Refinement and Parametrization of COSMO-RS

Andreas Klamt,* Volker Jonas,[†] Thorsten Bürger, and John C. W. Lohrenz

Bayer AG, IM-FA, Q18, D-51368 Leverkusen, Germany

Received: October 29, 1997; In Final Form: February 4, 1998

The continuum solvation model COSMO and its extension beyond the dielectric approximation (COSMO-RS) have been carefully parametrized in order to optimally reproduce 642 data points for a variety of properties, i.e., ΔG of hydration, vapor pressure, and the partition coefficients for octanol/water, benzene/water, hexane/water, and diethyl ether/water. Two hundred seventeen small to medium sized neutral molecules, covering most of the chemical functionality of the elements H, C, N, O, and Cl, have been considered. An overall accuracy of 0.4 (rms) kcal/mol for chemical potential differences, corresponding to a factor of 2 in the equilibrium constants under consideration, has been achieved. This was using only a single radius and one dispersion constant per element and a total number of eight COSMO-RS inherent parameters. Most of these parameters were close to their theoretical estimate. The optimized cavity radii agreed well with the widely accepted rule of 120% of van der Waals radii. The whole parametrization was based upon density functional calculations using DMol/COSMO. As a result of this sound parametrization, we are now able to calculate almost any chemical equilibrium in liquid/liquid and vapor/liquid systems up to an accuracy of a factor 2 without the need of any additional experimental data for solutes or solvents. This opens a wide range of applications in physical chemistry and chemical engineering.

1. Introduction

Chemical equilibria between different liquid phases, or between liquids and vapors, control almost all biological and industrial chemistry. Therefore, understanding and, even more importantly, predicting such equilibria is of tremendous importance for the control and optimization of chemical products and processes.

Dielectric continuum solvation models[1,2] (CSMs) like PCM[3] or COSMO[4] have turned out to be elegant and efficient methods for the inclusion of solvent effects in quantum chemical calculations. At costs comparable to gas-phase calculations, they are capable of giving a surprisingly good description of the properties and energetics of molecules in various solvents, especially in water. Parametrizations of such models have been reported[5] that allow for the calculation of ΔG_{hydr} with an accuracy of about 0.5 kcal/mol, which corresponds to an uncertainty of a factor 2.3 for the associated equilibrium constant, i.e., for Henry’s law constant.

Despite the considerable success of the dielectric CSMs, they are hardly justifiable from a theoretical point of view. This is because the electric fields on the molecular surfaces of fairly polar solutes are so strong that the major part of the solvent polarizability, i.e., the reorientation of static dipoles, no longer behaves linearly, as it does in the macroscopic limit, but it is almost at saturation. Although the solvent water appears to behave almost linear up to surprisingly strong electric fields, there cannot be any doubt that dielectric theory does not account for this situation in general and that, even for water, major deviations from linearity occur[6−8] (also see Appendix 1). Starting from this insight, one of us (A.K.) has proposed a novel and very fruitful concept called COSMO-RS[9] (conductor-like screening model for real solvents), which avoids the questionable dielectric approach. This theory takes the ideally screened molecules as a starting point for the description of molecules in solution. The deviations from ideal screening, which unavoidably occur in any solvent, are described as pairwise misfit interactions of the ideal screening charges on contacting parts of the molecules in the fluid. A detailed description of the COSMO-RS concept will be given in section 3. This concept describes solvent and solute on the same footing, i.e., starting from COSMO calculations for all molecules appearing in the system under consideration. It finally leads to the fact that the solvent water has the unique ability to almost ideally screen a solute due to its broad and well-balanced distribution of screening charge density on its surface. Thus not only does COSMO-RS give an answer to the question “why are CSMs quite successful in the treatment of the solvent water' it even represents a tremendous generalization of the CSM approach. This is because it no longer depends on experimental data or any parametrization for the solvent. Finally, it describes mixed solvents as well as pure ones. As soon as the COSMO calculations are available, it efficiently enables the calculation of the chemical potential of almost any solute in almost any solvent. Thus it is capable of treating almost the entire equilibrium thermodynamics of fluid systems and should become a powerful alternative to fragment-based methods like UNIFAC.[10]

In this article we will describe a careful and sound optimization of the relatively few parameters within COSMO-RS. Due to the much broader range of properties accessible by COSMO-RS compared to usual CSMs, a large data set was available for the optimization. In order to obtain reliable electrostatic potentials, we based the optimization on density functional theory (DFT), using the program DMol.[11,12] DFT calculations yield considerably more reliable molecular potentials than the semiempirical methods, which have been used within the program MOPAC[13,14] in the original COSMO and COSMO-RS papers. The optimization of the final 18 parameters turned out as rather sophisticated. This was partly due to the strongly nonlinear behavior of the problem, which yields multiple

[†] Present address: University of Zürich, Switzerland.

S1089-5639(98)00017-6 CCC: $15.00
Published on Web 06/05/1998

minima, and partly due to the appearing need for some conceptual improvements of the DMol/COSMO[15] implementation. These improvements involved a correction for outlying charges[16] and an improvement of the cavity construction, which will be described in section 2. Nevertheless, we obtained a consistent and satisfactory parametrization, which allows for the calculation of chemical potential differences with an accuracy of about 0.4 kcal/mol. This corresponds to an uncertainty factor of 2 in the associated equilibrium constants.

In order to present a complete picture of the refined COSMO-RS method, we first present a survey of COSMO and its implementation in DMol (section 2), a slightly modified rederivation of the COSMO-RS theory (section 3), and a description of the data set and of the optimization procedure (section 4). Results and discussion will be presented in section 5. A summary and an outlook are given in section 6. Finally, the full concept of COSMO-RS is summarized as a recipe in Appendix 2. We are aware that sections 2 and 3 involve some redundance with earlier papers; however we make a substantial presentation to aid understanding of the method and its refinements that are described in this paper.

2. COSMO and Its DMol Implementation

The basic idea of COSMO, compared to other CSMs, is the use of the boundary condition for the total potential

$$0 = \Phi_{\text{tot}} = \Phi^X + \Phi(q^*) \tag{1}$$

for the calculation of the screening charges q^* appearing on the cavity of a solute X, when embedded in a conductor, and to scale these charges by a factor

$$f(\epsilon) = \frac{\epsilon - 1}{\epsilon + 0.5} \tag{2}$$

to approximately yield the screening charges q at a finite dielectric constant ϵ. This replaces the direct use of the corresponding, but more complicated and numerically less stable, dielectric boundary condition for the electric field:

$$4\pi\sigma = E_{\text{tot}} n = (E^X + E(q))n \tag{3}$$

In these equations q and q^* denote the sets of screening charges that appear on the surface segments of a sufficiently fine discretization of the cavity surface. σ is the corresponding local screening charge density on one of these segments, and n denotes the outward normal vector of this segment. With the advantage of simplicity and numerical stability, the COSMO approximation has proven to be sufficiently close (i.e., within about 10%) to the exact results as resulting from eq 3 at the lower end of dielectric constants of solvents ($\epsilon \approx 2$), while it asymptotically coincides with the dielectric results at high dielectric constants, being safely within 0.5% error at the dielectric constant of water ($\epsilon \approx 80$).

As mentioned above, we do not consider the dielectric model to be relevant for the description of the screening behavior of solvents on a molecular scale. In the following, COSMO will exclusively be used for the self-consistent calculation of geometries, energies, and screening charge densities of molecules at their ideally screened state, i.e., with $f(\epsilon) = 1$. At this state COSMO is by definition exact, and although they should asymptotically be able to yield identical results, the truely dielectric CSMs such as PCM are evidently less suited for this task.

As described in the original COSMO paper,[4] a molecular shaped, van der Waals type cavity is constructed for each solute. The density of the basis grid is kept at its default of 1082 points per unit sphere, but compared to the original MOPAC implementation a slightly improved, i.e., slightly more homogeneous, grid is used. For the segment construction, a density of NSPA = 92, which means approximately 92 segments per unit sphere, is used. This turned out to be sufficiently fine to keep the discretization error safely below the final uncertainty of the method.

During the parametrization process of COSMO-RS, a closure of the originally open parts of the surface along the intersection lines of atomic spheres turned out to be useful. This avoids artificially large screening charge densities on small and isolated surface fragments, which otherwise appeared in rare but important cases, e.g., for ethers and amines. Thus a straightforward algorithm has been developed and implemented, which smoothly closes the originally open regions by sets of triangles. Thereby the total number of segments m increases by approximately 50%. The resulting increase of the costs of the COSMO algorithm, which partly scales with m^3, is not critical in combination with density functional calculations in the current DMol implementation. Details of the cavity closure will be published elsewhere. It should be pointed out here that the energetic implications of this closure turned out to be almost negligible, i.e., within 2% for most molecules. But the screening charge densities now are free of artifacts. Thus the open cavity used in the original COSMO is a reliable, time-saving approximation as long as screening densities are not explicitly required.

In order to avoid, as much as possible, interferences with insufficiencies of the underlying quantum chemical method, we decided to use a density functional method (DFT). DFT is known to be able to yield ground-state properties, especially ground-state charge distributions, i.e., densities, as reliable as Hartree−Fock (HF) calculations with higher order correlation corrections, but at much lower costs. The actual code used is DMol,[11,12,15] which has the additional advantage of using numerical atomic basis sets. These, even at the default level, have sufficiently good tails to reliably reproduce quantities such as dipole moments and polarizabilities. These properties are of crucial importance for any solvation calculation. As a validation of the suitability of DMol, gas-phase dipole moments for a representative set of 64 molecules composed of the elements H, C, and O have been calculated using the semiempirical Hamiltonian AM1,[17] density functional theory (DMol: SVWN[18]/DNP and BPW91[19−21]/DNP; Gaussian94:[22] BP86[19,23]/6-31G(d), BP86/SVP,[24,25] and B3LYP[26,27]/6-311G(d)), and ab initio Hartree−Fock methods (Gausssian94: HF/6-31G(d), HF/6-311G(d,p), and MP2/6-31G(d)). All calculations were single-point calculations using DMol:BPW91/DNP optimized geometries. The resulting analysis, regarding the accuracy of these methods, is presented in Table 1.

As soon as reasonably good basis sets are chosen, the different DFT methods, as well as MP2 calculations, yield good agreement with experimental dipole moments, with a standard deviation of about 0.15 D. The slope of the best regression line for these methods is almost identical to unity. It should be noted that some of the major deviations are common to all of these methods. We take this as an indication that a considerable part of the error may arise from conformational averaging in the experimental data and/or from experimental errors. Thus we may conclude that an accuracy of about 0.1 D can be achieved with any of the mentioned methods. It is

TABLE 1: Accuracy of Dipole Moments Calculated with Different Quantum Chemical Methods for 64 Compounds of Elements H, C, and O

method	program	basis set	rms unscaled	rms scaled	slope
DFT					
S-VWN[18]	DMol[11,12,15]	DNP	0.1649	0.1585	1.0200
BPW91[19−21]	DMol	DNP	0.1380	0.1300	1.0228
BP86[19,23]	Gaussian94[22]	6-31G(d)	0.1500	0.1483	0.9755
BP86	Gaussian94	SVP[24,25]	0.1374	0.1379	0.9927
B3LYP[20,21]	Gaussian94	6-31G(d)	0.1479	0.1447	1.0124
ab initio					
HF	Gaussian94	6-31G(d)	0.3293	0.1682	1.1630
HF	Gaussian94	6-311G(d,p)	0.3137	0.1401	1.1648
MP2	Gaussian94	6-31G(d)	0.1953	0.1919	1.0124
semiempirical					
AM1	Gaussian94		0.1902	0.1909	0.9769

remarkable that even the semiempirical AM1 method yields the right slope and almost the same accuracy as the DFT and MP2 methods on this data set, which contains only the three elements C, H, and O. Nevertheless, AM1, as well as most other semiempirical Hamiltonians, miscalculates the dipole moments of the compounds containing nitro or cyano groups by about 0.4 D. Thus they are less suited for general use. HF dipole moments are worse on our data set, but they can be brought into the same range of accuracy as the DFT methods by a scaling factor of 0.8. Nevertheless, the introduction of a scaling factor or alternatively the correction of the systematically overestimated dipole moments by larger cavity radii in CSM calculations would result in serious conceptional problems as soon as other multipole contributions, such as monopoles or quadrupoles, become important. Thus, the use of uncorrelated HF methods in CSM is not recommended. The choice of the DFT functional is of minor importance. Altogether, our validation clearly shows that DMol with the BPW91 functional yields reliable densities, and it should be a sound basis for the optimization of a solvation model. Nevertheless, at least in one case (cyclohexanone) we have to realize a substantial overestimation of the dipole moment by DMol, and as a result, this compound turns out to be one of the few larger outliers in our final results.

The use of any ab initio or DFT code inevitably introduces an additional complication into the concept of CSMs: the existence of some small part of the electronic density that is located outside of the cavity. As a part of the refinement and parametrization project, the problem of this outlying charge has been carefully analyzed by two of the authors.[12] It turned out that COSMO is considerably less sensitive to outlying charges than the true dielectric approach. Nevertheless, at reasonable cavity radii the outlying charge error is up to 25% for anions and neutral compounds, while it is much smaller for cations. As a result of our investigation, we have developed and implemented a rigorous algorithm for the removal of such outlying charge error by the introduction of an auxiliary cavity lying approximately 1 Å further outside the main cavity. Energies as well as screening charges now are reliably corrected for outlying charge effects. Considering the magnitude of the effect and its strong radii dependence, it is evident that any radii optimization without such a rigorous outlying charge correction must be subject to serious artifacts.

The gas-phase reference energies for all the structures of the data set were obtained from both S-VWN and BPW91 gas-phase optimizations applying the DNP basis of DMol. All structures were then reoptimized in a continuum conductor, i.e., with COSMO and $f(\epsilon) = 1$, using NSPA = 92, the closed cavity option, and the outlying charge correction. Structures for which the minimum conformation in solution differs from that of the gas phase were removed from the data set. Such structures, although in principle treatable by COSMO-RS, are less suited for the parametrization. For all other compounds the geometric changes generally were small.

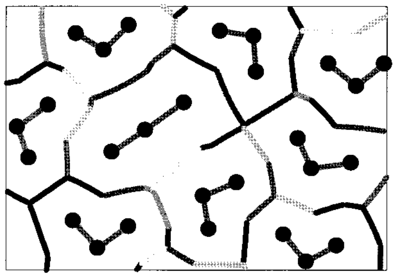

Figure 1. Schematic construction of molecular cavities.

The resulting ideal net screening energy gains

$$\Delta^X = E^X_{\text{gas}} - E^X_{\text{COSMO}} \tag{4}$$

of all molecules X are highly correlated with the bare screening energies E^X_{diel}. The latter is called dielectric energy in the COSMO nomenclature, and it is defined as half of the electrostatic interaction energies of the ideally screened and self-consistently polarized solutes with their screening charges:

$$E^X_{\text{diel}} = \frac{1}{2}\sum_\nu \Phi^X_\nu q^*_\nu = \frac{1}{2}\sum_\nu \Phi^X_\nu s_\nu \sigma^*_\nu \tag{5}$$

Here s_ν and σ_ν denote the area and the ideal screening charge density on a segment ν, respectively. E^X_{diel} and Δ^X are remarkably well correlated ($r^2 = 0.99$) with a slope of 0.80.

The performance of the DMol/COSMO calculations is comparable to that of gas-phase calculations. For single-point calculations the COSMO overhead in average is about 10%, while for geometry optimization with COSMO convergence is somewhat worse due to small inaccuracies in the gradients.

3. Basic Theory and Refinements of COSMO-RS

3.1. Concept of Misfit Relative to the Ideally Screened State. As discussed in the Introduction, the macroscopic dielectric theory is untenable as an explanation for the success of CSMs. A surprising, and extremely fruitful, explanation arises from the following consideration of initially ideally screened molecules: Imagine a snapshot of an ensemble of molecules in a condensed medium as schematically illustrated in Figure 1. All molecules are touching their neighbors at distances corresponding to the vdW radii of atoms. Now let us divide the entire volume of the system into molecular cavities which are defined as the union of all those points that have a smaller relative distance to an atom of the molecule under consideration than to other molecules. Here the relative distance is defined as the ratio of distance and vdW radius of the entire atom. This construction, yielding polyhedral cavities with slightly curved faces, is schematically illustrated in Figure 1. Since the closest points of such cavities are about a vdW radius away from the nearest atom, the mean distance of the cavity is somewhat larger. A detailed analysis yields that it corresponds to about 120% of the vdW radii. Because in a fluid the position of neighbor molecules fluctuates in time, the average molecular cavity of a solute is not such a pseudo-polyhedron, but it is considerably smeared out and it corresponds to something like a solvent-accessible surface constructed with vdW radii increased by about 20%, i.e., typical cavities as used in CSMs. As a consequence of the above construction, for each individual snapshot the set of polyhedra is space filling. Hence the volumes of the averaged cavities must be space filling as well; that is, they have volumes close to the molecular volumes as derived from the densities.

Now let us make an auxiliary assumption that has nothing to do with reality but that is the key for the following steps. We assume that all of the cavity surfaces are perfect, grounded conductors. Each molecule finds itself in a situation as described by the COSMO model at infinite dielectric constant, i.e., being enclosed in a conducting cavity of about 120% of the vdW radii. Using the averaged cavity for all molecules of the same species instead of each specific one, the energy of such molecules as well as the screening charges appearing on the cavities are quite well evaluated by a standard COSMO calculation. Thus we have an efficient way to calculate the total energy of this artificial ensemble of molecules, which are seperated by conducting interfaces, by just performing a COSMO calculation for each different type of molecule in the ensemble.

In order to get back to the real state of the condensed medium, we have to get rid of the conductor again. As a first step, let us consider the screening charges as well as the molecular polarizations as frozen in their ideally screened state. This does not change anything for the moment. We now remove, one after the other, small pieces of surface, each having an area a_{eff}, which is something like the effective contact area of atoms. Each of these surface patches is carrying a specific screening charge density

$$\sigma_{\mathrm{res}} = \sigma + \sigma' \tag{6}$$

where σ and σ' are the local screening charge densities of the two molecules sharing the surface patch under consideration. Obviously, if σ is the negative equivalent of σ', there is no screening charge density left and the conductor can be removed without changing the situation. The two neighbors screen each other on such a part of the contact surface as well as the conductor did before. In the general case of nonvanishing σ_{res} we have to prepare a suitable piece of surface, having the negative of the residual screening charge density, and place it at the position of the patch. Then it just cancels the residual screening charge density of that patch, and the situation is equivalent to having no conductor on that piece of contact surface. Having compensated the residual screening charge density of all surface patches in this way, the energy of the system is composed of four contributions: (a) the energy of the ideally screened system as considered before; (b) the interaction energy of the compensation patches with the ideally screened system; (c) the interaction energy of the compensation patches with each other; (d) the sum of the self-energies of the compensation patches. Since the electrostatic potential of the ideally screened system on the cavity is zero by definition, the contribution (b) is zero. Under the assumption that the residual charge densities on the patches are not correlated, contribution (c) should also be zero due to the random sign of the different summands. Thus the total energy is given by energy of the ideally screened system plus the sum of the self-energies of the compensation patches, each of which is positive and given by

$$E_{\mathrm{misfit}}(\sigma,\sigma') = \frac{\alpha}{2}(\sigma + \sigma')^2 \tag{7}$$

with

$$\alpha = \frac{0.3}{\epsilon_0} a_{\mathrm{eff}}^{\ 3/2} \tag{8}$$

Since this energy results from the misfit of the contacting ideal screening charge densities, in the following we will call it misfit energy. The constant α is easily derived from simple electrostatics.[4,9]

Now we have removed all conductor screening charges from the system, and the situation is closer to reality, again. But the polarization of the molecules so far remains frozen in the state of ideal screening, although the electrostatic situation has changed meanwhile due to the removal of the residual screening charges, i.e., the addition of the compensation patches. In reality the molecules will respond to this change by their electronic polarizability. This results in a reduction of the misfit energy. Since the overall electronic polarizability is well represented by a homogeneous dielectric medium of $\epsilon = n^2 \approx 2$, where n is the refraction index of the solvent, the reduction of the misfit energy approximately corresponds to a factor

$$f_{\mathrm{pol}} = 1 - f(\epsilon{=}n^2) = 1 - \frac{(n^2 - 1)}{(n_2 + 1/2)} \approx 0.6 \tag{9}$$

where the dielectric scaling factor of COSMO has been applied. This factor will be subsumed within the misfit energy constant α, which we further on call α′. Thus we end up with the result that the energy of an ensemble of molecules in the condensed state is quite well approximated by the sum of the energies of all molecules in their ideally screened state plus the sum of all misfit energies resulting from contacts of surface patches:

$$E_{\mathrm{condensed}}^{\mathrm{tot}} = \sum_X E_{\mathrm{ideal}}^X + \frac{\alpha'}{2}\sum_\nu (\sigma_{\nu 1} + \sigma_{\nu 2})^2 \tag{10}$$

Here the indices X and ν denote the molecules and surface patches, respectively, and $\sigma_{\nu 1}$ and $\sigma_{\nu 2}$ are the two ideal screening charge densities contributing to patch ν. Equation 10 is remarkable in that the electrostatic interaction of the molecules in the ensemble, including polarization, is expressed as a simple summation over the contact surface. If the molecule in vacuum is taken as reference point, the energy E_{ideal}^X is composed of the net electrostatic energy gain Δ^X of the molecule in the transition from vacuum to the ideal conductor, including back-polarization and eventual contributions from geometry relaxation, and a dispersion term $\gamma_k A_k^X$, where the γ_k are element-specific constants and the A_k^X are the corresponding portions of the surface area. Thus all input for eq 10 is available from the initial COSMO calculation with the exception of the exact value of the effective contact area a_{eff} and the polarization factor f_{pol}, both of which are subsumed in α′. These parameters, together with the dispersion constant γ, have to be finally fixed by fitting to experimental data.

The screening charge densities σ_ν, which appear in the above consideration, here and further on are understood as mean values over surface patches. They can be derived from the COSMO output by averaging of the original ideal screening charge densities σ_ν* over a region of radius r_{av}. For this task we have employed the following averaging algorithm:

$$\sigma_\nu = \sum_\mu \sigma_\mu^* \frac{r_\mu^{\ 2} r_{\mathrm{av}}^{\ 2}}{r_\mu^{\ 2} + r_{\mathrm{av}}^{\ 2}} \exp\left\{-\frac{d_{\mu\nu}^{\ 2}}{r_\mu^{\ 2} + r_{\mathrm{av}}^{\ 2}}\right\} \Bigg/ \sum_\mu \frac{r_\mu^{\ 2} r_{\mathrm{av}}^{\ 2}}{r_\mu^{\ 2} + r_{\mathrm{av}}^{\ 2}} \exp\left\{-\frac{d_{\mu\nu}^{\ 2}}{r_\mu^{\ 2} + r_{\mathrm{av}}^{\ 2}}\right\} \tag{11}$$

Here $d_{\mu\nu}$ is the distance of segments μ and ν, and r_μ is the mean radius of segment μ, i.e., $r_\mu^{\ 2} = s_\mu/\pi$. This averaging procedure

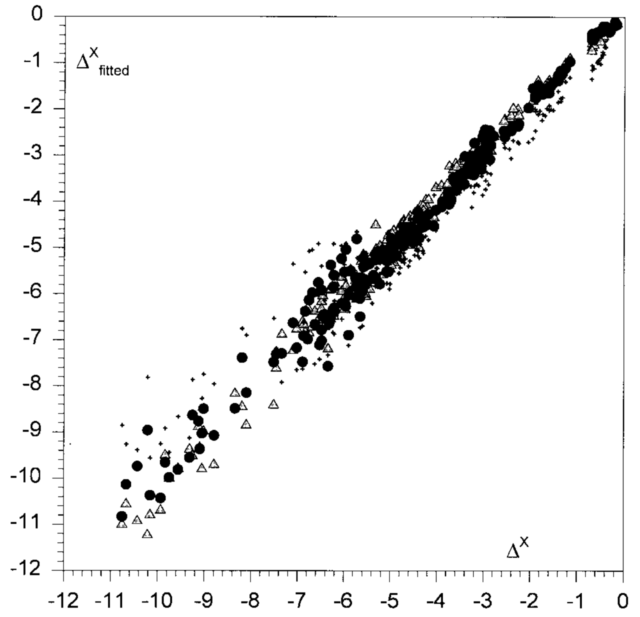

Figure 2. Quality of the fit of the ideal screening energy Δ'^X (in kcal/mol) by different descriptors: (a) $\Delta'^X_{\text{fitted}} = 0.8E'^X_{\text{diel}}$ (triangles, $r^2 = 0.995$), (b) fit according to eq 13 (small crosses, $r^2 = 0.985$), (c) fit according to eq 15 (filled circles, $r^2 = 0.995$).

is necessary because we have assumed constant charge density on each surface patch. Ideally we would expect r_{av} to be equal to the radius r_{eff} corresponding to the effective contact area a_{eff} of the independent surface patches, but—for reasons we could not identify yet—it turned out that an independent optimization of both parameters yields a considerable improvement of the fit, with r_{av} being about a factor 3 smaller than r_{eff}. As long as r_{av} is smaller than the correlation length of the screening charge density on the cavity surfaces, the effect of this averaging process is rather small. Nevertheless, there is a small energetic shift from the original ideally screened state to the new averaged ideally screened state. Since we want to take the averaged ideally screened state as a starting point for further considerations, we redefine the net ideal electrostatic screening energy as

$$\Delta'^X = \Delta^X + 0.8(E'^X_{\text{diel}} - E^X_{\text{diel}}) = \Delta^X + 0.4\sum_\nu s_\nu \Phi^X_\nu \sigma_\nu - 0.4\sum_\nu s_\nu \Phi^X_\nu \sigma^*_\nu \quad (12)$$

This energy is plotted versus the corresponding averaging corrected dielectric energy E'^X_{diel} (see second term of the right side of eq 12) in Figure 2. Like their original correspondences, both quantities are highly correlated ($r^2 = 0.996$), and the regression constant is 0.8. Starting from the self-consistent ideally screened state of solute X, we would have to raise the energy $-E'^X_{\text{diel}}$ for the transfer of X into vacuum if the polarization of X would be frozen. Allowing for electronic relaxation of the solute, this transfer energy would reduce to $-\Delta'^X$.

Since in this case only the solute is polarizable, while its environment, i.e., the vacuum, is not, the ratio of Δ'^X and E'^X_{diel} should correspond to $f_{\text{pol}}{}^{1/2}$. Thus we get the estimate $f_{\text{pol}} \approx 0.8^2 = 0.64$ from this consideration, being rather compatible with the previous estimate of $f_{\text{pol}} \approx 0.6$ in eq 9.

Before we procede with the derivation of the final COSMO-RS formulas we should make the following consistency consideration for a consolidation of the presently achieved status: In view of COSMO-RS a solute embedded in a virtual ensemble of nonpolar and nonpolarizable molecules should be electrostatically equivalent to a molecule in the vacuum. Since in such an ensemble the only nonvanishing screening charges are those of the solute, the residual screening charges are identical to the solute screening charges, and the total electrostatic energy of the electronically frozen solute in vacuum relative to its ideally screened state must be given by

$$E^{\text{frozen}\ X}_{\text{vacuum}} = \frac{\alpha}{2}\sum_{\nu\in X} s_\nu \sigma_\nu{}^2 \cong -E'^X_{\text{diel}} \cong -f_{\text{pol}}{}^{-1/2}\Delta'^X \quad (13)$$

Apparently this should be just the negative of the dielectric energy E'^X_{diel}. Since the latter is linearly related to the net ideal screening energy Δ'^X (vide infra), we expect a strong correlation between the sum in eq 13, which in the limit of small segment areas s_ν is the surface integral of the squared screening charge density, and Δ'^X. Indeed, this correlation is crucial for the entire COSMO-RS method. Both quantities are considerably correlated ($r^2 = 0.985$, see Figure 2), but the standard deviation still is about 0.6 kcal/mol, which is more than the anticipated accuracy of the COSMO-RS method. There is an obvious systematics in the residuals: Δ'^X is overestimated for compounds having large exposed areas of high polarity, like carbonyls or nitriles, while it is too low for compounds with small polar hot spots on the surface, as they typically appear on sp^3-oxygen or sp^3-nitrogen atoms. The reason for this is the high correlation of the screening charge densities over the relatively large surface of the sp^2-oxygen, while sp^3-oxygens usually have much smaller solvent-accessible surfaces and hence show less correlation in the screening charge densities. Such correlation contradicts the preconditions made in the derivation of COSMO-RS. Thus, a better description of the dielectric energy should be achievable if correlation is taken into account to some degree. This can be done by using a second screening charge density $\sigma_\nu{}^\circ$, which is derived from the original screening charge densities σ^*_ν by averaging over an area of radius $2r_{\text{av}}$ instead of r_{av}. Although the σ_ν and $\sigma_\nu{}^\circ$ are quite correlated, we can construct an independent descriptor σ''_ν from σ'_ν by orthogonalizing it over the entire data set, yielding

$$\sigma_\nu{}^\perp = \sigma_\nu{}^\circ - 0.816\sigma_\nu \quad (14)$$

$\sigma_\nu{}^\perp$ now is a descriptor for the correlation between the screening charge density on the segment ν with its surrounding. Hence the energy of each screening charge density σ_ν now should be corrected for the interaction with its surrounding, and we expect a relationship as expressed in eq 15 instead of eq 13. The

$$\Delta'^X \cong -\frac{\alpha}{2} f_{\text{pol}}{}^{1/2}\sum_{\nu\in X} s_\nu \sigma_\nu(\sigma_\nu + f_{\text{corr}}\sigma_\nu{}^\perp) = -\frac{\alpha}{2} f_{\text{pol}}{}^{1/2}\Big[\sum_{\nu\in X} s_\nu \sigma_\nu{}^2 + f_{\text{corr}}\sum_{\nu\in X} s_\nu \sigma_\nu \sigma_\nu{}^\perp\Big] \quad (15)$$

optimal value of f_{corr} can easily be determined by bilinear regression of Δ'^X with respect to the two sums on the right-hand side of eq 15. Obviously the exact value of f_{corr} depends on the averaging radius r_{av}, but in the range of the final optimum of $r_{\text{av}} = 0.5$ Å we got $f_{\text{corr}} = 2.4$. The total slope, i.e., $-1/2\alpha f_{\text{pol}}{}^{1/2}$, comes out as 1110 kcal/mol Å^2/e^2. By the introduction of this screening charge correlation correction the correlation coefficient improves to $r^2 = 0.996$ (see Figure 2), and the standard deviation decreases to 0.3 kcal/mol, being within the anticipated accuracy.

3.2. Statistical Thermodynamics and Chemical Potentials.

Although eq 10 is a considerable simplification compared to the standard evaluation of the energy of an ensemble of molecules, it is of limited use, since it requires the knowledge of all the neighborhood relations in the ensemble, i.e., the full information about the coordination of the molecules. This usually is known for molecules in crystals, and thus eq 10 may be an interesting way for the expression of interaction energies in organic crystals. But in the case of liquids or other disordered systems, which are our primary focus here, this information is not easily available. Even more, these neighborhoods are rapidly changing in time. An appropriate statistical average could only be generated by demanding thermodynamic sampling of the entire ensemble, but such would make most calculations unacceptably expensive.

In order to achieve an enormous reduction of the complexity of the problem, we now introduce an initially quite daring approximation, which in the end turns out to be rather accurate and extremely fruitful. Realizing that in eq 10 not the full geometric information is required but only the information on the neighborhood of screening charge densities, we may virtually cut all cavities into their effective contact segments and postulate that the statistical averaging can be done for the resulting ensemble of separated surface patches, each carrying its ideal screening charge density as known from the COSMO calculations. Let us assume that these patches are thermodynamically independent entities with the only boundary condition that they have to form pairs in order to represent the situation in a condensed medium, having almost no free surface in the bulk. Thus our initial ensemble of molecules is replaced by the corresponding ensemble of pairwise interacting surface patches. Obviously all properties of such an ensemble must be a function of the composition of the ensemble, i.e., of the amount of patches or surface area having the same properties. Since so far the screening charge density σ is the only property of the patches, the ensemble is sufficiently characterized by the distribution of the patches with respect to σ. We call this distribution a σ-profile of the ensemble and abbreviate it as $p_S(\sigma)$, where the lower index S denotes the ensemble, or the solvent. All σ-profiles are assumed to be normalized to one molecule. Apparently the σ-profile of an ensemble of molecules is composed of the σ-profiles of its components X_i:

$$p_S(\sigma) = \frac{\sum_i x_i p^{X_i}(\sigma)}{\sum_i x_i} \tag{16}$$

Here the x_i denote the molar fractions of the different components, and $p^X(\sigma)$ is the σ-profile of a single molecule X. Obviously, for pure solvents consisting of a single component the solvent σ-profile $p_S(\sigma)$ is identical with the σ-profile of a single solvent molecule $p^X(\sigma)$. Nevertheless, it is of great practical importance that σ-profiles of mixed fluids are easily derived from the σ-profiles of the components. A few σ-profiles of representative solvents are shown in Figure 3.

After these considerations we are now ready to consider the statistical thermodynamics of the ensemble of surface patches characterized by a σ-profile $p_S(\sigma)$. The chemical potential $\mu'_S(\sigma)$ of an additional patch with charge density σ in one mole of patches of this ensemble is exactly given by the implicit equation

$$\mu'_S(\sigma) = -kT \ln[\int d\sigma' p'_S(\sigma') \exp\{(-E(\sigma,\sigma') + \mu'_S(\sigma'))/kT\}] \tag{17}$$

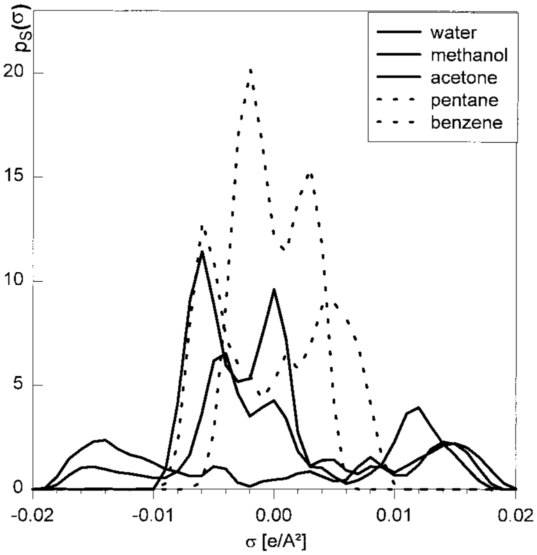

Figure 3. Four representative σ-profiles.

We later call the function $\mu'_S(\sigma)$ the σ-potential of the solvent S. In eq 17 $p'_S(\sigma)$ denotes the normalized σ-profile, i.e., $p_S(\sigma)/A^X$, and $E(\sigma,\sigma')$ is the interaction energy of the patches with screening charge densities σ and σ', respectively. Here we assume that $E(\sigma,\sigma')$ is given by the misfit energy as expressed in eq 7, except that α is replaced by α'. The derivation of this central equation is somewhat sophisticated, and we refer the interested reader to the original COSMO-RS article.[9] Using eq 17, the σ-potential $\mu'_S(\sigma)$ is easily iterated to self-consistency, starting from the initial guess $\mu'_S(\sigma) = 0$ in the integrand and updating it iteratively by new values yielded from integration. The whole procedure takes milliseconds on modern computers. We may conclude that the σ-potential $\mu'_S(\sigma)$, which parametrically depends on the temperature T, is almost exactly available from the corresponding σ-profile $p_S(\sigma)$ at negligible costs. The σ-potential is the key to all interesting thermodynamical properties of the solvent S. It tells us how much the solvent likes additional surface with screening charge density σ, i.e. surface of certain polarity. It includes the free energy necessary to remove the patches of the solvent molecules from their former partners, and it automatically covers cavitation energy as well.

For the optimization of the parameters it is useful to express the chemical potential of a patch per unit area in energy units of kT, i.e., $\tilde{\mu}_S(\sigma) = \beta^{-1}\mu'_S(\sigma)$ with $\beta = kT/a_{\mathrm{eff}}$. Using the corresponding definitions for the interaction energy, i.e., $\tilde{E}(\sigma,\sigma') = \beta^{-1}E(\sigma,\sigma')$, eq 17 simplifies to

$$\tilde{\mu}_S(\sigma) = -\ln[\int d\sigma' p'_S(\sigma') \exp\{-\tilde{E}(\sigma,\sigma') + \tilde{\mu}_S(\sigma')\}] \tag{18}$$

For any molecule X the standard chemical potential at unimolar concentration of X in solvent S, expressed relative to the ideally screened state, can now be calculated by integration of the σ-potential of the solvent weighted by the σ-profile of the solute. Thus we get

$$\mu_S^{*X} = \int d\sigma\, p^X(\sigma)\, \mu'_S(\sigma) - \lambda kT \ln A^S = \beta \tilde{\mu}_S^X - \lambda kT \ln A^S \tag{19}$$

with

$$\tilde{\mu}_S^X = \int d\sigma\, p^X(\sigma)\, \tilde{\mu}_S(\sigma) \tag{20}$$

From the exact treatment of this ensemble we would find the factor λ to be the number of effective contact patches of the solute X, i.e., A^X/a_{eff}. But, as discussed ref 9, this is an artifact from considering each of these patches as an independent entity. Instead $\lambda = 1$ appears to be much more reasonable for coupled sets of patches. There are some additional influences of the molecular size of the solvent on the chemical potential of solutes known as combinatorial factors in the terminology of the chemical engineers,[10] which cannot be expected to be adequately represented within this simplified model ensemble of decoupled surface patches, but which should be roughly proportional to $kT \ln Y^s$, where Y^s is some size-dependent molecular descriptor like the surface area A^s. Thus, we consider the factor λ to be adjustable but general during the parametrization of COSMO-RS.

Apparently the standard chemical potential is hypothetical in almost any real case, since usually the real molar concentration x of the solute X in the solvent S is smaller than 1. Thus in order to get the real chemical potential of X in S, the standard chemical potential has to be corrected for the true concentration by addition of $-kT \ln x$.

With respect to the ideally screened state the standard chemical potential of a molecule in the gas phase at a partial pressure of 1 bar is given by

$$\mu'^X_{\text{gas}} = -\Delta'^X - \sum_k \gamma_k A^X_k - \omega n^X_{\text{ra}} - \eta RT \tag{21}$$

where the first term is the negative of the ideal screening energy. The second term, in which the index k refers to the different elements occurring in solute X, and with A^X_k being the exposed surface area of element k in molecule X, represents the dispersion or van der Waals energy gain of the solute going along the transfer from gas phase to a condensed phase. Although being assumed to arise mainly from dispersion, other free energy contributions, which are correlated with the molecular size and thus with surface area, may be involved in this term as well. Such contribution might be the solvent-induced change in vibrational free energy, which is not accounted for otherwise. The nature of the third term, in which n^X_{ra} is the number of ring atoms in molecule X, is not yet understood, but this ring correction is highly significant. It consistently removes the problems with ring compounds, as they have been reported by Marten et al.[7] The last term accounts for the entropy of the molecule in the gas phase and for the adjustment to the special reference state chosen for the gas phase.

In summary, eqs 19 and 21 allow for the general description of chemical equilibria between two liquid phases or between a liquid and the gas phase, without the need of any experimental data, neither for the solute nor for the solvent. Only a few adjustable parameters have to be determined for this really general task, which are element-specific γ_k and the four parameters β, λ, ω, and η explicitly appearing in these equations, as well as the parameters implicitly used in the COSMO-RS algorithm, i.e., the cavity radii, which we assume to be element specific, the optimal value for the averaging radius r_{av}, and the exact value of the polarizability factor f_{pol}.

3.3. Generalization for Hydrogen Bonding. Hydrogen bonds are important interactions in condensed media. To some degree hydrogen bonds are electrostatic interactions between the strongly positively polar hydrogens of the donor molecule and the strongly negatively polar parts, i.e., the lone pairs of the acceptor. This electrostatic part of hydrogen bonding is very well treated by the COSMO-RS algorithm as derived so far. But the extra energy gain that arises from the mutual penetration of the electron densities of the donor and acceptor is not caught by COSMO-RS so far. During the optimization procedure it turned out that it is necessary to account for this extra hydrogen-bonding contribution. In an approximate sense, this can quite elegantly be introduced in COSMO-RS by simply changing the interaction energy operator $E(\sigma,\sigma')$, which presently only covers the misfit energy. It is reasonable to assume that a hydrogen bond is formed between a sufficiently polar piece of surface of the donor and the acceptor, respectively, and that the bond is stronger the more polar these pieces are. Such behavior can be described by the following function:

$$E_{\text{hb}}(\sigma,\sigma') = c_{\text{hb}} \max[0, \sigma_{\text{acc}} - \sigma_{\text{hb}}] \min[0, \sigma_{\text{don}} + \sigma_{\text{hb}}] \tag{22}$$

Here σ_{acc} and σ_{don} denote the larger and smaller value of σ and σ', respectively. This energy is zero, unless both σ-values are of opposite sign and exceed the thresholds σ_{hb} and $-\sigma_{\text{hb}}$ for acceptors and donors, respectively. The introduction of this hydrogen bond term reduced the rms of the fit on C, H, O compounds by a factor 2.

3.4. Generalization to Additional Descriptors. In order to take advantage of the improvement in the description of the interaction energies of surface patches by additional descriptors such as the correlation screening charge density σ''_ν, we have to generalize the above presented one-dimensional COSMO-RS theory to a multidimensional theory, where the dimensionality is meant with respect to the number of descriptors per surface patch. Up to now we have considered only one descriptor for each surface patch, i.e., the averaged ideal screening charge density σ.

The straightforward generalization of the COSMO-RS algorithm to an arbitrary number n of descriptors considered to be represented as an n-dimensional vector d and used in the energy expression $\tilde{E}(d,d')$ would consist in the extension of the presently one-dimensional histograms with respect to σ, i.e., of the σ-profiles, to n-dimensional histograms and the corresponding replacement of all one-dimensional integrals to n-dimension integrals with respect to d. But, although still being managable, this would considerably increase the numerical expense for the iterative solution of the multidimensional equivalent of eq 18. It is more efficient to replace the n-dimensional integral by an appropriately weighted sum over all the segments of the molecules making up the solvent. Thus eq 18 becomes

$$\tilde{\mu}_{\text{S}}(d) = - \ln[W^{-1} \sum_i x_i \sum_{\nu \in X_i} s^i_\nu \exp\{-\beta^{-1}\tilde{E}(d,d^i_\nu) + \tilde{\mu}_{\text{S}}(d^i_\nu)\}] \tag{23}$$

with s^i_ν and d^i_ν being the area and the descriptors of segment ν of the ith molecular component of the solvent, respectively, x_i being the corresponding molarity and the normalization factor W_ν being defined as

$$W_{\text{S}} = \sum_i x_i \sum_{\nu \in X} s^i_\nu \tag{24}$$

Equation 23 first has to be iterated to self-consistence for the entire set of segments appearing in the solvent, and the resulting set of $\tilde{\mu}_{\text{S}}(d^i_\nu)$ has to be stored. Then the chemical potential for any surface patch with descriptors d as appearing in a solute can be calculated from eq 23 in a single step, and the equivalent of eq 20 is easily evaluated for any solute X:

$$\tilde{\mu}^X_{\text{S}} = \sum_{\nu \in X} s_\nu \, \tilde{\mu}_{\text{S}}(d\nu) \tag{25}$$

4. Data Set and Optimization Procedure

The full data set for the radii optimization and parametrization of COSMO-RS for the elements H, C, N, O, and Cl covers 217 molecules and altogether about 642 independent data points for the six properties ΔG_{hydr}, which is equivalent to Henry's law constant for the water/air system, vapor pressure, and the partition coefficients for octanol/water, benzene/water, hexane/water, and diethyl ether/water. The latter three properties, i.e., the partition coefficients hexane/water and benzene/water, and diethyl ether/water are less well investigated than the first properties, but altogether about 150 data points are experimentally available.

The compounds have been collected under the aspect of data availability, especially for ΔG_{hydr}, diversity of chemical functionality, size, and conformational simplicity. A limited molecular size is required, because each change in the cavity radii implies new DMol/COSMO calculations for each of the compounds and because a large number of different radii combinations in the five-dimensional space of the cavity radii of the elements had to be tested. Approximately 15 000 DMol/COSMO calculations have been run throughout the entire optimization procedure. Although multiple conformers and solvation-induced conformational changes are treatable by the COSMO-RS approach, conformational simplicity, i.e., the existence of a single dominant conformation, which is stable under solvation apart from minor changes in bond lengths and angles, considerably simplifies the parametrization process. The full data set is given as Supporting Information.

Experimental data were taken from different sources, the most important of which are the Thor database[28] for all kinds of partition coefficients, the CRC[29] and D'Ans-Lax[30] handbooks, and a collection of Henry's law constants by Meylan,[31] which to a large degree are calculated from the ratio of vapor pressure and water solubility. In some cases the mean value of different references has been used.

In order to use eqs 19 and 21, we had to convert most of the experimental data to the appropriate reference systems, that is, ΔG_{hydr} had to be corrected by 4.28 kcal/mol, which is RT ln(number of moles of water in the standard gas-phase molar volume), for a conversion from 1 bar and 1 mol/mol as reference state in the gas phase and liquid phase, respectively, to the common reference states of 1 mol/L in both phases. Distribution coefficients, which usually are considered as ratios of concentrations in units of mol/L, had to be converted to ratios of concentrations in mol/mol, i.e., by multiplication with $\mathrm{MW}_1\,\mathrm{D}_2/\mathrm{MW}_2\,\mathrm{D}_1$, where the MW_i and D_i denote the molecular weights and the densities of the involved solvents.

In order to keep the dimensionality of the optimization problem small and to reduce the expense for a single-radius point, i.e., a single combination of radii, most of the parametrization was done on a reduced data set of H, C, O compounds (Table 2, Supporting Information). In addition the geometries of the ideally screened compounds have been updated only a few times, while for most radii points only single-point calculations have been performed using geometries from nearby radii points, since the effect of small radii changes on the geometries turned out to be negligible. The radii for the other elements N and Cl were optimized after the radii for H, C, and O had been fixed. The other COSMO-RS parameters were finally readjusted based on the entire data set.

Before we present the results of the optimization procedure in the next section, we would like to discuss some general aspects we became aware of during the optimization and some wrong tracks we have followed.

A major part of the time during the optimization process we spent with the analysis and the proper correction of the outlying charge error until we finally found the rigorous correction algorithm described in ref 16. The proper correction of the outlying charge error is of crucial importance for any radii optimization in CSMs. The error, which is up to 25% of the whole solvation energy, exponentially decays just in the relevant radii region, and hence it influences the final radii optimum. Since the outlying charge error is very sensitive to the basis set,[32] such optimum is not transferable between different basis sets.

As proposed in the original COSMO-RS paper, local polarizability, i.e. the linear polarization answer of the perfectly screened solute to a local misfit charge, has been considered as an additional local descriptor for each patch, in order to replace the global and general polarization factor f_{pol} in the misfit energy expression. The gain in accuracy by using local polarizability as a second descriptor was surprisingly small, and it did not justify the large additional numerical expense necessary to calculate the local polarizabilities. A second approach using element-specific polarizabilities instead of the general one did not yield a significant improvement of the fit either.

The introduction of the correlation screening charge density $\sigma_v{}^{\perp}$ into the misfit energy expression significantly improved the standard deviation of the fit by about 5%. For this eq 7 was replaced by

$$E_{\mathrm{misfit}}((\sigma,\sigma^{\perp}),(\sigma',\sigma^{\perp\prime})) = \frac{\alpha'}{2}(\sigma + \sigma^{\perp})[(\sigma + \sigma') + f_{\mathrm{corr}}(\sigma^{\perp} + \sigma^{\perp\prime})] \quad (26)$$

Taking the value of f_{corr} from the regression with repect to the dielectric energy, this methodological improvement does not introduce another adjustable parameter.

In order to remove certain systematic deviations occurring for alcohols and ethers on the one side and carbonyls on the other side, we intermediately introduced atom type specific radii, i.e., different radii for polar and nonpolar hydrogens as well as for sp^2- and sp^3-oxygens. For a while this appeared to yield significant gains in accuracy, especially the differentiation between two types of hydrogen. Fortunately, in the end we found a comparably accurate parameter set with only one radius for each element. Apart from the general advantage of having less parameters, this is of special importance for the applicability of the approach to less common situations, in which the classification of an atom may be less obvious, and to reactions, during which the atom type may change.

A hydrogen bond term in the energy expression as given by eq 22 turned out to be highly significant for the C, H, O data set. The best fit without such a term had about twice the standard deviation of our final optimum, i.e., 0.8 kcal/mol instead of 0.4 kcal/mol. It is notable that Marten et al. report almost the same decrease from 0.8 kcal/mol to 0.4 kcal/mol standard deviation by addition of first-shell hydrogen-bonding corrections to their SCRF-GVB[8] method. It turned out that the details of the functional form of the hydrogen bond term in COSMO-RS are less important, as long as it exhibits the characteristic behavior of hydrogen bonding, i.e., being almost zero for nonpolar or moderately polar interactions, but becoming important for strongly polar surface contacts. Different reasonable functional forms led to almost identical fit results. We finally took the simplest of these approaches. Attempts to introduce an upper bound for the hydrogen bond interaction did not yield a better fit. Unfortunately this term failed to adequately describe the acceptor behavior of nitrogen in neutral amines, especially if these are multiply substituted with methyl, ethyl, or even more bulky groups. Problems with a correct description of amines in CSMs are well-known.[8,33] A good collection and comparison of the results of different methods on amines is given by Marten et al.[8] In our opinion the problems of COSMO-RS and other CSMs with neutral amines in protic solvents result from the fact that for these amines the nitrogen lone pair is to a large degree hidden by the substituents, if one considers the vdW surface or the COSMO cavity. Only if a donor comes closer to the nitrogen does the full attractive potential of the lone pair become perceptible. Therefore the hydrogen bond approximation of eq 22, which is an estimate based on the screening charge density, i.e., on the polarity visable on the COSMO surface, fails in this special situation. For this reason we dropped all data points involving amines in water from the data set and took only the vapor pressure into account. The latter is well described, since hydrogen bonding does not play an important role in the pure amines due to the poor donor behavior of amine hydrogens. It should be noted that the hydrogen bond correction works well for all other nitrogen compounds.

Out of the six equilibrium constants chosen as representative goal properties in our parametrization, five of them involve water as the solvent. On the one hand this represents well the overwhelming importance of water as a solvent, but on the other hand the dominance of water in the goal properties may cause some bias of the parameter set toward an optimal description of the solvent water, which due to its extraordinarily strong interactions and high degree of internal structure, is a rather unusual fluid. In order to remove this bias, we temporarily allowed for a special description of the solvent water by an mth order Taylor series approach for the σ-potential of water, i.e.,

$$\mu'_{\text{water}}(\sigma) = \sum_{i=0}^{m} \mu^{i}_{\text{water}} \sigma^{i} \tag{27}$$

instead of using eq 17. In this case eq 19 simplifies to

$$\mu^{*X}_{\text{water}} = \sum_{i=0}^{m} \mu^{i}_{\text{water}}\, M_i^X - \lambda kT \ln A^{\text{S}} \tag{28}$$

with M_i^X being the ith σ-moment of solute X, i.e.,

$$M_i^X = \int \mathrm{d}\sigma\, p^X(\sigma)\sigma^i \tag{29}$$

As discussed in the original COSMO-RS paper, M_0^X is nothing else than the molecular surface area A^X, M_1^X is the negative total solute charge and hence zero throughout our parametrization, since only neutral species were considered, and M_2^X is highly correlated with the screening energy Δ^X. Due to the disappearance of the first moment, a Taylor series up to fourth order corresponds to four additional adjustable parameters in our model, which can easily be determined in the multilinear regression part of the fit, if the σ-moments of the solutes are supplied as descriptors. It turned out that these additional four parameters did not significantly influence the other parameters, and the gain in accuracy was less than 3% with respect to the standard deviation. We take this as proof of the robustness of the COSMO-RS theory and of the final parameter set. Obviously, the small improvement of the fit is not sufficient to permanently keep this exception rule for the solvent water.

Unfortunately we had to remove all data points where water acts as the solute because ΔG_{hydr} of water is calculated to be 2.3 kcal/mol too low. The reason for this error probably arises from the fact that, due to the neglect of all steric restraints in the COSMO-RS approach, there is no problem for a water molecule to form four hydrogen bonds, while in reality the formation of four hydrogen bonds per molecule implies a high degree of order, which goes along with crystallization. In liquid water on average only a smaller number of hydrogen bonds can be formed.

5. Results and Discussion

5.1. Results for Parameters. The optimization of the cavity radii and the other model parameters led to the following results: The radii are 1.30 Å for H, 2.00 Å for C, 1.72 Å for O, 1.83 Å for N, and 2.05 Å for Cl. Except for hydrogen, these radii are 13−18% larger than the corresponding van der Waals radii and thus agree reasonably with the widely accepted “van der Waals plus 20% rule” for dielectric CSMs.

Let us now consider the parameters needed for the free energy of transfer from gas phase to the ideally screened condensed phase, as expressed by eq 21. The dispersion constants come out as $\gamma_{\text{H}} = -0.041$, $\gamma_{\text{C}} = -0.037$, $\gamma_{\text{O}} = -0.042$, $\gamma_{\text{N}} = -0.027$, and $\gamma_{\text{Cl}} = -0.052$ (in kcal/(mol Å^2)). These values correspond to about −1.8 kcal/mol for a water molecule and about −5 kcal/mol for octane. The dispersion parameters for H, C, and O are quite close to each other, which initially drove us to the assumption that a single universal dispersion constant would be sufficient. But for nitrogen and chlorine the need for element-specific dispersion constants became obvious.

The exact value of the ring correction coefficient ω is −0.21 kcal/mol. For a six-membered ring this corresponds to −1.26 kcal/mol and it reflects the difference in ΔG_{hydr} between hexane and cyclohexane. As mentioned in section 4 the physical origin of this contribution still is an open question. From application to larger ring systems with up to 16 ring atoms we found that it works well even for rings of such size.

The constant η, which corresponds to the entropy difference of a molecule between the standard state in gas phase (1 bar) and in the liquid state (1 mol/mol) comes out to be −9.15, i.e., ηkT is −5.4 kcal/mol at room temperature.

The best value for the averaging radius r_{av} turns out to be 0.5 Å. This is considerably less than the initially assumed value of about 1 Å, which had been derived from the consideration of the correlation length of the screening charge density on the cavity surfaces. But since we have introduced the correlation correction (see eq 15), we should not be surprised that the optimal value now is smaller than the correlation length.

The optimal value of the the scaling parameter β for the chemical potentials in eq 19 is $\beta = kT/a_{\text{eff}} = 0.0832$ kcal/(mol Å^2). With $kT = 0.592$ kcal/mol at room temperature we thus have $a_{\text{eff}} = 7.1$ Å^2, i.e., a radius $r_{\text{eff}} = 1.5$ Å. Interpreting a_{eff} as the average statistically independent surface unit, we get about 7 of such units on a water molecule, corresponding quite well with standard estimates of the number of nearest neighbor molecules in liquid water. Thus we obtained a plausible result for a_{eff}, although it has been treated as an adjustable parameter during the optimization.

For the misfit energy parameter α' we find $\alpha' = \alpha f_{\text{pol}} = 1288$ kcal/(mol Å^2)/e^2. The correlation correction factor is $f_{\text{corr}} = 2.4$. It should be pointed out that the latter has been derived from a fit to model inherent data and thus is not a free parameter of the model. Comparing α' with the slope of −1360 kcal/(mol Å^2)/e^2 from the correlation of the dielectric energy with the second σ-moments, which according to eq 13 should be $-\alpha/2$, we find $f_{\text{pol}} = 0.48$, i.e., somewhat smaller but still in reasonable agreement with our previous estimates of $f_{\text{pol}} = 0.64$ and $f_{\text{pol}} = 0.6$ (cf. section 3.1), respectively.

The hydrogen bond parameters introduced in eq 21 are c_{hb} = 7400 kcal/(mol Å^2)/e^2 and σ_{hb} = 0.0082 e/Å^2. In order to check the reliability of this extra hydrogen bond term, we performed DMol/COSMO calculations for the water dimer. The total energy gain from the formation of an H-bonded dimer is −3.84 kcal/mol, while only −1.74 kcal/mol is gained if the bond distance is fixed at a vdW distance instead of the optimized distance of 1.7 Å. In the sense of our COSMO-RS treatment we should interpret the difference of 2.1 kcal/mol as the extra hydrogen bond energy. Using eq 21 together with the optimized parameters, we find a value of about 1.9 kcal/mol for the extra energy of one H-bond of water in the solvent water, which is in good agreement with the directly calculated value.

Finally, the best value of the parameter λ is 0.14. As mentioned above, we would have expected a value of 1 from the consideration that each solute may choose just one partner surface patch independently, while the choice of the rest is considerably constrained by neighborhood relations of the solvent patches. On the other hand, the number of constraints is smaller in solvents composed of smaller molecules. Therefore the degeneracy of a solute in such solvents is larger, causing an opposite trend. Our result of λ = 0.14 implies that both tendencies almost cancel. Within the accuracy of the method we could as well set λ to zero and hence drop this parameter. But in order to state that the question of degeneracy, which is known as the combinatorial factor in activity coefficients in the nomenclature of chemical engineers,[8] has been considered, we keep it in the formalism. Thus our result corresponds to a combinatorial factor of

$$\gamma_{comb} = (A^S/A^X)^{0.14} \qquad (30)$$

where A^S is the mean surface area of all the components of the solvent. This expression is relatively simple compared to the heuristic expressions for γ_{comb} which are routinely used by chemical engineers. Thus it might be that some improvement of the COSMO-RS approach will be achieved by a more sophisticated combinatorial factor. Some indication for the need for further improvements at this point might be the constant correction of −0.5 kcal/mol, which we needed to avoid an average overestimation of the diethyl ether/water partition coefficients. Nevertheless all other solvents, including the 171 solvents considered in the vapor pressure data by the calculation of the chemical potential of the molecules in their own fluid, are well described by our degeneracy term.

5.2. Results for Goal Properties. The finally achieved agreement between calculated and experimental data for the six goal properties is presented in Figure 4a−f. The calculated residuals are plotted against the experimental values. The corresponding data (about 1300 experimental and calculated values) are given in Table 2, which has been deposited as Supporting Information, in order to keep this article reasonably comprehensive. The overall standard deviation for chemical potential differences is 0.40 kcal/mol, corresponding to 0.3 log units or a factor 2 for the corresponding equilibrium constants. The error is rather homogeneous with respect to the different goal properties, with a slight increase for the three less well represented properties, i.e., the hexane/water, benzene/water, and diethyl ether/water partition coefficients. For these we can safely assume a larger experimental error, because due to the small number of data points, we had to accept almost any value documented in the Thor database[28] without being able to check their reliability.

The standard deviation achieved for the 163 values for ΔG_{hydr}, which cover a range of 14 kcal/mol, corresponding to 10 log units for the Henry coefficient, is 0.37 kcal/mol. Apart from the error of −2.3 kcal/mol for water, which has been discussed before, all errors are within 1 kcal/mol. The largest negative error is −0.83 for H_2, while the largest positive errors (1 kcal/mol) occur for dimethylpyridine and methylpyrazine. As expected, ΔG_{hydr} of cyclohexanone is significantly overestimated (−0.76 kcal/mol), consistent with the too large dipole moment calculated by DMol (vide infra).

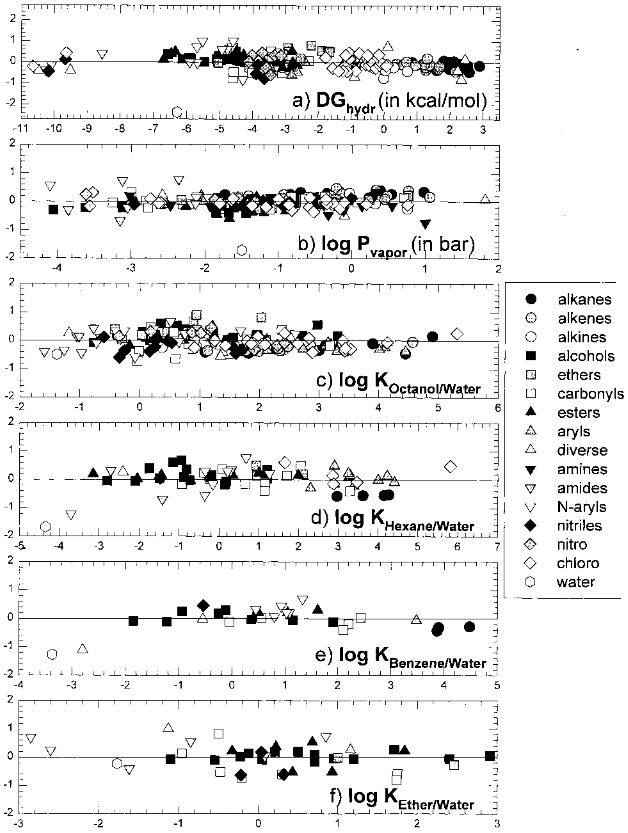

Figure 4. Residuals of the six fitted goal properties vs experimental data: different classes of compounds are marked by different symbols as given in the legend. The detailed data are presented in Table 2.

The 171 vapor pressures, covering about 6 log units, are best reproduced by the COSMO-RS results. The standard deviation is 0.32 kcal/mol. Again water is the largest outlier (−1.7 log units), while the others stay within an error of 1 kcal/mol (0.75 log units). Among these, NH_3 has the largest negative deviation, while two amides are the largest positive outliers.

The 170 data points for the octanol/water partition coefficient are spread over 7 log units. The standard deviation of the residuals is 0.41 kcal/mol. All errors are within 0.8 log units (1.1 kcal/mol), with two ethers, i.e. dipropyl ether and methyl *tert*-butyl ether, being the largest positive outliers, while acetaldehyde is the largest negative one. It is remarkable that the error for water is only −0.5 log units. This may have to do with the fact that the octanol phase offers hydrogen bond donors and acceptors as well. Thus there should be some cancellation of errors between the two phases.

For the partition coefficient betwen hexane and water we could collect 68 data points. It should be noted that in order to increase the data basis we used data for pentane, cyclohexane, heptane, and octane as well, since the experimental partition coefficients for these solvent systems turned out to be identical within about 0.2 log units, i.e., within the experimental error. The standard error achieved is 0.48 kcal/mol. Here again, water is the largest outlier. The error is −1.7 log units, as for the vapor pressure (and for ΔG_{hydr}, respectively). This clearly demonstrates that the error is caused by hydrogen bonding, while the other contributions, i.e., electrostatics and dispersion, which are present in hexane as well, appear to be well described. The second largest outlier is imidazole, with -1.2 log units. Since there is only a single experimental value for imidazole in these alkanes, we tend to ascribe this outlier to an experimental error. All other results are within 0.7 log units. It should be noted that the data for this partition coefficient covers more than 10 log units.

Only 30 data points were available for the benzene/water partition coefficient, for which we achieve a standard deviation of 0.45 kcal/mol. Again water is the largest outlier (-1.3 log units), followed by hydrogen peroxide, with -1.1 log units. The reason for the latter deviation should be similar to that of water. All other points come out quite well, with the largest positive outlier being dimethylpyridine, with an error of 0.7 log units. The range of the benzene/water data covers 8 log units.

For the diethyl ether/water partition coefficient we found 40 data points for our set of compounds. With a standard deviation of 0.6 kcal/mol this is the worst reproduced goal property, especially if we take into account that in contrast to all other properties we added an additional regression constant in this case, which came out as -0.5 kcal/mol. The error for water is surprisingly small again (-0.23 log units). This indicates that there is error cancellation between the solvents diethyl ether and water with respect to the solute water, although diethyl ether has no hydrogen bond donors. Hence we may conclude that the problems with the solute water can be ascribed to its donor behavior, which is overestimated to some degree. The largest outlier is hydrogen peroxide, this time with a positive deviation of 1 log unit. This does not seem very plausible, considering the negative deviations for water and hydrogen peroxide in all other partition coefficients. Thus we tend to assume an experimental error in this case. The other major deviations are random. Considering the fact that most of the experimantal data points are single values, some part of the scatter for this property may arise from experimental uncertainties.

6. Summary and Outlook

By careful parametrization of the COSMO-RS theory, which takes the ideally screened states of molecules as a starting point for subsequent solvation calculations, we have achieved a model that allows for the calculation of the chemical potential of almost any neutral solute X. This can be done in almost any organic solvent without using any experimental data for the solute or the solvent. An accuracy of about 0.4 kcal/mol can be achieved if the underlying COSMO calculations for the ideally screened states are performed using DFT methods.

Only eight general parameters are used. These are an averaging radius r_{av} for the screening charge density, an effective contact area a_{eff}, the electrostatic interaction coefficient α', two hydrogen-bonding parameters, a ring correction, a degeneracy difference between gas phase and liquid state, and a size dependence coefficient, as well as two parameters per element, i.e., the cavity radius and the dispersion coefficient. So far the elements H, C, O, N, and Cl have been considered. Additional common elements like F, Br, and I, as well as S and P will be parametrized soon. The renunciation of atom type specific parameters makes the presented method generally applicable.

Apart from the removal of the questionable dielectric approximation for solvents on a molecular scale, the special advantage of the COSMO-RS approach compared to other continuum solvation methods is its ability to treat the solvent on the same footing as the solute. Almost any solvent, even mixtures, can be handled, and the temperature dependence is include in a natural way. This allows a wide range of applications, especially in the area of chemical engineering.

Because the scope of this paper was the parametrization of the model, we concentrated on a core region out of the much broader range of applications of COSMO-RS. We only considered a suite of well-investigated room-temperature equilibrium parameters between more or less pure solvents for conformational simple, neutral solutes. Thus several additional aspects will be subjects of forthcomming papers; these are the application to mixtures and to varying temperatures, the study of ionic solutes, the treatment of multiple conformations, and the consideration of properties that are not directly related to chemical potentials, such as surface tensions or heats of transfer.

The achieved accuracy of 0.4 kcal/mol is satisfying. For many properties, this accuracy, which corresponds to deviations of a factor 2 in the equilibrium constants, is almost within experimental error. Considering the fact that the inaccuracy in the quantum chemical calculation of the electrostatics of the solute, which we found to be about 0.1 D for dipole moments, causes errors of this magnitude in the ideal screening energy, no dramatic increase of the accuracy of COSMO-RS can be expected. Nevertheless, some room is left for further improvements, especially in the heuristic hydrogen bond term, which presently is an estimate of the gain of hydrogen bond energy based on the screening charge density on the COSMO surface. This should benefit from introduction of an auxiliary screening charge density, which is evaluated on a surface about 0.5 Å closer to the atoms, i.e., at a distance much more characteristic for the hydrogen bond contacts. Due to the generalization of COSMO-RS to multiple descriptors, the inclusion of a third descriptor into the algorithm is straightforward. We hope to overcome the problems with amines by such modifications.

Appendix 1: Saturation of Reorientational Polarizability

The saturation of the reorientational polarizability at typical electric field strengths on the molecular surfaces of polar solutes can be easily proven by the following consideration for water as the solute. The static dipole moment of water is $\mu = 1.9$ D $= 0.4$ e Å. The average radius of a water molecule, as derived from its molecular volume of 30 Å^3, is $R = 1.9$ Å. Thus, for the electric field in the direction of the dipole moment we find a value of $E = 2\mu/R^3 = 0.12$ e/Å^2. The polarization of a dielectric medium necessary for an almost perfect compensation of this field is given by $P = E/(4\pi) = 0.010$ e/Å^2. On the other hand, the maximum polarization of a medium by perfectly ordering all of its permanent dipole moments is given by $P_{\max} = \mu'/V$, where μ' is the strength of a permanent dipole and V is the molecular volume. For water we find $P_{\max} = 0.013$ e/Å^2. Thus the solvent water could be able to screen the electric field of a water molecule almost perfectly by reorientational polarizability, but only if this is ordered up to 80% saturation. It is unlikely that the reorientational polarizability really behaves linearly up to this degree of saturation. Electric fields stronger than 0.16 e/Å?, as they occur on molecular surfaces of small ions, can definitely no longer be efficiently screened by reorientational polarizability. On the other hand, other relatively strong dielectrics such as acetone or methanol, which according to the dielectric theory should be able to screen 89% and 92%, respectively, of the electric field of a dipolar solute by their reorientational polarizabilities, have a lower value of $P_{\max}$ of 0.005 e/Å^2. Thus, even with perfect ordering of the static dipole moments they are only able to screen 50% of the maximum

TABLE 3: Element-Specific Parameters

element k	cavity radius R_k [Å]	dispersion constant γ_k [kcal/(mol Å^2)]
H	1.30	−0.041
C	2.00	−0.037
N	1.83	−0.027
O	1.72	−0.042
Cl	2.05	−0.052

TABLE 4: General COSMO-RS Parameters

symbol	value
r_{av}	0.5 Å
a′	1288 kcal/(mol Å^2)/e^2
f_{corr}	2.4
c_{hb}	7400 kcal/(mol Å^2)/e^2
σ_{hb}	0.0082 e/Å^2
a_{eff}	7.1 Å^2
λ	0.14
ω	−0.21 kcal/mol
η	−9.15

field of a water solute by reorientation. Hence, they are definitely not able to compensate the electric field of a water molecule by static polarizability as efficiently as they compensate macroscopic electric fields.

Appendix 2: Recipe for a COSMO-RS Calculation

In the following we give a recipe for the calculation of the chemical potential μ_S^{*X} of a solute X in a solvent S, which is composed of a set of n components X_i, $i = 1, ..., n$.

(1.1) This step is necessary only if any gas-phase-related property of X is desired: Do a DMol gas-phase geometry optimization using BPW91/DNP for X.

(1.2) Do a DMol/COSMO geometry optimization using BPW91/DNP and $\epsilon = \infty$ (ideal screening) for X. The cavity radii are given in Table 3. In addition use the parameters $R_{solv} = R_H$, NSPA = 92, and DISEX = 10. Trigger the outlying cavity correction and use the corrected results for energies and screening charge densities.

(1.3) Repeat 1.2 for each of the solvent molecules X_i.

(2.1) For each of the molecules X and X_i do the averaging of the screening charge densities according to eq 11. Use $r_{av} = 0.5$ Å and $r_{av} = 1$ Å to get the σ_ν and σ_ν°, respectively. Calculate $\sigma_\nu{}^\perp$ as $\sigma_\nu{}^\perp = \sigma_\nu^\circ - 0.816\sigma_\nu$.

(2.2) Only for X and only if gas-phase calculation is done, calculate the averaging corrected energy difference of gas phase and ideally screened state Δ'^X according to eq 12. Calculate the chemical potential of X in the gas phase according to eq 21, using the values for the element-specific dispersion paramters γ_k as given in Table 3, as well as $\omega = -0.21$ kcal/mol and $\eta = -9.15$.

(3.1) For the entire set of segments ν occurring in the the set of solvent components X_i iterate the $\tilde{\mu}_S(d_\nu^i)$ to self-consistency using eqs 23 and 24, starting with $\tilde{\mu}_S(d_\nu^i) = 0$ on the right side of eq 23. Use 0.001 kcal/mol as a convergence criterion. The set of descriptors d_ν for each segment is given by the two screening charge densities σ_ν and $\sigma_\nu{}^\perp$. The energy functional $\tilde{E}(d,d')$ is composed of two contributions E_{misfit} (eq 26) with α′ = 1288 kcal/mol Å^2/e^2 and $f_{corr} = 2.4$ for the electrostatic misfit and E_{hb} (eq 22) with $c_{hb} = 7400$ kcal/mol Å^2/e^2 and $\sigma_{hb} = 0.0082$ e/Å^2 for hydrogen bonding. The scaling parameter for $\tilde{E}(d,d')$ is $\beta = kT/a_{eff} = 0.0832$ kcal/(mol Å^2).

(3.2) Now the $\tilde{\mu}_S(d_\nu)$ are calculated for all segments ν of the solute X, using the $\tilde{\mu}_S(d_\nu^i)$ from 3.1) on the right side of eq 23.

(3.3) The chemical potentialμ_S^{*X} of the compound X in the solvent S can be calculated as

$$\mu_S^{*X} = \beta \sum_{\nu \in X} s_\nu \tilde{\mu}_S(d_\nu) - \lambda kT \ln \sum_i x_i A^{X_i}$$

with $\lambda = 0.14$.

Supporting Information Available: Table of experimental and calculated data for the six goal properties and 217 compounds (3 pages). Ordering information is given on any current masthead page.

References and Notes

(1) Cramer, C. J.; Truhlar, D. G. *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1995; Vol. 6.

(2) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(3) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(4) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* *1993*, 799.

(5) Giesen, D. J.; Gu, M. Z.; Cramer, C. J.; Truhlar, D. G. *J. Org. Chem.* **1996**, *61*, 8720.

(6) King, G.; Warshel, A. *J. Chem. Phys.* **1990**, *93*, 8682.

(7) Åquist, J.; Hansson, T. *J. Phys. Chem.* **1996**, *100*, 9512.

(8) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, N.; Sittkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775.

(9) Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224.

(10) Fredenslund, A.; Gmehling, J.; Rasmussen, P. *Vapor Liquid Equilibria Using UNIFAC*; Elsevier: Amsterdam, 1977.

(11) Delley, B. *J. Chem. Phys.* **1990**, *92*, 508; **1991**, *94*, 7245.

(12) DMol, version 950, Biosym Technologies, San Diego, CA, 1995.

(13) Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*,

(14) Stewart, J. J. P. MOPAC program package, QCPE-No. 455, 1993.

(15) Andzelm, J.; Kölmel, C.; Klamt, A. *J. Chem. Phys.* **1995**, *103*, 9312.

(16) Klamt, A.; Jonas, V. *J. Chem. Phys.* **1996**, *92*, 9972.

(17) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(18) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

(19) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(20) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

(21) We employed the local VWN correlation functional together with Becke's gradient corrected exchange functional[18] and Perdew and Wangs correlation functional.[19]

(22) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A.; *Gaussian 94*, Revision D.4; Gaussian, Inc.: Pittsburgh, PA, 1995.

(23) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822; *Phys. Rev. B* **1986**, *34*, 7406.

(24) The SVP basis set was taken from Turbomole5.0,[25] where it is the default basis set for ri-DFT calculations. For first-row elements the splitting is (511/31/1).

(25) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(26) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(27) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(28) Thor Database, Daylight Chemical Information Systems: Irvine, CA, 1991.

(29) Lide, D. R. *Handbook of Chemistry and Physics*; CRC Press: Boca Raton, 1994.

(30) Lax, E. *Taschenbuch fuer Chemiker und Physiker, Band I*; Springer: Berlin, 1967.

(31) Meylan, W. Private communication, 1997.

(32) Baldridge, K.; Klamt, A. *J. Chem. Phys.* **1997**, *106*, 6622.

(33) Morgantini, P.-Y.; Kollman, P. J. *J. Am. Chem. Soc.* **1995**, *117*, 6057.