



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Pharmacy 44*

Computational Analysis of Aqueous Drug Solubility – Influence of the Solid State

CAROLA WASSVIK



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006

ISSN 1651-6192
ISBN 91-554-6728-8
urn:nbn:se:uu:diva-7334

Dissertation presented at Uppsala University to be publicly examined in B41, Uppsala Biomedicinska Centrum BMC, Husargatan 3, Uppsala, Friday, December 8, 2006 at 13:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English.

Abstract

Wassvik, C. 2006. Computational Analysis of Aqueous Drug Solubility – Influence of the Solid State. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy* 44. 66 pp. Uppsala. ISBN 91-554-6728-8.

Aqueous solubility is a key parameter influencing the bioavailability of drugs and drug candidates. In this thesis computational models for the prediction of aqueous drug solubility were explored. High quality experimental solubility data for drugs were generated using a standardised protocol and models were developed using multivariate data analysis tools and calculated molecular descriptors. In addition, structural features associated with either solid-state limited or solvation limited solubility of drugs were identified.

Solvation, as represented by the octanol-water partition coefficient ($\log P$), was found to be the dominant factor limiting the solubility of drugs, with solid-state properties being the second most important limiting factor.

The relationship between the chemical structure of drugs and the strength of their crystal lattice was studied for a dataset displaying $\log P$ -independent solubility. Large, rigid and flat molecules with an extended ring-structure and a large number of conjugated π -bonds were found to be more likely to have their solubility limited by a strong crystal lattice than were small, spherically shaped molecules with flexible side-chains.

Finally, the relationship between chemical structure and drug solvation was studied using computer simulated values of the free energy of hydration. Drugs exhibiting poor hydration were found to be large and flexible, to have low polarisability and few hydrogen bond acceptors and donors.

The relationship between the structural features of drugs and their aqueous solubility discussed in this thesis provide new rules-of-thumb that could guide decision-making in early drug discovery.

Keywords: intrinsic solubility, solubility prediction, drug solubility, solid state, melting point, enthalpy of melting, entropy of melting, solvation, free energy of hydration, QSPR, multivariate analysis, PCA, PLS

Carola Wassvik, Department of Pharmacy, Box 580, Uppsala University, SE-75123 Uppsala, Sweden

© Carola Wassvik 2006

ISSN 1651-6192

ISBN 91-554-6728-8

urn:nbn:se:uu:diva-7334 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7334>)

To my familiy

Papers discussed

This thesis is based on the following Papers, which will be referred to by the Roman numerals assigned below:

- I Bergström, C.A.S.; Wassvik, C.M.; Norinder, U.; Luthman, K. and Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. Reproduced with permission from *Journal of Chemical Information and Computer Sciences*, **2004**, 44(4), 1477-1488. © 2004 American Chemical Society.
- II Wassvik, C.M.; Holmén, A.G.; Bergström, C.A.S.; Zamora I. and Artursson, P. Contribution of Solid-State Properties to the Aqueous Solubility of Drugs. Reprinted from *European Journal of Pharmaceutical Science*, **2006**, 29(3-4), 294-305. © 2006, with permission from Elsevier.
- III Wassvik, C.M.; Holmén, A.G.; Draheim, R. and Artursson, P. Log*P*-Independent Solubility of Drugs – Molecular Descriptors for Solid-State Limited Solubility. *Submitted*.
- IV Wassvik, C.M.; Holmén, A.G. and Artursson, P. Free Energy of Hydration of Drugs – Molecular Descriptors for Solvation Limited Solubility. *In manuscript*.

Contents

1. Introduction.....	11
1.1. Solubility in drug discovery and development.....	11
1.2. Thermodynamics of dissolution of solids	13
1.2.1. Estimation of solubility from thermodynamic properties	13
1.2.2. Computational estimation of thermodynamic properties.....	17
1.3. Experimental estimation of drug solubility	20
1.3.1. Definition of intrinsic solubility	20
1.3.2. Factors influencing solubility experiments.....	21
1.3.3. Kinetic methods.....	23
1.3.4. Thermodynamic methods	23
1.4. Computational estimation of drug solubility	24
1.4.1. Model development – QSPR	24
1.4.2. Dataset selection	25
1.4.3. Molecular descriptors	25
1.4.4. Experimental data	26
1.4.5. Statistical and mathematical methods.....	27
1.4.6. Available computational solubility models	28
2. Aims of the thesis.....	30
3. Materials and methods	31
3.1. Selection of dataset.....	31
3.2. Structural diversity and drug-likeness	31
3.3. Chemicals and drugs	32
3.4. Crystal structures	32
3.5. Differential scanning calorimetry (DSC)	32
3.6. Solubility determinations – the shake-flask method	33
3.7. Hydration free energy calculations.....	33
3.8. Molecular descriptor generation.....	34
3.8.1. 2D – Selma and Molconn-Z	34
3.8.2. 3D – Surface area descriptors and VolSurf	34
3.9. Statistical analysis	35
3.9.1. Linear regression	35
3.9.2. Multivariate analysis.....	35
4. Results and discussion	37

4.1. Training drug solubility models	37
4.1.1. Quality of experimental data (Papers I-III).....	37
4.1.2. Diversity and drug-likeness (Papers I, II and IV)	38
4.2. Validating drug solubility models (Paper I)	39
4.3. Global versus local models (Paper I)	41
4.4. Experimental properties influencing drug solubility	42
4.4.1. Solvation properties – $\log P$ (Paper II)	42
4.4.2. Solid-state properties (Papers II and III).....	43
4.5. Application of semi-empirical equations on drugs.....	45
4.5.1. The general solubility equation (GSE) (Paper II)	45
4.5.2. The Dannenfelser equation (Paper II).....	47
4.6. Molecular descriptors for drug solubility	48
4.6.1. Two dimensions or three? (Papers I and IV)	48
4.6.2. The solid state – $\log P$ -independent solubility (Paper III).....	49
4.6.3 Solvation – The free energy of hydration (Paper IV)	52
5. Conclusions.....	54
6. Acknowledgements.....	56
7. References and notes.....	58

Abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
AAE	Average absolute error
ADMET	Absorption, distribution, metabolism, elimination/excretion, toxicity
A_{MS}	The molecular surface area
ANN	Artificial neural network
BCS	Biopharmaceutics classification system
ChemGPS	Chemical global positioning system
CLOGP	Calculated log P
ΔH_m	Change in enthalpy of melting
DSC	Differential scanning calorimetry
ΔH_{sub}	Change in enthalpy of sublimation
ΔG	Change in Gibbs' free energy
ΔG_{cav}	ΔG due to cavitational forces
ΔG_{ele}	ΔG due to electrostatic forces
ΔG_{hyd}	Change in free energy of hydration
ΔG_{int}	ΔG due to interaction forces
ΔS_m	Change in entropy of melting
ΔG_{vdw}	ΔG due to van der Waals forces
E_{LJ}	Lennard-Jones interaction energy
E_C	Coulomb interaction energy
F	Test statistic from the t-test
FL	Fluorescence (detection)
γ	Gamma – surface tension
GSE	General solubility equation
HPLC	High Pressure Liquid Chromatography
log P	Octanol-water partition coefficient (the logarithm of)
M	Molar
MC	Monte Carlo
MD	Molecular dynamics
MLR	Multiple linear regression

MM	Molecular mechanics
MMFF	Merck molecular force field
MS	Mass spectrometry (detection)
N	Number of observations
NPSA	Non-polar surface area
NTP	Normal temperature and pressure
PCA	Principal components analysis
pH	Negative logarithm of the proton concentration (M)
ϕ	Phi – the flexibility number
pK_a	Acid dissociation constant
PLS	Projection to latent structures by means of partial least squares
PSA	Polar surface area
PTSA	Partitioned total surface area
π_2^H	Polarisability / dipolarity
Q^2	Cross-validated R^2
QM	Quantum mechanics
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
r	Correlation coefficient
R	The universal gas constant
R_2	The excess molar refractivity
R^2	Coefficient of determination
RMSE	Root mean square error
rpm	Revolutions per minute
σ	Sigma – the rotational symmetry number
s	Standard deviation
S	Molar solubility in water; aqueous solubility
S_0	Intrinsic solubility
SA	Surface area
SE	Standard error
S_{pH}	Total solubility at a given pH
S^O	Aqueous solubility of the supercooled liquid
S^S	Aqueous solubility of the solid
$\Sigma\alpha_2^H$	Summation of hydrogen bond acidity
$\Sigma\beta_2^H$	Summation of hydrogen bond basicity
T_m	Melting point
TSA	Total surface area
UV	Ultra violet (detection)
V_x	McGowen characteristic volume

1. Introduction

1.1. Solubility in drug discovery and development

The number of candidate drugs in pharmaceutical development that make it to the market has markedly decreased over the last few decades. The average success rate for new compounds (from first-in-man to registration) was only one in nine compounds for all therapeutic areas during a ten year period (1991-2000)¹. This contributes to the staggering expenditure in the pharmaceutical industry. The cost of the discovery and development of one drug was estimated to be in the order of \$804 million in 2001². In 1991 the most important cause of compound attrition in clinical development was related to inadequate pharmacokinetic profiles and poor bioavailability (accounting for 40% of attrition). Since adequate solubility is a prerequisite for drug absorption from the gastrointestinal tract, it plays a significant role for the resulting bioavailability of orally administered drugs. The figure of 40% had dropped to 10% in the year of 2000, indicating an improvement in the pharmacokinetic properties of the compounds being put forward as candidate drugs. However, judging from the increase in the scientific literature on the subject of poor solubility – the search string “poorly soluble” generated 80 hits in 1995, 162 hits in 2000 and 261 hits in 2005 on PubMed – poor solubility clearly constitute an important issue in contemporary drug discovery and development.

With candidate drugs in development becoming increasingly poorly soluble^{3,4}, formulators are presented with considerable technical challenges⁵. With the purpose of facilitating decision making in drug development, the biopharmaceutics classifications system (BCS), was designed by Amidon *et al.*⁶. The BCS correlates the *in vitro* drug solubility and permeability to the *in vivo* bioavailability and it applies the following four classes to candidate drugs: (I) high solubility and high permeability, (II) low solubility and high permeability, (III) high solubility and low permeability and (IV) low solubility and low permeability (Fig. 1). Characteristically, modern candidate drugs belong to Class II or IV of the BCS. The bioavailability of class II compounds can potentially be improved by the development of new, sophisticated and often expensive formulation designs, while class IV compounds are most likely returned to the lead optimisation phase for improvement of the physicochemical properties⁵.

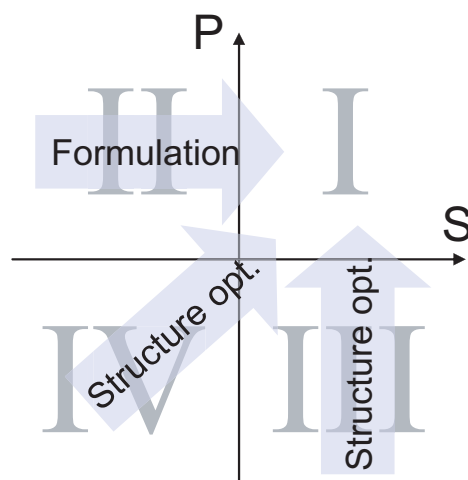


Figure 1. Improvement strategies based on BCS Classification. The solubility (S) increases in the direction of the x-axis, while the permeability (P) increases in the direction of the y-axis. Modified from reference⁵.

A more appealing approach would be to address the issue of poor solubility already in the early stages of drug discovery through *in vitro* (experimental) and *in silico* (computational) screening, in line with the “fail early, fail cheaply” principle. Typically, the drug discovery process goes through the stages of (i) *target identification* – in which the mechanisms responsible for a disease state are investigated in detail and one or several proteins are identified and validated as suitable targets, (ii) *hit identification* – in which large company libraries or smaller focused libraries are screened to find compounds that exert pharmacological activity towards the selected target, (iii) *hit-to-lead* – in which the hits are explored with regards to potency, selectivity and absorption, distribution, metabolism, excretion and toxicity (ADMET) properties with the intention being to prioritise and select a few lead series, (iv) *lead optimization* – in which pharmacological activity, physicochemical and ADMET properties of lead compounds are optimised to produce up to a few candidate drugs that can enter into the development phase. Computational⁷ and experimental⁸ methods are used in parallel throughout all of the Phases (i) to (iv) with an emphasis being placed on computational methods in the earlier stages and experimental in the later stages.

Rules and cut-off values are assigned for optimal potency, physicochemical properties, pharmacokinetic profiles and toxicity with the intention of identifying lead compounds that are not only active, but also exhibit adequate bioavailability and are free from serious side-effects⁹. An example of a simple set of such rules was provided by Lipinski’s “rule of 5”¹⁰, stating that a compound is likely to exhibit poor absorption if it has >5 hydrogen bond

donors, >10 hydrogen bond acceptors, the molecular weight (Mw) is >500 g mol⁻¹ and the calculated log P (CLOGP) >5. A more stringent rule for “lead-likeness” was proposed by Teague *et al.* stating that Mw <350 g mol⁻¹, CLOGP <3 and affinity ~0.1 μM are more appropriate cut-offs for successful leads since the lead optimisation process often results in larger and more lipophilic molecules¹¹. The identification of chemical features associated with drugs, drug-like compounds and leads is important to improve the drug discovery process¹²⁻¹⁸.

In the pursuit of lead compounds with the desired pharmacokinetic profiles there has been an increasing demand for computational models for the prediction of ADMET properties directly from chemical structure. As a result, a range of software for this purpose is now commercially available⁷. The predictive ability of these models for pharmaceutically interesting compounds has been questioned. Two important reasons for their lack of predictivity have been identified. Firstly, the experimental data used for model development is not of sufficiently high quality resulting in the introduction of error of unknown size in the models¹⁹. Secondly, the compounds used as training sets in the model development do not reflect the compounds that are to be predicted²⁰.

To find predictive computational models for ADMET properties that are free from the above-mentioned limitations presents a challenge for people involved in pharmaceutical research. This challenge was the driving force for the start of the work undertaken in this thesis. Focus was set on aqueous solubility.

1.2. Thermodynamics of dissolution of solids

1.2.1. Estimation of solubility from thermodynamic properties

The dissolution process was divided into three steps to simplify the examination of the contribution from the different interaction energies^{21,22}. If this division is applied to the dissolution of solids in water, the first step would represent the removal of one molecule from the crystal lattice, the second step the formation of a cavity in the water large enough to accommodate the molecule, and the third step the transfer of the molecule into the water (Fig. 2). Steps one and two would be energetically unfavourable, since they involve the breaking of intermolecular bonds in the crystal and the water, respectively; but energy is gained in the third step from favourable interactions between the solute (i.e. the molecule) and the water. For dissolution to occur, the energy gained in the third step has to be numerically larger than the energetic cost of steps one and two.

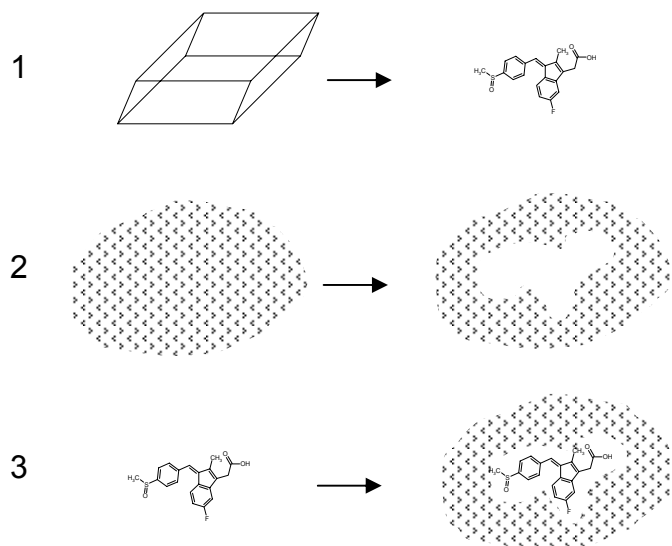


Figure 2. The dissolution process. 1. The removal of a molecule from its crystal lattice. 2. The creation of a cavity in the solvent. 3. The insertion of the molecule into the cavity. Energy is consumed by steps 1 and 2, while energy is gained in step 3. Modified from Fig. 2.16 on p. 47 of Solubility Behaviour of Organic Compounds²³.

From a thermodynamic perspective, there are two alternative routes from the pure crystalline compound to the saturated water solution. On the left-hand side of the schematic diagram (Figure 3), the crystalline solid is melted, cooled down to the temperature of the water to form a supercooled liquid (which is a hypothetical state for solids) and then transferred from the supercooled liquid to the water. On the right-hand side, the compound is removed from the crystal and transferred to the gas phase and then transferred from the gas phase into the water. Each step in the thermodynamic cycle is associated with a change in the Gibbs' free energy. The bottom half of Fig. 3 can be regarded as describing the stability of the solid state, while the top half can be regarded as being related to the solvation. The solubility of a compound depends on the balance between the two and, subsequently, it would be possible to estimate the solubility from a combination of the properties on the bottom half and the properties of top half of the figure. The approach to estimate solubility from experimental properties representing any of the individual steps in Fig. 3 has been taken by several research groups as outlined in Section 1.2.1.1. below. However, the ultimate approach for the estimation of drug solubility lies perhaps with free energy calculations of each of the separate steps that would allow for the estimation of solubility *ab initio* i.e. from first principles.

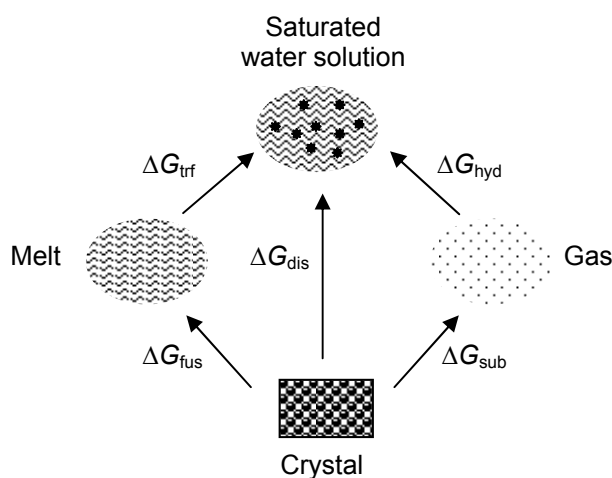


Figure 3. A thermodynamic cycle for dissolution of a crystalline drug. The change in Gibbs' free energy associated with each transition is denoted as follows: fusion (ΔG_{fus}), transfer from the melt to the saturated water solution (ΔG_{trf}), sublimation (ΔG_{sub}), hydration (ΔG_{hyd}) and dissolution (ΔG_{dis}). Modified from Fig. 8.1 on p. 373 of Solubility Behaviour of Organic Compounds²³.

1.2.1.1. Classical solubility models

The oldest rule for solubility is perhaps that of *similia similibus solvuntur* or “like dissolves like” first proposed by medieval alchemists. This rule suggests that a non-polar compound will dissolve in a non-polar solvent, as a result of their similar chemical structure, but not in a polar solvent. The like dissolves like rule can be exemplified by the classical work of Corwin Hansch²⁴ and co-workers from 1968. They correlated the aqueous solubility of 156 organic liquids to their experimental or calculated $\log P$ with excellent results.

$$\log \frac{1}{S} = 1.339 \log P - 0.978 \quad (1)$$

$r = 0.935 \quad s = 0.472$

In Eq. 1, S is the molar solubility in water, P is the octanol-water partition coefficient, r is the correlation coefficient and s is the standard deviation (in log units of S). In 1965, Irmann²⁵ proposed the following equation for the difference in aqueous solubility of solids and their corresponding super-cooled liquids.

$$\log \frac{S^S}{S^O} = -0.0095(T_m - 25) \quad (2)$$

In Eq. 2, S^S is the aqueous solubility of the solid, S^O is the aqueous solubility of the supercooled liquid and T_m is the melting point in °C. The constant -0.0095 arises from the approximation introduced by Walden,²⁶ that the ΔS_m of most organic compounds is close to 54.4 Jmol⁻¹K⁻¹. Equations 1 and 2 were combined by Yalkowsky and Valvani²⁷ in 1980 to form the general solubility equation (GSE), which was later (2001) revised by Jain and Yalkowsky²⁸. In this revised equation (Eq. 3), the T_m term represents the contribution to solubility from the solid state and the $\log P$ term represents the contribution to the solubility from the solvation. Hence, the GSE uses the left-hand route in the thermodynamic cycle in Fig. 3 above. The revised version of the GSE states that;

$$\log S = 0.5 - 0.01(T_m - 25) - \log P \quad (3)$$

where S is the aqueous solubility of a solid organic non-electrolyte, T_m is the melting point in °C and P is the octanol-water partition coefficient. For liquids the melting point term is omitted. Admittedly, the GSE is a simplification of the thermodynamic contributions to the solubility and there are several assumptions associated with it. The most important of which are the assumptions that the solubility is low enough to assume ideal solubility, that the supercooled liquid is completely miscible with *n*-octanol and that the ΔS_m of organic compounds is constant (Walden's rule). In spite of these assumptions, the GSE has proven to be able to predict aqueous solubility of many different classes of organic compounds, including drug-like ones, with good accuracy²⁹.

Abraham and Lee³⁰ proposed the following relationship for the solubility for 659 organic solids and liquids:

$$\begin{aligned} \log S = & 0.518 - 1.004R_2 + 0.771\pi_2^H + 2.168\Sigma\alpha_2^H + 4.238\Sigma\beta_2^H - \\ & 3.362\Sigma\alpha_2^H\Sigma\beta_2^H - 3.987V_x \\ R^2 = & 0.920 \quad s = 0.557 \end{aligned} \quad (4)$$

where S is the aqueous solubility, R_2 is the excess molar refractivity, π_2^H is the dipolarity/polarisability, $\Sigma\alpha_2^H$ is the overall or summation hydrogen bond acidity, $\Sigma\beta_2^H$ is the overall or summation hydrogen bond basicity and V_x is the McGowan characteristic volume³¹. The product $\Sigma\alpha_2^H \times \Sigma\beta_2^H$, reflects the strength of the interactions in the crystal. R_2 and V_x can be calculated from structure, whereas π_2^H , $\Sigma\alpha_2^H$ and $\Sigma\beta_2^H$ were experimental data in Eq.4, but can now be calculated with the program ABSOLV³².

In addition to the above-mentioned equations, the UNIFAC^{33,34}, AQUAFAC³⁵ and Mobile Order Theory^{36,37} provide estimates of the solubility of organic compounds from ΔS_m , T_m or the solubility in a hydrocarbon solvent (e.g. *n*-octane). These methods, the above-mentioned Equations 1-3 included, all require one or more experimental parameters. This limits the utility of these methods in early drug discovery since experimental data for the pure compound is rarely available. One way to overcome this restriction is to calculate these properties from the chemical structure. If the quality of these estimations is satisfactory, the predicted values could replace the experimental ones enabling the estimation of solubility already before a compound is synthesised.

1.2.2. Computational estimation of thermodynamic properties

1.2.2.1. Octanol-water partition coefficient ($\log P$)

Being the, by far, most frequently used property in drug design, $\log P$ has received a great deal of attention and many methods exist for its calculation directly from chemical structure. Five distinct categories of methods have been identified^{38,39}. They are (i) the π -substituent method⁴⁰, (ii) fragment based methods⁴¹, (iii) atomic contribution and/or surface area methods⁴², (iv) molecular properties methods⁴³ and (v) solvatochromatic parameters methods⁴⁴. Being so popular, numerous commercial software packages based on these methods are available for the calculation of $\log P$. The predictive ability of three such programs⁴⁵⁻⁴⁷ was evaluated on a diverse set of 300 drugs and drug-like molecules³⁹. It was found that the programs were equally accurate with a standard deviation of around 0.65 log units and it was pointed out by the authors that the error in the experimental data might be larger than previously suggested. When evaluating calculated $\log P$ for 2569 AstraZeneca⁴⁸ in-house data, root mean square errors (RMSEs) ranged from 0.84-1.20⁴⁹⁻⁵¹ and for 640 legacy Pharmacia compounds the RMSEs obtained⁵² ranged from 1.14-1.46^{49,51,53}. From these investigations it can be concluded that a certain measure of uncertainty (at least 0.5 log units) must be expected when using commercial software for the estimation of $\log P$ and it is reasonable to believe that it will be even larger for new drug-like compounds.

1.2.2.2. Melting properties (T_m), (ΔH_m) and (ΔS_m)

T_m is the temperature at which a phase transition from the solid to liquid form of the pure compound occurs. At this temperature, the solid and liquid phases co-exist at equilibrium. Hence, the “change” in Gibbs’ free energy is zero, which gives us the following relationship for T_m :

$$\Delta G = \Delta H_m - T_m \Delta S_m = 0 \quad (5)$$

$$T_m = \frac{\Delta H_m}{\Delta S_m} \quad (6)$$

In Equations 5 and 6, ΔG is the change in Gibbs' free energy, ΔH_m is the enthalpy of melting (kJmol^{-1}), T_m is the melting point (K) and ΔS_m is the entropy of melting ($\text{Jmol}^{-1}\text{K}^{-1}$). Most models for the estimation of the melting point directly from chemical structure are based on small homologous series of compounds. Only recently have models appeared that are based on structurally diverse sets of organic compounds⁵⁴⁻⁵⁶ and drugs^{57,58}. On average these models estimates the melting point of a diverse set of drugs compiled by Bergström *et al.*⁵⁹ from the MERCK Index with a RMSE of about 40°C. The reason for the lack of accuracy probably lies in the complex (possibly non-linear) relationship between the molecular structure and melting point. Important descriptors identified in the above-mentioned models include measures of the size, shape, polarisability, flexibility and hydrogen bond potential. The ability to form intermolecular hydrogen bonds results in a higher melting point than does intramolecular hydrogen bonds. This feature can be difficult to capture with descriptors derived from the single molecule, rather than from studying the interaction *per se*.

An alternative approach to the estimation of T_m was recently explored by Jain and Yalkowsky⁶⁰. They assessed T_m by making separate estimations of its components ΔH_m and ΔS_m in Eq. 6. The ΔH_m were predicted using a group contribution method and ΔS_m was predicted from the molecular symmetry and flexibility with a semi-empirical equation (Eq. 7) developed by Dannenfelser *et al.*^{61,62}.

$$\Delta S_m = 50 - R \ln \sigma + R \ln \phi \quad (7)$$

In Eq. 7, ΔS_m is the entropy of melting ($\text{Jmol}^{-1}\text{K}^{-1}$), R is the universal gas constant $8.31 \text{ Jmol}^{-1}\text{K}^{-1}$, σ is the rotational symmetry number and ϕ is the flexibility number. Equation 7 estimated the ΔS_m of 1799 organic compounds with an average absolute error (AAE) of $12.3 \text{ Jmol}^{-1}\text{K}^{-1}$ ⁶³. A group contribution method⁶⁴ estimated the ΔS_m for the same dataset with an accuracy of $10.4 \text{ Jmol}^{-1}\text{K}^{-1}$. The T_m of 2230 compounds was predicted from ΔH_m and ΔS_m with an AAE of 30.1°C. Since the AAE is normally a smaller number than the RMSE, this represents more or less the same accuracy as for the other methods for estimation of the melting point.

The accuracy that can be expected from present methods for the estimation of T_m directly from chemical structure is around $\pm 40^\circ\text{C}$ for organic compounds and drugs alike. In most cases, this is not satisfactory for precise predictions of T_m of single compounds. It could, however, provide a means

to apply the GSE (or other methods that include T_m) as a simple filter for a crude estimation of solubility without the need for experimental data.

1.2.2.3. Enthalpy of sublimation (ΔH_{sub})

The lattice energy is the energy that is released when one mole of crystal is formed from one mole of molecules in the gas phase. The enthalpy of sublimation (ΔH_{sub}) is the heat that is absorbed in the reverse process, i.e the energetic cost to remove one mole of molecules from the crystal lattice and transfer it to the gaseous state. The experimental determination of the ΔH_{sub} involves the determination of vapour pressure over the crystal, which is a time-consuming experiment for drugs^{65,66}. As a result of this, very few methods for the prediction of the ΔH_{sub} of drugs exist. There are, however, examples of such models for small organic compounds^{67,68}.

The future challenge in this field lies in the ability to predict the structure of crystals from the molecular structure of a compound. Because of the complexity of large organic compounds and drugs, only crystal structures of small organic, organometallic and inorganic compounds have been successfully solved by this technique⁶⁹. To inspire progress in this field, the Cambridge Crystallographic Data Centre has hosted three blind tests of crystal structure prediction⁷⁰⁻⁷².

1.2.2.4. Free energy of hydration (ΔG_{hyd})

The free energy of solvation can be calculated using high level computer simulations of the solute-solvent system of interest. These simulations are computationally demanding and time consuming and are therefore, not suitable for the screening of large compound libraries. They do, however, offer an excellent means for investigating the solute-solvent interactions more closely for a limited number of compounds. If the solvent is water, the free energy of solvation is termed the free energy of hydration (ΔG_{hyd}). However, the methods described here are applicable to any solute-solvent system. The methods available for the calculation of the free energy of solvation were recently reviewed⁷³. They can be roughly divided into two types, *discrete methods*⁷⁴⁻⁷⁶ and *continuum models*. In the discrete methods, the solvent molecules, as well as solute, are treated implicitly by means of Monte Carlo (MC) or molecular dynamics (MD) simulations. As the solute needs to be embedded in a large number of solvent molecules, all of which are represented at a (i) quantum mechanical (QM) level, (ii) a purely classical molecular mechanical (MM) level or (iii) a mixture of the two, they are highly computer intensive. *Continuum models*⁷⁷⁻⁸¹ are less computer intensive than discrete methods since they differ from the latter in the respect that they do not treat the solvent molecules individually, but rather as a dielectric continuum using the theory of polarisable fluids. The solute is still modelled by either QM or MM.

When performing the calculations, it is convenient to divide the solvation process into three additive parts: (i) cavitation forces (ΔG_{cav}), (ii) dispersion and repulsion (van der Waals) forces (ΔG_{vdw}) and (iii) electrostatic forces (ΔG_{ele}) (Eq. 8).

$$\Delta G_{\text{hyd}} = \Delta G_{\text{cav}} + \Delta G_{\text{vdw}} + \Delta G_{\text{ele}} \quad (8)$$

Since the cavity formation depends largely on the size of the solute, the van der Waals term and the electrostatic term can be grouped together as one interaction term; $\Delta G_{\text{int}} = \Delta G_{\text{vdw}} + \Delta G_{\text{ele}}$ (Eq. 9).

$$\Delta G_{\text{hyd}} = \Delta G_{\text{cav}} + \Delta G_{\text{int}} \quad (9)$$

If the free energy of one compound is simulated for two different solvents, for e.g. *n*-octanol and water, the free energy of transfer between those solvents can be obtained. In the case of *n*-octanol and water, this corresponds directly to the octanol-water partition coefficient of the compound.

The free energy of hydration can also be estimated by a quantitative structure-property relationship (QSPR) approach. One of the first examples of this was provided by Hine and Mokarjee⁸² in 1975, but several others have followed since^{76,83-85}. The RMSE of the predicted ΔG_{hyd} with these models ranges from about 3-6 kJmol⁻¹ which corresponds to approximately 10% of the range of ΔG_{hyd} of the respective training set. Unfortunately, the datasets that were used for the training of these models are restricted to organic compounds which are not drug-like. It can be expected that the error will be larger for drug molecules, since they are large, often flexible and complex chemical structures in comparison to the small organic molecules used for model development.

1.3. Experimental estimation of drug solubility

1.3.1. Definition of intrinsic solubility

The intrinsic solubility (S_0) of a drug is the concentration of a saturated water solution of the neutral form of that drug in equilibrium with its solid.

In practice, this means that the drug should be in its free form (i.e., not a salt or solvate), that the preferred solvent is water, that the pH should be adjusted so that only the neutral form of the drug exists in the solution and that there is excess solid present. It would also be desirable to start from a pure sample of the most stable polymorph at the relevant temperature and pressure.

1.3.2. Factors influencing solubility experiments

When determining aqueous solubility, very different values may be obtained depending on the experimental set-up and the methods used. Care must be taken to ensure that the conditions used in the experiment correspond to the desired endpoint. For example, consider the different endpoints of kinetic and thermodynamic methods for the determination of solubility described in Section 1.3.3. and 1.3.4.

1.3.2.1. pH

Drugs often contain ionisable groups such as amines, carboxylic acids and sulphonamides. The ratio of ionised and unionised compound will vary with the pH of the solution, depending on whether the protolytic function is an acid or a base, and on the pK_a of that group. The ionised form has a higher solubility than does the unionised form. It follows that the solubility of acids and bases is pH dependent. The Henderson-Hasselbalch equation describes the relationship between the pH of a solution and the fraction of deprotonated and the protonated species of the drug as a function of its pK_a . For a base, the Henderson-Hasselbalch equation becomes Eq. 10, which can be rewritten into a more practical form, describing the solubility of a base at a given pH as a function of its intrinsic solubility and pK_a (Eq. 11).

$$pH = pK_a + \log\left(\frac{[B]}{[HB^+]}\right) \quad (10)$$

$$S_{pH} = S_0 \cdot \left(1 + 10^{(pK_a - pH)}\right) \quad (11)$$

In Eq. 10, $[B]$ is the molar concentration of the base in neutral form and $[HB^+]$ is the molar concentration of the base in its ionised form. In Eq. 11, S_{pH} is the total solubility of the base at a given pH (i.e., the sum of the neutral and ionised forms at that pH) and S_0 is the intrinsic solubility of the base. The Henderson-Hasselbalch equation is only valid for ideal solutions, for which it predicts a tenfold increase in solubility with each unit decrease in pH below (for a base) the pK_a up to infinity. The reverse is true for an acid. In a practical situation, solutions do not behave ideally and the solubility is not infinite, but will, instead, be limited by ion-pairing and aggregation. Regardless of this, a change in pH will have a significant impact on the solubility, making a reliable pK_a value important when planning experiments.

1.3.2.2. Solid-state form

The physical form of the solid will influence the solubility. Salts, solvates and polymorphs of the same compound exhibit different solubilities. For some compounds this effect can be large, while it is less pronounced for

others. Typically, the difference in solubility between two polymorphs of the same compound is around a factor of two⁸⁶. There are several other aspects related to the physical form that will influence the solubility; some examples follow. The crystalline form of a compound is generally less soluble than the amorphous form of the same compound. Hydrates usually exhibit a lower solubility in water than their corresponding anhydrous form. When dissolved, any metastable form will transform into the thermodynamically stable form at the relevant temperature and pressure. Furthermore, the purity of the compound influences the solubility, with this effect being more pronounced for poorly soluble compounds²³.

1.3.2.3. Ionic strength

The solubility of a salt will decrease upon the addition of a *common ion*, i.e. any of the ions making up that salt. Oppositely, addition of a non-common ion will result in an increased solubility of a sparingly soluble salt. For non-electrolytes, on the other hand, an increase in the ionic strength will result in a decrease in the solubility. Consequently it would be advisable to control the ionic strength and to keep track of the ions present in solution. Using buffer as a solvent may result in multiple equilibria between salts in the solid form and their corresponding ions in solutions, with the result that the solubility will be unpredictable.

1.3.2.4. Temperature

Dissolution in water is an endothermic process (i.e., heat is absorbed) for most compounds. This results in an increase in the equilibrium solubility with an increase in temperature. The increase in solubility can be explained by Le Châtelier's principle, which can be summarised as: *'If a chemical system at equilibrium experiences a change in concentration, temperature, or total pressure, the equilibrium will shift in order to minimize that change.'* For an endothermic process this means dissolving more compound to avoid an increase in temperature. Temperature will have a minor effect on the estimation of solubility provided that the change in temperature is small ($\pm 2^\circ\text{C}$).

1.3.2.5. Organic solvent

The addition of a co-solvent (i.e. a water-miscible organic solvent) will increase the aqueous solubility of hydrophobic compounds; the larger and more non-polar the solute (i.e. compound), the greater the effect of the co-solvent. For this reason, co-solvent is often used to facilitate the determination of poorly soluble compounds or when determining solubility by automated methods starting from a dimethyl sulfoxide (DMSO) stock solution. Using co-solvent will change the experimental conditions, resulting in a different solubility value than if pure water and/or solid compound had been used.

1.3.3. Kinetic methods

Two main types of methods for the determination of drug solubility can be identified: (i) *kinetic methods* that estimate the non-equilibrium solubility in a high throughput mode and (ii) *thermodynamic methods* that estimate the solubility at equilibrium.

Examples of methods with a high capacity (50-300 compounds per day) used by pharmaceutical companies for fast estimation of the solubility in buffer are provided by (i) *the turbidimetric method*¹⁰ and (ii) *the nephelometric method*⁸⁷.

In (i), aliquots of 1 μL of a 10 $\mu\text{g mL}^{-1}$ DMSO solution are added at 1 min intervals to a pH 7 phosphate buffer. Precipitation is detected as an increase in UV absorbance by light scattering in the 600-820 nm range. A total of 14 dilutions is made, resulting in a range in solubility being covered from 4 $\mu\text{g mL}^{-1}$ to 56 $\mu\text{g mL}^{-1}$.

Method (ii) starts with a 10 mM DMSO solution being diluted 20-fold in phosphate-buffered saline (PBS), at pH 7.4, and then serially diluted 10 times across a 96-well plate with PBS containing 5% DMSO. The concentration at which the compound precipitates is detected by light scattering by using a nephelometer with a laser light source at 633 nm.

The advantage of using these methods is their speed and the fact that they estimate the solubility under conditions similar to those used in biological assays. Storing compound libraries in DMSO simplifies the handling, although several issues have arisen in conjunction with the interpretation of assay results, demonstrating that this storage method is not completely reliable⁸⁸.

1.3.4. Thermodynamic methods

Thermodynamic methods result in intrinsic solubility (S_0), since the solubility is measured at equilibrium between the solid and the saturated solution. However, depending on the set-up, these methods do not always strictly conform to the constraints associated with S_0 as outlined in Section 1.3.1. The most commonly used method is the *shake-flask method*^{59,89}. It has been used extensively in chemistry, for instance for the determination of $\log P$. Generally, excess solid material of the compound is added to 0.5-3 mL of pure water. The pH is adjusted with diluted acid or base to guarantee that the entire sample is in its neutral form, whereupon it is agitated until equilibrium is reached (normally for 24-72 hrs). The saturated solution is separated from the excess solid through filtration or centrifugation and the concentration of the compound is determined with HPLC-UV or HPLC-MS.

The throughput of thermodynamic methods is less than that of the kinetic methods, typically about 5-10 compounds per day. However, the extra time needed should be weighed against the high quality of the solubility data.

1.4. Computational estimation of drug solubility

1.4.1. Model development – QSPR

To estimate and optimise physicochemical and pharmacokinetic properties has become an equally important and integrated part of the work in early drug discovery as has been the estimation and optimisation of pharmacological activity. With statistical and mathematical tools it is possible to relate chemical structure to any measured activity or property and then use that relationship to predict that property for new unknown compounds. The relationship then becomes a model. These models are commonly referred to as quantitative structure-activity relationships (QSAR) and QSPR. Model development can be divided into four steps: (i) selection of the dataset that is going to be used for the training of the model, (ii) generation of a description of the chemical structures for the compounds in the training set, (iii) compilation of literature data or generation of experimental data for the desired property and (iv) the use of statistical or mathematical tools to relate the description of the chemical structure to the experimental data. The model is then validated and tested before it is used for the prediction of new, to the model, unknown compounds. Model validation procedures are discussed in Sections 3.9.2. and 4.2. The steps (i) to (iv) in model development are outlined in Fig. 4 and considerations taken in each one of them are discussed in the four sections below.

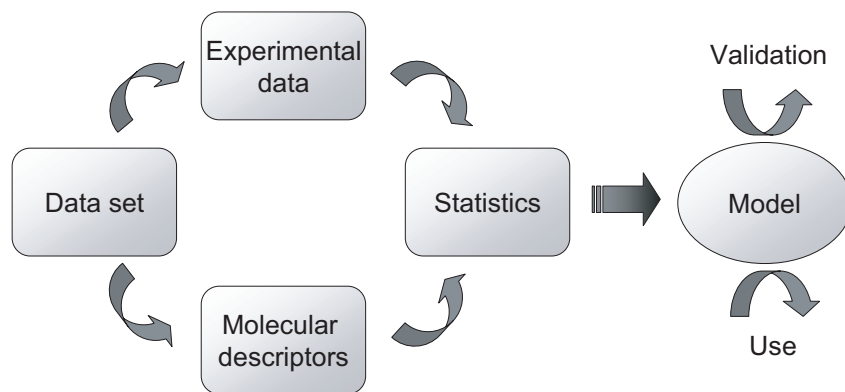


Figure 4. The four steps in the development of a QSPR model. (i) Selecting a dataset, (ii) calculating molecular descriptors of the chemical structures, (iii) generating experimental data and (iv) connecting the experimental data to the chemical structure by statistical or mathematical tools. The model is validated before use.

1.4.2. Dataset selection

When considering the compounds that should make up the training set it is important to have the future application of the model in mind. The first question to ask is if the model should be global, that is representative of most chemical structures; or if a local model, restricted to a homologous series of compounds, is more appropriate. In order to achieve a global model the training set needs to be chemically and structurally diverse. The diversity can be estimated by a number of tools⁹⁰. For a local model, the diversity within that restricted volume of chemical space⁹¹ need to be considered. Furthermore, if the model is to be used for the prediction of drugs, then drugs must also make up the majority of the training set.

1.4.3. Molecular descriptors

The way in which the chemical structure can be represented ranges from simple measures of molecular weight and counts of elements calculated from the molecular formula, often referred to as one-dimensional (1D) molecular descriptors, through measures of branching and connectivity that are calculated from or a two-dimensional (2D) structure of the molecule, to surface and volume descriptors calculated from the three-dimensional (3D) molecular structure. The 3D structure takes into account different possible conformations of the molecule. The optimisation of the conformation of a molecule can be achieved by faster techniques, such as MM, or by more sophisticated and computationally demanding QM techniques that consider the electron distribution of the molecule. The three groups are outlined in Fig. 5.

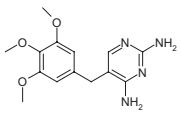

$C_{14}H_{18}N_4O_3$	1D: Molecular weight, Element counts, Number of atoms
	2D: Hydrogen bond donor and acceptors, connectivity, topology
	3D: Surface area, Volume, Electron distribution, Charges

Figure 5. Three types of molecular descriptors of ranging complexity can be identified. They are referred to as 1D, 2D and 3D molecular descriptors from their respective method of calculation. Examples of each descriptor type are given in the figure.

An important feature of molecular descriptors is that they are interpretable. It should be possible to for a person with knowledge of chemistry, for e.g. a medicinal chemist, to intuitively translate the values of the molecular descriptors selected as important in a model into a change in chemical structure of a lead compound resulting in improved properties of that compound.

Alternatively a fragment or atom-based approach can be taken for the representation of the chemical structure. The compounds in the training set are then divided into their constituent fragments or atoms. Each such fragment or atom is considered to make a contribution of a certain size to the property studied. The size and direction of that contribution is assigned through regression analysis of the fragments or atoms present in the training set. For new unknown compounds, the values for all fragments or atoms present in that molecule are summed up to yield a predicted value.

1.4.4. Experimental data

At first glance, measuring the equilibrium solubility of drugs may be perceived as a straightforward, routine experiment. In reality, though, the variability in the solubility data found in the literature is quite large. In 1984, a ring test performed by 17 laboratories in Japan came back with the discouraging results that the equilibrium solubility for anthracene at 20° C was ranging from -7.08 to -6.22 (83 nM to 603 nM) which is equal to a sevenfold difference in solubility. For fluoranthene, the solubility ranged from -6.38 to -5.94 (0.42 µM to 1.15 µM) which corresponds to a threefold difference in solubility⁹². Considering that a standardised protocol was used for the experiments performed by the 17 laboratories, this was a surprisingly large difference. The values reported for same compounds in the AQUASOL dATABASE⁹³ displayed an even larger range: more than 1.5 log units (a 36-fold difference) for anthracene (-6.79 to -5.23; with 22 determinations being made) and almost 0.9 log units (an eightfold difference) for fluoranthene (-6.23 to -5.35; for the 13 determinations made)⁹⁴. Anthracene and fluoranthene are both neutral compounds that are mainly used in the production of dyes and agrochemicals. Their chemical structures are simple in comparison to the highly functionalised and complex molecules handled by the modern drug discovery, which should make the determination of their solubility a relatively easy task.

In agreement with the above-mentioned findings, data compiled from the literature for seven common drugs (carbamazepine, diazepam, hydrocortisone, ketoprofen, naproxen, prednisolone and progesterone) supplied by Loftsson *et al.*⁹⁵ in 2006 also displayed a large difference in experimental values. The average difference observed in the published solubility values was 0.57 log units. The largest difference was 1.5 log units, observed for carbamazepine (-4.30 to -2.80), and the smallest difference was 0.14 log units, observed for hydrocortisone (-3.10 to -2.96). These values

support the statement made by Jorgensen and Duffy⁹⁶ that, ‘*the average uncertainty in experimental logS measurements for a reasonably complex organic molecule is likely no better than 0.6 log units*’.

This has two major implications for solubility modelling: firstly, the accuracy of computational models trained on data from the literature can never exceed this value, or they are likely to be overfitted; and secondly, if values from the literature are used for modelling, it is important that a thorough investigation be made of the quality of those values before such data is used in the development of computational models. The latter can sometimes be difficult, since details about the experiments, such as the pH, temperature, solvent, purity of the compound, equilibrium time and polymorphic form are often not reported in the literature. The most appealing strategy would be to use solubility data generated under standardised conditions from one single laboratory for model development.

1.4.5. Statistical and mathematical methods

Many techniques exist for relating the chemical structures in the training set to their corresponding experimental values of the property under investigation. These can be in the form linear or non-linear regression methods such as multiple linear regression⁹⁷⁻⁹⁹ (MLR) or projection to latent structures by means of partial least squares¹⁰⁰⁻¹⁰² (PLS); artificial neural networks^{103,104} (ANN) that are intrinsically non-linear in nature and classification models, such as decision trees. The transparency of the chosen method will greatly affect the interpretability of the resulting model.

Neural networks can be trained to very accurately predict the training set, but are often difficult to interpret in terms of the changes needed to be done to the chemical structure in order to increase or decrease solubility. As a consequence, an ANN model is often referred to as being a black box model. They also suffers from the risk of being overfitted, i.e. they are so well fitted to the observations of the training set that they lose the predictive abilities when compounds not part of the training set are considered. Therefore it becomes extremely important that the experimental data used for training neural networks are of high quality or the model would be predicting the experimental noise.

Linear regression methods have the advantage of being simple and transparent. The influence (both size and direction) of a selected variable is easily identified through the coefficient for that variable, which ensures interpretability of the model. On the other hand, they have the drawback of not being able to represent non-linear relationships between the molecular descriptors and the response variable.

For this thesis, multivariate methods¹⁰⁵ principal components analysis¹⁰⁶ (PCA) and PLS were chosen for the development of models for S_0 and ΔG_{hyd} . These are augmented linear regression methods that do not suffer

from the limitations of MLR which can not handle covariance among the variables, nor data matrices that are “short and fat”, i.e. contain a large number of variables (molecular descriptors) in comparison to the number of observations (compounds). Briefly, PCA and PLS are projection methods that have the ability to extract a few principal components or latent variables that contain the majority of the information related to the variation provided by hundreds of variables. PCA is used for the mapping of data, diversity analysis and to find clusters and trends in data. It only considers the molecular descriptors, while PLS tries to fit the variance in the response variable (for e.g. S_0) to the variance in the molecular descriptors.

1.4.6. Available computational solubility models

Numerous scientific articles have provided a wide range of computational models for the prediction of drug solubility over the last decade. They can be divided into three main categories according to the methods used for model development: (i) semi-empirical or regression equations based on experimental data, (ii) fragment or atom-based methods and (iii) models that relate molecular descriptors to the solubility through statistical or mathematical methods. The first category is discussed in Section 1.2.1.1. and will not be considered further here.

The method in the second category of dividing the chemical structure into fragments, either at a functional group level or at an atomic level, allows for direct interpretation of the contribution of a particular chemical fragment to S_0 . A drawback of these methods is that large datasets, containing as many fragments as possible, are needed in order to cover all potential fragments of new compounds to be predicted by the model. One of the first attempts to estimate water solubility based on fragments was made by Kühne *et al.*¹⁰⁷ in 1994. They used 694 non-electrolytes of environmental and pharmaceutical interest and noted that predictions were generally better for liquids than for solids. They therefore advocated the inclusion of a melting point term for solids. Several studies using group^{108,109} and atom contributions¹¹⁰ have followed using larger dataset containing few or no drugs. In general, they achieved standard errors (SE) of 0.5-0.6 log units for the training sets and SE of around 1 log unit for test sets.

In the third category of solubility models, ANN trained on large datasets either predominantly comprised of organic compounds¹¹¹ or augmented with drugs¹¹²⁻¹¹⁴ and drug-like compounds have been popular. In particular one dataset, first compiled by Huuskonen *et al.*¹¹⁵ from databases AQUASOL⁹³ and PHYSPROP¹¹⁶, has been re-used several times¹¹⁷⁻¹²⁰. Unfortunately, several errors and misprints have been discovered for this dataset¹²¹. MLR^{122,123}, PLS^{124,125}, support vector machines (SVM)¹²⁶ and decision trees^{127,128} have been applied to similar datasets with similar results. Typically, these models show RMSEs for both the training and test sets of ap-

proximately 0.5-0.7 log units. However when tested on drug-like compounds, the prediction made by these models in the pharmaceutically interesting region of solubility (-9 to -3 on a log M scale) was considerably less accurate.

A few models based on smaller datasets mainly comprised of drugs or drug-like structures have been developed. Unfortunately the solubility data was compiled from the literature rather than generated under standardised conditions¹²⁹⁻¹³¹. Models proposed by McFarland¹³², Raevsky *et al.*¹³³ and Jørgensen and Duffy¹³⁴ were accompanied by extensive discussions around the results and interpretation of the models at a physicochemical level, which is generally missing in the models referred to above. Prediction of the solubility in phosphate buffer pH 7.4 for a large number of compounds was attempted with limited success by Göller *et al.*¹³⁵. Promising alternative approaches to the methods described above include the prediction of solubility through free energy of solvation from quantum chemical calculations¹³⁶⁻¹³⁸ and the study of the change in solubility within series of structurally similar pairs of molecules¹³⁹.

Many of the above-mentioned models suffer from one or several of the following drawbacks that have been identified as key issues in solubility modelling^{96,121,140-143}. Firstly, the datasets used for training the models do not contain any or only a small fraction of drugs. Secondly, the training sets often cover a large range of solubility, from pico-molar (that is 10^{-12} M) to 100 M, or even 10 000 M^{112,113}. To accurately estimate solubility values in the pico-molar range appears nearly impossible and the meaning of a solubility value of 100 M is difficult to comprehend, especially considering that the solubility of pure water in itself is around 55 M. Training models on this large range of solubility results in a decreased accuracy for the smaller, but from a pharmaceutical perspective highly interesting region of nanomolar (10^{-9} M) to millimolar (10^{-3} M).

Regardless of statistical method, the use of solubility data from literature will evidently introduce error of unknown size into the model. As discussed in Section 1.4.4., the smallest expected experimental variation in literature data is around 0.5 log units. Ideally, data of intrinsic solubility generated under standardised conditions from one single laboratory should be used for model development.

The aims of this thesis were formulated with the objective to address several of the current issues in solubility modelling. The undertaken investigations strived to understand the solubility behaviour of drugs at a physicochemical level and to incorporate that knowledge into accurate and predictive means to estimate aqueous solubility purely from the chemical structure.

2. Aims of the thesis

The general objective of this thesis was to devise computational methods to predict the solubility of drugs and candidate drugs directly from chemical structure. The emphasis was on the quality of experimental solubility data (Papers I-III), model development and validation (Paper I), experimental properties influencing drug solubility (Papers II and III) and structural features influencing the solid state (III) as well as solvation (IV). The specific aims were:

- ◇ To develop computational models for solubility comprising high-quality experimental data for drugs and drug-like molecules.
- ◇ To assess and compare the quality of computational models based on both structurally diverse (global) datasets and homologous (local) series of compounds.
- ◇ To investigate which experimental properties contribute to the solubility and, in particular, the influence of solid-state properties on the solubility of crystalline drugs.
- ◇ To identify structural features of drugs, in the form of molecular descriptors, related to solid-state limited solubility.
- ◇ To identify structural features of drugs, in the form of molecular descriptors, related to solvation limited solubility.

3. Materials and methods

3.1. Selection of dataset

The strategy for the selection of the datasets used in Papers I-IV differed depending on the aim of the study in question. In Paper I, the goal was to achieve a structurally diverse dataset that covered as much of the oral drug space (see Section 3.2.) as possible, whilst taking care to only include high quality experimental data. High quality solubility data from collaborators within the pharmaceutical industry was added to in-house solubility data to form a database of diverse, drug-like compounds.

In Papers II and III, it was considered to be of great importance not only that the datasets reflected the structures of orally administered drugs, but also that they highlighted the contribution solid-state properties made to solubility. In Paper III, this was achieved through the selection of a dataset that displayed a wide range of solubility (S_0) and a narrow range of $\log P$. Hence, the solubility of the dataset used in Paper III was truly independent of lipophilicity.

In Paper IV, the hydration free energy of drugs was studied. Since this is not an experimental property, we searched the literature to find data of drugs obtained with a thoroughly validated simulation method.

3.2. Structural diversity and drug-likeness

The structural diversity of the datasets used in Papers I, II and IV was assessed by ChemGPS¹⁴⁴ methodology using the 2D molecular descriptors calculated with the program Selma¹⁴⁵, AstraZeneca R&D Mölndal, Sweden (see Section 3.8.1.). In Paper III the main objective for the selection was to achieve a dataset for which the S_0 was independent of $\log P$. Because of this, the structural diversity was not examined for this set of compounds. ChemGPS is a navigation tool that provides a “map” of the chemical space onto which a dataset of interest can be projected. From a PCA of a set of reference compounds and the selected molecular descriptors ChemGPS extract the first three principal components ($t[1]$, $t[2]$ and $t[3]$) and use them as the x, y and z-axes in a 3D co-ordinate system. The compounds under investigation can then be visualised in this co-ordinate system through PCA projection. In the ChemGPS analysis, the drug-likeness was gauged through

comparison of the chemical space covered by the compounds in our dataset with that covered by the compounds of an AstraZeneca R&D Mölndal in-house database comprising 456 orally administered drugs. The axes used to create the oral drug space were the first three principal components of the PCA and they mainly represent size (t[1]), polarity (t[2]) and flexibility (t[3]), respectively.

3.3. Chemicals and drugs

All chemicals and drugs used in the studies were either of analytical grade or of high purity. The drugs had a purity exceeding 98%, with the exception of griseofulvin (96%), which is a natural product. Compounds were generally in their free form, i.e. there were no salts or solvates, and possible polymorphic forms were investigated with differential scanning calorimetry. Chemicals and drugs were mainly purchased from Sigma-Aldrich, Stockholm, Sweden.

3.4. Crystal structures

The required crystal structures were retrieved from Cambridge structural database (CSD) version 5.27 from The Cambridge Crystallographic Data Centre (CCDC), Cambridge, UK. The search engine ConQuest version 1.8 (CCDC, Cambridge, UK) was used to query the database. Once the structures were retrieved, intermolecular forces were evaluated in the program Mercury version 1.4.1, CCDC, Cambridge, UK.

3.5. Differential scanning calorimetry (DSC)

Melting point (T_m), enthalpy of melting (ΔH_m) and entropy of melting (ΔS_m) for many of the investigated drugs were determined using differential scanning calorimetry (DSC). Triplicate samples of 1-3 mg were accurately weighed in sealed and pierced aluminium pans. Generally, samples of each compound were heated from room temperature to approximately 50 K above the melting temperature at a rate of 10 Kmin⁻¹. If any anomalies, were detected, such as for example asymmetric peak shape, multiple melting endotherms or re-crystallisation exotherms, the samples were rerun at a heating rate of 2 Kmin⁻¹ to enable a closer investigation.

3.6. Solubility determinations – the shake-flask method

The S_0 (expressed as the $\log S_0$ in M) of crystalline compounds was determined in quadruplicate according to the shake-flask method described by Bergström *et al.*⁵⁹. First of all, a rough estimation of the expected value of S_0 was made from previous determinations found in the literature and/or from $\log P$ and T_m using the GSE. At least three times excess weight of solid was weighed into 1.5 mL Eppendorf tubes, 1 mL of distilled water was added and the samples were thoroughly mixed on a vortex to achieve maximum wetting of the solid. For weak bases and weak acids, the pH was adjusted to at least 2 pH units above (bases) or below (acids) the pK_a with 0.01 M NaOH or 0.01 M HCl to ensure that all of the molecules were present in their neutral form. The pH was not adjusted for neutral (including bases with $pK_a < 2$ or acids with $pK_a > 12$) and zwitterionic compounds. The samples were agitated on an orbital plate shaker at 300 rpm for at least 24 hrs at room temperature ($21 \pm 0.5^\circ\text{C}$). The pH was then measured and the presence of undissolved material was confirmed before the samples were centrifuged in an Eppendorf centrifuge model 5403 for 15 min at a relative centrifugal acceleration of $23\,500 \times g$ to separate the saturated solution from the solid. After centrifugation, 0.5 mL of the supernatant was carefully pipetted into 2 mL HPLC autosampler glass vials using a Pasteur glass pipette and the samples were analysed by HPLC-UV, HPLC-FL or HPLC-MS-MS.

3.7. Hydration free energy calculations

Values for free energy of hydration (ΔG_{hyd}) for 48 drugs from Westergren *et al.*¹⁴⁶ were used in Paper IV. Briefly, in their method they use molecular simulations and interpret the results with the following equation:

$$\Delta G_{\text{hyd}} = A_{MS}\gamma + E_{LJ} + \frac{E_c}{2} \quad (12)$$

where A_{MS} is the molecular surface area, γ is the water-vacuum surface tension of the TIP4P model (63.5 mN m^{-1} , which is different from the macroscopic surface tension of 71.8 mN m^{-1} , as discussed by Westergren *et al.*¹⁴⁶) and E_{LJ} and E_C are the solute-water Lennard-Jones and Coulomb interaction energies, respectively. Standard NTP (fixed temperature and pressure) Monte Carlo simulations in the program BOSS Version 4.6¹⁴⁷ were used to obtain the interaction energies (E_{LJ} and E_C) and molecular volumes. A box of approximately 500 TIP4P¹⁴⁸ water molecules surrounding one drug molecule was prepared. The molecular surface area (A_{MS}) was calculated in BOSS by letting a sphere probe the Lennard-Jones potential energy surface around the

solute. The solute molecules were modelled by the OPLS-AA¹⁴⁹ force field for fully flexible molecules and partial charges were calculated using AM1¹⁵⁰ and CM1A¹⁵¹. The structures were optimized in vacuum and had a net charge of zero. In Paper IV the ΔG_{cav} is represented by $A_{\text{MS}} \gamma$, ΔG_{wdv} by E_{LJ} and ΔG_{ele} by $0.5 E_{\text{C}}$, respectively (from Eq. 8 in Section 1.2.2.4. and Eq. 12 above).

3.8. Molecular descriptor generation

3.8.1. 2D – Selma and Molconn-Z

Smiles were used as 2D structural input format for the calculation of molecular descriptors by the AstraZeneca in-house program Selma¹⁴⁵. A total of 93 2D descriptors that were related to molecular size, polarity, flexibility, charge distribution and connectivity were calculated. They included a range of well-known and commonly used 2D descriptors from different commercial sources. Selma descriptors were used in Papers I-IV.

The program Molconn-Z¹⁵² was used to calculate atom-type electrotopological state indices from Smiles. Briefly, the electrotopological state indices for a particular atom are values resulting from its topological and electronic environment. The indices encode the electronegativity as well as the local topology of each atom by considering perturbation effects from neighbouring atoms. Descriptors from Molconn-Z were used in Paper I.

3.8.2. 3D – Surface area descriptors and VolSurf

For the surface area descriptors low energy 3D conformers were obtained with the program MacroModel version 6.5. A 500-step Monte Carlo conformational search was performed using Merck Molecular Force Field (MMFF) in a simulated water environment on a Silicon Graphics Octane workstation. The in-house computer program MAREA¹⁵³ was used to calculate the free surface area (SA) of each atom as well as the molecular volume of the conformation with the lowest energy. The polar surface area (PSA) was defined as the area occupied by oxygen and nitrogen atoms and hydrogen atoms attached to these heteroatoms. The non-polar surface area (NPSA) was defined as the total surface area (TSA) minus the PSA. The SA was divided into the partitioned total surface areas (PTSAs). Each PTSA corresponds to the surface of a certain type of atom. For example, the NPSA originating from carbon atoms can be partitioned into the surface areas of sp-, sp²-, and sp³-hybridised carbon atoms. Both the absolute SAs and the surface areas relative to the TSAs were calculated.

Smiles were used as the 2D structural input format and were converted into 3D structures using the program Corina version 3.20^{154,155}. From a 3D representation of the molecules, 94 descriptors were calculated with the DRY (hydrophobic probe), OH2 (water), and O (carbonyl) probes in Vol-Surf^{156,157} version 4.0.1. These descriptors describe surface properties related to size, shape, hydrophobic-hydrophilic balance, amphiphilic moment and capacity factors.

3.9. Statistical analysis

3.9.1. Linear regression

Regression analysis was used in Papers II and III to (i) assess how much of the variance in solubility that could be ascribed to $\log P$ alone and (ii) to test how much of the variance in either the residuals (II) or $\log S_0$ (III) was related to any of the experimental solid-state properties T_m , ΔH_m , and ΔS_m . The fit of the regression was assessed from the coefficient of determination (R^2) and by calculating the RMSE according to Eq. 13:

$$\text{RMSE} = \sqrt{\frac{\sum (\text{obs} - \text{pred})^2}{N}} \quad (13)$$

where *obs* and *pred* were the observed and the predicted values for the observations (compounds), respectively, and *N* was the number of observations.

3.9.2. Multivariate analysis

Multivariate data analysis tools *principal components analysis* (PCA) and *projection to latent structures by means of partial least squares* (PLS), as implemented in Simca-P version 11.0.0.1 (Umetrics AB, Umeå, Sweden), were used for Papers I-IV. All variables (molecular descriptors) were mean centred and scaled to unit variance. Descriptors displaying skewness outside the range of ± 1.5 or a variance close to zero were either excluded directly or cubic root transformed to acquire a normal distribution for the observations in the dataset. The model predictivity was judged by the R^2 and the RMSE, calculated according to Eq. 14.

$$\text{RMSE} = \sqrt{\frac{\sum (\text{obs} - \text{pred})^2}{N - 1 - p}} \quad (14)$$

where *obs* and *pred* were the observed and the predicted values of the observations, respectively, and N was the number of observations and p was the number of latent variables used by the PLS model.

Each PLS model was validated by all or any of the following: (i) leave-many-out (4 or 7 groups) cross validation (Q^2), (ii) a permutation test (100 iterations) in which the values of the response variable were randomised and the PLS analysis repeated in order to detect the risk of chance correlations and (iii) using an external test set. All three methods of validation were used in Paper I, while Q^2 and permutation test were used for Papers II, III and IV.

Simple variable selection was applied to decrease the complexity of the models and alleviate interpretation, because of the high co-variance between some of the descriptors. One of two methods was used: (i) If the exclusion of the least important variable resulted in a model with a higher Q^2 , then that descriptor was permanently left out of the model. This procedure was repeated until no further improvement of the model was achieved. Or (ii) first, the bottom 50% of the variables exhibiting the lowest level of importance was excluded. Second, overlapping variables (residing in the same area of the PLS loading plot) were excluded to leave only a few variables representing the key descriptors encoding the predominance of the information related to the response variable. The aim with the variable selection was to maintain the predictivity and increase the robustness of the model by removing information not directly related to the response variable (i.e. removing the noise).

4. Results and discussion

4.1. Training drug solubility models

4.1.1. Quality of experimental data (Papers I-III)

Experimental variability in the data used for the training of a computational model introduces noise into that model. It has been a priority throughout this thesis to include only high quality experimental data produced under standardised conditions to achieve as high a prediction-to-noise ratio as possible.

In Paper I, a compilation of high quality data from collaborators in the industry was used with data generated in-house to form a large and diverse training set. Only those data meeting the following criteria were considered: (i) the solubility value should be determined at a pH resulting in measurement of the intrinsic solubility, (ii) the solubility value should be obtained under equilibrium conditions (iii) the solubility should be determined at room temperature. However, despite the precautions taken, there is probably some experimental variability in this dataset. With the intention of quantifying the experimental variability, the correlation between different methods used in-house for solubility determination was calculated. Excellent agreement was observed between the small-scale shake-flask method and potentiometric titration ($R^2=0.95$) (data not shown). Similarly, our shake-flask method has previously been shown to reproduce values found in the literature with high accuracy ($R^2=0.98$)⁵⁹. Given this, the highest possible predictivity (R^2 and Q^2) for the global model is more likely to be 0.9 rather than 1.0, once the experimental error has been accounted for.

In Paper II, a comparison was made between the use of filtration or centrifugation as methods to separate excess solid from the saturated solution when equilibrium has been achieved. Excellent agreement ($R^2=0.998$) was obtained for the 15 drugs investigated (Fig. 6). In this study, glass fibre filter was used for the filtration and the samples for both centrifugation and filtration originated from the same vial.

Since the purity, crystallinity and differences between polymorphic forms of drugs greatly influence the solubility, the solid-state properties for all of the compounds used in Papers II and III were characterised. How these properties contribute to the intrinsic solubility is discussed in Section 4.4. of this thesis.

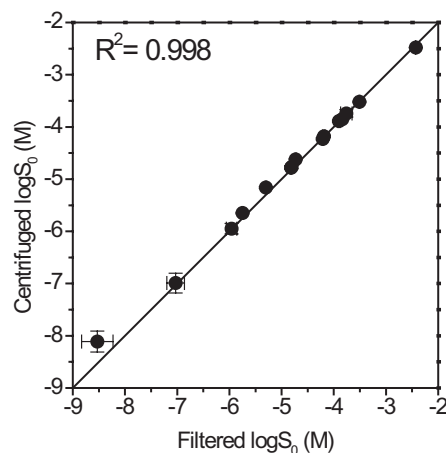


Figure 6. Correlation between experimental solubility of 15 drugs using either centrifugation or filtration to separate the excess solid material from the saturated solution at the end of the experiment. The agreement between the two methods was excellent.

Performing the above analyses of the experimental variability ensured that the solubility data used throughout Papers I-III were of the highest quality possible. This, together with the fact that the solid-state properties of the starting material used for the solubility determinations were characterised for the majority of the compounds investigated gives confidence in the results and in the conclusions drawn from them. In addition, the solubility data included in this thesis provides an excellent external test set for the validation of future solubility models.

4.1.2. Diversity and drug-likeness (Papers I, II and IV)

Since drug molecules generally differ from organic compounds by being larger, more lipophilic and more complex with multiple functional groups, drugs were prioritised over non-drugs for the investigations performed in this thesis. Generally, we concentrated on orally administered drugs because these comprise the majority of the pharmaceutical market. Furthermore, an effort was made to include compounds from as many structural and therapeutic groups as possible to make the models generally applicable to new molecules.

ChemGPS¹⁴⁴ methodology was applied to datasets in Papers I, II and IV to assess the structural diversity as well as the drug-likeness (Section 3.2.). The selection criterion for compounds in Paper III is discussed in Sections 3.1. and 4.4.2. Despite the large difference in the size of the datasets used (N=85, 26 and 48 in Papers I, II and IV, respectively), all three showed

a satisfactory spread in physicochemical properties as well as in the volume of the oral drug space that they covered (Fig. 7a-c). The results from these papers can, therefore, be regarded as being more representative and generally applicable to drugs rather than to organic compounds at large.

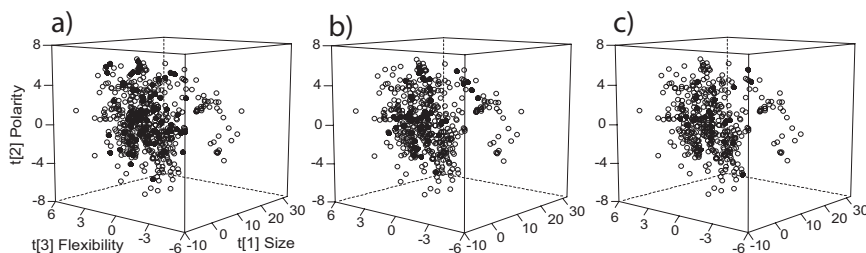


Figure 7. ChemGPS analysis of diversity and drug-likeness for the datasets used in (a) Paper I (N=85), (b) Paper II (N=26) and (c) Paper IV (N=48). Compounds included in studies I, II and IV (filled circles) are projected together with 456 orally administered drugs (open circles) from an AstraZeneca in-house database that served as a reference of the oral drug space. The axes mainly represent size ([t1]), polarity ([t2]) and flexibility ([t3]).

4.2. Validating drug solubility models (Paper I)

Model validation is probably the most important step in the development of reliable prediction tools. Without a thorough validation, the accuracy and robustness of the prediction of new compounds can not be judged. In general PLS was used to derive models for S_0 (I-III), T_m (III) and ΔG_{hyd} (IV). The choices available for model validation for PLS are the use of Q^2 , permutation test and external test set. The most powerful determinate of model accuracy and robustness is the application of an external test set, i.e. a set of compounds that was not included in the model training.

Table 1. Statistics for global solubility models

Model	R^2	Q^2	RMSE _{tr}	R^2_{te}	RMSE _{te} ^a	$R^2_{ext\ te}$	RMSE _{ext\ te} ^a
2D	0.75	0.68	0.92	0.62	0.86 (1.00)	0.56	0.80 (1.01)***
3D	0.57	0.53	1.20	0.67	0.94 (1.04)	0.52	0.89 (1.06)**
2D+3D	0.78	0.71	0.86	0.54	1.04 (1.10)	0.54	0.93 (1.07)*
Consensus	0.80		0.90	0.71	0.83 (0.93)	0.59	0.82 (0.95)*

^a RMSE values (log units) including the data for compounds with solubility values outside the solubility range (-8.8 to -1.2 log units) covered by the training set are given in parentheses. Statistically significant differences between the RMSE values before and after the exclusion of solubility data outside the solubility range covered by the training set are denoted with asterisks; $p < 0.05 = *$, $p < 0.01 = **$, $p < 0.001 = ***$.

In Paper I, a set of 207 drugs and drug-like compounds commonly used for model development^{130,134} was applied as an external validation of the models. The results (displayed in Table 1) showed that the global models (based on 2D, 3D, 2D+3D descriptors or a consensus of the three) predicted the external test set with the same accuracy as, or higher than, the training set and the test set comprised of in-house data (based on the RMSE after the removal of solubility data outside the solubility range of the training set).

When R^2 and the observed versus predicted plots for the four models were considered, it seems that even though the average prediction error was roughly the same as for the training set, the external test set comprised more compounds that were mispredicted by several log units. These compounds were predominantly those with a high solubility, with experimental values outside the range covered by the training set (-8.8 to -1.2) (Fig. 8a). Furthermore, very few low solubility compounds were present in the external test set (Fig. 8a).

When comparing the range of physicochemical properties (molecular descriptors) covered by the training set and the external test set studied in Paper I, it becomes clear that the external test set only partly covers the properties of the training set, making this validation representative of compounds residing in that area only (Fig. 8b).

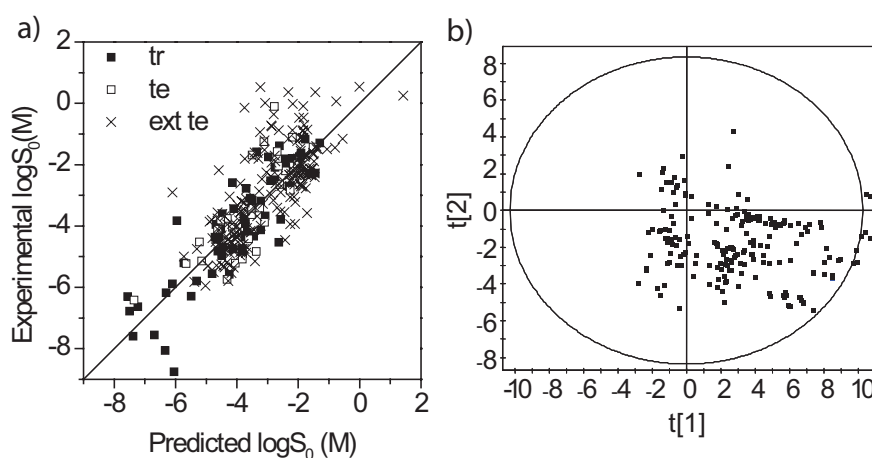


Figure 8. External test set: (a) observed versus predicted $\log S_0$ from the 2D descriptor model for the dataset in Paper I, ■ = training set, □ = test set and × = external test set. (b) the 207 compounds of the external test set used in Paper I is projected on the model space as defined by the first two principal components (PCs) in a PCA for the in-house training and test set data (N=85) using both 2D and 3D molecular descriptors. The ellipse (b) shows the 95% confidence interval limits.

For the local models, which are discussed in greater detail below (Section 4.3.), comprised of bases, acids and ampholytes, respectively; the pre-

diction of the external test set was generally inferior to that of the training set. This result can mainly be ascribed to large homologous series (e.g. barbituric acids and xanthenes) present in the external test sets with no structural homologues in the corresponding training sets.

An evaluation of whether the external test set is appropriate for use, demands that the solubility range and the descriptor range of that dataset is studied in relation to the properties of the training set. This highlights the importance of the transparency of new models; the user needs to be provided with information about the range of the physicochemical properties and the S_0 values of the training set in order to be able to evaluate the applicability domain of the model.

4.3. Global versus local models (Paper I)

Several large homologous series of compounds were identified within the external test set used for model validation in Paper I. It is generally considered to be less demanding to find highly predictive models for datasets of limited structural diversity than for datasets comprising compounds of greater structural diversity. To test this theory, local models were constructed for structural subsets found in the external test set for barbituric acids, xanthenes and steroids. In addition, a set of β -receptor antagonists was compiled from the training and test sets used in Paper I and added to the compounds for which the solubility had previously been determined in our laboratory¹⁵⁸. The resulting RMSE values for each of the training sets are shown in Fig. 9 together with the global consensus model and local models for acids, ampholytes, non-protolytes and bases. It is clear that the prediction error is generally smaller for the homologous series than for the diverse global dataset.

A global model based on a diverse set of drugs will be universally applicable to new drug-like molecules. However, the accuracy of the predicted S_0 will not be as great as it could be. On the other hand, a local model based compounds that are structurally similar to one another will give predicted S_0 values of high accuracy, but it will be restricted to the model range defined by that particular series of compounds. Thus the two types of model serve different purposes and should be used accordingly.

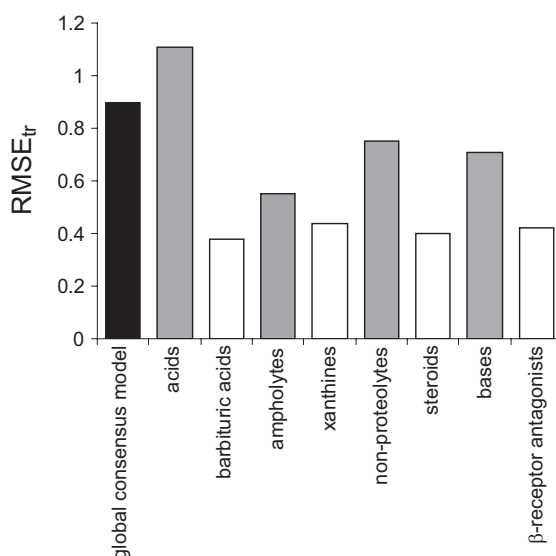


Figure 9. Root mean square errors (RMSEs), in log units, of the prediction for models based on local datasets (white and grey bars) as compared to the consensus model (black bar) for the global dataset in Paper I.

4.4. Experimental properties influencing drug solubility

4.4.1. Solvation properties – $\log P$ (Paper II)

Already in 1968 Corwin Hansch formulated a series of regression equations that described how experimental $\log P$ values influenced the water solubility of organic liquids²⁴ (see Section 1.2.1.1.). Even though $\log P$ is just a measure of the ability of a molecule to be solvated in octanol relative to the ability of the same molecule to be solvated in water, it is often regarded as a crude measure of the solvation of drugs in water. A more appropriate property for the estimation of the degree of solvation would be the ΔG_{hyd} , which is discussed in more detail in Section 1.2.1. and 4.6.3. Judging from the popularity of $\log P$ and its calculated counterparts as descriptors in models of S_0 it is clear that the octanol-water partition coefficient is closely related to the solubility of both organic compounds and drugs. Note, for instance, that in all of the global models presented in Paper I, CLOGP is rated as the most important descriptor (Fig 6, Paper I).

A regression analysis with $\log S_0$ and CLOGP as variables was performed for 270 drugs and drug-like compounds constituting the training, test and external test set in Paper I to estimate the extent to which solubility is related

to solvation for a normal set of drugs and drug-like compounds (Paper II). CLOGP was found to explain 54% of the variability in $\log S_0$ of this dataset (Eq. 15).

$$\log S_0 = -1.91(\pm 0.07) - 0.617(\pm 0.02)\text{CLOGP} \quad (15)$$

$N = 270, R^2 = 0.54, F = 308.8, \text{RMSE} = 1.12$

The regression parameters are indicated with ± 1 standard deviation and the statistical parameters given are as follows: N is the number of compounds, R^2 is the coefficient of determination, F is the test statistic from the F-test and RMSE is the root mean square error. When using Eq. 15 to predict the solubility of a set of 26 structurally diverse drugs it was found that it explained 67% of the variability of the solubility in this dataset (Fig. 10a). The increased degree of correlation between $\log S_0$ and CLOGP for the smaller set of drugs can be explained by the fact that the dependency of solubility on lipophilicity is highly dataset dependent.

The above analysis suggests that solubility is strongly dependent on solvation. However, almost half (46%) of the variability in the solubility of the compounds under consideration (for the larger dataset, $N=270$) remains unexplained and subsequently must be related to other properties, such as for instance those descriptive of the solid state.

4.4.2. Solid-state properties (Papers II and III)

The solubility of a crystalline solid is governed by the balance between its ability to interact with the solvent (i.e. water) and its ability to make stable interactions with itself in the crystalline state. It therefore seems natural to investigate if, and to what degree, solubility is related to experimental solid state properties such as T_m , ΔH_m and ΔS_m . In 1980 Yalkowsky and Valvani showed that in addition to $\log P$, T_m was important for the solubility of solid organic compounds²⁷ (see Section 1.2.1.1.).

The influence of experimental solid-state properties on solubility was investigated on two occasions in the work conducted for this thesis (Papers II and III). On the first the residuals (observed $\log S_0$ -predicted $\log S_0$) from Eq. 15 were related to experimental T_m , ΔH_m and ΔS_m by regression analysis of 26 compounds (Paper II). In addition, the improvement of the value obtained for the solubility as predicted by a combination of CLOGP, ΔH_m and ΔS_m , relative to the solubility as predicted by CLOGP alone (Eq. 15), was studied for the same 26 compounds. On the second occasion, $\log S_0$ for a $\log P$ -independent dataset comprised of 20 compounds were related to experimental values for T_m , ΔH_m and ΔS_m (Paper III).

The result of the analysis of the residuals in Paper II showed that they were mainly related to ΔH_m ($R^2=0.26$) and to a lesser extent related to

T_m ($R^2=0.09$) and ΔS_m ($R^2=0.09$). Furthermore, the overall prediction of $\log S_0$ was improved by 0.3 log units when ΔH_m and ΔS_m were considered in combination with CLOGP relative to the solubility as predicted by CLOGP alone (Fig. 10a-b). For some compounds (i.e. astemizole, glyburide, fenbufen, gliclazide and griseofulvin) the improvement was larger than one log unit.

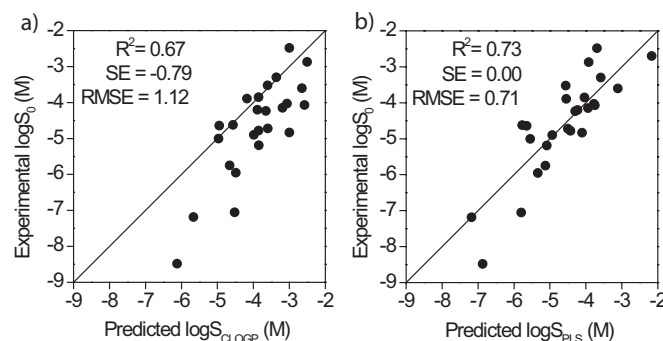


Figure 10. Observed $\log S_0$ versus $\log S_0$ as predicted by (a) CLOGP alone (Eq. 15) and (b) CLOGP together with ΔH_m and ΔS_m for the 26 drugs investigated in Paper II.

For the $\log P$ -independent dataset, $\log S_0$ was highly correlated to T_m ($R^2=0.70$) and ΔH_m ($R^2=0.71$), but to a lesser extent to ΔS_m ($R^2=0.31$), as displayed in Fig. 11b-d.

From the above displayed results it was concluded that the stability of the crystal as quantified by solid-state properties T_m , ΔH_m and ΔS_m influence the intrinsic solubility of drugs. To what extent these properties contribute to drug solubility is highly compound-specific.

The two above-mentioned examples of datasets constitute one “normal” dataset ($N=26$, Paper II) of drugs; normal in the sense that $\log P$ explains the greater part of the variability in S_0 of this set. Hence, for most compounds in this dataset, poor solvation is the limiting factor for the solubility. However, for a few compounds, $\log P$ alone is not enough – solid-state properties are also needed for a proper estimation of S_0 to be made since it is the strong interactions in the crystal that represent the major limiting factor for the solubility of these compounds. In the “normal” situation the solvation-limited compounds are in majority and the solid-state limited compounds are in minority. Because of this the overall effect of the solid-state properties on the “normal” dataset is small when regarding the dataset as a whole, although it is significant for individual compounds. The real challenge for the future lies in the ability to confidently distinguish between these two groups of poorly soluble compounds, since they need different treatment.

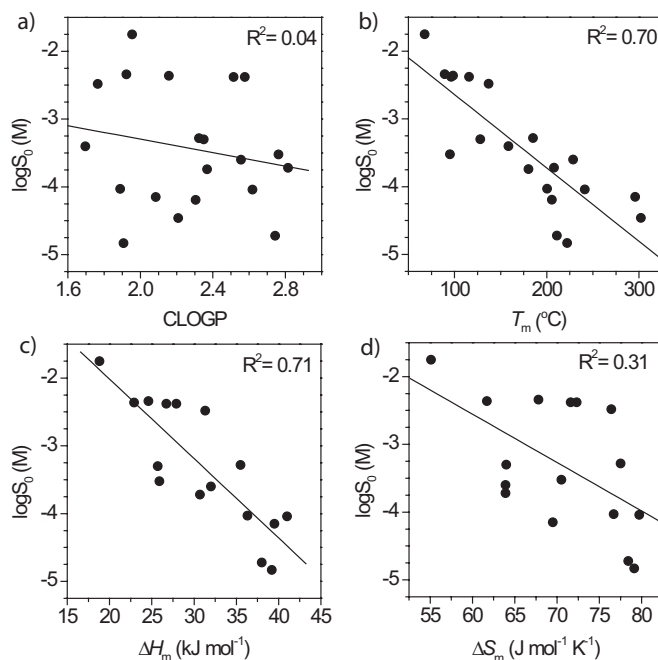


Figure 11. The experimental values for $\log S_0$ for 20 compounds investigated in Paper III correlated to (a) CLOGP, (b) T_m , (c) ΔH_m and (d) ΔS_m .

The second example is an “extreme” dataset (N=20, Paper III) for which the variability in S_0 is totally independent of $\log P$. As a result, the solubility of these compounds must be related entirely to other properties. It was shown in this study that those properties are to a large extent related to the solid state. In particular T_m and ΔH_m were important contributors to S_0 for these compounds. This “extreme” dataset constitute an example of compounds whose solubility is mainly governed by the strength of the interactions in the crystal.

4.5. Application of semi-empirical equations on drugs

4.5.1. The general solubility equation (GSE) (Paper II)

In several studies the GSE (Eq.3 in Section 1.2.1.1.) has proven to be predictive for the solubility of organic compounds and drugs²⁹. The GSE is simple in nature and relies on sound physicochemical reasoning. With simplicity being its main advantage, it has the potential of providing the field of drug discovery with highly accurate and easily interpretable predictions of drug

solubility. Its main drawback is the need for experimental properties ($\log P$ and T_m) as input parameters.

It was considered to be of interest to test the predictivity of GSE on a dataset comprised exclusively of drugs. For this purpose, GSE was used to predict S_0 for 26 drugs (Paper II) for which T_m had been experimentally determined. The results were compared with experimentally determined S_0 . Figure 12a shows the observed (i.e. experimental) $\log S_0$ plotted against the predicted $\log S_0$ (i.e. $\log S_{\text{GSE}}$). The RMSE was 0.9 log units, which is within the range of most solubility models (commonly between 0.7 and 1 log unit) (see Section 1.4.6.). Interestingly, we noted that the standard error (SE) had a negative sign, indicating that the solubility of most of the compounds was overestimated by the GSE. This is an undesirable feature in a solubility model since it increases the risk of insoluble compounds being advanced in drug discovery projects.

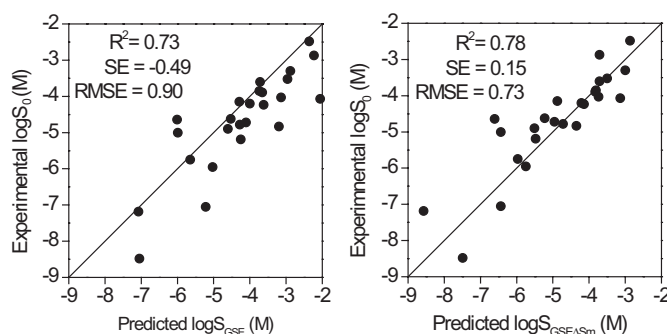


Figure 12. $\log S_0$ as predicted by (a) the GSE (Eq. 3) and (b) the GSE with experimental ΔS_m (Eq. 16) for the 26 compounds in Paper II.

We speculated on whether the reason for the overestimation might be the low weighting of the T_m term (-0.01) compared to the $\log P$ term (-1) in the GSE, which stems from the assumption of constant ΔS_m (see Section 1.2.1.1.) for organic compounds. This results in the solid-state properties having little impact on the predicted solubility. To alleviate the impact of unbalanced weighting, experimentally determined values of ΔS_m were used in the place of the constant value of $56.5 \text{ Jmol}^{-1}\text{K}^{-1}$ in Eq. 16, that was obtained through the combination of Eq. 14 and Eq. 26 from Jain and Yalkowsky's derivation of the GSE²⁸.

$$\log S_0 = 0.5 - \frac{\Delta S_m}{5705.85}(T_m - 25) - \log P \quad (16)$$

The prediction of $\log S_0$ (i.e. $\log S_{\text{GSE}\Delta S_m}$) did indeed improve by using experimental values for ΔS_m and the solubility of these compounds was no longer overpredicted (Fig. 12b). This was believed to be the result of the more appropriate weighting of the T_m term achieved by Eq. 16 as compared to the original version of the GSE.

4.5.2. The Dannenfelser equation (Paper II)

Naturally, the modified form of the GSE (Eq. 16) evaluated above suffers from the same drawback of requiring experimental input parameters as does the original GSE. In an attempt to get around this problem Dannenfelser and Yalkowsky have proposed a semi-empirical equation for the estimation of ΔS_m from two parameters obtained directly from the chemical structure^{62,159} (Eq. 7, Section 1.2.2.2.).

Since this equation had not been validated for drugs before, it was applied to the 26 compounds in Paper II and the results were compared to the experimentally obtained values. The observed versus predicted plot is shown in Figure 13. On average, the ΔS_m was underestimated by $15 \text{ J mol}^{-1} \text{ K}^{-1}$, although the ΔS_m of five compounds (chlorpropamide, glyburide, probenecid, piroxicam and diethylstilbestrol) was overestimated.

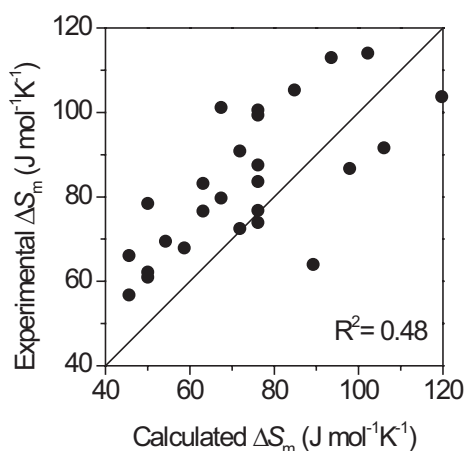


Figure 13. ΔS_m for the 26 drugs investigated in Paper II as predicted by the Dannenfelser semi-empirical equation (Eq. 7).

The Dannenfelser equation provides a simple means for the estimation of ΔS_m directly from molecular structure of a drug. However, it would be desirable to achieve predictions with slightly higher accuracy if it is to be really useful.

4.6. Molecular descriptors for drug solubility

4.6.1. Two dimensions or three? (Papers I and IV)

The information encoded by 2D and 3D molecular descriptors differs in the sense that the effect of the molecular conformation is taken into account by 3D descriptors, but not by 2D ones. Consequently 3D descriptors are more computationally demanding, i.e. they take a longer time to generate. However, the increased complexity of 3D molecular descriptors is not a guarantee that the information content will be higher than for 2D ones. Both 2D and 3D molecular descriptors were used for the model development in each of the Papers I and IV. The aim was to determine whether one or the other of the descriptor types was more predictive and more informative for S_0 and ΔG_{hyd} .

In Paper I, 2D Selma and Molconn-Z descriptors were used as well as 3D surface area descriptors (see Section 3.8.) to develop solubility models for both global and local datasets (see Section 4.3.). In the case of the global dataset, the 2D Selma and Molconn-Z descriptors turned out to be superior to the 3D surface area descriptors, although the most predictive model was achieved by combining both 2D and 3D descriptors in a consensus approach (Table 1). In the case of the local datasets, the 2D Selma descriptors outperformed the 3D surface area descriptors for acids and for ampholytes, however for bases the 3D surface area descriptors were superior.

In Paper IV, 2D Selma descriptors and 3D VolSurf descriptors were used to build models for ΔG_{hyd} . These models were composed by adding together the various contributions from the different forces, that, together, make up ΔG_{hyd} . Thus, ΔG_{hyd} was composed from the respective contributions from cavitation forces (ΔG_{cav}), dispersive and repulsive forces (ΔG_{vdw}), electrostatic forces (ΔG_{ele}) and interaction forces ($\Delta G_{\text{int}} = \Delta G_{\text{vdw}} + \Delta G_{\text{ele}}$) (as described in Section 1.2.2.4.). For all four energy terms, the 2D Selma descriptors gave more predictive models than did the 3D VolSurf ones (Fig. 14).

The reason for the apparent superiority of the 2D descriptors over the 3D descriptors is not immediately apparent. Brown and Martin showed that descriptors of 2D structure produced better separations than those of 3D structure when they were used to distinguish between active and inactive compounds¹⁶⁰. The same authors also found the 2D descriptors to be more useful than 3D ones for predicting $\log P$ and $\text{p}K_{\text{a}}$ ¹⁶¹. The information overlap between 3D ALMOND^{162,163} descriptors and 2D Selma descriptors was investigated by Oprea¹⁶⁴. He found that the first PLS component (which mainly encoded molecular size) showed 60% correlation, while information in the following components (mainly encoding information related to pharmacophoric patterns and hydrogen bonding) was only captured by the 3D ALMOND ones. A similar result was obtained by the same author and others when using 2D Selma and 3D VolSurf descriptors in combination with

ChemGPS to map the medicinal chemistry space with respect to permeability and solubility¹⁶⁵. They recommend the use of 3D VolSurf descriptors for the mapping of properties related to the biopharmaceutics classification system (BCS)⁶.

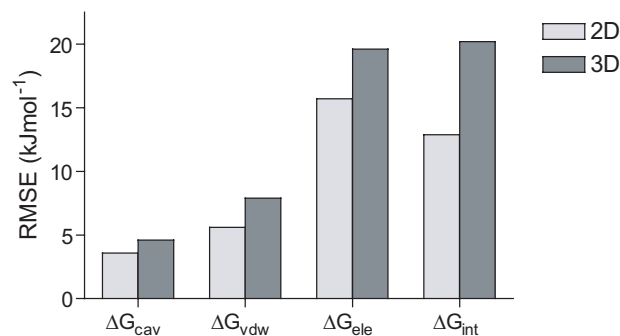


Figure 14. Root mean square error (RMSE) for models of the respective energy terms contributing to the ΔG_{hyd} as modelled by 2D and 3D molecular descriptors.

From the above studies it seems that the 2D molecular descriptors are generally more information rich and have a greater ability to discriminate between closely related chemical structures, but that the 3D molecular descriptors better capture specific information related to receptor-ligand interactions. It is, therefore, advisable to use both 2D and 3D molecular descriptors for the prediction of properties of drugs, as is supported by the results in Paper I where the most predictive global solubility models were achieved for the consensus approach where both 2D and 3D molecular descriptors were used.

4.6.2. The solid state – Log*P*-independent solubility (Paper III)

As mentioned previously, for certain compounds, the solubility is governed by solid-state properties rather than by solvation properties. Paper III was devoted to identifying structural features that describe the solubility of those compounds. This was achieved through the selection of a dataset for which log*P* explained none ($R^2=0.04$; Fig 11a) of the variance in solubility and instead T_m and ΔH_m explained a large proportion ($R^2=0.70$ and $R^2=0.71$; Fig. 11b-c) of the variance in solubility. This dataset is interesting for two reasons. Firstly, it illustrates that the solid-state properties are important determinants of drug solubility, and secondly, it provides us with a tool for the identification of molecular descriptors that are related to solid-state limited solubility.

For the 20 compounds in our log*P*-independent dataset, PLS models were built to describe log*S*₀ using 2D Selma descriptors. With only five descriptors, the best model was able to explain 76% of the variability in log*S*₀ of

this dataset (Fig. 15a). The descriptors were (in decreasing order of importance): the number of rigid bonds, the Balaban index, the number of rigid fragments, the second smallest eigenvalue (Min eV #2) and the third largest eigenvalue (Max eV #3) (Fig. 15c).

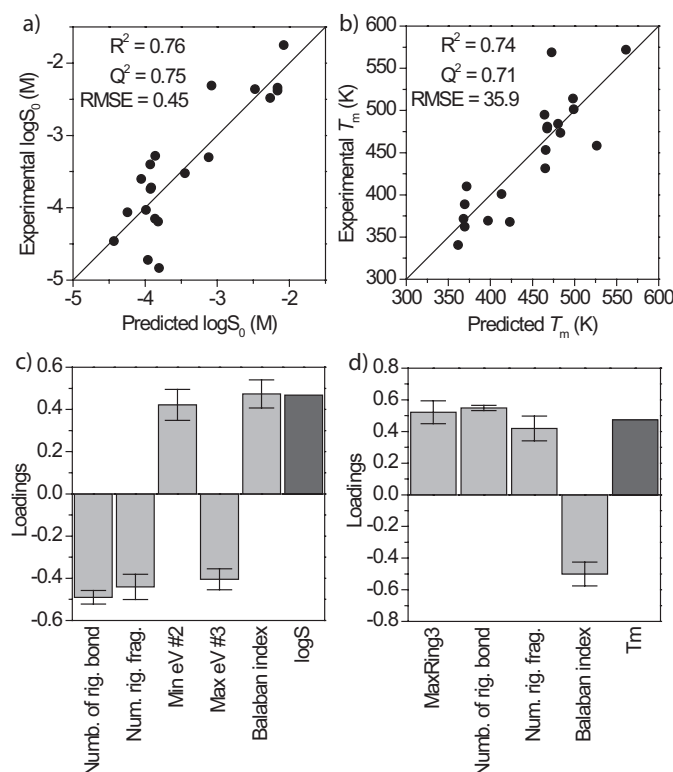


Figure 15. PLS models of $\log S_0$ and T_m for the 20 compounds studied in Paper III. (a) Observed versus predicted $\log S_0$, (b) observed versus predicted T_m , (c) loadings for molecular descriptors included in the final $\log S_0$ model and (d) loadings of molecular descriptors included in the final T_m model. Error bars represent the 95% confidence intervals.

The selected molecular descriptors were interpreted as being related to the rigidity (i.e. lack of flexibility) and to the aromaticity of the molecule. The model predicted that large, flat molecules with an extended ring-structure and conjugated π -systems would be poorly soluble, while small spherically shaped molecules with flexible side-chains would be highly soluble.

In addition, PLS models for the T_m of this dataset were constructed since T_m is such an important factor controlling the solubility (Fig. 15b). They identified descriptors similar to the ones identified by the $\log S_0$ models (Fig. 15d). By examining the intermolecular interactions present in the crys-

tal structures for the compounds in the dataset, we were able to better understand and explain the solubility behaviour predicted by the model. This analysis showed that even though the model considered the effect of non-specific van der Waals interactions present in the crystal, it did not account for the effect of highly specific hydrogen bonding. Consequently there is a need for new molecular descriptors that not only consider single molecules, but also capture the intermolecular interactions. This should enable the effects of hydrogen bonding in crystals to be taken into account.

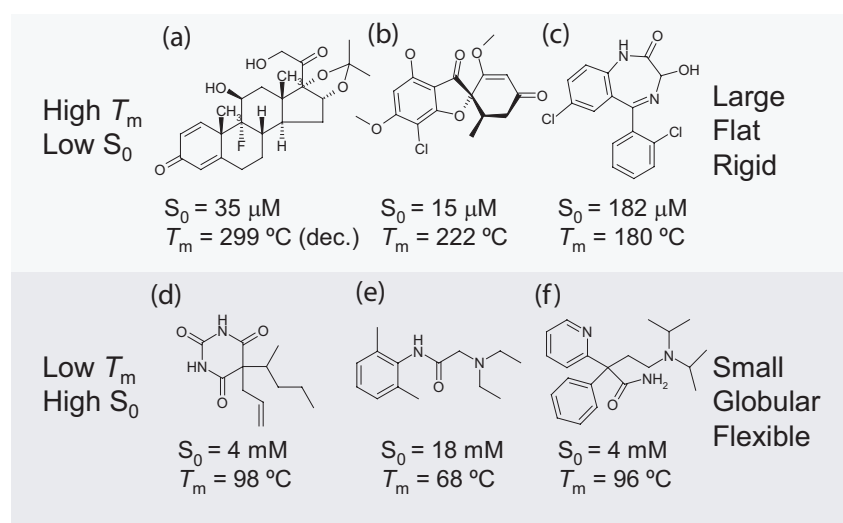


Figure 16 Compounds containing structural features that are indicative of low solubility due to high melting point (top panel) and compounds containing structural features that are indicative of high solubility due to low melting point (bottom panel). Compounds are (a) triamcinolone acetate, (b) griseofulvin, (c) lorazepam, (d) secobarbital, (e) lidocaine and (f) disopyramide.

These results suggest that compounds with a solid-state limited solubility could be identified using molecular descriptors related to rigidity (or the lack of flexibility) and aromaticity. Examples of compounds containing (top panel), and not containing (bottom panel), such structural features are given in the in Fig. 16.

4.6.3 Solvation – The free energy of hydration (Paper IV)

For many reasons it would be appealing to be able to calculate S_0 from first principles by adding the contributions from ΔG_{hyd} and ΔG_{sub} (see Section 1.2.1.). With the accuracy of free energy calculations steadily improving and the advancement of computer technology resulting in ever decreasing calculation times, this might be feasible in the not too distant future.

In Paper IV the possibility of developing QSPR models for simulated ΔG_{hyd} values of 48 drugs was investigated. It was also of interest to identify the structural features that govern the hydration process of drugs. In order to better understand how the chemical structure was related to the different energy terms (ΔG_{cav} , ΔG_{vdw} , ΔG_{ele} and ΔG_{int}) making up ΔG_{hyd} the terms were modelled separately. We found that it was possible to develop models of high accuracy for ΔG_{cav} ($R^2=0.98$), ΔG_{vdw} ($R^2=0.94$) and ΔG_{int} ($R^2=0.91$), while the accuracy was lower for ΔG_{ele} ($R^2=0.75$) (data not shown). The resulting QSPR for ΔG_{hyd} fulfilled our expectations, with a RMSE of 12.0 kJ mol^{-1} , which is equal to 10% of the range. These results are displayed in Fig. 17. Improvements in the ΔG_{ele} model should be expected by incorporating descriptors specifically designed to take into account the electrostatic properties of the molecule, like for instance electrotopological state indices^{166,167}.

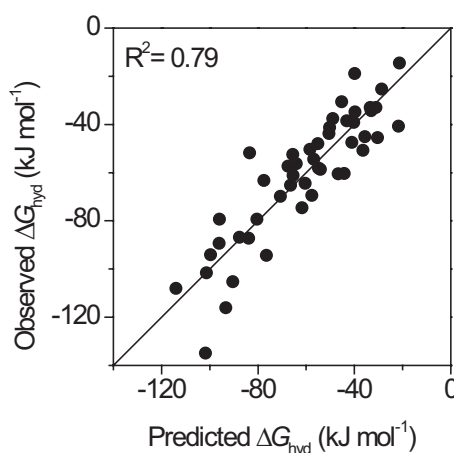


Figure 17. Observed versus predicted ΔG_{hyd} for the 48 drugs modelled in Paper IV. Predicted ΔG_{hyd} values were calculated by summing the predicted ΔG_{cav} and ΔG_{int} values (Eq. 9).

The structural features associated with poor hydration can be summarised as large size, high flexibility, low polarisability and a lack of possibilities to form hydrogen bonds. Molecular size works in opposite direction for ΔG_{cav} and ΔG_{int} , because, while it is energetically more expensive to incorporate a large molecule into the solvent, it is quite comprehensible for drug mole-

cules that an increased molecular size correlates with an increased number of heteroatoms, which are both polarisable and can engage in hydrogen bond interactions. The models were interpreted as selecting small, rigid and polarisable molecules containing hydrogen bond acceptors and donors for favourable interactions with the surrounding water molecules. Consequently these structural features (small, rigid and polarisable) resulted in large negative values of ΔG_{hyd} . In direct contrast to this, large, flexible, non-polarisable molecules with few hydrogen bond acceptors and donors interact poorly with water and, subsequently, have smaller negative values of ΔG_{hyd} .

Solubility depends on the balance between hydration and sublimation. The value of ΔG_{hyd} has to be numerically larger than the value of ΔG_{sub} for dissolution to occur. It is therefore not possible to draw conclusions regarding the solubility of a compound from the value of ΔG_{hyd} alone; ΔG_{hyd} must be considered in connection with the value of ΔG_{sub} for that molecule. However, the molecular descriptors identified above can be used to flag potentially poorly soluble on the grounds of poor hydration.

5. Conclusions

The research conducted for this thesis has involved the development of computational models of aqueous drug solubility. The work presented here includes a presentation of these models, as well as providing an analysis of several aspects related to model development and interpretation. It highlights the importance of selecting a training set that is representative of the chemical space to which the solubility model should apply. Furthermore, solvation was found to be the dominating factor limiting the solubility, although experimental solid-state properties were found to contribute significantly to the solubility of drugs. Calculated molecular descriptors that were associated with $\log P$ -independent solubility were identified, with the intention of making it possible to incorporate the solid-state properties in future solubility models. Finally, the solvation properties of drugs were studied and molecular descriptors related to the poor hydration of drugs were revealed. The specific conclusions were:

- ◇ Global models provide solubility estimations with lower accuracy that are universally applicable, while local models are more accurate, but are restricted to a specific chemical series of compounds.
- ◇ External test sets used for model validation should represent the range of S_0 and the physicochemical properties of the training set if they are to constitute a fair validation.
- ◇ The major limiting factor for drug solubility in general was shown to be the solvation, as quantified by $\log P$. However, solid-state properties indisputably play an important role. The relative importance of solvation and solid state contributions to the solubility was highly compound specific.
- ◇ The contribution to the solubility of drugs from their solid state can be modelled by calculated molecular descriptors related to size, flexibility and aromaticity.
- ◇ The contribution to drug solubility from the free energy of hydration can be modelled by calculated molecular descriptors related to the size, flexibility, polarisability and hydrogen bond potential of the drugs.

Through the identification of groups of molecular descriptors associated with either solid-state limited or solvation limited solubility, it is expected that the results presented here will contribute to the development of computational models for the classification of compounds according to their solubility behaviour and, thereby, the design of rules-of-thumb that could be applied as computational filters. Such filters could provide valuable guidance for decision-making in the early drug discovery.

6. Acknowledgements

Avhandlingsarbetet har utförst vid Institutionen för farmaci på Uppsala universitet med finansiellt stöd ifrån AstraZeneca R&D i Mölndal, IFs stiftelse, Stiftelsen Bengt Lundqvists Minne, Apotekare CD Carlssons stiftelse och Apotekarsocietets resestipendium.

Jag önskar särskilt framföra min stora tacksamhet till:

Mina handledare, **Professor Per Artursson** för dina aldrig sinande idéer, ditt stöd i skrivandet och för att du alltid för fram vår forskning med största tänkbara entusiasm. **Dr. Anders Holmén** för ditt engagemang och stöd, din vetenskapliga strävan efter att veta sanningen och för att du alltid med ett smittande gott humör tagit emot mig på mina vistelser i Mölndal. Särskilt vill jag tacka er båda för er uthållighet och för att ni styrde Titanic på kurs undan isbergen.

Professor Martin Malmsten och **Professor Göran Alderborn** för tillhandahållande av ändamålsenliga lokaler och en god arbetsmiljö.

Mina medförfattare; **Dr. Christel Bergström** för otaliga löslighetsdiskussioner, din enorma framåtanda och problemlösarförmåga och för att du delar mitt (ovanliga?!) intresse för kemiska strukturer, kristaller, och framförallt att du inser vikten av ljudliga skrattsalvor. **Dr. Ismael Zamora** (my third supervisor and Barcelona host) for making multivariate analysis and computational modelling seem simple and comprehensible, for your excellent sense of humour and great support. **Dr. Ulf Norinder** för ingående statistiska diskussioner, en stor portion av bittsk Göteborgshumor och bullar till prediktionsgruppmötena. **Rieke Draheim**, for being the most independent of students, for your intuitiveness and your persistence with the HPLC that never worked!

Alla nya och gamla medlemmar av prediktionsgruppen och cellgruppen för, studsmattetävlingar, bowlingkvällar och grillfester; särskilt tack till **Johan Gråsjö** (Förste forskningsingenjör) för din filosofiska inställning till matematiska och statiska frågor, **Pär Matsson** för dina snygga bilder av den kemiska rymden, **Lucia Lazorova** vår allvetande labbchef med det största

hjärtat i världen, **Dr Eva Ragnarsson** för sällskap på promenader hem efter en långa dagar på jobbet och **Dr. Ina Hubatsch** – vårt HPLC-orakel.

Fyskemgruppen, absorptionsgruppen och alla andra ”kollegor” på Astra-Zeneca i Mölndal för att ni hjälpt mig tillrätta på labbet, för trevligt lunch och fikasällskap samt att ni släppt in mina analyser i redan fulltecknade kölistor. Min förra chef i absorptionsgruppen, **Dr. Anna-Lena Ungell**, utan dig hade det inte blivit något av med doktorerandet för min del. **Dr. Lennart Lindfors** för många, långa och intressanta diskussioner om löslighetsteori och hydratiseringsenergi, **Göran Runberger** för hjälp med programmering.

Mina studenter: **Nina Ginman**, **Lina Stillemark**, **Karin Sundström** och **Kajsa Johansson** för ett hårt jobb i labbet och roliga stunder tillsammans.

Professor Gabriele Criciani and every one at **Laborio di Chemiometria** for making my stay in Perugia so interesting and fun, in particular my flat-mate **Rikke Bergmann** for dinner company, and the never-ending search for a salsa party.

Mina rumskompisar; **Frauke Fichtner** för dina försök att lära mig tyska och Feuerzangenbowle, ”**Farmacie doktorsexamen, Uppsala universitet (Dr.) Sibylle Neuhoff**”, for endless talks in the office, for laughter, cries and for always having a bottle of champagne on ice, whenever needed.

Nuvarande och gamla medarbetare på **Institutionen för farmaci**, särskilt doktorandkollegor för innebandy, finlandsresor, konferenssällskap, Borrel, kräftskivor, fisketävlingar och en härlig stämning; medlemmar in **WC-klubben** och särskilt mina fantastiska vänner **Dr. Christer Tannergren** för otaliga citat och dåliga ordvitsar och **Dr. Niclas Petri** för ett enormt bidrag till min musiksamling och trevligt konsertsällskap.

Mina extrafamiljer i Stockholm: **Linda och Joakim** och **Maria och Martin** för underbar vänskap, rekreation och ett stort stöd samt för att jag alltid får bo hos er när jag är i stan.

Friluftsflickorna på västkusten; **Ulrika Krave**, **Sofi Nielsen** och **Dr. Linda Eriksson** för skidåkning, kajakturer och seglatser; alltid med Gourmet meny.

Min underbara familj; mamma **Mona-Lisa**, pappa **Göran** och bror **Peter** för ert enorma stöd, för att ni alltid stöttat och trott på mig i alla mina val och för att ni alltid ställer upp och flyttar mina möbler över halva landet; släkterna **Wassvik** och **Hermansson** för att ni lärt mig hur pinfärska ”widding” smakar och för att ni försett mig med en stor portion envishet och styrka samt en krokig näsa. Skoltiden är över nu – Jag lovar!

7. References and notes

1. Kola I, Landis J 2004. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 3(8):711-715.
2. DiMasi JA, Hansen RW, Grabowski HG 2003. The price of innovation: new estimates of drug development costs. *J Health Econ* 22(2):151-185.
3. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 46(1-3):3-26.
4. Lipinski CA 2000. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* 44(1):235-249.
5. Pouton CW 2006. Formulation of poorly water-soluble drugs for oral administration: Physicochemical and physiological issues and the lipid formulation classification system. *Eur J Pharm Sci* 29(3-4):278-287.
6. Amidon GL, Lennernäs H, Shah V, P., Crison J, R. 1995. A Theoretical Basis for a Biopharmaceutic Drug Classification: The Correlation of *in Vitro* Drug Product Dissolution and *in Vivo* Bioavailability. *Pharm Res* V12(3):413.
7. van de Waterbeemd H, Gifford E 2003. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov* 2(3):192-204.
8. Di L, Kerns EH 2003. Profiling drug-like properties in discovery research. *Curr Opin Chem Biol* 7(3):402-408.
9. Wunberg T, Hendrix M, Hillisch A, Lobell M, Meier H, Schmeck C, Wild H, Hinzen B 2006. Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov Today* 11(3-4):175-180.
10. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 23(1-3):3.
11. Teague SJ, Davis AM, Leeson PD, Oprea T 1999. The Design of Leadlike Combinatorial Libraries. *Angew Chem Int Ed Engl* 38(24):3743-3748.
12. Keller TH, Pichota A, Yin Z 2006. A practical view of 'druggability'. *Current Opinion in Chemical Biology* 10(4):357.
13. Lajiness MS, Vieth M, Erickson J 2004. Molecular properties that influence oral drug-like behavior. *Curr Opin Drug Discov Devel* 7(4):470-477.
14. Vieth M, Siegel MG, Higgs RE, Watson IA, Robertson DH, Savin KA, Durst GL, Hipskind PA 2004. Characteristic physical properties and structural fragments of marketed oral drugs. *J Med Chem* 47(1):224-232.
15. Wenlock MC, Austin RP, Barton P, Davis AM, Leeson PD 2003. A comparison of physiochemical property profiles of development and marketed oral drugs. *J Med Chem* 46(7):1250-1256.
16. Walters WP, Murcko MA 2002. Prediction of 'drug-likeness'. *Adv Drug Deliv Rev* 54(3):255-271.
17. Viswanadhan VN, Balan C, Hulme C, Cheetham JC, Sun Y 2002. Knowledge-based approaches in the design and selection of compound libraries for drug discovery. *Curr Opin Drug Discov Devel* 5(3):400-406.

18. Oprea TI, Davis AM, Teague SJ, Leeson PD 2001. Is there a difference between leads and drugs? A historical perspective. *J Chem Inf Comput Sci* 41(5):1308-1315.
19. Stouch TR, Kenyon JR, Johnson SR, Chen XQ, Doweiko A, Li Y 2003. In silico ADME/Tox: why models fail. *J Comput Aided Mol Des* 17(2-4):83-92.
20. Davis AM, Riley RJ 2004. Predictive ADMET studies, the challenges and the opportunities. *Curr Opin Chem Biol* 8(4):378.
21. Scatchard G 1931. Equilibria in non-electrolyte solutions in relation to the vapor pressures and densities of the components. *Chem Rev* 8:321-333.
22. Hildebrand JH, Scott RL. 1950. Solubility of Nonelectrolytes. New York: Reinhold.
23. Grant DJW, Higuchi T. 1990. Solubility Behaviour of Organic Compounds. New York: Wiley Interscience. p 386-387.
24. Hansch C, Quinn JE, Lawrence GL 1968. The Linear Free-Energy Relationship between Partition Coefficients and the Aqueous Solubility of Organic Liquids. *J Org Chem* 33(1):347-350.
25. Irmann F 1965. Eine einfache Korrelation zwischen Wasserlöslichkeit und Struktur von Kohlenwasserstoffen und Halogenkohlenwasserstoffen. *CHEM-ING-TECH* 37(8):789-798.
26. Walden P 1908. Über die schmelzwärme, spezifische kohäsion und molekulargröße bei der schmelztemperatur. *Z Electrochem* 14:713-724.
27. Yalkowsky SH, Valvani SC 1980. Solubility and partitioning I: Solubility of nonelectrolytes in water. *J Pharm Sci* 69(8):912-922.
28. Jain N, Yalkowsky SH 2001. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J Pharm Sci* 90(2):234-252.
29. Ran Y, He Y, Yang G, Johnson JL, Yalkowsky SH 2002. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere* 48(5):487-509.
30. Abraham MH, Le J 1999. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J Pharm Sci* 88(9):868-880.
31. Abraham MH, McGowan JC 1987. The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. *Chromatographia* 23(4):243-246.
32. Abraham MH. ABSOLV. ADMEBoxes, Pharma Algorithms, Toronto, Ontario, Canada <http://www.ap-algorithms.com/absolv.htm>.
33. Banerjee S 1985. Calculation of Water Solubility of Organic Compounds with UNIFAC-Derived Parameters. *Environ Sci Technol* 19(4):369-370.
34. Kan AT, Tomson MB 1996. UNIFAC Prediction of Aqueous and Nonaqueous Solubilities of Chemicals with Environmental Interest. *Environ Sci Technol* 30(4):1369-1376.
35. Myrdal PB, Ward GH, Simamora P, Yalkowsky SH 1993. Aquafac: aqueous functional group activity coefficients. *SAR QSAR Environ Res* 1:53-61.
36. Ruelle P, Kesselring UW 1998. The hydrophobic effect. 2. Relative importance of the hydrophobic effect on the solubility of hydrophobes and pharmaceuticals in H-bonded solvents. *J Pharm Sci* 87(8):998-1014.
37. Ruelle P, Kesselring UW 1998. The hydrophobic effect. 1. A consequence of the mobile order in H-bonded liquids. *J Pharm Sci* 87(8):987-997.
38. Pliska V, Testa B, Van de Waterbeemd H editors. 1996. Lipophilicity in Drug Action and Toxicology. In *Methods and Principles in Medicinal Chemistry*. 4. Mannold R, Kubinyi H, Timmerman H editors., Weinheim, Germany: VCH.

39. Erös D, Kövesdi I, Örfi L, Takács-Novák K, Ascády G, Kéri G 2002. Realability of logP Predictions Based on Calculated Molecular Descriptors: A Critical Review. *Curr Med Chem* 9:1891-1829.
40. Fujita T, Iwasa J, Hansch C 1964. A New Substituent Constant, π , Derived from Partition Coefficients. *J Am Chem Soc* 86(23):5175-5180.
41. Mannhold R, Rekker RF 2000. The hydrophobic fragmental constant approach for calculating log P in octanol/water and aliphatic hydrocarbon/water systems. *Perspect Drug Discov* 18(1):1-18.
42. Iwase K, Komata K, Hirono S, Nakagawa S, Moriguchi I 1985. Estimation of hydrophobicity based on the solvent-accessible surface area of molecules. *Chem Pharm Bull* 33(5):2114-2121.
43. Bodor N, Gabanyi Z, Wong C-K 1989. A New Method for the Estimation of Partition Coefficient. *J Am Chem Soc* 111(11):3783-3786.
44. Kamlet M, Abboud J-L, Abraham MH, Taft RW 1983. Linear Solvation Energy Relationships. 23. A Comprehensive Collection of the Solvatochromic Parameters, π^* , α , and β , and Some Methods for Simplifying the Generalized Solvatochromic Equation. *J Org Chem* 48(17):2877-2887.
45. Hansch C, Leo A. 1979. Substituent constants for correlation analysis in chemistry and biology., New York: John Wiley & Sons.
46. Meylan W. 1993-1997. KOWWIN for Microsoft Windows, Version 1.60. Syracuse, NY, USA: Syracuse, Research Corp.
47. 1999. SciLogP Ultra Version 1.5. Scivision, www.scivision.com.
48. Tetko IV, Bruneau P 2004. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *J Pharm Sci* 93(12):3103-3110.
49. ACD Labs logP suit is available at <http://acdlabs.com>.
50. CLOGP v. 4.71 (<http://www.biobyte.com>).
51. Tetko IV. ALOGPS is available free on-line at <http://www.vcclab.org>.
52. Tetko IV, Poda GI 2004. Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds. *J Med Chem* 47(23):5601-5604.
53. Tsantili-Kakoulidou A, Panderi I, Csizmadia F, Darvas F 1997. Prediction of distribution coefficient from structure. 2. Validation of Prolog D, an expert system. *J Pharm Sci* 86(10):1173-1179.
54. Jain N, Yalkowsky SH 1999. UPPER III: unified physical property estimation relationships. Application to non-hydrogen bonding aromatic compounds. *J Pharm Sci* 88(9):852-860.
55. Karthikeyan M, Glen RC, Bender A 2005. General melting point prediction based on a diverse compound data set and artificial neural networks. *J Chem Inf Model* 45(3):581-590.
56. Godavarthy SS, Robinson RL, Gasem KAM 2006. An Improved Structure-Property Model for Predicting Melting-Point Temperatures. *Ind Eng Chem Res* 45(14):5117-5126.
57. Bergström CAS, Norinder U, Luthman K, Artursson P 2003. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J Chem Inf Comput Sci* 43(4):1177-1185.
58. Modarresi H, Dearden JC, Modarress H 2006. QSPR correlation of melting point for drug compounds based on different sources of molecular descriptors. *J Chem Inf Model* 46(2):930-936.
59. Bergström CAS, Norinder U, Luthman K, Artursson P 2002. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm Res* 19(2):182-188.

60. Jain A, Yalkowsky SH 2006. Estimation of melting points of organic compounds-II. *J Pharm Sci* 95(12):2562-2618.
61. Dannenfelser R-M, Surendran N, Yalkowsky SH 1993. Molecular symmetry and related properties. *SAR QSAR Environ Res* 1:273-292.
62. Dannenfelser RM, Yalkowsky SH 1996. Estimation of entropy of melting from molecular structure - a non-group contribution method. *Ind Eng Chem Res* 35(4):1483-1486.
63. Jain A, Yang G, Yalkowsky SH 2004. Estimation of Total Entropy of Melting of Organic Compounds. *Ind Eng Chem Res* 43(15):4376-4379.
64. Chickos JS, Acree JWE, Liebman JF 1999. Estimating Solid-Liquid Phase Change Enthalpies and Entropies. *J Phys Chem Ref Data* 28(6):1535.
65. Perlovich GL, Kurkov SV, Hansen LKR, Bauer-Brandl A 2004. Thermodynamics of sublimation, crystal lattice energies, and crystal structures of racemates and enantiomers: (+)- and (+/-)-ibuprofen. *J Pharm Sci* 93(3):654-666.
66. Perlovich GL, Volkova TV, Bauer-Brandl A 2006. Towards an understanding of the molecular mechanism of solvation of drug molecules: A thermodynamic approach by crystal lattice energy, sublimation, and solubility exemplified by paracetamol, acetanilide, and phenacetin. *J Pharm Sci* 95(10):2158-2169.
67. Charlton MH, Docherty R, Hutchings MG 1995. Quantitative structure-sublimation enthalpy relationship studied by neural networks, theoretical crystal packing calculations and multilinear regression analysis. *J Chem Soc Perk T* 2 (11):2023-2030.
68. Ouvrard C, Mitchell JBO 2003. Can we predict lattice energy from molecular structure? *Acta Crystallogr B* 59(5):676-685.
69. Datta S, Grant DJ 2004. Crystal structures of drugs: advances in determination, prediction and engineering. *Nat Rev Drug Discov* 3(1):42-57.
70. Lommerse JP, Motherwell WD, Ammon HL, Dunitz JD, Gavezzotti A, Hofmann DW, Leusen FJ, Mooij WT, Price SL, Schweizer B, Schmidt MU, van Eijck BP, Verwer P, Williams DE 2000. A test of crystal structure prediction of small organic molecules. *Acta Crystallogr B* 56(Pt 4):697-714.
71. Motherwell WD, Ammon HL, Dunitz JD, Dzyabchenko A, Erk P, Gavezzotti A, Hofmann DW, Leusen FJ, Lommerse JP, Mooij WT, Price SL, Scheraga H, Schweizer B, Schmidt MU, van Eijck BP, Verwer P, Williams DE 2002. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallogr B* 58(Pt 4):647-661.
72. Day GM, Motherwell WDS 2006. An Experiment in Crystal Structure Prediction by Popular Vote. *Cryst Growth Des* 6(9):1985-1990.
73. Orozco M, Luque FJ 2000. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem Rev* 100(11):4187-4226.
74. Wan S, Stote RH, Karplus M 2004. Calculation of the aqueous solvation energy and entropy, as well as free energy, of simple polar solutes. *J Chem Phys* 121(19):9539-9548.
75. Nagy PI 2004. Effects of the Solute Model and Concentration on the Calculated Free Energy of Hydration in Explicit Solvent Solution. *J Phys Chem B* 108(30):11105-11117.
76. Duffy EM, Jorgensen WL 2000. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J Am Chem Soc* 122(12):2878-2888.
77. Luque FJ, Curutcheta C, Muñoz-Muriedasa J, Bidon-Chanal A, Soterasa I, Morrealeb A, Gelpib JL, Orozco M 2003. Continuum solvation models: Dissecting the free energy of solvation. *Phys Chem Chem Phys* 5:3827-3836.

78. Shimizu K, Freitas AA, Farah JP, Dias LG 2005. Predicting hydration free energies of neutral compounds by a parametrization of the polarizable continuum model. *J Phys Chem A* 109(49):11322-11327.
79. Kelly CP, Cramer CJ, Truhlar DG 2005. SM6: A Density Functional Theory Continuum Solvation Model for Calculating Aqueous Solvation Free Energies of Neutrals, Ions, and Solute-Water Clusters. *J Chem Theory Comput* 1(6):1133-1152.
80. Curutchet C, Bidon-Chanal A, Soteras I, Orozco M, Luque FJ 2005. MST continuum study of the hydration free energies of monovalent ionic species. *J Phys Chem B* 109(8):3565-3574.
81. Jorgensen WL, Ulmschneider JP, Tirado-Rives J 2004. Free Energies of Hydration from a Generalized Born Model and an All-Atom Force Field. *J Phys Chem B* 108(41):16264-16270.
82. Hine J, Mookerjee PK 1975. Structural effects on rates and equilibria. XIX. Intrinsic hydrophilic character of organic compounds. Correlations in terms of structural contributions. *J Org Chem* 40(3):292-298.
83. Viswanadhan VN, Ghose AK, Singh UC, Wendoloski JJ 1999. Prediction of Solvation Free Energies of Small Organic Molecules: Additive-Constitutive Models Based on Molecular Fingerprints and Atomic Constants. *J Chem Inf Comput Sci* 39(2):405-412.
84. Mansson RA, Frey JG, Essex JW, Welsh AH 2005. Prediction of properties from simulations: a re-examination with modern statistical methods. *J Chem Inf Model* 45(6):1791-1803.
85. Katritzky AR, Oliferenko AA, Oliferenko PV, Petrukhin R, Tatham DB, Maran U, Lomaka A, Acree WE, Jr. 2003. A general treatment of solubility. 1. The QSPR correlation of solvation free energies of single solutes in series of solvents. *J Chem Inf Comput Sci* 43(6):1794-1805.
86. Pudipeddi M, Serajuddin AT 2005. Trends in solubility of polymorphs. *J Pharm Sci* 94(5):929-939.
87. Bevan CD, Lloyd RS 2000. A high-throughput screening method for the determination of aqueous drug solubility using laser nephelometry in microtiter plates. *Anal Chem* 72(8):1781-1787.
88. Di L, Kerns EH 2006. Biological assay challenges from compound solubility: strategies for bioassay optimization. *Drug Discov Today* 11(9-10):446-451.
89. Glomme A, Marz J, Dressman JB 2005. Comparison of a miniaturized shake-flask solubility method with automated potentiometric acid/base titrations and calculated solubilities. *J Pharm Sci* 94(1):1-16.
90. Kitchen DB, Stahura FL, Bajorath J 2004. Computational techniques for diversity analysis and compound classification. *Mini Rev Med Chem* 4(10):1029-1029.
91. Dobson CM 2004. Chemical space and biology. *Nature* 432(7019):824-828.
92. Kishi H, Hashimoto Y 1989. Evaluation of the procedures for the measurement of water solubility and n-octanol/water partition coefficient of chemicals results of a ring test in Japan. *Chemosphere* 18(9-10):1749-1759.
93. Yalkowsky SH, Dannenfelser R-M 1991. *AQUASOL dATABASE of Aqueous Solubility*, 5th edition.
94. Myrdal PB, Manka AM, Yalkowsky SH 1995. AQUAFAC 3: aqueous functional group activity coefficients; application to the estimation of aqueous solubility. *Chemosphere* 30(9):1619-1637.
95. Loftsson T, Hreinsdottir D 2006. Determination of aqueous solubility by heating and equilibration: a technical note. *AAPS PharmSciTech* 7(1):E1-E4.

96. Jorgensen WL, Duffy EM 2002. Prediction of drug solubility from structure. *Adv Drug Deliv Rev* 54(3):355-366.
97. Legendre AM. 1805. Sur la Méthode des moindres quarrés. In *Nouvelles méthodes pour la détermination des orbites des comètes*.
98. Gauss CF. 1809. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*.
99. Draper NR, Smith H. 1998. *Applied Regression Analysis*. In *Wiley Series in Probability and Statistics*, New York, USA: Wiley. p 736.
100. Wold H. 1966. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, Krishnaiah PR, editor, New York: Academic Press. p 391-420.
101. Wold S, Sjöström M, Eriksson L. 1999. PLS in Chemistry. In *The Encyclopedia of Computational Chemistry*, Schleyer PvR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR, editors., Chichester, UK: John Wiley & Sons. p 2006-2020.
102. Wold S, Sjöström M, Eriksson L 2001. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58(2):109.
103. McCulloch W, Pitts W 1943. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 7:115 - 133.
104. Leardi R editor 2003. *Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks*. In *Data Handling in Science and Technology*. 23, Amsterdam, The Netherlands: Elsevier.
105. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. 2001. *Multi- and Megavariate Data Analysis - Principles and Applications*. Umeå: Umetrics AB. p 533.
106. Pearson K 1901. On lines and planes of the closest fit to systems of points in space. *Phil Mag* 6(2):559-572.
107. Kuhne R, Ebert RU, Kleint F, Schmidt G, Schuurmann G 1995. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* 30(11):2061.
108. Klopman G, Zhu H 2001. Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J Chem Inf Comput Sci* 41(2):439-445.
109. Clark M 2005. Generalized Fragment-Substructure Based Property Prediction Method. *J Chem Inf Model* 45(1):30-38.
110. Hou TJ, Xia K, Zhang W, Xu XJ 2004. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J Chem Inf Comput Sci* 44(1):266-275.
111. McElroy NR, Jurs PC 2001. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J Chem Inf Comput Sci* 41(5):1237-1247.
112. Gao H, Shanmugasundaram V, Lee P 2002. Estimation of Aqueous Solubility of Organic Compounds with QSPR Approach. *Pharm Res* 19(4):497-503.
113. Catana C, Gao H, Orrenius C, Stouten PFW 2005. Linear and nonlinear methods in modeling the aqueous solubility of organic compounds. *J Chem Inf Model* 45(1):170-176.
114. Bruneau P 2001. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J Chem Inf Model* 41(6):1605-1616.
115. Huuskonen J 2000. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J Chem Inf Comput Sci* 40(3):773-777.

116. PHYSPROP. 1994. Physical/Chemical Property Database. SRC, Environmental Science Center, Syracuse, NY, USA <http://www.syrres.com/esc/physprop.htm>.
117. Tetko IV, Tanchuk VY, Kasheva TN, Villa AEP 2001. Estimation of aqueous solubility of chemical compounds using E-state indices. *J Chem Inf Comput Sci* 41(6):1488-1493.
118. Liu R, So S-S 2001. Development of Quantitative Structure-Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J Chem Inf Comput Sci* 41(6):1633-1639.
119. Yan A, Gasteiger J 2003. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J Chem Inf Comput Sci* 43(2):429-434.
120. Yan A, Gasteiger J, Krug M, Anzali S 2004. Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *J Comput Aided Mol Des* 18(2):75-87.
121. Balakin KV, Savchuk NP, Tetko IV 2006. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions. *Curr Med Chem* 13(2):223-241.
122. Butina D, Gola JM 2003. Modeling aqueous solubility. *J Chem Inf Comput Sci* 43(3):837-841.
123. Delaney JS 2004. ESOL: Estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci* 44(3):1000-1005.
124. Cheng A, Merz KM, Jr. 2003. Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships. *J Med Chem* 46(17):3572-3580.
125. Votano JR, Parham M, Hall LH, Kier LB, Hall LM 2004. Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem Biodivers* 1(11):1829-1841.
126. Lind P, Maltseva T 2003. Support vector machines for the estimation of aqueous solubility. *J Chem Inf Comput Sci* 43(6):1855-1859.
127. Xia X, Maliski E, Cheetham J, Poppe L 2003. Solubility prediction by recursive partitioning. *Pharm Res* 20(10):1634-1640.
128. Yamashita F, Itoh T, Hara H, Hashida M 2006. Visualization of large-scale aqueous solubility data using a novel hierarchical data visualization technique. *J Chem Inf Model* 46(3):1054-1059.
129. Huuskonen J, Salo M, Taskinen J 1997. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J Pharm Sci* 86(4):450-454.
130. Huuskonen J, Salo M, Taskinen J 1998. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J Chem Inf Comput Sci* 38(3):450-456.
131. Chen X-Q, Cho SJ, Li Y, Venkatesh S 2002. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *J Pharm Sci* 91(8):1838-1852.
132. McFarland JW, Avdeef A, Berger CM, Raevsky OA 2001. Estimating the Water Solubilities of Crystalline Compounds from Their Chemical Structures Alone. *J Chem Inf Comput Sci* 41(5):1355-1359.
133. Raevsky OA, Raevskaja OE, Schaper K-J 2004. Analysis of Water Solubility Data on the Basis of HYBOT Descriptors. Part 3. Solubility of Solid Neutral Chemicals and Drugs. *QSAR & Comb Sci* 23(5):327-343.
134. Jorgensen WL, Duffy EM 2000. Prediction of drug solubility from Monte Carlo simulations. *Bioorg Med Chem Lett* 10(11):1155-1158.

135. Göller AH, Hennemann M, Keldenich J, Clark T 2006. In Silico Prediction of Buffer Solubility Based on Quantum-Mechanical and HQSAR- and Topology-Based Descriptors. *J Chem Inf Model* 46(2):648-658.
136. Klamt A, Eckert F, Hornig M, Beck ME, Burger T 2002. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J Comput Chem* 23(2):275-281.
137. Katritzky AR, Tulp I, Fara DC, Lauria A, Maran U, Acree WE, Jr. 2005. A general treatment of solubility. 3. Principal component analysis (PCA) of the solubilities of diverse solutes in diverse solvents. *J Chem Inf Model* 45(4):913-923.
138. Thompson JD, Cramer CJ, Truhlar DG 2003. Predicting aqueous solubilities from aqueous free energies of solvation and experimental or calculated vapor pressures of pure substances. *J Chem Phys* 119(3):1661-1670.
139. Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, MacFaul P, Wood JM, Colclough N, Law B 2006. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J Med Chem*.
140. Delaney JS 2005. Predicting aqueous solubility from structure. *Drug Discov Today* 10(4):289-295.
141. Bergstrom CA 2005. Computational models to predict aqueous drug solubility, permeability and intestinal absorption. *Expert Opin Drug Metab Toxicol* 1(4):613-627.
142. Bergström CAS 2005. In silico predictions of drug solubility and permeability: two rate-limiting barriers to oral drug absorption. *Basic Clin Pharmacol Toxicol* 96(3):156-161.
143. Johnson SR, Zheng W 2006. Recent Progress in the Computational Prediction of Aqueous Solubility and Absorption. *AAPS Journal* 8(1):E27-E40.
144. Oprea TI, Gottfries J 2001. Chemography: The art of navigating in chemical space. *J Comb Chem* 3(2):157-166.
145. Olsson T, Sherbuhkin V. Synthesis and Structure Administration. AstraZeneca R&D Mölndal, Sweden.
146. Westergren J, Lindfors L, Höglund T, Lüder K, Nordholm S, Kjellander R 2006. In Silico Prediction of Drug Solubility - 1. Free Energy of Hydration. *J Phys Chem B*, Submitted.
147. Jorgensen WL, Tirado-Rives J 2005. Molecular modeling of organic and biomolecular systems using *BOSS* and *MCPRO*. *J Comput Chem* 26(16):1689-1700.
148. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML 1983. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926-935.
149. Jorgensen WL, Maxwell DS, Tirado-Rives J 1996. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J Am Chem Soc* 118(45):11225-11236.
150. Dewar MJS, Zebisch EG, Healy EF, Stewart JJP 1985. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J Am Chem Soc* 107(13):3902-3909.
151. Storer JW, Giesen DJ, Cramer CJ, Truhlar DG 1995. Class IV charge models: A new semiempirical approach in quantum chemistry. *J Comput Aided Mol Des* 9(1):87.
152. Molconn-Z Version 3 15S. Quincy, MA, USA: Hall Associates Consulting.

153. MAREA Version 2.4. The program MAREA is available from the authors upon request. It is provided free of charge for academic users. Contact Johan Gråsjö (e-mail: johan.grasjo@farmaci.uu.se).
154. Sadowski J. 2004. 3D Structure Generation. In Handbook of chemoinformatics: from data to knowledge, Gasteiger J, editor, Weinheim: Wiley-VCH. p 231-261.
155. Gasteiger J, Rudolph C, Sadowski J 1990. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comp Method* 3(6, Part 3):537-547.
156. Cruciani G, Pastor M, Clementi S. 2000. Handling information from 3D grid maps for QSAR studies. In *Molecular Modeling and Prediction of Bioactivity*, Gundertofte K, Jorgensen FS, editors.: Springer - Verlag. p 73-82.
157. Cruciani G, Crivori P, Carrupt PA, Testa B 2000. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *THEOCHEM* 503(1-2):17.
158. No homologous series of bases was found within the external test set to generate a model for a series of homologous bases in Paper I. Hence, the solubility values of β -receptor antagonists found in the training and test sets combined with shake flask solubility values of such analogues determined in our laboratory were used. Except for the 8 compounds in the training and test sets the following compounds and solubility values (given as logS0 (M)) were used: bunitrolol -1.57, carvedilol -6.14, celiprolol -1.94, oxprenolol -2.44, pafenolol -2.78, timolol -1.45.
159. Dannenfelser RM, Yalkowsky SH 1999. Predicting the total entropy of melting: Application to pharmaceuticals and environmentally relevant compounds. *J Pharm Sci* 88(7):722-724.
160. Brown RD, Martin YC 1996. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J Chem Inf Comput Sci* 36(3):572-584.
161. Brown RD, Martin YC 1997. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J Chem Inf Comput Sci* 37(1):1-9.
162. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S 2000. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* 43(17):3233-3243.
163. ALMOND. Molecular Discovery Ltd., 215 Marsh Road, 1st Floor HA5 5NE, Pinner, Middlesex, UK; http://www.moldiscovery.com/soft_almond.php.
164. Oprea TI 2002. On the Information Content of 2D and 3D Descriptors for QSAR. *J Braz Chem Soc* 13(6):811-815.
165. Oprea TI, Zamora I, Ungell AL 2002. Pharmacokinetically based mapping device for chemical space navigation. *J Comb Chem* 4(4):258-266.
166. Kier LB, Hall LH 1990. An electrotopological-state index for atoms in molecules. *Pharm Res* 7(8):801-807.
167. Hall LH, Kier LB 1995. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valance State Information. *J Chem Inf Comput Sci* 35(6):1039-1045.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Pharmacy 44*

Editor: The Dean of the Faculty of Pharmacy

A doctoral dissertation from the Faculty of Pharmacy, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-7334



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006