Sigma Profile Database for Predicting Solid Solubility in Pure and Mixed Solvent Mixtures for Organic Pharmacological Compounds with COSMO-Based Thermodynamic Methods

Eric Mullins,[†] Y. A. Liu,* Adel Ghaderi,[‡] and Stephen D. Fast[§]

SINOPEC/FPCC/AspenTech Center of Excellence in Process System Engineering, Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Thermodynamic methods based on COSMO (COnductor-like Screening MOdels), such as COSMO-RS (Real Solvent) and COSMO-SAC (Segment Activity Coefficient), represent significant and recent developments of solvation thermodynamics and computational quantum mechanics. These are a priori prediction methods based on molecular structures and a few parameters that are fixed for all of the compounds. They require no experimental data and rely on sigma profiles specific to each molecule as their only input. A sigma profile is the probability distribution of a molecular surface segment having a specific charge density. Generating sigma profiles by quantum mechanical calculations represents the most time-consuming and computationally expensive aspect of using COSMO-based methods. This article presents a free, web-based VT-2006 Solute Sigma Profile Database for large, pharmaceutical-related solutes, to supplement the published VT-2005 Sigma Profile Database for solvents and small molecules (www.design.che.vt.edu). Together, these databases contain sigma profiles for 1645 unique compounds, enabling the users to predict binary and multicomponent vapor−liquid equilibrium (VLE) and solid−liquid equilibrium (SLE), as well as other thermodynamic properties. We validate the VT-2006 Solute Sigma Profile Database by solid solubility predictions in pure solvents for 2434 literature solubility values, which include 194 solutes, 160 solvents, and 1356 solute−solvent pairs. We also compare solubility predictions for mixed solvents to literature values for 39 systems. By comparison with experimental data, we find a root-mean-squared error (RMSE) of 0.7419 between experimental and predicted solute mole fractions (x_{sol}) on a $\log_{10}$ (x_{sol}) scale for solubilities in pure solvents. This article also presents examples investigating the effects of conformational isomerism on solubility predictions of small, medium-sized, and large drug molecules. To provide better understanding of accuracy, we compare a priori COSMO-SAC solubility predictions, which use molecule-specific sigma profiles, to those by the non-random two-liquid segment activity coefficient (NRTL-SAC) model, which uses regressed molecule-specific parameters, for 17 solutes and 258 experimental solubility values. We find that NRTL-SAC, which contains regressed parameters based on experimental data, is a more accurate method for predicting SLE behavior than the COSMO-SAC model for many of the systems studied. Finally, this article presents a set of guidelines for applying the COSMO-SAC model for solubility predictions for new drug molecules when no experimental data are available.

1. Introduction

Predictive thermodynamic models are in high demand in the current engineering practice. Their need is significant enough that engineers are willing to accept inaccuracies for their promised benefits such as time and cost savings. For example, in pharmaceutical drug development and manufacturing, we use solvents as a medium for the synthesis reaction or to separate and purify the desired components from unwanted byproducts. Even after the drug is synthesized, researchers must spend considerable amounts of time to identify suitable solvents and to scale up from a laboratory bench-scale to high-volume manufacturing. Reducing both time and cost with a predictive property method could significantly enhance the success of developing and manufacturing a new drug.

Solvent selection is just one potential application for predictive models. There are many instances in which researchers need phase-equilibrium data, and the quickest way to obtain new values is through a predictive thermodynamic model. Currently, group-contribution methods like UNIFAC and activity-coefficient models like NRTL and NTRL-SAC also perform these functions. NTRL-SAC, developed by Chen and Song from the polymer-NRTL model use regressed molecule-specific parameters to predict SLE behavior. However, these methods require regressed parameters from experimental data, and therefore have little or no applicability to compounds with new functional groups (in the case of UNIFAC) or new compounds (in the case of NRTL) without substantial experimentation. Solvation thermodynamics is another approach to characterize molecular interactions and to account for liquid-phase nonidealities. COSMO-based thermodynamic models such as COSMO-RS by Klamt and his colleagues[1−4] and COSMO-SAC by Lin and Sandler[5,6] are two a priori models, which predict intermolecular interactions based on molecular structure and a few adjustable parameters. COSMO-RS is the first extension of a dielectric continuum−solvation model to liquid−phase thermodynamics, and COSMO-SAC is a variation of COSMO-RS. COSMO-

* To whom correspondence should be addressed. Tel.: (540) 231-7800; Fax: (540) 231-5022; E-mail: design@vt.edu.

[†] Present address: Eastman Chemical Company, P.O. Box 511, Kingsport, TN 37662.

[‡] Present address: Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139.

[§] Present address: ExxonMobil Research and Engineering Co., 2800 Decker Drive MOB-621, Baytown, TX 77522.

10.1021/ie0711022 CCC: $40.75
Published on Web 02/01/2008

based models predict liquid-phase activity coefficients, which we use in this work for SLE calculations.

Both COSMO-based models require input in the form of a sigma profile, a molecule-specific distribution of the surface-charge density. This is similar to the way UNIFAC requires parameter databases with one exception. Sigma profiles for COSMO-based methods are molecule-specific, whereas UNIFAC binary interaction parameters are specific to each functional group. We generate sigma profiles from a single structure by performing quantum-mechanical calculations. These calculations represent the most time-consuming task involved in using COSMO-based methods, with over 90% of the computational effort devoted to this step. Until recently, researchers were hindered by the lack of an open-literature sigma-profile database. Hopefully, our published sigma profiles[7] will help in improving the situation.

We present an open-literature database, the VT-2006 Solute Sigma Profile Database, which we use in conjunction with the VT-2005 Sigma Profile Database.[7,8] Our work contains sigma profiles for 1645 unique compounds, composed of the following 10 atoms only: hydrogen, carbon, nitrogen, oxygen, fluorine, phosphorus, sulfur, chlorine, bromine, and iodine. The VT-2005 Sigma Profile Database includes many common solvents and is the larger of the two databases. The VT-2006 Solute Sigma Profile Database includes 206 solutes and 32 solvents and focuses primarily on larger pharmacological compounds and their derivatives. Both databases are available free of charge from our website (www.design.che.vt.edu). We continue to update the sigma profiles as new results become available. We refer to specific compounds from both databases throughout this work with the following nomenclature: VT-(Index No.) refers to a compound from the VT-2005 Sigma Profile Database[7,8] with a four-digit index number, and VTSOL-(Index No.) refers to a compound from the VT-2006 Solute Sigma Profile Database with a three-digit index number.

We validate the VT-Solute Sigma Profile Database by comparing our predictions of solid solubility in pure and mixed solvents to experimental data from the literature. We discuss several factors that affect the accuracy of COSMO-SAC solubility predictions, such as conformational isomerism, and sensitivity to melting-point temperature and latent heat of fusion. Through our validation effort and literature review, we present a set of guidelines for applying the COSMO-SAC method. Our current work focuses on solid solubility in pure and mixed solvents. Previous work[6−14] has already demonstrated the applicability of COSMO-based methods to predict pure-component vapor pressure, binary and multicomponent VLE behavior, and other physical properties.

2. Theory

This section summarizes the COSMO-SAC model, focusing on how to calculate a sigma profile and on the governing equations for SLE. This summary is necessary as we shall compare the accuracy of using different exchange-energy expressions required to calculate the sigma profile.

2.1. Summary of the COSMO-SAC Model. The basic principle behind COSMO-based thermodynamic models is the solvent-accessible surface of a solute molecule.[2,15] Conceptually, COSMO-based models create a cavity with the exact size of a molecule within a homogeneous medium, or solvent, of a dielectric constant ϵ and then place the molecule inside the cavity. Figure 1 illustrates the ideal solvation process with COSMO-based methods.

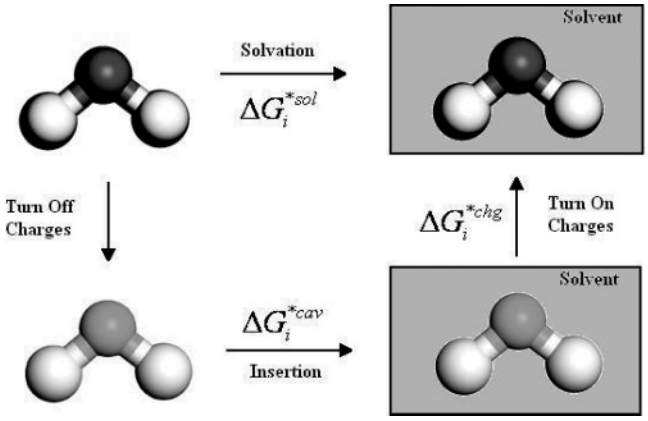

Figure 1. Conceptual diagram of solvation process with a COSMO-based model.

In the figure, the solvation free energy $\Delta G_i^{*\,sol}$ represents the change in Gibbs free energy associated with moving a molecule i from a fixed position in an ideal gas to a fixed position in a solution S. The cavity-formation free energy $\Delta G_i^{*\,cav}$ represents the change in Gibbs free energy required to form a cavity within a solution S of the exact size of the molecule i. The charging free energy $\Delta G_i^{*\,chg}$ represents the Gibbs free energy required to remove the screening charges from the surface of the molecular cavity. We determine the solvation free energy from the sum of the cavity-formation free energy and the charging free energy in eq 1.

$$\Delta G_i^{sol} = \Delta G_i^{cav} + \Delta G_i^{chg} \tag{1}$$

Lin and Sandler[6] define the activity coefficient of molecule i in a solution S, $\gamma_{i/S}$, in terms of the difference in the free energies of restoring the charges around a pure species i in a solution S, $\Delta G_{i/s}^{*\,res}$, and restoring the charges around a pure species i, $\Delta G_{i/i}^{*\,res}$. They[11,16] later suggest that using the Staverman−Guggenheim combinatorial term, $\ln \gamma_{i/S}^{SG}$, to account for molecular size and shape effects and to improve the calculation of the cavity-formation free energy.

$$\ln \gamma_{i/S} = \frac{\Delta G_{i/S}^{*\,chg} - \Delta G_{i/i}^{*\,chg}}{RT} + \ln \gamma_{i/S}^{SG} \tag{2}$$

Treating the homogeneous medium as a perfect conductor,[1] Lin and Sandler[6] define the charging free energy as the sum of two terms: (1) the ideal solvation energy, ΔG^{IS}, and (2) the restoring free energy, ΔG^{*res}, which represents the free energy necessary to remove the screening charges from the cavity surface after placing the solute in the cavity.

Lin and Sandler[6] define the restoring free energy as the sum of the products of the sigma profile and the natural log of the segment activity coefficients over all surface segments.

$$\frac{\Delta G_{i/s}^{res}}{RT} = \sum_{\sigma_m} \left[n_i(\sigma_m) \frac{\Delta G_{\sigma_m/s}^{*\,res}}{RT} \right] = n_i \sum_{\sigma_m} p_i(\sigma_m) \ln \Gamma_s(\sigma_m) \tag{3}$$

where $n_i(\sigma_m)$ is the number of segments with a surface-charge density σ_m; n_i is the total number of surface segments around the molecular cavity; $p_i(\sigma_m)$, the sigma profile for a molecule i, is the probability of finding a segment with a surface-charge density σ_m; $G_s(\sigma_m)$ is the activity coefficient for a segment m of charge density, σ_m. We calculate the segment activity coefficient for the segment in a solution $\Gamma_S(\sigma_m)$ and in a pure liquid

$\Gamma_i(\sigma_m)$ as derived rigorously using statistical mechanics.[6]

$$\ln \Gamma_s(\sigma_m) = -\ln\left\{\sum_{\sigma_n} p_s(\sigma_n)\Gamma_s(\sigma_n)\exp\left[\frac{-\Delta W(\sigma_m, \sigma_n)}{RT}\right]\right\}$$

$$\ln \Gamma_i(\sigma_m) = -\ln\left\{\sum_{\sigma_n} p_i(\sigma_n)\Gamma_i(\sigma_n)\exp\left[\frac{-\Delta W(\sigma_m, \sigma_n)}{RT}\right]\right\} \quad (4)$$

The exchange energy $\Delta W(\sigma_m, \sigma_n)$ is the energy required to obtain one pair of segments from a neutral pair of segments. It contains contributions from electrostatic interactions or misfit energy E_{mf}, hydrogen-bonding interactions E_{hb}, and nonelectrostatic interactions E_{ne} of the segment pairs.[4,5]

$$\Delta W(\sigma_m, \sigma_n) = \left(\frac{\alpha'}{2}\right)(\sigma_m + \sigma_n)^2 + c_{hb}\max[0, \sigma_{acc} - \sigma_{hb}]\min[0, \sigma_{don} + \sigma_{hb}]\ [\equiv]\ \text{kcal·mol}^{-1} \quad (5)$$

The first term of eq 5 describes the electrostatic contribution, and the second term represents the hydrogen-bonding contribution to the exchange energy. The nonelectrostatic energy contribution is assumed constant and, therefore, does not appear in eq 5. The electrostatic contribution or misfit energy is a correction for induced surface charges, which screen any polarization effects by placing the molecule in an ideal conductor rather than next to another molecular cavity. Klamt and coworkers[2,4] fit the misfit energy constant α', the hydrogen-bonding constant c_{hb}, and the hydrogen-bonding sigma cutoff value σ_{hb} to experimental data. a_{eff} is the effective area of a standard surface segment, representing the contact area between different segments, a theoretical bonding site. The misfit energy constant α' is the product of the polarizability factor f_{pol} and the model constant α. Generally, the polarizability factor is a function of the dielectric constant of the medium, but Klamt suggests setting this factor to a constant 0.64.

$$f_{pol} = \frac{\epsilon - 1}{\epsilon + {}^1/_2} = 0.64 \quad (6)$$

$$\alpha = \frac{0.3a_{eff}^{3/2}}{\epsilon_0}\ [\equiv]\ \text{kcal·Å}^4\text{·mol}^{-1}\text{·e}^{-2} \quad (7)$$

$$\alpha' = f_{pol}\alpha\ [\equiv]\ \text{kcal·Å}^4\text{·mol}^{-1}\text{·e}^{-2} \quad (8)$$

Mathias et al.[17] suggest another model for the exchange energy, which introduces a new term to account for hydrogen bonding, ΔW_{hb}. We show the equations for this new definition of the exchange energy below.

$$\Delta W(\sigma_m, \sigma_n) = \frac{\alpha'}{2}(\sigma_m + \sigma_n)^2 + \Delta W_{hb}(\sigma_m, \sigma_n)\ [\equiv]\ \text{kcal·mol}^{-1} \quad (9)$$

$$\Delta W_{hb} = -c_{hb}\{\max[0, |\sigma_m - \sigma_n| - \sigma_{hb}^n]\}^2\ [\equiv]\ \text{kcal·mol}^{-1} \quad (10)$$

For our work concerning solubility, we use both eq 5 and eqs 9 and 10 for the exchange-energy definition and compare the result of each definition to gain knowledge concerning acceptable cases for both.

Finally, Lin and Sandler[6] arrive at the following expression for the activity coefficient of a species i in a mixture s as a function of the pure component and mixture segment activity coefficient.

$$\ln \gamma_{i/s} = n_i \sum_{\sigma_m} p_i(\sigma_m)[\ln \Gamma_s(\sigma_m) - \ln \Gamma_i(\sigma_m)] + \ln \gamma_{i/s}^{SG} \quad (11)$$

As in Mullins et al.,[7] we use a slightly different definition of our sigma profiles for pure components. Specifically, we modify eq 4 by substituting the area-weighted sigma profile $p_i'(\sigma)$ in place of the standard sigma profile $p_i(\sigma)$, which is analogous to the equation used by Klamt.[4] Each term in eq 12 has the same definition as that in eq 4. We also redefine eq 11 to incorporate the change in the definition of the sigma profile seen in eq 13.

$$\ln \Gamma_i(\sigma_m) = -\ln\left\{\sum_i \frac{p_i'(\sigma_n)}{A_i}\Gamma_i(\sigma_n)\exp\left[\frac{-\Delta W(\sigma_m, \sigma_n)}{RT}\right]\right\} \quad (12)$$

$$\ln \gamma_{i/s} = \frac{1}{a_{eff}}\sum_{\sigma_m} p_i'(\sigma_m)[\ln(\Gamma_s(\sigma_m)) - \ln(\Gamma_i(\sigma_m))] + \ln \gamma_{i/s}^{SG} \quad (13)$$

We refer to eq 13 as the COSMO-SAC model from this point forward.

2.2. Calculation of Sigma Profiles. A sigma profile is a probability distribution of the surface-charge density of a molecule or a mixture. COSMO-based models construct the molecular shaped cavity within the perfect conductor[1] according to a specific set of rules and atom-specific dimensions. Then, the molecule's dipole and higher moments draw charges from the surrounding medium to the surface of the cavity to screen, or cancel, the electric field both inside the conductor and tangential to the surface, allowing the molecule to move freely within the system without altering the system's overall energy. We calculate the induced charges on the solute surface in discretized space from Poisson's equation and the zero total potential boundary condition.

$$\Phi_{tot} = \Phi_i + \Phi(q^*) = \Phi_i + Aq^* = 0 \quad (14)$$

In eq 14, Φ_{tot} is the total potential on the cavity surface, Φ_i is the potential due to the charge distribution of the solute molecule i, $\Phi(q^*)$ is the potential as a function of the ideal screening charge q^*. We set $\Phi(q^*)$ equal to the product of the ideal screening charge q^* and coulomb interaction matrix A, which describes potential interactions between surface-charges and is a function of the cavity geometry.[4] The surface-charge distribution in a finite dielectric solvent is well-approximated by a simple scaling of the surface-charge density in a conductor σ^*.

We average these segment surface-charge densities σ^* from COSMO calculation output and obtain a new surface-charge density $\sigma = q_{avg}/a_{eff}$, where q_{avg} is the average screening charge for a given segment. Klamt sets the adjustable parameter, a_{eff}, to 7.1 Å.[2] Klamt[1] then defines the sigma profile $p_i(\sigma)$ for a molecule i as the probability of finding a segment with a surface-charge density using the following equations.

$$p_i(\sigma) = n_i(\sigma)/n_i = A_i(\sigma)/A_i \quad (15)$$

$$n_i = \sum_\sigma n_i(\sigma) = A_i/a_{eff} \quad (16)$$

$$A_i = \sum_\sigma A_i(\sigma)\ [\equiv]\ \text{Å}^2 \quad (17)$$

Here, $n_i(\sigma)$ is the number of segments with a surface-charge density σ, n_i is the total number of surface segments around the molecular cavity, A_i is the surface area of the molecular cavity, and $A_i(\sigma)$ is the total surface area of all of the segments with a particular charge density σ. $A_i(\sigma)$ and $n_i(\sigma)$ are proportional by a_{eff}, $A_i(\sigma) = a_{eff}n_i(\sigma)$, as defined by Lin and Sandler.[6] We use area-weighted sigma profiles $p_i'(\sigma)$ in both databases according to eq 18.

$$p_i'(\sigma) = p_i(\sigma)A_i = A_i(\sigma)\ [\equiv]\ \text{Å}^2 \quad (18)$$

The sigma profile of a mixture is also a weighted average of the pure-component sigma profiles. In principle, the mixture sigma profile $p_S(\sigma)$ is not limited to a specific number of components.

$$p_S(\sigma) = \frac{\sum_i x_i n_i p_i(\sigma)}{\sum_i x_i n_i} = \frac{\sum_i x_i A_i p_i(\sigma)}{\sum_i x_i A_i} = \frac{\sum_i x_i p_i'(\sigma)}{\sum_i x_i A_i} \quad (19)$$

Lin and Sandler[6] also define the averaging algorithm for the segment surface-charge densities σ^* to calculate the new surface-charge densities σ.

$$\sigma_m = \frac{\sum_n \sigma_n \frac{r_n^2 r_{eff}^2}{r_n^2 + r_{eff}^2} \exp\left(-\frac{d_{mn}^2}{r_n^2 + r_{eff}^2}\right)}{\sum_n \frac{r_n^2 r_{eff}^2}{r_n^2 + r_{eff}^2} \exp\left(-\frac{d_{mn}^2}{r_n^2 + r_{eff}^2}\right)}\ [\equiv]\ \text{e/Å}^2 \quad (20)$$

In the equation, σ_m is the average surface-charge density on segment m, the summation is over *n* segments from the COSMO output, and r_n is the radius of the actual surface segment, which have an assumed circular geometry. The effective radius, $r_{eff} = \sqrt{a_{eff}/\pi}$, is an adjustable parameter in this model, and d_{mn} is the distance between the two segments m and n[2,6]. The paired segments m and n have segment charge densities σ_m and σ_n, respectively. We use an averaging radius,[3] $r_{av} = 0.81764$ Å, for the sigma-averaging algorithm in place of the effective segment radius r_{eff}. This corresponds to the average segment surface area of $a_{av} = \pi r_{av}^2 = 2.100265$ Å^2. We use the averaging algorithm in eq 21 to calculate the average surface-charge density σ_m.

$$\sigma_m = \frac{\sum_n \sigma_n \frac{r_n^2 r_{av}^2}{r_n^2 + r_{av}^2} \exp\left(-\frac{d_{mn}^2}{r_n^2 + r_{av}^2}\right)}{\sum_n \frac{r_n^2 r_{av}^2}{r_n^2 + r_{av}^2} \exp\left(-\frac{d_{mn}^2}{r_n^2 + r_{av}^2}\right)}\ [\equiv]\ \text{e/Å}^2 \quad (21)$$

Our averaging algorithm is identical to those presented by Lin and Sandler[6] and Klamt and co-workers,[1,2] except that we use a different value for r_{av}. Klamt and co-workers report using averaging radii ranging between 0.5 and 1.0 Å, stating that "the best value for the averaging radius r_{av} turns out to be 0.5 Å. This is less than the initially assumed value of about 1 Å".[2] The deliverables for our two databases include results from the density-functional theory (DFT) calculations, thus enabling future work in the optimization of r_{av}. We use eq 21 when generating all of the sigma profiles for the VT-2006 Solute Sigma Profile Database. Each sigma profile contains 50 segments, ranging from −0.025 to 0.025 e/Å^2 with a step size of 0.001 e/Å^2. Given the sigma profiles and the COSMO-based models, we compute various physical properties, such as partition coefficients, infinite-dilution activity coefficients, and phase-equilibrium behavior, etc.[1,14,18] Figure 2 shows examples of area-weighted sigma profiles from the VT-2005 Sigma Profile Database.

Wang et al.[19] discuss an alternative method for generating sigma profiles by essentially dividing the sigma profile into two parts, a hydrogen-bonding sigma profile and a nonbonding sigma profile. They show improvement in some cases where hydrogen

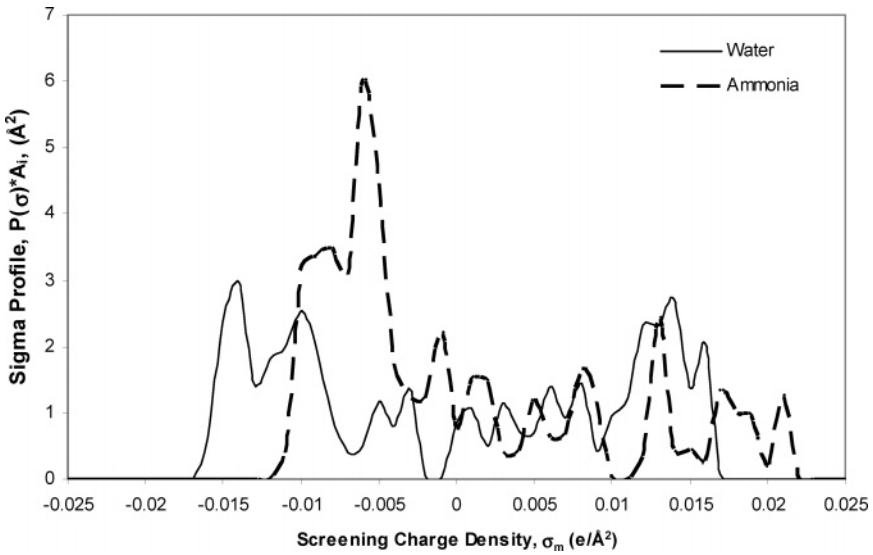

Figure 2. Water (VT-1076) and Ammonia (VT-1070) sigma profiles available in the VT-2005 Sigma Profile Database. Both sigma profiles have peaks above the hydrogen-bonding cutoff, signifying these compounds likely form hydrogen bonds.

bonding is prevalent; however, all of the sigma profiles discussed in this work have not undergone such treatment.

The sigma-profile generation procedure makes two crucial assumptions. The first assumption is that the optimized geometry from the DMol3 calculation in the vapor phase is identical to the optimal geometry in the condensed phase. This assumption can save large amounts of computational time for calculations on larger molecules. Factors such as solvent polarity, molecule size, and solvent−solute interactions could affect a solute molecule's structural conformation, leading to different structures in a condensed phase than in an ideal gas. The second assumption requires that the molecule is in the lowest-energy conformation once optimized, but several low-energy structural conformations may exist because of the freedom in choosing dihedral angles. Each conformation results in a slightly different sigma profile and therefore may affect property predictions.

2.3. Solubility Theory. Ben-Naim[20] defines "... the solvation process of a molecule s in a fluid l as the process of transferring the molecule s from a fixed position in an ideal gas phase g into a fixed position in the fluid or liquid phase l. The process is carried out at a constant temperature T and pressure P. Also, the composition of the system is unchanged." The molecule s serves as our solute and the fluid l as our solvent in the conventional sense. From this definition, we also infer the idea that solubility should focus on interactions between the solvated molecule and its surroundings. COSMO-based thermodynamic methods incorporate these molecular interactions by calculating "the chemical potential of any species in a mixture"[6] and therefore are potential tools for predicting solubility for a wide range of solutes. We use the solubility equation derived from ideal solubility in Prausnitz et al.,[21] which we show in two forms below in eq 22.

$$\ln x_{\mathrm{sol}} = \frac{\Delta H_{\mathrm{fus}}^{T_{\mathrm{m}}}}{RT_{\mathrm{m}}}\left(1 - \frac{T_{\mathrm{m}}}{T}\right) - \ln \gamma_{\mathrm{sol}}$$

$$\ln x_{\mathrm{sol}} + \ln \gamma_{\mathrm{sol}} = \ln K_{\mathrm{sp}} = \frac{\Delta H_{\mathrm{fus}}^{T_{\mathrm{m}}}}{RT_{\mathrm{m}}}\left(1 - \frac{T_{\mathrm{m}}}{T}\right) \qquad (22)$$

The second form of the solubility equation is useful in industrial applications. Using experimental solubility data and an activity coefficient model to regress K_{sp} is an alternative to measuring the latent heat of fusion and the melting temperature experimentally.

3. Computational Methods and Geometry Optimization

3.1. Sigma-Profile Generation and Geometry Optimization. Our procedure to calculate the sigma profile includes one optional step and three essential steps. The optional step is using a pre-optimization tool such as *Amber8*[32,33] or Accelrys *MS Forcite Plus*,[34] to provide an initial guess for the optimum low-energy geometry of a molecule. For the first essential step, we calculate the optimum low-energy geometry of an individual molecule in the ideal gas-phase based on the Hamiltonian energy using DFT. The next step is to calculate the charge and position of each segment on the surface of the geometrically optimized molecule in the condensed phase using both DFT and COSMO calculations. We assume that the low-energy optimal geometry does not change from the ideal gas phase to the condensed phase. We believe that this is a weak assumption for some molecules. However, we do not presently have a better alternative to selecting an appropriate condensed phase conformation without exhaustive computational effort. The third step is to average the charge surface segments using eq 21 to generate the sigma profile. We present a more detailed procedure of the three essential steps in Mullins.[8]

We use two supplemental software packages: (1) *Amber8*, developed by Pearlman et al.,[33] which simulates energy minimization as well as several other tasks; and (2) *MS Forcite Plus*,[34] a simulation module for annealing dynamics, in conjunction with our DMol3 calculations.

Our calculation settings and procedures for using both software tools for geometry optimization are available in Mullins.[8] The final result from using *Amber8* is a structure with a minimized molecular mechanics energy, which serves as an input structure for our sigma profile generation. *MS Forcite Plus* outputs the lowest-energy conformation of each anneal cycle, and we use the lowest-energy conformation from all of the anneal cycles as our initial guess for an optimized molecular structure. We find that the resulting low-energy conformations are not necessarily unique for each anneal cycle, and several low-energy conformations are identical, indicating that we have a higher probability of a global low-energy conformation.

For our purposes, we use *Amber8* optimized conformations as input structures for all of the compounds in the VT-2006 Solute Sigma Profile Database.

3.2. Convergence Method for Solubility Predictions. To ensure convergence of the solute mole fraction, we dampen the solution to the solubility of eq 22 by a damping factor $\omega = {}^1/_3$, which increases the total computation time but is more stable than Newton's method. We also normalize all of the mole fractions prior to the next iteration step to prevent calculating mole fractions outside of the possible physical range. When normalizing mole fractions for a system with more than two solvents, we keep the mole fraction ratio of those two solvents constant throughout each iteration for a given temperature. Equation 23 shows the convergence scheme we use for generating the initial guess for the solute mole fraction for the $i + 1$ iteration based on a damping factor ω, the value for the solute mole fraction from the previous iteration i, and the solution to eq 22.

$$x_{\mathrm{sol}}^{\mathrm{i}+1} = (\omega)x_{\mathrm{sol}}^{\mathrm{Eqn\ 22}} + (1 - \omega)x_{\mathrm{sol}}^{\mathrm{i}} \qquad (23)$$

This scheme requires the user to supply an initial guess for the solute mole fraction, and we use the literature value for validation purposes; however, if a literature value is not available, we recommend using the ideal solubility as an initial guess.

4. VT-2006 Solute Sigma Profile Database

In this section, we discuss the VT-2006 Solute Sigma Profile Database. We describe this database by its contents and deliverables, validation criteria, and error quantification. We present examples of the effect of conformational changes on solubility predictions for pure and mixed solvent systems. We also compare COSMO-SAC as an a priori tool for predicting solubility to other solubility prediction methods. In previous work concerning solubility modeling, researchers present a variety of model comparison studies and other solubility studies focusing on specific compounds.[3,36−42] In the majority of these publications, the authors use different methods to quantify the prediction errors. In some cases, the authors do not publish a listing of compounds by name, but by classification, thus preventing duplication of their work. Still other researchers use smaller sets of compounds, which do not lend themselves to generalizations about a model's overall accuracy. Chen and

Song[41] correlate and predict solid solubility using the NRTL-SAC model, they develop and report parameters for several solute and solvent compounds, and thus provide the means for a meaningful comparison. Therefore, we compare solubility predictions in pure solvents using both the NRTL-SAC and COSMO-SAC models.

4.1. Resources for VT-2006 Solute Sigma Profile Database. Our free, web-based database (www.design.che.vt.edu) includes several resources as follows:

1. Indices of the VT-2005 Sigma Profile Database, containing 1432 compounds, and the VT-2006 Solute Sigma Profile Database, containing 206 compounds, which are searchable by CAS-RN, chemical formula, and name. The VT-2005 Sigma Profile Database index also includes the normal boiling point temperature and the pure component vapor pressure as predicted by a revised COSMO-SAC-BP model.[11] The VT-2006 Solute Sigma Profile Database index includes the latent heat of fusion, normal melting temperature, and the Hildebrand solubility parameter.[22,23]

2. Procedures for generating sigma profiles using Accelrys' MS and using the *FORTRAN* programs to predict VLE and SLE. These procedures include screen captures and sample outputs for reference.

3. Calculation outputs from the DMol3 module geometry optimization and energy calculation tasks and the COSMO calculation. These files include atomic coordinates, surface charges, etc. in *.OUTMOL and *.COSMO files and are viewable as simple text.

4. Executable *FORTRAN90* programs and source code for calculation programs for sigma profile averaging and activity coefficient prediction for use with VLE and SLE systems.[7] Mullins[8] contains the *FORTRAN* source code for executable programs listed above that use the COSMO-SAC model to calculate VLE and SLE behavior.

4.2. Validation of VT-2006 Solute Sigma Profile Database. Here, we present several representative cases and a summary of the validation effort regarding the application of COSMO-SAC to solubility modeling. We compare the solubility prediction error by using exchange energies defined by Lin and Sandler,[6] eq 5, and by Mathias et al.,[17] eqs 9−10. Because the hydrogen-bonding term differs in the two exchange-energy definitions, we use shorthand notation, W_{hb} (Lin 2002) and W_{hb} (Mathias 2002), to designate which of these two exchange energies we use for a particular COSMO-SAC prediction. We also study the conformational effects of solutes and solvents on COSMO-SAC solubility predictions in pure and mixed solvents.

4.2.1. Error Calculation Methodology and Quantification. We compare experimental and predicted solute mole fractions on a logarithmic base 10 scale and quantify prediction error using the root-mean-squared error of $\log_{10}(x_{sol})$, RMSE, that is similar to Chen and Song.[41] We also calculate the absolute average relative error, AA%E, to compare predicted and experimental solute mole fractions in percent mole fraction (% M.F.). We use the definitions in eqs 24 and 25 to calculate the prediction error, where x_{sol}^{exp} represents the experimental solute mole fraction, x_{sol}^{pred} is the predicted solute mole fraction, and n is the number of solubility points.

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i}^{n}\left(\log_{10}\left(x_{sol_i}^{exp}\right) - \log_{10}\left(x_{sol_i}^{pred}\right)\right)^2\right]^{1/2} \quad (24)$$

$$\mathrm{AA\%E} = \frac{1}{n}\sum_{i}^{n}\left|\frac{x_{sol_i}^{exp} - x_{sol_i}^{pred}}{x_{sol_i}^{exp}}\right| \times 100\% = \frac{1}{n}\sum_{i}^{n}\left|1 - \frac{x_{sol_i}^{pred}}{x_{sol_i}^{exp}}\right| \times 100\% \; [\equiv] \; \%\ \mathrm{M.F.} \quad (25)$$

Because we have access to more data for some compounds, we use several schemes to ensure that we give equal weight for each calculation. When calculating the overall error for a given solute, we weight each solute−solvent pair equally by essentially taking an average of the average error. We calculate the average error, either RMSE or AA%E, for a given system over the full range of temperatures, and then average all of the systems for the given solute. When we calculate the error for all of the systems, we first calculate the average error per solute, weighting each solute−solvent pair equally, and then average the error over all of the solutes. This method magnifies the effects of outlying predictions; however, we believe it provides a fair analysis.

Because of the physical constraints of the composition domain [0:1], we observe very large and very small error measurements for which we discuss a proper frame of reference. Simply put, a model may do one of two things, under-predict or over-predict, when predicting a single value, such as solute mole fraction. In this case, under-prediction is a result of an over-prediction of the solute activity coefficient by the model and vice versa. It follows that when we want to compare an over-prediction with an under-prediction, we need a method for systematic quantification. Over-prediction and under-prediction generate very different AA%E values. The question becomes, which model prediction is better if one model over-predicts and the other model under-predicts solute mole fractions? We next ask, how do we determine which prediction is best from these error measurements? The root-mean-squared error of the logarithm of the solute mole fraction is linearly proportional to the ratio of the predicted to experimental solute mole fraction $x_{sol}^{pred}/x_{sol}^{exp}$ and is equivalent in value to the difference in the order-of-magnitude of the predicted and experimental solubilities. For example, if the experimental solute mole fraction equals 0.01, regardless of whether the model predicts a value of 0.001 m.f. or 0.1 m.f., the RMSE is 1.0. Essentially, RMSE is a tool for comparing over-prediction to under-prediction. However, the user must decide which prediction to use.

We see a different behavior as a result of under-prediction or over-prediction when calculating AA%E. The standard formula for the absolute average relative percent error is on the left-hand side of eq 25. However, when we rearrange this formula to the form on the right-hand-side, we can elucidate the nature of the equation. As the ratio of the predicted mole fraction to the experimental mole fraction, $x_{sol}^{pred}/x_{sol}^{exp}$, approaches zero, AA%E becomes 100%. This represents the severe case of under-prediction. In the opposing case, as the ratio of predicted solute mole fraction to experimental solute mole fraction increases due to model over-prediction, the AA%E rapidly increases, especially if the solute mole fraction is less than 0.1, which corresponds to 80% of the systems in this study.

Figure 3 gives a graphical representation of this discussion. This explanation hopefully places our error analysis of COSMO-SAC solubility modeling in proper perspective. The RMSE values are not affected by the differences of under- or over-prediction, whereas AA%E values differ greatly. In the under-

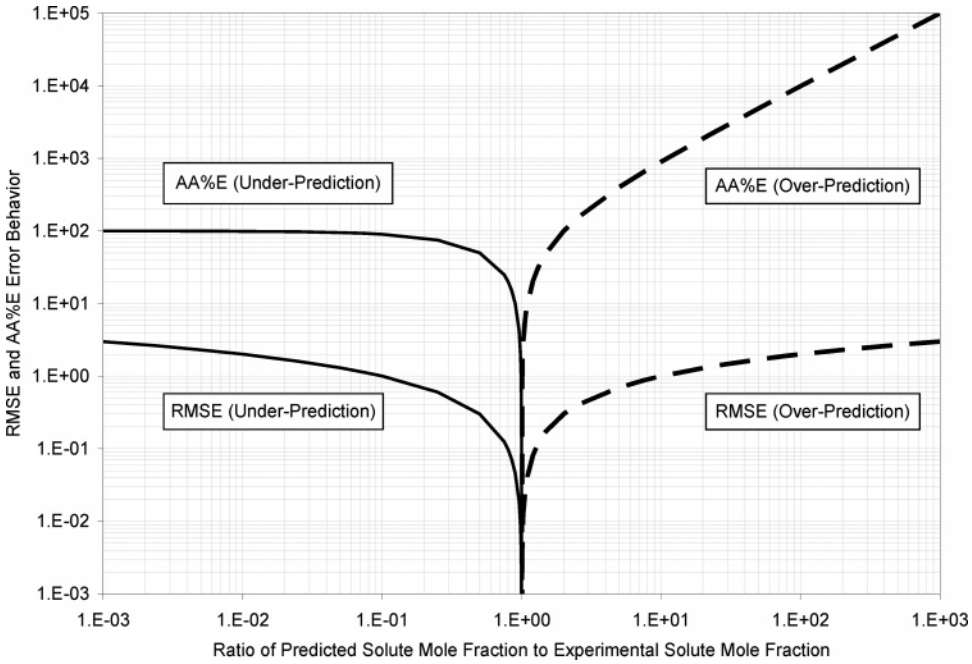

Figure 3. RMSE and AA%E behavior in relation to solubility modeling over-prediction and under-prediction on a logarithmic base 10 scale.

Table 1. Overall Error for Predicted Benzoic Acid Solubility in 50 Pure Solvents

exchange energy	W_{hb} (Lin 2002)	W_{hb} (Lin 2002)	W_{hb} (Mathias 2002)
error/solute	VTSOL-044	VT-0610	VTSOL-044
RMSE	0.2811	0.3554	0.3502
AA%E	125.2	222.7	1604.2

Table 2. Error Summary, RMSE, and AA%E of Predicted Solubility in Pure Solvents in Comparison to Experimental Values for the Entire Sample Set and Select Solutes That Exclude Outliers

sample set	literature points	solute−solvent pairs	no. solutes	W_{hb} (Lin2002) RMSE	W_{hb} (Lin2002) AA%E	W_{hb} (Mathias 2002) RMSE	W_{hb} (Mathias 2002) AA%E
1	2434	1356	194	0.9054	5314.7	1.1139	21215.8
2	2366	1295	171	0.7421	669.3	0.6346	2743.7
3	2205	1180	166	0.7487	711.8	0.9426	9499.4
4	2053	1114	149	0.7181	664.1	0.8254	980.9

prediction case, AA%E is limited to a maximum of 100%, but in the over-prediction case, AA%E is unbounded. We use AA%E to compare model error when model behavior is consistent. For example, this would be appropriate when we compare two solubility predictions that over-predict the solute mole fraction.

4.2.2. Solid Solubility Predictions in Pure Solvents. We use the solubility equation, eq 22, and the COSMO-SAC model to predict solubility in pure solvents at a given temperature *T*. We predict the activity coefficient of the solute using the COSMO-SAC model and published pure solute properties (heat of fusion and normal melting-point temperature)[22] to generate our predicted values. We assume that a single sigma profile describes a given compound's solid structure, or polymorph. We know that this assumption is often questionable, as many solids have multiple polymorphs; however, for our solubility study, this assumption is unavoidable. Figure 4 compares COSMO-SAC solubility predictions with published experimental solubilities in pure solvents with solute mole fractions for benzoic acid in 50 different solvents at 25 °C. Benzoic acid is one of 25 of compounds listed in both databases, VT-0610 and VTSOL-044 in the VT-2005 Sigma Profile Database and the VT-2006 Solute Sigma Profile Database, respectively. The difference in the sigma profiles between the two databases is that we generate the sigma profile for benzoic acid in the VT-2006 Solute Sigma Profile Database from energy minimized structures from *Amber8* output, whereas in the VT-2005 Sigma Profile Database, we do not use a pre-optimization tool for benzoic acid.

Weighting each solvent equally, we calculate the RMSE and AA%E for predicted benzoic acid solubility using the sigma profiles from both databases (VTSOL-044, VT-0610) and both exchange-energy expressions in Table 1. Using the Mathias et al.[17] exchange-energy expression shows a small improvement in the error for 12 of the 50 solvents, but the overall error is greater than the predictions using the Lin and Sandler[6] exchange-energy expression. We see an increase in error when we use the sigma profile from the VT-2005 Sigma Profile Database (VT-0610) over using the VTSOL-044 sigma profile, but this increase in overall error is smaller than the increase in overall error from using the Mathias et al.[17] exchange-energy expression in this case.

As stated above, we predict solubility in pure solvents for comparison with 2434 literature solubility points. This sample size includes 1356 solute−solvent pairs, 194 solutes, and 160 solvents. The types of solutes in this study include a wide range of functional groups and a combination of functional groups, which we break down into several subsets in our validation. To calculate the overall prediction error in comparison to all of literature values, we weight each solute−solvent pair equally. We calculate the RMSE and AA%E for the entire sample set and several smaller sample sets, which exclude select solutes and solute−solvent pairs. We set a cutoff value for exclusion at 5000% AA%E to remove outliers. We summarize these error calculations in Table 2.

We divide the entire data set into several sample sets to demonstrate the changes in overall error as we remove specific outlying solute−solvent pairs and outlying solutes as a whole from the error calculation. Sample set no. 1 includes all of the literature solubility points and has the greatest values for RMSE and AA%E. Sample set no. 2 excludes all of the solute−solvent pairs (61) with an error greater than the error cutoff value. Removing these outliers drastically reduces the overall error to

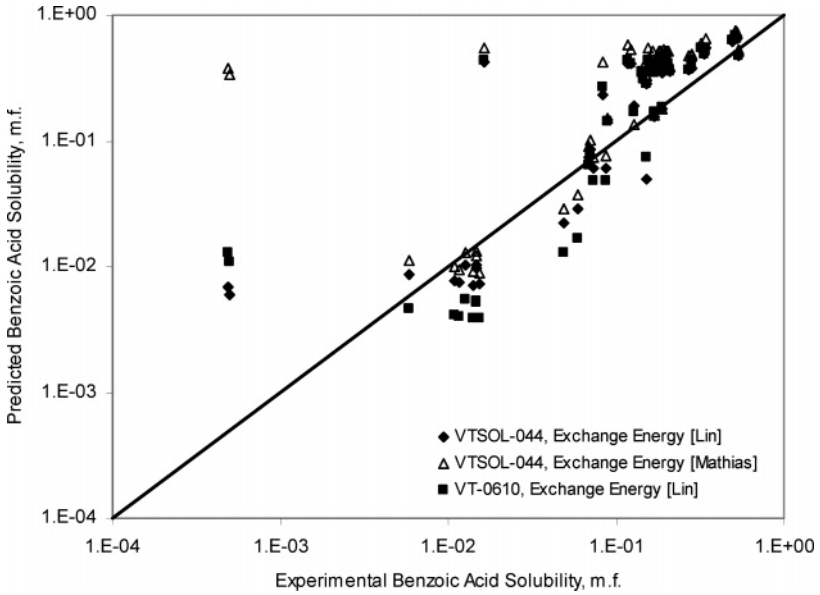

Figure 4. Predicted benzoic acid (VTSOL-044, VT-0610) solubility in 50 pure solvents at 298.15 K using both definitions for the exchange energy compared with experimental values on a logarithmic base 10 scale.[22]

Table 3. Nineteen Solutes with an Overall AA%E Exceeding the Error Cutoff Value of 5000% AA%E

index No.	solute name	formula	solvents	RMSE	AA%E
6	3-hydroxy-5-nitropyridine	$C_5H_4N_2O_3$	1	2.8851	76 661.0
7	5-chloro-3-pyridinol	C_5H_4ClNO	1	1.7924	6100.8
11	2-amino-2-nitropyridine	$C_5H_5N_3O_2$	1	1.9094	8017.9
15	2-aminopyridine	$C_5H_6N_2$	1	2.2439	17 433.8
20	2,4,6-triiodophenol	$C_6H_3I_3O$	2	1.6793	6517.5
25	2-hydroxynicotinic acid	$C_6H_5NO_3$	1	2.1987	15 701.8
47	3-hydroxybenzoic acid	$C_7H_6O_3$	1	2.8961	78 620.1
52	*p*-hydroxybenzamide	$C_7H_7NO_2$	1	2.7791	60 034.2
54	4-aminosalicylic acid	$C_7H_7NO_3$	1	1.8747	7633.0
59	4-hydroxybenzyl alcohol	$C_7H_8O_2$	1	3.0487	111 776.5
62	2,6-pyridinedimethanol	$C_7H_9NO_2$	1	3.2640	183 567.1
77	phenoxyacetic acid	$C_8H_8O_3$	1	2.0036	9982.3
79	3-hydroxy-4-methyloxybenzoic acid	$C_8H_8O_4$	1	1.9305	8420.3
85	acyclovir	$C_8H_{11}N_5O_3$	2	1.6588	6144.1
108	sulfapyridine	$C_{11}H_{11}N_3O_2S$	2	1.6406	22 005.2
146	mitotane	$C_{14}H_{10}Cl_4$	4	2.4328	101 776.2
155	morphine	$C_{17}H_{19}NO_3$	8	1.5885	25 827.0
191	desoxycorticosterone acetate	$C_{23}H_{32}O_4$	1	2.5366	34 306.0
196	norethindrone enanthate	$C_{27}H_{38}O_3$	1	1.9127	8078.3

669% AA%E and 0.742 RMSE. Sample set no. 3 excludes all of the solute−solvent pairs for 28 solutes whose overall error exceeds cutoff value when using the exchange energy defined by Lin and Sandler,[6] and sample set no. 4 excludes all of the solute−solvent pairs for 45 solutes whose overall error exceeds the cutoff value when using the exchange energy defined by Mathias et al.[17]

When we sort the predicted solubilities by their respective literature values, we see that the AA%E value rapidly increases as the literature solute composition in a pure solvent drops below 10 mole percent due to model over-prediction. We also see that on average, the exchange energy defined by Lin and Sandler[6] generates more accurate solubility predictions in pure solvents for the majority of the systems. On the basis of the exclusions in sample sets no. 2−4 in Table 2, we caution users when using the COSMO-SAC model to predict solubility with the 19 solutes listed in Table 3. This follows because each solute generates an error greater than the cutoff value for more than half of the solvents studied without further model improvements or in depth research. However, we must note that literature values for only one solvent are available for most of these solutes. In our experience, the COSMO-SAC model generally over-predicts the solute mole fraction; however, there are exceptions. The current COSMO-SAC parameter set is not parametrized with a data set that includes SLE data, which may contribute to the error in these predictions.

4.2.3. Effects of the Conformational Isomerism on Solubility Predictions. We study the effect of the changes in molecular conformations on solubility predictions. For our case study, we use five common pharmaceuticals and bio-related compounds, including: two small molecules (acetaminophen, $C_8H_9NO_2$, and aspirin, $C_9H_8O_4$), two medium-sized molecules (ibuprofen, $C_{13}H_{18}O$, and lidocaine, $C_{14}H_{22}N_2O$), and a large molecule (cholesterol, $C_{27}H_{46}O$). In each case, we present three to six different conformations, depending on the complexity of each compound. We also follow the geometry optimization procedure, whether we draw the conformation manually or we import the conformation from the *Amber8* output. We generate conformation A, in each case study, from *Amber8* output and draw the remaining conformations manually. We compare and rank each conformation based on several criteria: (1) the condensed phase energy, (2) the relative difference in the peak height of each conformation's sigma profile using the conformation with the lowest condensed phase energy as a reference, (3) error calculations (RMSE and AA%E) from the experimental mole fraction weighting each solvent equally, and (4) the number of

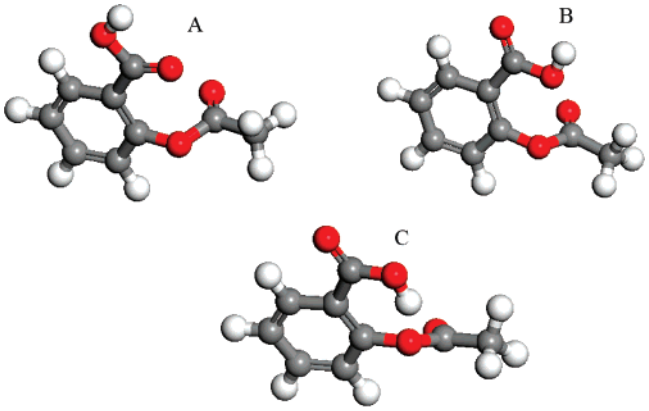

Figure 5. Three optimized aspirin (VTSOL-088) conformations with variations in the relative positions of the carboxyl group in proximity to the ester linkage.

solvents for which a conformation predicts the most accurate results relative to the other conformations. We define peak height as the relative difference between the sigma value, $p(\sigma)$, between two sigma profiles over the range of screening charge densities, σ_m, [−0.025, 0.001, 0.025] e/Å.[2] We refer to the number of solvents (criterion no. 4) as best case in the following summary tables in this section.

In the following subsections, we discuss the conformational effects for a small molecule (aspirin), a medium-sized molecule (lidocaine), and a large molecule (cholesterol). Detailed results for other drug molecules are available in Mullins.[8]

4.2.3.1. Aspirin (VTSOL-088, VT-1422, CAS-RN: 50−78-2, $C_9H_8O_4$). Aspirin is a common over-the-counter analgesic and antipyretic. It is also a nonsteroidal anti-inflammatory drug and a derivative of benzoic acid. Shown in Figure 5, we rotate the ester group by 45° between conformation A and C about the C−O bond, we rotate the carboxylic acid group by 180° between conformations A and B, and rotate the hydroxyl hydrogen atom by 180° between conformation B and C. Conformations A and B show how these functional groups might align when unaffected by other functional groups, whereas conformation C may better describe the interaction of these functional groups when in close proximity.

We calculate the condensed phase energy for each conformation in the COSMO-calculation. These are (−648.9150609 E_h), (−648.9149189 E_h), (−648.9148635 E_h), for conformations A,

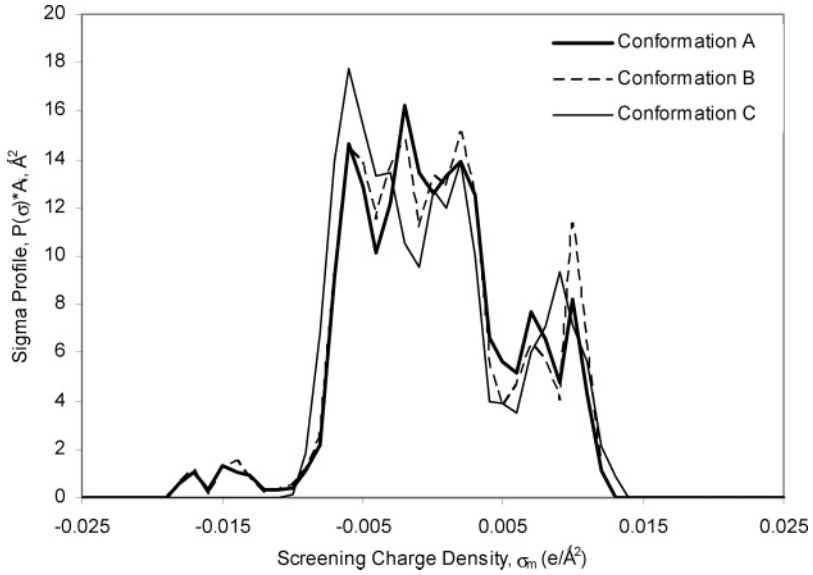

Figure 6. Sigma profiles for aspirin conformations A, B, and C.

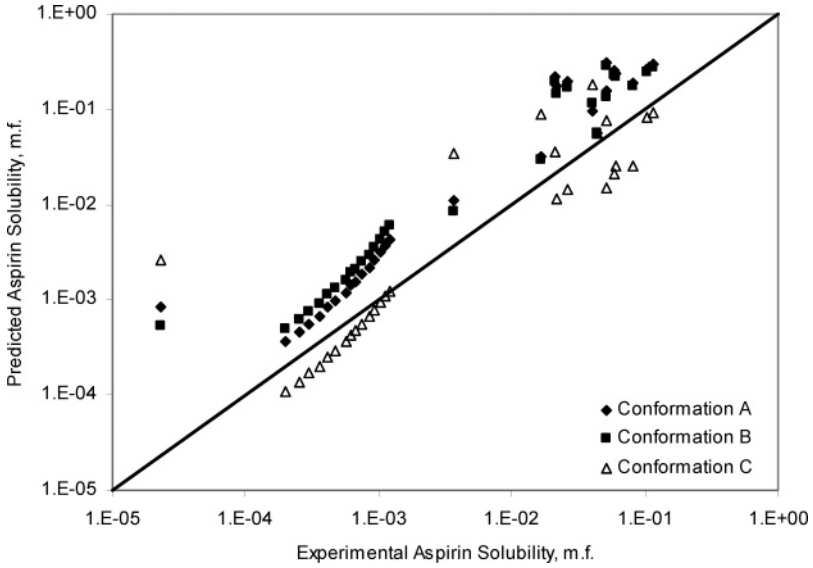

Figure 7. Predicted aspirin solubility in 15 pure solvents compared with experimental values[22] at various temperatures for each conformation using the exchange energy defined by Lin and Sandler[6] on a logarithmic base 10 scale.

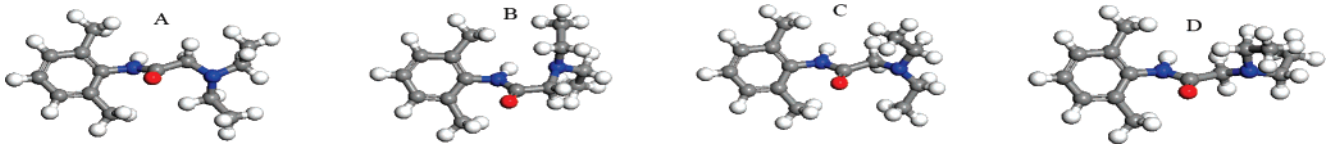

Figure 8. Four lidocaine conformations and their variations in structure concerning the relative positions of the amide and amine functional groups.

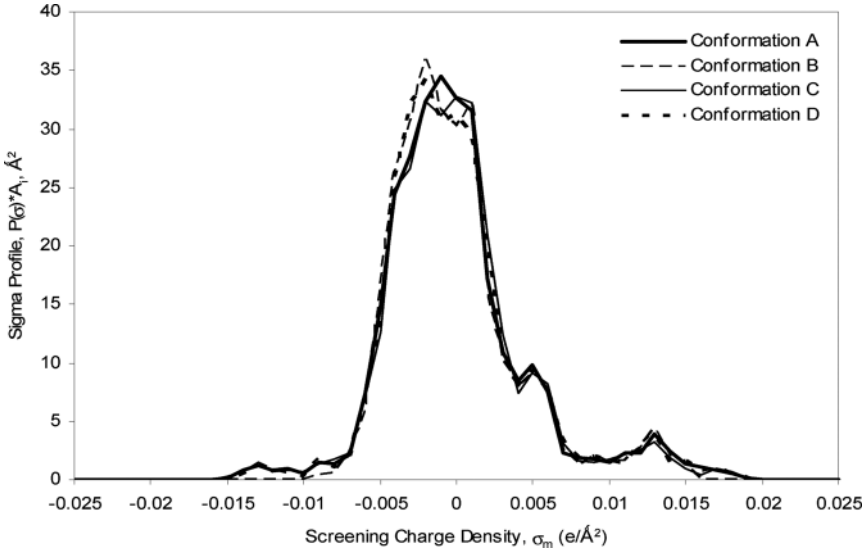

Figure 9. Sigma profiles for lidocaine conformations A, B, C, and D.

Table 4. Error Summary, RMSE, and AA%E, for Aspirin Conformations A, B, and C in 15 Pure Solvents at Various Temperatures Comparing Both Exchange-Energy Expressions

exchange energy	W_{hb} (Lin 2002)			W_{hb} (Mathias 2002)		
conformation	RMSE	AA%E	best case	RMSE	AA%E	best case
A	0.5857	496.8	1	0.7586	1216.3	2
B	0.5487	383.0	3	0.7346	1203.5	3
C	0.4947	895.4	11	0.4759	903.1	10

Table 5. Error Summary for Lidocaine Conformations A, B, C, and Din 20 Pure Solvents Comparing Both Exchange-Energy Expressions to Experimental Values[26]

	W_{hb} (Lin 2002)			W_{hb} (Mathias 2002)		
exchange energy conformation	RMSE	AA%E	best case	RMSE	AA%E	best case
A	0.1241	37.5	7	0.1265	42.5	4
B	0.1355	32.7	10	0.1214	33.3	11
C	0.1357	34.9	3	0.1386	40.9	3
D	0.1284	38.9	0	0.1293	44.4	2

B, and C, respectively. Conformation A has the lowest condensed phase energy, and the resulting sigma profiles of conformations B and C differ in peak height by 17 and 56% from the sigma profile of conformation A. The sigma profile of conformation C does not have peaks below −0.010 e/Å², whereas both conformations A and B have two peaks beyond this point (Figure 6). Table 4 summarizes the calculated errors for each conformation. We summarize the error calculations for aspirin solubility predictions in each solvent in Mullins.[8]

We find that conformation C generates predictions with the lowest RMSE regardless of which exchange-energy expression we use. Conformation C generates better predictions for 11 of 15 pure solvents when using the exchange energy defined by Lin and Sandler[6] and 10 of 15 solvents when using the exchange energy defined by Mathias et al.[17] From the difference in the RMSE between each conformation per solvent, we find that aspirin is sensitive to conformational effects. Figure 7 illustrates the sensitivity to conformational effects of the COSMO-SAC predicted solubilities.

4.2.3.2. Lidocaine (VTSOL-148, CAS-RN: 137−58-6, $C_{14}H_{22}N_2O$). Lidocaine is a local anesthetic and anti-arrhythmic drug. It functions by limiting nervous signals by blocking the sodium ion channels in the cellular membrane. Molecular lidocaine also consists of more atoms than molecular aspirin, and we calculate that a lidocaine molecule cavity occupies slightly more volume than the ibuprofen molecule cavity. We classify lidocaine as a medium-sized molecule, similarly to ibuprofen. We compare COSMO-SAC solubility predictions in pure solvents of four slightly different conformations in twenty solvents with experimental solubilities.[26] Conformation A results from using *Amber8* output as an initial structure, and we manually draw the remaining conformations by rotating the tertiary amine into different positions relative to the aromatic ring and the amide. Figures 8 and 9 illustrate the differences in the four conformations and the differences in the sigma profiles, respectively. We rank each conformation by lowest to highest condensed phase energy. Conformation B has the lowest condensed phase energy (−731.6647081 E_h), followed by conformation A (−731.6594249 E_h), D (−731.6588885 E_h), and C (−731.6559627 E_h) in that order. The relative differences of the peak height in the sigma profiles between conformation B and conformations A, D, and C are 13.6, 13.5, and 15.9%, respectively.

We compare lidocaine solubility for twenty pure solvents and summarize the resulting deviation from experimental values in Table 5. We present an itemized listing of calculated errors for each conformation and solvent in Mullins.[8] The exchange-energy expression defined by Mathias et al.[17] generates improved predictions for 12 of the 20 solvents studied, some of which are statistically significant.

We find that conformation B has the lowest prediction error for both exchange-energy expressions and the most accurate prediction for at least half of the solvents when using either exchange-energy expression. We also find that predictions of lidocaine solubility in pure solvents are significantly affected by variations in sigma profiles due to conformational changes

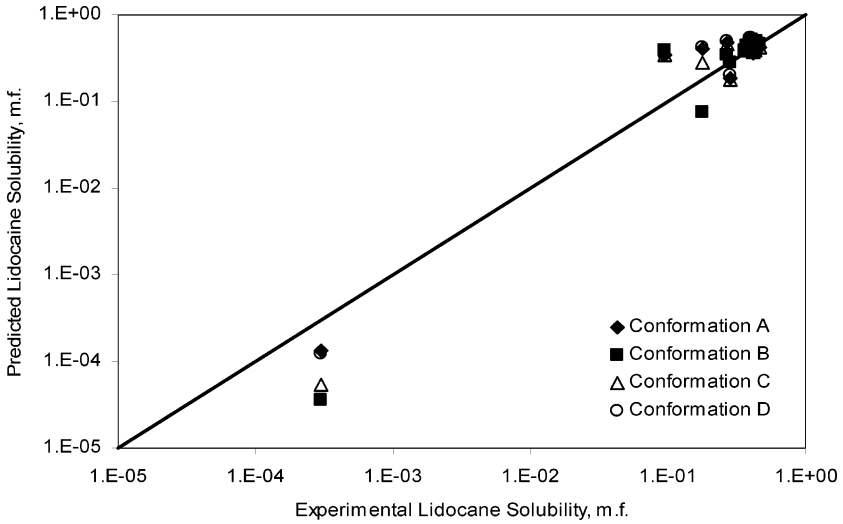

Figure 10. Predicted lidocaine solubility in 20 pure solvents using the exchange-energy expression defined by Lin and Sandler[6] compared to their experimental value at 298.15 K on a logarithmic base 10 scale.[26]

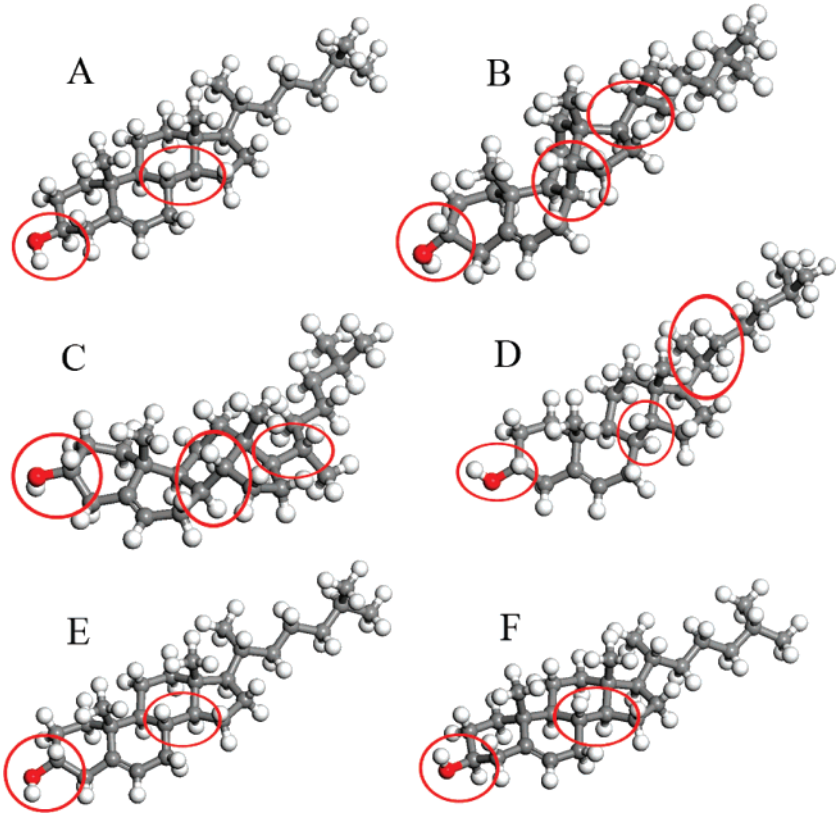

Figure 11. Six cholesterol conformations with circled regions to emphasize their relative differences. Conformations A and F have similar atomic positions for all of the atoms except the hydroxyl hydrogen placement.

for some solvents, such as 1,2-propanediol, 1,3-propanediol, and triacetin, while being fairly insensitive to conformational changes for other solvents. Figure 10 compares COSMO-SAC-predicted lidocaine solubilities in pure solvents to their experimentally determined values.[26]

4.2.3.3. Cholesterol (VTSOL-197, CAS-RN: 57-88-5, $C_{27}H_{46}O$). Cholesterol is an important bio-molecule and comes from several sources. Much research has gone into reducing and controlling the cholesterol levels in the body. Cholesterol is also a relatively large molecule when compared to the previous four cases. Figure 11 illustrates the relative differences in six optimized cholesterol conformations. Conformation F has the lowest condensed phase energy, followed by conformations A, E, B, C, and D ranked from lowest to highest condensed phase energy. However, conformations F and D are only separated by 0.025 E_h. By using conformation F as a reference, the remaining sigma profiles differ in peak height by 25, 17.5, 20, 52, and 20% for conformations A, B, C, D, and E, respectively. We study the COSMO-SAC solubility predictions of cholesterol in 60 pure solvents at various temperatures. Conformation A is the resulting structure from using *Amber8* as a pre-optimization tool.

We observe relatively small variations in the sigma profiles for the different conformations (Figure 12), which is likely the result of the conformational freedom of the molecule. Although cholesterol is the largest molecule we use to study conformational effects, it also has the largest fused-ring structure, thus inhibiting its conformational freedom. Despite the cholesterol

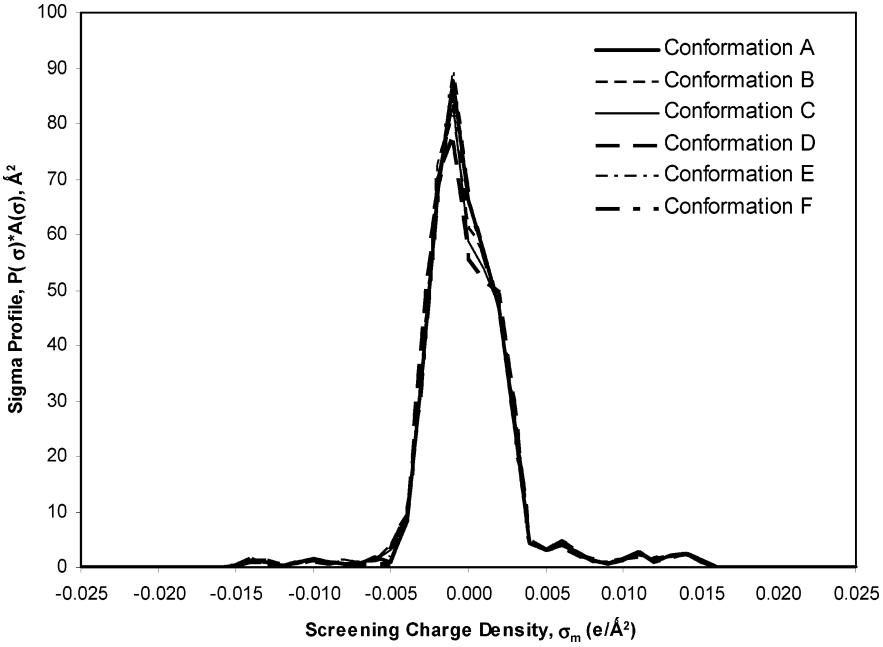

Figure 12. Sigma profiles for six cholesterol conformations.

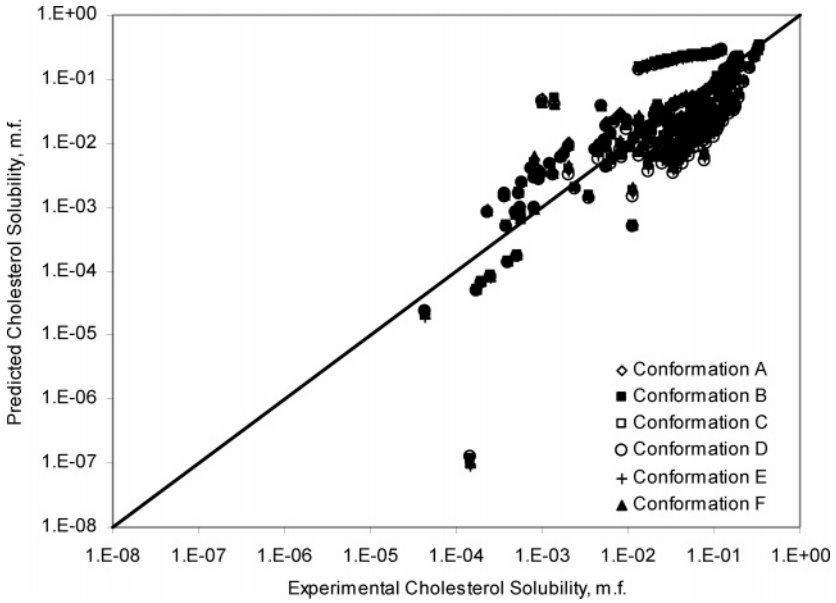

Figure 13. Predicted cholesterol solubility in 60 pure solvents for each conformation using the exchange-energy equation defined by Lin and Sandler[6] at various temperatures compared with their experimental values on a logarithmic base 10 scale.[26]

conformations exhibiting several differences in geometry, the variations in their sigma profiles are not significant. We see in Table 6 that using conformation A, with the exchange energy defined by Lin and Sandler,[6] generates the lowest overall RMSE and the most accurate solubility prediction for the majority of solvents. However, there is only a small variation between the best and worst predictions, which leads us to conclude that cholesterol is not sensitive to conformational changes. We present the error calculations for each conformation and each solvent in Mullins.[8]

Noting that conformations A, E, and F are similar with the exception of the hydroxyl group hydrogen atom position and the position of a methyl group on the tail of the molecule, there is still an uneven distribution when we compare the number of solvents for which each conformation generates the lowest RMSE. Figure 13 illustrates the relative insensitivity in cholesterol solubility predictions as a function of conformational changes.

Table 6. Error Summary for Each Cholesterol Conformation in 60 Pure Solvents Comparing Both Exchange-Energy Expressions to Experimental Values[26]

exchange energy conformation	W_{hb} (Lin 2002)			W_{hb} (Mathias 2002)		
	RMSE	AA%E	best case	RMSE	AA%E	best case
A	0.4602	214.5	31	0.5140	245.9	19
B	0.4760	206.4	7	0.5294	228.7	4
C	0.4771	207.4	3	0.5315	228.5	6
D	0.4911	202.9	6	0.5554	228.8	3
E	0.4658	208.0	5	0.5207	243.5	8
F	0.4650	213.5	8	0.5168	236.7	20

4.2.3.4. Summary of Conformational Effects. In the five cases studied, we look at molecules with various sizes and degrees of conformational freedom, from cholesterol and ibuprofen, which show little variation in their sigma profiles as a result of conformational changes, to acetaminophen and aspirin, which have the most conformational freedom. We find that the conformations with the highest calculated condensed

Table 7. Mixed-Solvent Solubility Error for Acetaminophen[45] and Naphthalene[46] Using Both Exchange-Energy Expressions from Experimental Solute Mole Fraction

					W_{hb} (Lin 2002)		W_{hb} (Mathias et al. 2002)	
solvent name(s)				solute name	RMSE	AA%E	RMSE	AA%E
water	acetone	toluene		acetaminophen-A	0.2431	67.2	0.2825	86.6
water	acetone	toluene		acetaminophen-B	0.2265	60.8	0.2820	87.2
water	acetone	toluene		acetaminophen-C	0.1792	42.3	0.2308	62.7
methanol	1-propanol	water		naphthalene	1.2985	1889.5	1.6321	4188.4
methanol	1-butanol	water		naphthalene	1.2714	1768.4	1.6049	3927.6
methanol	1-propanol	1-butanol	water	naphthalene	4.4427	2 903 815.5	4.4537	2 978 292.0

phase energy for acetaminophen, aspirin, and ibuprofen generate the best solubility predictions in pure solvents. We also note that the energy difference between the highest and lowest condensed phase energy is small, usually less than 5.0×10^{-2} E_h. Acetaminophen and aspirin are the smallest molecules in the above cases, and their predicted solubilities are also very sensitive to conformational variations. Ibuprofen is a larger molecule, but its predicted solubilities are relatively less sensitive as well as less accurate than acetaminophen and aspirin.

We find that the conformation with the lowest calculated condensed phase energy for lidocaine generates the best overall predicted solubilities. Lidocaine is a medium-sized molecule, and its predicted solubility sensitivity depends on the solute−solvent pair. Lidocaine solubility predictions are only sensitive for a small number of the solvents we study.

Cholesterol is the largest molecule and also the least conformationally flexible. We find that cholesterol conformation A generates the best overall solubility predictions. This conformation is second, having a lower calculated condensed phase energy than conformation F, but only by a small margin. Conformations A and F are very similar, except for the hydroxyl hydrogen atom placement and the arrangement of the isopropyl group on the tail of the molecule. However, cholesterol is not sensitive to conformational effects. The overall error, when excluding outliers for the best conformation for each molecule, generates an error less than the overall error for all of the solute−solvent pairs excluding outliers.

The key learning from our conformational studies is that, whereas conformational changes may yield slightly different sigma profiles and solubility predictions, they do not provide significant improvements on the overall quality of the COSMO-SAC model predictions.

4.2.4. Solid Solubility in Binary and Mixed Solvents. We now discuss COSMO-SAC-predicted solubility for selected cases in mixed solvents. Because the COSMO-SAC model is not limited to a fixed number of compounds, we study its accuracy as a function of the number of compounds present. First, we discuss the model predictions for 14 solutes in 14 binary solvents, representing 37 solute−solvent systems.[44−48] Several of the systems include ethanol and cyclohexane as one or both components in the binary solvents; in these cases, we look at all of the possible combinations of conformations for these solvents. Next, we analyze solubility predictions for three ternary solvent systems and one quaternary solvent system. During our iterative calculation scheme, we keep the ratio of the solvent composition constant to that reported in the experimental data for all of the systems.

For the binary solvent systems, all of the solutes are either small- or medium-sized molecules. The largest solute is sulfamethazine (VTSOL-130), which has two separate aromatic rings and a molecular weight of 278.33 g/mol. We exclude the system of (1) water, (2) ethanol, and (3) propyl-4-hydrobenzoate from our error calculations because its AA%E exceeds the error cutoff value of 5000%. We calculate an RMSE of 0.8839 and an AA%E of 986.7% using the exchange-energy expression defined by Lin and Sandler[6] for 36 binary solvent systems, using the best prediction in each case involving multiple conformations. The calculated RMSE and AA%E when using the exchange-energy expression defined by Mathias et al.[17] for 36 binary solvent systems are 1.0418 and 1779.3%, respectively. The Mathias et al.[17] exchange-energy expression presents a significant improvement in accuracy for only 4 of the 36 systems. The overall RMSE error for these systems is greater than the overall RMSE error for all of the solubility predictions in pure solvents. For the systems studied, we find that the error in the COSMO-SAC solubility predictions in binary solvents is greater than the overall model error for the solubility in pure solvents, excluding outliers.

Acetaminophen, cyclohexane, and ethanol are the only compounds with multiple conformations in this study. We predict solubility in a binary solvent system of (1) cyclohexane and (2) ethanol for 15 of the 37 systems. We use the two conformations for ethanol and cyclohexane and three conformations for acetaminophen in our studies. Interested readers may refer to Mullins[8] where we list the error measurements, RMSE, and AA%E for each binary solvent system, including multiple conformations, using both exchange-energy expressions.

Literature data for solubilities in ternary and quaternary solvents are limited, but we predict mixed solvent solubility for acetaminophen over a range of temperatures[45] and for naphthalene at 298.15 K[46] comparing both exchange-energy expressions. As with pure and binary solvents, the original exchange-energy expression by Lin and Sandler[6] generates more accurate predictions. We summarize the calculated error for each system in Table 7.

COSMO-SAC grossly over-predicts the naphthalene mole fraction for the quaternary solvent, but predicts solute mole fractions with comparable accuracy to pure and binary solvent systems for the ternary solvent systems. The sample size for these systems is not large enough to make general statements regarding model accuracy as a function of the number of components.

4.3. Comparison of COSMO-SAC and NRTL-SAC Solubility Predictions. The NRTL-SAC model developed by Chen and Song[39−41] is a variation of the original NRTL model, which incorporates the segment-based method similar to the polymer NRTL model.[49] There is a key difference between the COSMO-SAC[5,6] and NRTL-SAC[41] methods. Specifically, the NRTL-SAC model requires experimental data to regress the necessary parameters for each compound, whereas COSMO-SAC requires quantum-mechanical calculations which can be time-consuming, but not experimental data. COSMO-SAC uses a few adjustable parameters that are fixed for all compounds. We compare predicted solubilities in pure solvents by both methods.

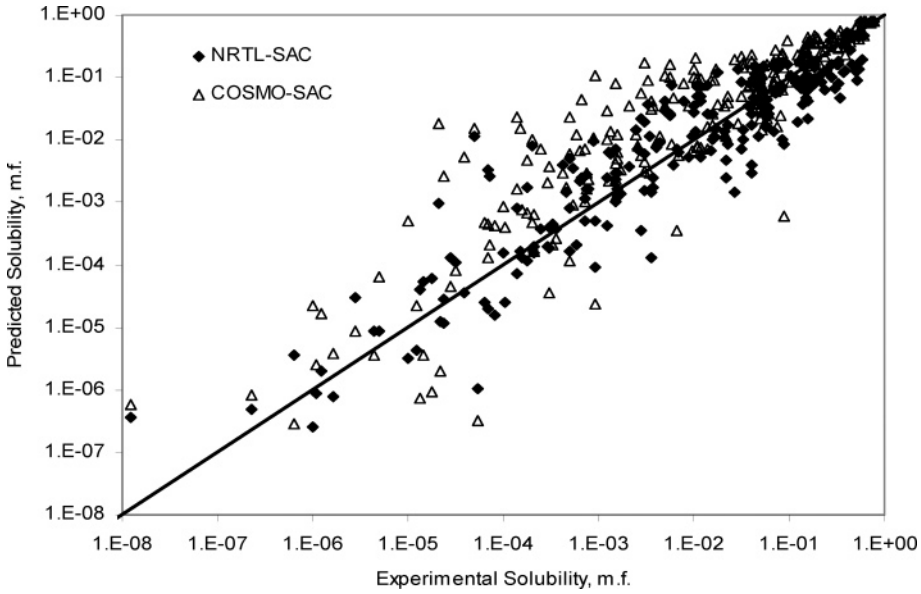

Figure 14. NRTL-SAC and COSMO-SAC predicted solubilities for 17 solutes comprising 258 experimental solubility points at 298.15 K, except acetaminophen solubility at 303.15 K, on a logarithmic base 10 scale.[22]

For this comparison, we use NRTL-SAC binary parameters (τ_{12}, τ_{21}, α_{12}, and α_{21}) and molecular segment parameters (hydrophobicity X, polarity types Y^- and Y^+, and hydrophilicity Z) for a set of 15 solutes published by Chen and Song.[41] These solutes include benzoic acid, 4-aminobenzoic acid, theophylline, methyl paraben, acetaminophen, aspirin, sulfadiazine, ephedrine, camphor, lidocaine, piroxicam, morphine, estrone, estriol, testosterone, haloperidol, and hydrocortisone with various solvents. We cannot compare the solubility predictions for both models for the same set of solvents for each solute due to incomplete literature data sets. Compliments of Dr. Chen,[50] we obtain the necessary NRTL-SAC molecular segment parameters for 128 solvents, all of which are included in our databases. From these 16 solutes and 128 solvents, we create a comparison set of 258 solubility points from the literature.[22] For acetaminophen and lidocaine, we use the best conformation from our conformational study with the COSMO-SAC model for this comparative study. We also use the Lin and Sandler[6] definition for the exchange energy in this comparison.

Finally, we use literature values of the latent heat of fusion and the normal melting-point temperature[22] to calculate K_{sp} from eq 22; however, we duplicate the results in Chen and Song[41] using their published K_{sp} values to ensure consistency. All of the literature values in this comparison are at 298.15 K, except acetaminophen at 303.15 K, which is the temperature at which Chen and Song[41] regress the binary interaction parameters. We also predict solubilities for several solute−solvent systems not included in the regression data set by Chen and Song. Therefore, we present different overall error values.

Figure 14 shows the predicted solubilities from both models for the entire data set relative to their respective experimental values. We see that the COSMO-SAC model over-predicts solubility for the majority of systems as a result of an under-prediction of the solute activity coefficient, and NRTL-SAC predicts a more even scattering of the values. We find that NRTL-SAC, which contains regressed parameters based on experimental data, is a more accurate method for predicting SLE behavior than the COSMO-SAC model for many of the systems studied.

For this set of points, the NRTL-SAC model is more accurate at predicting solubilities in pure solvents than the COSMO-SAC model with average RMSE values of 0.43 and 0.74, respectively. This is slightly greater than the average RMSE value of 0.37 that Chen and Song[41] report for NRTL-SAC for 14 solutes, but still in general agreement considering the larger comparison set. The COSMO-SAC model generally over-predicts the solute mole fraction with an RMSE value of 0.74, which is comparable to the overall RMSE for predicting solubilities in pure solvents reported in Table 2. Table 8 summarizes the error calculations for each solute and model, using the Lin and Sandler[6] definition for the exchange energy with the COSMO-SAC model.

5. Application Guidelines and Heuristics of COSMO-SAC Solid Solubility Predictions

To summarize our work and literature reports, we discuss several trends and guidelines for applying the COSMO-SAC model to solubility predictions. We consider aqueous and *n*-octanol solubilities, solvent conformational effect, nitrogen-containing compound solubility, exchange-energy expression, multicomponent system, general accuracy, and model sensitivity factors.

5.1. Aqueous Solubility. Klamt[51] reports satisfactory solubility predictions of hydrocarbons in aqueous binary systems using the *COSMOtherm* program developed by COSMO*logic* and released in 2002. The test set included aqueous systems of alkanes, alkylbenzenes, alkylcyclohexanes, and alkenes. Given this observation, we study aqueous solubilities in pure solvents for 147 of the 206 solutes in the VT-2006 Solute Sigma Profile Database. This includes 458 literature values of aqueous solubility in pure solvents. We find that all but two experimental values for aqueous solubility are below 10 solute mole percent. As one might expect from this composition range, the COSMO-SAC model predicts aqueous solubility less accurately than the average accuracy for all of the predicted solubilities. Using the exchange energy defined by Lin and Sandler,[6] we calculate an RMSE of 1.0409, which is slightly less than a 2-fold increase in the average error. We observe an even greater increase in error with an RMSE of 1.4303 when we use the Mathias et al.[18] exchange energy. We predict aqueous solubility for 14 of 19 solutes listed in Table 3. Most of the solutes in this database

Table 8. Comparison Error Summary for NRTL-SAC and COSMO-SAC Predicted Solubilities for 17 Solutes Comprising 258 Solute−Solvent Pairs at 298.15 K, except Acetaminophen at 303.15 K

VT-2006 index No.	solute name	CAS-RN	no. pure solvents	NRTL-SAC		COSMO-SAC	
				RMSE	AA%E	RMSE	AA%E
44	benzoic acid	65−85−0	40	0.5025	133.7	0.3138	96.6
53	4-aminobenzoic acid	150−13−0	7	0.3545	49.5	0.7323	551.9
60	theophylline	58−55−9	13	0.6306	161.0	0.8945	396.6
74	methylparaben	99−76−3	29	0.5004	65.7	0.4616	181.3
82	acetaminophen	103−90−2	24	0.4501	88.6	0.6709	370.4
88	aspirin-1	50−78−2	14	0.3887	50.5	0.7169	957.3
88	aspirin-2	50−78−2	14	0.4850	59.2	0.7169	957.3
100	sulfadiazine	68−35−9	2	0.2492	76.8	1.1812	2287.5
105	ephedrine	299−42−3	9	0.0636	12.4	0.4280	23.9
106	camphor	76−22−2	8	0.1411	21.9	0.2156	50.1
148	lidocaine	137−58−6	9	0.1365	33.2	0.3747	57.7
149	piroxicam	36 322−90−4	22	1.0149	2004.5	1.4503	7432.0
155	morphine	57−27−2	1	0.2410	42.6	1.0325	90.7
158	estrone	53−16−7	11	0.3033	73.9	0.5187	246.6
160	estriol	50−27−1	10	0.5916	50.8	1.1082	2304.8
163	testosterone	58−22−0	16	0.3731	119.0	0.4645	200.1
173	haloperidol	52−86−8	16	0.6629	368.5	1.2222	2067.5
181	hydrocortisone	50−23−7	13	0.7106	510.9	0.8267	1381.8
	average			0.4303	227.3	0.7419	1099.8

Table 9. Summary and Comparison of Solubility Prediction Error in *n*-octanol Using Both Exchange-Energy Expressions for 52 Solutes at Various Temperatures[22]

exchange energy	W_{hb} (Lin 2002)	W_{hb} (Mathias 2002)
RMSE	0.5640	0.6605
AA%E	312.4	582.5

Table 10. COSMO-SAC Predicted Pure-Solvent Solubility Error Summary for Nitrogen-Containing and Nitrogen Free Solutes[a]

sample set	literature points	solute solvent pairs	no. solutes	W_{hb} (Lin 2002)		W_{hb} (Mathias 2002)	
				RMSE	AA%E	RMSE	AA%E
nitrogen-Free	1770	919	101	0.7851	4461.9	0.9843	24 610.1
nitrogen-containing	664	437	93	1.0360	6240.7	1.2547	17 529.5

[a] Compare with sample set no. 1 in Table 2.

are large and hydrophobic and contain both polar and nonpolar functional groups. Therefore, it is difficult to discern trends solely based on one factor without the influence of another.

5.2. *n*-Octanol Solubility. We have a smaller sample size, 52 solutes, for our study of *n*-octanol solubility than our aqueous solubility study, but we observe far fewer outliers. Only one solute, niflumic acid (VTSOL-136), has an AA%E error exceeding the cutoff value. Summarizing the prediction error of the other solutes in *n*-octanol, we find that the RMSE and AA%E values are comparable to the overall model error when excluding outliers (Table 9).

We find that COSMO-SAC predicts the solubility in *n*-octanol more accurately than aqueous solubility, but we study fewer solutes for *n*-octanol than for water.

5.3. Solvent Conformational Effects on Solubility in Ethanol and Cyclohexane. Using the multiple conformations of ethanol and cyclohexane from the VT-2005 Sigma Profile Database, we investigate the effect of the conformational isomerism of the solvent on the predicted SLE behavior. We use the COSMO-SAC model and the solubility equation, eq 22, to predict ethanol solubility for 50 solutes and to predict cyclohexane solubility for 39 solutes, both at varying temperatures. Prediction accuracy improves for 39 of 50 solutes in ethanol and 29 of 39 solutes in cyclohexane, but solubility predictions in cyclohexane are much less sensitive to conformational changes than solubility predictions in ethanol. We also see that solubility predictions in cyclohexane are fairly insensitive to using different exchange-energy expressions, but solubility predictions in ethanol improve when we use the exchange-energy expression defined by Lin and Sandler.[6]

5.4. Nitrogen-Containing Compound Solubility. Ninety-three compounds in the VT-2006 Solute Sigma Profile Database contain nitrogen in some form, whether as an amide, amine, pyridine-derivative, nitrile, nitro functional group, or some combination thereof. Other researchers document issues concerning the accuracy of nitrogen-containing compound predictions with COSMO-based methods.[1,6,52] We analyze the effect of the presence of nitrogen atoms on the accuracy of COSMO-SAC solubility predictions in pure solvents. Nearly all of the solutes in the VT-2006 Solute Sigma Profile Database have more than one functional group, and approximately 60% of the 93 solutes have multiple nitrogen-containing functional groups as well as other functional groups.

When we compare the RMSE and AA%E values of the nitrogen-containing and nitrogen-free solutes, we find that COSMO-SAC predicts solubility for nitrogen-containing solutes significantly less accurately than solubility of nitrogen-free solutes. In Table 10, we see an increase in error, similar to the error difference in Table 2, when using the Mathias et al.[18] exchange energy for most solutes. Using the exchange-energy expression defined by Lin and Sandler[6] generally improves solubility predictions for the majority of the nitrogen-containing solutes (73 of 93). However, there are exceptions. For example, 7 of the 10 sulfur-containing compounds show more accurate results using the exchange-energy expression defined by Mathias et al.[18] There are significantly more outliers in the nitrogen-containing solute set, which contributes to its greater than average calculated error.

To further determine how a specific nitrogen functional group behaves, we categorize the 93 solutes by their nitrogen-containing functional group: amides, amines, nitriles, nitro, pyridines, aminosulfonyls, and multiple nitrogen functional groups. After calculating the RMSE and AA%E for each functional group category, we find that the prediction error for solutes with a single amide, amine, or nitro functional group is comparable to the average error for all of the literature values (Table 11). We calculate the greatest errors for molecules with pyridine derivatives, aminosulfonyl groups, and multiple nitrogen functional groups.

5.5. Comparison of Exchange-Energy Expressions of Lin and Sandler[6] and Mathias et al.[17] Of the 2434 solubility literature points in pure solvents and 1356 solute−solvent pairs,

Table 11. Error Summary of the Predicted Solubility for Each Nitrogen-containing Solute Categorized by Functional Group. Each Error Measurement Weights Each Solute−solvent Pair Equally, and Each Solute Equally. Compare with the Sample Set No. 1 Error Values from Table 2.

functional group	solutes	literature points	W_{hb} (Lin 2002) RMSE	W_{hb} (Lin 2002) AA%E	W_{hb} (Mathias 2003) RMSE	W_{hb} (Mathias 2003) AA%E
amine	12	151	0.7557	1091.2	1.0184	5204.2
amide	5	120	0.7787	845.2	1.3008	21 636.1
nitrile	1	3	0.1606	37.7	0.4517	183.2
nitro	5	27	0.7908	1030.2	1.4277	6954.6
pyridine	14	42	1.3167	15 532.0	1.6651	31300.1
aminosulfone	11	42	0.9376	3448.9	0.9320	3222.4
multiple groups	57	324	0.9782	4869.5	1.1334	17 204.9

we see an improvement in predicting solute mole fractions for 444 solute−solvent pairs by using the exchange-energy expression of Mathias at al.[17] These pairs include 106 of the 160 total solvents and 102 of 194 total solutes. We identify several solvents and solutes that have a higher probability for consistent improvement for predicted solubilities in pure solvents using the Mathias et al.[17] exchange-energy definition. Each solvent generates improved predictions for more than 50% of the solute−solvent pairs and constitutes a 5% improvement or greater in the overall RMSE per solvent or solute. We recommend using the Mathias et al.[18] exchange-energy definition for 11 solutes and 14 solvents (listed in Table 12).

5.6. Effect of Multicomponent Systems on Accuracy of Solubility Predictions. With the given sample set of 37 binary solvent systems, COSMO-SAC predicts solute mole fractions for 36 systems within comparable accuracy to the model predictions for solubilities in pure solvents. The model accuracy is also similar for ternary solvent systems, but we only study three systems, which is not a large enough sample to make a general statement concerning accuracy. See Table 7 for the summarized error values of solubility predictions in mixed solvents. We find that conformational effects of both solutes and solvents play a role in the overall accuracy of the model for binary and ternary solvent systems, similar to their effects on solubility predictions in pure solvents.

5.7. Accuracy of the COSMO-SAC Model as a Solubility Predictor. The COSMO-SAC model systematically overpredicts the solute mole fraction, but this model is an improvement over using an ideal solubility and is a useful method for a priori solubility predictions for compounds without experimental data. Refer to Table 2 for an overall error summary of COSMO-SAC solubility predictions in pure solvents. We recommend looking at a similar representative system to the system of interest before using COSMO-SAC to model solubility and evaluating this model for individual solute−solvent pairs. For example, if a new compound contains an ester and an amine functional group, find a similar molecule from our databases to serve as a representative chemical for how solubility predictions with this new compound may behave. Although the overall average error for a particular solute may be poor, particular solute−solvent pairs may generate predictions with acceptable accuracy or vice versa.

5.8. Effect of Conformation Pre-optimization with Amber 8 on Solubility Predictions. We recommend *Amber8* for generating initial molecular geometries for smaller molecules with conformational freedom for COSMO-SAC solubility modeling applications. However, we note that only one *Amber8*-optimized conformation of the five conformational variations in our drug case study presented in Section 4.3.3.3 generates the most accurate solubility predictions in pure solvents. The best conformation does not necessarily have the lowest condensed phase energy for these cases as well. In general, we recommend that users consider manually drawn conformations as well as optimized initial structures from *Amber8*, *MS Forcite Plus*, or other pre-optimization tools when studying molecules with greater conformational freedom.

6. Conclusions

With the completion of this work, we have provided the means to predict VLE and SLE behavior using the COSMO-SAC model developed by Lin and Sandler.[5,6] We have produced molecule-specific sigma profiles for 1670 organic compounds, including 1464 solvents and 206 pharmacologically related solutes, and validated the compound's sigma profile by predicted pure-component vapor pressure in the case of VLE and predicted solubility in pure solvents in the case of SLE. We have studied conformational isomerism, or the existence of alternative, low-energy, stable variations in molecular structure, and their effects on thermodynamic property prediction. We separate these compounds into two databases, VT-2005 Sigma Profile Database[8] and the VT-2006 Solute Sigma Profile Database.

The COSMO-SAC model predicts VLE behavior more accurately than SLE behavior with the current model parameters, but this model's strength lies in its ability to generate a priori property predictions for compounds without experimental data. Excluding 23 of the total 194 solutes for which we compare solubility data in pure solvents to literature values as outliers, COSMO-SAC predicts, a priori, without fitted parameters, solute mole fractions with an average RMSE of 0.742 ($\log(x_{sol})$ units), which is greater than the average RMSE of 0.430 for the NRTL-SAC model, with parameters regressed from experimental data.

Table 12. Recommended Solutes and Solvents for Use with the Mathias et al.[17] Exchange-Energy Expression

VT-2006 Solute Sigma Profile Database			VT-2005 Sigma Profile Database		
index no.	solute name	CAS-RN	Index No.	solvent name	CAS-RN
088	acetylsalicylic acid	50−78-2	0242	benzene	71−43−2
089	sulfamethoxazole	723−46-6	0243	toluene	108−88−3
101	ephedrine	299−42-3	0583	acetic-acid	64−19−7
105	sulfisomidine	515−64-0	0584	propionic-acid	79−09−4
115	sulfamethazine	57−68-1	0638	methyl-acetate	79−20−9
126	thioxanthone	492−22-8	0639	ethyl-acetate	141−78−6
129	9,10-anthraquinone	84−65-1	0641	*N*-butyl-acetate	123−86−4
130	lidocaine	137−58-6	0728	1,4-dioxane	123−91−1
134	prostaglandin	363−24−6	0749	Triethylene-glycol-dimethyl-ether	112−49−2
135	progesterone	57−83−0	0785	dichloromethane	75−09−2
138	hydrocortisone	50−23−7	0786	chloroform	67−66−3
			0807	chlorobenzene	108−90−7
			0962	pyridine	110−86−1
			1388	formamide	75−12−7

The accuracy of the COSMO-SAC model depends largely on the molecular conformation and the exchange-energy expression. The use of pre-optimization tools, such as *Amber8* and *MS Forcite Plus*, to generate an initial molecular structure improves COSMO-SAC solubility predictions, especially with small, flexible molecules. However, a compound's sensitivity to changes in its own sigma profile and other sigma profiles in the system as a result of conformational variations is system-specific. Our study shows that variations in sigma profiles caused by conformational differences may or may not provide some improvement to the overall quality of COSMO-SAC predictions. Therefore, we suggest that there is probably no gain to be made in trying to find the "best" conformation until future improvements in computational power or calculation methods make such an effort justifiable with improved prediction results. We have confirmed the issues other researchers have observed regarding COSMO-SAC model applications involving nitrogen-containing compounds, and we have studied the effects of individual nitrogen-containing functional groups on predicted solubility. We have also discussed COSMO-SAC applicability as a solubility predictor in several common solvents: water, ethanol, cyclohexane, and *n*-octanol. Ultimately, the guidelines and resources we provide should help future researchers and practitioners improve the applicability and accuracy of the COSMO-SAC model.

Acknowledgment

We thank Dr. Chau-Chyun Chen, Vice President for Technology at Aspen Technology, Inc. for his sound advice on this work. We thank Alliant Techsystems (particularly Ken Dolph, Vice President), Aspen Technology (particularly Larry Evans, Founder, and Mark Fusco, President), China Petroleum and Chemical Corporation (particularly Tianpu Wang, President, and Xianghong Cao, Chief Technical Officer), Formosa Petrochemical Corporation (particularly Wilfred Wang, Chairman and President), and Milliken Chemical (particularly John Rekers, President; Jean Hall, Business Manager; Jack Miley, Director of R &D and Technology) for supporting our educational programs in computer-aided design and process systems engineering at Virginia Tech.

Symbols

A = Coulomb interaction energy matrix
AA%E = absolute average relative percent error
a_{av} = average segment surface area, Å^2
a_{eff} = effective segment surface area, Å^2
A_i = total molecular cavity surface area, Å^2
$A_i(\sigma_m)$ = area of segments with charge density σ, e/Å^2
a.u. = atomic unit, Bohr radius, 5.2918×10^{-11} Å
c_{hb} = hydrogen-bonding constant, kcal Å^4 mol^{-1} e^{-2}
d_{mn} = distance between surface segment m and n, Å
e = elementary charge, 1.6022×10^{-19} coulomb
E_h = Hartree, an atomic unit of energy, $4.35974417 \times 10^{-18}$ J
f_{pol} = polarizability factor, 0.64
ΔG = Gibbs free energy change, kcal/mol
ΔG^{IS} = Gibbs ideal solvation energy, kcal/mol
ΔG^{*cav} = cavity formation free energy, kcal/mol
ΔG^{*chg} = charging free energy, kcal/mol
ΔG^{*res} = restoring free energy, kcal/mol
ΔG^{*sol} = solvation free energy, kcal/mol
l_i = Staverman−Guggenheim (SG) combinatorial term parameter
n_i = total number of segments on the surface of the molecular cavity
$n_i(\sigma)$ = number of segments with charge density σ
$p_i(\sigma)$ = sigma profile, probability of segment i having a charge density σ
$p_i'(\sigma)$ = area-weighted sigma profile of component i, Å^2
$p_S(\sigma)$ = sigma profile of mixture S
q = standard area parameter, 79.53 Å^2
q^* = ideal screening charge, e
q_{avg} = average screening charge, e
q_i = normalized surface area parameter for SG combinatorial term
r = standard volume parameter, 66.69 Å^2
r_{av} = average segment radius, Å
r_{eff} = effective segment radius, Å
r_i = normalized volume parameter for SG combinatorial term
r_n = circular segment radius, Å
R = ideal gas constant, 0.001987 kcal mol^{-1} K^{-1}, 8.314 kJ kmol^{-1} K^{-1}
V_i = Molecular cavity volume, Å^3
$\Delta W(\sigma_m, \sigma_n)$ = exchange-energy between segments σ_m and σ_n, kcal mol^{-1}
ΔW_{hb} = hydrogen-bonding contribution of the exchange-energy, kcal mol^{-1}
x_{sol} = solute mole fraction, m.f.
x_i = mole fraction of component i
$x_{j,I}$ = segment mole fraction of component i
z = coordination number

Greek Symbols

α = model constant, Å^4 kcal e^{-2} mol^{-2}
α' = misfit energy constant, Å^4 kcal e^{-2} mol^{-2}
γ_i = activity coefficient of solute i
γ_I^C = combinatorial contribution to NRTL-SAC activity coefficient
γ_I^R = residual contribution to NRTL-SAC activity coefficient
$\gamma_{i/S}^{SG}$ = Staverman−Guggenheim activity coefficient of solute i in a solvent S
γ_{sol} = solute activity coefficient
$\Gamma_i(\sigma_m)$ = segment activity coefficient of segment σ_m in a pure liquid i
$\Gamma_m^{lc,I}$ = segment local composition interaction contribution of component I
Γ_m^{lc} = segment local composition interaction contribution
$\Gamma_s(\sigma_m)$ = segment activity coefficient of segment σ_m in a solvent S
ϵ = dielectric constant
θ_o = permittivity of free space, 2.395E-04
θ_i = composition-weighted volume fraction
σ = charge density, e/Å^2
σ^* = surface segment charge density from COSMO calculation output
σ^r = molecular rotational symmetry number
σ_{acc} = hydrogen acceptor segment
σ_{don} = hydrogen donor segment
σ_{hb} = hydrogen-bonding cutoff value, 0.0084 e/Å^2
σ_{hb}^n = new hydrogen-bonding cutoff value from Mathias defined exchange-energy, 0.0084 e/Å^2
σ_m = segment charge density of segment m
ϕ_i = composition-weighted surface area fraction
Φ_i = potential due to the charge distribution of the solute i
$\Phi(q^*)$ = potential as a function of the ideal screening charge q^*

Φ_{tot} = total potential on the cavity surface

Literature Cited

(1) Klamt, A. Conductor-Like Screening Model for Real Solvents: A New Approach to the Quantitative Calculations of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224.

(2) Klamt, A. COSMO and COSMO-RS. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Chichester, 1998.

(3) Eckert, F.; Klamt, A. Fast Solvent Screening via Quantum Chemistry: COSMO-RS Approach. *AIChE J.* **2002**, *48*, 369.

(4) Klamt, A.; Eckert, F. COSMO-RS: A Novel and Efficient Method for the a priori Prediction of Thermophysical Data of Liquids. *Fluid Phase Equilib.* **2000**, *172*, 43.

(5) Lin, S. T. Quantum Mechanical Approaches to the Prediction of Phase Equilibria: Solvation Thermodynamics and Group Contribution Methods. Ph.D. Dissertation, University of Delaware, Newark, DE, 2000.

(6) Lin, S. T.; Sandler, S. I. A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model. *Ind. Eng. Chem. Res.* **2002**, *41*, 899.

(7) Mullins, E.; Oldland, R.; Liu, Y. A.; Wang, S.; Sandler, S. I.; Chen, C. C.; Zwolak, M.; Seavey, K. C. Sigma-Profile Database for Using COSMO-Based Thermodynamic Methods. *Ind. Eng. Chem. Res.* **2006**, *45*, 4389.

(8) Mullins, E. Application of COSMO-SAC to Solid Solubility in Pure and Mixed Solvent Mixtures for Organic Pharmacological Compounds. Thesis, M. S. Virginia Polytechnic Institute and State University, Blacksburg, 2007.

(9) Lin, S. T.; Chang, J.; Wang, S.; III, W. A. G.; Sandler, S. I. Prediction of Vapor Pressures and Enthalpies of Vaporization Using a COSMO Solvation Model. *J. Phys. Chem.* **2004**, *108*, 7429.

(10) Lin, S. T.; Sandler, S. I. Prediction of Octanol-Water Partition Coefficients Using Group Contribution Solvation Model. *Ind. Eng. Chem. Res.* **1999**, *38*, 4081.

(11) Wang, S.; Lin, S. T.; Chang, J.; Goddard, W. A.; Sandler, S. I. Application of the COSMO-SAC-BP Solvation Model to Predictions of Normal Boiling Temperatures for Environmentally Significant Substances. *Ind. Eng. Chem. Res.* **2006**, *45*, 5426.

(12) Constantinescu, D.; Klamt, A.; Geanač, D. Vapor-Liquid Equilibrium Prediction at High Pressure Using Activity Coefficients at Infinite Dilution from COSMO-Type Methods. *Fluid Phase Equilib.* **2005**, *231*, 231.

(13) Eckert, F.; Klamt, A. Validation of the COSMO-RS Method: Six Binary Systems. *Ind. Eng. Chem. Res.* **2001**, *40*, 2371.

(14) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. First Principles Calculations of Aqueous pKa Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the pKa Scale. *J. Phys. Chem.* **2003**, *107*, 9380.

(15) Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans.* **1993**, *2*, 799.

(16) Lin, S. T.; Sandler, S. I. Infinite Dilution Activity Coefficients from *Ab Initio* Solvation Calculations. *AIChE J.* **1999**, *45*, 2606.

(17) Mathias, P.; Lin, S. T.; Song, Y.; Chen, C.-C.; Sandler, S. I. In *Phase-Equilibrium Predictions for Hydrogen-Bonding Systems from a New Expression for COSMO Solvation Models*, AIChE Annual Meeting, Indianapolis, IN, 2002. Accessible from www.aspentech.com/publication_files/TP50.pdf.

(18) Putnam, R.; Klamt, A.; Taylor, R.; Eckert, F.; Schiller, M. Prediction of Infinite Dilution Activity Coefficients Using COSMO-RS. *Ind. Eng. Chem. Res.* **2003**, *42*, 3635.

(19) Wang, S.; Stubbs, J. M.; Siepmann, J. I.; Sandler, S. I. Effects of Conformational Distributions on Sigma Profiles in COSMO Theories. *J. Phys. Chem.* **2005**, *109*, 11285.

(20) Ben-Naim, A. *Solvation Thermodynamics*; Plenum Press: New York, 1987.

(21) Prausnitz, J. M.; Lichtenthaler, R. N.; Azevedo, E. G. d., *Molecular Thermodynamics of Fluid-Phase Equilibria*, 3rd ed.; Prentice Hall: New Jersey, 1999.

(22) Marrero, J.; Abildskov, J. *Solubility and Related Properties of Large Complex Chemicals, Part 1: Organic Solutes Ranging from C4 to C40*. DECHEMA: 2003; Vol. XV.

(23) Joback, K. G. A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques. Thesis M.S. Massachusetts Institute of Technology Dept. of Chemical Engineering, 1984.

(24) Joback, K. G.; Reid, R. C. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Commun.* **1987**, *57*, 233.

(25) Constantinou, L.; Gani, R. New Group-Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 1697.

(26) Constantinou, L.; Gani, R.; Oconnell, J. P. Estimation of the Acentric Factor and the Liquid Molar Volume at 298-K Using a New Group-Contribution Method. *Fluid Phase Equilib.* **1995**, *103*, 11.

(27) Yalkowsky, S. H.; Dannenfelser, R. M.; Myrdal, P.; Simamora, P.; Mishra, D. Unified Physical Property Estimation Relationships (Upper). *Chemosphere* **1994**, *28*, 1657.

(28) Krzyzaniak, J. F.; Myrdal, P. B.; Simamora, P.; Yalkowsky, S. H. Boiling-Point and Melting-Point Prediction for Aliphatic, Non-Hydrogen-Bonding Compounds. *Ind. Eng. Chem. Res.* **1995**, *34*, 2530.

(29) Zhao, L. W.; Yalkowsky, S. H. A Combined Group Contribution and Molecular Geometry Approach for Predicting Melting Points of Aliphatic Compounds. *Ind. Eng. Chem. Res.* **1999**, *38*, 3581.

(30) Chickos, J. S.; Acree, W. E. Estimating Solid-Liquid Phase Change Enthalpies and Entropies. *J. Phys. Chem. Ref. Data* **1999**, *28*, 1535.

(31) Chickos, J. S.; Nichols, G.; Ruelle, P. The Estimation of Melting Points and Fusion Enthalpies Using Experimental Solubilities, Estimated Total Phase Change Entropies, and Mobile Order and Disorder Theory. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 368.

(32) Case, D. A.; Darden, T. A.; III, T. E. C.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Godlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *Amber 8*, University of California, San Francisco: San Francisco, CA, 2004.

(33) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. Amber, a Package of Computer-Programs for Applying Molecular Mechanics, Normal-Mode Analysis, Molecular-Dynamics and Free-Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1.

(34) Accelrys, *MS Modeling Getting Started, Release 4.0*; Accelrys Software Inc.: San Diego, CA, January, 2007.

(35) Wang, S. Private Communication. University of Delaware: Newark, DE, 2005.

(36) Klamt, A. Prediction of the Mutual Solubilities of Hydrocarbons and Water with COSMO-RS. *Fluid Phase Equilib.* **2003**, *206*, 223.

(37) Klamt, A.; Eckert, F.; Hornig, M.; Beck, M.; Bürger, T. Prediction of Aqueous Solubility of Drugs and Pesticides with COSMO-RS. *J. Comput. Chem.* **2001**, *23*, 275.

(38) Oleszek-Kudlak, S.; Grabda, M.; Shibata, E.; Eckert, F.; Nakamura, T. Application of the Conductor-Like Screening Model for Real Solvents for Prediction of the Aqueous Solubility of Chlorobenzenes Depending on Temperature and Salinity. *Environ. Toxicol. Chem.* **2005**, *24*, 1368.

(39) Chen, C. C.; Crafts, P. A. Correlation and Prediction of Drug Molecule Solubility in Mixed Solvent Systems with the Nonrandom Two-Liquid Segment Activity Coefficient (NRTL-SAC) Model. *Ind. Eng. Chem. Res.* **2006**, *45*, 4816.

(40) Chen, C. C.; Song, Y. H. Extension of Nonrandom Two-Liquid Segment Activity Coefficient Model for Electrolytes. *Ind. Eng. Chem. Res.* **2005**, *44*, 8909.

(41) Chen, C. C.; Song, Y. H. Solubility Modeling with a Nonrandom Two-Liquid Segment Activity Coefficient Model. *Ind. Eng. Chem. Res.* **2004**, *43*, 8354.

(42) Grensemann, H.; Gmehling, J. Performance of a Conductor-Like Screening Model for Real Solvents Model in Comparison to Classical Group Contribution Methods. *Ind. Eng. Chem. Res.* **2005**, *44*, 1610.

(43) Bustamante, P.; Romero, S.; Pena, A.; Escalera, B.; Reillo, A. Enthalpy-Entropy Compensation for the Solubility of Drugs in Solvent Mixtures: Paracetamol, Acetanilide, and Nalidixic Acid in Dioxane-Water. *J. Pharm. Sci.* **1998**, *87*, 1590.

(44) Manzo, R. H.; Ahumada, A. A. Effects of Solvent Medium on Solubility .5. Enthalpic and Entropic Contributions to the Free-Energy Changes of Disubstituted Benzene-Derivatives in Ethanol Water and Ethanol Cyclohexane Mixtures. *J. Pharm. Sci.* **1990**, *79*, 1109.

(45) Granberg, R. A.; Rasmuson, A. C. Solubility of Paracetamol in Binary and Ternary Mixtures of Water Plus Acetone Plus Toluene. *J. Chem. Eng. Data* **2000**, *45*, 478.

(46) Dickhut, R. M.; Andren, A. W.; Armstrong, D. E. Naphthalene Solubility in Selected Organic Solvent-Water Mixtures. *J. Chem. Eng. Data* **1989**, *34*, 438.

(47) Romero, S.; Reillo, A.; Escalera, B.; Bustamante, P. The Behavior of Paracetamol in Mixtures of Amphiprotic and Amphiprotic-Aprotic Solvents. Relationship of Solubility Curves to Specific and Nonspecific Interactions. *Chem. Pharm. Bull.* **1996**, *44*, 1061.

(48) Bustamante, P.; Ochoa, R.; Reillo, A.; Escalera, J. B. Chameleonic Effect of Sulfanilamide and Sulfamethazine in Solvent Mixtures - Solubility Curves with 2 Maxima. *Chem. Pharm. Bull.* **1994**, *42*, 1129.

(49) Chen, C. C. A Segment-Based Local Composition Model for the Gibbs Energy of Polymer-Solutions. *Fluid Phase Equilib.* **1993**, *83*, 301.

(50) Chen, C. C. Private Communication. Aspen Technology, Inc.: Cambridge, MA, 2006.

(51) Klamt, A. Prediction of the Mutual Solubilities of Hydrocarbons and Water with COSMO-RS. *Fluid Phase Equilib.* **2003**, *206*, 223.

(52) Panayiotou, C. Equation of State Models and Quantum Mechanical Calculations. *Ind. Eng. Chem. Res.* **2003**, *42*, 1495.

Received for review August 13, 2007
Revised manuscript received October 23, 2007
Accepted October 26, 2007

IE0711022