Prediction of Aqueous Solubility of Drugs and Pesticides with COSMO-RS

ANDREAS KLAMT,[1] FRANK ECKERT,[1] MARTIN HORNIG,[1] MICHAEL E. BECK,[2] THORSTEN BÜRGER[2]

[1] *COSMOlogic GmbH&Co.KG, Burscheider Str. 515, 51381 Leverkusen, Germany*

[2] *Bayer AG, Agricultural Centre, Alfred-Nobel Str. 50, 40789 Monheim, Germany*

Received 10 May 2001; Accepted 3 July 2001

Abstract: The COSMO-RS method, originally developed for the prediction of liquid–liquid and liquid–vapor equilibrium constants based on quantum chemical calculations, has been extended to solid compounds by addition of a heuristic expression for the Gibbs free energy of fusion. By this addition, COSMO-RS is now capable of *a priori* prediction of aqueous solubilities of a wide range of typical neutral drug and pesticide compounds. Only three parameters in the heuristic expression have been fitted on a data set of 150 drug-like compounds. On these data an rms deviation of 0.66 log-units was achieved. Later, the model was tested on a set of 107 pesticides, which have been critically selected based on two experimental data sources and by a crosscheck with an independent HQSAR model. On this data set an rms of 0.61 log-units was achieved, without any adjustments to the structurally extremely diverse pesticides. This result verifies the ability of this extended COSMO-RS to predict aqueous solubilities of drugs and pesticides of almost arbitrary structural classes. The new method is COSMO-RSol.

 J Comput Chem 23: 275–281, 2002

Key words: aqueous solubility; drugs; pesticides; COSMO-RS; prediction; ADME

Introduction

Aqueous solubility (S_{aq}) is a key property for all chemical compounds considered as potential agents in life science, for example, for drugs or pesticides. S_{aq} largely determines the availability of the compounds *in vivo*, as well as the environmental behavior of the compounds if released to the environment.

Due to its importance, there have been many attempts to find reliable methods for the prediction of S_{aq} of new compounds[1–5] because such methods are of great value in the process of finding promising candidates for new agents. But within the different properties considered for the characterization of absorption, distribution, metabolism, and environmental fate (ADME) of agents, S_{aq} is one of the hardest target for prediction. This is due to the fact that the prediction of solubility by theoretical computational models involves the chemical potential of a compound in its pure state, i.e., in its pure liquid or crystal. In contrast to the chemical potential μ_S^X of a compound X in a given solvent S, for example, in water, octanol, etc., which usually can be approximated quite well by a sum of fragment contributions, the chemical potential μ_X^X of a pure compound X involves the compound as solute and as solvent. This causes a strong nonlinearity of μ_X^X with respect to its structural composition. For example, addition of an acceptor group may increase or decrease the solubility of a new compound compared to a similar precursor, depending on a delicate balance between donors and acceptors. Therefore, all attempts to develop linear incremental schemes for log S_{aq}, similar to the very successful increment methods for log P_{ow},[6, 7] had only limited success. Even linear regression methods based on more complex "fingerprints" of the compounds typically are only applicable to certain substance classes. A neural network approach was reported,[3] which appears to be slightly more successful, probably due to its ability to catch some of the nonlinearity of the problem. But due to the lack of any physical model, neural networks require a large training set, and they are always in danger of overfitting the data, i.e., to learn from all the noise and error covered in the typically not very clean experimental data sets for S_{aq}.

On the other hand, the situation for a rigorous calculation of the chemical potentials or free energies of compounds in water and in their pure state appears to be rather hopeless. While the calculation of the free energy of a compound X in water may be possible with substantial, but finite effort by molecular dynamics (MD) or Monte Carlo (MC) free-energy perturbation methods starting from an equilibrated water ensemble, the calculation of the free energy in the pure state of X is extremely expensive. If X is a liquid, this requires the generation and full equilibration of a large ensemble of molecules by MD/MC methods. But typically drugs and pesticides are crystalline, and hence, the predictive calculation of the free energy of the pure compound would require simultaneous pre-

Correspondence to: A. Klamt; e-mail: andreas.klamt@cosmologic.de

DOI 10.1002/jcc.1168

diction of the crystal structure together with an accurate calculation of the corresponding total free energy. Although successful crystal structure predictions have been reported for a few relatively simple drug and dye compounds, such procedures are extremely time consuming, and they appear to be far away from becoming routinely applicable methods.[8] Finally, it should be noted that any force-field-based MD/MC method requires a quantum chemical calculation of the compound X as initial step to yield reasonable partial charges.

In this article we want to introduce a very efficient and physically well-founded alternative approach, which involves the COSMO-RS method[9–12] as the central source of chemical potentials. COSMO-RS is a combination of the continuum solvation method COSMO[13] with a very efficient and accurate statistical thermodynamics of interacting surfaces. Starting from the surface polarization charge densities σ from density functional (DFT) COSMO calculations, COSMO-RS considers all interactions in a liquid system as contact interactions of the molecular surfaces. The interaction energies, especially the electrostatic interactions and hydrogen bonding, are written as pair interactions of the respective polarization charge densities σ and σ' of contacting surfaces. Then the ensemble of interacting molecules is replaced by the corresponding system of geometrically independent, surface segments. Under the condition that all of these surface segments have to form pairs, i.e., that there is no free surface in the system, the statistical thermodynamics of this ensemble can be solved exactly within milliseconds, practically independent of the size of the system. This leads to the chemical potential μ_S^X of any solvent or solute molecule X in the liquid ensemble S. Thus, given the DFT-COSMO calculations for a molecule X the chemical potential of X in water and in the pure liquid X can be calculated in milliseconds. In combination with a simple linear regression for the heat of crystallization (the free enthalpy of fusion ΔG_{fus}) the COSMO-RS method can be applied to crystalline compounds.

The COSMO-RS kind of thermodynamics of pair-wise interacting surface segments is unusual in computational chemistry, but it has been successfully used by the chemical engineers over 3 decades, because it is the basis of models like UNIFAC and UNIQUAC.[14, 15] If we accept, that at least for electronically demanding molecules the initial step for a good calculation has to be a quantum chemical (QC) calculation for the electrostatics, which may be a HF/6-31-G** calculation as often used in the context of force fields, or a DFT/SVP/COSMO calculation for COSMO-RS, then the relation between both approaches becomes clearer. The second step in MD/MC is the reduction of the real quantum chemical system to an ensemble of pair-wise interacting spheres, having certain interaction parameters that are derived from the initial QC step. Instead, in COSMO-RS we represent the system by surface pieces, with the interaction parameters derived from QC. Both approaches require a pair-wise interaction functional. In the first case, this functional typically is of force field type, while in COSMO-RS it is represented by very simple interaction formulae for electrostatic misfit and hydrogen bonding. We should be aware that in either case these interaction functionals are only an approximation to the real physics of interaction. MD/MC has to solve the statistical thermodynamics of the problem by an extremely demanding, exhaustive sampling of the phase space of a large ensemble of molecules, taking into account periodic boundary conditions to avoid artifacts. Instead, COSMO-RS has an exact, algebraic solution for the thermodynamics of such ensemble of pair-wise interacting surface segments.

Theory

General COSMO-RS Theory

The theory of COSMO-RS has been described in detail in several article.[9–12] Therefore, we only will give a short survey of the basic concept here, and refer the interested reader to these articles for details.

The starting point of COSMO-RS is the state of a molecule X in its ideally (electrostatically) screened state, i.e., the state of X embedded in a perfect conductor. This state can be calculated with reasonable effort with dielectric continuum solvation methods. Apparently, the conductor-like screening model COSMO is optimally suited for this task, because it is just derived from the limiting case of a molecule in a conductor. Density functional theory (DFT), combined with COSMO, allows for good accuracy of the relevant electrostatics. If efficiently implemented, DFT/COSMO calculations have less than 50% computational overhead compared to the corresponding DFT vacuum calculations.

The only empirical parameters required in this first step are the atomic radii used in the construction of the cavity separating the molecular interior from the conductor outside. Within the framework of COSMO and COSMO-RS we operate with element-specific radii, which are about 17% larger than the corresponding van der Waals (vdW) radii given by Bondi.[10, 16] At the level of dielectric continuum solvation theory,[17, 18] these radii are purely empirical, while the COSMO-RS theory has the basic requirement that the cavities should be space-filling, which just explains an increase of the radii by about 20% relative to vdW-radii.[10]

As soon as a reasonable set of radii is chosen, the state of a molecule in a conductor is well defined at any quantum chemical SCF level (HF or DFT). Note that the artifacts arising from tails of the molecular electron density reaching outside the cavity are quite small within the COSMO theory, in contrast to original dielectric continuum solvation theory. The residual small effects are very accurately corrected by appropriate algorithm.[19] As a result of a DFT/COSMO calculation we do not only yield the total energy of X in its self-consistent state in the conductor, but we also gain the polarization charge density σ, which the conductor places on the cavity to screen the electric field of the molecule. This polarization charge density is a very good local descriptor of the polarity on the molecular surface.

In the next step we consider the molecules of a liquid as if each molecule was swimming in a conductor, i.e., we do a DFT/COSMO calculation for each of the molecules. Then we virtually compress the ensemble to squeeze out the conductor between the molecules. Only a thin film of conductor is left at the interface of the pair-wise contacting molecules, and the polarization charge density on this film is just the sum of the polarization charge densities σ and σ' of the two contacting molecules.

When we remove the conductor between the molecules we replace the artificial interaction of the molecules with the conductor by the real interaction of the molecules. Removing the conductor

piece by piece, the energy change resulting from removal of a certain piece of conductor can be interpreted as a local contact energy of the initially perfectly screened molecular surfaces. As shown in the previous articles, this local contact energy is very well approximated by the expression:

$$E_{\text{cont}}(\sigma,\sigma') = E_{\text{misfit}}(\sigma,\sigma') + E_{\text{hb}}(\sigma,\sigma') = a_{\text{eff}}\left[\frac{\alpha'}{2}(\sigma+\sigma')^2 + c_{\text{hb}}\min(\sigma\sigma' + \sigma_{\text{hb}}^2, 0)\right] \quad (1)$$

where a_{eff} is the area of an effectively independent thermodynamic contact. The first term represents the electrostatic contact interaction energy, which results from the misfit of the two contacting polarization charge densities. Note that the misfit energy is zero, if the two polarization charge densities compensate each other, i.e., if $\sigma = -\sigma'$. The coefficient α' in the misfit energy expression can be derived from basic electrostatics. The second term accounts or the extra energy of hydrogen bonding, if two very polar pieces of molecular surface with opposite signs interact. This empirical formula gives a reasonable description of hydrogen bond energy, if the two parameters c_{hb} and σ_{hb} are appropriately adjusted.

Based on this quantitative expression for the interaction energy of molecules in a condensed state as local contact energies of molecular surfaces, the thermodynamics of the liquid system is evaluated using a model of pair-wise interacting surface segments of size a_{eff}. If $p_{\text{S}}(\sigma)$ and $p_{\text{X}}(\sigma)$ are the surface compositions functions with respect to the polarization charge density σ (the σ-profile) of a solvent S and of a solute X, respectively, then the (pseudo-)chemical potential $\mu_{\text{S}}^{\text{X}}$ of X in S is given by

$$\mu_{\text{S}}^{\text{X}} = \int \mu_{\text{S}}(\sigma)\, p^{\text{X}}(\sigma)\, d\sigma + \mu_{\text{S,comb}}^{\text{X}} \quad (2)$$

where the last term is a simple and small contribution taking into account size effects of solute and solvents. Chemical engineers know this term as the combinatorial contribution.[15] Note, that the pseudochemical potential as introduced by Ben-Naim[20] is just the true chemical potential minus the trivial concentration term $kT \ln x$. We will use the term chemical potential in the sense of pseudochemical potential throughout this article. The thermodynamics of molecular interactions is included in the first part of eq. (2). This is just the integration of a solvent specific function $\mu_{\text{S}}(\sigma)$ (the σ-potential) over the surface of the solute X. The σ-potential expresses the free energy of a piece of surface of polarity σ in an ensemble of composition $p_{\text{S}}(\sigma)$. Within the assumption of pair-wise interacting surfaces it can be exactly calculated from the equation

$$\mu_{\text{S}}(\sigma) = -kT \ln\left[\int \frac{p_{\text{S}}(\sigma')}{A_{\text{S}}} \exp\left\{-\frac{E(\sigma,\sigma') - \mu_{\text{S}}(\sigma')}{kT}\right\} d\sigma'\right] \quad (3)$$

which has to be solved recursively due to the appearance of the unknown σ-potential in the exponent of eq. (3).

The capability of COSMO-RS to predict the chemical potential $\mu_{\text{S}}^{\text{X}}$ of any solute X in any pure or mixed solvent S at variable temperature T enables the calculation of any thermodynamic liquid–liquid equilibrium. Of special importance for the subject of this article are the chemical potentials $\mu_{\text{W}}^{\text{X}}$ of a compound X in pure water (W) and the chemical potential $\mu_{\text{X}}^{\text{X}}$ of the compound in its pure liquid state.

The few adjustable parameters of the COSMO-RS method have been carefully fitted to a large number of thermodynamic data. The accuracy of COSMO-RS is about 1.5 kJ/mol for large chemical potential differences like those typically involved in octanol–water partition coefficients or in water solubility. This corresponds to about 0.27 log-units or slightly less than a factor 2 for equilibrium constants at room temperature.

Solubility

As mentioned above the solubility S_{S}^{X} of a liquid compound X in a solvent S is related to the difference $\Delta_{\text{S}}^{\text{X}} = \mu_{\text{S}}^{\text{X}} - \mu_{\text{X}}^{\text{X}}$ of the chemical potentials of X in S and in pure X. If S_{S}^{X} is sufficiently small so that the solvent behavior of the X-saturated solvent S is not significantly influenced by the solute X, then the decadic logarithm of the solubility is given by

$$\log S_{\text{S}}^{\text{X}} = \log\left(\frac{MW^{\text{X}}\rho_{\text{S}}}{MW_{\text{S}}}\right) - \frac{\ln(10)}{kT}\Delta_{\text{S}}^{\text{X}} \quad (4)$$

In the case of high solubility (S_{S}^{X} greater than 10 mass%) eq. (4) becomes approximate, and the true solubility would have to be derived from a detailed search for a thermodynamic equilibrium of a solvent-rich and a solute-rich phase. But, in general, at least for the purpose of estimating drug solubility, eq. (4) is sufficiently accurate. Because the molecular weights MW and the solvent density ρ are known, eq. (4) is sufficient for the prediction of the solubility of compounds that are liquid at room temperature.

Unfortunately, most drugs are solid at room temperature. Because the solid state of a compound X is related to its liquid state by the free energy difference $\Delta G_{\text{fus}}^{\text{X}}$, which is negative in the case of solids, a more general expression for solubility reads

$$\log S_{\text{S}}^{\text{X}} = \log\left(\frac{MW^{\text{X}}\rho^{\text{S}}}{MW_{\text{S}}}\right) + \frac{\ln(10)}{kT}\left[-\Delta_{\text{S}}^{\text{X}} + \min\left(0, \Delta G_{\text{fus}}^{\text{X}}\right)\right] \quad (5)$$

Because for liquids $\Delta G_{\text{fus}}^{\text{X}}$ is positive, eq. (5) reduces to eq. (4) in this case.

For the precise calculation of $\Delta G_{\text{fus}}^{\text{X}}$ it is necessary to evaluate the free energy of a molecule of compound X in its crystal, i.e., the crystal structure has to be known. In general, crystal structure prediction for drugs has to be considered as an unsolved problem. Thus, there is no viable way to a fundamental model. Hence, we treat $\Delta G_{\text{fus}}^{\text{X}}$ by a QSPR approach in this article. For a physically sound and stable QSPR model we first identified a small set of descriptors of potential significance for $\Delta G_{\text{fus}}^{\text{X}}$. From physical intuition we considered the quantities size, rigidity, polarity, and number of hydrogen bonds as plausible driving forces of crystallization. Molecular size can almost equally well be described by the molecular surface area A^{X} and by the cavity volume V^{X}. Because both are available in the framework of COSMO-RS, we tried both descriptors in our QSPR study. Molecular rigidity of drugs to a large degree is caused by ring structures. Therefore, we used the number of ring atoms $N_{\text{ringatom}}^{\text{X}}$ as a descriptor of rigidity. We also found that $N_{\text{ringatom}}^{\text{X}}$ can be replaced by the number of rotatable bonds (as a measure of flexibility) without change in regression

quality. From our experience with COSMO we consider the dielectric COSMO energy E^X_{diel} as a good descriptor of polarity. The chemical potential μ^X_W of a compound X in water is a combined measure of polarity and hydrogen bonding. Hence, in combination with E^X_{diel} this should be able to reflect hydrogen bonding ability of a compound X.

Unfortunately a direct measurement of ΔG^X_{fus} is impossible, because the supercooled melt of a solid compound is usually inaccessible. Hence, no experimental data for ΔG^X_{fus} are available. To perform a QSPR with respect to ΔG^X_{fus} we choose an indirect way. We took a reasonable experimental data set of aqueous solubilities of 150 common organic and drug compounds from the study published by Duffy and Jorgensen[5] for their QikProp method (called QikProp data set further on). The QikProp method uses partly nonlinear models with six and seven adjusted parameters, respectively, which are based on a set of 11 molecular descriptors, which in turn, are derived from a fast, simplified force-field Monte-Carlo simulation.

Here, DFT/COSMO calculations were performed for all compounds (for details of these calculations see the Computational Details section), followed by a COSMO-RS calculation using the COSMO*therm* program. From these calculations we got values for the free energy differences Δ^X_S of the compounds in aqueous solution and in their liquid state. It should be noted that for the calculation of Δ^X_S a correction of -10.5 kJ/mol was added to the chemical potentials μ^X_W of 15 poly-substitued aliphatic amines in water to account for a known systematic error of COSMO-RS for secondary and tertiary amine groups.[10] Taking the experimental values of $\log S^X_S$ and the calculated values of Δ^X_S we solved eq. (5) for ΔG^X_{fus}. Doing this we allowed for positive values of ΔG^X_{fus} in the first iteration of the QSPR, i.e., we neglected the minimum function in eq. (5). We tested the performance of different combinations of descriptors in multilinear regression. To account for the fact that positive values of ΔG^X_{fus} do not show up in solubility, for each descriptor combination we did a second iteration, in which we omitted those compounds X yielding a positive ΔG^X_{fus}.

Finally, it turned out, that the descriptor combination V^X, $N^X_{ringatom}$, and μ^X_W is best suited for the regression of ΔG^X_{fus}. The polarity descriptor E^X_{diel} did not achieve any significance. Even the regression constant was insignificant, leading to the final regression equation

$$\Delta G^X_{fus} = 12.2V^X - 0.76N^X_{ringatom} + 0.54 * \mu^X_W \qquad (n = 127,\ r^2 = 0.71,\ s = 0.65) \quad (6)$$

where ΔG^X_{fus} and μ^X_W are in kJ/mol and V^X in nm^3. Substitution of this expression into eq. (5) for $\log S^X_W$ and using all 150 compounds yields an correlation coefficient of $r^2 = 0.90$ and a standard deviation of $s = 0.66$ log-units. Although in some cases ΔG^X_{fus} is as large as 20–25 kJ/mol, the overall contribution of ΔG^X_{fus} to the variance of $\log S^X_W$ is only about 25% of the variance arising from Δ^X_W. Hence, ΔG^X_{fus} is much less important than Δ^X_W.

In accordance with physical intuition, the negative coefficient of $N^X_{ringatom}$ indicates that larger number of ring atoms supports fusion. The positive coefficient of μ^X_W means that negative chemical potential in water, i.e., higher polarity, supports fusion, which is physically intuitive as well. Nevertheless, we have problems to understand the positive coefficient of the volume descriptor. The usual assumption would be that increasing size or volume of molecules should support crystallization, while the present result indicates the opposite. This can only be understood as a result of some intercorrelation of the size descriptor and the polarity descriptor, because the latter is size intensive as well. Because the coefficient in front of the volume is highly significant and turns out to be very stable even for the pesticide data set discussed below, we decided to accept this disturbing fact for the present study on drugs and pesticides.

The COSMO-RS-based solubility prediction method consisting of eqs. (5) and (6) will be refered to as COSMO-RSol, further on.

Applications

Computational Details

The molecular geometries of all of the compounds in the training set and the test sets have been sketched as two-dimensional structures and subsequently converted to three-dimensional geometries. To obtain the lowest energy conformations for each compound a molecular dynamics (MD) calculation has been done with molecular modeling program package Alchemy.[21] The MM3 force field has been used to obtain the potential energy during the MD calculation, using an overall MD run time of 5 ps, a time step of 0.001 ps, and a initial temperature of 293 K. From the geometries created by the MD calculation up to five significant lowest energy conformations have been picked for each molecule. Special care has been taken in choosing conformations of molecules that are able to build internal hydrogen bonds, because the polarization charge densities σ computed in the subsequent QC-COSMO calculations (and thus also COSMO-RS' chemical potentials μ^X_S) critically depend upon the correct representation of such hydrogen bonds. The geometry of the chosen conformations has been optimized by the molecular mechanics method MM3 as implemented in Alchemy.[21] Subsequently, the geometries of all conformations have been optimized by the semiempirical AM1/COSMO method using the MOPAC2000 program.[22] Using the geometries thus optimized, the COSMO polarization charge densities σ of the molecular surfaces have been computed on the *ab initio* QC level with the Turbomole program package using the B-P density functional theory with the SVP quality basis set.[23] All of the COSMO-RS calculations have been done using the COSMO*therm* program.[24] For each compound only the one conformation lowest in energy on the Turbomole-BP/SVP/COSMO QC level has been used in the COSMO-RS calculations—with the exception of salicylic acid, succinic acid, 2-acetyloxybenzoic acid, and alanine from the training data set, where two relevant conformations had to be taken into account in COSMO-RS. The relative contribution of each conformer was determined by an iterative procedure using the Boltzmann-weight of the free energies of the conformers in the liquid.[25] This iterative procedure is automated in the COSMO*therm* program.

The Training Set

By adjustment of only three parameters the QikProp data set of 150 aqueous solubilities has been described with an accuracy of

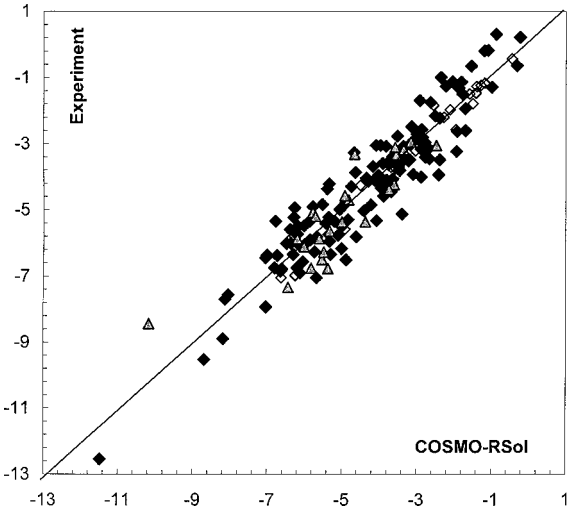

Figure 1. Water solubility log(xH20) calculated COSMO-RSol: diamonds = QikProp training set (filled = solid and open = liquid); triangles = test data set by McFarland.

$s = 0.66$ log-units. The quality of this description may be compared with the standard deviations of $s = 0.88$ and $s = 0.72$ log-units, which was achieved by Jorgensen and Duffy with the QikProp method[5] on the same data.

Calculated and experimental data for the 150 compounds are shown in Figure 1. The corresponding table compiling all experimental and calculated data together with the descriptors is available as supplementary material. The scatter plot clearly shows a rather homogeneous error distribution. The 23 compounds, which had a calculated value of $\Delta G^{X}_{fus} > 0$ are marked by open symbols. Apparently these liquids show a much smaller error. This may be due to better experimental data and due to avoidance of the ambiguities arising from the heuristic treatment of ΔG^{X}_{fus}.

Due to the fact that solubility measurements are far from being trivial, the quality of the experimental data is, in general, quite variable. Experimental errors of 0.5 log-units or more easily arise due to problems with measurement of small concentrations or due to ambiguities arising from different crystal modifications and from solvate crystals. In four cases we could find independently measured data. In two of these cases the values differed by more than half a log-unit. In both cases the calculation error was considerably smaller with respect to these new data. In addition, two experimental data have been included in the original data set, which clearly belong to miscible systems, for which no meaningful value of solubility can be assigned. If we remove these four systems with apparently questionable experimental values from the data set, the rms error reduces to $s = 0.63$. Further indication for the existence of considerable experimental error is a correlation of $r = 0.56$ between the deviations of the COSMO-RS method and the QikProp method. Because both methods are absolutely independent, the most plausible source for such correlation is experimental error. Thus, it appears plausible that a considerable part of the observed error of $s = 0.66$ is due to experimental inaccuracies, and that the intrinsic error of the COSMO-RS is $s = 0.5$ or less.

Test on a High Quality Data Set

McFarland provided us with a data set of 24 aqueous solubilities of drugs that have been measured by the pSol-method.[26–28] The first half of the data has been validated by independent shake flask measurements by the FDA. Hence, the first 12 values can be considered as a high quality data set. The quality of the second half is slightly less sure. On the first half we achieve an rms-deviation of $s = 0.64$ using COSMO-RS solubility method as parameterized before. On the second half we yield $s = 0.86$. It is remarkable, that McFarland et al.[28] achieved values of rms = 0.55 and 0.70 on the two parts of the data set by a QSPR method (HYBOT) trained on these 24 data. Hence, there is clear indication that the first part really is of higher quality than the second. Furthermore, for all six compounds of the McFarland data set, which were in the QikProp data set as well, the COSMO-RS result was closer to the McFarland data, i.e., to the data from the test set. This indicates a higher consistency of this test set data. Summarizing this test, the COSMO-RS solubility method achieved the same accuracy on an a test set of relatively large drug molecules as it had on the training set.

Large Pesticide Test Set

Finally, we tested the COSMO-RS solubility prediction method on a large data set of 548 pesticide molecules with experimental solubilites extracted from the Pesticide Manual.[29] A diverse subset containing 405 compounds was generated allowing for a maximal Tanimotosimilarity computed on Unity fingerprints of 70%.[30a] This diverse subset was used for training of a HQSAR model using the Sybyl suite.[30b] The remainder of the complete data set, i.e., 143 compounds, was used as a test set, see below. The best model was obtained with a holographic length of 763 and a fragmentation rule allowing for fragments containing one to five atoms. The number of latent dimensions (components) of this model is 10, yielding an r^2 of 0.84 and a standard error of 0.88 logarithmic units. The number of latent dimensions was chosen to minimize the standard error as obtained from cross-validation using 10 crossvalidation groups. The q^2 values obtained from this procedure are usually much smaller than those derived from leave-one-out, but the latter are often too forgiving with respect to overfitting and thus are not a good measure of predictivity. We found an q^2 of 0.60 and a crossvalidated standard error of 1.3.

A critical test on predictivity is provided via applying both methods on the 143 compounds of the test set, which were not part of the training set. To remove as much of experimental error as possible, we first eliminated 14 compounds for which the solubilities from the pesticide handbook disagree by more than 0.6 log-units from those reported in the PhysProp database.[31] In a next step 22 compounds have been separated for which both prediction methods, i.e., HQSAR and COSMO-RS deviate from the experimental value by more than one log-unit in the same direction. The removal of these 22 compounds appears to be statistically justified, because only about 1.5 compounds should be such outliers in the same direction based on the rms values reported below assuming normal error distributions. The remainder of 107 pesticides was considered as final test set.

The results of COSMO-RSol and of the HQSAR model on this pesticide data set are shown in Figure 2. COSMO-RSol has an rms

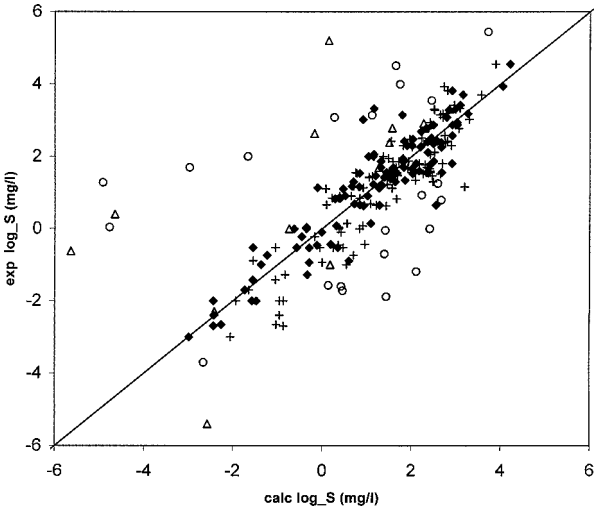

Figure 2. Results for pesticide data set: filled diamonds = COSMO-RSol; crosses = HQSAR; triangles = data excluded due to exp. discrepancies; circles = data excluded due to common deviation of predictions.

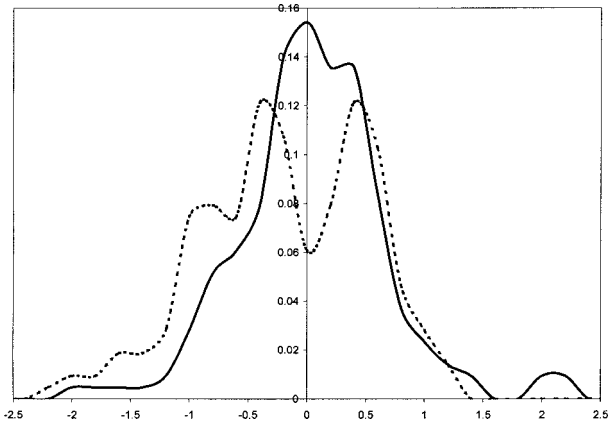

Figure 3. Error distribution for the pesticide data set: solid line = COSMO-RSol; dashed line = HQSAR.

error of 0.61 and a mean unsigned error of 0.45 log-units. The mean signed error was only 0.06, indicating the high degree of transferability of the model. The HQSAR method yields rms = 0.72 and a mean unsigned error of 0.59 log-units, which is in line with the results of the crossvalidation procedure. Thus, although the HQSAR method was specially trained on a large training set of the same source, it is significantly less accurate in prediction on the test set than the purely predictive COSMO-RSol method, which was never trained on pesticides before. It is noteworthy that a refit of the three parameters of COSMO-RSol only yields a reduction of the rms of 0.01 log-units, which is absolutely insignificant. Use of the PhysProp data instead of the data from the Pesticide Manual reduces the rms by 0.01 for both methods, indicating a slightly better quality of the PhysProp data. For the 14 compounds for which the experimental data show major deviations, the PhysProp data show better agreement with both predictions than the Pesticide Manual data. This indicates better quality of the PhysProp data, too. The two largest outliers of the COSMO-RSol method are two very large pesticides of extremely complex chemical structure. Both are quite similar, and the relative solubility of the structures is correctly predicted (see Fig. 2). The error distribution of the two models on the 107 pesticides used as test set is shown in Fig. 3. COSMO-RSol clearly shows an almost Gaussian error distribution, whereas a strange oscillation of period 0.6 log.-units appears to be present in the error distribution of the HQSAR model. Presently, we do not have an explanation for the strange error distribution of the HQSAR model.

Summary and Outlook

The COSMO-RSol method has been introduced as a novel prediction method for aqueous solubility of solid and liquid drug-like compounds. Although not being fully *ab initio*, COSMO-RSol has a rather sound physico-chemical basis compared to all presently available prediction methods for aqueous solubility. This enables COSMO-RSol to achieve even better predictions for aqueous solubility than other state-of-the-art methods on data sets used for the development and parameterization of the other methods. Even on a structurally most demanding and diverse data set of pesticides COSMO-RSol achieved a very satisfying predictive accuracy.

The average accuracy of COSMO-RSol on the data sets considered so far is about 0.65 log-units (rms). There is a strong indication that a significant part of this error is due to experimental error. Thus, it appears to be justified to assume an intrinsic prediction error of 0.5 log-units for the COSMO-RSol method. Considering the average accuracy of 0.3 log-units of COSMO-RS for liquid compounds and the additional approximations involved in the estimation of ΔG_{fus} it is unlikely that the intrinsic error of COSMO-RSol is much less than 0.5 log-units. For a definite assessment of the intrinsic predictive error of the method a broad data set of high quality experimental data would be required.

It should be noted, that in contrast to all other prediction methods for aqueous solubility COSMO-RSol is able to predict solubility in almost arbitrary solvents and solvent mixtures due to the capability of COSMO-RS to predict the chemical potential of a compound X in arbitrary liquids. A validation of COSMO-RSol on non-aqueous solubilities will be given in a forthcoming paper. Another advantage of this new method is that based on the same COSMO calculations used for aqueous solubility many other physico-chemical properties like partition coefficients, vapor pressures, Henry constants, etc. are easily available by COSMO-RS. Even physiological partition behavior can be calculated based on COSMO-RS.[32, 33]

As all other solubility prediction methods COSMO-RSol is restricted to the solubility of pure, neutral, nonionic compounds. For compounds with known pK values correction for dissociation or protonation can be trivially made. Simultaneous prediction of pK values is presently not routinely doable with COSMO-RS, although the first promising results have been reported.[34] Due to the sound physical basis of COSMO-RSol an extension of the method to cocrystals and even to salts appears to be achievable with reasonable effort if a data set of good experimental solubilities can be collected. The first steps in this direction are being planed.

Presently, the greatest disadvantage of the COSMO-RSol method compared with many other solubility prediction methods is its relatively high time demand of approximately 2 CPU hours on a 1-GHz PC processor. At reasonable turn-around times this limits the methods to data sets in the order of 1000 compounds, and prohibits applications in high throughput screening (HTS), where typically several ten-thousands or hundred-thousands of compounds have to be treated within days. A slightly more approximate method (COSMO*frag*), which derives the required σ-profiles of new compounds from a large database of about 10,000 precalculated COSMO-files of drug-like structures by a similarity analysis of fragments is being developed. This method will reduce the calculation times of COSMO-RSol to seconds, hopefully at a small loss of accuracy.

References

1. Yalkowsky, S. H. Solubility and Solubilization in Aqueous Media; Oxford University Press: Oxford, 1999.
2. Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. J Chem Inf Comput Sci 2000, 40, 1.
3. Huuskonen, J.; Salo, M.; Taskinen, J. J Chem Inf Comput Sci 1998, 38, 450.
4. Mitchell, B. E.; Jurs, P. C. J Chem Inf Comput Sci 1998, 38, 489.
5. Duffy, E. M.; Jorgensen, W. L. J Am Chem Soc 2000, 122, 2878.
6. Hansch, C.; Leo, A. J. Substituent Parameters for Correlation Analysis in Chemistry and Biology; Wiley: New York, 1979; and CLOGP-Program, Daylight CIS, Irvine CA.
7. Meylan, W.; Howard, P. Users Guide for LOGKOW; Syracuse Research Corporation, Syracuse, NY, 1994.
8. Verwer, P.; Leusen, F. J. J. In Reviews in Computational Chemistry; Lipkowitz, K. B.; Boyd, D. B., Eds.; Wiley-VCH: New York, 1998; p. 327, vol. 12.
9. Klamt, A. J Phys Chem 1995, 99, 2224.
10. Klamt, A.; Jonas, V.; Buerger, T.; Lohrenz, J. C. W. J Phys Chem 1998, 102, 5074.
11. Klamt, A.; Eckert, F. Fluid Phase Equilibria 2000, 172, 43.
12. Klamt, A. In Encyclopedia of Computational Chemistry; Schleyer, P. v. R.; Allinger, L., Eds.; Wiley: New York, 1998; vol. 2.
13. Klamt, A.; Schüürmann, G. J Chem Soc Perkin Trans 1993, 2, 799.
14. Fredenslund, A.; Gmehling, J.; Rasmussen, P. Vapor Liquid Equilibria Using UNIFAC; Elsevier: Amsterdam, 1977.
15. Abrams, D. S.; Prausnitz, J. M. AIChE J 1975, 21, 116.
16. Bondi, A. J Phys Chem 1964, 68, 441.
17. Tomasi, J.; Persico, M. Chem Rev 1994, 94, 2027.
18. Cramer, C. J.; Truhlar, D. G. In Reviews in Computational Chemistry; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH Publishers: New York, 1995; p. 1, vol. 6.
19. Klamt, A.; Jonas, V. J Chem Phys 1996, 92, 9972.
20. Ben Naim, A. Solvation Thermodynamics; Plenum Press: New York, 1987.
21. Alchemy32, Version2.0.5 (8/04/98), Tripos, Inc., St. Louis, MO (1998).
22. MOPAC2000, Fujitsu Corp., Tokyo (2000).
23. Schäfer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. Phys Chem Chem Phys 2000, 2, 2187.
24. Klamt, A.; Eckert, F. COSMO*therm*, Version C1.1-Revision 01.01; COSMO*logic*, GmbH&CoKG, Leverkusen, Germany, 2001; see also URL: http://www.cosmologic.de.
25. Eckert, F.; Klamt, A., in preparation.
26. McFarland, J. W., personal communication.
27. Avdeef, A.; Berger, C. M.; Brownell, C. Pharmaceut Res 2000, 17, 85.
28. McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. Poster presented at QSAR2000, Düsseldorf, Germany, 2000.
29. Tomlin, C. D. S., Ed. The Pesticide Manual; British Crop Protection Council: Farnham, Surrey, UK, 1997; 11th ed.
30. (a) James, C. A.; Weininger, D. Daylight Theory Manual; Daylight Chemical Information Systems, Inc.: 1995; UNITY Reference Manual; Tripos Inc.: St. Louis, MO, 1995; (b) Sybyl 6.7.1; Tripos Inc.: St. Louis, MO, 2001.
31. Howard, P.; Meylan, W. PHYSPROP DATABASE; Syracuse Research Corp.: Syracuse, NY, 2000.
32. Klamt, A.; Eckert, F. In Rational Approaches to Drug Design; Prous Science: Barcelona, 2001.
33. Klamt, A.; Eckert, F.; Hornig, M. J Comp-Aid Mol Design 2001, 15, 355.
34. Beck, M. E.; Bürger, Th., to be published.