molecular pharmaceutics

articles

GSSI, a General Model for Solute−Solvent Interactions. 1. Description of the Model

Felix Deanda,*,† Karl M. Smith,‡ Jie Liu,‡ and Robert S. Pearlman*,‡,§

Computational, Analytical and Structural Sciences, GlaxoSmithKline, Five Moore Drive, Research Triangle Park, North Carolina 27709, Laboratory for the Development of Computer-Assisted Drug Discovery Software, College of Pharmacy, University of Texas, Austin, Texas 78712, and Optive Research, Inc., 7000 Mopac Expressway, Austin, Texas 78731

Received September 17, 2003

Abstract: A novel, semiempirical approach for the general treatment of solute−solvent interactions (GSSI) was developed to enable the prediction of solution-phase properties (e.g., free energies of desolvation, partition coefficients, and membrane permeabilities). The GSSI approach is based on the principle that all solution-phase processes can be modeled in terms of one or more gas-to-solution transfer processes. The free energy of each gas-to-solution transfer process is calculated as the sum of the free energy of cavity formation and the free energy of solute−solvent interaction. The solute's contributions to these free energies are modeled on the basis of various quantities computed from the solute's three-dimensional (3D) structure, whereas the solvent's contributions are modeled by empirically determined regression coefficients. More specifically, the free energy of cavity formation is modeled on the basis of the total solvent-accessible surface area of the solute. The enthalpy of solute−solvent interaction is modeled on the basis of intermolecular interaction potentials calculated at many uniformly distributed points on the solvent-accessible surface of the solute. The entropy of solute−solvent interaction is modeled on the basis of an effective number of rotatable bonds in the solute and by the regression coefficients characteristic of the solvent. The potential utility of the GSSI approach was demonstrated by modeling the free energy of gas-to-solution transfer for 111 solutes in water, 250 solutes in hexadecane, and 84 solutes in octanol.

Keywords: Solution-phase properties; solute−solvent interactions; solvation model; free energy of gas-to-solution transfer

1. Introduction

Accurate and efficient models of solution-phase properties can play important roles in computer-assisted drug design (CADD). Estimates of the free energy of partial desolvation of the drug and receptor are critically important for modeling the overall free energy of drug−receptor interaction. Estimates of a partition coefficient and/or permeability can be of great value in predicting ADME properties of a potential drug candidate. Therefore, it is not surprising that solute−solvent interactions have been modeled for more than 100 years at various levels of theoretical rigor.

At an empirical level, the solute is represented in relatively crude terms and its interaction with the solvent characterized in a purely empirical fashion. For example, in group-contribution methods, the solute is described by counting the number of occurrences of predefined chemical substructures or "fragments" for which presumably additive contributions to the property of interest have been empirically determined.[1−13] In principle, the fragment values must

* To whom correspondence should be addressed. F.D.: telephone, (919) 483-9482; fax, (919) 483-6053; e-mail, fd69145@gsk.com. R.S.P.: telephone, (512) 471-3383; fax, (512) 471-7474; e-mail, pearlman@list.phr.utexas.edu.

† GlaxoSmithKline.

‡ University of Texas.

§ Optive Research, Inc.

(1) Rytting, J. H.; Huston, L. P.; Higuchi, T. *J. Pharm. Sci.* **1978**, *67*, 615−618.

10.1021/mp034009u CCC: $27.50
Published on Web 10/23/2003

characterize not only the nature of the corresponding solute fragments but also the nature of the solvent in the vicinity of the fragments and the nature of all types of intermolecular interactions that might occur between the fragments and the solvent. Typically, entropic contributions to the free energy of solute−solvent interaction are not explicitly addressed. As a result, the fragment values must also somehow account for changes in the conformational entropy of rotatable bonds between fragments (despite the fact that a given fragment might be bonded to a wide variety of other fragments).

At a far more rigorous level, molecular dynamics (MD) has been applied to the study of solute−solvent interactions.[14−18] Not only is the solute represented in a more rigorous fashion compared to empirical methods, but the solvent is also treated explicitly as an ensemble of individual molecules. In principle, molecular dynamics is the most appropriate way to account for solvation effects in the calculation of solution-phase properties. In practice, however, this approach suffers numerous limitations. Most obvious is the fact that MD simulations require extremely large amounts of computer resources [i.e., central processing unit (CPU) time, memory, and disk storage]. The solvent ensemble usually includes hundreds or thousands of solvent molecules, depending on the size of the solute, requiring very extensive calculations of solute−solvent and solvent−solvent interactions. Approximations have been introduced into the potential energy functions to help reduce computational cost. However, this has led to interaction energies that are of questionable quantitative value. Finally, while entropic effects can be addressed in a qualitative fashion, they are essentially impossible to model in a quantitatively reliable manner. In light of these limitations, molecular dynamics cannot be used to model solution-phase properties for CADD purposes.

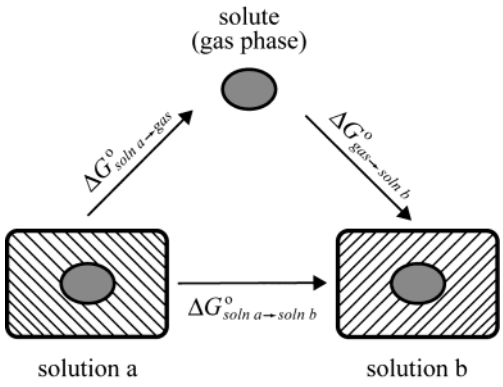
Our semiempirical general treatment of solute−solvent interactions (GSSI) represents a useful compromise between the insufficient rigor of empirical approaches and the excessive rigor of MD methods. Our approach is similar to molecular dynamics with respect to modeling the enthalpic aspects of solute−solvent interaction and attempts to model the entropic contributions in a more practical manner. GSSI describes the solute's electronic distribution and potential for intermolecular interaction in relatively rigorous terms, indeed, far more rigorously than in MD simulations. Also, the solvent molecules in the primary solvation layer are treated explicitly, much like in molecular dynamics. Solvent molecules further from the solute surface are characterized implicitly.

The GSSI approach is based on the principle that all solution-phase processes can be modeled in terms of one or more gas-to-solution transfer processes.[19−23] Figure 1 illustrates this fundamental principle using the partitioning process as an example. In the first step, the solute is transferred from solution *a* into the gas phase. In the second step, the solute is transferred from the gas phase into solution *b*. Since free energy is a path-independent, thermodynamic state function, the partitioning process can be modeled as the difference between two free energies of gas-to-solution transfer.

$$\Delta G^{\circ}_{\text{soln } a \rightarrow \text{soln } b} = \Delta G^{\circ}_{\text{soln } a \rightarrow \text{gas}} + \Delta G^{\circ}_{\text{gas} \rightarrow \text{soln } b} = \Delta G^{\circ}_{\text{gas} \rightarrow \text{soln } b} - \Delta G^{\circ}_{\text{gas} \rightarrow \text{soln } a} \quad (1)$$

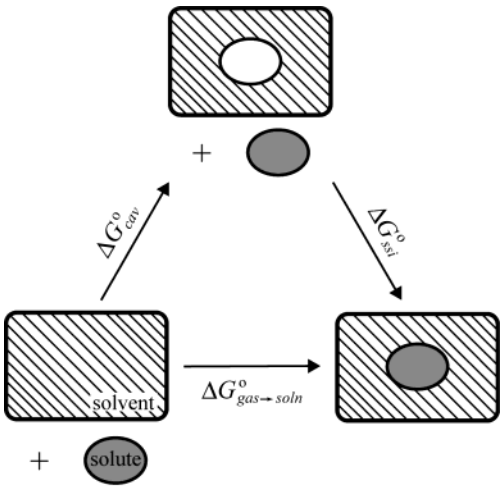
Thus, if we can reliably calculate the free energy change associated with gas-to-solution transfer, we should, in principle, be able to predict solution-phase properties such as the free energy of desolvation, partitioning, and membrane permeability. In this paper, we consider the physical basis

(2) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563−595.
(3) Hine, J.; Mookerjee, P. K. *J. Org. Chem.* **1975**, *40*, 292−298.
(4) Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 405−412.
(5) Ghose, A.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7*, 565−577.
(6) Abraham, R. J.; Hudson, B. D.; Kermode, M. W.; Mines, J. R. *J. Chem. Soc., Faraday Trans. 1* **1988**, *84*, 1911−1917.
(7) Suzuki, T.; Kudo, Y. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 155−198.
(8) Sangster, J. Methods of Calculating Partition Coefficients. In *Octanol−Water Partition Coefficients: Fundamentals and Physical Chemistry*; Fogg, P. G. T., Ed.; John Wiley & Sons: New York, 1997; pp 113−122.
(9) Altomare, C.; Carotti, A.; Trapani, G.; Liso, G. *J. Pharm. Sci.* **1997**, *86*, 1417−1425.
(10) Nuñez, F. A. A.; Yalkowsky, S. H. *J. Pharm. Sci.* **1997**, *86*, 1187−1189.
(11) Kakoulidou, A. T.; Panderi, I.; Csizmadia, F.; Darvas, F. *J. Pharm. Sci.* **1997**, *86*, 1173−1179.
(12) Masuda, T.; Jikihara, T.; Nakamura, K.; Kimura, A.; Takagi, T.; Fujiwara, H. *J. Pharm. Sci.* **1997**, *86*, 57−63.
(13) Meylan, W. M.; Howard, P. H. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(14) Lybrand, T. P. Computer Simulation of Biomolecular Systems Using Molecular Dynamics and Free Energy Perturbation Methods. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 295−320.
(15) Grant, G. H.; Richards, W. G. Statistical Mechanics. In *Computational Chemistry*, 1st ed.; Grant, G. H., Richards, W. G., Eds.; Oxford University Press: Oxford, England, 1995; pp 46−61.
(16) Rossky, P. J.; Karplus, M. *J. Am. Chem. Soc.* **1979**, *101*, 1913−1937.
(17) Lee, S. H.; Rossky, P. J. *J. Chem. Phys.* **1994**, *100*, 3334−3345.
(18) Linse, P. *J. Am. Chem. Soc.* **1990**, *112*, 1744−1750.
(19) Pearlman, R. S. Molecular Surface Area and Volume: Their Calculation and Use in Predicting Solubilities and Free Energies of Desolvation. In *Partition Coefficient: Determination and Estimation*; Dunn, W. J., III, Block, J. H., Pearlman, R. S., Eds.; Pergamon Press: New York, 1986; pp 3−20.
(20) Pearlman, R. S. Molecular Surface Areas and Volumes and Their Use in Structure/Activity Relationships. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980; pp 321−347.
(21) Skell, J. M. Software Tools for Computer-Assisted Molecular Design. Ph.D. Thesis, The University of Texas, Austin, TX, 1993.
(22) Escobar-Valderrama, J. L. A General Model for the Treatment of Solute−Solvent Interactions. Ph.D. Thesis, The University of Texas, Austin, TX, 1996.
(23) Amidon, G. L.; Pearlman, R. S.; Anik, S. T. *J. Theor. Biol.* **1979**, *77*, 161−170.



$$\Delta G^{o}_{soln\ a \rightarrow soln\ b} = \Delta G^{o}_{soln\ a \rightarrow gas} + \Delta G^{o}_{gas \rightarrow soln\ b}$$
$$= \Delta G^{o}_{gas \rightarrow soln\ b} - \Delta G^{o}_{gas \rightarrow soln\ a}$$

Figure 1. Thermodynamic analysis of the partitioning process. Partitioning modeled as the difference between two gas-to-solution transfer processes.



$$\Delta G^{o}_{gas \rightarrow soln} = \Delta G^{o}_{cav} + \Delta G^{o}_{ssi}$$

Figure 2. Thermodynamic analysis of the gas-to-solution transfer process. The first step involves the formation of a cavity within the solvent. The second step involves placing the solute in the cavity to form the solution.

for modeling the free energy of gas-to-solution transfer. To validate our approach, we apply it to the prediction of free energies of gas-to-aqueous solution transfer, free energies of gas-to-hexadecane transfer, and free energies of gas-to-octanol transfer.

2. Theory of the Gas-to-Solution Transfer Process

The transfer of solute from gas to solution can, itself, be modeled as a two-step process.[19−23] As illustrated in Figure 2, the first step involves the formation of a cavity within the solvent that is sufficient in size to accommodate the solute. The second step involves placing the solute into the cavity to form the solution. Such a hypothetical two-step process is, once again, based on the fact that free energy is a thermodynamic state function and is, therefore, path-independent. Accordingly, the free energy of gas-to-solution transfer, $\Delta G^{\circ}_{gas \rightarrow soln}$, can be calculated as the sum of the free energy of cavity formation, ΔG°_{cav}, and the free energy of solute−solvent interaction, ΔG°_{ssi}.

$$\Delta G^{\circ}_{gas \rightarrow soln} = \Delta G^{\circ}_{cav} + \Delta G^{\circ}_{ssi} \quad (2)$$

2.1. Free Energy of Cavity Formation. Cavity formation is a thermodynamically unfavorable process. Enthalpically, cavity formation is an unfavorable process because solvent−solvent interactions must be broken to form a cavity within the solvent. Clearly, the enthalpy change of cavity formation, ΔH°_{cav}, will be related to the average strength of the solvent−solvent interaction and the size of the cavity (computed on the basis of the solvent-accessible surface of the solute). Entropically, cavity formation is also an unfavorable process. Solvent molecules near the surface of the cavity become quasi-structured when they experience an asymmetric intermolecular force field. This asymmetric field results from having "bulk" solvent on one side of the solvent molecules near the cavity surface and an empty cavity on the other. The end result is an overall decrease in the entropy of the system. Like ΔH°_{cav}, the entropy change of cavity formation, ΔS°_{cav}, will depend to some extent on the strength of the solvent−solvent interaction, but primarily upon the size of the cavity (solute).

Although ΔH°_{cav} and ΔS°_{cav} are difficult to model individually, modeling ΔG°_{cav} is a much simpler task. Note that ΔH°_{cav} and ΔS°_{cav} are both related to the size of the solute, suggesting that ΔG°_{cav} could be modeled on the basis of either the solute's surface area or volume. In fact, Hermann[24] has shown that the free energy of cavity formation for a microscopic, nonspherical cavity can be estimated as

$$\Delta G^{\circ}_{cav} = C^{cav}\mathrm{TSA}^{acc} \quad (3)$$

where TSA^{acc} is the total solvent-accessible surface area of the solute and C^{cav} represents the effective, curvature-corrected surface tension of the solvent and accounts for the effects of solvent−solvent interaction. TSA^{acc} is computed within the GSSI software package using the Savol3 program[21,25] developed by Pearlman, Skell, and Deanda at the University of Texas. The Savol3 program computes the van der Waals or solvent-accessible surface area and/or volume (and atomic contributions thereto) of a solute in a given conformation.

2.2. Free Energy of Solute−Solvent Interaction. Various empirical and semiempirical solvation models have been developed in which ΔG°_{ssi} is modeled as a sum of atomic or group contributions.[2,3,5−7] However, the nonadditivity of solvation effects for neighboring charged atoms or polar

(24) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754−2759.
(25) Pearlman, R. S.; Skell, J. M.; Deanda, F. *SAVOL3: A Program for the Atomic Partitioning of the Surface Area and Volume of Molecular Structures*; The University of Texas: Austin, TX, 1999.

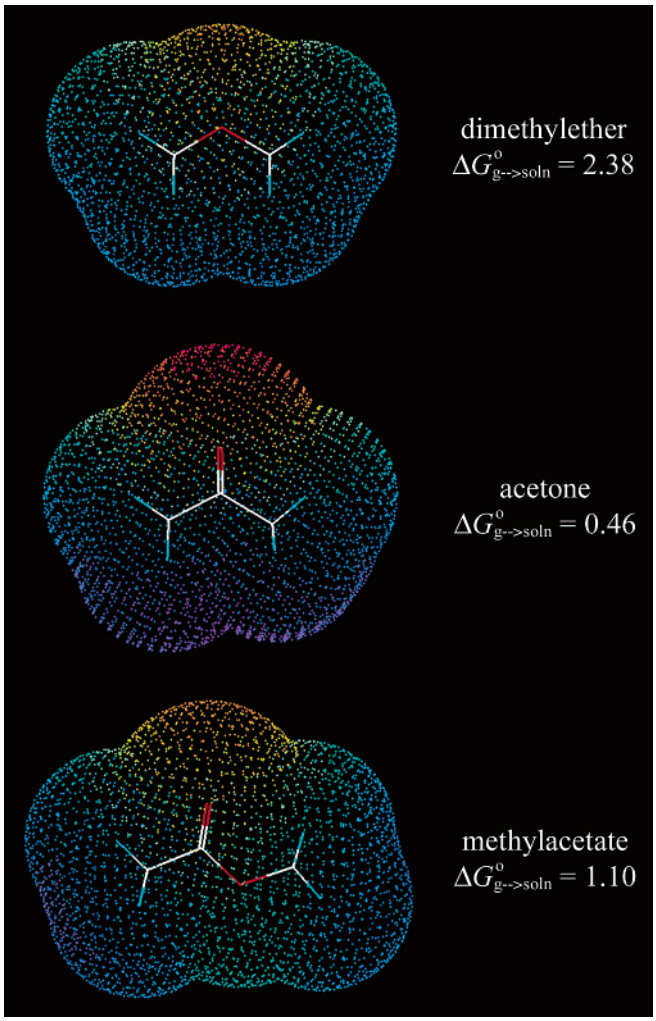


Figure 3. Electrostatic potentials color-coded onto the solvent-accessible surfaces of dimethyl ether, acetone, and methyl acetate. The free energies of gas-to-aqueous solution transfer are in kilocalories per mole.

functional groups has required that these additivity schemes introduce many correction terms to account for "proximity effects".[2,3] As an example, consider the hydrophilic characters of dimethyl ether, acetone, and methyl acetate (Figure 3). If we accept that the hydrophilicity of a solute is determined, to a large extent, by its number of potential H-bond forming atoms or functional groups, we can reasonably expect methyl acetate (with its two electronegative oxygen atoms) to be the most hydrophilic among the three solutes. However, because of proximity effects, the hydrophilic character of methyl acetate is only intermediate between that of acetone and that of dimethyl ether.

In efforts to understand the relative hydrophilicities of dimethyl ether, acetone, and methyl acetate, we conducted a simple analysis of the electrostatic potentials calculated at the solvent-accessible surfaces of the solutes. These surface potentials were globally scaled and color-coded with colors toward red corresponding to negative electrostatic potential and colors toward blue corresponding to positive electrostatic potential (Figure 3). Note, in particular, the surface potentials over the oxygen atoms of the solutes. Acetone has the most negative surface potential, while dimethyl ether has the least negative. Significantly, the surface potential over the carbonyl oxygen atom of methyl acetate is intermediate relative to those of the two other solutes. On the basis of these observations, we chose to abandon an atom-centered approach to modeling ΔG°_{ssi} within GSSI in favor of a solvent-accessible dot surface model, which implicitly accounts for proximity effects.

In modeling ΔG°_{ssi}, GSSI takes into account not only the enthalpic aspects (ΔH°_{ssi}) of solute−solvent interaction (often the only aspects explicitly considered by other solvation models) but also the various entropic contributions (ΔS°_{ssi}) from both the solute and the solvent. These entropic contributions are discussed in detail in section 2.2.2. For the moment, we turn our attention to ΔH°_{ssi} or, more specifically, to the energy change associated with the solute−solvent interaction, ΔE°_{ssi}, given that the Δ(PV) component of ΔH°_{ssi} is essentially constant for a series of solutes.[19−22]

2.2.1. Energy Change of Solute−Solvent Interaction. In the quantum mechanical treatment of intermolecular interactions, the interaction energy can be calculated by the application of Rayleigh−Schrödinger perturbation theory. One of the most significant features of perturbation methods is that, in a natural way, these methods decompose the interaction energy into an infinite sum of perturbation corrections, each having defined physical meaning. Typically, first- and second-order terms not involving electron exchange (E^{10} and E^{20}) are the only terms explicitly considered. As a result, the modes of intermolecular attraction that are generally taken into account include the electrostatic, polarization, and dispersion interactions.[26−30] Intermolecular (exchange) repulsion is addressed in the first non-zero (first second-order) exchange term (E^{12}). Pearlman[26] demonstrated how the exchange repulsion energy could be conveniently computed directly from E^{12}, but it is normally approximated using the familiar, empirical r^{-12} functional form.

In our GSSI approach, ΔE°_{ssi} is calculated by adding contributions from four modes of interaction: the electrostatic, the polarization of solute by solvent, the polarization of solvent by solute, and the dispersion interactions. Significantly, these four modes of interaction primarily account

(26) Pearlman, R. S. Intermolecular Interaction Energies. Ph.D. Thesis, University of Michigan, Ann Arbor, MI, 1975.
(27) Claverie, P.; Rein, R. *Int. J. Quantum Chem.* **1969**, *3*, 537−551.
(28) Ratajczak, H.; Orville-Thomas, W. J. On Some Problems of Molecular Interactions. In *Molecular Interactions*; Ratajczak, H., Orville-Thomas, W. J., Eds.; John Wiley & Sons: New York, 1980; Vol. 1, pp 1−20.
(29) Rao, C. N. R.; Sudhindra, B. S.; Ratajczak, H.; Orville-Thomas, W. J. Semi-Empirical Quantum Mechanical Studies of Molecular Complexes. In *Molecular Interactions*; Ratajczak, H., Orville-Thomas, W. J., Eds.; John Wiley & Sons: New York, 1980; Vol. 1, pp 67−87.
(30) Rigby, M.; Smith, E. B.; Wakeham, W. A.; Maitland, G. C. The Nature of Intermolecular Forces. In *The Forces between Molecules*; Rigby, M., Smith, E. B., Wakeham, W. A., Maitland, G. C., Eds.; Oxford Science Publications: New York, 1986; pp 1−35.

for H-bond interactions, and therefore theoretically, H-bonding should not require special attention. However, unavoidable errors in approximating the four nonexchange terms coupled with the effects of higher-order terms in the perturbation expansion require a semiempirical correction to account for H-bonding. We do not explicitly address solute−solvent repulsion because, to a very good approximation, the fluid solvent will pack around all solutes at a van der Waals contact distance. Thus, to a very good approximation, the solute−solvent repulsion will be quite small and will be essentially the same for all solutes in a given solvent. Thus, the exchange repulsion energy will not be useful in distinguishing solution-phase properties of one solute from another. The only exception to the foregoing argument occurs for H-bonding between solute and solvent and is the primary reason for the semiempirical H-bonding correction mentioned above.

To calculate the energy of solute−solvent interaction, the solute is described as rigorously as possible while the solvent is treated empirically. The solute's electronic distribution is represented using "multipoint" charges and polarizabilities, rather than atom-centered charges and polarizabilities as in other QSPR models explicitly addressing electrostatic interactions. The multipoints, which reflect the orbital electron densities and polarizabilities in a bond-directed hybrid basis (as opposed to the usual s, p_x, p_y, p_z, ..., AO basis), are provided by our HSCF program[31] using the AM1 Hamiltonian. The multipoints are positioned at the expectation value of each hybrid orbital lobe and at the nucleus of each atom. Hence, there are nine points per non-hydrogen atom and two points per hydrogen atom of the solute. In contrast, we assume that the electronic properties of the solvent can be adequately represented by an effective dipole moment with a magnitude of $|\vec{\mu}|$ and an effective polarizability with a magnitude of α_s positioned at the center of a "solvent sphere" of appropriate radius.

(a) Solvent Configuration. Before we describe the development of the functional forms for the different modes of interaction, we must first introduce the concept of "solvent configuration". In the calculation of the solute−solvent interaction energies, solvent molecules are first positioned at the solvent-accessible surface of the solute. The many possible solvent positions are represented by "dots" (as in Figure 3) on a specially constructed, uniform molecular dot surface, which is generated within GSSI by the Uniform Dot Surface (UDS) algorithm of Kim and Pearlman.[32,33] Although the exact number of dots on the solute surface depends on the size of the solute and the user-specified dot density, the total number of dots, N_{dots}, will typically be in the range of 2000−3000 for small drug molecules.

Each dot on the solute surface represents a position that could possibly be occupied by a solvent molecule. Since the dots are closely spaced and since the solvent sphere is (usually) on the order of 3.0 Å in diameter, only small subsets of dots can be occupied at any given time (by closely packed but nonoverlapping solvent spheres). To address this issue, we developed an algorithm which identifies N_{dots} such subsets of solvent-occupied dots which we term solvent configurations. The algorithm uses each dot on the solvent-accessible surface as an initial starting point for construction of a solvent configuration. Thus, a total of N_{dots} solvent configurations are generated for a given solute. Clearly, GSSI will need to evaluate ΔE°_{ssi} by averaging over all solvent configurations.

(b) Electrostatic Interaction. To derive the functional form of the electrostatic interaction, we first consider a solvent dipole $\vec{\mu}_d(\theta,\phi)$ positioned at dot d on the solute surface with fixed orientation (θ,ϕ). The interaction energy associated with this fixed dipole and the solute is given by

$$E^{elec}_{d,(\theta,\phi)} = \vec{\mu}_d(\theta,\phi)\cdot\vec{\xi}_d \tag{4}$$

where $\vec{\xi}_d$ is the electric field generated by the solute charges at solvent position d, and is calculated as

$$\vec{\xi}_d = \sum_i \frac{q_i \vec{r}_{di}}{||\vec{r}_{di}||^3} \tag{5}$$

In eq 5, q_i is the ith multipoint charge of the solute, $\vec{r}_{di}$ is the vector from q_i to solvent position d, and the sum over index i is over all solute multipoints.[26]

Next, we consider the case where the solvent dipole is free to rotate about solvent position d. If the charge−dipole interaction is averaged over all possible solvent dipole orientations, treating them all as equally probable, the net result for the interaction energy is obviously zero. However, such averaging is unrealistic, since not all orientations are equally probable. Indeed, each solvent orientation (θ,ϕ) occurs with a probability that is proportional to the associated Boltzmann factor, $\exp[\Delta E^{elec}_{d,(\theta,\phi)}/RT]$, and the Boltzmann-averaged electrostatic energy $\langle E^{elec}_d \rangle$ reflects this

$$\langle E_d{}^{elec} \rangle = \frac{\sum_\theta \sum_\phi E^{elec}_{d,(\theta,\phi)} \exp[-\Delta E^{elec}_{d,(\theta,\phi)}/RT]}{\sum_\theta \sum_\phi \exp[-\Delta E^{elec}_{d,(\theta,\phi)}/RT]} \tag{6}$$

In this equation, $\Delta E^{elec}_{d,(\theta,\phi)} = E^{elec}_{d,(\theta,\phi)} - E^{elec}_{min}$, where E^{elec}_{min} is the minimum (i.e., most attractive) interaction energy between the solute and the solvent molecule. Also, R is the universal gas constant, and T is the absolute temperature of the system in kelvin.

Since there are an infinite number of solvent dipole orientations, the Boltzmann-averaged electrostatic energy is

(31) Smith, K. M.; Pearlman, R. S. *HSCF: A Standalone and C-callable MO Information-server*; The University of Texas: Austin, TX, 1998.

(32) Brusniak, M.-Y. K. Development and Application of Software for CADD. Ph.D. Thesis, The University of Texas, Austin, TX, 1996.

(33) Brusniak, M.-Y. K.; Pearlman, R. S. *UDS/QDS: Novel Algorithms for the Generation of Uniform or Quick Molecular Dot Surfaces*; The University of Texas: Austin, TX, 1999.

not really a ratio of sums but, rather, a ratio of double integrals over the spherical polar coordinates, θ and ϕ. If we multiply eq 6 by 1 in the form of $(r^2 \sin\theta\ \Delta\theta\ \Delta\phi)/(r^2 \sin\theta\ \Delta\theta\ \Delta\phi)$, where r is half the distance separating the dipole partial charges, we introduce the surface area element needed to integrate the equation. If we then substitute eq 4 into eq 6 and take the limit as $\Delta\theta$ and $\Delta\phi \rightarrow 0$, we can write

$$\langle E_d^{\text{elec}} \rangle = \frac{\int_\theta \int_\phi \vec{\mu}_d(\theta,\phi)\cdot\vec{\xi}_d \exp\{-[\vec{\mu}_d(\theta,\phi)\cdot\vec{\xi}_d - E_{\min}^{\text{elec}}]/RT\} r^2 \sin\theta \, d\theta \, d\phi}{\int_\theta \int_\phi \exp\{-[\vec{\mu}_d(\theta,\phi)\cdot\vec{\xi}_d - E_{\min}^{\text{elec}}]/RT\} r^2 \sin\theta \, d\theta \, d\phi} \tag{7}$$

Evaluating eq 7 yields

$$\langle E_d^{\text{elec}} \rangle = RT - ||\vec{\mu}||\ ||\vec{\xi}_d|| \coth(||\vec{\mu}||\ ||\vec{\xi}_d||/RT) \tag{8}$$

where $||\vec{\mu}||$ is the magnitude of the solvent dipole and $||\vec{\xi}_d||$ is the magnitude of the electric field generated by the solute charges at solvent position d.

Note that, to simplify the evaluation of $\langle E_d^{\text{elec}} \rangle$, we made the assumption that the orientation of the solvent dipole at dot d is independent of the relative position and orientation of all other solvent molecules. Obviously, this assumption leads to a rather rough approximation of $\langle E_d^{\text{elec}} \rangle$. In an effort to include the effects of solvent−solvent interactions, we developed the following novel approach for the approximate treatment of solvent "structuring" around the solute surface. First, for each solvent position d of a given solvent configuration g, the Boltzmann-averaged electrostatic energy is computed using eq 8. Next, the $N_{\text{solv},g}$ solvent positions (or dots) of solvent configuration g are ranked on the basis of the energy from most to least attractive and then categorized as either "seeds" or "neighbors" according to the following simple procedure.

(1) Label the most attractive dot from the given solvent configuration as a seed.

(2) Identify the dots that are neighbors of the given seed. Label these dots as neighbors. A seed and its neighbors are termed a cluster.

(3) Select the next most attractive dot not already categorized as a seed or neighbor. Label this dot as a seed.

(4) Repeat steps 2 and 3 until all $N_{\text{solv},g}$ dots of the solvent configuration have been categorized.

In our treatment of solvent−solvent interactions, only the solute charges determine the orientations of the seeds. In contrast, both the solute charges and the orientations of the dipoles of neighboring solvent molecules determine the orientations of the neighbors. Details with regard to how the dipole orientations of the neighbors are computed will be provided shortly. For now, we turn our attention to the Boltzmann-averaged dipole orientation $\langle \vec{\mu}_d \rangle$ of each seed which is calculated as

$$\langle \vec{\mu}_d \rangle = \frac{\int_\theta \int_\phi \vec{\mu}_d(\theta,\phi) \exp\{-[\vec{\mu}_d(\theta,\phi)\cdot\vec{\xi}_d - E_{\min}^{\text{elec}}]/RT\} r^2 \sin\theta \, d\theta \, d\phi}{\int_\theta \int_\phi \exp\{-[\vec{\mu}_d(\theta,\phi)\cdot\vec{\xi}_d - E_{\min}^{\text{elec}}]/RT\} r^2 \sin\theta \, d\theta \, d\phi} \tag{9}$$

As with the Boltzmann-averaged electrostatic energy given by eq 7, each solvent dipole orientation is weighted according to its associated Boltzmann factor. Evaluating eq 9 yields

$$\langle \vec{\mu}_d \rangle = ||\vec{\mu}|| \left[\coth\left(\frac{||\vec{\mu}||\ ||\vec{\xi}_d||}{RT} \right) - \frac{RT}{||\vec{\mu}||\ ||\vec{\xi}_d||} \right] \left[\frac{-\vec{\xi}_d}{||\vec{\xi}_d||} \right] \tag{10}$$

where all variables are as previously defined. Note that we actually represent the Boltzmann-averaged orientation of a seed by adjusting the magnitude of a vector positioned antiparallel to $\vec{\xi}_d$. The reason (and theoretical justification) for this is that the Boltzmann-averaged orientation of a solvent dipole cannot be expressed in terms of the two spherical polar coordinates, θ and ϕ, because of the degeneracy in the electrostatic energy as a function of the polar coordinate angle ϕ. However, since we are actually interested in the *effect* of averaging the orientation and not interested in the averaged orientation *per se*, we can express the effect by adjusting the magnitude of the solvent dipole.

For the solvent positions categorized as neighbors, we first consider the set of neighbors grouped with the most attractive seed. Once their Boltzmann-averaged orientations are established, we then proceed to the next set of neighbors grouped with the second-most attractive seed, and so on. For any given neighbor positioned at dot d', its Boltzmann-averaged orientation is calculated by taking into account not only the electric field $\vec{\xi}_{d'}$ generated by the solute charges but also the electric field $\vec{\zeta}_{d'}^{\text{s}}$ generated by all seeds and the electric field $\vec{\zeta}_{d'}^{\text{n}}$ generated by all neighbors from clusters previously evaluated (i.e., clusters whose entire set of neighbors have had their Boltzmann-averaged orientations established). Thus, the total electric field $\vec{\xi}_{d'}^{\text{tot}}$ at dot d' is simply computed as the vector sum $\vec{\xi}_{d'}^{\text{tot}} = \vec{\xi}_{d'} + \vec{\zeta}_{d'}^{\text{s}} + \vec{\zeta}_{d'}^{\text{n}}$.

The electric field $\vec{\zeta}_{d'd}$ at dot d' generated by a solvent dipole $\vec{\mu}_d(\theta,\phi)$ positioned at dot d with orientation (θ,ϕ) can be computed as

$$\vec{\zeta}_{d'd} = \frac{\vec{\mu}_d(\theta,\phi)}{||\vec{r}_{d'd}||^3} - \frac{3\vec{r}_{d'd}[\vec{\mu}_d(\theta,\phi)\cdot\vec{r}_{d'd}]}{||\vec{r}_{d'd}||^5} \tag{11}$$

where $\vec{r}_{d'd}$ is the vector from solvent position d to d'.[26] Thus, it follows that $\vec{\zeta}_{d'}^{\text{s}}$ can be calculated as

$$\vec{\zeta}_{d'}^{\text{s}} = \sum_{d \in \{s\}} \vec{\zeta}_{d'd} \tag{12}$$

where $\{s\}$ is the set of all seeds. Similarly, $\vec{\zeta}_{d'}^{\text{n}}$ can be

calculated as

$$\vec{\xi}^{\,\mathrm{n}}_{d'} = \sum_{d\in\{n\}} \vec{\xi}_{d'd} \tag{13}$$

where $\{n\}$ is the set of all previously considered neighbors.

Once $\vec{\xi}^{\,\mathrm{tot}}_{d'}$ has been calculated, the Boltzmann-averaged dipole orientation $\langle\vec{\mu}_{d'}\rangle$ of the given neighbor can be determined in a manner very similar to eq 9 to yield

$$\langle\vec{\mu}_{d'}\rangle = ||\vec{\mu}||\left[\coth\left(\frac{||\vec{\mu}||\ ||\vec{\xi}^{\,\mathrm{tot}}_{d'}||}{RT}\right) - \frac{RT}{||\vec{\mu}||\ ||\vec{\xi}^{\,\mathrm{tot}}_{d'}||}\right]\left(\frac{-\vec{\xi}^{\,\mathrm{tot}}_{d'}}{||\vec{\xi}^{\,\mathrm{tot}}_{d'}||}\right) \tag{14}$$

After establishing the Boltzmann-averaged orientations for all solvent dipoles of solvent configuration *g*, we then calculate the electrostatic energy between the solute and each solvent as

$$E^{\mathrm{elec}}_{d} = \langle\vec{\mu}_{d}\rangle\cdot\vec{\xi}_{d} \tag{15}$$

where $\langle\vec{\mu}_{d}\rangle$ is as defined by eq 10 or 14 depending on whether the solvent was categorized as a seed or neighbor and $\vec{\xi}_{d}$ is as defined by eq 5. Note that, since the dipole orientations for the seeds are determined only by the solute charges, their associated electrostatic energies do not actually need to be recomputed when encountered as seeds in subsequent solvent configurations. Finally, the total electrostatic interaction energy between the solute and solvent configuration *g* is calculated as

$$E^{\mathrm{elec}}_{g} = \sum_{d\in g}^{N_{\mathrm{solv},g}} E^{\mathrm{elec}}_{d} \tag{16}$$

where the sum over index *d* is over the $N_{\mathrm{solv},g}$ solvent molecules of solvent configuration *g*.

(c) Polarization of the Solvent by the Solute. The polarization of the solvent by the solute results from the fact that the solute's charge distribution generates an electric field which polarizes the neighboring solvent molecules (i.e., solvent molecules in the first solvation shell). Thus, the polarization energy arises from the interaction between the induced dipole moments of the neighboring solvents and the solute's charge distribution. Having already calculated $\vec{\xi}_{d}$, we can calculate the polarization interaction energy E^{polv}_{d} associated with a solvent molecule positioned at dot *d* as

$$E^{\mathrm{polv}}_{d} = -\frac{1}{2}\alpha_{\mathrm{s}}||\vec{\xi}_{d}||^2 \tag{17}$$

where α_{s} is the effective polarizability of the solvent and $\vec{\xi}_{d}$ is as defined by eq 5.[26] Note that E^{polv}_{d} is independent of the solvent dipole orientation and attractive regardless of that orientation. To calculate the polarization energy E^{polv}_{g} associated with solvent configuration *g*, we simply sum the contributions from the $N_{\mathrm{solv},g}$ solvents.

$$E^{\mathrm{polv}}_{g} = \sum_{d\in g}^{N_{\mathrm{solv},g}} E^{\mathrm{polv}}_{d} \tag{18}$$

(d) Polarization of the Solute by the Solvent. The polarization of the solute by the solvent arises from the fact that the solvent's charge distribution generates an electric field which polarizes the solute. The net result is an attractive interaction between the induced dipole moment of the solute and the solvent's charge distribution. Recall that the Boltzmann-averaged dipole orientation for every solvent molecule of solvent configuration *g* has already been established. Thus, the electric field $\vec{\xi}_{id}$ at solute multipoint *i* generated by the Boltzmann-averaged solvent dipole $\langle\vec{\mu}_{d}\rangle$ positioned at dot *d* can be calculated as

$$\vec{\xi}_{id} = \frac{\langle\vec{\mu}_{d}\rangle}{||\vec{r}_{id}||^3} - \frac{3\vec{r}_{id}(\langle\vec{\mu}_{d}\rangle\cdot\vec{r}_{id})}{||\vec{r}_{id}||^5} \tag{19}$$

where $\vec{r}_{id}$ is the vector from solvent position *d* to multipoint *i*.[26] To calculate the electric field $\vec{\xi}^{\,g}_{i}$ at solute multipoint *i* generated by the solvent dipoles of solvent configuration *g*, we simply sum the contributions from each of the $N_{\mathrm{solv},g}$ solvents.

$$\vec{\xi}^{\,g}_{i} = \sum_{d\in g}^{N_{\mathrm{solv},g}} \vec{\xi}_{id} \tag{20}$$

Finally, the energy due to polarization of the solute by solvent configuration *g*, E^{polu}_{g}, is calculated as

$$E^{\mathrm{polu}}_{g} = -\frac{1}{2}\sum_{i}\alpha_{i}||\vec{\xi}^{\,g}_{i}||^2 \tag{21}$$

where α_{i} is the polarizability associated with multipoint *i* of the solute.[26] Note that the sum over index *i* is over all solute multipoints, and not over the solvent molecules of the solvent configuration as in eqs 16 and 18. Also note that E^{polu}_{g}, like E^{polv}_{d}, is attractive regardless of the solvent dipole orientation.

(e) Dispersion Interaction. The dispersion interaction is a quantum mechanical phenomenon arising from electron correlation effects and is often discussed in terms of the interaction between induced or instantaneous dipoles. The dispersion interaction has been shown to be a major contributor to the attractive interaction between both nonpolar and polar molecules.[28−30] From London's approximate formula,[30,34] we can write the dispersion interaction energy between the solute and a solvent molecule positioned at dot *d* as

$$E^{\mathrm{disp}}_{d} = -\frac{3}{2}E_{\mathrm{ca}}\sum_{i}\frac{\alpha_{\mathrm{s}}\alpha_{i}}{||\vec{r}_{di}||^6} \tag{22}$$

where α_{s}, α_{i}, and $\vec{r}_{di}$ are as previously defined.[26] E_{ca}, which has units of energy, is the closure approximation factor and can be approximated as

(34) London, F. *Trans. Faraday Soc.* **1937**, *33*, 8−26.

$$E_{ca} = I_v I_u / (I_v + I_u) \quad (23)$$

where I_u and I_v are the ionization potentials of the solute and solvent, respectively.[27] The ionization potential of the solute is calculated by our HSCF program,[31] while the effective (i.e., approximate) ionization potential of the solvent must be provided as an input parameter to GSSI. Once the dispersion energies are computed for all solvent molecules of solvent configuration g, the overall dispersion energy, E_g^{disp}, is calculated as

$$E_g^{disp} = \sum_{d \in g}^{N_{solv,g}} E_d^{disp} \quad (24)$$

(f) H-Bond Interaction. Quantum mechanical methods have long been used by investigators in efforts to understand the electronic nature of the so-called H-bond. As before, perturbation theory has provided a means of decomposing the H-bond energy into physically meaningful components. On the basis of results from perturbational calculations, the H-bond energy has been found to include contributions not only from the electrostatic, polarization, and dispersion interactions but also from exchange repulsion and charge-transfer interactions.[28,29,35−37] Although the relative importance of each component varies from one H-bond complex to another, it has been shown that each component makes a significant contribution to the H-bond energy.

The electrostatic, polarization, and dispersion interactions have already been addressed. The two remaining terms arise from the fact that the intermolecular overlap between an H-bond donor (A−H) and acceptor (B) is not negligible. Indeed, experimental evidence has shown that the atoms involved in the H-bond approach each other, closer than the sum of their van der Waals radii. As a result, modes of intermolecular interactions involving electron exchange can no longer be ignored. Unfortunately, exchange repulsion and charge-transfer interactions are *extremely* difficult to treat in a rigorous manner. If we evaluate these modes of interaction using *ab initio* MO methods, the results would be not only computationally expensive but also fairly sensitive to basis set choice and, therefore, somewhat arbitrary. Molecular mechanics force fields address these deficiencies as well as errors in the nonexchange terms by empirically adjusting the van der Waals radii of H-bonded atoms.[38−41] In a similar spirit, we have compiled a set of "effective" radii for atoms involved in H-bonds to account for the proximity observed between H-bond donors and acceptors.

Table 1. Average Interatomic Distances (and standard deviations) between H-Bond Heavy Atoms and H-Bond Hydrogen Atoms and Acceptors

H-bond donor (A−H)	H-bond acceptor (B)	r_{AB} (Å)	r_{HB} (Å)	N_{obs}[a]
$OH_{alcohol}$	O_{water}	2.76 ± 0.08	1.91 ± 0.15	489
$OH_{carbox.acid}$	O_{water}	2.59 ± 0.07	1.69 ± 0.14	152
$NH_{amine/amide}$	O_{water}	2.90 ± 0.12	2.02 ± 0.16	341
$NH_{pyrrole}$	O_{water}	2.80 ± 0.15	1.87 ± 0.18	4
OH_{water}	$N_{3'-amine}$	2.92 ± 0.11	2.08 ± 0.24	69
OH_{water}	$N_{nitrile}$	2.97 ± 0.12	2.21 ± 0.25	13
OH_{water}	$N_{pyridine}$	2.91 ± 0.12	2.06 ± 0.19	98

[a] N_{obs} is the total number of observations found for a given H-bond complex in the CSDS database.[42,43]

Since one of our objectives was to better model experimental free energies of gas-to-aqueous solution transfer for solutes containing H-bond donor/acceptor groups, the Cambridge Structural Database System (CSDS)[42,43] was used to retrieve and analyze crystallographic data for H-bond complexes involving water and compounds from the following chemical classes: alcohols, carboxylic acids, amines, amides, pyrroles, nitriles, and pyridines. Reported in Table 1 are the average interatomic distances (and standard deviations) between H-bond heavy atoms and H-bond hydrogen atoms and acceptors. Note that the amines and amides were combined into a single H-bond donor group. This was due to the average interatomic distances being essentially identical for H-bond complexes involving these classes of compounds.

On the basis of the results from our analysis of the CSDS crystallographic data, we then calculated the effective radii of H-bond donor heavy atoms and H-bond acceptors using

$$R_{A|B}^{eff} = r_{AB} - R_{water} \quad (25)$$

where R_{water} is the effective radius of a water molecule and was assigned a value of 1.50 Å.[19−22] r_{AB} represents one of two possible distance parameters, depending on whether water is the H-bond donor or acceptor. In the case where water is the H-bond acceptor, r_{AB} is the average distance between an H-bond donor heavy atom (A) and the water's oxygen atom (B), and the value calculated from eq 25 is the effective radius of the H-bond donor heavy atom, R_A^{eff}. If, on the other hand, water is the H-bond donor, then r_{AB} is

(35) Jeffrey, G. A. Nature and Properties. In *An Introduction to Hydrogen Bonding*; Truhlar, D. G., Ed.; Oxford University Press: New York, 1997; pp 11−32.
(36) Scheiner, S. Quantum Chemical Framework. In *Hydrogen Bonding, A Theoretical Perspective*; Truhlar, D. G., Ed.; Oxford University Press: New York, 1997; pp 3−51.
(37) Schaad, L. J. Theory of the Hydrogen Bond. In *Hydrogen Bonding*; Joesten, M. D., Schaad, L. J., Eds.; Marcel Dekker: New York, 1974; pp 53−154.
(38) *Tripos Force Field*, version 6.4.2; Tripos, Inc.: St. Louis, MO, 1998.
(39) Allinger, N. L.; Rahman, M.; Lii, J. *J. Am. Chem. Soc.* **1990**, *112*, 8293−8307.
(40) Schmitz, L. R.; Allinger, N. L. *J. Am. Chem. Soc.* **1990**, *112*, 8307−8315.
(41) Allinger, N. L.; Zhu, Z. S.; Chen, K. *J. Am. Chem. Soc.* **1992**, *114*, 6120−6133.
(42) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.
(43) *Cambridge Structural Database System (CSDS)*, version 4; Cambridge Crystallographic Data Center: Cambridge, U.K., 1995.

Table 2. Effective Atomic Radii (R^{eff}) for H-Bond Donor Heavy (A) and Hydrogen Atoms (H) and H-Bond Acceptors (B)[a]

H-bond donor	R_A^{vdw} (Å)	R_A^{eff} (Å)	R_H^{vdw} (Å)	R_H^{eff} (Å)	H-bond acceptor	R_B^{vdw} (Å)	R_B^{eff} (Å)
$OH_{alcohol}$	1.66	1.26	1.08	0.41	$N_{3'-amine}$	1.82	1.42
$OH_{carbox.acid}$	1.69	1.09	1.08	0.19	$N_{nitrile}$	1.72	1.47
$NH_{amine/amide}$	1.82	1.40	1.14	0.52	$N_{pyridine}$	1.78	1.41
$NH_{pyrrole}$	1.80	1.30	1.14	0.37			

[a] Escobar's OPT2A atomic radii[22] (R^{vdw}) are listed for comparison.

the average distance between an H-bond acceptor (B) and the water's oxygen atom (A), and the value calculated from eq 25 is the effective radius of the H-bond acceptor, R_B^{eff}. For H-bond donor hydrogen atoms, the effective atomic radii were calculated using

$$R_H^{eff} = r_{HB} - R_{water} \tag{26}$$

where r_{HB} is the average distance between the H-bond donor hydrogen atom (H) and the oxygen atom (B) of water. The effective atomic radii calculated from the H-bond donor/acceptor groups in our CSDS data set are listed in Table 2 along with Escobar's OPT2A van der Waal radii[22] for comparison. The OPT2A radii were derived by fitting atom-centered spheres to the 0.002 au isodensity surface calculated at the HF/6-31G* level. Note that the effective atomic radii are significantly smaller than the corresponding OPT2A radii.

On the basis of experimental and theoretical studies, it has been shown that H-bond donors and acceptors undergo significant redistribution of electron density upon H-bond formation.[35−37] Indeed, the hydrogen atom has been shown to lose a significant amount of electron density. While both heavy atoms of the H-bond gain electron density, the H-bond donor heavy atom gains more electrons compared to the H-bond acceptor. To account for the redistribution of electron density, as a first approximation, we modify the atomic charges of H-bond donors according to the following strategy. Since the H-bond donor strengths are at least *qualitatively* proportional to their differences in electronegativity (F−H > O−H > N−H), we remove electron density from the H-bond donor hydrogen in a manner that is dependent upon the electronegativity of both the hydrogen and H-bond donor heavy atom. The amount of electron density removed from the hydrogen, Δq, is given by

$$\Delta q = q_H\left(\frac{\chi_A}{\chi_A + \chi_H} - \frac{1}{2}\right) \tag{27}$$

where q_H is the original electron density on the hydrogen and χ_A and χ_H are the electronegativities of the H-bond donor heavy atom and hydrogen, respectively. To keep the molecule neutral, Δq is added to the heavy atom. Given that the electron density of the H-bond donor hydrogen has been reduced, it is also now less polarizable. To account for this, we reduce the polarizability of the hydrogen in a manner identical to that for its charge.

(g) Solute−Solvent Interaction Energy. Having derived expressions for the four modes of intermolecular interaction, we calculate the solute−solvent interaction energy for a given solvent configuration g, E_g^{ssi}, as

$$E_g^{ssi} = C^{elec}E_g^{elec} + C^{polv}E_g^{polv} + C^{polu}E_g^{polu} + C^{disp}E_g^{disp} \tag{28}$$

Recall that the equations for the modes of interaction involve the effective dipole moment and polarizability of the solvent. Just as we could not use the bulk solvent interfacial tension as the effective interfacial tension for the free energy of microscopic cavity formation, we cannot use the nominal values of the bulk solvent dipole moment or polarizability. We argue that those nominal values are sufficiently accurate for use when computing the energies needed for Boltzmann averaging of and within solvent configurations, but when computing the solute−solvent interaction energy, we will express the effective solvent dipole moment and effective polarizability by multiplying the nominal values (provided as input to GSSI) by scaling factors included in the regression coefficients C^{elec}, C^{polv}, C^{polu}, and C^{disp} determined for each mode of interaction.

In a manner analogous to but *much* faster than molecular dynamics, ΔE_{ssi}° is calculated by Boltzmann averaging the energies of all solvent configurations.

$$\Delta E_{ssi}^{\circ} = \langle \sum_m C^m E_g{}^m \rangle_g = \sum_m C^m \langle E_g{}^m \rangle_g \tag{29}$$

The sum over index m is over the four modes of interaction, and the broken brackets followed by the subscript g represent Boltzmann averaging over all solvent configurations. It is significant to note that, although the Boltzmann average in eq 29 should be over all solvent configurations, careful analysis reveals that averaging over just the 100−150 most attractive configurations yields results essentially identical to those obtained by averaging over all configurations. It is also significant to note that it was *not* sufficient to simply consider the single most attractive configuration or the configuration seeded at the most attractive solvent position.

2.2.2. Entropy Change of Solute−Solvent Interaction. Perhaps one of the most significant features of GSSI is that it also addresses the entropic contributions to the free energy of gas-to-solution transfer. As the solute molecule is placed into the solvent cavity, there are changes in entropy associated with both the solute and the solvent.

$$\Delta S_{ssi}^{\circ} = \Delta S_{ssi}^{\circ}(\text{solute}) + \Delta S_{ssi}^{\circ}(2',\text{solvent}) \tag{30}$$

For the solute, there is a considerable loss of translational (ΔS_{trans}°), rotational (ΔS_{rot}°), vibrational (ΔS_{vib}°), and confor-

mational (ΔS°_{conf}) entropy.

$$\Delta S^\circ_{ssi}(\text{solute}) = \Delta S^\circ_{trans} + \Delta S^\circ_{rot} + \Delta S^\circ_{vib} + \Delta S^\circ_{conf} \quad (31)$$

The largest loss is in ΔS°_{trans}, approximately equal to R ln(V_{free}/V°_{gas}), where V°_{gas} is the volume occupied by the solute as an ideal gas at standard temperature and pressure and V_{free} is the free volume of the liquid solvent. Noting that $V^\circ_{gas} \gg V_{free}$, that V_{free} is approximately constant for many liquids, and that V°_{gas} is actually a constant, the loss of translational entropy is essentially constant for all solutes in a given solvent. In general, the loss of rotational and vibrational entropy is relatively small compared to the loss of translational entropy and is also relatively constant for all solutes.[19−22] Intuitively, the loss of conformational entropy is related to the number of rotatable bonds in the solute and to the strength of the solute−solvent interaction. As such, it can be effectively modeled as

$$\Delta S^\circ_{conf} = C^{conf} N_{rot} \quad (32)$$

where N_{rot} is the number of rotatable (single, acyclic) bonds in the solute and C^{conf} is a regression coefficient. On the basis of the arguments presented here, we can estimate the entropy change of the solute, ΔS°_{ssi}(solute), to be

$$\Delta S^\circ_{ssi}(\text{solute}) = C^{conf} N_{rot} + C^{trv} \quad (33)$$

where C^{trv} is a constant which represents the essentially constant loss of translational, rotational, and vibrational entropy for the solute.

In addition to the loss of solute entropy, there is also a change in entropy associated with the solvent molecules near the surface of the cavity as the solute molecule is placed in the cavity. Recall that these solvent molecules had previously experienced a considerable loss of entropy when the solvent cavity was first created. This entropy change was identified as ΔS°_{cav} and constitutes the primary entropy change of the solvent [$\Delta S^\circ_{cav} = \Delta S^\circ(1',\text{solvent})$]. However, there is a secondary entropy change of the solvent that occurs once the solute has been placed in the cavity. If the solute is chemically similar to the solvent, placing the solute in the cavity will reduce the asymmetry of the intermolecular force field and the entropy of the solvent molecules at the cavity surface will be increased. If, on the other hand, the solute−solvent interaction is stronger than the solvent−solvent interaction, the entropy of the solvent molecules at the cavity surface will be further reduced. Clearly, the magnitude of the secondary entropy change of the solvent will depend on the strength of the solute−solvent interaction. Thus, we can argue that the secondary entropy change of the solvent, $\Delta S^\circ_{ssi}(2',\text{solvent})$, is approximately proportional to ΔE°_{ssi} and can be approximated as

$$\Delta S^\circ_{ssi}(2',\text{solvent}) = C^{2',solv} \Delta E^\circ_{ssi} \quad (34)$$

where $C^{2',solv}$ is the proportionality constant that relates ΔE°_{ssi} to $\Delta S^\circ_{ssi}(2',\text{solvent})$.

Since we cannot directly validate eq 34, the following strategy was developed to indirectly validate it on the basis of available experimental values for $\Delta H^\circ_{gas\rightarrow soln}$ and $\Delta S^\circ_{gas\rightarrow soln}$. First, note that $\Delta S^\circ_{gas\rightarrow soln}$ can be calculated as the sum of the entropies of cavity formation and solute−solvent interaction.

$$T\Delta S^\circ_{gas\rightarrow soln} = T[\Delta S^\circ_{cav} + \Delta S^\circ_{ssi}(\text{solute}) + \Delta S^\circ_{ssi}(2',\text{solvent})] \quad (35)$$

Recalling that ΔS°_{cav} is related to the size of the solute and that ΔS°_{ssi}(solute) can be modeled via eq 33, we can rewrite eq 35 as

$$T\Delta S^\circ_{gas\rightarrow soln} = T[f'(\text{TSA}^{acc}) + C^{conf} N_{rot} + C^{trv} + \Delta S^\circ_{ssi}(2',\text{solvent})] \quad (36)$$

where f' is a function of TSA^{acc} and all other terms are as previously defined. Similarly, $\Delta H^\circ_{gas\rightarrow soln}$ can be calculated as the sum of the enthalpies of cavity formation and solute−solvent interaction.

$$\Delta H^\circ_{gas\rightarrow soln} = \Delta H^\circ_{cav} + \Delta HH^\circ_{ssi} \quad (37)$$

Again, recalling that ΔH°_{cav} is related to the size of the solute and that the Δ(PV) component of ΔH°_{ssi} is essentially constant, we can rewrite eq 37 as

$$\Delta H^\circ_{gas\rightarrow soln} = f''(\text{TSA}^{acc}) + \Delta E^\circ_{ssi} + C^{PV} \quad (38)$$

where f'' is also a function of TSA^{acc}.

Assuming that $T\Delta S^\circ_{ssi}(2',\text{solvent})$ is linearly related to ΔE°_{ssi} (i.e., a $y = mx + b$ relationship exits between the two thermodynamic functions), we can write the following expression:

$$T\Delta S^\circ_{ssi}(2',\text{solvent}) = m[\Delta H^\circ_{gas\rightarrow soln} - f''(\text{TSA}^{acc}) - C^{PV}] + b = m[\Delta H^\circ_{gas\rightarrow soln} - f''(\text{TSA}^{acc})] + b' \quad (39)$$

Substituting eq 39 into eq 36, we obtain an equation, which enables us to calculate $T\Delta S^\circ_{gas\rightarrow soln}$ in terms of $\Delta H^\circ_{gas\rightarrow soln}$, TSA^{acc}, and N_{rot}.

$$T\Delta S^\circ_{gas\rightarrow soln} = T[f'(\text{TSA}^{acc}) + C^{conf} N_{rot} + C^{trv}] + m[\Delta H^\circ_{gas\rightarrow soln} - f''(\text{TSA}^{acc})] + b' \quad (40)$$

Assuming that both f' and f'' are linear functions of TSA^{acc} and that T is constant, we can simplify eq 40 to yield

$$T\Delta S^\circ_{gas\rightarrow soln} = C^{H}_{T\Delta S} \Delta H^\circ_{gas\rightarrow soln} + C^{cav}_{T\Delta S} \text{TSA}^{acc} + C^{conf}_{T\Delta S} N_{rot} + C_{T\Delta S} \quad (41)$$

Thus, a multiple linear regression analysis can be conducted wherein a set of $T\Delta S^\circ_{gas\rightarrow soln}$ values is regressed against corresponding sets of $\Delta H^\circ_{gas\rightarrow soln}$, TSA^{acc}, and N_{rot} values. On the basis of statistical results, we can assess the validity of our assumption that $T\Delta S^\circ_{ssi}(2',\text{solvent})$ is approximately linearly related to ΔE°_{ssi}. Such statistical results are discussed in section 3.1.

2.3. Free Energy of Gas-to-Solution Transfer. Given that we have derived expressions for ΔG°_{cav}, ΔE°_{ssi}, ΔS°_{ssi}(solute), and $\Delta S^\circ_{ssi}(2',\text{solvent})$, the free energy of the gas-to-solution transfer is calculated as

$$\Delta G^\circ_{gas\rightarrow soln} = \Delta G^\circ_{cav} + \Delta G^\circ_{ssi} = C^{cav}_{gas\rightarrow soln}\text{TSA}^{acc} + \langle \sum_m C^m_{gas\rightarrow soln} E_g{}^m \rangle_g + C^{conf}_{gas\rightarrow soln} N_{rot} + C_{gas\rightarrow soln} \quad (42)$$

The solute descriptors are as previously defined, and all coefficients are determined by multiple linear regression for the solvent of interest. Note that the coefficients $\{C_{gas\rightarrow soln}{}^m\}$ not only characterize the solvent but also include the proportionality constant for the approximate relationship between $\Delta S^\circ_{ssi}(2',\text{solvent})$ and ΔE°_{ssi} as expressed in eq 34. Also note that $C_{gas\rightarrow soln}$ includes the essentially constant ΔS°_{trans}, ΔS°_{rot}, ΔS°_{vib}, and $\Delta(PV)$ terms. In addition, the regression coefficients will also include the $-RT$ factor if log equilibrium constant values are being regressed rather than free energies. Note that there are a total of seven regression coefficients in the model expressed by eq 42.

3. Results and Discussion

To validate the basic GSSI approach, we applied it to the prediction of free energies of gas-to-solution transfer. Since the GSSI approach is focused on the gas-to-solution transfer process, validation of our basic model was relatively straightforward and uncomplicated by differences between two or more solvents. Two different sets of experimental data were collected from the primary literature. One data set was comprised of a series of solutes with measured free energies of gas-to-aqueous solution transfer. The second data set was comprised of a series of solutes with measured free energies of gas-to-hexadecane transfer. To demonstrate the utility of the GSSI approach for a solvent possessing both polar and nonpolar characteristics, a third data set was also collected from the literature. However, unlike the first two data sets, this set was comprised of a series of solutes for which the free energies of gas-to-octanol transfer had been derived from experimentally measured octanol−water partition coefficients and free energies of gas-to-aqueous solution transfer.

The current implementation of GSSI uses a single, Concord-generated structure[44] from which the solute descriptors are calculated. Ideally, we would Boltzmann average over solute conformations; however, the energy required for such averaging would need to include the solvation energy which leads to obvious difficulties. Instead, we attempt to identify the most probable conformation of a solute in a given solvent, which should provide a reasonable estimate of the Boltzmann-averaged results. To this end, we first generated a set of maximally diverse conformers for each solute in each of the data sets using Confort.[45] The energy range associated with a given set of conformers was kept to less than 10 kcal/mol of the lowest-energy conformer identified for that set. Next, the solvent-accessible surface areas of the conformers were partitioned into polar and nonpolar contributions. The partitioning was based on the crude but popular definition of "polar" surface, which is that surface associated with nitrogen and oxygen atoms and hydrogen atoms bonded to these heteroatoms.[46−48] For the aqueous solution data set, a conformer was selected for each solute having the maximum exposed polar surface and minimum exposed nonpolar surface. The converse was true for the hexadecane data set in which conformers with maximally exposed nonpolar surfaces and minimally exposed polar surfaces were selected. For the octanol data set, given that octanol has both polar and nonpolar characteristics, the conformers that were selected were those with maximally exposed polar and nonpolar surfaces.

In the identification of the most probable solute conformations, Savol3[25] was used to compute the atomic contributions to the solvent-accessible surface areas, the sum of which yields TSA^{acc}. The atomic radii used in the Savol3 calculations were Escobar's OPT2A radii.[22] To calculate the intermolecular interaction potential components, UDS[33] was used to generate the uniform molecular dot surfaces. The electronic properties (i.e., multipoint charges, polarizabilities, etc.) required by GSSI were computed using our HSCF program.[31] For each training set, a GSSI model was derived by regressing the free energies of gas-to-solution transfer against the calculated solute descriptors. To assess the predictive ability of our GSSI models, cross validation was performed by the leave-one-out procedure.

3.1. Approximate Linear Relationship between $T\Delta S^\circ_{ssi}(2',\text{solvent})$ and ΔE°_{ssi}. Before the utility of the GSSI approach in modeling free energies of solvation can be demonstrated, our assumption that a linear relationship exists between $T\Delta S^\circ_{ssi}(2',\text{solvent})$ and ΔE°_{ssi} must be validated. To this end, a set of experimentally measured enthalpies ($\Delta H^\circ_{gas\rightarrow aq}$) and entropies ($T\Delta S^\circ_{gas\rightarrow aq}$) of gas-to-aqueous solution transfer were collected from the literature for 93 diverse solutes.[49,50] The $T\Delta S^\circ_{gas\rightarrow aq}$ values covered a range of 12.4 kcal/mol, with maximum and minimum values of −7.52 and −19.92 kcal/mol, respectively. For $\Delta H^\circ_{gas\rightarrow aq}$, the range of values was 15.4 kcal/mol, with maximum and minimum values of −3.49 and −18.89 kcal/mol, respectively.

A multiple linear regression model was developed for $T\Delta S^\circ_{gas\rightarrow aq}$ using $\Delta H^\circ_{gas\rightarrow aq}$, TSA^{acc}, and N_{rot} as predictor variables. In the calculation of TSA^{acc}, the solute conformers

(44) *CONCORD*, version 4.0.2; Tripos, Inc.: St, Louis, MO, 1998.
(45) *Confort*, version 3.93; Tripos, Inc.: St. Louis, MO, 2001.

(46) Palm, K.; Luthman, K.; Ungell, A.; Strandlund, G.; Artursson, P. *J. Pharm. Sci.* **1996**, *85*, 32−39.
(47) Clark, D. *J. Pharm. Sci.* **1999**, *88*, 815−821.
(48) Osterberg, T.; Norinder, U. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1408−1411.
(49) Abraham, M. H.; Whiting, G. S.; Fuchs, R.; Chambers, E. J. *J. Chem. Soc., Perkins Trans. 2* **1990**, *1*, 291−300.
(50) Abraham, M. H. *J. Chem. Soc., Faraday Trans. 1* **1984**, *80*, 153−181.

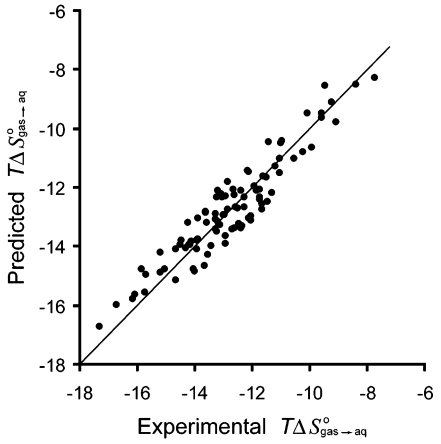


Figure 4. Predicted vs experimental $T\Delta S^{\circ}_{gas\rightarrow aq}$ values (kilocalories per mole) for a set of 93 diverse solutes.

Table 3. Regression Statistics for the Four-Parameter Model of $T\Delta S^{\circ}_{gas\rightarrow aq}$

predictor	coefficient	standard deviation	*t* ratio
constant	−5.519	0.455	−12.13
$\Delta H^{\circ}_{gas\rightarrow aq}$	0.307	0.024	12.76
TSA^{acc}	−0.0099	0.0017	−5.78
N_{rot}	−0.347	0.045	−7.68

that were used were those having maximally exposed polar surfaces and minimally exposed nonpolar surfaces. Also, an effective solvent radius of 1.50 Å was used for water. For N_{rot}, the values that were used were those reported by GSSI. Listed in Table 3 are the values for the four regression coefficients along with their corresponding standard deviations and *t* ratios; note that all terms are significant. The r^2 value for the regression equation was equal to 0.880 with a standard error of prediction of 0.644 kcal/mol. The average unsigned error between predicted and experimental $T\Delta S^{\circ}_{gas\rightarrow aq}$ values was 0.53 kcal/mol. The cross-validated r^2 value was equal to 0.878, which is essentially identical to the r^2 value. A plot of predicted versus experimental $T\Delta S^{\circ}_{gas\rightarrow aq}$ values is shown in Figure 4. Clearly, there is excellent agreement between predicted and experimental values, supporting our assumption that a linear relationship exists between $\Delta S^{\circ}_{ssi}(2',\text{solvent})$ and ΔE°_{ssi}.

3.2. Free Energy of Gas-to-Aqueous Solution Transfer. We validated our GSSI approach by applying it to the prediction of free energies of gas-to-aqueous solution transfer ($\Delta G^{\circ}_{gas\rightarrow aq}$). A training set of experimentally measured $\Delta G^{\circ}_{gas\rightarrow aq}$ values for 111 diverse solutes was compiled from the primary literature.[1−4,49−53] The data covered a range of roughly 9.0 kcal/mol with maximum and minimum values of 6.95 and −2.07 kcal/mol, respectively. To calculate the GSSI solute descriptors, the following solvent parameters

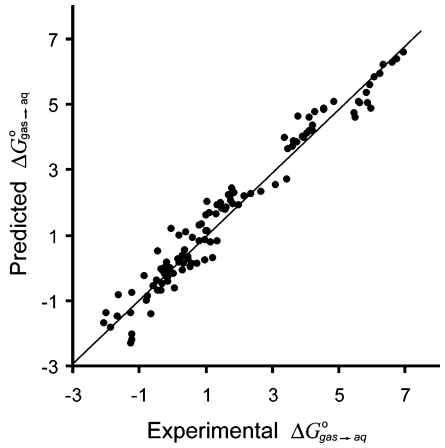


Figure 5. Predicted vs experimental free energies (kilocalories per mole) of gas-to-aqueous solution transfer for the seven-parameter GSSI model.

Table 4. Regression Statistics for the Seven-Parameter GSSI Model of Gas-to-Aqueous Solution Transfer

predictor	coefficient	standard deviation	*t* ratio
constant	10.916	0.698	15.64
TSA^{acc}	−0.059	0.007	−7.84
$\langle E_g^{elec}\rangle_g$	−0.289	0.011	−26.10
$\langle E_g^{polu}\rangle_g$	1.123	0.020	56.75
$\langle E_g^{disp}\rangle_g$	−0.517	0.075	−6.89
$\langle E_g^{polv}\rangle_g$	3.666	0.057	64.79
N_{rot}	0.503	0.022	22.71

for water were used: an effective solvent radius of 1.50 Å, a dipole moment of 1.85 D, a molecular polarizability of 1.45 Å^3, and an ionization potential of 12.6 eV.

The experimental $\Delta G^{\circ}_{gas\rightarrow aq}$ values were regressed against the calculated solute descriptors to generate a GSSI model. As a result of very careful algorithm design and software engineering, the time required to calculate the descriptors and generate the model was only 3 CPU minutes on a single SGI R12000 processor. Table 4 lists the values of the seven regression coefficients for the GSSI model along with their corresponding standard deviations and *t* ratios. Illustrated in Figure 5 is a plot of predicted versus experimental $\Delta G^{\circ}_{gas\rightarrow aq}$ values. The r^2 value for the regression equation was equal to 0.959, and the standard error of prediction was 0.476 kcal/mol. The average unsigned error between predicted and experimental $\Delta G^{\circ}_{gas\rightarrow aq}$ values was 0.38 kcal/mol. The cross-validated r^2 value was equal to 0.952, which is nearly identical to the r^2 value. The sets of regression coefficients from the cross validation did not differ significantly from those values reported in Table 4, thus indicating the stability of the GSSI model.

Although it is instructive to compare the experimental error with the standard error of prediction, such a comparison is not possible for the $\Delta G^{\circ}_{gas\rightarrow aq}$ data set since an estimate of the experimental error has not been reported in the literature. In addition, an assessment of the experimental error would be very difficult if not impossible. Not only are the $\Delta G^{\circ}_{gas\rightarrow aq}$ values taken from numerous literature sources

(51) Abraham, M. H.; Grellier, P. L.; McGill, R. A. *J. Chem. Soc., Perkins Trans. 2* **1987**, *5*, 797−803.
(52) Fuchs, R.; Chambers, E. J.; Stephenson, W. K. *Can. J. Chem.* **1987**, *65*, 2624−2627.
(53) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720−725.

Table 5. Cross-Correlation Analysis of Solute Descriptors Used in Developing the GSSI Model for Gas-to-Aqueous Solution Transfer

	TSA^{acc}	$\langle E_g^{elec}\rangle_g$	$\langle E_g^{polu}\rangle_g$	$\langle E_g^{disp}\rangle_g$	$\langle E_g^{polv}\rangle_g$	N_{rot}
TSA^{acc}	1.000	–	–	–	–	–
$\langle E_g^{elec}\rangle_g$	0.089	1.000	–	–	–	–
$\langle E_g^{polu}\rangle_g$	0.057	0.910	1.000	–	–	–
$\langle E_g^{disp}\rangle_g$	−0.985	−0.043	0.003	1.000	–	–
$\langle E_g^{polv}\rangle_g$	0.076	0.979	0.846	−0.044	1.000	–
N_{rot}	0.580	−0.196	−0.125	−0.520	−0.251	1.000

where typically only a single experimental value is reported, but also the solvation free energies have been determined either directly or indirectly through various analytical techniques which have different levels of accuracy.

We can also assess the validity of our model by considering the signs of the various regression coefficients (and descriptors) relative to what we expect on the basis of our understanding of the thermodynamic aspects of the corresponding components of the gas-to-solution process. For example, since cavity formation is both enthalpically and entropically unfavorable and since TSA^{acc} is positive, we expect that C^{cav} will also be positive, thereby yielding a positive (unfavorable) contribution to $\Delta G^{\circ}_{gas\rightarrow soln}$. It is important to recall that the interaction potential components (i.e., electrostatic, polarization, and dispersion) are calculated as negative values and that TSA^{acc} and N_{rot} are positive values. Thus, since the interaction between solute and any solvent is more energetically favorable than noninteracting solute in the gas phase, we would expect positive signs for the regression coefficients associated with the interaction energies, indicating that the gas-to-solution process is favorable. However, recall that the coefficients of the solute−solvent interaction terms also account for the typically favorable secondary entropy change of the solvent that occurs once the solute has been placed into the cavity. This favorable entropic contribution corresponds to a decrease (negative component) of the free energy of solvation and, hence, might result in negative coefficients of the solute−solvent interaction terms. Finally, we expect C^{conf} to be positive, reflecting the fact that the loss of conformational entropy hinders the gas-to-solution process.

From Table 4, we see that the only regression coefficient not having the expected sign was C^{cav}. However, the reversed sign for this coefficient can be attributed to the fact that the cavity and dispersion terms are highly correlated ($r = -0.985$), as shown in Table 5. This is not surprising considering the fact that cavity formation and dispersion interactions are both related to the size of the solute. Also note from Table 5 that $\langle E_g^{elec}\rangle_g$ and $\langle E_g^{polv}\rangle_g$ are highly correlated ($r = 0.979$) as well. This correlation results from the fact that $\langle E_g^{elec}\rangle_g$ and $\langle E_g^{polv}\rangle_g$ are calculated in terms of the electric field generated by the solute's charge distribution. The relatively high cross correlations between $\langle E_g^{elec}\rangle_g$ and $\langle E_g^{polu}\rangle_g$ ($r = 0.910$) and $\langle E_g^{polu}\rangle_g$ and $\langle E_g^{polv}\rangle_g$ ($r = 0.846$) were at first somewhat surprising. However, this can easily be understood if we recall that the electric field generated by a solvent configuration g is calculated on the basis of the Boltzmann-averaged orientations of the solvent dipoles, which in turn are calculated on the basis of the electrostatic interaction between the solvent and the solute.

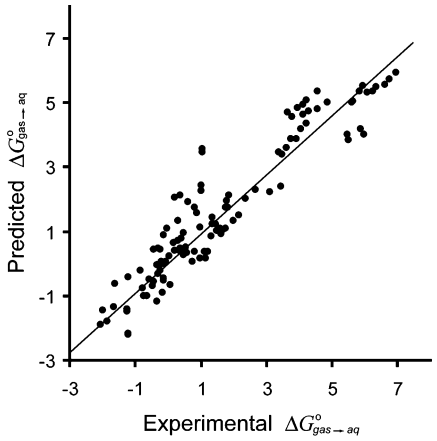


Figure 6. Predicted vs experimental free energies (kilocalories per mole) of gas-to-aqueous solution transfer for the four-parameter GSSI model.

Table 6. Regression Statistics for the Four-Parameter GSSI Model of Gas-to-Aqueous Solution Transfer

predictor	coefficient	standard deviation	*t* ratio
constant	7.191	0.371	19.41
TSA^{acc}	−0.011	0.001	−8.58
$\langle E_g^{elec}\rangle_g$	0.197	0.002	105.61
N_{rot}	0.469	0.014	32.74

Given that TSA^{acc} and $\langle E_g^{disp}\rangle_g$ are highly correlated and that $\langle E_g^{elec}\rangle_g$, $\langle E_g^{polv}\rangle_g$, and $\langle E_g^{polu}\rangle_g$ are also correlated, we reduced the number of descriptors in our GSSI model to include only TSA^{acc}, $\langle E_g^{elec}\rangle_g$, and N_{rot}. Thus, the cavity term must account not only for cavity formation but also for the dispersion interaction. Likewise, $\langle E_g^{elec}\rangle_g$ must account not only for the electrostatic interaction but also for the two types of polarization interactions. Table 6 lists the values of the regression coefficients for the four-parameter GSSI model along with their corresponding standard deviations and *t* ratios. Note that all terms are significant with the most significant being $\langle E_g^{elec}\rangle_g$. This is to be expected with a polar solvent like water. A plot of predicted versus experimental $\Delta G^{\circ}_{gas\rightarrow aq}$ values is illustrated in Figure 6. A reasonable fit is still achieved with an r^2 value of 0.885 and a standard error of prediction of 0.801 kcal/mol. The cross-validated r^2 value was equal to 0.865, which is again very similar to the r^2 value. The average unsigned error between predicted and experimental $\Delta G^{\circ}_{gas\rightarrow aq}$ values was 0.61 kcal/mol. An analysis of the solutes with the larger errors revealed that the most problematic compounds were the secondary and tertiary amines, which were predicted not to be sufficiently hydrophilic. This may indicate that we need to reevaluate the adjustments made to these amines to better account for their H-bonding potential.

3.3. Free Energy of Gas-to-Hexadecane Transfer. To validate the GSSI approach for nonpolar solvents, we applied

Table 7. Regression Statistics for the Seven-Parameter GSSI Model of Gas-to-Hexadecane Transfer

predictor	coefficient	standard deviation	*t* ratio
constant	8.587	0.407	21.08
TSA^{acc}	−0.026	0.003	−8.66
$\langle E_g^{elec}\rangle_g$	−5.056	1.610	−3.14
$\langle E_g^{polu}\rangle_g$	3.298	0.101	32.67
$\langle E_g^{disp}\rangle_g$	0.180	0.069	2.61
$\langle E_g^{polv}\rangle_g$	9.632	2.710	3.55
N_{rot}	0.475	0.024	19.46

our solvation model to the prediction of free energies of gas-to-hexadecane transfer ($\Delta G^\circ_{gas\rightarrow hex}$). A set of experimentally measured $\Delta G^\circ_{gas\rightarrow hex}$ values was taken from a compilation of data from Abraham et al.[49] The data set included 250 structurally diverse solutes as evidenced by the wide range of free energies. The maximum and minimum values of $\Delta G^\circ_{gas\rightarrow hex}$ were 2.41 and −5.24 kcal/mol, respectively. The solvent parameters used to compute the GSSI descriptors were derived from a methylene (CH_2) unit of hexadecane, the portion of the solvent most likely to make van der Waals contact with the solutes. The one exception was the molecular ionization potential, which was calculated by HSCF to be 10.78 eV. The effective solvent radius, dipole moment, and polarizability of CH_2 were calculated to be 1.92 Å, 0.48 D, and 1.85 Å^3, respectively. The solvent radius was calculated on the basis of the volume occupied by CH_2, which in turn was calculated by Savol3 using Escobar's OPT2A atomic radii. The charges and polarizabilities used to calculate the dipole moment and polarizability of CH_2 were obtained from HSCF.

The experimental $\Delta G^\circ_{gas\rightarrow hex}$ values were regressed against the calculated solute descriptors to generate a GSSI model. The time required to calculate the descriptors and generate the model was approximately 7.50 CPU minutes on a single SGI R12000 processor. Table 7 lists the values of the regression coefficients for the GSSI model along with their corresponding standard deviations and *t* ratios. The predicted $\Delta G^\circ_{gas\rightarrow hex}$ values are plotted against the experimental values in Figure 7. Agreement between both sets of values was excellent with an r^2 value of 0.964 and a standard error of prediction of 0.277 kcal/mol. The average unsigned error between the predicted and experimental $\Delta G^\circ_{gas\rightarrow hex}$ values was 0.22 kcal/mol. The cross-validated r^2 was equal to 0.961, which is essentially identical to the r^2 value. For most of the regression coefficients, there was little difference in the values obtained from cross validation compared to those reported in Table 7. The one exception was the coefficient for $\langle E_g^{polu}\rangle_g$. Its value was significantly affected when diisopropyl sulfide, 1,2-dibromoethane, 1,3-dichlorobenzene, methoxyflurane, 3-nitrotoluene, or 2-methylbutan-2-ol was excluded.

While Abraham et al.[49] do not specifically report an experimental error for the $\Delta G^\circ_{gas\rightarrow hex}$ values, they do state that "the expected error in the log L_H values (the log of the Ostwald solubility coefficient of solutes in hexadecane) is very small, probably no more than 0.03 log unit." This translates to an error of roughly ±0.05 kcal/mol for $\Delta G^\circ_{gas\rightarrow hex}$. One obvious concern with a regression model is that it may be overfitting the experimental data. A second concern is that if the standard error of prediction is much larger than the experimental error, then the model may not be very useful. If we accept the assessment of the experimental error for $\Delta G^\circ_{gas\rightarrow hex}$, the GSSI model has certainly not overfit the data. Moreover, the value of the standard error of prediction is not unreasonable given the strong correlation ($r^2 = 0.961$) between predicted and experimental values.

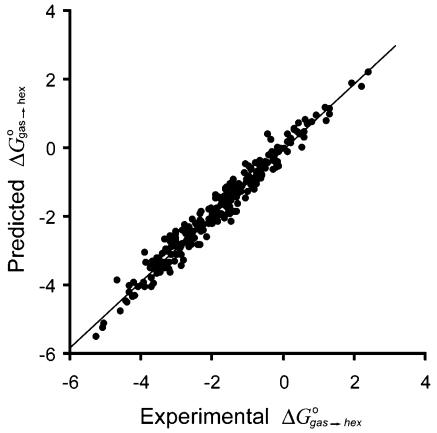


Figure 7. Predicted vs experimental free energies (kilocalories per mole) of gas-to-hexadecane transfer for the seven-parameter GSSI model.

Table 8. Cross-Correlation Analysis of Solute Descriptors Used in Developing the GSSI Model for Gas-to-Hexadecane Transfer

	TSA^{acc}	$\langle E_g^{elec}\rangle_g$	$\langle E_g^{polu}\rangle_g$	$\langle E_g^{disp}\rangle_g$	$\langle E_g^{polv}\rangle_g$	N_{rot}
TSA^{acc}	1.000	–	–	–	–	–
$\langle E_g^{elec}\rangle_g$	−0.029	1.000	–	–	–	–
$\langle E_g^{polu}\rangle_g$	−0.048	0.871	1.000	–	–	–
$\langle E_g^{disp}\rangle_g$	−0.985	−0.012	0.008	1.000	–	–
$\langle E_g^{polv}\rangle_g$	−0.047	0.997	0.859	−0.005	1.000	–
N_{rot}	0.765	−0.148	−0.077	−0.685	−0.180	1.000

Given our previous arguments with regard to the signs of the regression coefficients, we can see from Table 7 that C^{cav} was the only coefficient not to have the expected sign. However, again, this can be attributed to the fact that TSA^{acc} and $\langle E_g^{disp}\rangle_g$ are highly correlated ($r = -0.985$), as shown in Table 8. From this table, we also see that $\langle E_g^{elec}\rangle_g$ and $\langle E_g^{polv}\rangle_g$ are highly correlated ($r = 0.997$), as well as $\langle E_g^{elec}\rangle_g$ and $\langle E_g^{polu}\rangle_g$ ($r = 0.871$), and $\langle E_g^{polv}\rangle_g$ and $\langle E_g^{polu}\rangle_g$ ($r = 0.859$). These cross correlations result for the same reasons outlined in the previous section.

As with the previous data set, we reduced the number of terms in our GSSI model for $\Delta G^\circ_{gas\rightarrow hex}$ to include only TSA^{acc}, $\langle E_g^{elec}\rangle_g$, and N_{rot}. Table 9 lists the values of the regression coefficients for the four-parameter GSSI model along with their corresponding standard deviations and *t* ratios. Note that the most significant term is TSA^{acc}, which also accounts for the dispersion interaction. This is to be

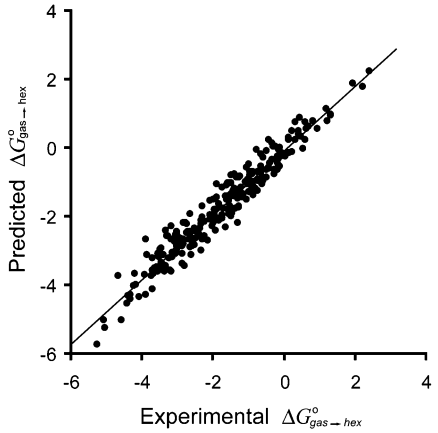


Figure 8. Predicted vs experimental free energies (kilocalories per mole) of gas-to-hexadecane transfer for the four-parameter GSSI model.

Table 9. Regression Statistics for the Four-Parameter GSSI Model of Gas-to-Hexadecane Transfer

predictor	coefficient	standard deviation	*t* ratio
constant	8.730	0.095	91.58
TSA^{acc}	−0.035	0.0003	−111.36
$\langle E_g^{elec}\rangle_g$	1.037	0.021	49.14
N_{rot}	0.455	0.054	85.04

expected with a nonpolar solvent like hexadecane. Illustrated in Figure 8 is a plot of predicted versus experimental $\Delta G^{\circ}_{gas\rightarrow hex}$ values. The r^2 value for the regression model was equal to 0.939, which was essentially identical to the cross-validated r^2 value of 0.937. The standard error of prediction for the model was equal to 0.357 kcal/mol. The average unsigned error between the predicted and experimental $\Delta G^{\circ}_{gas\rightarrow hex}$ values was 0.28 kcal/mol.

3.4. Free Energy of Gas-to-Octanol Transfer. Finally, to assess the validity of the GSSI approach for solvents having both polar and nonpolar characteristics, we applied our solvation model to the prediction of free energies of gas-to-octanol transfer ($\Delta G^{\circ}_{gas\rightarrow oct}$). Although experimentally measured $\Delta G^{\circ}_{gas\rightarrow oct}$ values for anhydrous octanol are preferred, such data exist for only a small number of solutes.[54,55] To develop a basic GSSI model, we need accurately measured $\Delta G^{\circ}_{gas\rightarrow oct}$ values for a training set of at least 20 structurally diverse compounds. Unfortunately, to the best of our knowledge, such experimental data are unavailable in the literature for anhydrous octanol. Instead, we used the log $P_{oct/gas}$ values for a set of 85 solutes as reported by Duffy and Jorgensen.[54] The data refer to water-saturated octanol and were derived from experimental values of $\Delta G^{\circ}_{gas\rightarrow aq}$ and octanol−water partition coefficients. Only one compound was eliminated from the original set. That compound was acetic acid, which undergoes extensive dimerization in the vapor phase leading to significant error in the experimental $\Delta G^{\circ}_{gas\rightarrow aq}$ value.[50,55] In addition, problems inherent in measuring octanol−water partition coefficients for neutral organic acids have been reported, creating additional uncertainty in the $\Delta G^{\circ}_{gas\rightarrow oct}$ value.[55] For consistency with our previous data sets, the log $P_{oct/gas}$ values were converted to $\Delta G^{\circ}_{gas\rightarrow oct}$ values. The range of values for $\Delta G^{\circ}_{gas\rightarrow oct}$ was roughly 12.00 kcal/mol, with maximum and minimum values of 0.49 and −11.50 kcal/mol, respectively.

Table 10. Regression Statistics for the Seven-Parameter GSSI Model of Gas-to-Octanol Transfer.

predictor	coefficient	standard deviation	*t* ratio
constant	7.268	0.695	10.48
TSA^{acc}	−0.042	0.005	−8.28
$\langle E_g^{elec}\rangle_g$	0.077	0.035	2.20
$\langle E_g^{polu}\rangle_g$	1.458	0.134	10.92
$\langle E_g^{disp}\rangle_g$	−0.064	0.047	−1.35
$\langle E_g^{polv}\rangle_g$	−0.174	0.278	−0.63
N_{rot}	0.546	0.006	85.51

Molecular dynamics simulations of various solutes in octanol have illustrated the dual nature of this solvent. As expected, the hydroxyl group of octanol tends to associate around the polar regions of a solute, while the hydrophobic alkyl portion tends to associate around the nonpolar regions.[55] Thus, the solvent parameters used to compute the GSSI descriptors were those calculated for the hydroxyl and one of the methylene units of octanol. The dipole moment and polarizability of the hydroxyl were calculated to be 1.35 D and 0.91 Å^3, respectively, on the basis of the HSCF-derived charges and polarizabilities. The dipole moment and polarizability of CH_2 were the same as those calculated for hexadecane. All four solvent parameters were tried in all possible combinations. The combination that yielded the best statistical results was the dipole moment of the hydroxyl and the polarizability of CH_2. The ionization potential of octanol was estimated to be 9.77 eV from an empirical relationship between ionization potential and the inverse number of non-hydrogen atoms for a series of alkanols (methanol through heptanol).[56,57] We reasoned that the effective solvent radius of octanol must fall somewhere between 1.50 Å (the radius of a hydroxyl) and 1.92 Å (the radius of CH_2). We explored numerous values from 1.50 to 1.90 Å in increments of 0.05 Å. The effective solvent radius that yielded the best results was 1.60 Å.

Once again, the $\Delta G^{\circ}_{gas\rightarrow oct}$ values were regressed against the calculated solute descriptors to generate a GSSI model. The time required to calculate the descriptors and generate the model was roughly 2 CPU minutes on a single SGI R12000 processor. Table 10 lists the values of the regression coefficients along with their corresponding standard deviations and *t* ratios. A plot of predicted versus experimental $\Delta G^{\circ}_{gas\rightarrow oct}$ values is illustrated in Figure 9. When the quality

(54) Duffy, M. L.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 2878−2888.
(55) Best, S. A.; Merz, K. M., Jr.; Reynolds, C. H. *J. Phys. Chem. B* **1997**, *101*, 10479−10487.
(56) Holmes, J. L.; Lossing, F. P. *Org. Mass Spectrom.* **1991**, *26*, 537−541.
(57) *CRC Handbook of Chemistry and Physics*, 82nd ed.; CRC Press LLC: Boca Raton, FL, 2001; Section 10.

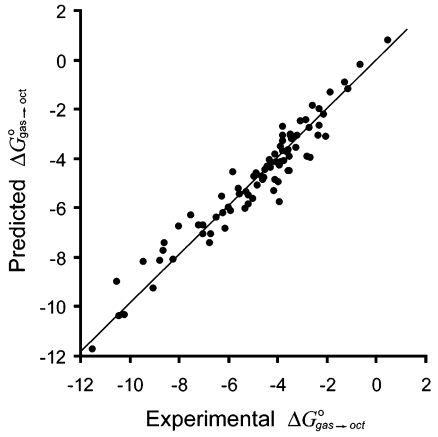


Figure 9. Predicted vs experimental free energies (kilocalories per mole) of gas-to-octanol transfer for the seven-parameter GSSI model.

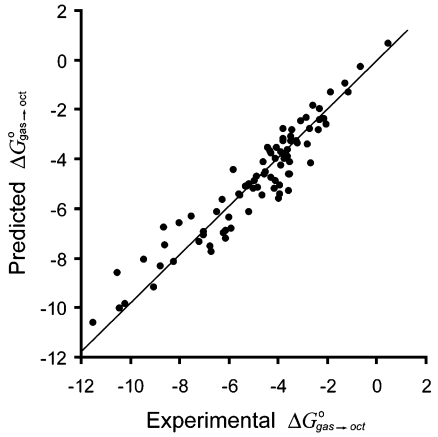


Figure 10. Predicted vs experimental free energies (kilocalories per mole) of gas-to-octanol transfer for the four-parameter GSSI model.

Table 11. Cross-Correlation Analysis of Solute Descriptors Used in Developing the GSSI Model for Gas-to-Octanol Transfer

	TSA^{acc}	$\langle E_g^{elec}\rangle_g$	$\langle E_g^{polu}\rangle_g$	$\langle E_g^{disp}\rangle_g$	$\langle E_g^{polv}\rangle_g$	N_{rot}
TSA^{acc}	1.000	–	–	–	–	–
$\langle E_g^{elec}\rangle_g$	0.252	1.000	–	–	–	–
$\langle E_g^{polu}\rangle_g$	0.240	0.940	1.000	–	–	–
$\langle E_g^{disp}\rangle_g$	−0.919	−0.002	0.050	1.000	–	–
$\langle E_g^{polv}\rangle_g$	0.321	0.983	0.945	−0.052	1.000	–
N_{rot}	0.087	−0.278	−0.195	−0.109	−0.248	1.000

of the data is considered, there is a fairly strong correlation between the predicted and experimental values with an r^2 value of 0.927 and a standard error of prediction of 0.642 kcal/mol. The average unsigned error between predicted and experimental $\Delta G^{\circ}_{gas\rightarrow oct}$ values was 0.49 kcal/mol. The cross-validated r^2 value was equal to 0.912, which is similar to the r^2 value. The values of regression coefficients from the cross validation did not differ significantly from those values reported in Table 10, thus indicating the stability of the GSSI model.

From Table 10, we see that, once again, C^{cav} was the only coefficient not to have the expected sign. If we consider the cross-correlation results listed in Table 11, then the reversed sign can be attributed to the high degree of correlation between TSA^{acc} and $\langle E_g^{disp}\rangle_g$ ($r = -0.919$). As we saw with both the water and hexadecane data sets, $\langle E_g^{elec}\rangle_g$ and $\langle E_g^{polv}\rangle_g$ were also highly correlated ($r = 0.983$), as well as $\langle E_g^{elec}\rangle_g$ and $\langle E_g^{polu}\rangle_g$ ($r = 0.940$) and $\langle E_g^{polv}\rangle_g$ and $\langle E_g^{polu}\rangle_g$ ($r = 0.945$). Clearly, the high degree of correlation seen among these descriptors also accounts for the small values of the t ratios associated with $\langle E_g^{polv}\rangle_g$, $\langle E_g^{disp}\rangle_g$, and $\langle E_g^{elec}\rangle_g$.

A four-parameter GSSI model was generated for $\Delta G^{\circ}_{gas\rightarrow oct}$. The three solute descriptors that were used were TSA^{acc}, $\langle E_g^{elec}\rangle_g$, and N_{rot}. Table 12 lists the values of the regression coefficients along with their corresponding standard deviations and t ratios. Illustrated in Figure 10 is a plot of predicted versus experimental $\Delta G^{\circ}_{gas\rightarrow oct}$ values. With the four-parameter model, we still achieve a reasonable fit between predicted and experimental values with an r^2 value of 0.896 and a standard error of prediction of 0.764 kcal/mol. The average unsigned error between predicted and experimental $\Delta G^{\circ}_{gas\rightarrow oct}$ values was 0.59 kcal/mol. The cross-validated r^2 value was equal to 0.884, which again is similar to the r^2 value.

Table 12. Regression Statistics for the Four-Parameter GSSI Model of Gas-to-Octanol Transfer

predictor	coefficient	standard deviation	t ratio
constant	6.866	0.310	22.13
TSA^{acc}	−0.037	0.0001	−38.74
$\langle E_g^{elec}\rangle_g$	0.199	0.003	60.53
N_{rot}	0.605	0.005	113.63

4. Summary

We have developed a novel, semiempirical approach for the general treatment of solute−solvent interactions, which we call GSSI. Our GSSI approach is based on the principle that all solution-phase processes can be modeled in terms of one or more gas-to-solution transfer processes. Thus, if we can reliably calculate the free energy of gas-to-solution transfer, we can predict the free energies of desolvation, partition coefficients, and membrane permeabilities. The free energy of each gas-to-solution transfer process is calculated as the sum of the free energy of cavity formation and the free energy of solute−solvent interaction. The free energy of cavity formation is modeled in terms of the total solvent-accessible surface area of the solute. Both the enthalpic and entropic contributions to the free energy of solute−solvent interaction are explicitly addressed. The enthalpy of solute−solvent interaction is modeled on the basis of four modes of intermolecular interaction calculated at many points on the solute's solvent-accessible surface. The entropy of solute−solvent interaction is modeled on the basis of the effective number of rotatable bonds in the solute. While the solute's contributions are modeled explicitly, the solvent is characterized empirically by regression coefficients obtained by fitting a set of experimental data.

We have validated the GSSI approach by applying it to the prediction of free energies of gas-to-solution transfer for 111 solutes in water, 250 solutes in hexadecane, and 84 solutes in octanol. There was excellent agreement between predicted and experimental values for each data set with the seven-parameter GSSI model. For the aqueous solution data set, the GSSI model was able to explain 95.9% of the variance in the experimental data with a standard error of prediction of 0.476 kcal/mol. For the hexadecane and octanol data sets, the respective GSSI models were able to explain 96.1 and 92.7% of the variance in the experimental data with standard errors of prediction of 0.277 and 0.642 kcal/mol, respectively. Cross-correlation analysis confirmed our expectation that, for theoretical reasons, several of the GSSI solute descriptors were highly correlated. Thus, we were able to reduce the number of descriptors from six to three and yet still generate models that provided good agreement between calculated and experimental free energies of solvation. On the basis of the theoretical discussion in section 3.2 and for obvious statistical reasons, the GSSI software (distributed by Optive Research) does automatically eliminate cross-correlated descriptors (with the lesser variance) from all regression models.

Acknowledgment. This work was supported in part by the American Foundation for Pharmaceutical Education, to whom we are grateful.

MP034009U