing fragments.

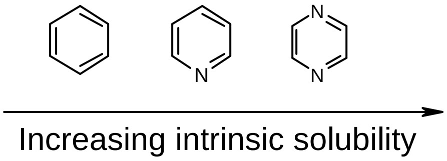
In case both hexagonal rings have similar $pK_a$ values, the compounds with aliphatic fragments are more soluble than those with aromatic ones, e.g.

 has higher intrinsic solubility than 

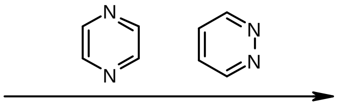
Furthermore, strong basic and acidic fragments provide lower intrinsic solubilities than similar neutral ones, because of the formation of intermolecular hydrogen bonds, e.g.

 has higher intrinsic solubility than 

Fragments with high polar surface are solubility enhancing, e.g.



Moreover, fragments with high dipole moment are more soluble than fragments with low dipole moment. e.g.



Increasing intrinsic solubility

In conclusion, property-based solubility rules were deduced by comparison of the influence of different rings on the solubility. They partly reflect the already existing knowledge in that field as to summarized:

1 Compounds containing aliphatic fragments are more soluble than aromatic ones.

2 Dipole moment enhances solubility.

3 Compounds containing polar fragments are more soluble than non polar ones.

4 Compounds containing strong basic and acid fragments have lower intrinsic solubility than neutral ones.

In contrast to the property based solubility rules, structure based solubility rules can be more conveniently used by medicinal chemist as a guideline to improve the structures of leads to achieve higher solubility. Hence, a more diverse data set should be collected in the near future to optimize the developed solubility model and to extend the structural based rules by addition of further structural fragments.

### 4.3.5 The impact of solid state on solubility

Aqueous solubility of a compound is governed by three major factors[98]:

- intermolecular interactions in the crystal lattice
- the difference between the solute-water adhesive interaction and the sum of the solute-solute and water-water interactions
- the entropy of mixing (solute/solvent)

In order to study the impact of solid state on solubility, two data sets containing compounds with measured 3D crystal structures and melting points were collected. (chapter 4.2.1.1)

Melting point is considered as an important parameter for assessing the cohesive energy of solid state. The relationship between melting point and the formation of intermolecular hydrogen bonds was confirmed by the descriptors used in the prediction of melting point for the second data set. The number of hydrogen donors was identified as the most important descriptor for the prediction of melting point. Furthermore, the influence of solid state properties on solubility was confirmed by solubility models for the first and second data set, because the number of hydrogen donors, melting point and lipophilicity belonged to the most important parameters. (Eq. 20 and Eq. 21) The VIP plots in the related solubility studies identified lipophilicity as a more important descriptor in the prediction of solubility than the number of hydrogen bonds or melting points. This leads to the conclusion that the solubility of drug-like compounds depends more on the solvation process than the cohesive energy in solid state.

In order to evaluate the extent of the influence of solid state on solubility, polymorphs were evaluated. Differences in the solubilities of polymorphic forms can be assumed to be only dependent on differences in the crystal packing.

Pudipeddi[27] collected a solubility data set of 72 compounds with diverse polymorphs. 2 to 3 fold differences in solubilities were observed for most of the collected cases. Larger differences were described for premafloxacin[28] (~30 fold), codeine[29] (~13 fold) and cyclopenthiazide[30] (~4 fold). According to those results, it can be assumed that differences in solubilities which are based on different crystal packings and resulting in polymorphs are usually low and are often in the range of the experimental error of the solubility measurements.

Polymorphs can be divided into two categories, resp. enantiotropic and monotropic. The differences between both categories are described in Table 30.

| Enantiotropic | Monotropic |
|---|---|
| Transition temperature < melting temperature of I | Transition temperature > melting temperature of I |
| I is stable above transition temperature; II is stable below transition temperature | I always stable |
| Transition reversible | Transition irreversible |
| Solubility of I higher than II below transition temperature; solubility of II higher above transition temperature | Solubility of I always lower than II |
| Transition from II to I endothermic<br>$\Delta H^{I}_{f} < \Delta H^{II}_{f}$ | Transition from II to I exothermic<br>$\Delta H^{I}_{f} > \Delta H^{II}_{f}$ |
| Density I < density II | Density I > density II |

Table 30: Thermodynamic rules for enantiotopic and monotropic phase transitions[99]. I is the higher melting form.

Enantiotropic polymorphs can be interconverted below the melting point of each polymorph, because different enantiotopic forms are stable under different conditions, while monotropic polymorphs behave differently. Thus, for a monotropic polymorphic pair, only one thermodynamically stable form under all attainable conditions does exist. However, the unstable form of a monotropic polymorphic pair can still be useful, because the activation energy for the conversion to the stable form is high, and under this situation, the meta stable compound can be formed. Table 31 shows characteristic properties of known polymorphs collected from different literature sources[30,100-106].

| Name | Polymorphic forms | S (µg/mL) | Solution | Melting point (°C) | Transformation by heating | Comment |
|---|---|---|---|---|---|---|
| Cyclopenthiazide[30] | I | 34.7 | Water | 238 |  | Forms I and II showed only a single melting point at 238 and 225°C. Form III melts at 181°C and then recrystallizes to form I. |
|  | II | 61.80 |  | 225 |  |  |
|  | III | 17.15 |  |  | $\mathrm{III} \xrightarrow{181°C} \mathrm{I}$ (238°C) |  |
| Premafloxacin[100] | I | 3230 | Ethyl acetate |  | $\mathrm{I} \xrightarrow[\text{Endotherm}]{140\text{-}150°C} \mathrm{II} \xrightarrow[\text{Exotherm}]{165\text{-}180°C} \mathrm{III}$ |  |
|  | II |  |  |  |  |  |
|  | III | 140 |  | 198-202 |  |  |
| MK571[101] | I | 1240 | Methyl ethyl ketone | 164 |  | No conversion is observed between form I and II. |
|  | II | 2400 |  | 152 |  |  |
| Auranofin[102] | I | 600 | 25% polyethylene glycol 200 |  | $\mathrm{I} \xrightarrow{385°C} \mathrm{II}$ (389°C) |  |
|  | II | 1300 |  | 389 |  |  |
| Seratrodast[103] | I | 543 | 50 mM phosphate buffer at pH=8 |  | $\mathrm{I} \xrightarrow{83.4°C} \mathrm{II}$ |  |
|  | II | 817 |  |  |  |  |
| Acetazolamide[103] | A | 2040 | 50 mM phosphate buffer at pH=8 |  | $\mathrm{A} \xrightarrow{78.4°C} \mathrm{B}$ |  |
|  | B | 2280 |  |  |  |  |
| Carbamazepine[103] | I | 11560 | 50 mM phosphate buffer at pH=8 |  | $\mathrm{III} \xrightarrow{73°C} \mathrm{I}$ |  |
|  | III | 9680 |  |  |  |  |
| Indomethacin[103] | a | 576 | 50 mM phosphate buffer at pH=8 | 157 |  | No conversion is observed between form a and β. |
|  | β | 432 |  | 163 |  |  |
| Mefenamic acid[104] | I | 6090 | 50 mM phosphate buffer at pH=8 |  | $\mathrm{I} \xrightarrow{89°C} \mathrm{II}$ |  |
|  | II | 7930 |  |  |  |  |
| Sulfathiazole[104] | I | 677 | 50 mM phosphate buffer at pH=8 |  | $\mathrm{I} \xrightarrow{112.6°C} \mathrm{II}$ |  |
|  | II | 1118 |  |  |  |  |

| | | | | | | |
|---|---|---|---|---|---|---|
| Proscar[105] | I | 160 | Water at 25 °C | | $\mathrm{II} \xrightarrow{165^{\circ}\mathrm{C}} \mathrm{I}$ | |
| | II | 59 | | | | |
| Sulfuno[106] | I | 205-211 | Water at pH = 3.8 T = 20 °C | 91.7 mg% | $\mathrm{II} \xrightarrow{188\text{-}195^{\circ}\mathrm{C}} \mathrm{I}$ | |
| | II | | | 84.6 mg% | | |
| Tromexan[106] | I | 172-182 | Water at pH = 3.8 T = 20 °C | 8.9 mg% | | No automatic conversion is observed between form I and form II |
| | II | 153-160 | | 15.3 mg% | | |

Table 31: Polymorphs with their corresponding transition temperatures. The enantiotropic polymorphs are the I and III forms of cyclopenthiazide, the I and II forms of premafloxacin, the I and II forms of auranofin, the I and II forms of seratrodast, the A and B forms of acetazolamide, the III and I forms of carbamazepine, the I and II forms of mefenamic acid, the I and II forms of sulfathiazole, the I and II forms of proscar, the I and II forms of sulfuno. The monotropic polymorphs are the I and II forms of cyclopenthiazide, the II and III forms of premafloxacin, the I and II forms of MK571, the a and β forms of indomethacin and the I and II forms of tromexan.

Most polymorphs in Table 31 are enantiotropic, which crystallize according to the empirical Ostwald's law in stages[107]. Take cyclopenthiazide as an example. The form I and III of cyclopenthiazide are enantiotropic. By heating, form III transforms to form I before it melts. Thus, 181°C is considered as the conversion temperature of form III, which is lower than the melting point of form I (238°C). Usually, the higher the melting point, the lower the solubility. However, form I has a higher solubility value than form III, although its melting point is high. The higher solubility value of form I can be explained by measuring the reaction energy required by the transformation. From form III to form I, heating is needed for the conversion. Therefore, form III is the most stable and the least soluble form at room temperature. However, the melting point of form III is probably so high, that is not measurable. Therefore, in case of enantiotopic forms, no direct comparison of melting point and solubility can be performed for these two related polymorphic forms.

The I and II form of cyclopenthiazide are monotropic. Both of them occupy a single melting point at 238 and 225°C. Thus, no automatic conversion can be expected between these two related forms. Furthermore, relationship is observed between melting points and solubilities of these two monotropic forms. The most stable form, resp. form I has a higher melting point and lower solubility than the less stable form, resp. form II.

Beside the diazepam derivates described in the chapter 4.2.1.2.3.1 and the sulfadiazine in the chapter 4.3.2, the halogen analogues of deoxyuridine derivates in the first data set can be used as an additional example for studying the influence of solid state on solubility.

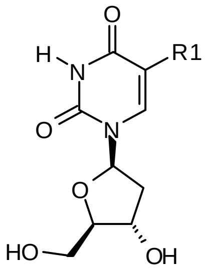
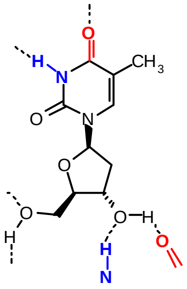
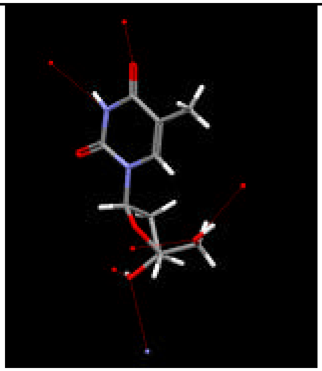
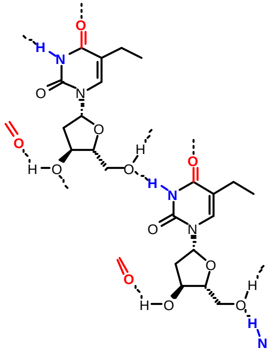
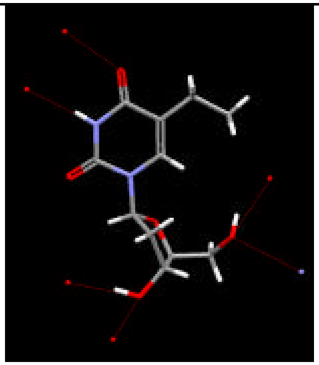
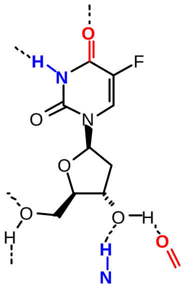
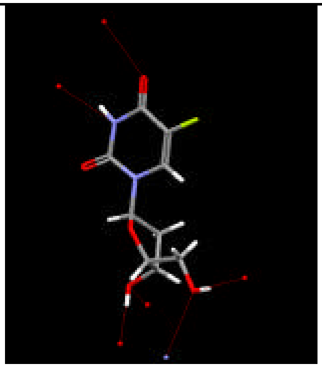
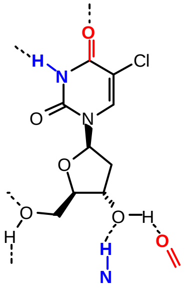
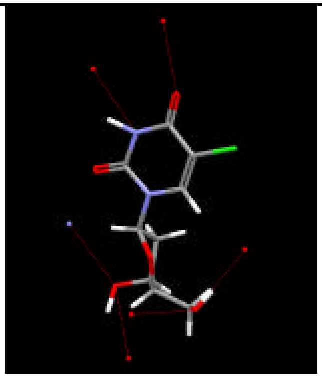
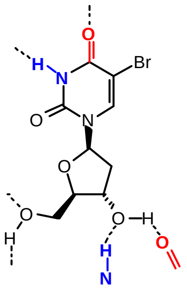
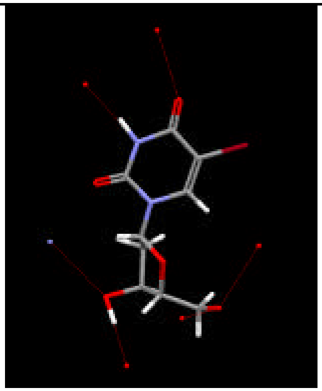


Figure 55: The scaffold of deoxyuridine derivates.

| R1 | Name | 2D Structure | 3D Crystal Structure | MW | ClogP | LogP[89] (exp) | MP (°C) | pK$_a$ | S (µg/mL) | pH | Log1/S$_0$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 2'-deoxyuridine |  |  | 228 | -1.884 | -1.467 | 165-167 | 9.16 | 1695276 | 6.5 | -0.87 |
| CH3 | thymidine |  |  | 242 | -1.385 | -1.177 | 186-188 | 9.55 | 64298 | 6.5 | 0.58 |
| C2H5 | 5-ethyl-2'-deoxyuridine |  |  | 256 | -0.856 | -0.646 | 152-153 | 9.57 | 71750 | 6.6 | 0.55 |
| CF3 | a,a,a-trifluorothymidine |  |  | 296 | -0.413 | 0.009 | 178-180 | 7.5 | 39128 | 6.3 | 0.91 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| F | 5-fluoro-2'-deoxyuridine |  |  | 246 | -1.405 | -1.2 | 148-150 | 7.42 | 502278 | 5.5 | -0.3 |
| Cl | 5-chloro-2'-deoxyuridine |  |  | 262 | -0.835 | -0.937 | 176-177 | 7.74 | 58862 | 6.3 | 0.66 |
| Br | 5-bromo-2'-deoxyuridine |  |  | 307 | -0.685 | -0.572 | 191-194 | 7.78 | 14752 | 6.5 | 1.34 |

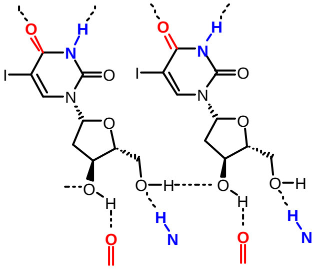
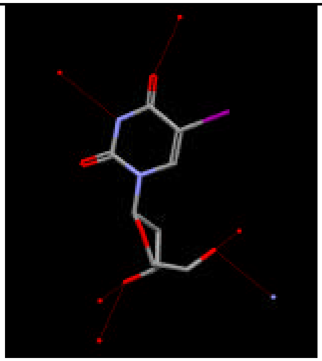
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| I | 5-iodo-2'-deoxyuridine |  |  | 354 | -0.425 | -0.282 | 155-180 | 8.09 | 1900 | 6.5 | 2.28 |

Table 32: The influence of halogen atoms and crystal lattice on solubility for selected compounds from data set 1.

Table 32 shows that the melting points of the halogen analogues increase with ascending atomic weights, until iodine atom is added to the scaffold. The observed reduction of melting point by substituting the bromine with iodine atom can be explained by comparing the crystal packing of halogen analogues with methyl and ethyl derivates. The methyl derivate occupies the same conformational space as F, Cl and Br derivates. The uracil hydrogen group forms an intermolecular hydrogen bond with the hydroxyl group on the furan ring. Furthermore, there are differences in size between the ethyl and methyl moiety. Substituting the methyl with an ethyl group, crystal packing with lower density is possible; resp. the uracil hydrogen group forms an intermolecular hydrogen bond with the methoxyl group, but not with the hydroxyl moiety. The lower melting point of ethyl derivate results in the similar solubility values for ethyl and methyl derivates, although the lipophilicity and molecular weight of the ethyl derivate is higher. Therefore, melting point is an important parameter for assessing the influence of crystal cohesive energy on solubility. Within a series of compounds with similar structures, the influence of solid state on solubility can be especially high, when modification of substituents causes a change of crystal packing.

An external data set of (4S)-7-(4-amino-2-substituted-pyrrolidin-1-yl)quinolone-3-carboxylic acids[93] is used in the following to demonstrate the impact of solid state on solubility.

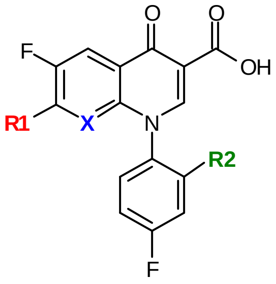


Figure 56: The scaffold of (4S)-7-(4-amino-2-substituted-pyrrolidin-1-yl)quinolone-3-carboxylic acids[93].

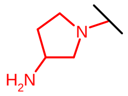
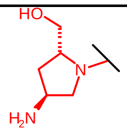
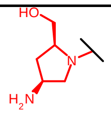
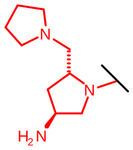
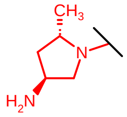
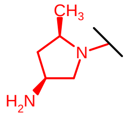
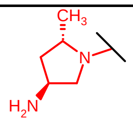
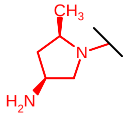
| Nr | Substituent | | | pKa | | | MW | logP | S (µg/mL) | Log1/S₀ | MP (°C) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | R1 | X | R2 | pKa1 | pKa2 | pKa3 | | | | | |
| 1 |  | **N** | **F** | 4 | 8 | | 404 | -0.46 | 8 | 4.80 | |
| 2 |  | **CH** | **F** | 4 | 8 | | 433 | -0.98 | 60 | 3.96 | 214-217[93] |
| 3 |  | **CH** | **F** | 4 | 8 | | 433 | | 60 | 3.96 | >240[93] |
| 4* |  | **CH** | **F** | 4 | 8 | 9 | 486 | -0.56 | 680 | 3.55 | 245[93] |
| 5 |  | **CH** | **F** | 4 | 8 | | 417 | 0.085 | 53 | 3.99<br>**3.34 fold** | 206-210[93] |
| 6 |  | **CH** | **F** | 4 | 8 | | 417 | 0.03 | 182 | 3.46 | >150[93] |
| 7 |  | **N** | **F** | 4 | 8 | | 418 | -0.2 | 150 | 3.54<br>**2.26 fold** | 231-234[93] |
| 8 |  | **N** | **F** | 4 | 8 | | 418 | -0.11 | 340 | 3.19 | 294-296[93] |

Table 33: The solubility of (4S)-7-(4-amino-2-substituted-pyrrolidin-1-yl)quinolone-3-carboxylic acids[93] measured using 0.05 M phosphate buffer at pH = 7.4.

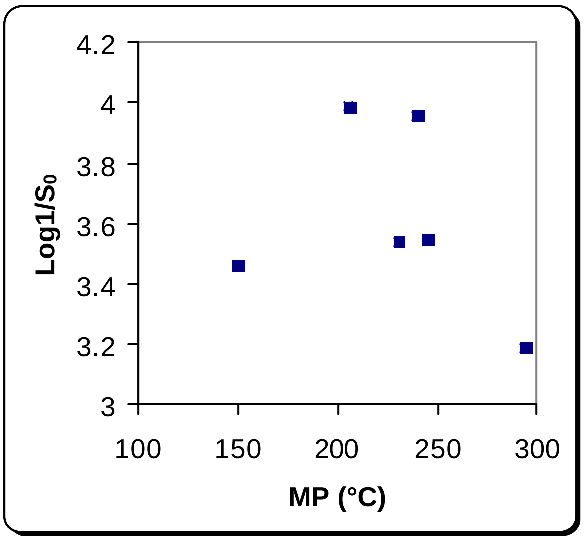
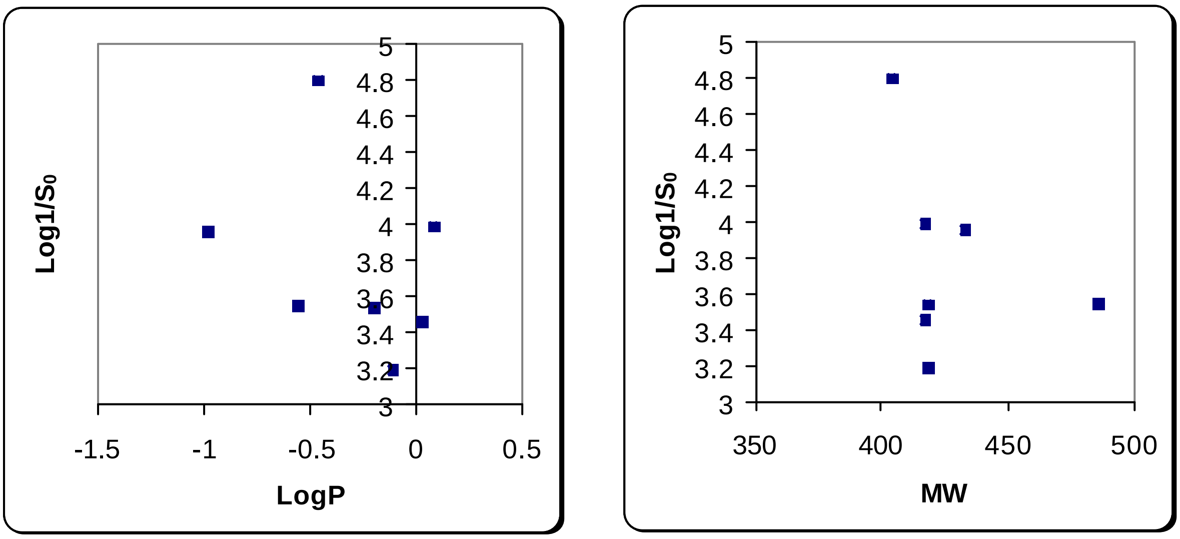


Figure 57: The relationship between melting points and solubilities of (4S)-7-(4-amino-2-substituted-pyrrolidin-1-yl)quinolone-3-carboxylic acids[93].

Table 33 shows, although melting point is an important parameter for solubility, it can not be used independently. For example, no direct correlation is observed between solubilities and melting points in this series. (Figure 57) Furthermore, the solubility of compound 8 is higher than compound 2, in spite of its higher melting point. Nevertheless, melting point is still an useful parameter in describing solubility variation. Table 33 shows, that the solubility to a certain degree is dependent on the stereo chemistry. The higher melting points of stereoisomers 7 and 8 lead to minor improvement in their solubility in comparison to stereoisomers 5 and 6.



a)

b)

Figure 58: a) The relationship between solubility and lipophilicity; b) the relationship between solubility and molecular weight of (4S)-7-(4-amino-2-substituted-pyrrolidin-1-yl)quinolone-3-carboxylic acids[93].

In case of no remarkable relationship between lipophilicity, molecular weight and solubility (Figure 58), fragmental based tool is often considered as an useful method to predict the solubility.

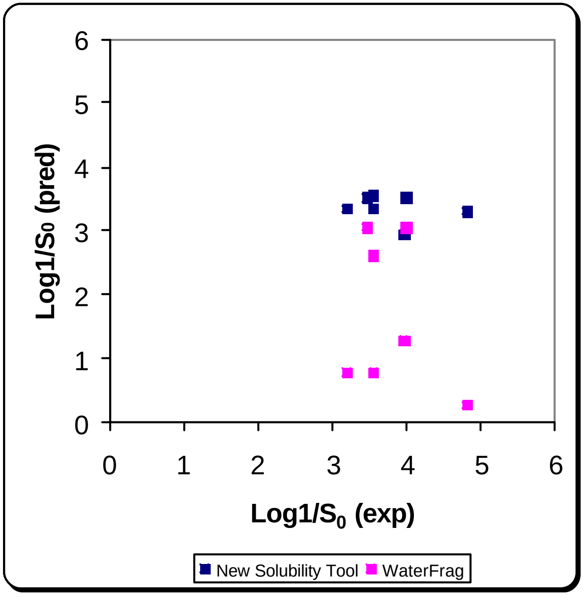
<table>
<tr><th>Nr</th><th>$Log1/S_0$ (exp)</th><th>$Log1/S_0$ (pred)</th><th>$Log1/S_0$ (pred WaterFrag)</th><td rowspan="9">
</td></tr>
<tr><td>1</td><td>4.80</td><td>3.31</td><td>0.27</td></tr>
<tr><td>2</td><td>3.96</td><td>2.95</td><td>1.28</td></tr>
<tr><td>3</td><td>3.96</td><td>2.95</td><td>1.28</td></tr>
<tr><td>4</td><td>3.55</td><td>3.55</td><td>2.62</td></tr>
<tr><td>5</td><td>3.99</td><td>3.52</td><td>3.05</td></tr>
<tr><td>6</td><td>3.46</td><td>3.52</td><td>3.05</td></tr>
<tr><td>7</td><td>3.54</td><td>3.35</td><td>0.79</td></tr>
<tr><td>8</td><td>3.19</td><td>3.35</td><td>0.79</td></tr>
</table>

Table 34: Comparison of the solubility of (4S)-7-(4-amino-2-substituted-pyrrolidin-1-yl)quinolone-3-carboxylic acids[93] predicted with the newly developed solubility tool and WaterFrag.

Table 34 shows, that the newly developed solubility tool can not differentiate between the solubility of stereoisomers, because it is based on 2D structures. (e.g. compound 5 and 6, compound 7 and 8). Nevertheless, the tool developed in this study shows significant better solubility prediction results than the commercial product WaterFrag[96] for all eight compounds listed in Table 34.

In conclusion, the impact of solid state on solubility was studied by collection of compounds with diverse polymorphic forms, measured 3D crystal structures and melting points. When polymorphs are monotropic, a direct comparison of melting point and solubility could be performed and crystal state related information was identified as an important factor for solubility. However, in comparison with lipophilicity, the influence of crystal lattice on solubility is restricted. Solubility differences caused by diverse crystal packing of polymorphs are usually 2 to 3 fold, which is in the range of the experimental error of the solubility measurements. Within a series of compounds with similar structures, the often described rules that the higher melting point related to lower solubility can not be always confirmed. Nevertheless, the influence of solid state on solubility can be especially high, when modification of substituents cause a change of crystal packing. Therefore, melting point is an useful parameter in describing solubility variation. Furthermore, in case, when solubility values can not be reasonably predicted using properties like lipophilicity, molecular weight and crystal state related information, fragmental based solubility prediction tool can be considered as an alternative. Thus, the influence of substituents on solubility was carefully studied using high quality solubility data of drug-like compounds in congeneric series. A general solubility model of high accuracy was developed, which showed significantly higher predictive power for drug-like compounds in comparison to commercially available tools. The derived fragment contributions to solubility can guide the decision processes in the synthesis of more soluble drug candidates.

# 5 CONCLUSIONS AND OUTLOOK

Aqueous solubility of drug-like compounds was studied from two aspects, in vitro and in silico. A new crystallization method based on saturated solution was developed, which avoids disadvantages of usual methods such as formation of amorphous materials from supersaturation. The invented method was applied for patent EU 05018750.9, in order to secure the achieved intellectual property for future pharmaceutical application. The experimental results showed crystalline forms of weak acids or bases could be more easily obtained using the newly developed crystallization method. Compounds with high tendency to form amorphous materials usually had higher HT solubility values than equilibrium solubilities. In the development of improved in silico solubility models, lipophilicity was confirmed as the major driving factor and crystal information related descriptors as the second important factor for solubility. Reasons for the limited precision of commercial solubility prediction tools were identified. A general solubility model of high accuracy was obtained for drug-like compounds in congeneric series when lipophilicity was used as descriptor in combination with the structural fragments. Rules were derived from the prediction models of solubility which could be used by chemists or interested scientists as a rough guideline on the contribution of structural fragments on solubility: Aliphatic and polar fragments with high dipole moments are always considered as solubility enhancing. Strong acids and bases usually have lower intrinsic solubility than neutral compounds. In summary, an improved solubility prediction method for congeneric series was developed using high quality solubility results of drugs and drug precursors as input parameter. The derived model overcomes difficulties of commercially available solubility prediction tools by focusing on structurally related series and showed a much higher predictive power for drug-like compounds in comparison to commercially available tools.

The theoretical solubility model obtained in this thesis has an average error of ± 0.42 log units. Practical solubility measurements showed average error of ± 0.143 log units. Thus the newly developed model meets well or exceeds the precision of commercially available prediction tools which have a typical deviation of about 1 to 2 log units[66]. Hence the new model can be smoothly used for aqueous solubility predictions of drug-like compounds in congeneric series and for evaluating the substituental effect on the solubility. However, there is still room for improvement and there are a number of different options to further improve the success of the prediction. Firstly, the predictive power could be enhanced by using the measured ionization constant to consider the pH dependence of solubility. Secondly, incorporating information reflecting solid state, e.g. crystal structure and density, melting point and information about polymorphic forms, would result in better solubility predictions. Thirdly, the model flexibly allows extending the initial data set by addition of further structural fragments in order to enhance the predictive power. Fourthly, solubility predictions could be extended from aqueous to other solvent systems, when high quality data in such systems are available. Solubility in other solvent systems is considered as important as aqueous solubility, because different solvent and solvent mixtures are frequently used in the pharmaceutical industry for formulation and crystallization. However, complications may occur in solubility measurements, because formation of aggregates or micelles in the solution can cause shifts in solubility which are related to shifts in $pK_a$ value. Therefore, in order to develop a more advanced prediction tool, new procedures should be developed to allow the determination of solubility in different solvent systems precisely. In conclusion, the generation of high quality data containing useful information is regarded as the crucial step in future model development. Today's rapid advances in fully automated or robot-driven measurement systems as well as in high-speed and high-precision measurement technologies offer a promising perspective on the future of solubility models.

# 6 ABBREVIATIONS

| | |
|---|---|
| % C | The percentage of aliphatic carbon in a molecule |
| % aromatic atom | The percentage of aromatic atoms in a molecule |
| % NO | The percentage of the sum of nitrogen and oxygen atoms in a molecule |
| A | Amphiphilic moment |
| Ap$K_a$ | Acid p$K_a$ |
| Bp$K_a$ | Base p$K_a$ |
| CP | Critical packing and defined as volume (hydrophobic)/[surface(hydrophilic)*length hydrophobic] |
| d0 | The “thickness” of the structure and defined as the root mean square deviation of the atom positions from the plane defined by the maximum and medium principal axes |
| d1 | The “width” of the structure and defined as d0 but for maximum and minimum axes |
| d2 | The “length” of the structure and defined as d0 but for the medium and minimum axes |
| Emin1-3 | The interaction energy between water and molecule at 3 best local minima |
| GC | Group Contribution |
| HL1-2 | Hydrophilic-lipophilic balance, which is defined as ratio of hydrophilic (-3, -4 kcal/mol)/lipophilic (-0.6, -0.8 kcal/mol) |

| | |
|---|---|
| MLR | Multiple linear regression |
| MP | Melting point |
| NN | Neutral network |
| Ovality | Defined as total surface area/surface area of a shere with volume equal to the total volume. This quantity must be larger than or equal to one |
| PCA | Principal component analysis |
| PLS | Projection to latent structure |
| Rmse | Root mean square error of the fit for observations in the data set |
| Rg | Radius of gyration, which is defined as root mean square distance of the atoms from the centroid |
| S | Solubility |
| $S_0$ | The molarity of the unionized molecular species |
| Log1/$S_0$ | The logarithmic transformed form of $S_0$ |
| V | Molecular volume |
| VlP | Variable Influence on Projection |

# 7 REFERENCES

(1) Du-Cuny, L.; Fischer, H.; Huwyler, J.; Kansy, M. Method for crystallization of a weakly acidic and/or weakly basic compound. EP Patent Appln. No.05018750.9 filed August 30, **2005**.

(2) Kansy, M. AAPS meeting 19. May, 2003. **2003**.

(3) Chan, O. H.; Stewart, B. H. Physicochemical and drug-delivery considerations for oral drug bioavailability. *Drug Discovery Today* **1996**, *1*, 461-473.

(4) Avdeef, A. Physicochemical profiling (solubility, permeability and charge state). *Current Topics in Medicinal Chemistry (Hilversum, Netherlands)* **2001**, *1*, 277-351.

(5) Abraham, M. H.; Le, J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *JOURNAL OF PHARMACEUTICAL SCIENCES* **1999**, *88*, 868-880.

(6) Meylan, W. M.; Howard, P. H. Estimating log P with atom/fragments and water solubility with log P. *Perspectives in Drug Discovery and Design* **2000**, *19*, 67-84.

(7) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from Monte Carlo simulations. *Bioorganic & Medicinal Chemistry Letters* **2000**, *10*, 1155-1158.

(8) Yalkowsky, S. H. Aqueous Solubility. Methods of estimation for organic compounds. **1992**.

(9) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 1488-1493.

(10) Klamt, A.; Eckert, F.; Hornig, M.; Beck, M. E.; Burger, T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *Journal of Computational Chemistry* **2002**, *23*, 275-281.

(11) Garside, J.; Tavare, N. S. Recent progress in solids processing: recent advances in industrial crystallization research. *Chemical Engineering Research and Design* **1986**, *64*, 77-79.

(12) Smakula, A. *Einkristalle Wachstum, Herstellung und Anwendung*; Springer-Verlag / Berlin. Göttingen. Heidelberg, **1962**.

(13) Gray, R. J.; Hou, W. B.; Kudryavtsev, A. B.; DeLucas, L. J. A new approach to the measurement of protein solubility by Michaelson interferometry. *Journal of Crystal Growth* **2001**, *232*, 10-16.

(14) Wiencek, J. M. New strategies for protein crystal growth. *Annual Review of Biomedical Engineering* **1999***, 1*, 505-534, 501 Plate.

(15) Berisio, R.; Lamzin, V. S.; Sica, F.; Wilson, K. S.; Zagari, A. et al. Protein titration in the crystal state. *Journal of Molecular Biology* **1999***, 292*, 845-854.

(16) Baird, J. K. Theory of protein crystal nucleation and growth controlled by solvent evaporation. *Journal of Crystal Growth* **1999***, 204*, 553-562.

(17) Stewart, P. D. S.; Baldock, P. F. M. Practical experimental design techniques for automatic and manual protein crystallization. *Journal of Crystal Growth* **1999***, 196*, 665-673.

(18) McPherson, A. Crystallization of proteins by variation of pH or temperature. *Methods in Enzymology* **1985***, 114*, 125-127.

(19) Wang, F.; Berglund, K. A. Monitoring pH Swing Crystallization of Nicotinic Acid by the Use of Attenuated Total Reflection Fourier Transform Infrared Spectrometry. *Industrial & Engineering Chemistry Research* **2000***, 39*, 2101-2104.

(20) Zhu, J.; Garside, J. Controlled batch crystallization by pH variation. *Jubilee Research Event, a Two-Day Symposium, Nottingham, UK, Apr. 8-9, 1997* **1997***, 1*, 449-452.

(21) Shekunov, B. Y.; York, P. Crystallization processes in pharmaceutical technology and drug delivery design. *Journal of Crystal Growth* **2000***, 211*, 122-136.

(22) Avdeef, A. pH-metric solubility. 1. Solubility-pH profiles from Bjerrum plots. Gibbs buffer and pKa in the solid state. *Pharmacy and Pharmacology Communications* **1998***, 4*, 165-178.

(23) Avdeef, A.; Comer, J. E. A. Measurement of pKa and log P of water-insoluble substances by potentiometric titration. *Trends QSAR Mol. Modell. 92, Proc. Eur. Symp. Struct.-Act. Relat.: QSAR Mol. Modell., 9th* **1993**, 386-387.

(24) Box, K.; Bevan, C.; Comer, J.; Hill, A.; Allen, R. et al. High-Throughput Measurement of pKa Values in a Mixed-Buffer Linear pH Gradient System. *Analytical Chemistry* **2003***, 75*, 883-892.

(25) Allen, R. I.; Box, K. J.; Comer, J. E.; Peake, C.; Tam, K. Y. Multiwavelength spectrophotometric determination of acid dissociation constants of ionizable drugs. *JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS* **1998***, 17*, 699-712.

(26) Kansy, M. Strategies for improving solubility. *Oct. 2-3, 2003, Conference at Marriott Hotel, Brussels, Belgium* **2003**.

(27) Pudipeddi, M.; Serajuddin, A. T. M. Trends in solubility of polymorphs. *Journal of Pharmaceutical Sciences* **2005**, *94*, 929-939.

(28) Schinzer, W. C.; Bergren, M. S.; Aldrich, D. S.; Chao, R. S.; Dunn, M. J. et al. Characterization and interconversion of polymorphs of premafloxacin, a new quinolone antibiotic. *JOURNAL OF PHARMACEUTICAL SCIENCES* **1997**, *86*, 1426-1431.

(29) El-Gindy, N. A.; Ebian, A. R. Codeine crystal forms. II. Thermodynamics, stabilization and tabletting. *Scientia Pharmaceutica* **1978**, *46*, 8-16.

(30) Gerber, J. J.; VanderWatt, J. G.; Loetter, A. P. Physical characterization of solid forms of cyclopenthiazide. *International Journal of Pharmaceutics* **1991**, *73*, 137-145.

(31) Ebian, A. R.; El-Gindy, N. A. Codeine crystal forms. I. Preparation, identification and characterization. *Scientia Pharmaceutica* **1978**, *46*, 1-7.

(32) Durbin, S. D.; Feher, G. Protein Crystallization. *Annual Review of Physical Chemistry* **1996**, *47*, 171-204.

(33) Biscans, B.; Laguerie, C. Determination of induction time of lysozyme crystals by laser diffraction. *Journal of Physics D: Applied Physics* **1993**, *26*, B118-B122.

(34) Palmer, R. A.; Niwa, H. X-ray crystallographic studies of protein-ligand interactions. *Biochemical Society transactions* **2003**, *31*, 973-979.

(35) Saridakis, E. E.; Stewart, P. D.; Lloyd, L. F.; Blow, D. M. Phase diagram and dilution experiments in the crystallization of carboxypeptidase G2. *Acta crystallographica. Section D, Biological crystallography* **1994**, *50*, 293-297.

(36) Sirius Analytical Instruments Ltd.

solubility Training Course, pSOL model 3 Noyes-Whitney template titrator. **2003**.

(37) Hancock, B. C.; Parks, M. What is the true solubility advantage for amorphous pharmaceuticals? *Pharmaceutical Research* **2000**, *17*, 397-404.

(38) PHYSPROP Syracuse Research Coporation. **1994**.

(39) Stahura, F. L.; Godden, J. W.; Bajorath, J. Differential Shannon Entropy Analysis Identifies Molecular Property Descriptors that Predict Aqueous Solubility of Synthetic Compounds with High Accuracy in Binary QSAR Calculations. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 550-558.

(40) Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. *Journal of Chemical Information and Computer Sciences*, ACS ASAP.

(41) Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved method for estimating water solubility from octanol/water partition coefficient. *Environmental Toxicology and Chemistry* **1996***, 15*, 100-106.

(42) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *Journal of Chemical Information and Computer Sciences* **2003***, 43*, 429-434.

(43) Engkvist, O.; Wrede, P. A Fuzzy ARTMAP Based on one- and two-dimensional descriptors. *Journal of Chemical Information and Computer Sciences, 42*, 1247-1249.

(44) Cheng, A.; Zell, A. Prediction of aqueous solubility of diverse set of compounds using quantitative structure-property relationships. *Abstracts of Papers - American Chemical Society* **2001***, 43*, COMP-137.

(45) McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. Estimating the Water Solubilities of Crystalline Compounds from Their Chemical Structures Alone. *Journal of Chemical Information and Computer Sciences* **2001***, 41*, 1355-1359.

(46) Liu, R.; So, S.-S. Development of Quantitative Structure-Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *Journal of Chemical Information and Computer Sciences* **2001***, 41*, 1633-1639.

(47) Engkvist, O.; Wrede, P. High-throughput, in silico prediction of aqueous solubility based on one- and two-dimensional descriptors. *JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES* **2002**, *42*, 1247-1249.

(48) Wegner Jorg, K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES* **2003**, *43*, 1077-1084.

(49) Klamt, A.; Eckert, F.; Li, Y.; Venkatesh, S. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *Journal of Computational Chemistry* **2002***, 23*, 275-281.

(50) Bergstroem, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and solubilization properties of drug-like molecules. *Journal of Chemical Information and Computer Sciences* **2004***, 44*, 1477-1488.

(51) Kier, L. B.; Hall, L. H. The nature of structure-activity relationships and their relation to molecular connectivity. *European Journal of Medicinal Chemistry* **1977***, 12*, 307-312.

(52) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *Journal of Chemical Information and Computer Sciences* **1995***, 35*, 1074-1080.

(53) Yalkowsky, S. H.; Valvani, S. C. Solubility and partitioning. I: Solubility of nonelectrolytes in water. *Journal of Pharmaceutical Sciences* **1980***, 69*, 912-922.

(54) Chen, X.-Q.; Cho, S. J.; Li, Y.; Venkatesh, S. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *Journal of Pharmaceutical Sciences* **2002***, 91*, 1838-1852.

(55) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *Journal of Chemical Information and Computer Sciences* **2001***, 41*, 1237-1247.

(56) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *Journal of Chemical Information and Computer Sciences*, ACS ASAP.

(57) Cheng, A.; Merz, K. M., Jr. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure-Property Relationships. *Journal of Medicinal Chemistry* **2003***, 46*, 3572-3580.

(58) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A Fuzzy ARTMAP Based on Quantitative Structure-Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *Journal of Chemical Information and Computer Sciences* **2001***, 41*, 1177-1207.

(59) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water-Air Partition Coefficients. *Journal of Chemical Information and Computer Sciences* **1998***, 38*, 720-725.

(60) Wanchana, S.; Yamashita, F.; Hashida, M. Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least squares method. *Pharmazie* **2002***, 57*, 127-129.

(61) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Advanced Drug Delivery Reviews* **2002***, 54*, 355-366.

(62) Delgado, E. J. Predicting aqueous solubility of chlorinated hydrocarbons from molecular structure. *Fluid Phase Equilibria* **2002***, 199*, 101-107.

(63) Nikolic, S.; Trinajstic, N. Modeling the aqueous solubility of aliphatic alcohols. *SAR and QSAR in Environmental Research* **1998**, *9*, 117-126.

(64) Yin, C.; Liu, X.; Guo, W.; Lin, T.; Wang, X. et al. Prediction and application in QSPR of aqueous solubility of sulfur-containing aromatic esters using GA-based MLR with quantum descriptors. *Water Research* **2002**, *36*, 2975-2982.

(65) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Internet software for the calculation of the lipophilicity and aqueous solubility of chemical compounds. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 246-252.

(66) Le, J. The prediction of aqueous solubility for organic compounds. *Post-doctoral project funded by F. Hoffmann-La Roche* **2002**.

(67) Yakowsky, S. H.; Dannenfelser, R. M. dATAbASE of Aqueous Solubility. **1990**.

(68) Gerber, P. R.; Mueller, K. MAB, a generally applicable molecular force field for structure modeling in medicinal chemistry. *Journal of Computer-Aided Molecular Design* **1995**, *9*, 251-268.

(69) CallistoGen AG CallistoGen.

(70) Eriksson, L.; Johansson, E. Intoduction to multi- and megavariate data analysis using projection methods (PCA &PLS). **1999**.

(71) Sun, H. A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 748-757.

(72) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 266-275.

(73) Bergstroem, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and local computational models for aqueous solubility prediction of drug-like molecules. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1477-1488.

(74) *Daylight*; Daylight Chemical Information Systems Inc.: Aliso Viejo.

(75) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O. et al. The development of versions 3 and 4 of the Cambridge Structural Database System. *Journal of Chemical Information and Computer Sciences* **1991**, *31*, 187-204.

(76) Hasselbalch, K. A. Die Berechnung der Wasserstoffzahl des Blutes aus der freien und gebunden Kohlensäure desselben, und die Sauerstoffbindung des Blutes als Funktion der Wasserstoffzahl. *Die Biochem. Z.* **1916***, 78*, 112-144.

(77) *ClogP*; BioByte Corp.: Claremont, USA.

(78) *ACD*; Advanced Chemistry Development Inc.: Toronto, Canada.

(79) Cruciani, G.; Meniconi, M.; Carosati, E.; Zamora, I.; Mannhold, R. VOLSURF: a tool for drug ADME-properties prediction. *Methods and Principles in Medicinal Chemistry* **2003***, 18*, 406-419.

(80) Rauhut, G.; Clark, T. Multicenter point charge model for high-quality molecular electrostatic potentials from AM1 calculations. *Journal of Computational Chemistry* **1993***, 14*, 503-509.

(81) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chemical Reviews (Washington, DC, United States)* **1993***, 93*, 2567-2581.

(82) Frisch, A.; Frisch, M. J. Gaussian98 Users Reference. **1998**.

(83) Longfils, G.; Ooms, F.; Wouters, J.; Olivier, A.; Sevrin, M. et al. A comparison of ab initio, semi-empirical, and molecular mechanics approaches to compute molecular geometries and electrostatic descriptors of heteroatomic ring fragments observed in drugs molecules. *Molecular Modeling and Prediction of Bioactivity, [Proceedings of the European Symposium on Quantitative Structure-Activity Relationships: Molecular Modeling and Prediction of Bioactivity], 12th, Copenhagen, Denmark, Aug. 23-28, 1998* **2000**, 482-483.

(84) Lee, J. E.; Choi, W.; Mhin, B. J. DFT calculation on the electron affinity of polychlorinated dibenzo-p-dioxins. *Bulletin of the Korean Chemical Society* **2003***, 24*, 792-796.

(85) Maekelae, N. I.; Knuuttila, H. R.; Linnolahti, M.; Pakkanen, T. A.; Leskelae, M. A. Activation of Racemic Ethylene-Bridged Bis(indenyl)-Type Siloxy-Substituted Zirconocenes with Methylaluminoxane. A Combined UV/vis Spectroscopic and ab Initio Hartree-Fock Study. *Macromolecules* **2002***, 35*, 3395-3401.

(86) Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *The Properties of Gases and Liquids*, **1987**; 741 pp.

(87) Neely, W. B.; Blau, G. E.; Editors *Environmental Exposure from Chemicals, Vol. 1*, **1985**; 245 pp.

(88) Goosen, C.; Laing Timothy, J.; du Plessis, J.; Goosen Theunis, C.; Flynn Gordon, L. Physicochemical characterization and solubility analysis of

thalidomide and its N-alkyl analogs. *PHARMACEUTICAL RESEARCH* **2002**, *19*, 13-19.

(89) *MedChem database*; BioByte Corp.: Claremont, USA.

(90) Bergstroem, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs. *Journal of Chemical Information and Computer Sciences*, ACS ASAP.

(91) *SRC WsKow*; Syracuse Research Corporation: North Syracuse.

(92) Bavetsias, V.; Skelton, L. A.; Yafai, F.; Mitchell, F.; Wilson, S. C. et al. The Design and Synthesis of Water-Soluble Analogues of CB30865, a Quinazolin-4-one-Based Antitumor Agent. *Journal of Medicinal Chemistry* **2002**, *45*, 3692-3702.

(93) Rosen, T.; Chu, D. T.; Lico, I. M.; Fernandes, P. B.; Marsh, K. et al. Design, synthesis, and properties of (4S)-7-(4-amino-2-substituted-pyrrolidin-1-yl)quinolone-3-carboxylic acids. *JOURNAL OF MEDICINAL CHEMISTRY* **1988**, *31*, 1598-1611.

(94) Edwards, P. D.; Andisik, D. W.; Strimpler, A. M.; Gomes, B.; Tuthill, P. A. Nonpeptidic Inhibitors of Human Neutrophil Elastase. 7. Design, Synthesis, and in Vitro Activity of a Series of Pyridopyrimidine Trifluoromethyl Ketones. *Journal of Medicinal Chemistry* **1996**, *39*, 1112-1124.

(95) Bernstein, P. R.; Andisik, D.; Bradley, P. K.; Bryant, C. B.; Ceccarelli, C. et al. Nonpeptidic inhibitors of human leukocyte elastase. 3. Design, synthesis, X-ray crystallographic analysis, and structure-activity relationships for a series of orally active 3-amino-6-phenylpyridin-2-one trifluoromethyl ketones. *Journal of medicinal chemistry* **1994**, *37*, 3313-3326.

(96) *WaterFrag*; Syracuse Research Corporation: North Syracuse.

(97) *Kowwin*; Syracuse Research Corporation: North Syracuse.

(98) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *Journal of Chemical Information and Computer Sciences* **1992**, *32*, 474-482.

(99) Streng, W. H. Physical chemical characterization of drug substances. *Drug Discovery Today* **1997**, *2*, 415-426.

(100) Schinzer, W. C.; Bergren, M. S.; Aldrich, D. S.; Chao, R. S.; Dunn, M. J. et al. Characterization and interconversion of polymorphs of premafloxacin, a new quinolone antibiotic. *JOURNAL OF PHARMACEUTICAL SCIENCES* **1997**, *86*, 1426-1431.

(101) Ghodbane, S.; McCauley, J. A. Study of the polymorphism of 3-[[[3-[2-(7-chloro-2-quinolinyl)-(E)-ethenyl)phenyl][[3-(dimethylamino-3-oxopropyl)thio]methyl]thio]propanoic acid (MK571) by DSC, TG, XRPD and solubility measurements. *International Journal of Pharmaceutics* **1990**, *59*, 281-286.

(102) Lindenbaum, S.; Rattie, E. S.; Zuber, G. E.; Miller, M. E.; Ravin, L. J. Polymorphism of auranofin. *International Journal of Pharmaceutics* **1985**, *26*, 123-132.

(103) Urakami, K.; Shono, Y.; Higashi, A.; Umemoto, K.; Godo, M. A novel method for estimation of transition temperature for polymorphic pairs in pharmaceuticals using heat of solution and solubility data. *Chemical & Pharmaceutical Bulletin* **2002**, *50*, 263-267.

(104) Urakami, K.; Shono, Y.; Higashi, A.; Umemoto, K.; Godo, M. Estimation of transition temperature of pharmaceutical polymorphs by measuring heat of solution and solubility. *Bulletin of the Chemical Society of Japan* **2002**, *75*, 1241-1245.

(105) McCauley, J. A. Detection and characterization of polymorphism in the pharmaceutical industry. *AIChE Symposium Series* **1991**, *87*, 58-63.

(106) Brandstaetter-Kuhnert, M.; Martinek, A. Effect of polymorphism on the solubility of pharmaceuticals. *Mikrochimica et Ichnoanalytica Acta* **1965**, 909-919.

(107) Holleman, W. Anorganic chemistry.

# 8 APPENDIX

| Nr. | Name | Connection Environment | SMART | $b_i$ | Coeff (ClogP) | Fragment contribution to $\log 1/S_0$ |
|---|---|---|---|---|---|---|
| Frag1 | Tertiary Amine | AZZ | AN(Z)Z | -0.505804 | -2.2 | -0.78789 |
| Frag2 | Tertiary Amine | AAa | A[N&X3](A)a | 0.246447 | -1.12 | 0.102841 |
| Frag3 | Tertiary Amine | AAA | AN(A)A | 0.112752 | -2.37 | -0.19113 |
| Frag4 | Tertiary Amine | AAZ | AN(A)Z | 0.03326 | -1.98 | -0.22062 |
| Frag5 | Secondary amine | AA | A[NH]A | 0.283721 | -1.77 | 0.056772 |
| Frag6 | Secondary amine | Aa | A[NH]a | -0.0704758 | -1.03 | -0.20254 |
| Frag7 | Secondary amine | aa | a[NH]a | 0.0130401 | -0.09 | 0.0015 |
| Frag8 | Secondary amine | AZ | A[NH]Z | 0.320647 | -1.69 | 0.103955 |
| Frag9 | Secondary amine | Za | [NH](Z)a | -0.0950652 | -1.15 | -0.24252 |
| Frag10 | Secondary amine | ZZ | [NH](Z)Z | -0.0924969 | -2.1 | -0.36176 |
| Frag11 | Primary Amine | A | A[NH2] | 0.88698 | -1.54 | 0.689521 |
| Frag12 | Primary Amine | Z | [NH2]Z | 0.800753 | -1.35 | 0.627656 |
| Frag13 | Primary Amine | a | a[NH2] | -0.293951 | -1 | -0.42217 |
| Frag14 | Acid Hydrazide-NH | aa | a[NH][NH]C(a)=O | -0.21722 | -2.3 | -0.51213 |
| Frag15 | Aromatic Amide | aa | a[nH]c(a)=O | -0.148826 | -2 | -0.40527 |
| Frag16 | Acid Imide | Aza | AN(C(Z)=O)C(a)=O | -1.05145 | -1.72 | -1.27199 |
| Frag17 | Amide | AAA | AN(A)C(A)=O | -0.568526 | -3.19 | -0.97755 |
| Frag18 | Amide | AAa | AN(A)C(a)=O | -0.328139 | -2.82 | -0.68972 |
| Frag19 | Amide | AaA | AN(a)C(A)=O | -0.495233 | -1.4 | -0.67474 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Frag20 | Amide | aaa | aN(a)C(a)=O | -0.0949459 | -0.33 | -0.13726 |
| Frag21 | Amide | Aaa | AN(a)C(a)=O | 0.0412647 | -2.09 | -0.22672 |
| Frag22 | Amide | AaZ | AN(a)C(Z)=O | -0.459037 | -2.12 | -0.73086 |
| Frag23 | Amide | AAZ | AN(A)C(Z)=O | -0.860587 | -2.99 | -1.24396 |
| Frag24 | Amide | AZA | AN(Z)C(A)=O | -0.105184 | -2.99 | -0.48856 |
| Frag25 | Amide | AZa | AN(Z)C(a)=O | -0.3433 | -2.2 | -0.62538 |
| Frag26 | Amide | AZZ | AN(Z)C(Z)=O | -0.533844 | -2.87 | -0.90184 |
| Frag27 | Formylamine | AA | AN(A)[CH]=O | -0.103085 | -2.67 | -0.44543 |
| Frag28 | NH-Amide | AA | A[NH]C(A)=O | -0.260609 | -2.71 | -0.60809 |
| Frag29 | NH-Amide | Aa | A[NH]C(a)=O | -0.148322 | -1.81 | -0.3804 |
| Frag30 | NH-Amide | aA | a[NH]C(A)=O | -0.0877811 | -1.51 | -0.28139 |
| Frag31 | NH-Amide | aa | a[NH]C(a)=O | 0.0820121 | -1.06 | -0.0539 |
| Frag32 | NH-Amide | AV | A[NH]C(V)=O | -0.416968 | -2.26 | -0.70675 |
| Frag33 | NH-Amide | AZ | A[NH]C(Z)=O | -0.414681 | -2.51 | -0.73651 |
| Frag34 | NH-Amide | aZ | a[NH]C(Z)=O | -0.124235 | -1.54 | -0.32169 |
| Frag35 | NH-Amide | aV | a[NH]C(V)=O | -0.530449 | -1.3 | -0.69714 |
| Frag36 | NH-Amide | ZA | [NH](Z)C(A)=O | -0.58768 | -2.25 | -0.87618 |
| Frag37 | NH-Amide | Za | [NH](Z)C(a)=O | -0.0663947 | -1.41 | -0.24718 |
| Frag38 | Formamine-NH | a | a[NH][CH]=O | 0.0694213 | -0.75 | -0.02674 |
| Frag39 | Urea (tetrasub) | AAAA | AN(A)C(=O)N(A)A | -0.598173 | -3.01 | -0.98412 |
| Frag40 | 1,1,3-Urea | Aaa | A[NH]C(=O)N(a)a | -0.085475 | -2.16 | -0.36243 |
| Frag41 | 1,1,3-Urea | aAA | a[NH]C(=O)N(A)A | -0.282532 | -2.77 | -0.6377 |
| Frag42 | 1,1,3-Urea | aAZ | a[NH]C(=O)N(A)Z | -0.492871 | -2.09 | -0.76085 |
| Frag43 | N,N' Urea | Aa | A[NH]C(=O)[NH]a | -0.137129 | -1.57 | -0.33843 |
| Frag44 | N,N' Urea | Za | [NH](Z)C(=O)[NH]a | 0.298589 | -1.37 | 0.122928 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Frag45 | NH-Urea | a | a[NH]C([NH2])=O | -0.374615 | -1.07 | -0.51181 |
| Frag46 | NH-Carbamate | aA | a[NH]C(=O)OA | 0.288844 | -1.06 | 0.152931 |
| Frag47 | NH-Carbamate | aZ | a[NH]C(=O)OZ | 1.01701 | -1.06 | 0.881097 |
| Frag48 | NH2-Amide | a | aC([NH2])=O | 0.0208799 | -1.26 | -0.14068 |
| Frag49 | NH2-Amide | A | AC([NH2])=O | -0.411489 | -1.99 | -0.66665 |
| Frag50 | NH2-Amide | Z | C(Z)([NH2])=O | -0.131347 | -1.99 | -0.3865 |
| Frag51 | Thioamide-NH | aA | a[NH]C(A)=S | -0.420511 | -0.96 | -0.5436 |
| Frag52 | Thioamide-NH2 | A | AC([NH2])=S | 0.0493423 | -1.13 | -0.09555 |
| Frag53 | Ester | AA | AOC(A)=O | -0.275125 | -1.45 | -0.46104 |
| Frag54 | Ester | Aa | AOC(a)=O | -0.203746 | -0.56 | -0.27555 |
| Frag55 | Ester | AY | AOC(Y)=O | -0.086602 | -0.96 | -0.20969 |
| Frag56 | Ester | AZ | AOC(Z)=O | -0.124924 | -1.38 | -0.30187 |
| Frag57 | Ester | Za | O(Z)C(a)=O | -1.40291 | -0.3 | -1.44138 |
| Frag58 | Carboxy (ZW-) | A | AC([OH])=O | 0.509152 | -1.07 | 0.371957 |
| Frag59 | Carboxy (ZW-) | a | aC([OH])=O | 1.08236 | -0.03 | 1.078513 |
| Frag60 | Carboxy | Z | C(Z)([OH])=O | -0.295626 | -1.03 | -0.42769 |
| Frag61 | Carbonyl | Aa | AC(a)=O | 0.104273 | -1.09 | -0.03549 |
| Frag62 | Carbonyl | aa | aC(a)=O | -0.0367111 | -0.53 | -0.10467 |
| Frag63 | Carbonyl | AA | AC(A)=O | -0.61826 | -1.84 | -0.85418 |
| Frag64 | Aldehyde | a | a[CH]=O | 0.00267908 | -0.42 | -0.05117 |
| Frag65 | Ether | AA | AOA | -0.12718 | -1.82 | -0.36054 |
| Frag66 | Ether | Aa | AOa | -0.00697339 | -0.61 | -0.08519 |
| Frag67 | Ether | aa | aOa | 0.167299 | 0.53 | 0.235256 |
| Frag68 | Ether | AY | AOY | -0.0610615 | -1.3 | -0.22775 |
| Frag69 | Ether | AZ | AOZ | 0.242133 | -1.28 | 0.078011 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Frag70 | Ether | aZ | aOZ | 0.153457 | -0.41 | 0.100887 |
| Frag71 | Alcohol or Hydroxy | A | A[OH] | -0.372527 | -1.64 | -0.58281 |
| Frag72 | Alcohol or Hydroxy | a | a[OH] | -0.331856 | -0.44 | -0.38827 |
| Frag73 | Alcohol or Hydroxy | Z | [OH]Z | 0.00490034 | -1.34 | -0.16691 |
| Frag74 | Sulfide | AA | A[S&X2]A | 0.415272 | -0.7 | 0.325518 |
| Frag75 | Sulfide | Aa | A[S&X2]a | -0.093221 | 0.03 | -0.08937 |
| Frag76 | Sulfide | aa | a[S&X2]a | 0.14817 | 0.77 | 0.246899 |
| Frag77 | Sulfide | AZ | A[S&X2]Z | -0.613268 | -0.35 | -0.65815 |
| Frag78 | Sulfide | VV | V[S&X2]V | -0.416968 | 0.18 | -0.39389 |
| Frag79 | Sulfide | Za | [S&X2](Z)a | 0.591607 | 0.03 | 0.595454 |
| Frag80 | Azo | A | AN=[N+]=[N-] | -0.479841 | 0.62 | -0.40034 |
| Frag81 | Nitro | a | a[N+](=O)[O-] | 0.00594268 | -0.03 | 0.002096 |
| Frag82 | Nitrile | a | aC#N | 0.255116 | -0.34 | 0.211521 |
| Frag83 | Nitrile | A | AC#N | -0.0841117 | -1.27 | -0.24695 |
| Frag84 | Nitrile | Z | C(Z)#N | 0.742307 | -0.88 | 0.629473 |
| Frag85 | Fluoride | A | AF | -0.0522532 | -0.38 | -0.10098 |
| Frag86 | Fluoride | a | aF | 0.05092 | 0.37 | 0.098361 |
| Frag87 | Fluoride | Z | FZ | 0.019207 | -0.18 | -0.00387 |
| Frag88 | Chloride | a | aCl | 0.0639783 | 0.94 | 0.184505 |
| Frag89 | Chloride | Z | ClZ | 0.304997 | 0.26 | 0.338334 |
| Frag90 | Bromide | a | aBr | 0.364067 | 1.09 | 0.503827 |
| Frag91 | Iodide | a | aI | 0.0205602 | 1.35 | 0.193657 |
| Frag92 | Sulfoxide | AA | A[S&X3](A)=O | -0.5968 | -3.01 | -0.98274 |
| Frag93 | Sulfonyl | AA | AS(A)(=O)=O | 0.0359061 | -3.01 | -0.35004 |
| Frag94 | Sulfonyl | Aa | AS(a)(=O)=O | 0.633814 | -2.17 | 0.355577 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Frag95 | Sulfonamide | AAa | AN(A)S(a)(=O)=O | -0.179356 | -2.09 | -0.44734 |
| Frag96 | Sulfonamide | Aaa | AN(a)S(a)(=O)=O | 0.122288 | -1.67 | -0.09184 |
| Frag97 | Sulfonamide | AAA | AN(A)S(A)(=O)=O | -0.388641 | -1.37 | -0.5643 |
| Frag98 | Sulfonamide | AAZ | AN(A)S(Z)(=O)=O | -0.295066 | -2.76 | -0.64895 |
| Frag99 | Sulfonamide | AZa | AN(Z)S(a)(=O)=O | -0.362315 | -1.89 | -0.60465 |
| Frag100 | NH-Sulfonamide | Aa | A[NH]S(a)(=O)=O | -0.150474 | -1.75 | -0.37486 |
| Frag101 | NH-Sulfonamide | aA | a[NH]S(A)(=O)=O | -2.31194 | -1.72 | -2.53248 |
| Frag102 | NH-Sulfonamide | aa | a[NH]S(a)(=O)=O | -0.213922 | -1.13 | -0.35881 |
| Frag103 | NH-Sulfonamide | aZ | a[NH]S(Z)(=O)=O | -0.228931 | -1.6 | -0.43408 |
| Frag104 | NH-Sulfonamide | AA | A[NH]S(A)(=O)=O | -1.21692 | -2.5 | -1.53747 |
| Frag105 | NH-Sulfonamide | AZ | A[NH]S(Z)(=O)=O | -0.181414 | -2.42 | -0.49171 |
| Frag106 | NH-Sulfonamide | Za | [NH](Z)S(a)(=O)=O | -0.542045 | -1.55 | -0.74079 |
| Frag107 | NH2-Sulfonamide | a | aS([NH2])(=O)=O | 0.035859 | -1.61 | -0.17058 |
| Frag108 | tetrasubst. Sulfamide | AAAA | AN(A)S(=O)(=O)N(A)A | -0.287803 | -4.05 | -0.80709 |
| Frag109 | Sulfondiamide, trisubs. | AAA | A[NH]S(=O)(=O)N(A)A | -0.761375 | -3.4 | -1.19732 |
| Frag110 | Sulfondiamide,trisubs | aAA | a[NH]S(=O)(=O)N(A)A | 0.428803 | -2.043 | 0.16685 |
| Frag111 | Sulfondiamide,trisubs | ZAA | [NH](Z)S(=O)(=O)N(A)A | -0.0114224 | -1.545 | -0.20952 |
| Frag112 | Thiadiazoledioxide | AA | A[NH]S(=O)(=O)[NH]A | -0.942546 | -1.775 | -1.17014 |
| Frag113 | N-carboxysulfonamide | aa | aC(=O)[NH]S(a)(=O)=O | -1.69519 | -0.97 | -1.81956 |
| Frag114 | sulfonylurea,N(disubst-amino) | AAa | AN(A)[NH]C(=O)[NH]S(a)(=O)=O | 0.0325044 | -4.34 | -0.52397 |
| Frag115 | 1-Sulfonyl-3-Urea | Aa | A[NH]C(=O)[NH]S(a)(=O)=O | -0.519879 | -2.26 | -0.80966 |
| Frag116 | FragA | AA | A[NH]S(=O)(=O)[NH]C(=O)OA | -1.24472 | -1.745 | -1.46846 |
| Frag117 | FragB | AAa | AN(A)C=NS(a)(=O)=O | -0.318364 | -1.745 | -0.54211 |
| Frag118 | FragC | Aaaa | An(a)c(=O)n(a)S(a)(=O)(=O) | -1.04836 | -2.728 | -1.39814 |
| Frag119 | FragD | Zaaa | n(Z)(a)c(=O)n(a)S(a)(=O)(=O) | 0.774524 | -2.728 | 0.42474 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Frag120 | FragE | aaa | [nH](a)c(=O)n(a)S(a)(=O)(=O) | -0.154406 | -2.424 | -0.46521 |
| Frag121 | Thiophosporothioate | AAA | AOP(=S)(OA)SA | -0.051248 | 0.1 | -0.03843 |
| Frag122 | FragF | A | AOP([OH])([OH])=O | 1.15179 | -2.174 | 0.87304 |
| Frag123 | FragG | AAa | A[N+](A)(a)[O-] | -2.09328 | -1.349 | -2.26625 |
| Frag124 | cyanoguanidyl #1 | aAA | a[NH]C(=NC#N)N(A)A | -1.56706 | -1.104 | -1.70861 |
| Frag125 | Oxanilic ester | aA | a[NH]C(=O)C(=O)OA | 0.249812 | -1.72 | 0.029274 |
| Frag126 | Amidine | a | aC([NH2])=[NH] | 0.46317 | -1.27 | 0.300331 |
| Frag127 | FragH | aa | aC([NH2])=NC(a)=O | 0.537292 | -1.137 | 0.391506 |
| Frag128 | FragI | a | aC([NH2])=NO | -0.0988768 | -0.891 | -0.21312 |
| Frag129 | Dicarbonylhydrazine (sym) | aa | aC(=O)[NH][NH]C(a)=O | -0.392074 | -1.49 | -0.58312 |
| Frag130 | Acid Hydrazide-NH2 | A | AC(=O)[NH][NH2] | 0.212569 | -2.5 | -0.10798 |
| Frag131 | N,N-carboxamide,alpha-keto | AZA | AN(Z)C(=O)C(A)=O | 0.046828 | -3.105 | -0.3513 |
| Frag132 | Formocarboxamide | Aa | AN[CH]=O)C(a)=O | 0.18389 | -1.43 | 0.000535 |
| Frag133 | Acid Imide | Aaa | AN(C(a)=O)C(a)=O | 0.00251297 | -1.05 | -0.13212 |
| Frag134 | Tertiary Imine | Aaa | AN=C(a)a | -0.28748 | -1.65 | -0.49904 |
| Frag135 | Carbamate, N, N | AAA | AOC(=O)N(A)A | -0.00739098 | -1.95 | -0.25742 |
| Frag136 | N-carboxyguanidyl | Aa | AOC(=O)N=C(a)[NH2] | 0.357515 | -1.5 | 0.165185 |
| Frag137 | Carbonate | AA | AOC(=O)OA | 0.0425553 | -1.93 | -0.20491 |
| Frag138 | Iminoxy | Aa | AON=Ca | 0.0416457 | -0.6 | -0.03529 |
| Frag139 | FragJ | aa | a[n+](a)[O-] | -0.740466 | -1.745 | -0.96421 |
| Frag140 | 1-Pyrrole | Aaa | An(a)a | -0.0874169 | -1.09 | -0.22718 |
| Frag141 | 1-Pyrrole | aaa | an(a)a | -0.167806 | -0.56 | -0.23961 |
| Frag142 | 1-Pyrrole | Zaa | n(Z)(a)a | -0.181305 | -0.89 | -0.29542 |
| Frag143 | Ring amide, N-subst. | aaa | an(a)c(a)=O | -0.754877 | -2.35 | -1.05619 |
| Frag144 | Ring amide, N-subst. | aZa | an(Z)c(a)=O | -0.525923 | -2.39 | -0.83237 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Frag145 | Arom.1-(3H)Diazo-2,4-dioxo | Zaa | n(Z)(a)c(=O)[nH]c(a)=O | -1.09721 | -2.79 | -1.45494 |
| Frag146 | disubst.pyrimidin-dione | Zaaa | n(Z)(a)c(=O)n(a)c(a)=O | -0.0664098 | -1.91 | -0.31131 |
| Frag147 | FragK | AAaa | AN(A)n(a)c(a)=O | -0.0620699 | -3.297 | -0.48481 |
| Frag148 | 1-amino-2-pyridone | aa | an([NH2])c(a)=O | -0.842206 | -1.6 | -1.04736 |
| Frag149 | Tetrazolyl | Ya | n1(Y)annn1 | 0.174292 | -1.77 | -0.05266 |
| Frag150 | 2_3_4_trisubst_urazole | ZZa | n1(Z)n(Z)c(=O)n(a)c1=O | -0.708604 | -2.207 | -0.99159 |
| Frag151 | 2-tetrazolyl | Aa | An1nann1 | -0.651707 | -1.65 | -0.86327 |
| Frag152 | 2-tetrazolyl | Za | n1(Z)nann1 | -0.0943726 | -1.65 | -0.30594 |
| Frag153 | 2-pyrimidinone | aZa | a[n&X2]c(=O)n(Z)a | 0.643812 | -3.12 | 0.243766 |
| Frag154 | Triazole | aaa | annn(a)a | 0.293345 | -1.25 | 0.13307 |
| Frag155 | Isoxazolyl | aa | a[n&X2]oa | -0.237556 | -0.95 | -0.35937 |
| Frag156 | Isothiazole #1 | aa | a[n&X2]sa | -0.342363 | -0.2 | -0.36801 |
| Frag157 | 134triazinone | a | O=c1[nH]an[nH]1 | 1.36702 | -1.01 | 1.237518 |
| Frag158 | Aromatic Diazo (TYPE 2) | aa | a[n&X2][n&X2]a | -0.483229 | -2.16 | -0.76018 |
| Frag159 | Diazole-N-subst. | aaa | a[n&X2]n(a)a | -0.0462078 | -1.1 | -0.18725 |
| Frag160 | Diazole-N-subst. | aYa | a[n&X2]n(Y)a | -0.0872313 | -1 | -0.21545 |
| Frag161 | Diazole-N-subst. | aAa | a[n&X2]n(A)a | -0.289211 | -1.69 | -0.5059 |
| Frag162 | Diazole-N-subst. | aZa | a[n&X2]n(Z)a | -0.363835 | -1.69 | -0.58053 |
| Frag163 | Aromatic NH | aa | a[nH]a | -0.0163266 | -0.68 | -0.10352 |
| Frag164 | Aromatic oxygen | aa | a[o&X2]a | 0.108018 | -0.11 | 0.093914 |
| Frag165 | Thiophenyl | aa | a[s&X2]a | -0.0664319 | 0.36 | -0.02027 |
| Frag166 | Aromatic_nitrogen_TYPE2 | aa | a[n&X2]a | 0.031963 | -1.14 | -0.11421 |
| Frag167 | Aliphatic carbon | | [C;!$(*=,#[!#6])] | 0.0163697 | 0.195 | 0.041373 |
| Frag168 | Aromatic carbon | | [c;!$(*=,#[!#6])] | 0.0435754 | 0.13 | 0.060244 |
| Frag169 | NH-Amide | ZZ | Z[NH]C(Z)=O | | | |

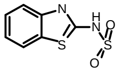
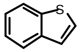
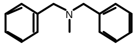
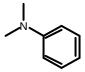
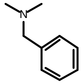
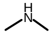
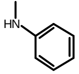
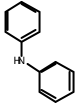
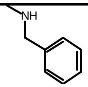
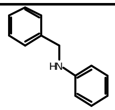
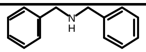
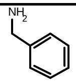
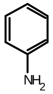
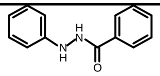
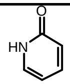
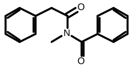
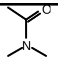
| | | | | | | |
|---|---|---|---|---|---|---|
| Frag170 | Tertiary Imine | aAa | aN=C(A)a | | | |
| CorrFrag1 | Aliphatic ring | | | 0.0503823 | | |
| CorrFrag2 | Trifluoromethyl | | C(F)(F)F | 0.0649662 | | |
| CorrFrag3 |  | a | aS(=O)(=O)[NH]c1sc2ccccc2n1 | 1.24005 | | |
| CorrFrag4 |  | | s1ccc2ccccc12 | 0.528395 | | |

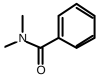
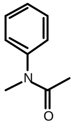
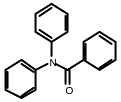
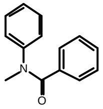
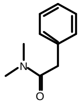
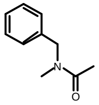
Table 35: Fitting 170 fragments and 4 correction factors to the solubility data of 2473 drug-like compounds in 81 congeneric series. $b_i$ is the coefficient of the fragments described in the equation of $Log1/S_0 = 0.12822 * C\log P + \sum_{i=1}^{174} b_i * frag_i + \sum_{i=1}^{n=81} c_i * f_{series,i} + 3.81803$. The last two fragments Z[NH]C(Z)=O and aN=C(A)a are major components of scaffolds, thus, their fragmental constants were statistically not well validated.

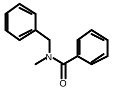
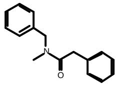
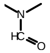
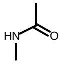
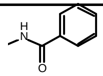
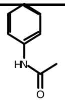
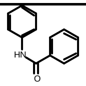
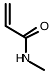
| Nr. | $c_i$ | Nr. | $c_i$ | Nr. | $c_i$ | Nr. | $c_i$ |
|---|---|---|---|---|---|---|---|
| Scaffold1 | -0.82929 | Scaffold21 | -0.92209 | Scaffold41 | -0.36708 | Scaffold61 | 0.061132 |
| Scaffold2 | -0.36232 | Scaffold22 | -0.28748 | Scaffold42 | 0.85059 | Scaffold62 | -0.37454 |
| Scaffold3 | -0.88836 | Scaffold23 | 0.016066 | Scaffold43 | -0.0925 | Scaffold63 | 0.117622 |
| Scaffold4 | -0.97198 | Scaffold24 | -0.38336 | Scaffold44 | -0.61775 | Scaffold64 | -0.64903 |
| Scaffold5 | -0.22905 | Scaffold25 | 0.977961 | Scaffold45 | 0.138562 | Scaffold65 | 0.170775 |
| Scaffold6 | -0.01408 | Scaffold26 | -0.43312 | Scaffold46 | -0.946 | Scaffold66 | 0.041265 |
| Scaffold7 | -0.39203 | Scaffold27 | -0.17257 | Scaffold47 | -1.06197 | Scaffold67 | -0.36912 |
| Scaffold8 | -0.24554 | Scaffold28 | -0.3399 | Scaffold48 | -0.18632 | Scaffold68 | -0.52412 |
| Scaffold9 | -0.21366 | Scaffold29 | -1.97686 | Scaffold49 | -0.18391 | Scaffold69 | 0.506033 |
| Scaffold10 | 0.381965 | Scaffold30 | -0.33184 | Scaffold50 | 0.192332 | Scaffold70 | 1.08843 |
| Scaffold11 | -0.26606 | Scaffold31 | -0.38843 | Scaffold51 | 0.360716 | Scaffold71 | 0.230182 |
| Scaffold12 | 0.032504 | Scaffold32 | -0.45904 | Scaffold52 | 0.833598 | Scaffold72 | -0.17462 |
| Scaffold13 | -0.51988 | Scaffold33 | -0.06644 | Scaffold53 | 1.04617 | Scaffold73 | -0.07553 |
| Scaffold14 | -0.80882 | Scaffold34 | -0.0866 | Scaffold54 | -0.15332 | Scaffold74 | 2.09032 |
| Scaffold15 | -3.35E-05 | Scaffold35 | 0.521279 | Scaffold55 | -0.2185 | Scaffold75 | -2.2763 |
| Scaffold16 | 0.430987 | Scaffold36 | 1.31949 | Scaffold56 | -0.22962 | Scaffold76 | 0.316835 |
| Scaffold17 | -0.17408 | Scaffold37 | -0.11513 | Scaffold57 | -0.08548 | Scaffold77 | -0.26243 |
| Scaffold18 | -0.22399 | Scaffold38 | -0.30322 | Scaffold58 | -0.69876 | Scaffold78 | -0.10436 |
| Scaffold19 | -0.3978 | Scaffold39 | 0.058895 | Scaffold59 | 0.143422 | Scaffold79 | -0.44006 |
| Scaffold20 | -0.06641 | Scaffold40 | -0.21242 | Scaffold60 | -0.95461 | Scaffold80 | 0.185193 |
|  |  |  |  |  |  | Scaffold81 | -0.21854 |

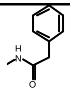
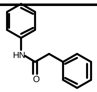
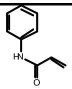
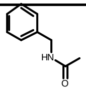
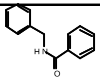
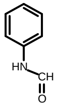
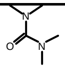
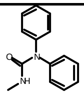
Table 36: Fitting 170 fragments and 4 correction factors to the solubility data of 2473 drug-like compounds in 81 congeneric series. $c_i$ is the coefficient of the congeneric series indices described in the equation of $\mathbf{Log1/S_0 = 0.12822 * C\log P + \sum_{i=1}^{174} b_i * frag_i + \sum_{i=1}^{n=81} c_i * f_{series,i} + 3.81803}$.

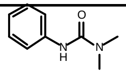
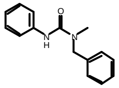
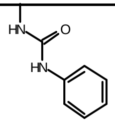
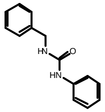
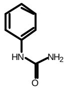
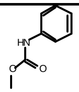
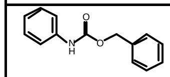
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

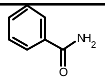
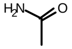
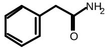
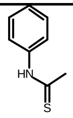
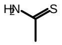
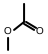
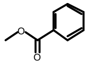
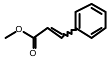
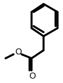
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | $-NH_2$ | | | | | | | |
| |  | | | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | | |  | | | | | |
| | | | | |  | | | | |

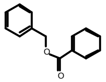
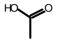
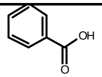
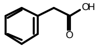
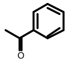
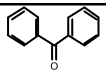
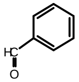
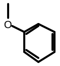
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | |  | | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |
| | | | | | |  | | | |
| | | | |  | | | | | |

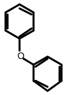
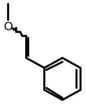
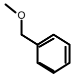
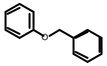
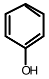
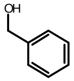
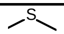
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | |  | | | | | |
| | | | |  | | | | | |
| | | | | |  | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | | | | |  | | | |

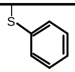
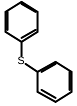
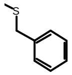
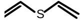
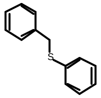
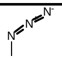
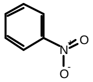
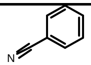
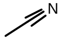
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | | |  | | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |
| | | | | |  | | | | |
| | | |  | | | | | | |

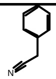
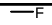
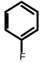
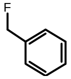
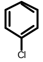
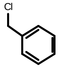
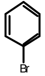
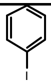
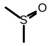
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | | |  | | | | | |
| | |  | | | | | | | |
|  | | | | | | | | | |

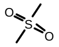
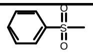
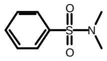
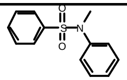
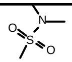
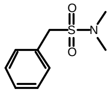
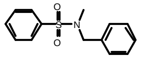
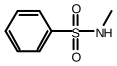
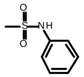
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | |  | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | | |  | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |

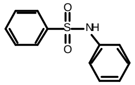
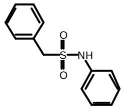
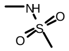
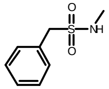
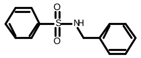
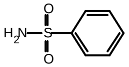
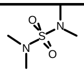
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | |  | | | |
| | | |  | | | | | | |
| |  | | | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | | | | | | | | |
| | | |  | | | | | | |
| | | | | | | | | | |
| | | |  | | | | | | |

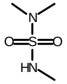
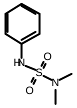
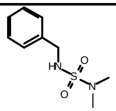
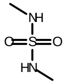
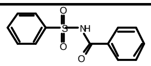
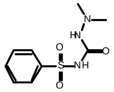
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | |  | | | | | | | |
| | | | | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |

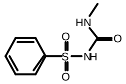
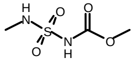
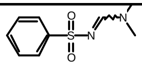
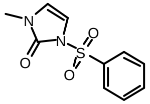
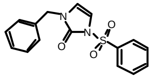
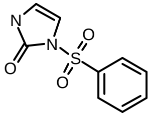
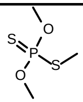
|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

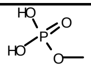
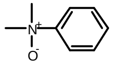
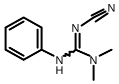
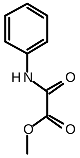
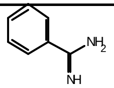
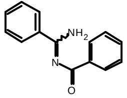
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| |  | | | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | |  | | | | | | | |
| | |  | | | | | | | |
| | |  | | | | | | | |
| | | | | |  | | | | |

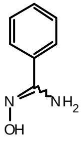
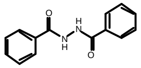
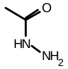
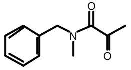
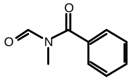
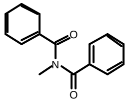
| | | | |  | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | |  | | | | | | | |
| | | | |  | | | | | |
| | |  | | | | | | | |
| | | | |  | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | | | | | | |  |

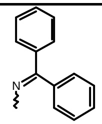
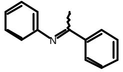
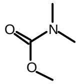
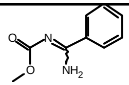
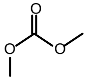
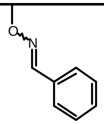
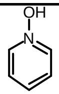
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | | | | |  | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | | |  | | | | |

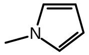
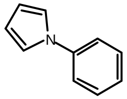
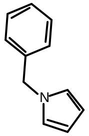
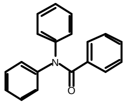
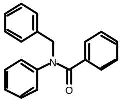
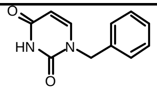
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | |  | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | |  | | | | | | | |
| | | |  | | | | | | |
| | | | | | | |  | | |
| | | | | | | |  | | |
| | | | |  | | | | | |

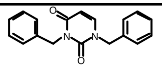
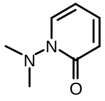
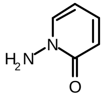
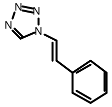
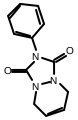
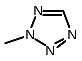
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | | |  | | | |
| | | | |  | | | | | |
| | | | | |  | | | | |
| |  | | | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |

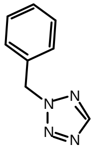
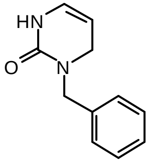
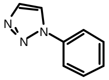
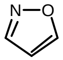
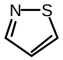
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | | | |  | | |
| | | | | | |  | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| |  | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | |  | | | | | | | |
| | | | |  | | | | | |
| | |  | | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | | | |  | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| |  | | | | | | | | |
| | | | | |  | | | | |

| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

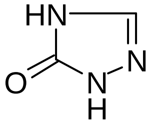
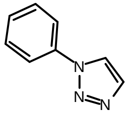
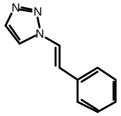
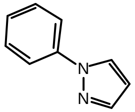
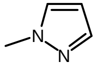
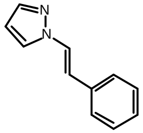
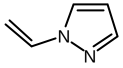
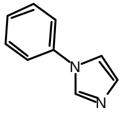
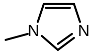
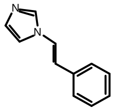
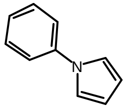
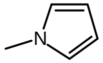
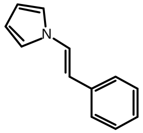
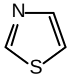
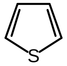
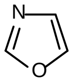
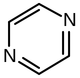
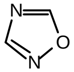
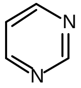
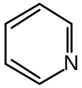
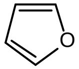
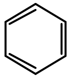
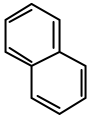
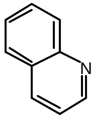
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| |  | | | | | | | | |

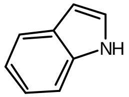
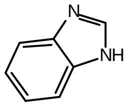
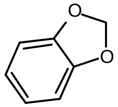
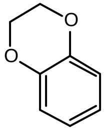
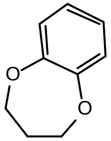
Table 37: Fragments are scaled according to its contribution to the $\mathrm{LogS}_0$. Solubility increases with higher $\mathrm{LogS}_0$ value.

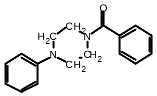
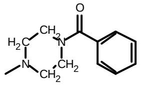
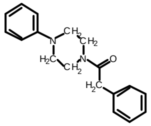
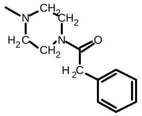
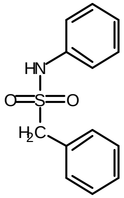
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |

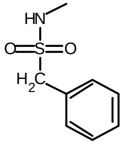
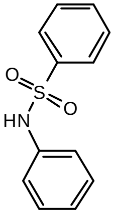
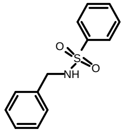
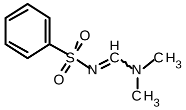
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | 0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |

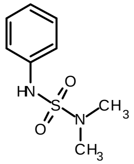
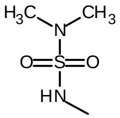
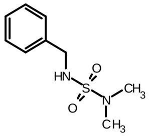
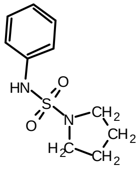
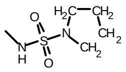
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |

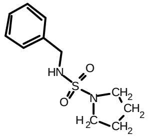
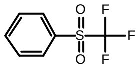
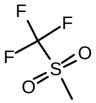
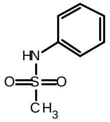
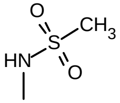
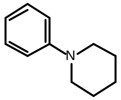
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |

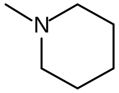
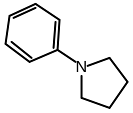
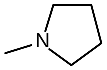
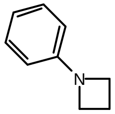
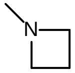
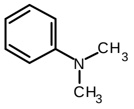
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

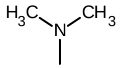
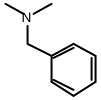
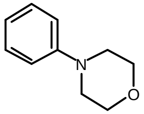
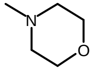
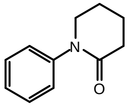
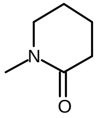
|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

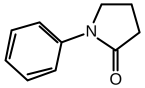
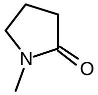
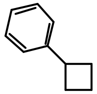
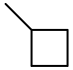
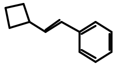
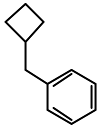
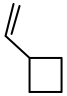
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | | |  | | | | | |

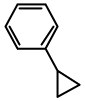
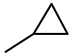
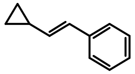
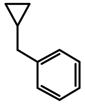
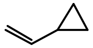
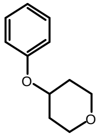
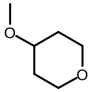
|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

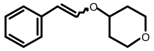
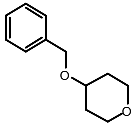
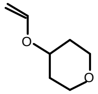
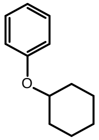
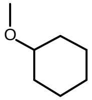
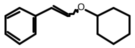
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | | | |  | | | | |
| | | | | | | | | |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | | |  | | | |
| | |  | | | | | | | |

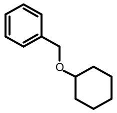
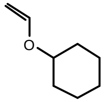
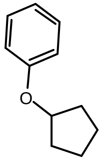
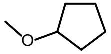
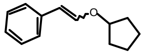
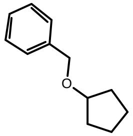
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | |  | | | | | | | |

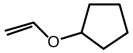
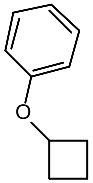
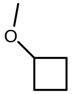
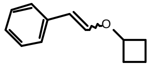
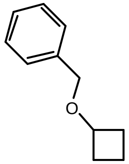
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | | |  | | | | | |
| | | | | |  | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |

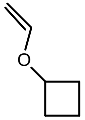
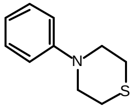
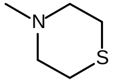
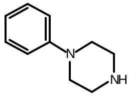
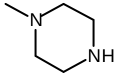
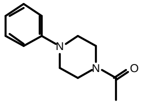
|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

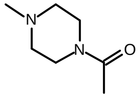
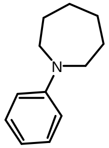
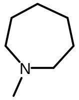
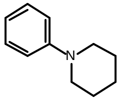
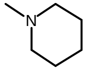
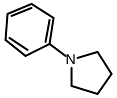
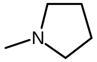
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

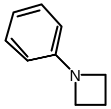
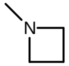
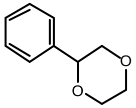
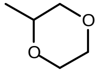
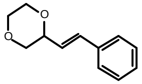
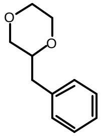
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

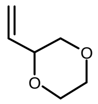
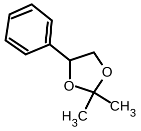
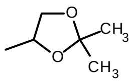
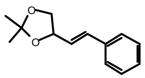
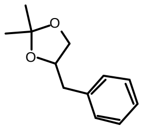
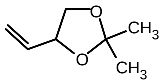
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | |  | | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |

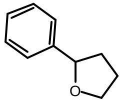
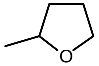
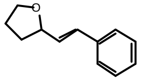
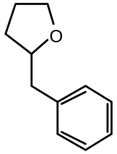
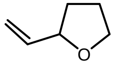
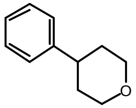
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | |  | | | | | | | |

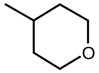
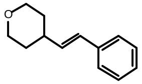
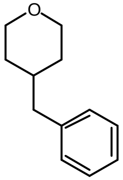
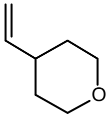
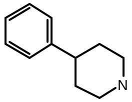
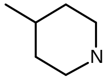
|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |

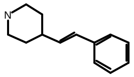
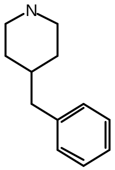
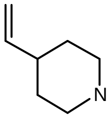
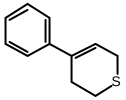
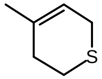
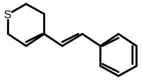
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | |  | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |

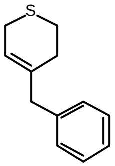
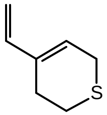
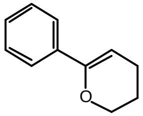
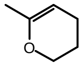
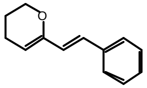
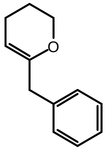
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

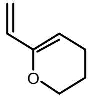
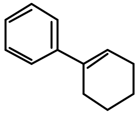
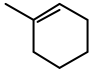
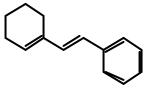
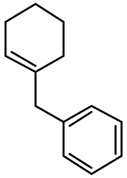
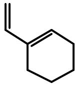
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |

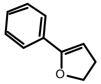
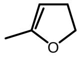
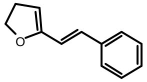
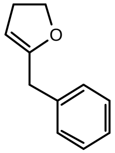
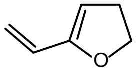
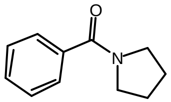
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| | | | |  | | | | | |
| | |  | | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | |  | | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |

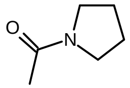
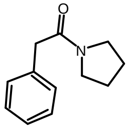
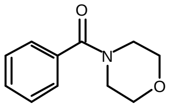
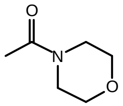
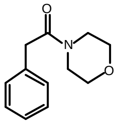
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | |  | | | | | | | |
| | | |  | | | | | | |

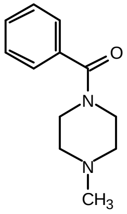
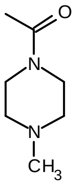
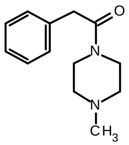
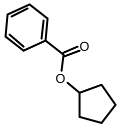
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| -2. | -1.5 | -1 | -0.5 | 0. | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | |  | | | | | | | |
| |  | | | | | | | | |

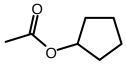
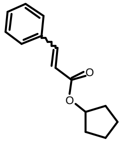
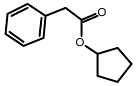
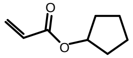
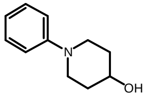
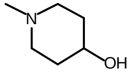
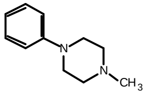
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| | |  | | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | |  | | | | | | | |

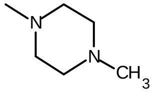
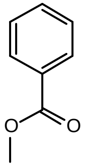
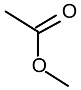
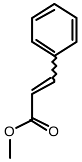
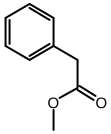
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

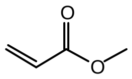
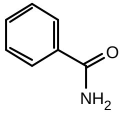
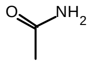
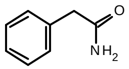
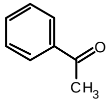
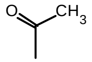
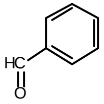
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

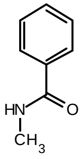
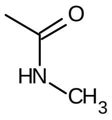
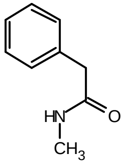
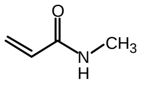
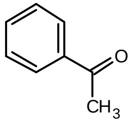
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | |  | | | | |
| | | | | | |  | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | |  | | | | | |
| | | | | | |  | | | |
| | | | | |  | | | | |

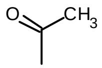
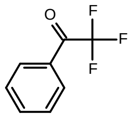
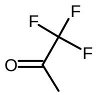
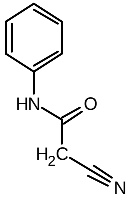
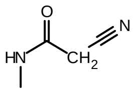
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | | | |  | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | |  | | | | |
| | | |  | | | | | | |

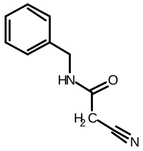
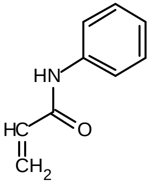
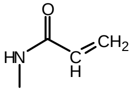
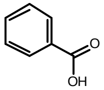
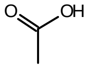
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | |  | | | | | | | |

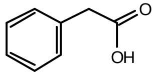
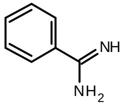
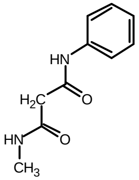
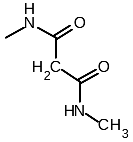
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |

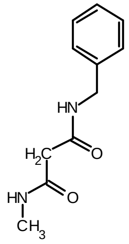
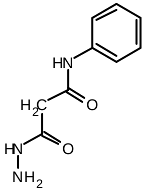
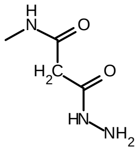
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | | |  | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | |  | | | | |
| | | |  | | | | | | |

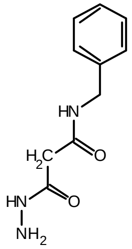
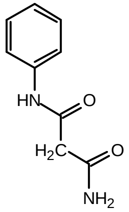
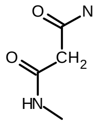
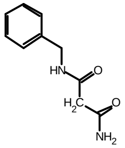
|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |

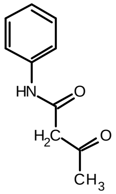
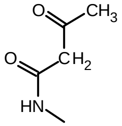
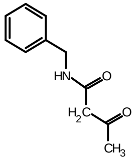
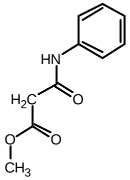
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | |  | | | | |
| | | |  | | | | | | |
| | | | | | |  | | | |
| | | | |  | | | | | |
| | | | | |  | | | | |

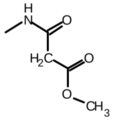
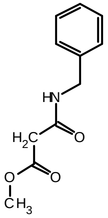
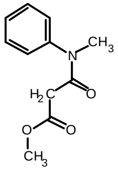
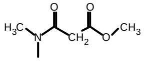
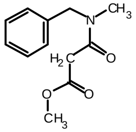
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | |  | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | |  | | | | | |
| | | | | |  | | | | |
| |  | | | | | | | | |
| | | |  | | | | | | |

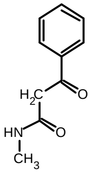
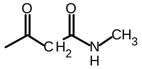
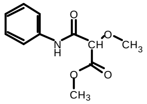
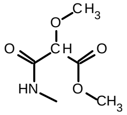
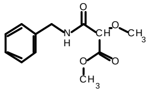
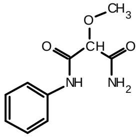
|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

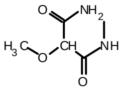
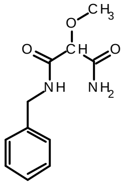
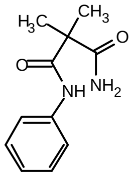
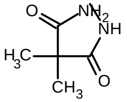
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | |  | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | |  | | | | |

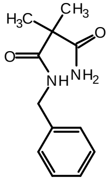
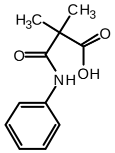
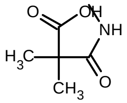
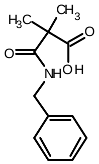
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | |  | | | | |
| | | | | |  | | | | |
| | | | | | |  | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | | |  | | | |

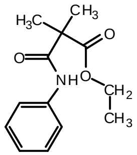
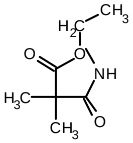
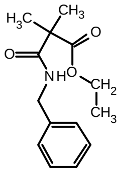
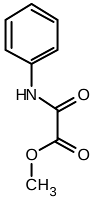
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | |  | | | | |
| | | | | | |  | | | |
| | | | | | |  | | | |
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |

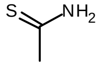
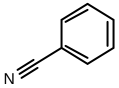
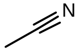
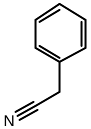
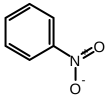
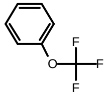
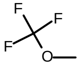
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | |  | | | | |
| | | | | |  | | | | |
| | | | | |  | | | | |
| | | | | | |  | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |

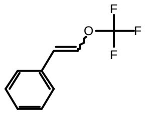
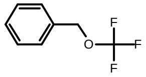
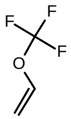
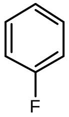
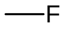
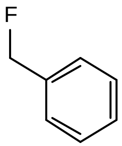
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | | | | |  | | | |
| | | | | |  | | | | |
| | | | | | |  | | | |
| | | | | | |  | | | |
| | | | | |  | | | | |

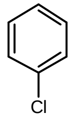
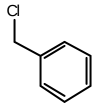
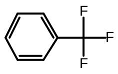
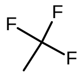
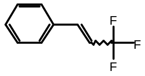
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

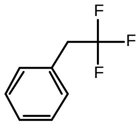
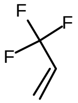
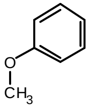
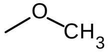
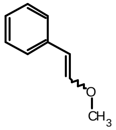
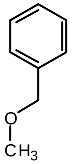
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | |  | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | |  | | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |

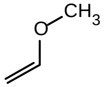
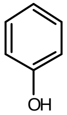
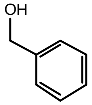
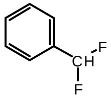
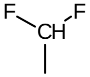
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | | | | |  | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | |  | | | | |
| | | |  | | | | | | |

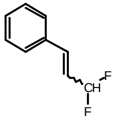
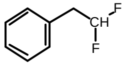
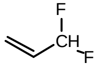
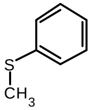
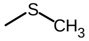
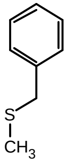
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| | |  | | | | | | | |
| | | | |  | | | | | |
| |  | | | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | | |  | | | | |

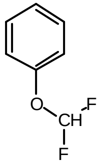
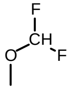
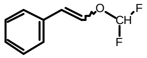
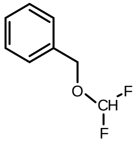
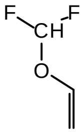
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | |  | | | | | |
| | | |  | | | | | | |

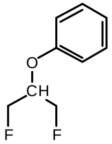
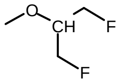
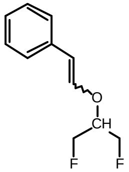
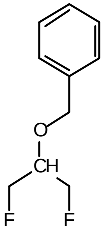
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | |  | | | | | | | |
| | | | —Cl | | | | | | |
| | |  | | | | | | | |
| | |  | | | | | | | |
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |

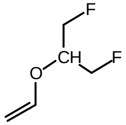
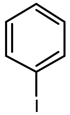
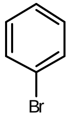
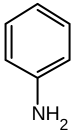
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |

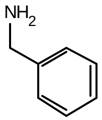
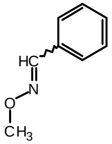
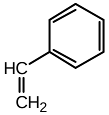
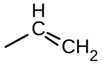
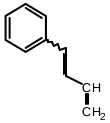
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | —OH |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

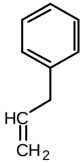
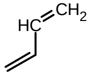
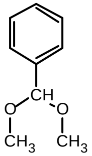
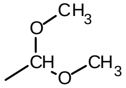
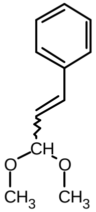
|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

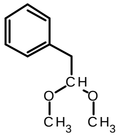
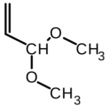
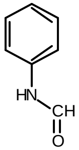
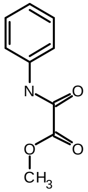
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | |  | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | | |  | | | | | |

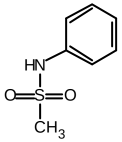
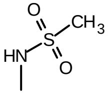
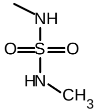
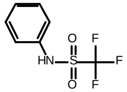
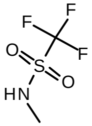
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | | |  | | | | | |
| | | |  | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | |  | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | |  | | | | | | | |
| | |  | | | | | | | |
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | —NH$_2$ | | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| |  | | | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | |  | | | | | | |
| | |  | | | | | | | |
| | | | | |  | | | | |
| | | | |  | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | |  | | | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | |  | | | | |
| | | |  | | | | | | |
| | | |  | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | |  | |
| | | | | | |  | | | |
| -2 | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 |
| | | | | | |  | | | |
| | | | | | | | |  | |
| | | | | | | |  | | |

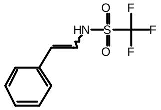
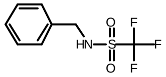
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | |  | | | | | | |
| | | |  | | | | | | |

Table 38: Fragments are scaled according to its contribution to the $\mathrm{LogS}_0$. Solubility increases with higher $\mathrm{LogS}_0$ value.

# Curriculum Vitae

| | |
|---|---|
| 1977 | Born on June 11 in Shanghai, China |
| 1995 | University-entrance diploma (‘Abitur’) at the Hongkou Comprehensive Secondary School, Shanghai, China |
| 1995-1996 | Applied Chemistry at the Tongji-University, Shanghai, China |
| 1996-1997 | Intensive German course at the Studienkolleg of Technical University Darmstadt, Germany |
| 1997-2002 | Chemistry study at Technical University Darmstadt, Germany |
| 1999 | Intermediate diploma in chemistry |
| 2002 | Diploma in chemistry |
| 2003-2006 | Ph.D. studies in pharmaceutical chemistry under the direction of Prof. Michael Wiese at the Department of Pharmacy, Rheinische Friedrich Wilhelms University Bonn, Germany. The studies were performed at F. Hoffmann-La Roche Ltd. In Basel, Switzerland, under the direction of Dr. Manfred Kansy. |
| 2006 | Final examination to obtain the degree of Doctor of Natural Sciences, Rheinische Friedrich Wilhelms University Bonn, Germany |