Journal of Computer-Aided Molecular Design, **15:** 355–365, 2001.
KLUWER/ESCOM

COSMO-RS: A novel view to physiological solvation and partition questions

Andreas Klamt*, Frank Eckert & Martin Hornig
COSMOlogic GmbH&CoKG, Burscheider Str. 515, 51381 Leverkusen, Germany

Received 8 August 2000; accepted 22 October 2000

Key words: continuum solvation, COSMO, COSMO-RS, partition coefficients, QSAR, solubility

Summary

Both, dielectric continuum solvation models as well as surface or group based methods using polarity and lipophilicity parameters have been proven to be useful tools for the analysis of solvation and partition questions. For the first time, COSMO-RS provides an integrated theory, which combines the aspects of continuum solvation and surface interactions, and which ends up with chemical potentials of molecules in almost arbitrary solvents and mixtures. Due to its sound theoretical basis, COSMO-RS does not only provide a new quantitative access to solvation and partition properties in well defined solvents, but it also opens a novel view and gives a better understanding of the general problem of solvation. Finally, this allows for a generalisation of COSMO-RS to sophisticated physiological partition problems involving as complex phases as blood, brain, or cell membranes. The use of COSMO-RS for drug discovery and design is demonstrated by applications to blood-brain partition coefficients, and water solubility.

Computational approaches to partition and solvation problems

The solvation and partition behaviour of chemical compounds is of overwhelming importance for technical chemistry and especially for biochemistry. Because solvation and partition usually are considered in thermodynamic equilibrium, the necessary quantity required for their calculation is the chemical potential μ_S^X of a compound X in a solvent S, at a given temperature T. For physiological properties T is usually assumed room temperature. Using the pseudo-chemical potential μ^{*X}_S according to Ben-Naim [1]

$$\mu^{*X}_S = \mu^X_S - RT \ln x^X_S, \tag{1}$$

where x^X_S is the molar concentration of compound X in solvent S, the equilibrium condition of equal chemical potentials of X in two phases S and S' reads

$$\mu^{*X}_S - RT \ln x^X_S = \mu^{*X}_{S'} - RT \ln x^X_{S'}. \tag{2}$$

*To whom correspondence should be addressed.

Converting molar concentrations x_S to volume concentrations c_S, which usually are used in partition and solubility studies, by

$$c^X_S = \frac{x^X_S}{v_S}, \tag{3}$$

where v_s is the molar volume of the solvent S, Equation 2 easily leads to the expression

$$\begin{aligned} \log K^X_{S,S'} &= 0.4343 \ln K^X_{S,S'} = 0.4343 \ln \frac{c^X_S}{c^X_{S'}} \\ &= -\frac{0.4343}{RT}\left[\mu^{*X}_S - \mu^{*X}_{S'} + RT \ln \frac{v_S}{v_{S'}}\right] \end{aligned} \tag{4}$$

for the logarithmic partition coefficient of compound X in solvents S and S'. The constant 0.4343 is just 1/ln(10). In the same way the solubility of compound X in solvent S

$$\log S^X_S = -\frac{0.4343}{RT}\left[\mu^{*X}_S - \mu^{*X}_X + RT \ln v_S\right] \tag{5}$$

can be defined. In these equations the μ^{*X}_{S} are the pseudo-chemical potentials of solute X in pure S. Hence, μ^{*X}_{X} denotes the pseudo-chemical potential of X in pure X. Because the molarity of X in pure X is $x = 1$, we just have $\mu^{*X}_{X} = \mu^{X}_{X}$. Note that Equation 5 is valid only in the limit of small solubilities ($S \ll 0.01$), but this is no severe limitation since in drug design solubilities typically are far below that limit.

Thus we see, that for the calculation of partition coefficients or solubilities we need to be able to calculate the pseudo-chemical potential of a compound X in a pure solvent S, or at least the difference between the pseudo-chemical potentials of X in two different solvents. Please note that the pseudo-chemical potential is closely related to the free energy of compound X in solvent S and that many workers in that area prefer to use the expression 'free energy'.

Unfortunately, the calculation of a pseudo-chemical potential is a very sophisticated task, because we do not only have to calculate the interaction energy of a solute X in a solvent S, but we have also to take into account the change in the entropy and in the interactions of the solvent molecules caused by the presence of the solute molecule X.

Molecular dynamics (MD) or Monte Carlo (MC) methods are the most straight-forward procedures to compute the change in free energy of an ensemble of solvent molecules S by insertion of a solute X. But in order to get reasonably accurate numbers one has to consider a very large ensemble of solvent molecules and to apply periodic boundary conditions in order to avoid surface effects. Nowadays such calculations can be done routinely based on force-field pair-potentials [2, 3], but one should be aware that such classical, and typically non-polarisible, force-fields provide a reasonable, but still rather approximate description of the real interactions, which are of quantum-chemical nature. Still such calculations are quite time-consuming. Recently, Jorgensen and Duffy introduced two shortcuts of the MD/MC approach (BOSS and QikProp) [4, 5], which are based on averaged interaction descriptors derived from rapid simulations in reference solvents and combined with a QSAR analysis with respect to the logarithmic partition property.

Conversely, the computationally fastest, but chemically least detailed approach to the estimation of partition coefficients is the fragment- or group-based increment approach, also known as linear free energy relationship (LFER) approach. The most prominent out of this class of methods is the CLOGP method of Leo and Hansch [6, 7] for the calculation of octanol-water partition coefficients. For a recent review of these methods see [8]. The basic assumption of these methods is that the change in free energy (or in pseudo-chemical potential) of a solute X between solvents S and S' can be split into independent contributions of the chemical groups of X. In this case, according to Equation 4 the logarithmic partition coefficient becomes a sum of group contributions, which can be fitted by linear regression from a sufficiently large set of experimental data. Despite of the fact that the LFER approach appears to be extremely successful for octanol-water, its big disadvantage is that a data set of several thousand partition coefficients is required in order to fit the large number of group-parameters. For the octanol-water system about 20 000 experimental data values are available, but even for the next best measured system the number of available data is approximately 1000. Thus in general there is not sufficient information available to satisfactorily fit a LFER for partition coefficients involving less common solvents.

In order to overcome this drawback, Abraham has introduced a set of rather general partitioning descriptors, which themselves are composed of group contributions [9]. Thus he only has to fit these group contributions to appropriately chosen experimental data. Then most diverse partitioning problems can be accessed with small effort. Nevertheless, even this retains the basic disadvantage of being unable to resolve molecular details like isomeric or conformational effects, and it is hard to apply for sets of compounds with most diverse functional groups and heterocycles due to the problem of missing fragment contributions.

Chemical engineers are used to even more rigorous group contribution models (GCMs), in which solute and solvent are represented by groups. UNIFAC [10] is the most popular GCM. In such models the chemical potentials of the compounds are derived from an approximate statistical thermodynamics of pair-wise interacting surface pieces. For each pair of functional groups an interaction parameter has to be fitted to experimental thermodynamic data in a large non-linear fitting procedure. The advantage of this approach compared with the LFER approach is its ability to treat any solvent or solvent mixture as well as complete binary phase diagrams, provided the interaction parameters for all groups involved in the system are known. A disadvantage is the fact that in GCMs the number of required parameters scales squared with the number of different groups, while in an LFER it scales

linearly. Therefore, UNIFAC can afford a much less detailed definition of groups (about 100 times less) than CLOGP, and thus it is not reliable for typical drug compounds.

A very different approach is the explicit representation of the solute combined with a continuum representation of the solvent. Most of these continuum solvation models (CSM) [for a review 11, 12] concentrate on the electrostatic behaviour of the solvent. Either by solution of the dielectric boundary conditions or by solution of the Poisson-Boltzmann equations, both of which represent the same physics in non-ionic solvents, the solute is treated as if it is embedded in a dielectric medium. Usually the macroscopic dielectric constant of the solvent is used. The conductor-like screening model (COSMO) [13] is just one model out of this class, which by a slight approximation achieves superior efficiency and robustness compared with others. The advantage of such CSMs is that the solute can be treated with high rigor, typically at a quantum chemical level. If supplemented with a bunch of surface specific descriptors characterising dispersive interactions, cavitation energies, and other non-electrostatic contributions, the results of such CSMs appear to be capable of describing quite diverse liquid partition properties and coefficients. The SM_x models of Cramer et al. [14], which are based on semi-empirical quantum chemical calculations, are most widely parameterised for the calculation of partition coefficients. Due to the relative success of CSMs, most workers in that field do not realise that the continuum description of the solvent is based on a wrong physical concept. The macroscopic dielectric continuum theory is a linear response theory, which describes the first-order response of a medium to the application of weak electric fields. But the electric fields in the surrounding of an even modestly polar molecule are so strong that the permanent dipole moments of the polar solvents get almost perfectly ordered. This leads to saturation effects which are absolutely incompatible with the assumptions of a linear response theory. Thus, the application of the dielectric theory to the situation of a molecule in a polar solvent is physically absolutely untenable. The relative success of CSMs must be caused by some other reasons [15].

To summarise, there are very different approaches to the calculation of free energies or pseudo-chemical potentials of molecules in solution, each of them covering different aspects of the problem in more detail. Despite of the usually assumed picture of pairwise and distance-dependent interactions of atoms, the relative success of MD/MC derived interaction parameters, of group interaction models, and of surface parameter supplemented CSMs suggests that many aspects of free energies of molecules in solution can be as well or even better described by a model of surface interactions without explicit knowledge of the atom positions of the solvent, i.e. by some kind of solvent surface continuum model.

COSMO-RS, a theory of interacting surfaces

The conductor-like screening model for real solvents (COSMO-RS) [15–18] is a novel theory which integrates the concepts of quantum theory, dielectric continuum models, and surface interactions. COSMO-RS considers the a liquid system as an ensemble N molecules, including solvent and solute molecules. The molecules can be classified into n compounds X_i with molarity x_i. Then the COSMO-RS concept is at best described by the following series of steps:

(1) For each compound X_i, i.e., for solute and solvent molecules, do a density functional (DFT) COSMO calculation in order to get the total energy E^i_{COSMO} and the polarisation (or screening) charge density (SCD) σ on the molecular surface. For the cavity construction in COSMO use a set of radii which ensure that the cavity volumes are approximately equal to the molar volumes of the compounds.

(2) Consider each individual molecule of the system as swimming around in a virtual conductor. The energies and SCDs of the molecules are known from the COSMO calculations and hence the total energy is known.

(3) Now compress the system to the true density of the liquid. This means, that the molecules are pressed together and get slightly deformed, but with no overall compression of the cavities, because the cavities just had the right molar volumes (see Figure 1).

(4) If we assume, that there resists a thin film of conductor between the molecules, we may assume that the total energy of the system is unchanged because each molecule still is embedded in a conductor and the shape of the cavities does not change significantly. Note, that each piece of molecular surface now is in close contact with another one. Each piece of contact surface has a net SCD $\sigma + \sigma'$ arising from the two touching molecules.

(5) In order to get back to the real system, in which no conducting surface is present between the molecules, we have to remove the conductor. If we do that

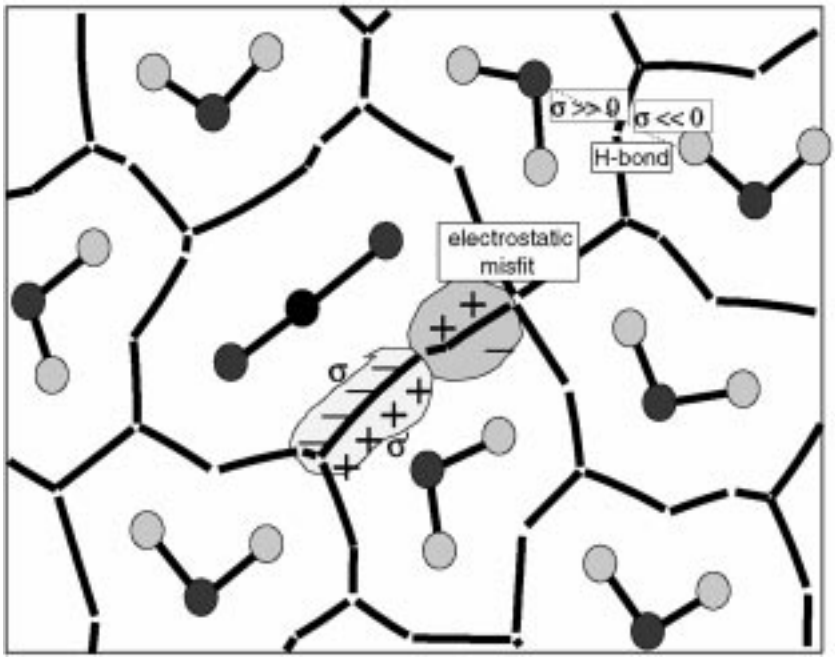

Figure 1. Schematic picture of COSMO-RS interactions.

piece by piece, we can define the energy change of the system caused by the removal of a piece of surface as the local contact interaction of the two contacting molecular surfaces.

(6) Now the electrostatic contribution to this interaction energy is given by

$$e_{\text{misfit}}\left(\boldsymbol{\sigma}, \boldsymbol{\sigma}'\right) = \frac{\boldsymbol{\alpha}'}{2}\left(\boldsymbol{\sigma} + \boldsymbol{\sigma}'\right)^2 \tag{6}$$

This is a specific interaction energy per unit area. The name 'misfit' results from the fact that at perfect fit of the two SCDs, i.e., for $\sigma = -\sigma'$, the electrostatic contact energy is just zero. Note, that at least to first order, this functional does include polarisability effects.

(7) Hydrogen bonding energy can also be described by the two adjacent SCDs. Because donors, and only donors, have a very strongly negative SCD, and because only acceptors have a very positive SCD, a functional of the form

$$e_{\text{hb}}\left(\boldsymbol{\sigma}, \boldsymbol{\sigma}'\right) = c_{\text{hb}} \min\{0, \boldsymbol{\sigma}\boldsymbol{\sigma}' + \boldsymbol{\sigma}_{\text{hb}}^2\} \tag{7}$$

yields a reasonable description of hydrogen bonding energies, if the two parameters c_{hb} and σ_{hb} are appropriately adjusted.

(8) Now the total energy of a concrete realisation of the ensemble can be calculated by taking the COSMO energies of molecules and adding all local contact interactions by integration over the contact surface.

(9) In order to avoid the cumbersome integration and the time-consuming thermodynamic averaging over a large number of realisations, we now make use of the fact, that in the COSMO-RS picture all interactions are local pair-wise interactions of molecular surfaces. This justifies the approximation to replace the ensemble of interacting molecules by the respective ensemble of interacting surface pieces.

(10) In order to describe the composition of the surface ensemble with respect to the interaction parameter σ we only need to know the distribution functions $p^X(\sigma)$ for all compounds *X*. These histograms will be called 'σ-profiles', further on. A few representative σ-profiles are given in Figure 2. Note, that the σ-profiles provide a detailed and vivid description of the polarity of the molecular surface (see next section).

(11) Now the statistical thermodynamics is extremely simplified. If the assume that all molecular surfaces are in close pair-wise contact, and that the interaction energy is given by the functional e(σ,σ′), the chemical potential of a piece of surface with SCD σ in an ensemble described by a distribution $p_S(\sigma)$ is exactly given by the formula

$$\mu_S(\sigma) = -\frac{RT}{a_{\text{eff}}} \ln\left\{\int d\sigma' p_S\left(\sigma'\right) \exp\left(\frac{a_{\text{eff}}}{RT}\left(\mu_S\left(\sigma'\right) - e\left(\sigma, \sigma'\right)\right)\right)\right\} \tag{8}$$

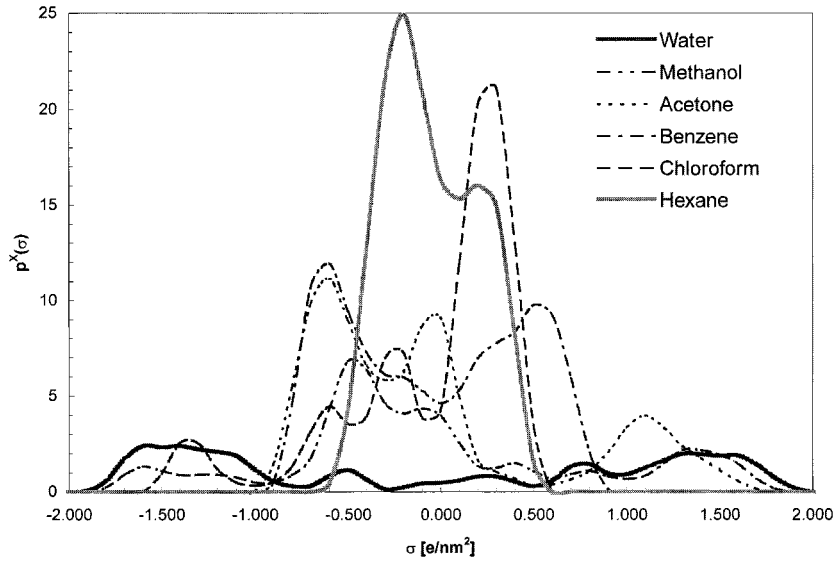

Figure 2. σ-profiles of representative solvents.

This implicit equation, in which a_{eff} denotes an effective, statistically independent piece of contact area, can be solved by iteration within milliseconds on a PC. The function $\mu_S(\sigma)$ tells us how much the solvent S does like surface of polarity σ. This is a characteristic function for each solvent. We call it the σ-potential of solvent *S.* A few examples are given in Figure 3.

(12) The final step is the calculation of the chemical potentials of the compounds. For each component X the pseudo-chemical potential is given by integration of the σ-potential $\mu_S(\sigma)$ over the surface of X. Making use of the distribution function (σ-profile) of X the surface integral reduces to

$$\mu^{*X}_S = \int p^X(\sigma)\,\mu_S(\sigma)\,d\sigma + \mu^X_{\text{comb S}}. \qquad (9)$$

The combinatorial contribution $\mu^X_{\text{comb,S}}$ to the PCB μ^{*X}_S of compound X in solvent S takes into account size and shape effects of solute and solvent [18]. Usually it is small compared to the first term in Equation 9 which results from the surface interactions. Here it is sufficient to consider it as a solvent specific constant.

As a result of this series of relatively simple steps, starting from a quantum chemical calculation for each compound we found an expression for the pseudo-chemical potential of an almost arbitrary chemical compound X in an almost arbitrary solvent S, which may be a pure compound or a mixture. According to Equations 4 and 5, the ability to calculate the pseudo-chemical potentials allows us to calculate any partition coefficient as well as solubility. Based on density functional COSMO calculations, the few parameters required in COSMO-RS, have been fitted to a large set of experimental data [17], covering 215 diverse chemical compounds and the properties ΔG_{hydr}, $\log P_{\text{vapor}}$, and the aqueous partition coefficients with octanol, hexane, benzene, and ether (see Figure 4). Note, that the properties ΔG_{hydr} and $\log P_{\text{vapor}}$ involve the gas-phase, which requires a small addendum to the steps given above that is not of interest here. However, since $\log S_{\text{aq}}$ is the difference of $\Delta G_{\text{hydr}}/RT$ and $\ln P_{\text{vapor}}$, aqueous solubility was implicitly taken into account in the parameterisation of COSMO-RS. The initial COSMO-RS parameterisation yielded a rms-error of 0.3 log-units for the diverse partition and solubility properties of small and medium sized molecules. In recent parameterisations the error has been reduced to about 0.23 log-units.

COSMO-RS is the least parameterised of all quantitative methods for the calculation and liquid partition and solubility. For the description of essentially the entire organic chemistry we need about 10 element specific radii as well as 5 (up to 10 in recent parameterisations) intrinsic parameters. The COSMO radii come out to be about 1.17(±0.02) times Bondi van der Waals radii [19] and they are in good agreement with the volume argument. Hence, essentially they are not really free parameters. In addition, at least two of the other parameters, i.e. the α' parameter of the misfit interaction as well as the effective contact area a_{eff}, come out to be very close to a previous physical

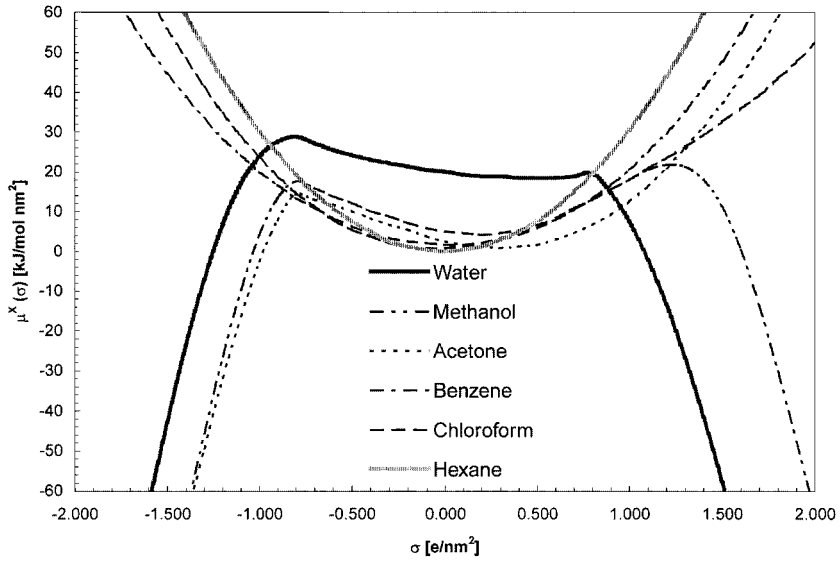

Figure 3. σ-potentials of representative solvents.

estimate. In the latter case we find $a_{\text{eff}} \cong 7\text{Å}^2$ which corresponds to about 6.5 independent contact of a water molecule in water. This number is almost identical with the number of nearest neighbours of water typically found in MD simulations. Thus, we have 10–20 adjustable parameters in COSMO-RS, depending on the individual taste of counting free parameters. This should be compared with the large number (several hundreds) of parameters hidden in the force-fields and in the remainder procedures of MC/MD calculations, and with the approximately 5000–10 000 parameters hidden in a good UNIFAC parameterisation. CLOGP, which in contrast to COSMO-RS, MC/MD, and UNIFAC is restrained to log K_{ow} at room temperature, i.e. to a single partition property, has even more than 3000 adjusted parameters.

Summarising, we conclude that due to its unusual, but sound physical basis, the COSMO-RS way to treat intermolecular interactions in solution as contact interactions of previously ideally screened molecules turns out to be capable of describing the general liquid-liquid equilibrium thermodynamics of small and medium sized, neutral organic compounds with a comparable accuracy, requiring much less parameters than other approaches. The disadvantage is that the computational costs of the DFT/COSMO calculation still are high compared to group contribution methods. But depending on the details, they typically are much less than the overall costs of good MC/MD calculations.

σ-profiles and σ-potentials

In Figures 2 and 3 σ-profiles and σ-potentials of a few representative organic compounds are given. In order to provide a rough feeling of the information content of these entities, a short discussion is given here.

Hexane is taken as a representative of alkanes. According to the non-polar character of alkanes, the σ-profile of hexane is a rather narrow distribution centred at $\sigma = 0$, but nevertheless two peaks arising from the exposed carbon and hydrogen surfaces, respectively, are clearly resolvable. They result from the slightly negative polarity of the carbon atoms, which results in a peak at slightly positive σ (!), and the weakly positive polarity of alkane hydrogens leading to a peak at $\sigma \cong -0.25$ e/nm^2. (Units e/nm^2 for σ will be used further.) The σ-potential of alkane, which characterises the solvent behaviour of hexane, is a simple parabola. Its reflects the dielectric constant of $\varepsilon = 2.1$, which results from electronic polarisability.

The σ-profile of benzene clearly exhibits two well separated peaks at $\sigma = \pm 0.5$, corresponding to the π-face and to the hydrogen ring of benzene, respectively. The σ-potential of benzene is very close to a parabola, again. But due to the much broader σ-profile, the curvature of this parabola is less than that of alkanes, reflecting the fact, that benzene is less repellent to solute polarity than alkane due to its considerable electrostatic quadrupole moment. Note, that a simple dielectric model would not resolve the difference be-

tween alkane and benzene, because both have almost identical macroscopic dielectric constants.

Water has a very broad and symmetric σ-profile with two pronounced peaks at about $\sigma = \pm 1.5$, resulting from the polar lone-pairs of oxygen and from the two polar hydrogens, respectively. Due to its very broad, but symmetric σ-profile, the σ-potential of water would be a rather constant line in a wide σ-range, if only electrostatic interactions were present. Such a constant line corresponds to a parabola of zero curvature and hence to a pseudo-dielectric behaviour of a strong dielectric, as it is usually assumed for water. Note, that COSMO-RS reproduces this common finding for the electrostatic behaviour of water without using the erroneous dielectric theory. The presence of hydrogen bonding causes the strong decay of the σ-potential of water for strongly positive and negative values of σ. These parts of the σ-potentials represent the affinity of the solvent for hydrogen bond donors and acceptors, respectively. An important feature in the σ-potential of water is its relatively high value at $\sigma = 0$. While all other solvents in this region are close to zero, water has a value of about 20 k/mol nm^2. In this way COSMO-RS quantitatively graspes hydrophobicity, i.e., the fact that non-polar molecular surface regions try to move into any other solvent but water. Hydrophobicity is just a phenomenon arising from the extremely strong and balanced electrostatic and hydrogen bond interactions in water.

Methanol has only one donor hydrogen, resulting in a donor peak at $\sigma = -1.5$ which is about half as high as in water. The oxygen peak is very similar to water. In addition, we have an alkane-like structure from the methyl group which is shifted to the left because the methyl group is polarised by the hydroxy group. In the σ-potential we see that due to the imbalance of donor and acceptor capacity methanol has a higher affinity for donors than water, but it act much less attractive to additional acceptors.

Acetone shoes a very asymmetric σ-profile. The carbonyl oxygen results in a peak quite different from that of the sp^3-oxygen peaks of water and methanol. There are no donors and the negative polarity of the oxygen is compensated by a large polarised alkane structure resulting from the two methyl groups. The electrostatic asymmetry expressed in the σ-profile and the resulting strong electrostatic misfit is the reason for the relatively high vapour pressure of acetone. In the σ-potential this asymmetry becomes most apparent by the pronounced donor-affinity and the strongly repulsive interaction with additional acceptors.

Finally, the σ-profile of chloroform shows a very strong peak at about $\sigma = 0.3$, resulting from the large amount of slightly polar chlorine surface. The single hydrogen atom has a small exposed surface leading to a peak of small area but very high SCD at about $\sigma = 1.4$. The inverse asymmetry, i.e., the complementarity, in the σ-profiles of acetone and chloroform is the reason for their strongly negative heat of mixing [18].

Altogether we may conclude, that σ-profiles give an interesting and detailed quantitative description of the polarity and hydrogen bonding features of solutes. On the other hand, the σ-potentials provide a quantitative and integral description of the solvent behaviour regarding electrostatics, hydrogen bonding, and hydrophilicity.

Consideration of a large number of different solvents led us to the finding that σ-potentials can be described very well by a Taylor-like expansion of the form

$$\boldsymbol{\mu}_S(\boldsymbol{\sigma}) \cong \sum_{i=-2}^{m} c_S^i f_i(\boldsymbol{\sigma}), \tag{10}$$

with

$$f_i(\boldsymbol{\sigma}) = \boldsymbol{\sigma}^i \quad \text{for} \quad i \geq 0, \tag{11}$$

and

$$\begin{aligned} f_{-2/-1}(\boldsymbol{\sigma}) &= f_{\text{acc/don}}(\boldsymbol{\sigma}) \\ &\cong \begin{cases} 0 & \text{if } \pm\boldsymbol{\sigma} < \boldsymbol{\sigma}_{\text{hb}}, \\ \mp\boldsymbol{\sigma} + \boldsymbol{\sigma}_{\text{hb}} & \text{if } \pm\boldsymbol{\sigma} > \boldsymbol{\sigma}_{\text{hb}}. \end{cases} \end{aligned} \tag{12}$$

The highest order required for a sufficient description typically is $m = 3$. In this sense, we may characterise each solvent (at fixed temperature, usually room temperature) by the set of σ-coefficients c_S^i. Obviously any difference between the σ-potentials of two solvents is of the same kind of expansion, with coefficients $c_{S,S'}^i$ being just the difference of the coefficients of the two solvents. Combining Equations 4 and 9 and using Equation 10 for $\mu_s(\sigma)$, we thus find that any partition coefficient between two solvents S and S' should be expressible in the form

$$\begin{aligned} \log K_{S,S'}^X &= -\frac{0.4343}{RT}\left[c_{S,S'} + \int p^X(\boldsymbol{\sigma})\left(\mu_{S'}(\boldsymbol{\sigma}) - \mu_S(\boldsymbol{\sigma})\right) \mathrm{d}\boldsymbol{\sigma}\right] \\ &\cong \tilde{c}_{S,S'} + \int p^X(\boldsymbol{\sigma}) \sum_{i=-2}^{m} \tilde{c}_{S,S'}^i f_i(\boldsymbol{\sigma})\, \mathrm{d}\boldsymbol{\sigma} \\ &= \tilde{c}_{S,S'} + \sum_{i=-2}^{m} \tilde{c}_{S,S'}^i M_i^X \end{aligned} \tag{13}$$

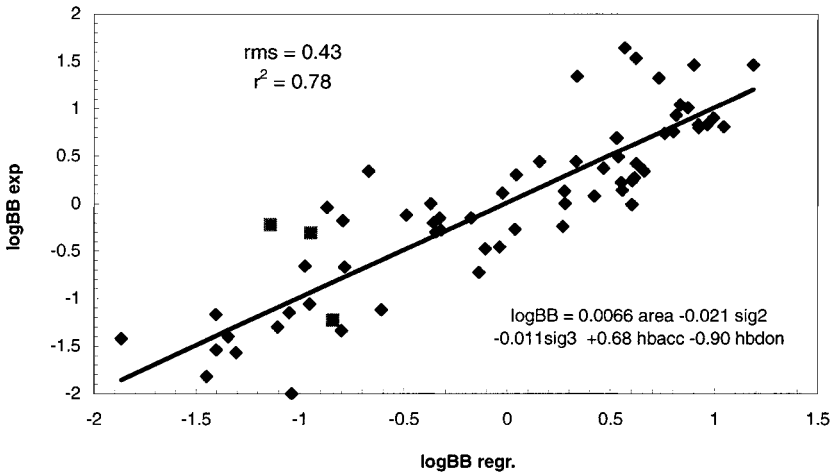

Figure 4. σ-moment regression for the blood-brain partition coefficient.

where the density terms as well as the combinatorial contributions have been subsumed in $\tilde{c}_{S,S'}$ and the σ-moments M_i^X of the solute X are defined by

$$M_i^X = \int p^X(\boldsymbol{\sigma})\, f_i(\boldsymbol{\sigma})\, \mathrm{d}\boldsymbol{\sigma}. \tag{14}$$

Equation 13 implies that any logarithmic partition coefficient can be represented as a linear combination of σ-moments. As a consequence, the set of σ-moments M_i^X, $i = 0, 2, 3$, complemented by the hydrogen bond moments M_{acc}^X $(= M_{-2}^X)$ and M_{don}^X $(= M_{-1}^X)$ should be a very good and almost complete set of molecular descriptors for a linear regression analysis of any partition problem. Note, that the first moment M_1^X usually is of no importance, because it is just the negative of the total charge of the molecule. Hence, for neutral compounds M_1^X trivially vanishes. By definition of the σ-profiles the zeroth moment M_0^X is identical with the molecular surface. The second moment is an excellent measure of the overall polarity of the solute, and the third moment is a measure of the asymmetry of the sigma profile. The hydrogen bond moments are quantitative measure of the acceptor and donor capacities of the compound X, respectively.

If the S and S' are simple solvents or mixtures of known chemical composition, we need not derive the partition coefficients from the linear regression approach of Equation 13, but can directly get the pseudo-chemical potentials in the two phases from the COSMO-RS thermodynamics. But in cases, in which one or both of the phases are of unknown composition or just too complex for a direct chemical specification, Equation 13 allows us to handle the problem by simple linear regression analysis, based on a small number of experimental data. Such cases are adsorption coefficients to absorbers like activated carbon [20], to polymers, or to complex physiological properties like membrane partition coefficients [21], blood-brain partition coefficients [22], or even the oral drug permeability, although the latter probably is a combined partition and transport phenomenon. For all of these cases the σ-moment regression approach has been successfully used. As an example, the application to blood-brain partition coefficients is shown in Figure 4.

Solubility

Aqueous solubility of drugs is of extreme importance for drug design. From Equation 5 we see that the main difference between the calculation of partition coefficients and solubilities results from the pseudo-chemical potential μ_X^X, i.e. the pseudo-chemical potential of a compound X in its pure state, while for partition coefficients only pseudo-chemical potentials of X in a two fixed phases S and S' are required. μ_X^X is a much more complex quantity than any μ_S^X in a fixed phase S, because here X operates as solute and matrix. While in the case of a fixed solute we can reasonably address a certain free energy increment to a functional group, there is no way find group increments for μ_X^X and hence for $\log S_{\mathrm{aq}}$, because the addition of a certain functional group can increase or decrease the solubility of a compound, depending on the remainder of the compound X. Thus $\log S_{\mathrm{aq}}$ is a strongly non-linear property, which can hardly be expressed by a linear

regression analysis, unless one of the descriptors does include most of the non-linear behaviour. Thus, there is no way to calculate solubility from a LFER method.

Unfortunately almost all drug molecules are solid at room temperature. Therefore the calculation of μ_X^X means the calculation of the free energy of the compound X in its own pure crystal. For this task the knowledge of the crystal structure is required. In general, the crystal structure will not be known in drug design, and predictive methods for crystal structure are extremely expensive and still far from being routinely applicable. Thus the direct way of calculating μ_X^X is not viable. On the other hand, the calculation of μ_X^X for the liquid state of X is possible with some computational methods. If we know μ_X^X in the liquid state, the missing piece is the free energy difference of X between the crystal and the liquid state, i.e., ΔG_{fus}^X. Typically ΔG_{fus}^X is small compared to μ_X^X. Hence it is reasonable to use μ_X^X of the liquid as a fundamental input for the calculation of $\log S_{\text{aq}}$ and to search for some plausible empirical approximation for ΔG_{fus}^X.

In principle, MD/MC methods are able to calculate μ_X^X for liquid compounds. But this requires the full set-up and equilibration of a sufficiently large box of X molecules. If not prohibitive, this is at least much more expensive than a free energy perturbation calculation applied to a pre-equilibrated box of a solvent S, which is required for partition coefficient calculation.

Group interaction methods such as UNIFAC can calculate the μ_X^X for liquid compounds very easily, because they routinely calculate the pure compound X as reference for activity coefficients. But these methods are of very limited applicability to typical drug compounds, because they are not capable of treating heterocycles or multifunctional aromatic rings.

Since COSMO-RS is treating solute and solvent on the same footing, it is very well able to calculate the chemical potential μ_X^X of a compound X in its pure liquid state. In a study on the aqueous solubility of 127 organic molecules, most of which were drugs, we found that a simple correlation of $\log S_{\text{aq}}^X$ vs. $\Delta\mu_{\text{aq}}^X = \mu_X^X - \mu_{\text{aq}}^X$ yields a correlation coefficient of $r^2 = 0.65$ and a r.m.s.-deviation of 1.2 log-units. The slope in this regression is close to the theoretical expectation. This clearly shows the great significance of the pseudo-chemical potentials directly calculated by COSMO-RS. In a second step we subtracted the theoretical values of $0.4343\Delta\mu_{\text{aq}}^X/RT$ from the experimental values $\log S_{\text{aq}}^X$ in order to get a reasonable data for ΔG_{fus}^X. Now we tried to find a simple linear expression for ΔG_{fus}^X based on available molecular descriptors. Thus we derived the equation

$$\begin{aligned}\log S_S^X &= \frac{0.434}{RT}\Delta\boldsymbol{\mu}_S^X + 0.0022V^X - 0.0032A^X \\ &\quad + 0.37\boldsymbol{\mu}_{\text{water}}^X - 0.116N_{\text{ringatom}}^X \\ &\quad - 0.79 \qquad (15)\end{aligned}$$

which is applicable to room-temperature solubilities of a wide range of solid organic solutes in arbitrary solvents S. Note that the descriptor μ_{water}^X is used as a polarity descriptor as a part of ΔG_{fus}^X. Hence this also appears in the solubility of solvents S. V^X is the COSMO volume, A^X the COSMO surface area. The number of ring atoms acts as a descriptor of molecular rigidity. On our aqueous solubility data set this equation shows a correlation of $r^2 = 0.87$ and has an rms error of 0.71 log-units. Considering the substantial experimental error of solubilities, we assume that the accuracy of Equation 15 is approximately 0.5 log-units. A detailed publication is in preparation [23].

Computational aspects and shortcuts

Within the Turbomole program [24, 25] used for the drug applications in this paper, the high quality COSMO parameterisation uses full geometry optimisation at the TZVP basis set, but for the relatively large drug compounds this would be too time-consuming. Here we use a slightly less accurate level (about 10% increase in rms) which combines single-point DFT/COSMO on SVP basis with geometries optimised on semi-empirical AM1/COSMO [13] level with MOPAC2000 [26]. For a typical drug molecule such calculation takes about 2 hours on a single 600 MHz-IntelPIII-CPU which is about a factor 20 faster than the optimal level. On a cheap cluster of 10 such CPUs a 120 compound study thus can be performed in just a day. The subsequent COSMO-RS calculations are done with the COSMO*therm* program [27] and took only milliseconds per compound. Note, that only one DFT/COSMO calculation is required for each compound, which enables the calculation of all the COSMO-RS properties based on σ-profiles, e.g., any partition coefficient and solubility.

In order to make the COSMO-RS approach suitable for very fast screening purposes, e.g., in the context of HTS studies, we are currently developing a fragment based approximation for σ-profiles (COSMO*frag*). The basic idea is to use a large database of about 10 000 pre-calculated σ-profiles of

medium-sized organic and drug-like compounds. For a new compound we then search for most similar substructures in the database, take the partial σ-profiles of the substructures and thus compose an approximate σ-profile for the new compound. If the cuts of fragments are done with chemical sense and augmented by some reasonable superposition rules on aromatic fragments, this procedure should only be about 0.2 or 0.3 log-units less accurate than the original method, at computational cost of a few seconds per compound. The advantages of COSMO*frag* compared with other fragment methods are that the flexibility of COSMO-RS regarding the properties is retained, that rather large fragments can be used since a large database of compounds is available, that the database can be easily supplemented by templates typical for the present study and that missing fragments can be added within a few hours by additional DFT/COSMO calculations.

COSMO-RS as a novel tool for modelling in drug design and development

The above considerations have shown that COSMO-RS on the one hand gives a novel view to partition and solubility problems, and on the other hand it provides a novel tool for molecular modelling in drug design and development. Although there may be many more ways to benefit from COSMO-RS in this area, the following list summarises the most important aspects according to our present knowledge:

(1) COSMO-RS introduces the screening charge density σ as a key parameter for local intermolecular interactions, including electrostatics, hydrogen bonding, as well as 'hydrophobic' effects. σ is better polarity descriptor than the electrostatic potential, since the latter is influenced too strongly by long range effects of ionic charges, while in accordance with the real situation of molecules in solution, σ shows a very local compensation of charges and multipoles.

(2) COSMO-RS introduces the vivid concept of σ-profiles for a qualitative and quantitative comparison of the polarity distribution on molecular surfaces.

(3) σ from COSMO-RS may be used as integral descriptor for electrostatics, hydrogen bonding, and hydrophobicity in molecular field analysis (MFA), whereas so far multiple fields are used.

(4) Being the key interaction parameter, σ can be visualised on the molecular surface, e.g., by Cerius^2 [28] based on DMol/COSMO files [29, 30].

(5) Several efforts have been reported which just try to refit partition coefficients from CLOGP to the molecular surface. Because in COSMO-RS any logarithmic partition coefficient is calculated as a surface integral of some function $\Delta\mu_{S,S'}(\sigma)$, partition coefficients can visualised as a local property on the molecular surface.

(6) Due to the quantum chemical basis of COSMO-RS, σ and all the derived thermodynamic properties can be studied as a function of conformation, substitution and isomerisation.

(7) Solubility, and especially relative solubility, in almost arbitrary solvents can easily be calculated. This allows for efficient solvent screening, which may be useful in various steps during drug development.

(8) Partition coefficients between arbitrary solvents of known chemical composition can directly be treated based on the fast statistical COSMO-RS thermodynamics.

(9) Quantitative equations for logarithmic partition and adsorption coefficients involving complicated physiological phases or other non-trivial phases can be easily developed by a linear regression analysis with respect to the σ-moments.

(10) The σ-moments are a rather orthogonal and complete set of descriptors for partitioning, which is valuable in any QSAR study on activities which involve partition behaviour.

(11) The ability to efficiently calculate the chemical potential μ_X^X of a compound in its pure state places COSMO-RS in a unique position regarding solubility (and vapour pressure) prediction.

(12) The concept of COSMO-RS can be generalised to membranes and membrane partitioning.

(13) As a future perspective, COSMO-RS may be considered an interesting novel approach for the description of receptor ligand binding, which automatically includes desolvation.

A great value of COSMO-RS is the ability to access most diverse local and global aspects of physicochemical and physiological partitioning and solubility based on the same molecular description, i.e. on the screening charge density σ. The time-demanding step of the COSMO calculation has to be done only once per molecule, in order to have the entire range of properties available. But it can be done at different computational levels, starting with a rapid COSMO*frag* composition for screening purposes, over an intermediate level based on semi-empirical geometries and a small basis set DFT/COSMO, up to highest level

full DFT/COSMO calculations for very accurate and detailed studies.

Summarising, we are convinced that beyond its primary application field of chemical engineering thermodynamics, COSMO-RS provides a powerful and valuable novel access to many important question in drug design and development.

References

1. Ben Naim, A., Solvation Thermodynamics, Plenum Press, New York, NY, 1987.
2. Jorgensen , W.L. , In Schleyer, P. V.R. and Allinger, L., (Eds), Computation of Free Energy Changes in Solution, Encyclopedia of Computational Chemistry , Vol. 2, Wiley, New York, NY, 1998, pp. 1061–1070.
3. Schäfer, H., van Gunsteren, W.F. and Mark, A.E, J. Comp. Chem., 20 (1999) 1604.
4. Duffy, E.M. and Jorgensen, W.L., J. Am. Chem. Soc., 122 (2000) 2878.
5. Jorgensen, W.L. and Duffy, E.M., Bioorg. Med. Chem. Lett., 10 (2000) in press.
6. Hansch, C. and Leo, A.J., Substituent Parameters for Correlation Analysis in Chemistry and Biology, Wiley, New York, NY, 1979.
7. CLOGP program, BioByte Corporation, Claremont, CA.
8. Duban, M.E., Bures, M.G., DeLazzer, J. and Martin, Y.C., Helv. Chem. Acta (2000) in press.
9. Abraham, M.H., Chem. Soc. Rev., 22 (1993) 72.
10. Gmehling, J., Fluid Phase Equilibria, 144 (1998) 37.
11. Cramer, C.J. and Truhlar, D.G., in Lipkowitz, K.B. and Boyd, D.B. (Eds), Reviews in Computational Chemistry, Vol. 6, VCH Publishers, New York, NY, 1995.
12. Tomasi, J. and Persico, M., Chem. Rev., 94 (1994) 2027.
13. Klamt, A. and Schüürmann, G., J. Chem. Soc. Perkin Trans., 2 (1993) 799.
14. Giesen, D.J., Gu. M.Z., Cramer, C.J. and Truhlar, D.G., J. Org. Chem., 61 (1996) 8720.
15. Klamt, A., COSMO and COSMO-RS, in Schleyer, P. v.R. and Allinger, L. (Eds.) Encyclopedia of Computational Chemistry , Vol. 2, Wiley, New York, NY, 1998, pp. 604–615.
16. Klamt, A., J. Phys. Chem., 99 (1995) 2224.
17. Klamt, A., Jonas, V., Buerger, T. and Lohrenz, J.C.W., J. Phys. Chem., 102 (1998) 5074.
18. Klamt, A. and Eckert, F., Fluid Phase Equilibria, 172 (2000) 43.
19. Bondi, A., Phys. Chem., 68 (1964) 441.
20. Mehler, K., Thesis, TU Münich (to be published).
21. Busalla, T., Diploma Thesis, Univ. Cologne, 1996.
22. Klamt, A., Eckert, F., Hornig, M. and Blake, J.F. (to be published).
23. Klamt, A., Eckert, F. and Hornig, M. (to be published).
24. Ahlrichs, R., Bär, M., Häser, M., Horn, H., and Kölmel, C., Chem. Phys. Letters, 162 (1989) 165.
25. Schäfer, A., Klamt, A., Sattel, D., Lohrenz, J.C.W. and Eckert, F., Phys. Chem. Chem. Phys., 2 (2000) 2187.
26. MOPAC2000 program, Fujitsu, Japan, 1999.
27. COSMOtherm program, COSMOlogic GmbH&CoKG, Leverkusen, Germany, 1999.
28. Cerius^2 program, Molecular Simulations, San Diego, CA.
29. DMol^3 program, Molecular Simulations, San Diego, CA.
30. Andzelm, J., Kölmel, C. and Klamt, A., J. Chem. Phys., 103 (1995) 9312.