Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium

3 AUTHORS, INCLUDING:

Christopher J Cramer
University of Minnesota Twin Cities
532 PUBLICATIONS   **23,458** CITATIONS

SEE PROFILE

Donald Truhlar
University of Minnesota Twin Cities
1,342 PUBLICATIONS   **81,415** CITATIONS

SEE PROFILE

Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium

Gregory D. Hawkins, Christopher J. Cramer,* and Donald G. Truhlar*

Department of Chemistry and Supercomputer Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431

The pairwise descreening approximation provides a rapid computational algorithm for the evaluation of solute shape effects on electrostatic contributions to solvation energies. In this article we show that solvation models based on this algorithm are useful for predicting free energies of solvation across a wide range of solute functionalities, and we present six new general parametrizations of aqueous free energies of solvation based on this approach. The first new model is based on SM2-type atomic surface tensions, the AM1 model for the solute, and Mulliken charges. The next two new models are based on SM5-type surface tensions, either the AM1 or the PM3 model for the solute, and Mulliken charges. The final three models are based on SM5-type atomic surface tensions and are parametrized using the AM1 or the PM3 model for the solute and CM1 charges. The parametrizations are based on experimental data for a set of 219 neutral solute molecules containing a wide range of organic functional groups and the atom types H, C, N, O, F, P, S, Cl, Br, and I and on data for 42 ions containing the same elements. The average errors relative to experiment are slightly better than previous methods, but—more significantly—the computational cost is reduced for large molecules, and the methods are well suited to using analytic derivatives.

1. Introduction

In previous papers[1−10] we and our co-workers have introduced a new approach to solvation modeling based on the following elements.

(1) Electrostatics are modeled by the generalized Born approximation[11−17] with atom-centered charges and dielectric descreening by other parts of the solute calculated from a conformationally sensitive solute-shape molecular model[1,18] based on overlapping atomic spheres. This electrostatic approach is called density descreening (DD) because it is based on integrating an approximation to the free energy density[19] of the polarized dielectric medium over the region not occupied by the descreening solute. The original applications[1−4,18] of DD used a rectangular quadrature rule for the radial integrations about each atomic center. This approximation in calculating the volume integral over the dielectric free energy density caused a systematic error which was compensated by parametrization.[1−4] (As discussed extensively in previous papers, and as is always true in semiempirical work, the parametrization also makes up in part for any systematic physical deficiencies of the model, e.g., for the approximation of the dielectric field in the free energy density by the Coulomb field.) Later work carried out this radial quadrature by a converged trapezoidal rule.[5−10]

(2) The atomic charges used in the electrostatic calculations are based on either class II[20] or class IV[21] charge models.

(3) First solvation shell effects (i.e., short-range effects) attributable to deviations of the solvation free energies from the predictions of a purely homogeneous continuum model of the solvent are calculated from atomic solvent-accessible surface areas[22,23] with micro-interfacial atomic surface tensions depending on the local nature of the solute.[24] The various functional forms of the surface tensions are denoted by SM*x*, in particular by the *x* in SM*x*, where *x* may be 1,[1,4] 1a,[1,4] 2,[2,4] 3,[3,4] 4,[6−9] or 5.[10] (The functional form of SM3 is the same as SM2.)

(4) Solute electronic and geometric relaxation in the presence of the solvent are modeled by including the solute−solvent interactions and the solute-induced changes in the solvent−solvent interactions self-consistently in a solute Fock operator,[1,12−16,25,26] and geometries are optimized in solution. Several general parametrizations of this kind of model have been proposed,[1−10] differing in the model used for the solute internal energy[27,28] and partial charges,[20,21] the quadrature method used for dielectric descreening calculations,[1,4,5,18] and the functional form used for the dependence of the surface tensions on the local nature of the solute.[1,2,6,10] Table 1 summarizes previous parametrizations. (The models for which the reference is "present" are new in this paper and are explained below.)

In a recent letter,[29] we introduced a computationally appealing modification of our previous approach in which the three-dimensional quadratures[1,5,19] in the original dielectric descreening calculations are replaced by a sum of pair terms involving scaled atoms, and we presented a preliminary set of parameters for solutes containing C, H, O, and N in aqueous solution. This electrostatic approach, which was inspired by the work of Schaefer and Froemmel,[30] is called pairwise descreening (PD). (The PD approach and the DD approach both utilize the approximation that the electric field is replaced by the Coulomb field in the free energy density.) The parameters required were the same as in the previous SM2[2] and SM2.1[5] models plus C, H, O, and N scale factors for the new step. This was called solvation model 2.2 because it closely resembles SM2[2] and SM2.1.[5]

A major component of the widespread usefulness of modern electronic structure theory is associated with the availability of a variety of "levels" differing in cost and reliability. Thus, various combinations of computational choices may be more or less useful depending on the problem being studied. For example, one has various levels of one-electron basis sets and various degrees of the inclusion of electron correlation. The various methods of treating solvation in our own previous work, e.g., SM2[2] and SM4,[6] have not been developed in a level framework. Rather, they reflect different choices in the functional forms used for modeling first solvation shell effects and different algorithms for calculating electrostatic descreening

---

TABLE 1: SM*x* Models

SM*x*	ref	solute energy	partial charges	dielectric descreening[a]	surface tension dependence	solutes no.[b]	solutes heteroatoms[c]	solvents
				Density Descreening				
SM1	1, 4	AM1	AM1 Mulliken	RD	Z^d	145	N O F P S Cl Br I	H_2O
SM1a	1, 4	AM1	AM1 Mulliken	RD	Z, MM type[e]	145	N O F P S Cl Br I	H_2O
SM2	2	AM1	AM1 Mulliken	RD	Z, bond orders	150	N O F P S Cl Br I	H_2O
SM2.1	5	AM1	AM1 Mulliken	DD	Z, bond orders	148	N O F P S Cl Br I	H_2O
SM3	3	PM3	PM3 Mulliken	RD	Z, bond orders	150	N O F P S Cl Br I	H_2O
SM3.1	5	PM3	PM3 Mulliken	DD	Z, bond orders	148	N O F P S Cl Br I	H_2O
SM4	6, 7	AM1 or PM3	CM1A or CM1P	DD	Z, bond orders	176	N O F S Cl Br I	alkanes
SM4-SRP (Claisen)	8	AM1	CM1A	DD	Z, bond orders	39	O	H_2O
SM4-SRP (sugar)	9	AM1	CM1A	DD	Z, bond orders	47	O	H_2O
SM5.4/U	10	unspecific[g]	accurate[g]	DD	Z, distances	215	N O F S Cl Br I	H_2O
SM5.4/A	10	AM1	CM1A	DD	Z, distances	215	N O F S Cl Br I	H_2O
SM5.4/P	10	PM3	CM1P	DD	Z, distances	214	N O F S Cl Br I	H_2O
				Pairwise Descreening				
SM2.2[f]	29	AM1	AM1 Mulliken	PD	Z, bond orders	139	N O	H_2O
SM2.2PD/A	present	AM1	AM1 Mulliken	PD	Z, distances	219	N O F P S Cl Br I	H_2O
SM5.2PD/A	present	AM1	AM1 Mulliken	PD	Z, distances	219	N O F P S Cl Br I	H_2O
SM5.2PD/P	present	PM3	PM3 Mulliken	PD	Z, distances	214	N O F S Cl Br I	H_2O
SM5.4/U	present	unspecific[g]	accurate[g]	PD	Z, distances	215	N O F S Cl Br I	H_2O
SM5.4PD/A	present	AM1	CM1A	PD	Z, distances	215	N O F S Cl Br I	H_2O
SM5.4PD/P	present	PM3	CM1P	PD	Z, distances	214	N O F S Cl Br I	H_2O

[a] DD denotes density descreening with converged quadratures, RD denotes rectangle rule, and PD denotes pairwise descreening. [b] This column gives the number of neutral solutes used in the parametrization. Ionic solutes were also used to check the reasonableness of each model, and in most cases ionic solutes were used in the parametrization as well. [c] All models are parametrized for solutes containing H and C plus one or more heteroatoms. [d] Z denotes atomic number. [e] MM type denotes "molecular mechanics type", e.g., molecular mechanics recognizes carbonyl and alcoholic oxygens as different types of atoms that have different parameters. [f] SM2.2 should be considered a preliminary version of the present SM2.2PD/A model and is not recommended for further use. [g] The U models have been parametrized using both AM1 and PM3 solute Hamiltonians with both CM1A and CM1P partial charges in an attempt to develop models with "unspecific" character that may be used with more than one type of partial charges. For example, SM5.4/A was parametrized using 215 neutral-solute data points, SM5.4/P was parametrized using 214 neutral-solute data points, and SM5.4/U was parametrized using 429 neutral-solute data points.

effects. The choices between pairwise descreening and density descreening or between class II and class IV charges are, however, comparable to choices available in electronic structure theory, and various combinations of such choices represent compromises between the faithfulness of the model to the physics of the situation and the cost of the computations. The new models introduced in the present paper may be considered as the systematic introduction of "levels" into our approach; in particular we introduce a pairwise-descreening level into the SM2[2] method, and we introduce both pairwise descreening and lower level charges into the SM5.4/U,[10] SM5.4/A,[10] and SM5.4/P[10] models.

The present article is a followup on the original pairwise descreening paper,[29] and it presents six new general parametrizations, all based on experimental training data. All five models use a single fixed intrinsic Coulomb radius for each atomic number Z (where Z equals 1, 6, 7, ... for H, C, N, ...). The first model uses the surface tension prescription of the SM2 approach, as well as Mulliken charges and the pairwise descreening approximation for electrostatics. The only extension of the SM2 surface tension functional forms is the addition of a new surface tension term that allows the model to treat disulfide molecules more accurately. The new parametrization is called SM2.2PD/A. Here the SM2 prefix denotes SM2 functional forms for the surface tensions, the 2 after the point denotes class II charges, the PD denotes the pairwise descreening algorithm, and the A after the slash signifies the AM1 solute Hamiltonian. Two other new models are called SM5.2PD/A and SM5.2PD/P. In this naming convention, the SM5 refers to the SM5 surface tension functional forms,[10] the presence of a decimal point means that the electrostatics are modeled using solute point charges, the 2 after the point again stands for class II charges, the PD again represents pairwise descreening, the A after the slash represents the AM1 Hamiltonian, and the P after the slash represents the PM3 Hamiltonian. Two other models presented here,

SM5.4PD/A and SM5.4PD/P, share the characteristics of the previously mentioned two models, except they are based on class IV[21] charges. Thus, SM5.4 implies the combination of the SM5 functional forms for surface tensions and class IV charges in the electrostatics. Finally we will present a model called SM5.4PD/U, which is parametrized with the same data and functional forms, but is proposed as a Hamiltonian-unspecific model. For further details and technical differences of the models, see section 3.

As described in a preceding paper,[10] our more recent models (SM4, SM5.4/U, SM5.4/A, and SM5.4/P) have a very significant advantage over the SM3.1[5] and earlier models[1−5] in that they are based on much more accurate atomic charges. Why then should we develop new models that use Mulliken charges; why not just models based on class IV charges? The answer is simplicity and computational efficiency. A similar motivation applies to the use of pairwise descreening. The pairwise descreening algorithm makes it possible to calculate solvation energies of very large molecules, and the SM2.2 and SM5.2 types of models are less computationally expensive than the SM5.4 models, allowing for treatment of the largest possible solutes. We will elaborate on this point in section 5.

As also described in the preceding paper,[10] the SM5.4-type models are more robust than SM2 and SM3 in that the same surface tensions may be used with either AM1 or PM3 wave functions with little loss of accuracy for most solutes. (These "unspecific" surface tensions are obtained as an intermediate step of the parametrization.) We will see below that the SM5.4 PD models share this robustness, which is not surprising since this desirable feature was attributed[10] to the use of class IV charges. Nevertheless, since there are many applications where getting the right answer is more important than elegance, we do not use the compromise surface tensions for our final parameters, but rather we use parameter sets optimized specifically for use with AM1 or PM3. In the present paper, we

present two parametrizations that utilize class IV charges and SM5-type surface tensions, namely, SM5.4PD/A and SM5.4PD/P; these models have identical parameters for the calculation of the electronic-nuclear-polarization free energy and differ only in the magnitudes of the atomic surface tensions, which were optimized to obtain the best possible results. In contrast, the universal parameters constitute a model called SM5.4PD/U.

Section 2 reviews the theory. Section 3 discusses the parametrizations, and section 4 gives results. Discussion is presented in section 5, and concluding remarks are in section 6.

2. Theory

In the SMx models, the standard-state free energy of solvation is written as[1−6]

$$\Delta G_{S}^{\circ} = \Delta G_{ENP} + G_{CDS} \qquad (1)$$

where

$$\Delta G_{ENP} = \Delta E_{EN} + G_{P} \qquad (2)$$

$$G_{CDS} = \sum_{k} \sigma_{k} A_{k} \qquad (3)$$

ΔE_{EN} is the change in the solute's internal free energy upon insertion in solution (approximated here, as usual, as the change ΔE_{EN} in the sum of the solute electronic total energy and nuclear repulsion energy), G_{P} is the electric polarization energy of the solvent (including both the solute−solvent and solvent−solvent contributions), G_{CDS} accounts for first solvation shell effects, σ_{k} is the atomic surface tension of atom k, and A_{k} is the solvent-accessible surface area of atom k. We use a standard state of one mole of solute per liter in both the gas phase and solution. The details of ΔE_{EN}, G_{P}, and σ_{k} are explained elsewhere;[1−6] for the present purposes, though, it is important to remind the reader that the SM2 and SM5-type solvation models contain several kinds of parameters:

(i) intrinsic Coulomb radii ρ_{k} in G_{P}; (ii) the parameter $d_{kk'}$ in the Still−Tempczyk−Hawley−Hendrickson[18] generalization of the Ohno−Klopman[31,32] formula, used in G_{P}, and the parameters in the $C_{kk'}$ term that we added[2,4] to this equation; (iii) van der Waals[33] and solvent[7,22] radii used in G_{CDS}; (iv) nonlinear parameters used in expressing the dependence[1,2,7,10] of atomic surface tensions on local molecular geometry or bond orders; (v) surface tension coefficients.

In addition to these parameters the present paper will involve a sixth type of parameter, namely, scaling factors[29] for the pairwise descreening step, in particular (vi) scale factors $S_{kk'}$ by which the radius of atom k is scaled when it descreens atom k'.

Notice that the free energy is linear in the surface tension coefficients but depends on all other parameters nonlinearly. This is important in devising a strategy for optimizing parameters.

With regard to parameters of type (iii) we emphasize that the calculation of solvent-accessible surface areas in all SM5 models is based on Bondi's values of atomic radii[33] and effective solvent radii at least as large as the radius of water. Although one might obtain a bit smaller mean errors by optimizing these radii to smaller values, that would change the interpretation. In the present work, parameters of types (iii) and (iv) are taken, when available, from previous models without change: For SM2.2PD/A the parameters were taken from the SM2 model, and for the SM5 PD models the parameters were taken from SM5.4/U[10] (note that these parameters are the same in SM2.1

as in SM2, and they are the same in SM5.4/A and SM5.4/P as in SM5.4/U). The only new parameters of this type that enter in the present work are required because of the addition of an SM5-type surface tension for sulfur interacting with sulfur in SM2.2PD/A (since this is a very minor detail, it is not reflected in the systematic notation) and because we developed SM5-type models for phosphorus.

For the six new models presented here, the number of parameters of type (i) have been reduced as compared to the number that were used in the SM2[2,4] and SM5.4 DD[10] models presented previously. In particular, each new model uses only a single fixed intrinsic Coulomb radius for each atomic number. The calculation of type (ii) and type (vi) parameters is discussed below. Finally, the surface tension coefficients (v) have been reoptimized for each model to produce a best fit to experimental results.

The polarization free energy of inserting a molecule in a medium of dielectric constant ϵ is expressed in the generalized Born approximation as

$$G_{P} = -\frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\sum_{k,k'} q_{k} q_{k'} \gamma_{kk'} \qquad (4)$$

where q_{k} is the partial charge on atom k, and $\gamma_{kk'}$ is a Coulomb integral. The Coulomb integrals $\gamma_{kk'}$ ($k \neq k'$) are modeled such that they reduce to the self Coulomb integrals γ_{kk} when the distance $R_{kk'}$ between identical atoms tends to zero, and they are asymptotic to $1/R_{kk'}$, where $R_{kk'}$ is the interatomic distance between k and k', at large distances.[1,10−18] To accomplish this, $\gamma_{kk'}$ for $k \neq k'$ is an explicit function of the diagonal elements γ_{kk} and the distances $R_{kk'}$. Still et al.[18] introduced an expression for $\gamma_{kk'}$ that accounts for the nonhomogeneous character of each atom's local surroundings. This expression for $\gamma_{kk'}$ for $k \neq k'$ was further modified for the SM1 and SM2 models, and in its most general form it is given by

$$\gamma_{kk'} = \left(R_{kk'} + \alpha_{k}\alpha_{k'}\left(\exp\left(\frac{-r_{kk'}^{2}}{d_{kk'}\alpha_{k}\alpha_{k'}}\right) + C_{kk'}\right)\right)^{-1/2} \qquad (5)$$

where α_{k} is an effective atomic radius, and $d_{kk'}$ is an empirically optimized constant. When $C_{kk'}$ is equal to zero, eq 5 becomes the equation prescribed by Still et al.,[18] and it also matches the equation used in the SM5 family of solvation models.

In SM2, $C_{kk'}$ is represented

$$C_{kk'} =$$
$$\begin{cases} d_{kk'}^{(1)} \exp\left(\dfrac{-d_{kk'}^{(2)}}{1 - [(R_{kk'} - R_{kk'}^{(1)})/R_{kk'}^{(2)}]^{2}}\right), & |R_{kk'} - R_{kk'}^{(1)}| < R_{kk'}^{(2)} \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

where $d_{kk'}^{(1)}$, $d_{kk'}^{(2)}$, $R_{kk'}^{(1)}$, and $R_{kk'}^{(2)}$ are constants (parameters) as described previously.[1−4] We note that $C_{kk'}$ was nonzero only for k, k' equal to N, H (or equivalently H, N) and O, O in SM2. These functions were introduced in the SM1−SM3 models to empirically correct for certain trends in the experimental data that were not treated accurately without them. These functions are not used in SM5 models. Although (as discussed further in sections 5.2 and 5.3) the partition of free energy into electrostatic and first solvation shell effects is not unique, some readers may consider that the SM5 electrostatics are more "physical" for this reason, but this is not experimentally verifiable, even in principle.

When the atom is totally surrounded by solvent (the monatomic case), γ_{kk} is equal to the reciprocal of the atom's intrinsic Coulomb radius ρ_{k}. However, when the atom is part of a solute

Aqueous Free Energies of Solvation

J. Phys. Chem., Vol. 100, No. 51, 1996 **19827**

with two or more atoms, the other solute atoms block it from the screening effect of the dielectric medium, and the extent of this descreening is determined by a volume integral over the dielectric free-energy density[18,19,29,30] in the space surrounding each atom. The form of the integral was originally based on numerically integrating this density in a shell of solvent around the solute atom being considered. In ref 29, we showed that this numerical integral can be replaced by an algorithm, similar to one employed earlier by Schaefer and Froemmel,[30] which assumes that each atom in the solute may be considered to make an independent contribution to the descreening of other atoms from the dielectric.

Using the approximations introduced by Still *et al.*,[18] the Coulomb integral, γ_{kk}, can be represented as the following dielectric descreening integral in atomic units:

$$\gamma_{kk} = \alpha_k^{-1} =$$
$$\int_{\rho_k}^{\infty}dr \int_0^{\pi}d\theta \int_0^{2\pi}d\phi \frac{\sin\theta}{4\pi r^4}W(r,\theta,\phi,\{R_{kk'},\rho_{k'}\}_{\text{all }k,k'}) \quad (7)$$

where α_k is the effective Born radius, W is unity if point r, θ, ϕ is in a region of dielectric constant ϵ and zero if it is in a region of dielectric constant unity (regions inside solute atoms have dielectric constant unity because their polarization is included explicitly in the self-consistent-field calculation), and the set of $R_{kk'}$ specifies the geometry of the system, which is assumed to consist of a set of spherical atoms with intrinsic radii ρ_k, one for each atomic number. Applying the pairwise descreening approximation with the scale factors $S_{kk'}$ to eq 7, the expression simplifies to

$$\alpha_k^{-1} = \rho_k^{-1} - \sum_{k'}\int_{\rho_k}^{\infty}\frac{dr}{r^2}H_{kk'}(R_{kk'},\rho_{k'}) \quad (8)$$

where $H_{kk'}$ is the fraction of the area of a sphere of radius r centered at atom k that is shielded by a sphere of scaled radius $S_{kk'}\rho_{kk'}$ at a distance $R_{kk'}$.

Equation 8 shows that using the pairwise descreening approximation to calculate effective Born radii eliminates the necessity of carrying out the numerical integration in eq 7 and hence greatly reduces the number of times the computationally expensive solvent accessible surface area calculation needs to be completed, but if the scale factors $S_{kk'}$ were not used, it would introduce its own systematic error. For example, suppose one wishes to calculate the effective Born radius for a sphere k in a set of spheres $\{k'\}$. When incorporating the pairwise approximation, the volume integral in eq 7 is exactly reproduced by eq 8 only if no two of the k' spheres overlap each other. If two of the k' spheres overlap, then the volume integral will be overestimated. This causes the effective Born radius to be too large, and ultimately the calculated polarization free energy will be underestimated with respect to the numerical integration method. Since van der Waals spheres of covalently bonded atoms overlap to a large extent, the systematic error could be large. To combat this systematic error, we have chosen to utilize a set of multiplicative scale factors $\{S_{kk'}\}$ for the radius of each of the k' spheres. By using $S_{kk'}\rho_{k'}$ instead of $\rho_{k'}$ in eq 8, we account for the overlap of the van der Waals spheres. The scale factors were determined (along with surface tension coefficients and intrinsic Coulomb radii) by fitting the model to experimental free energies of solvation; the scale factors were initially restricted to lie in a range between 0.5 and 1.0.

The simplification of the method, then, results from the fact that it is possible to calculate $H_{kk'}$ analytically.[29] In particular,

the modified eq 8 becomes

$$\alpha_k^{-1} = \rho_k^{-1} - \frac{1}{2}\sum_{k'}\int_{L_{kk'}}^{U_{kk'}}dr\left(\frac{1}{r^2} - \frac{R_{kk'}}{2r^3} - \frac{1}{2R_{kk'}r} + \frac{S_{k'k}^2\rho_{k'}^2}{2R_{kk'}r^3}\right)$$

$$= \rho_k^{-1} - \frac{1}{2}\sum_{k'}\left[\frac{1}{L_{kk'}} - \frac{1}{U_{kk'}} + \frac{R_{kk'}}{4}\left(\frac{1}{U_{kk'}^2} - \frac{1}{L_{kk'}^2}\right) + \right.$$
$$\left. \frac{1}{2R_{kk'}}\ln\frac{L_{kk'}}{U_{kk'}} + \frac{S_{k'k}^2\rho_{k'}^2}{4R_{kk'}}\left(\frac{1}{L_{kk'}^2} - \frac{1}{U_{kk'}^2}\right)\right] \quad (9)$$

where

$$L_{kk'} = \begin{cases} 1 & \text{if } R_{kk'} + S_{k'k}\rho_{k'} \leq \rho_k \\ \rho_k & \text{if } R_{kk'} - S_{k'k}\rho_{k'} \leq \rho_k < R_{kk'} + S_{k'k}\rho_{k'} \\ R_{kk'} - S_{k'k}\rho_{k'} & \text{if } \rho_k \leq R_{kk'} - S_{k'k}\rho_{k'} \end{cases} \quad (10)$$

and

$$U_{kk'} = \begin{cases} 1 & \text{if } R_{kk'} + S_{k'k}\rho_{k'} \leq \rho_k \\ R_{kk'} + S_{k'k}\rho_{k'} & \text{if } \rho_k < R_{kk'} + S_{k'k}\rho_{k'} \end{cases} \quad (11)$$

We note that the expression in eq 9 for the PD potential is, up to a constant factor, identical to the function F_4 given by Schaefer and Froemmel.[30]

3. Parametrization

3.1. Preliminaries. As previously reported,[29] the pairwise descreening approximation was incorporated into the AMSOL[34] computer code in the form of the expression in eq 9 for the effective Born radius. AMSOL, a semiempirical quantum chemistry program, performs SCF calculations using the NDDO approximation[35,36] including free energy of solvation terms added to the Fock operator to account for solvation effects.

A second preliminary step is the selection of a training set. For four of the new models, we employed the SM5.4 aqueous models neutral and ionic training sets developed in earlier work.[10] This set was selected[10] based on (i) availability of experimental data,[37-40] (ii) the existence of well-defined dominant conformations that are qualitatively reproduced by semiempirical molecular orbital theory, and (iii) a balanced representation of various functional groups. The neutral training sets contains 215 compounds, and the ionic training set contains 34 compounds. Together, these 249 molecules comprise the training set for the SM5.4PD/U and SM5.4PD/A models. The training set for the two PD/P models contain only 248 molecules because hydrazine, the structure of which is not accurately predicted by PM3 in the gas phase, was removed from the set.

Since the SM2.2PD/A and SM5.2PD/A models were parametrized for phosphorus-containing molecules, the SM5.4 aqueous models training set was augmented by four neutral and eight ionic phosphorus-containing molecules to make 261 molecules in the training sets for those models. The added neutral compounds are PH_3, $OP(OCH_3)_3$, $OP(OC_2H_5)_3$, and $OP(OC_3H_7)_3$, and the added ionic compounds are PH_2^-, PH_4^+, $MePH_3^+$, $Me_2PH_2^+$, Me_3PH^+, $H_2PO_4^-$, HPO_4^{2-}, and PO_4^{3-}.

We did not attempt to parametrize for P in SM5.2PD/P because our previous experiences showed that PM3 is quite inaccurate for partial charges at phosphorus centers.[41]

The sources of experimental data and conformations for the non-phosphorus compounds have been described previously.[10] The experimental result for phosphine was taken from Wagman

et al.,[42] while data for the three phosphate esters were available from Wolfenden.[43] Pearson provided an experimental free energy of solvation for the first ionic phosphorus compound and estimates for the free energy of solvation for the next four phosphorus ions.[39] ΔG_S^o is not actually available for the final three phosphorus ions, $H_2PO_4^-$, HPO_4^{2-}, PO_4^{3-}, but in an attempt to better constrain the results for phosphorus-containing ions to lie in a physical regime, we employed the known enthalpies of solvation[44] for these ions in place of experimental ΔG_S^o values.

3.2. Parameters to be Optimized. As mentioned above, for each of the six PD models presented here, the intrinsic Coulomb radius, ρ_k, in eqs 4−10 is a single constant parameter for each atomic number. This differs from the scheme used in the original SM2 model, where the Coulomb radius was a smooth sigmoidal function of the atomic charge,[4] and it differs from the SM5.4 DD models, where the Coulomb radius for hydrogen is adjusted as a function of geometry for molecules where the hydrogen is bonded to nitrogen or oxygen atoms.[10] Thus, since SM2.2PD/A is parametrized for H, C, N, O, F, P, S, Cl, Br, and I atoms, there are 10 intrinsic Coulomb radius parameters. The same is true for SM5.2PD/A. Since SM5.2PD/P and the models based on SM5.4 do not contain parameters for phosphorus, there are only nine Coulomb radius parameters for these models.

In both models the solvation free energy is given by eqs 1−3, but we must specify the form of the atomic surface tensions σ_k. In the SM2.2PD/A model, the atomic surface tensions are modeled as in SM2,[2,4] in particular,

$$\sigma_k = \sigma_k^{(0)} + \sigma_k^{(1)}[f(B_k^H) + a_k\hat{g}(B_k^H)] \qquad (12)$$

where $\sigma_k^{(0)}$, $\sigma_k^{(1)}$, and a_k are empirically derived surface tension coefficients. The coefficients are all zero for hydrogen atoms in the SM2 approach. Furthermore, $\sigma_k^{(0)}$ is nonzero for all other atoms, $\sigma_k^{(1)}$ is nonzero for C, N, O, and S, and a_k is nonzero for N and O. $f(B_k^H)$ and $\hat{g}(B_k^H)$ are functions of the sum of the bond orders of hydrogen to atom k, where

$$f(B_k^H) = \tan^{-1}(\sqrt{3}B_k^H) \qquad (13)$$

and

$$\hat{g}(B_k^H) = \begin{cases} \exp\left(\dfrac{-b_k}{1 - [(B_k^H - c_k)/w_k]^2}\right), & |B_k^H - c_k| < w_k \\ 0, & \text{otherwise} \end{cases} \qquad (14)$$

The sum of bond orders is defined by

$$B_k^H = \sum_{k'=H} B_{kk'} \qquad (15)$$

where $B_{kk'}$ is the bond order (defined elsewhere[45]) of atom k to atom k'. The constants b_k, c_k, and w_k were not changed from their values[2,4] in SM2. The proportionality constants (surface tension coefficients $\sigma_k^{(0)}$, $\sigma_k^{(1)}$, and a_k) used in eq 12 depend on the local nature of the solute for each atom's or group-of-atoms' interface with the solvent. Since $\sigma_k^{(0)}$ is nonzero for every atom type except hydrogen, and $\sigma_k^{(1)}$ is nonzero only for the C, N, O, and S atom types, there are 13 surface tension coefficients for SM2.2PD/A, excluding the two a_k parameters.

In the models employing SM5-type surface tensions, the functional dependencies of the σ_k are different for each atomic number. In particular, σ_k for phosphorus and halogens are

constant (called $\sigma_x^{(0)}$ where x is F, P, Cl, Br, or I), and

$$\sigma_k|_{k=H} = \sigma_H^{(0)} + \sum_{k'=C,O,N,S} \sigma_{Hk'}T_{Hk'}(R_{kk'}) \qquad (16)$$

$$\sigma_k|_{k=C} = \sigma_C^{(0)} + \sum_{i=1}^{2}\sigma_{CC}^{(i)}\sum_{\substack{k'=C \\ k'\neq k}} T_{Ck'}^{(i)}(R_{kk'}) \qquad (17)$$

$$\sigma_k|_{k=N} = \sigma_N^{(0)} + \sigma_{NC}\{\sum_{k'=C} T_{Nk'}(R_{kk'})[\sum_{\substack{k''\neq k \\ k''\neq k'}} T_{Nk''}(R_{k'k''})]^2\}^{1.3} \qquad (18)$$

$$\sigma_k|_{k=O} = \sigma_O^{(0)} + \sigma_{OC}\sum_{k'=C} T_{Ok'}(R_{kk'}) + \sigma_{OO}\sum_{\substack{k'=O \\ k'\neq k}} F_{Ok'}(R_{kk'}) +$$
$$\sigma_{ON}\sum_{k'=N} T_{Ok'}(R_{kk'}) + \sigma_{OP}\sum_{k'=P} T_{Ok'}(R_{kk'}) \quad (19)$$

and

$$\sigma_k|_{k=S} = \sigma_S^{(0)} + \sigma_{SS}\sum_{\substack{k'=S \\ k'\neq k}} T_{Sk'} \qquad (20)$$

where the $T_{kk'}$ and $F_{kk'}$ functions are given in ref 10. The parameters in these functions, except for those in T_{OP}, are not changed in this work and are also given in ref 10 (ref 10 did not include P). The choice of functional forms for P and the parameters in T_{OP} are discussed in the Supporting Information. There are 20 surface tension coefficients for each SM5 model, except SM5.2PD/A, which has two additional surface tensions coefficients, $\sigma_P^{(0)}$ and σ_{OP}, for phosphorus.

Notice that the parametrized scale factor $S_{kk'}$ introduced in conjunction with eq 8 depend on the descreened atom k as well as the descreening atom k'. We allowed this possibility because the average overlap with other atoms of the atoms of the type k' that are most important for descreening atoms of type k may differ from the average overlap with other atoms of the atoms of the type k' that are most important for descreening atoms of type k''. In preliminary work, we ascertained which pairwise interactions within the pairwise descreening approximation benefited significantly from individual scaling factors, $\{S_{kk'}\}$. To accomplish this, we used the parameters developed for the SM2.1 model and the pairwise descreening approximation. Since there are 10 supported atomic numbers within SM2.1, there were 100 possible scaling factors. From this number, we eliminated any potential scale factors that our parametrization set could not utilize. For example, since there is no molecule in our parametrization set that contains more than one phosphorus, we could not possibly determine a parameter for S_{PP}. (Thus in the final model S_{PP} is defined to be the same as S_{Pk} for all other k, except $k = H$, which has a special value, S_{PH}.) Then for each of the remaining scale factors, we ran two SCF calculations for each neutral training set molecule at the optimized SM2.1 geometry, the first one with a particular scale factor set to 1.00 and the second with it set to 0.90. (The other scale factors for such a pair of runs were set to 1.00.) In general, we found that the scale factors where hydrogen is the descreening or descreened atom were the parameters to which the accuracy is most sensitive. Thus, we decided to determine a unique scale factor for hydrogen descreening any other atomic number and for any other atomic number descreening hydrogen. In addition, we optimized a scale factor for every atomic number other than 1 (hydrogen) descreening nonhydrogenic atoms. Thus, for models containing phosphorus we optimized 28 scaling factors,

Aqueous Free Energies of Solvation

J. Phys. Chem., Vol. 100, No. 51, 1996 **19829**

and for models not including phosphorus we optimized 25 scaling factors.

The final parameters to be optimized are the $d_{kk'}$. In models SM1 through SM4, all $d_{kk'}$ were set equal to 4.0, the value used by Still *et al.*[18] In SM5.4/U, SM5.4/A, and SM5.4/P, we retained $d_{kk'} = 4.0$, except for d_{CH} ($\equiv d_{HC}$), which was given the value 4.2.[10] In the present work we treated d_{CH} as special again. We therefore optimized two values. One, which we will call d_{CH}, is used for $k' = H$, and for $k = H$, $k' = X$. The other, which we call d_0, is used for all other k, k' combinations.

3.3. Steps in the Parametrization. Our general strategy was to first consider the solutes in the training set containing at most C, H, and/or O and find the parameters, including d_0 and d_{CH}, affecting these atom types. We then froze these parameters and found the new parameters necessary to treat N, S, and P. Since no molecule in the training set contains two or more of these new types, these elements could be added one at a time. However, some of the halo compounds contain S, so we then froze C, H, O, and S and found the new parameters necessary to treat halogenated compounds. Thus there are five stages in the parametrization: (i) CHO, (ii) N, (iii) S, (iv) P, and (v) halogens.

At each stage of the parametrization, we considered both neutrals and ions simultaneously and simultaneously optimized the nonlinear ρ_k and $S_{kk'}$ parameters and the linear surface tension coefficients by a genetic algorithm[46] procedure discussed below. We note here that the genetic algorithm optimized nonlinear parameters by considering the mean unsigned error over the combined set of neutrals and ions, with ions weighted one-sixth of neutrals. In the genetic algorithm steps, surface tension coefficients (which are linear parameters) were always optimized for fixed values of the nonlinear parameters by minimizing the sum of the squares of the errors for neutral solutes. The parametrization steps were carried out using fixed geometries. (The *final* results of the method involve calculating solvation free energies by comparing calculations at solution-phase geometries optimized with the final parameters to gas-phase calculations at optimized gas-phase geometries.) The gas-phase geometries were taken from AM1 calculations for parametrizing PD/A calculations and for AM1 data points in parametrizing the PD/U method and from PM3 calculations for parametrizing PD/P calculations and for PM3 data points in parametrizing the PD/U method. The aqueous-phase geometries were taken from aqueous-phase SM2.1 calculations for the parametrization of SM2.2PD/A and from aqueous-phase SM5.4/A calculations for the parametrization of SM5.2PD/A, SM5.4PD/A, and the AM1 data points in parametrizing the SM5.4PD/U method. The aqueous-phase SM5.4/P geometries were used for parametrizing SM5.2PD/P, SM5.4PD/P, and the PM3 data points in SM5.4PD/U.

The parameters of SM2.2PD/A, SM5.2PD/A, SM5.2PD/P, and SM5.4PD/U were optimized independently of all other models. However, the nonlinear σ_k and $S_{kk'}$ parameters for SM5.4PD/A and SM5.4PD/P models were not optimized independently. We forced the nonlinear parameters in these models to be the same as in the SM5.4PD/U model. For the SM5.4PD/U model, we calculated the mean unsigned error over a combined set of 249 AM1-SM5.4 PD calculations on neutrals and ions and 248 PM3-SM5.4 PD calculations on neutrals and ions. The resulting parameters (including a set of compromise surface tension coefficients) are called the SM5.4PD/U parameters, where U denotes that they are "unspecific" as to whether AM1 or PM3 is used for the solute Hamiltonian and partial charges. The linear surface tension coefficients were then optimized for SM5.4PD/A by using the SM5.4PD/U nonlinear

parameters and considering the sum of the squares of the errors for the 215 neutral solutes in the training set for SM5.4PD/A. The linear surface tension coefficients for SM5.4PD/P were found similarly using the 214 neutral solutes in the training set for SM5.4PD/P.

The parametrization of d_0 and d_{CH} was undertaken after finding the other H, C, and O parameters in four of the models and was completed in conjunction with finding the H, C, and O parameters in the other two models as discussed in section 5.4.

3.4. Details of the Parametrization. We started with the 107 neutrals and 13 ions in the parametrization set that contains only H, and/or C, and/or O atoms. This group is called the HCO training subset. As mentioned in previous papers,[4,10] the uncertainty in the experimental free energies of solvation for ionic compounds is typically 5 kcal (in this paper all energies in kcal are on a per mole basis), which is about 25 times greater than the typical uncertainty in the neutral values. We therefore chose to weight the ions less heavily than the neutrals in devising an unfitness function for the genetic algorithm. After a small amount of experimenting with various weights, we eventually chose a weighting factor of 1/6 for the ions because this appeared (by trial-and-error) to be a good compromise for parametrization.

We used the genetic algorithm implemented in GAFORTRAN-version 1.5.1,[46] to find the best values for the HCO parameters. The genetic algorithm was defined to evolve a population in combined ρ_k and $S_{kk'}$ parameter space. For each point in this space we need to define an unfitness function that the genetic algorithm attempts to minimize by evolving these parameters, and this was done as follows. First, for a given set of $\{\rho_k, S_{kk'}\}$ parameters, we optimized the linear coefficients to minimize the sums of the squares of the errors in the neutrals. Then the unfitness was calculated:

$$U = \frac{\left(\sum_{n=1}^{N} |G_S^\circ(\exp_n) - G_S^\circ(\text{calc}_n)| + \frac{1}{6}(\sum_{i=1}^{I} |G_S^\circ(\exp_i) - G_S^\circ(\text{calc}_i)|) \right)}{N + I} \quad (21)$$

where N is the total number of neutral molecules in the training subset, I is the total number of ionic compounds, $\Delta G_S^\circ(\exp)$ is the experimental standard-state free energy of solvation, and $\Delta G_S^\circ(\text{calc})$ is the standard-state free energy of solvation calculated with a given set of parameters. For the HCO training subsets, there were three intrinsic Coulomb radii parameters and seven scaling factors. For each set of parameters tested, a linear regression was employed to find surface tension coefficients that minimized the sum of the squared errors between the calculated free energies of solvation and the experimental free energies of solvation for the neutrals. Using these coefficients, the G_{CDS} term was calculated for both the neutrals and ions then added to the calculated ΔE_{EN} and G_P terms to yield the calculated free energy of solvation, $\Delta G_S^\circ(\text{calc})$. The a_k parameters, which were mentioned in conjunction with SM2-type surface tensions, were set to zero for all atoms in this step.

For the genetic algorithm, appropriate jump mutation and creep mutation probabilities were determined by the formulas recommended in GAFORTRAN-version 1.5.1.[46] Elitism, uniform crossover, and niching were also used as recommended. Each set of parents was allowed one offspring, and there were 50 individuals per generation. To speed the convergence of the genetic algorithm, relatively few values of any single parameter were allowed. Typically, 25 or less evenly spaced values were allowed per parameter within a physically meaningful range.

19830 *J. Phys. Chem., Vol. 100, No. 51, 1996*

Hawkins et al.

TABLE 2: Optimized Scale Parameters by Atom Type

		$S_{kk'}$			
			SM5.2PD		SM5.4PD/
descreening atom	descreened atom	SM2.2PD/A	/A	/P	A, P, U[a]
H	H	0.50	0.50	0.71	0.84
H	C	0.91	1.00	1.00	1.00
H	N	0.96	0.50	1.00	0.81
H	O	0.50	0.50	0.50	0.50
H	F	0.97	0.50	0.50	0.50
H	P	1.00	1.00		
H	S	0.91	0.50	0.50	0.50
H	Cl	0.50	0.50	0.50	0.50
H	Br	0.50	0.50	0.50	1.00
H	I	1.00	0.50	1.00	1.00
C	H	0.77	0.77	0.72	0.76
N	H	0.88	1.00	0.90	0.75
O	H	0.91	1.00	0.50	0.50
F	H	0.67	0.66	0.50	0.65
P	H	0.64	0.98		
S	H	1.10	1.10	1.10	0.78
Cl	H	1.00	0.97	0.50	0.76
Br	H	0.87	0.82	0.50	0.59
I	H	0.96	1.00	0.50	0.63
C	all but H	0.50	0.59	0.50	0.67
N	all but H	0.50	0.50	0.50	0.52
O	all but H	0.96	1.00	1.00	1.00
F	all but H	1.00	0.50	0.50	1.00
P	all but H	0.50	0.50		
S	all but H	1.10	1.10	1.10	1.00
Cl	all but H	1.00	1.00	1.00	1.00
Br	all but H	1.00	1.00	1.00	0.86
I	all but H	1.00	1.00	1.00	1.00

[a] SM5.4PD/A, P, U is used in this table to denote a parameter that is the same in SM5.4PD/U, SM5.4PD/A, and SM5.4PD/P.

TABLE 3: Bondi's van der Waals Radii and Intrinsic Coulomb Radii (Å) by Atomic Number

			SM5.2PD		SM5.4PD
		SM2.2PD/A	/A	/P	/A, P, U[a]
k	R_k	$\rho_k^{(0)}$	$\rho_k^{(0)}$	$\rho_k^{(0)}$	$\rho_k^{(0)}$
H	1.20	1.16	1.17	1.17	1.17
C	1.70	2.07	1.99	1.96	1.89
N	1.55	1.96	1.79	1.90	1.94
O	1.52	1.50	1.64	1.67	1.66
F	1.47	1.51	1.50	1.50	1.50
P	1.80	2.36	2.47		
S	1.80	2.23	2.34	2.30	2.11
Cl	1.75	2.14	2.14	2.15	2.14
Br	2.02	2.33	2.33	2.31	2.30
I	2.60	2.68	2.69	2.67	2.67

[a] SM5.4PD/A, P, U is used in this table to denote a parameter that is the same in SM5.4PD/U, SM5.4PD/A, and SM5.4PD/P.

After the genetic algorithm began to converge, which typically took less than 50 generations, the best guess of the genetic algorithm was relaxed into a self-consistent minimum. Each parameter was adjusted individually until a best value was found. This cycle was then repeated until a self-consistent minimum was found. Once again, at each step of the adjusting process, appropriate surface tension parameters were determined by a linear regression that minimized the sum of the squares of the errors in the free energy of solvation for the neutrals. The final parameters are given in Tables 2–5.

For SM2.2PD/A, we then fixed all the parameters that were thus far determined and found a value for the a_O surface tension parameter. In the HCO training subset, the function $\hat{g}(B_k^H)$ affects only water, $CH_3OH_2^+$, and H_3O^+. The parameter a_O in eq 12 was chosen to minimize the unsigned error of the neutral molecule plus 1/6 the error for the two ions. The best value of a_O was found to be −1.32.

At this point we examined the reasonableness of the values we were obtaining for the polarization free energy for molecules in the HCO portion of our training set. As has been observed in some earlier models, the polarization free energy, G_P, was positive (which is unphysical) for several longer unbranched alkanes and a few branched alkanes and cycloalkanes. We then adjusted the d_0 and d_{CH} parameters of eq 5 to reduce the magnitude of this problem. For the SM2.2PD/A and SM5.4PD/U models, we held the HCO electrostatic parameters constant and used the genetic algorithm to minimize the mean unsigned deviation between the electronic-nuclear-polarization free energy of solvation for the HCO training subset calculated with SM5.4 DD models (which do not show the nonphysical behavior) and the values obtained with the PD models. Once again, a linear fit of the surface tension coefficients was completed for each population element in the genetic algorithm prior to calculating the mean unsigned error. We found that a best fit was achieved for SM2.2PD/A when d_{CH} was 3.8 and d_0 was 4.0. For, SM5.4PD/U we found the best fit for d_{CH} equal to 4.5 and d_0 equal to 4.0. Although this did not entirely eliminate the prediction of positive polarization free energies, SM2.2PD/A predicts a positive value for the polarization free energy for only three molecules of the 219 molecules in the training set, and the largest magnitude of this erroneous prediction is 0.10 kcal. The SM5.4PD/U model (consequently the SM5.4PD/A and SM5.4PD/P models) does not predict any positive polarization free energies of solvation when used with either the AM1 or the PM3 Hamiltonian for any of the molecules in the training set.

The determination of d_0 and d_{CH} for the SM5.2 PD models was accomplished by an alternate method. In these two models, the values for d_{CH} and d_0 were optimized by the genetic algorithm in conjunction with the other HCO parameters. For SM5.2PD/A, d_{CH} is 4.0 and d_0 is 3.2. In the SM5.2PD/P model d_{CH} is 4.1 and d_0 is 3.6. Both of the SM5.2 PD models predict only one unphysical positive polarization free energy for any of the molecules in the HCO training set. In both cases the troublesome molecule is methane, which is an atypical solute, and the magnitude of the prediction is less than 0.01 kcal for both models.

After all the HCO parameters were determined and fixed to appropriate values, we considered the 43 neutrals and the 12 ions from our training set that contain N, and we found the intrinsic Coulomb radius for nitrogen and the three scaling factors that involve nitrogen in the same way that we had found the parameters for the HCO training subset.

As in the previous step, the SM2 a_N surface tension parameter was set to zero while the other parameters were found. Of the molecules containing N in the parametrization set, the function $\hat{g}(B_k^H)$ affects only ammonia and $CH_3NH_3^+$. Initially, we chose a_N, in eq 12, to minimize the unsigned error of the neutral molecule plus 1/6 of the unsigned error for the ion, but this resulted in a value that was larger in magnitude than the value found for a_N in SM2. Since we didn't wish to increase the effect of the $\hat{g}(B_k^H)$ function, we fixed the value of a_N to −9.12, which was the value obtained for SM2.

A similar method was used to find the parameters for sulfur, except there are no a_k parameters. Additionally, a surface tension identical in form to the sulfur-bonded-to-sulfur surface tension, σ_{SS}, of eq 20 was added to the SM2.2PD/A model. This addition allows the model to adequately handle disulfides, whereas SM2 and SM2.1 were both quite poor for disulfides. When determining the sulfur parameters, we discovered that both the S_{Sk} and S_{SH} scaling factors were consistently pinned to our upper scaling factor limit of 1.0 when optimized by the

Aqueous Free Energies of Solvation

J. Phys. Chem., Vol. 100, No. 51, 1996 **19831**

TABLE 4: Single-Subscript Surface Tension Coefficients (cal mol⁻¹ Å⁻²) by Atomic Number

| | SM2.2PD/A | | SM5.2PD | | | | SM5.4PD | | | |
| | | | /A | | /P | | /A | | /P | |
k	$\rho_k^{(0)}$	$\rho_k^{(1)}$	$\rho_k^{(0)}$	$\sigma_{Hk'}$	$\rho_k^{(0)}$	$\sigma_{Hk'}$	$\rho_k^{(0)}$	$\sigma_{Hk'}$	$\rho_k^{(0)}$	$\sigma_{Hk'}$
H			26.54		26.65		27.13		25.95	
C	5.08	3.90	40.15	−25.51	48.75	−28.30	72.71	−29.50	65.68	−27.71
N	−31.56	−5.39	−32.74	−82.74	−26.87	−76.00	−1.20	−61.13	18.08	−59.59
O	8.94	−59.00	−77.54	−92.13	−102.92	−59.94	−53.89	−35.12	−75.53	0.16
P	3.67		−37.96							
S	−62.82	42.35	−63.97	56.87	−60.26	52.44	−51.35	53.50	−40.92	25.69
F	18.11		16.17		17.36		18.11		16.82	
Cl	−2.85		−4.51		−5.00		−1.41		−1.08	
Br	−9.24		−10.78		−7.29		−5.52		−5.75	
I	−11.96		−12.80		−9.45		−7.66		−9.40	

TABLE 5: Double-Subscript Surface Tension Coefficients (cal mol⁻¹ Å⁻²) by Atomic Number

| | | SM2.2PD/A | SM5.2PD | | | | SM5.4PD | | | |
| | | | /A | | /P | | /A | | /P | |
k	k'	$\sigma_{kk'}^{(1)}$	$\sigma_{kk'}^{(1)}$	$\sigma_{kk'}^{(2)}$	$\sigma_{kk'}^{(1)}$	$\sigma_{kk'}^{(2)}$	$\sigma_{kk'}^{(1)}$	$\sigma_{kk'}^{(2)}$	$\sigma_{kk'}^{(1)}$	$\sigma_{kk'}^{(2)}$
O	O		44.50		49.45		32.46		15.18	
O	P		−48.06							
O	N		122.89		181.20		129.91		165.48	
O	C		84.90		128.43		86.41		127.64	
C	C		−21.39	−10.87	−33.64	−10.63	−46.71	−20.89	−52.33	−12.75
S	S	32.74	29.58		33.16		37.31		36.46	
N	C		−25.60		−41.91		−26.56		−40.36	

genetic algorithm with all the SM*x*.2 PD models. We individually tested these parameters and discovered that the mean unsigned error in the calculated free energy of solvation dropped by more than 0.1 kcal if these scaling factors were increased to 1.1. Increasing the scaling factors beyond this value had a lesser effect on the results, so we fixed both these scaling factors to 1.1 for SM2.2PD/A and the SM5.2 PD models. This same sensitivity to S_{Sk} and S_{SH} was not present in the SM5.4 PD models, which use the more accurate class IV charges; thus these parameters retained the usual upper bound of 1.0.

All other scale factors that became pinned at 1.0 were examined individually, but the effects of allowing these scaling factors to increase to 1.1 always made changes of less than 0.1 kcal in the mean unsigned error of the calculated solvation free energy for the affected molecules in the training set. Thus, the upper limit of 1.0 for all other scaling factors was retained.

A methodological comment might be useful here concerning the use of 1.1 for S_{Sk}. At first sight this might appear unphysical if one accepted the argument that $S_{kk'}$ only makes up for overlapping van der Waals spheres. But, an advantage of the PD formalism is that it can make up for other possible systematic errors in our model such as replacing the continuous charge density by point charges, and such errors could conceivably be more important for large polarizable atoms like S than for, say, C, N, or O. Anyway, even if one accepted the oversimplified view of the role of the $S_{kk'}$ parameters, there is no conflict with $S_{kk'} = 1.1$ since one can obtain the same results with $S_{kk'} = 1.0$ if one increases ρ_k by 10% and scales the radius down in one-center terms instead of up in two-center terms. In any event, using radii that differ by 10% in one-center and two-center terms is well within the range that we accept as "physical" for a semiempirical theory.

All the halogen parameters were determined simultaneously using the same method used for finding the H, C, and O parameters with the parameters for all non-halogen atom types fixed to the values found in previous steps.

Finding phosphorus parameters proved particularly challenging. There is very little experimental free energy of solvation data for neutral molecules containing phosphorus, and our data

for phosphorus-containing ions included heats of solvation for $H_2PO_4^-$, HPO_4^{2-}, and PO_4^{3-}. Although we chose parameters for phosphorus following procedures similar to those outlined for nitrogen and sulfur, we do not feel that we have enough data for these parameters to be as reliable as those for other atoms. Further discussion of the phosphorus parameters is therefore relegated to the Supporting Information section of this article.

At this point the SM2.2PD/A, SM5.2PD/A, SM5.2PD/P, and SM5.4PD/U models were completed. Using the electrostatic parameters from the SM5.4PD/U model, we then finished the SM5.4PD/A and SM5.4PD/P models by finding Hamiltonian-specific surface tension coefficients in a stepwise fashion. The surface tension coefficients involving only H, C, and O were found first, then frozen. From there, the coefficients involving N were found and so forth.

The final optimized parameters are given in Tables 2−5. Table 3 also gives the van der Waals radii of Bondi[33] that were used in calculating G_{CDS}.

4. Results

The mean unsigned errors in the aqueous free energies of solvation of neutral solutes for the five new specific models presented here and selected previous SM*x* models are collected in Table 6. Excluding phosphorus, the mean unsigned error in the free energies of solvation is 0.46 kcal for SM5.4PD/A and 0.44 for SM5.4PD/P. This compares to a mean unsigned error of 0.50 kcal for SM5.4/A and 0.44 for SM5.4/P. Thus, for the neutral molecules in the training set, adding the pairwise descreening approximation as presented produces models that have slightly smaller mean errors than the models without pairwise descreening.

When we use the pairwise descreening approximation, but substitute class II charges (Mulliken charges), the resulting models, SM5.2PD/A and SM5.2PD/P, have mean unsigned errors over the training set neutrals excluding phosphorus of 0.49 and 0.51 kcal, respectively. These errors are comparable to those in the previous paragraph.

TABLE 6: Mean Unsigned Errors (kcal/mol) in Aqueous Free Energies of Solvation by Functional Group for Neutral Molecules and Selected SM*x* Models

functional group	number of molecules	SM2	SM2.2	SM2.2PD/A	SM5.2PD /A	SM5.2PD /P	SM5.4 PD/A	SM5.4 PD/P	SM5.4 /A[a]	SM5.4 /P[a]
Compounds Containing at Most C, H, and/or O										
unbranched alkanes	8	0.48	1.34	0.16	0.16	0.35	0.48	0.43	0.60	0.56
branched alkanes	5	0.53	1.24	0.21	0.17	0.47	0.29	0.44	0.65	0.69
cycloalkanes	5	0.52	0.26	0.23	0.26	0.17	0.20	0.19	0.21	0.13
alkenes	9	0.19	0.77	0.11	0.36	0.40	0.55	0.42	0.45	0.33
alkynes	5	0.44	0.95	0.26	0.15	0.14	0.20	0.19	0.25	0.21
arenes	8	0.84	0.54	0.13	0.23	0.17	0.18	0.23	0.14	0.18
alcohols	16	0.59	0.42	0.44	0.37	0.44	0.44	0.36	0.51	0.40
ethers	9	1.57	1.64	0.61	0.72	0.85	0.64	0.75	0.93	1.03
aldehydes	6	0.59	0.54	0.36	0.28	0.25	0.28	0.28	0.29	0.36
ketones	12	0.73	0.57	0.44	0.35	0.28	0.26	0.25	0.37	0.37
carboxylic acids	5	0.39	0.34	0.79	0.71	0.74	1.03	0.87	0.77	0.75
esters	12	0.53	0.60	0.65	0.59	0.39	0.68	0.55	0.44	0.41
bifunctional	5	0.53	0.66	0.98	0.72	0.52	0.58	0.54	0.32	0.34
water, H_2	2	1.23	1.48	1.35	2.33	2.07	2.34	1.82	1.59	1.17
subtotal	107	0.65	0.76	0.43	0.43	0.43	0.49	0.44	0.49	0.46
Compounds Containing N										
aliphatic amines	15	1.19	1.33	0.86	0.50	0.67	0.59	0.72	0.73	0.74
aromatic amines	10	1.09	0.98	0.72	0.55	0.83	0.36	0.67	0.39	0.56
nitriles	4	0.24	0.31	0.44	0.41	0.36	0.36	0.49	0.49	0.48
nitrohydrocarbons	6	0.76	0.71	1.44	0.41	0.26	0.37	0.23	0.45	0.14
amides	3	0.59	0.60	1.65	1.17	1.66	1.54	0.60	1.84	0.50
bifunctional	3	1.46	1.77	0.35	0.51	0.75	0.31	0.71	0.50	0.68
ammonia	1	0.01	0.52	0.32	2.21	1.87	0.48	1.29	2.03	0.75
subtotal	42	0.96	1.02	0.88	0.58	0.72	0.53	0.62	0.68	0.57
hydrazine	1	6.05	6.51	3.81	1.22	*b*	0.68	*b*	2.36	*b*
subtotal	43	1.08	1.15	0.95	0.59	*b*	0.53	*b*	0.72	*b*
Compounds Containing P										
phosphate esters, PH_3	4	5.99		0.64	1.20					
Compounds Containing S, but No Halogen										
thiols	4	0.65		0.59	0.59	0.49	0.38	0.31	0.30	0.24
organic sulfides, H_2S	5	0.82		0.77	0.65	0.54	0.42	0.36	0.40	0.35
organic disulfides	2	2.25		0.30	0.37	0.25	0.35	0.02	0.20	0.05
subtotal	11	1.02		0.62	0.58	0.47	0.39	0.28	0.33	0.26
Compounds Containing Halogens										
fluorinated hydrocarbons	3	0.6		0.29	0.35	0.59	0.36	0.42	0.54	0.47
chloroalkanes	8	0.25		0.39	0.39	0.61	0.26	0.20	0.27	0.23
chloroalkenes	5	0.27		0.35	0.63	0.94	0.73	0.60	0.65	0.55
chloroarenes	3	0.22		0.11	0.17	0.77	0.14	0.25	0.13	0.20
brominated hydrocarbons	10	0.43		0.55	0.53	0.17	0.19	0.18	0.25	0.24
Iodinated hydrocarbons	8	0.35		0.46	0.39	0.15	0.12	0.09	0.23	0.15
other halo molecules	17	0.56		0.63	0.66	0.69	0.53	0.43	0.48	0.41
subtotal	54	0.42		0.48	0.51	0.52	0.35	0.31	0.37	0.32
excluding hydrazine, P	214	0.66		0.54	0.49	0.51	0.46	0.44	0.49	0.44
excluding P	215	0.69		0.56	0.49		0.46		0.50	
entire set	219	0.79		0.56	0.49					

a Reference 10. *b* Hydrazine was not calculated using any PM3 method (see text, section 3.1).

The SM2.2PD/A model, which incorporates a pairwise descreening model with the SM2 surface tension functionals and the electrostatic cutoff-Gaussians inherent in the SM2 formalism, yields a mean unsigned error over the neutral training set of 0.56 kcal for SM2.2PD/A, which is an improvement of more than 0.1 kcal over the result obtained with the original SM2 model.

The mean signed errors broken down by functional group for the specific PD models are compiled in Table 7. The results for the ionic training set are given in Table 8. The mean signed and unsigned energy differences between calculated energy components for several SM*x* models are listed in Tables 9 and 10.

5. Discussion

5.1. Timings. As discussed in section 4, there is a slight improvement in accuracy for our models when incorporating the pairwise descreening approximation. However, we first

emphasize that the main impetus for pursuing this project was to reduce the complexity of the solvation calculation, ultimately reducing the computational cost and allowing the examination of much larger solutes than have previously been feasible. If, in the process of simplifying the calculation, we had increased the mean errors by, for example, "only" a factor of 1.5, we would have been satisfied. The fact that we accomplished the speedup while hardly affecting the mean errors is a bonus.

To quantify our success in speeding up the algorithm, we carried out a series of timings. These timings were carried out on an SGI Power Challenge L R8000 workstation, although we would expect similar results on other modern processors. First, we selected a representative set of 10 of the largest molecules in our neutral training set. (Each molecule contained 16 or more atoms.) Using the AMSOL code, we completed a single geometry cycle for each molecule. This type of run converges a self-consistent-field calculation for a given geometry and calculates the gradient by finite differences at that geometry; such a "cycle"

Aqueous Free Energies of Solvation

J. Phys. Chem., Vol. 100, No. 51, 1996 **19833**

TABLE 7: Mean Signed Errors (kcal/mol) in Aqueous Free Energies of Solvation by Functional Group for Neutral Molecules and Selected SM*x* Models

functional group	number of molecules	SM2	SM2.2	SM2.2PD/A	SM5.2PD		SM5.4			
					/A	/P	PD/A	PD/P	/A[a]	/P[a]
Compounds Containing at Most C, H, and/or O										
unbranched alkanes	8	−0.48	−1.34	−0.15	−0.15	−0.35	−0.39	−0.34	−0.60	−0.53
branched alkanes	5	−0.53	−1.24	−0.21	−0.17	−0.47	−0.27	−0.44	−0.65	−0.69
cycloalkanes	5	0.52	−0.11	0.20	0.26	0.17	0.09	0.11	−0.15	−0.06
alkenes	9	−0.05	−0.77	0.06	0.36	0.40	0.49	0.40	0.33	0.25
alkynes	5	−0.44	−0.95	−0.25	0.08	0.05	0.02	0.05	−0.04	−0.01
arenes	8	0.84	0.52	−0.03	−0.22	−0.10	−0.13	−0.03	0.10	0.12
alcohols	16	0.46	0.12	−0.17	0.10	0.23	0.22	0.15	0.29	0.16
ethers	9	1.29	1.29	−0.07	−0.02	0.07	−0.04	0.09	0.10	0.30
aldehydes	6	−0.59	−0.54	−0.07	−0.24	−0.23	−0.24	−0.28	−0.29	−0.36
ketones	12	0.68	0.55	0.21	−0.09	0.05	0.08	0.11	0.29	0.30
carboxylic acids	5	−0.10	−0.27	0.60	0.71	0.74	1.03	0.87	0.77	0.75
esters	12	−0.32	−0.54	−0.57	−0.59	−0.39	−0.68	−0.55	−0.44	−0.41
bifunctional	5	0.04	−0.27	0.13	0.41	0.18	0.24	0.17	0.01	0.01
water, H_2	2	−1.10	−1.48	−1.35	−1.51	−1.25	−1.61	−1.08	−1.56	−1.13
subtotal	107	0.16	−0.19	−0.09	−0.04	0.00	−0.03	−0.01	−0.02	−0.01
Compounds Containing N										
aliphatic amines	15	0.65	0.44	0.67	0.20	0.30	0.09	0.25	0.18	0.24
aromatic amines	10	0.69	0.26	0.24	−0.22	0.05	0.09	0.12	0.05	0.03
nitriles	4	−0.09	−0.27	−0.08	0.32	0.34	0.09	0.10	0.29	0.13
nitrohydrocarbons	6	−0.76	−0.71	1.44	−0.01	−0.01	0.01	−0.01	0.21	0.10
amides	3	0.32	0.60	−0.15	1.17	1.66	1.54	0.42	1.44	0.42
bifunctional	3	1.13	1.52	0.01	−0.08	−0.75	0.26	−0.70	−0.10	−0.68
ammonia	1	−0.01	−0.52	−0.32	−2.21	−1.87	−0.48	1.29	−2.03	0.75
subtotal	42	0.38	0.24	0.48	0.07	0.17	0.18	0.14	0.18	0.12
hydrazine	1	−6.05	−6.51	−3.81	1.22	*b*	0.68	*b*	2.36	*b*
subtotal	43	0.23	0.08	0.38	0.10	*b*	0.19	*b*	0.23	*b*
Compounds Containing P										
phosphate esters, PH_3	4	−5.99		0.62	1.18					
Compounds Containing S, but No Halogen										
thiols	4	0.33		0.21	−0.01	0.02	0.03	0.08	0.03	0.05
organic sulfides, H_2S	5	0.12		−0.10	0.14	0.09	0.09	0.02	0.09	0.04
organic disulfides	2	−2.25		−0.01	0.00	0.01	0.05	0.00	0.03	−0.01
subtotal	11	−0.24		0.03	0.06	0.05	0.06	0.04	0.06	0.03
Compounds Containing Halogens										
fluorinated hydrocarbons	3	0.60		−0.04	−0.18	0.59	0.25	0.15	0.54	0.47
chloroalkanes	8	0.00		0.24	0.20	0.44	−0.15	−0.08	−0.24	−0.20
chloroalkenes	5	0.04		0.19	0.50	0.71	0.73	0.60	0.63	0.52
chloroarenes	3	−0.22		−0.11	−0.17	−0.77	−0.10	−0.25	−0.10	−0.20
brominated hydrocarbons	10	0.12		0.21	0.14	−0.07	−0.15	−0.10	−0.14	−0.13
iodinated hydrocarbons	8	−0.25		0.11	−0.03	−0.06	−0.02	−0.01	−0.07	−0.04
other halo molecules	17	0.14		−0.05	−0.04	−0.02	0.01	0.06	−0.04	0.00
subtotal	54	0.05		0.08	0.07	0.09	0.03	0.03	0.00	0.00
excluding hydrazine, P	214	0.16		0.08	0.01	0.06	0.04	0.03	0.03	0.02
excluding P	215	0.13		0.06	0.02		0.04		0.04	
entire set	219	0.02		0.06	0.02					

a Reference 10. *b* Hydrazine was not calculated using any PM3 method (see text).

constitutes a single step in a geometry optimization process. We then calculated geometric mean speedup factors referenced to the SM5.4/A model, which was the slowest method tested and hence is assigned a baseline speedup factor of 1.0. The SM5.4PD/A model has a speedup factor of 1.2. The SM2 model, which has less computationally expensive class II charges (Mulliken charges) and less expensive surface tension functionals, has a speedup factor of 1.3. The SM5.2PD/A model has a geometric mean speedup factor of 3.4, while the SM2.2PD/A model has a speed up factor of 3.4. Together, these results show that there is a substantial reduction in the required computer resources for a given molecule when the pairwise descreening approximation is incorporated into the SM*x* models, particularly when the lower level class II charges are used. Proceeding to bigger molecules or minimizing overhead will increase all these speedups. Further, the simplicity that is inherent in the pairwise descreening model will allow the incorporation of further enhancements such as analytic gradients

into the solvation calculations, and the time savings when that code is available should be even more dramatic.

To avoid a possible misimpression, we emphasize that geometries *are* optimized in solution for all results in this paper, as well as in all our previous SM*x* models dating back to SM1.[1] This is accomplished by evaluating the required gradients (and—if and when required by the optimization algorithm—Hessians) by numerical differentiation. However, the PD models allow for straightforward analytic differentiation in *future* work, if desired. Analytic differentiation has two main advantages: it is faster, and it is more stable.

5.2. Trends in Solvation Energies and Coulomb Radii. Tables 6 and 7 show that not only are the solvation free energies remarkably accurate over the whole neutral training sets but in addition we can obtain rather uniform accuracy over the individual solute functional groups. The question always arises in semiempirical models whether the flexibility of the parametrization is so great that one is simply fitting the training set.

TABLE 8: Calculated and Experimental[a] Free Energies of Solvation (kcal/mol) for Ionic Solutes in the Training Set

solute	SM2	SM2.2PD/A	SM5.2PD		SM5.4				exp[b]
			/A	/P	PD/A	PD/P	/A[a]	/P[a]	
HC_2^-	−77.7	−74.3	−75.0	−77.0	−76.7	−78.6	−77.2	−79.2	−73
$CH_3OH_2^+$	−82.1	−78.3	−76.2	−84.7	−79.7	−79.8	−91.0	−89.7	−85
$(CH_3)_2OH^+$	−61.7	−64.5	−63.1	−69.9	−64.6	−65.1	−70.3	−71.1	−70
$CH_3C(OH)CH_3^+$	−58.5	−61.0	−60.1	−63.9	−61.6	−61.2	−65.7	−65.7	−64
H_3O^+	−98.7	−98.4	−98.3	−105.5	−102.8	−99.7	−114.9	−109.4	−104
CH_3O^-	−83.1	−93.8	−88.1	−92.1	−86.5	−90.6	−84.5	−88.2	−95
$CH_3CO_2^-$	−76.3	−79.2	−77.7	−78.7	−74.3	−76.9	−76.6	−78.6	−77
$CH_3COCH_2^-$	−71.2	−78.3	−73.3	−75.4	−71.1	−72.9	−71.7	−73.0	−81
$C_6H_5O^-$	−67.7	−74.2	−69.8	−72.0	−66.2	−68.1	−66.7	−68.3	−72
$C_6H_5CH_2^-$	−52.8	−59.3	−57.5	−59.6	−55.8	−56.5	−56.8	−57.6	−59
OH^-	−106.7	−117.0	−109.6	−108.6	−104.8	−106.5	−109.9	−111.8	−106
HO_2^-	−95.5	−95.9	−97.7	−95.6	−94.5	−96.0	−99.5	−101.0	−101
O_2^-	−88.1	−85.4	−94.2	−95.4	−90.4	−94.7	−96.7	−99.9	−87
$CH_3NH_3^+$	−70.5	−72.3	−72.6	−77.3	−70.6	−74.4	−76.0	−78.4	−70
$CH_3C(OH)NH_2^+$	−66.8	−63.9	−60.1	−67.2	−64.9	−64.9	−68.9	−67.8	−66
$(CH_3)_2NH_2^+$	−61.5	−60.7	−61.9	−66.8	−60.6	−64.2	−65.6	−68.7	−63
$(CH_3)_3NH^+$	−52.6	−53.2	−55.6	−59.0	−54.5	−56.8	−57.2	−59.5	−59
$C_5H_5NH^+$	−56.7	−50.2	−50.3	−57.5	−53.2	−55.5	−55.1	−57.0	−59
NH_4^+	−76.4	−88.1	−88.7	−89.7	−84.8	−85.5	−88.0	−87.1	−79
CN^-	−83.5	−79.6	−80.3	−77.4	−74.1	−73.7	−71.9	−71.6	−77
CH_2CN^-	−69.6	−69.9	−71.7	−70.8	−66.2	−67.3	−64.9	−66.3	−75
NH_2^-	−80.9	−89.1	−96.0	−89.7	−88.2	−84.4	−94.7	−88.1	−95
NO_2^-	−77.7	−77.9	−78.2	−76.8	−71.3	−73.8	−69.8	−72.0	−72
NO_3^-	−59.0	−54.2	−62.0	−58.9	−57.4	−60.4	−60.3	−63.2	−65
N_3^-	−75.4	−73.9	−79.3	−73.7	−76.7	−64.3	−69.4	−62.7	−74
$CH_3SH_2^+$	−73.7	−71.4	−68.4	−67.4	−66.9	−67.1	−65.2	−66.2	−74
$(CH_3)_2SH^+$	−70.5	−69.2	−73.2	−64.5	−61.0	−60.5	−58.5	−58.8	−61
HS^-	−76.4	−76.8	−76.0	−76.0	−81.4	−81.2	−87.5	−86.8	−76
$n\text{-}C_3H_7S^-$	−69.1	−73.1	−72.2	−73.4	−75.3	−75.5	−78.1	−78.8	−76
$C_6H_5S^-$	−66.0	−67.4	−65.0	−68.1	−66.2	−68.4	−69.3	−72.1	−67
F^-	−107.0	−106.7	−107.2	−107.1	−107.0	−107.1	−106.9	−106.9	−107
Cl^-	−77.0	−77.0	−77.3	−77.0	−76.8	−76.7	−77.3	−77.2	−77
Br^-	−72.0	−71.7	−72.0	−72.1	−72.1	−72.2	−72.0	−72.0	−72
I^-	−63.0	−62.9	−63.1	−63.0	−62.7	−63.0	−63.0	−63.2	−63
mean unsigned error	4.1	3.8	4.2	3.6	4.4	4.1	4.9	5.1	
mean signed error	2.2	1.0	1.9	0.5	3.5	2.7	0.0	−0.6	
rms error	5.5	5.2	5.4	5.1	5.5	5.5	6.1	6.4	

[a] Reference 10. [b] Reference 37.

We believe that this is not the case here; rather we are representing the true dependencies of physical solvation effects on the functional groups. There is no universally accepted way to demonstrate the physicality of a semiempirical model, but we simply remark that avoidance of overparametrization was a steady source of concern in the design process of the SM5 functional forms and in the present effort as well. Our experiences in varying the functional forms, freezing, unfreezing, or manually varying nonlinear parameters, and deleting selected molecules from the training set convince us that the final models are reasonably robust to such changes, as they should be. Although we could easily add to this paper a variety of quantitative tests (such as fitting a subset of the training set and reporting the error over the remaining molecules), we do not believe that quantitative measures are adequate to convey the type of molecule-by-molecule analysis that goes into our parametrizations. Clearly, there will be deficiencies of the model, but we believe they will result less from overparametrization than from underparametrization, i.e., from restricting ourselves to a simple model with two clearly defined aspects of the physics, namely, (1) electrostatics with approximate atomic charges and pairwise descreening and (2) first solvation effects with surface tensions calculated from predetermined radii and depending only on atomic number and on a few geometrical variables (interatomic distances only) where the available data are adequate to identify a systematic dependence.

When examining results for the training set molecules, in general, the pairwise descreening models have a slightly better mean unsigned error than the SM5.4 DD models for unbranched alkanes, branched alkanes, ethers, and aliphatic amines. The mean unsigned error for the PD models is slightly worse than the SM5.4 DD models for carboxylic acids, esters, and sulfur-containing molecules. The SM5.4 DD models and the PD models perform similarly for the cycloalkanes, alkenes, arenes, alcohols, aldehydes, ketones, aromatic amines, nitriles, nitrohydrocarbons, and singly halogenated compounds. For the classes of compounds where the SM5.4 DD models do differ from the PD models, the differences in the mean unsigned errors are of a relatively small magnitude and typically are about 0.2 kcal/mol for the class. The SM2.2PD/A model (which uses only class II charges and SM2-type surface tensions) makes some predictions for nitrogen-containing compounds that are qualitatively unlike those for the other PD models. For example, for the nitrohydrocarbons, SM2.2PD/A achieves a mean unsigned error of 1.44 kcal/mol, which is twice as large as any other model listed, including SM2.1.

In the HCO portion of the training set, the PD models have similar or improved mean signed errors for most of the molecule types when compared to the SM5.4 DD models. The alkenes and esters are the only exceptions to this trend, and in these cases the PD models have mean signed errors that are about 0.1-0.2 kcal/mol higher than the SM5.4 DD models. In the nitrogen-containing molecules, the SM5.4 DD and PD models perform similarly. An exception occurs with ammonia and hydrazine, where the SM5.4PD/A model performs significantly better than the SM5.4/A model.

Aqueous Free Energies of Solvation

J. Phys. Chem., Vol. 100, No. 51, 1996 **19835**

TABLE 9: Mean Signed Energy Differences[a] between Calculated Energy Components for the Neutrals in the Training Set[b] and Several SMx Models

solute	SM2	SM2.2-PD/A	SM5.2PD		SM5.4			
			/A	/P	PD/A	PD/P	/A	/P
Electronic-Nuclear-Polarization Free Energy								
SM2		1.3	1.2	0.5	1.9	1.4	1.6	1.2
SM2.2PD/A			−0.1	−0.8	0.6	0.1	0.3	−0.1
SM5.2PD/A				−0.7	0.7	0.1	0.4	0.0
SM5.2PD/P					1.4	0.9	1.1	0.7
SM5.4PD/A						−0.6	−0.3	−0.8
SM5.4PD/P							0.3	−0.2
SM5.4/A								−0.4
Cavitation-Dispersion-Structural Free Energy								
SM2		−1.2	−1.1	−0.4	−1.8	−1.2	−1.5	−1.0
SM2.2PD/A			0.1	0.8	−0.6	0.0	−0.3	0.2
SM5.2PD/A				0.7	−0.7	−0.2	−0.4	0.0
SM5.2PD/P					−1.4	−0.8	−1.1	−0.6
SM5.4PD/A						0.6	0.3	0.8
SM5.4PD/P							−0.3	0.2
SM5.4/A								0.5
Free Energy of Solvation								
experiment	−0.2	−0.1	0.0	−0.1	0.0	0.0	0.0	0.0
SM2		0.1	0.2	0.1	0.1	0.2	0.2	0.2
SM2.2PD/A			0.1	0.0	0.0	0.0	0.0	0.1
SM5.2PD/A				0.0	0.0	0.0	0.0	0.0
SM5.2PD/P					0.0	0.0	0.0	0.0
SM5.4PD/A						0.0	0.0	0.0
SM5.4PD/P							0.0	0.0
SM5.4/A								0.0

TABLE 10: Mean Unsigned Energy Differences between Calculated Energy Components for the Neutrals in the Training Set[a] and Several SMx Models

solute	SM2	SM2.2-PD/A	SM5.2PD		SM5.4			
			/A	/P	PD/A	PD/P	/A	/P
Electronic-Nuclear-Polarization Free Energy								
SM2		1.4	1.3	1.1	2.0	1.6	1.7	1.5
SM2.2PD/A			0.6	1.2	1.0	1.0	1.2	1.2
SM5.2PD/A				0.9	0.9	0.9	0.9	1.0
SM5.2PD/P					1.5	1.0	1.2	0.9
SM5.4PD/A						0.6	0.5	0.9
SM5.4PD/P							0.6	0.4
SM5.4/A								0.5
Cavitation-Dispersion-Structural Free Energy								
SM2		1.3	1.3	1.1	1.8	1.6	1.7	1.5
SM2.2PD/A			0.7	1.2	1.0	1.0	1.1	1.2
SM5.2PD/A				0.8	0.9	0.9	0.8	1.0
SM5.2PD/P					1.4	0.9	1.1	0.8
SM5.4PD/A						0.6	0.5	0.9
SM5.4PD/P							0.6	0.3
SM5.4/A								0.5
Free Energy of Solvation								
experiment	0.6	0.5	0.5	0.5	0.5	0.4	0.5	0.4
SM2		0.6	0.6	0.6	0.6	0.6	0.6	0.6
SM2.2PD/A			0.4	0.5	0.5	0.5	0.6	0.6
SM5.2PD/A				0.3	0.3	0.3	0.4	0.5
SM5.2PD/P					0.3	0.2	0.3	0.3
SM5.4PD/A						0.2	0.2	0.3
SM5.4PD/P							0.3	0.2
SM5.4/A								0.2

[a] The energy component differences are obtained by subtracting the energy component calculated by the model in the column from the energy component calculated by the model in the row. [b] The four phosphorus compounds were excluded because the SM5.4 models are unparametrized for phosphorus, the two disulfide molecules were also excluded because SM2 yields rather large errors for disulfides, and hydrazine was excluded because PM3 does not yield an accurate geometry. Thus, the results from 212 neutral molecules were used to construct this table.

[a] The four phosphorus compounds were excluded because the SM5.4 models are unparametrized for phosphorus, the two disulfide molecules were also excluded because SM2 yields rather large errors for disulfides, and hydrazine was excluded because PM3 does not yield an accurate geometry. Thus, the results from 212 neutral molecules were used to construct this table.

In developing the PD models, the results for the entire ionic training set were employed directly in determining the electrostatic parameters. This is the first time this type of approach was used in conjunction with an SMx model. Ultimately, this approach afforded a very modest improvement in the mean unsigned error for the solvation free energy for the ions. See Table 8. However, it is interesting to note, as can be seen in Table 3, that even though we did not constrain our intrinsic Coulomb radii to have certain size relationships with respect to each other, they roughly exhibit the appropriate size trends. (The two exceptions to this are in the SM5.4PD/U model, where the optimized radius for nitrogen is 0.05 Å larger than carbon's radius, and sulfur is 0.03 Å smaller than the optimized chlorine radius.) Further, the optimized radii are similar for each PD model. When comparing to Bondi's van der Waals radii, our optimized radii for nitrogen and sulfur show the largest deviations (excluding phosphorus) and are about 30% larger than the Bondi radii. Our radii for carbon and chlorine are about 23% larger, while bromine and oxygen are 15% and 10% larger than Bondi radii, respectively. Hydrogen, fluorine, and iodine deviate from the Bondi radii by at most 3%.

5.3. Magnitudes of Surface Tensions. One cannot directly interpret the surface tension coefficients because they must be combined with various functional forms[10] to obtain the actual surface tensions in any given solute molecule. Table 11 gives the actual surface tensions for three representative molecules. As we have stressed before, the critical requirement for surface tensions is that they be *consistent* with the electrostatic model with which they are used. Thus the surface tensions differ greatly among the models, and they have no meaning without

reference to the electrostatics and the whole functional form set (e.g., SM2 functional forms or SM5 functional forms). Furthermore the parameters differ depending on the molecule, and we believe that it would not be appropriate to average over various molecules to find, for example, an "average" oxygen surface tension.

Since this point may still be confusing to some readers, we amplify on it here. Suppose we increase the oxygen radius by 0.1 Å (which can easily be done without making the model "unphysical"). This moves electrostatic effects in a shell of width 0.1 Å out of the electrostatics calculation; thus these effects must be included in the first solvation shell calculation. Since the boundary between the solute (dielectric constant 1.0) and the solvent (bulk dielectric constant) is not really infinitely sharp, since it fluctuates, and since the electrostatics in the first solvation shell (wherever it may "begin") are not the same as in the bulk (because the water molecules have nonbulk electronic polarizabilities, nonbulk dipole moments and reorientational polarizabilities, nonbulk average number of hydrogen bonds, etc.), the Coulomb radii are intrinsically undefinable, and there is no unique partition of solvation free energies into ENP and CDS parts. (Another reason for this is that only their sum is a thermodynamic state function.) This nonunique partition is discussed further in the next subsection.

5.4. Partition of ΔG_S° into ΔG_{ENP} and G_{CDS}. The components ΔG_{ENP} and G_{CDS} of the solvation free energy ΔG_S are not state functions. Hence there is no unique way to partition the free energy into well-defined contributions of these two types. Anybody trying to design or use solvation models should be aware of this. Nevertheless from a methodological perspective, the partition is interesting and fruitful. Also, one can gain considerable physical insight from examining the origin of

TABLE 11: Atomic Surface Tensions (cal mol^{-1} Å^{-2}) for Selected Molecules

			SM5.2PD		SM5.4			
	SM2	SM2.2PD/A	/A	/P	PD/A	PD/P	/A	/P
methylamine								
N	−55.33	−38.47	−118.69	−168.96	−90.57	−120.27	−109.51	−126.98
H(4), H(5) on N	0.00	0.00	−37.71	−32.93	−21.20	−21.23	−25.96	−19.73
C	7.77	10.37	40.00	48.27	73.03	65.97	64.06	58.96
H(1) on C	0.00	0.00	7.51	5.31	5.22	5.02	4.06	4.43
H(2), H(3) on C	0.00	0.00	7.57	5.33	5.24	5.03	4.07	4.44
methanol								
O	−63.71	−52.01	−77.40	−102.03	−53.70	−74.86	−80.22	−96.67
H on O	0.00	0.00	−48.86	−26.68	−6.80	17.70	43.18	59.84
C	7.88	10.45	40.34	49.07	72.87	65.81	63.49	59.28
H(1) on C	0.00	0.00	7.44	5.28	5.14	4.98	4.28	4.39
H(2), H(3) on C	0.00	0.00	7.49	5.24	5.18	5.24	4.04	4.37
ethanal								
O	−29.71	1.71	−27.93	−23.98	−3.60	2.79	14.70	−7.71
C in carbonyl	6.81	9.04	26.54	27.53	43.31	32.63	40.78	31.35
H in carbonyl	0.00	0.00	7.35	5.28	4.98	5.27	4.11	4.41
C in methyl	7.75	10.49	26.71	27.88	43.15	33.08	40.90	31.15
H(1) in methyl	0.00	0.00	7.43	5.49	5.17	5.18	4.20	4.52
H(2), H(3) in methyl	0.00	0.00	7.41	5.40	5.03	5.10	4.15	4.50

TABLE 12: Mean Signed Differences (kcal/mol) in Electronic-Nuclear-Polarization Free Energy of Solvation between Density Descreening and Pairwise Descreening Methods for the Neutral Molecules of the Training Set

	number of	SM5.4/A minus PD/A value[a]			SM5.4/P minus PD/P value[a]	
functional group	molecules	SM2.2	SM5.2	SM5.4	SM5.2	SM5.4
Compounds Containing at Most C, H, and/or O						
unbranched alkanes	8	−0.07	0.24	0.10	0.01	0.04
branched alkanes	5	−0.04	0.23	−0.05	0.00	−0.01
cycloalkanes	5	0.30	0.42	−0.02	0.06	−0.03
alkenes	9	−0.18	0.20	0.03	0.01	−0.01
alkynes	5	−0.42	0.09	0.24	0.12	0.10
arenes	8	0.16	0.80	−0.03	0.84	−0.06
alcohols	16	−1.43	−1.81	−0.20	−2.36	−0.15
ethers	9	1.84	0.65	1.06	0.27	0.95
aldehydes	6	0.49	−0.41	0.72	−0.83	0.61
ketones	12	1.18	0.48	0.91	−0.20	0.81
carboxylic acids	5	−3.31	−2.94	−0.60	−2.42	−0.49
esters	12	0.40	0.21	1.20	0.03	0.78
bifunctional	5	−1.73	−2.58	−0.15	−2.29	−0.04
water, H$_2$	2	−2.85	−3.29	−0.70	−3.10	−0.63
subtotal	107	−0.16	−0.35	0.30	−0.60	0.23
Compounds Containing N						
aliphatic amines	15	1.59	0.43	1.26	−1.07	0.38
aromatic amines	10	−0.32	−0.04	0.49	−1.01	0.09
nitriles	4	−3.04	−2.40	0.05	−3.20	0.09
nitrohydrocarbons	6	−4.41	−1.18	0.67	0.76	0.50
amides	3	1.97	−0.81	1.03	−2.99	1.01
bifunctional	3	2.99	0.47	1.16	−0.81	0.87
ammonia, hydrazine	2[b]	−0.09	−2.66	0.53	−2.11	−0.66
subtotal	43[b]	−0.08	−0.39	0.83	−1.14	0.36
Compounds Containing S, but No Halogen						
thiols	4	−0.68	−0.65	−0.18	−0.56	−0.12
organic sulfides, H$_2$S	5	−0.76	−0.51	−0.10	−0.74	−0.01
organic disulfides	2	−1.77	−1.43	−0.25	−1.97	−0.18
subtotal	11	−0.92	−0.73	−0.16	−0.90	−0.08
Compounds Containing Halogens						
fluorinated hydrocarbons	3	0.10	0.70	0.27	−0.03	0.23
chloroalkanes	8	−0.67	−0.51	0.16	−1.01	0.10
chloroalkenes	5	−0.50	−0.34	0.14	−0.65	0.12
chloroarenes	3	−0.45	−0.05	−0.13	0.59	−0.04
brominated hydrocarbons	10	−0.68	−0.46	0.25	0.04	0.17
iodinated hydrocarbons	8	−0.46	−0.23	0.11	0.16	0.06
other halo molecules	17	−1.04	−1.02	−0.07	−0.85	−0.18
subtotal	54	−0.69	−0.51	0.08	−0.41	0.02
entire set excluding P	215[b]	−0.31	−0.42	0.33	−0.68	0.19

[a] In each case we subtract the PD result from the corresponding DD result. For example, the last column is the mean signed difference of SM5.4/P from SM5.4PD/P. [b] Hydrazine was not calculated using any PM3 method (see text). Therefore, the number of elements for each SM5-P method is one less than the value shown in the Number of Molecules column.

Aqueous Free Energies of Solvation

J. Phys. Chem., Vol. 100, No. 51, 1996 **19837**

TABLE 13: Components of the Calculated Free Energy of Solvation (kcal/mol) for Selected Neutral Solutes from the Training Suite and Selected SM*x* Models

| | | | SM5.2PD | | SM5.4 | | | | |
solute	SM2	SM2.2PD/A	/A	/P	PD/A	PD/P	/A	/P	exp
				Electronic-Nuclear-Polarization Free Energy					
1-butanol	−0.82	−2.55	−2.19	−1.28	−3.75	−3.60	−3.68	−3.56	
methyl propyl ether	−0.91	−3.89	−2.92	−1.81	−3.31	−2.57	−2.12	−1.56	
butanal	−2.50	−4.85	−3.87	−3.67	−5.05	−5.15	−4.30	−4.51	
butanone	−2.75	−5.73	−4.73	−4.28	−5.47	−5.41	−4.61	−4.63	
methyl propanoate	−1.75	−5.73	−5.45	−4.90	−6.35	−5.49	−5.15	−4.64	
1-iodobutane	0.09	−0.38	−0.71	−0.76	−0.91	−0.70	−0.88	−0.67	
				Cavitation-Dispersion-Structural Free Energy					
1-butanol	−3.57	−2.27	−2.46	−3.33	−0.78	−1.10	−0.72	−1.05	
methyl propyl ether	0.38	2.04	1.30	0.40	1.54	1.02	0.69	0.36	
butanal	−0.82	1.71	0.77	0.54	1.84	1.95	1.04	1.24	
butanone	−0.49	1.87	0.90	0.53	1.69	1.70	0.99	1.07	
methyl propanoate	−1.19	2.11	1.84	1.41	2.72	1.97	1.78	1.31	
1-iodobutane	−0.30	0.42	0.49	0.43	0.63	0.41	0.39	0.25	
				Free Energy of Solvation					
1-butanol	−4.39	−4.82	−4.65	−4.61	−4.53	−4.69	−4.40	−4.61	−4.72
methyl propyl ether	−0.54	−1.85	−1.62	−1.41	−1.77	−1.55	−1.43	−1.19	−1.66
butanal	−3.32	−3.14	−3.10	−3.13	−3.21	−3.21	−3.26	−3.27	−3.18
butanone	−3.24	−3.86	−3.83	−3.75	−3.78	−3.71	−3.62	−3.56	−3.64
methyl propanoate	−2.94	−3.62	−3.61	−3.49	−3.63	−3.51	−3.37	−3.33	−2.93
1-iodobutane	−0.21	0.04	−0.22	−0.34	−0.28	−0.29	−0.49	−0.42	−0.25

solvation free energy contributions in the model. Although the general trends in the electrostatics and overall solvation free energy of the new PD models seem to be relatively similar to those in the SM5.4 DD models, which are our least simplified models for electrostatics,[10] some systematic differences can be recognized. Table 12 presents the mean signed differences in the electrostatic contributions (ΔG_{ENP}) between the SM5.4 DD models and the new PD models.

The new models yield ΔG_{ENP} values similar to the DD ones for branched alkanes, unbranched alkanes, and alkenes. The cycloalkanes have larger electrostatic terms for the PD/A models based on class II charges, while the other new models' electrostatics match the SM5.4 DD models. SM2.2PD/A tends to have smaller ΔG_{ENP} values for alkynes. The SM5.2 PD models obtain systematically less negative free energy of solvation values for hydrocarbons than the SM5.4 DD models.

All the PD models based on class II charges significantly underestimate ΔG_{ENP} when compared to SM5.4 DD for the alcohols, carboxylic acids, bifunctional CHO compounds, and nitriles. These underestimates range from just less than 1.5 kcal to over 4 kcal. In each of these classes of compounds, the SM5.4 PD models perform similarly to the SM5.4 DD model.

The ΔG_{ENP} values calculated by the PD models for the ethers are less negative than the SM5.4 DD values, with a mean difference of about 1.0 kcal. The signed differences (PD vs DD) for esters and ketones are in the same direction, but the effects are slightly less pronounced, about 0.7 kcal.

Table 13 breaks down the free energy of solvation into its components for several molecules with four carbons and various functional groups. These results demonstrate the trends previously discussed.

These differences in electrostatic contributions are compensated by the surface tensions in each respective model, and the user should be aware of the different partitionings in the different models if attempting to make physical sense of the individual terms. These differences could be more important for charged systems where the electrostatics play a larger role. Even for neutrals, the partition of effects between ENP and CDS can have significant effects on the solute wave functions, induced changes in dipole moments, and other properties, so that the lower level models might be more suited for calculating total free energies of solvation than for calculating such properties.

Clearly the quantitative success of the present implementation of pairwise descreening depends on our strategy of scaling the atomic radii in two-center terms. Schaefer and Karplus[47] have suggested an alternative approach. Instead of adjusting effective radii using experimental data for free energies of solvation, they used effective average volumes of united-atom groups adjusted to experimental crystal structure data. In addition, they introduced a Gaussian to smooth the fluctuations of the molecular density. The Gaussian width parameter was set equal to a scale factor times the van der Waals radius, and this scale factor was adjusted to theoretical atomic solvation energies; the value was assigned as 1.6 or 1.8−2.0, depending in part on the parameter set. The former value was obtained for use with CHARMM parameter set 22 in order to minimize the average relative energy of the calculated atomic solvation energies from other theoretical values for 15 neutrals and two ions containing H, C, O, and N. The mean signed and unsigned deviations of the resulting molecular solvation energies, including two-center terms, for the 15 neutrals were both 1.12 kcal. In this parameter set the van der Waals radius of polar hydrogens is 0.2245 Å. Schaefer and Karplus introduced a geometry-dependent procedure for assignment of charge radii, which have discontinuous derivatives as functions of molecular geometry. These aspects are illustrative of the differences in the two approaches.

5.5. Rational Selection of Model. The flow chart in Figure 1 provides a rational scheme for choosing among current-generation-specific SM*x* models for aqueous solvation. The first issue is whether the solute contains P since some of the methods are not parametrized for phosphorus. When the solute does not contain P, the second choice is between AM1 and PM3. The choice between AM1 and PM3 can be made in several ways. In many cases one knows from previous experience or from gas-phase calculations on the system of interest whether AM1 or PM3 is the method of choice for a given range of solutes or class of reactions. The single most critical gas-phase attribute for making this choice is molecular geometry. When the gas-phase situation does not provide a choice, either because of insufficient data or because the two Hamiltonians lead to comparable errors, one may choose between AM1 and PM3 by consulting the tables of mean unsigned errors in free energies of solvation for various functional groups and choosing the
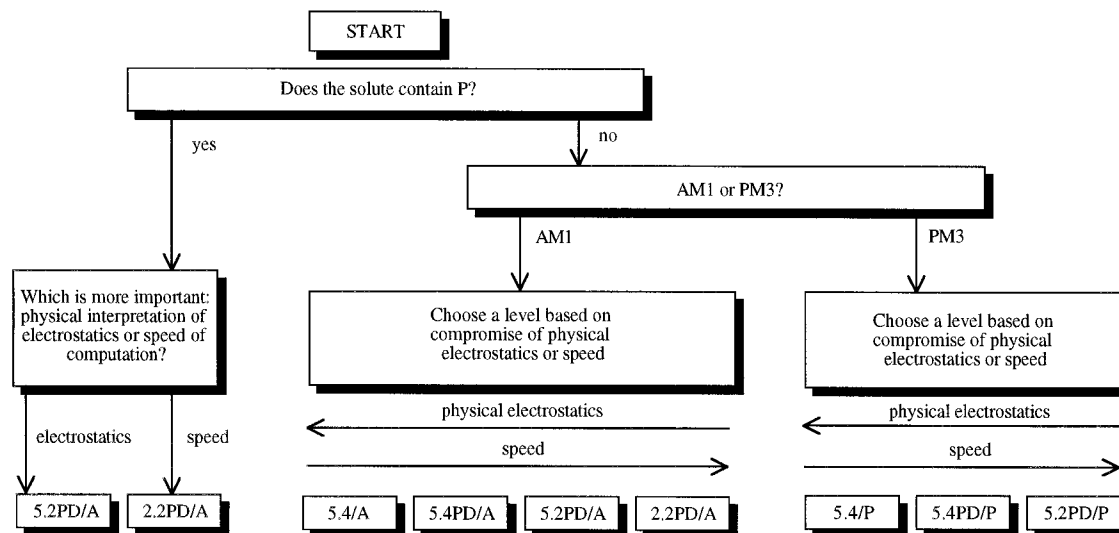
Figure 1. Flow chart for rational selection among current-generation SM*x* aqueous solvation models. See text for amplification of choice boxes.

Hamiltonian that happens, on the average, to lead to smaller errors for the group or groups of interest.

The final decision in the choice of model is a compromise between the theoretical justification of the model and the speed of computation. For the most part, using the more theoretically justified formulations, in which the formalism more directly attempts to faithfully simulate the real system, does not reduce the error and may even increase it. Nevertheless, it does allow for a more "physical" interpretation of the various terms, especially the separation into ENP and CDS contributions, which aspect is abbreviated as "physical electrostatics" in the flow chart. One may also wish to choose the methods with more physical electrostatics when the system of interest differs significantly from any molecules in the training set (an example would be transition-state modeling) since it is well-known that making up for simplifications in the physics of the model by a semiempirical choice of parameters is more trustworthy for solutes that are similar to those in the training set than for those that are not. The successive simplifications involved in the current-generation models may be summarized as follows:

5.4 → 5.4PD: replace density descreening by pairwise descreening
5.4PD → 5.2PD: replace class IV charges by class II charges
5.2PD → 2.2PD: replace SM5 functional forms by SM2 functional forms that contain empirical modifications of the electrostatic terms for O−O and N−H interactions

When one is willing to accept one or more of these simplifications, or when one requires the greater speed they provide, one sometimes obtains a "bonus" of a slightly smaller mean error, which results partly from the large number of parameters in PD models and partly from the use of genetic algorithms to make fine adjustments of the Coulomb radii in the present work.

The flow chart does not include the universal models, i.e., SM4 and /U models, nor does it include SM1, SM1a, SM2, SM2.1, SM2.2, SM3, or SM3.1. The SM4 models were also excluded because they are only available for alkanes, whereas the chart refers to aqueous models. The excluded aqueous models fall into one of the following classes:

(i) previous-generation models: SM2 and SM3

(ii) obsolete: SM1, SM1a, and SM2.2

(iii) designed to allow greater flexibility than SM2 and SM3 in quadrature grids for exploratory runs: SM2.1 and SM3.1

(iv) designed for use with other Hamiltonians: SM5.4/U and SM5.4PD/U

Models in classes (i) and (iv) and perhaps occasionally those in class (ii) may still be useful in difficult cases for obtaining a "second opinion" (in the medical sense of an independent estimation) or even a "third opinion." The alternatives in the flow chart can also be used for such purposes, although, since they have many similarities to each other, the estimation will often be only semi-independent. Two final notes: (1) The flexibility advantages of SM2.1 and SM3.1, which are discussed in ref 5 and in the AMSOL manual, are only with respect to SM2 and SM3. The new models presented in ref 10 and here all share this flexibility for exploratory runs. (2) Models SM3 and SM3.1 are no longer recommended for compounds containing N, and models SM2, SM2.1, SM3, and SM3.1 are no longer recommended for phosphorus.

6. Summary

We have presented six new parametrized models for aqueous free energies of solvation. They are slightly more accurate (in an empirical sense) than previous models, although not necessarily more reliable, but their main advantage is that they are significantly simpler in a computational sense. When this is fully exploited, e.g., by coding the analytic gradients, the new models should allow useful calculations on very large and complex systems.

Supporting Information Available: A discussion of how we obtained phosphorus parameters and results for several phosphorus-containing neutral and ionic solutes are presented. In addition, a table of the surface tension coefficients obtained for the SM5.4PD/U model is included. Finally, a table of the values of ΔG_S° calculated for all 219 neutral solutes by the

SM2.2PD/A, SM5.2PD/A, SM5.2PD/P, SM5.4PD/A, and SM5.4PD/P models is presented, and the results are compared to experimental data[37−39] for all molecules in the test suite (16 pages). Ordering information is given on any current masthead page.

References and Notes

(1) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 8305. Erratum: **1991**, *113*, 9901.

(2) Cramer, C. J.; Truhlar, D. G. *Science* **1992**, *256*, 213.

(3) Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **1992**, *13*, 1089.

(4) Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 629.

(5) Liotard, D. A.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **1995**, *16*, 422.

(6) Giesen, D. J.; Storer, J. W.; Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1995**, *117*, 1057.

(7) Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1995**, *99*, 7137.

(8) Storer, J. W.; Giesen, D. J.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G.; Liotard, D. A. *Am. Chem. Soc. Symp. Ser.* **1994**, *568*, 24.

(9) Barrows, S. E.; Dulles, F. J.; Cramer, C. J.; Freuch, A. D.; Truhlar, D. G. *Carbohydr. Rev.* **1995**, *276*, 219.

(10) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 16385.

(11) Hoijtink, G. J.; Boer, E. d.; Meij, P. H. v. d.; Weijland, W. P. *Recl. Trav. Chim. Pays-Bas.* **1956**, *75*, 487.

(12) Peradejordi, F. *Cah. Phys.* **1963**, *17*, 393.

(13) Jano, I. *C. R. Hebd. Seances Acad. Sci.* **1965**, *261*, 103.

(14) Tapia, O. In *Quantum Theory of Chemical Reactions*; Davdel, R., Pullman, A., Salem, L., Veillard, A., Eds.; Reidel: Dordrecht, 1981; Vol. 2, p 25.

(15) Contreras, R.; Aizman, A. *Int J. Quantum Chem.* **1985**, *27*, 293.

(16) Kozaki, T.; Morihashi, K.; Kikuchi, O. *J. Mol. Struct. (THEOCHEM)* **1988**, *168*, 265.

(17) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164.

(18) Still, W. A.; Tempczyk, A.; Hawley, R. C.; Hendrickson T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(19) Bucher, M.; Porter, T. L. *J. Phys. Chem.* **1986**, *90*, 3406.

(20) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.

(21) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87.

(22) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.

(23) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.

(24) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199.

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

(26) Cramer, C. J.; Truhlar, D. G. In *Solvent Effects and Chemical Reactivity*; Tapia, O., Bertran, J., Eds.; Kluwer: Dordecht, 1996; p 1.

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

(28) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221.

(29) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122.

(30) Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045.

(31) Klopman, G. *J. Am. Chem. Soc.* **1964**, *86*, 4550.

(32) Ohno, K. *Theor. Chim. Acta* **1964**, *2*, 219.

(33) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(34) Hawkins, G. D.; Lynch, G. C.; Giesen, D. J.; Rossi, I.; Storer, J. W.; Cramer, C. J.; Truhlar, D. G. *QCPE Bull.* **1996**, *16*, 11.

(35) Pople, J. A.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S129.

(36) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.

(37) Hine J.; Mookerjee, P. K. *J. Org. Chem.* **1975**, *40*, 292.

(38) Cabani, S.; Paolo, G.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563.

(39) Pearson, R. G. *J. Am. Chem. Soc.* **1986,** *108*, 6109.

(40) Abraham, M. H.; Whiting, G. S. *J. Chem. Soc., Perkin Trans.* **1990**, *2*, 291.

(41) Easton, E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93*, 281.

(42) Wagman, D. D.; Evans, W. H.; Parker, V. B.; Schumm, R. H.; Halow, I.; Bailey, S. M.; Churney, K. L.; Nuttall, R. C. *J. Phys. Chem. Ref Data* **1982**, *11*, Suppl. 2.

(43) Wolfenden, R.; Williams, R. *J. Am. Chem. Soc.* **1983**, *105*, 1028.

(44) George, P.; Witonsky, R. J.; Trachtman, M.; Wu, C.; Dorwart, W.; Richman, W.; Shurayh, F.; Lentz, B. *Biochim. Biophys. Acta* **1970**, *223*, 1.

(45) Armstrong, D. R.; Perkins, P. G.; Stewart, J. J. P. *J. Chem. Soc., Dalton Trans.* **1973**, *1973*, 838.

(46) Carroll, D. L. *AIAA J.* **1996**, *34*, 338.

(47) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578.

Discussion of phosphorus parameterization. As mentioned in Section 3 of the text, obtaining parameters for phosphorus poses several problems. The most significant of these is a paucity of experimental free energy of solvation data for neutrals. To parameterize the PD models for phosphorus, we used 4 neutral molecules, namely phosphine and three esters, $OP(OCH_3)_3$, $OP(OC_2H_5)_3$, and $OP(OC_3H_7)_3$. In addition, we used two ions for which we have good experimental data PH_2^- and $(CH_3)_3 PH_4^+$. Finally we used three more ions for which we had less reliable free energy of solvation data, and we used experimental heats of solvation for $H_2PO_4^-$, HPO_4^{2-}, and PO_4^{3-}. From this data, we determined values for the coulombic radius for phosphorus (ρ_P), for three scaling factors involving phosphorus (S_{HP}, S_{PH}, and S_{Pk}), and for the phosphorus surface tension coefficients.

To determine which surface tensions are necessary when parameterizing for phosphorus, we first considered surface tension types similar to those used for nitrogen, because phosphorus is in the same column of the periodic table as nitrogen. Thus, for the SM2.2PD/A model we considered using a surface tension like that defined in eq. (12) involving the surface tension coefficients $\sigma_P^{(0)}$ and $\sigma_P^{(1)}$. Following our usual method for determining the surface tension coefficients for each trial set of nonlinear parameters in the genetic algorithm by minimizing the sum of the squared errors for the neutrals, we found that including the $\sigma_P^{(1)}$ surface tension coefficient did not significantly improve the mean unsigned error for the neutral and ionic phosphorus compounds in our training set. Therefore, we used only a single nonzero surface tension coefficient of the type $\sigma_P^{(0)}$ when parameterizing for phosphorus in the SM2.2PD/A model.

To decide which surface tension coefficients to use when parameterizing the SM5.2PD/A model for phosphorus, we considered $\sigma_P^{(0)}$, σ_{HP}, σ_{OP}, and σ_{PC} because each coefficient has a counterpart in our treatment of nitrogen. Since the four neutrals in the phosphorus portion of our training set do not contain any molecule where phosphorus

is bonded to a carbon, we set the σ_{PC} coefficient to zero. Furthermore, we found that including σ_{HP} did not improve the fit. Thus we included only $\sigma_P^{(0)}$ and σ_{OP}. The function, T_{OP}, from eq. (18) is represented

$$T_{OP} = \begin{cases} \exp\left[-\left(\dfrac{\Delta R_{OP}}{\Delta R_{OP} - R_{OP} + \overline{R}_{OP}}\right)\right] & R_{OP} \leq \overline{R}_{OP} + \Delta R_{OP} \\[2em] 0 & \text{otherwise} \end{cases} \tag{S-1}$$

where ΔR_{OP} and $\overline{R}_{OP}$ are parameters. Following the same kind of procedure as used in SM5.4/U[10] to specify T_{ON}, we set ΔR_{OP} to the maximum width value, 0.3 Å, and we determined $\overline{R}_{OP}$ such that $\overline{R}_{OP} - \Delta R_{OP}$ is just a little larger than the largest O–P bond distance. This yielded $\overline{R}_{OP}$ equal to 2.0 Å.

For each of the two models parameterized for phosphorus, namely SM2.2PD/A and SM5.2PD/A, we used the genetic algorithm to optimize ρ_P and the three scaling parameters to minimize the mean weighted unsigned error (to within a hundredth of a kcal) between the calculated free energies of solvation and the experimental free energies of solvation for the 4 neutral molecules and 8 ions. The neutrals were weighted 1, the ion with experimental data was weighted 1/6, the four ions with estimated ΔG_S^0 were weighted 1/12, and $H_2PO_4^-$, HPO_4^{2-}, and PO_4^{3-} were weighted 1/12, 1/48, and 1/108, respectively. The latter two weights were arrived at by noting that the doubly and triply charged ions have solvation energies about 4 and 9 times larger than singly charged ions. During the minimization, all non-phosphorus parameters were fixed to the previously determined values, and the surface tension coefficients involving phosphorus were found to minimize the sum of the squared errors in the 4 phosphorus neutrals.

Once the genetic algorithm began to converge, we relaxed the parameters into a self consistent minimum. Upon fully optimizing the parameters, some of the solution-phase geometries in the SM5.2PD/A model changed substantially enough that it was necessary to refit the phosphorus parameters with the new geometries. Geometry changes are always a

possibility when optimizing parameters with a single SCF calculation from an approximate

solvated geometry. This type of change sometimes occurred (and was accounted for) when

the SM5.4/A and SM5.4/P models were developed,[10] but phosphorus is the only element

for which we encountered what we considered to be significant geometry changes

elsewhere during the course of the present project. The resulting final parameters for

phosphorus used in the SM2.2PD/A and SM5.2PD/A models are listed in Tables 2–5 of

the main text.

Table S1 contains the results for the 8 phosphorus ions. Even ignoring the ions for

which we have only heat of solvation experimental data, the error in the ions is larger than

we have experienced with other atom types.

The results for the 4 phosphorus neutrals are in Table S2, where they are compared

to results for some phosphorus containing molecules outside the training set and to the

carbon esters in the training set. Although we don't have experimental free energies for the

phosphorus-containing molecules outside the training set, they provide a test of the stability

of our predictions, and indeed, some instability of our predictions can be seen. Notice how

the results for the carbon esters are more stable from method to method. In the case of the

dimethyl ester of phosphonic acid, the range of calculated free energies is over 10 kcal/mol,

and for the dimethyl ester of methylphosphonic acid, the range of calculated free energies is

about 7 kcal/mol, which indicates that one should exercise caution when applying SMx

models to phosphorus-containing molecules. Although the parameterization for

phosphorus-containing compounds is much less reliable than for other kinds of solutes

(due to the scarcity of experimental data), we decided to present phosphorus parameters for

two of the models anyway, simply because there is so much interest in phosphorus

compounds.

Table S1. Calculated Free Energies of Solvation (kcal/mol) for Phosphorus Ions

Molecule	Weight[a]	SM2	SM2.1	SM3	SM3.1	SM2.2 PD/A	SM5.2 PD/A	Expt
In training set								
PH_2^-	1/6	-67	-67	-67	-67	-67	-73	-67[b]
PH_4^+	1/12	-60	-73	-59	-72	-67	-66	-73[b]
$MePH_3^+$	1/12	-57	-66	-60	-70	-51	-66	-66[b]
$Me_2PH_2^+$	1/12	-51	-61	-59	-67	-58	-70	-57[b]
Me_3PH^+	1/12	-47	-57	-56	-63	-65	-73	-53[b]
$H_2PO_4^-$	1/12	-111	-105	-86	-84	-84	-76	-76[c]
HPO_4^{2-}	1/48	-302	-292	-266	-260	-273	-252	-299[c]
PO_4^{3-}	1/108	-629	-614	-570	-559	-582	-538	-637[c]
Not in training set								
$Me_2PO_4^-$	0	-84	-81	-72	-69	-85	-75	

[a]Weight of unsigned error in unfitness function of genetic algorithm

[b]Ref. 39

[c]The experimental value shown for these ions in the training set is the heat of solvation
rather than the free energy of solvation. (Ref. 44)

Table S2. Calculated Free Energies of Solvation (kcal/mol) for Neutral Phosphorus-Containing Molecules

Molecule	SM2	SM2.1	SM3	SM3.1	SM2.2 PD/A	SM5.2 PD/A	Expt.
Phosphorus compounds in training set							
phosphine	0.6	0.6	0.5	-0.3	0.6	0.6	0.6
$OP(OCH_3)_3$	-20.9	-20.2	-12.3	-12.2	-7.3	-8.6	-8.7
$OP(OC_2H_5)_3$	-15.4	-13.9	-12.1	-11.3	-6.8	-7.2	-7.8
$OP(OC_3H_7)_3$	-10.2	-9.5	-10.2	-8.7	-6.0	-7.0	-6.1
Mean unsigned error	6.0	5.3	3.0	2.5	0.6	0.4	
Phosphorus compounds not in training set							
phosphonic acid, dimethyl ester	-12.3	-12.6	-14.8	-16.0	-5.3	-10.1	
dimethyl hydrogen phosphite	-11.3	-10.9	-11.8	-12.2	-13.7	-14.6	
methylphosphonic acid, dimethyl ester	-11.8	-11.5	-12.6	-12.1	-9.8	-5.5	
Carbon esters in training set							
methyl methanoate	-4.8	-4.8	-6.2	-6.3	-3.9	-3.1	-2.8
methyl ethanoate	-4.0	-3.9	-4.7	-4.7	-4.3	-4.1	-3.3
ethyl methanoate	-4.1	-4.2	-5.5	-5.6	-3.6	-2.9	-2.7
methyl propanoate	-2.9	-2.9	-3.6	-3.7	-3.6	-3.6	-2.9
ethyl ethanoate	-3.3	-3.3	-4.0	-4.0	-4.2	-3.9	-3.1
methyl butanoate	-2.6	-2.5	-3.2	-3.3	-3.0	-3.2	-2.8
propyl ethanoate	-3.1	-3.0	-3.8	-3.7	-3.7	-3.7	-2.9
methyl pentanoate	-2.3	-2.2	-3.1	-3.1	-2.6	-3.0	-2.6
butyl ethanoate	-2.9	-2.7	-3.5	-3.5	-3.3	-3.5	-2.6
methyl hexanoate	-2.1	-1.9	-2.8	-2.8	-2.3	-2.8	-2.5
pentyl ethanoate	-2.7	-2.4	-3.4	-3.2	-3.0	-3.3	-2.5
methyl octanoate	-1.6	-1.4	-2.4	-2.3	-1.8	-2.6	-2.0

Table S3. Microscopic Surface Tensions (cal mol^{-1} Å^{-2}) for SM5.4PD/U by Atom Type and Atom Pair

k	$\sigma_k^{(0)}$	$\sigma_{Hk'}$
H	27.69	–
C	70.00	-30.16
N	8.23	-62.24
O	-64.45	-18.88
F	17.51	–
S	-45.88	37.63
Cl	-1.26	–
Br	-5.52	–
I	-7.66	–

k	k'	$\sigma_{kk'}^{(1)}$	$\sigma_{kk'}^{(2)}$
O	O	24.07	–
O	N	146.77	–
O	C	106.54	–
C	C	-50.09	-17.42
S	S	37.09	–
N	C	-33.78	–

Table S4. Calculated and Experimental[a] Free Energies of Solvation, ΔG_S^0 (kcal/mol), for All Neutral Solutes Used in the Parameterizations of the Five New Models

Molecule	SM2.2PD	SM5.2PD		SM5.4PD		Expt.[a]
	/A	/A	/P	/A	/P	
Compounds containing at most C, H, and/or O						
Unbranched alkanes						
methane	1.39	2.05	1.99	2.34	2.36	2.00
ethane	1.61	1.73	1.44	1.38	1.43	1.83
propane	1.82	1.86	1.57	1.48	1.53	1.96
n-butane	2.02	2.00	1.74	1.64	1.70	2.08
n-pentane	2.22	2.14	1.90	1.80	1.86	2.33
n-hexane	2.43	2.28	2.08	1.97	2.03	2.49
n-heptane	2.65	2.42	2.25	2.13	2.20	2.62
n-octane	2.87	2.56	2.42	2.30	2.36	2.89
mean unsigned error	0.16	0.16	0.35	0.48	0.43	
mean signed error	-0.15	-0.15	-0.35	-0.39	-0.34	
root mean square error	0.24	0.18	0.38	0.48	0.43	0.35[b]
Branched alkanes						
2-methylpropane	2.04	2.10	1.80	1.84	1.80	2.32
2,2-dimethylpropane	2.32	2.42	2.00	2.24	2.04	2.50
2-methylpentane	2.39	2.37	2.10	2.18	2.11	2.52
2,4-dimethylpentane	2.51	2.53	2.30	2.54	2.36	2.88
2,2,4-trimethylpentane	2.75	2.81	2.50	2.91	2.59	2.85
mean unsigned error	0.21	0.17	0.47	0.29	0.44	
mean signed error	-0.21	-0.17	-0.47	-0.27	-0.44	
root mean square error	0.23	0.20	0.48	0.32	0.45	0.22[b]
Cycloalkanes						
cyclopropane	1.13	1.03	0.89	0.64	0.56	0.75
cyclopentane	1.40	1.40	1.31	1.03	1.20	1.20
cyclohexane	1.52	1.55	1.45	1.32	1.43	1.23
methylcyclohexane	1.63	1.75	1.72	1.75	1.77	1.71
cis-1,2-dimethylcyclohexane	1.77	2.03	1.93	2.19	2.06	1.58
mean unsigned error	0.23	0.26	0.17	0.20	0.19	
mean signed error	0.20	0.26	0.17	0.09	0.11	
root mean square error	0.25	0.29	0.20	0.29	0.25	0.34[b]
Alkenes						
ethene	1.12	1.88	1.93	2.32	2.06	1.27
propene	1.20	1.59	1.58	1.68	1.57	1.27
s-trans-1,3-butadiene	0.73	1.38	1.62	1.93	1.62	0.61
2-methylpropene	1.33	1.68	1.69	1.83	1.79	1.16

1-butene	1.46	1.72	1.72	1.81	1.71	1.38
cyclopentene	0.60	0.57	0.61	0.33	0.45	0.56
1-pentene	1.69	1.87	1.90	1.99	1.89	1.66
E-2-pentene	1.47	1.47	1.46	1.32	1.37	1.34
1-hexene	1.92	2.02	2.07	2.16	2.06	1.68
mean unsigned error	0.11	0.36	0.40	0.55	0.42	
mean signed error	0.06	0.36	0.40	0.49	0.40	
root mean square error	0.13	0.42	0.49	0.67	0.52	0.37[b]
Alkynes						
ethyne	-0.58	0.22	0.28	0.50	0.35	-0.01
propyne	-0.86	-0.49	-0.55	-0.67	-0.65	-0.31
1-butyne	-0.32	-0.05	-0.13	-0.22	-0.12	-0.16
1-pentyne	0.02	0.19	0.14	0.07	0.16	0.01
1-hexyne	0.28	0.35	0.31	0.25	0.32	0.29
mean unsigned error	0.26	0.15	0.14	0.20	0.19	
mean signed error	-0.25	0.08	0.05	0.02	0.05	
root mean square error	0.36	0.17	0.18	0.28	0.23	0.20[b]
Arenes						
benzene	-1.01	-1.11	-1.06	-1.30	-1.36	-0.87
toluene	-0.95	-1.05	-0.92	-1.04	-0.98	-0.89
ethylbenzene	-0.57	-0.76	-0.60	-0.65	-0.56	-0.80
o-xylene	-0.95	-1.09	-0.90	-0.93	-0.76	-0.90
m-xylene	-0.93	-1.00	-0.78	-0.79	-0.61	-0.84
p-xylene	-0.89	-0.97	-0.77	-0.77	-0.59	-0.81
naphthalene	-2.64	-2.92	-2.86	-2.92	-2.79	-2.39
anthracene	-4.05	-4.59	-4.65	-4.33	-4.30	-4.23
mean unsigned error	0.13	0.23	0.17	0.18	0.23	
mean signed error	-0.03	-0.22	-0.10	-0.13	-0.03	
root mean square error	0.15	0.27	0.24	0.26	0.27	1.16[b]
Alcohols						
methanol	-6.42	-5.75	-5.54	-5.73	-5.55	-5.11
ethanol	-5.75	-5.09	-4.90	-5.39	-5.12	-5.01
1,2-ethanediol	-10.92	-9.93	-9.85	-8.99	-9.20	-9.30
1-propanol	-5.29	-4.81	-4.61	-5.01	-4.78	-4.83
2-propanol	-4.89	-4.41	-4.16	-4.03	-4.21	-4.76
1-butanol	-4.82	-4.65	-4.61	-4.53	-4.69	-4.72
2-methyl-2-propanol	-4.39	-3.81	-3.38	-3.34	-3.47	-4.51
cyclopentanol	-4.81	-4.61	-4.24	-3.94	-4.26	-5.49
1-pentanol	-4.54	-4.48	-4.43	-4.35	-4.52	-4.47
phenol	-5.95	-5.89	-5.81	-6.23	-6.15	-6.62
1-hexanol	-4.47	-4.56	-4.61	-4.43	-4.76	-4.36
o-cresol	-5.74	-5.12	-5.37	-5.45	-5.64	-5.87
m-cresol	-5.88	-5.92	-5.72	-5.92	-5.80	-5.49

p-cresol	-5.84	-5.86	-5.73	-5.98	-5.83	-6.14
1-heptanol	-4.24	-4.43	-4.44	-4.28	-4.59	-4.24
1-octanol	-3.84	-4.09	-3.92	-3.88	-4.02	-4.09
mean unsigned error	0.44	0.37	0.44	0.44	0.36	
mean signed error	-0.17	0.10	0.23	0.22	0.15	
root mean square error	0.63	0.48	0.56	0.59	0.49	1.24[b]
Ethers						
dimethyl ether	-2.32	-2.47	-2.33	-2.44	-2.25	-1.92
tetrahydrofuran	-2.77	-3.35	-3.49	-3.42	-3.46	-3.47
1,4-dioxane	-4.38	-5.46	-6.16	-5.33	-5.69	-5.05
diethyl ether	-2.13	-1.24	-1.05	-1.77	-1.41	-1.76
methyl propyl ether	-1.85	-1.62	-1.41	-1.77	-1.55	-1.66
methyl isopropyl ether	-2.01	-1.52	-1.28	-1.53	-1.34	-2.01
1,2-dimethoxyethane	-4.21	-3.96	-3.66	-3.84	-3.48	-4.84
t-butyl methyl ether	-1.77	-1.12	-0.96	-1.01	-0.94	-2.21
anisole	-3.13	-3.42	-3.00	-3.19	-3.04	-1.04
mean unsigned error	0.61	0.72	0.85	0.64	0.75	
mean signed error	-0.07	-0.02	0.07	-0.04	0.09	
root mean square error	0.83	0.98	1.01	0.92	0.97	1.36[b]
Aldehydes						
ethanal	-4.29	-4.16	-3.99	-4.07	-4.07	-3.50
propanal	-3.63	-3.42	-3.46	-3.60	-3.57	-3.44
butanal	-3.14	-3.10	-3.13	-3.21	-3.21	-3.18
pentanal	-2.34	-3.06	-3.14	-2.90	-3.04	-3.03
benzaldehyde	-4.34	-4.64	-4.71	-4.54	-4.74	-4.02
octanal	-2.13	-2.55	-2.44	-2.57	-2.53	-2.29
mean unsigned error	0.36	0.28	0.25	0.28	0.28	
mean signed error	-0.07	-0.24	-0.23	-0.24	-0.28	
root mean square error	0.46	0.38	0.35	0.34	0.39	0.53[b]
Ketones						
propanone	-4.40	-4.51	-4.28	-4.24	-4.16	-3.85
butanone	-3.86	-3.83	-3.75	-3.78	-3.71	-3.64
cyclopentanone	-3.10	-4.03	-3.97	-3.77	-3.80	-4.68
2-pentanone	-3.42	-3.53	-3.41	-3.39	-3.35	-3.53
3-pentanone	-3.34	-3.16	-3.25	-3.35	-3.31	-3.41
2-hexanone	-2.85	-3.23	-3.05	-3.03	-2.98	-3.29
3,3-dimethylbutanone	-3.24	-3.54	-3.33	-3.08	-3.13	-2.89
2-heptanone	-2.88	-3.25	-3.07	-3.06	-3.01	-3.04
4-heptanone	-2.43	-2.60	-2.58	-2.58	-2.62	-2.93
methyl phenyl ketone	-4.84	-5.25	-4.93	-4.86	-4.76	-4.58
5-nonanone	-1.92	-2.39	-2.30	-2.32	-2.34	-2.67
2-octanone	-2.65	-3.12	-2.91	-2.92	-2.86	-2.88

mean unsigned error	0.44	0.35	0.28	0.26	0.25	
mean signed error	0.21	-0.09	0.05	0.08	0.11	
root mean square error	0.59	0.42	0.34	0.35	0.33	0.63^b

Carboxylic acids

ethanoic acid	-7.17	-6.61	-6.37	-6.14	-6.27	-6.70
propanoic acid	-6.24	-5.94	-5.91	-5.64	-5.75	-6.47
butanoic acid	-5.52	-5.47	-5.48	-5.17	-5.35	-6.36
pentanoic acid	-5.16	-5.27	-5.31	-5.00	-5.18	-6.16
hexanoic acid	-4.80	-5.07	-5.11	-4.81	-4.99	-6.21
mean unsigned error	0.79	0.71	0.74	1.03	0.87	
mean signed error	0.60	0.71	0.74	1.03	0.87	
root mean square error	0.89	0.80	0.79	1.07	0.91	0.20^b

Esters

methyl methanoate	-3.93	-3.12	-3.42	-4.05	-3.90	-2.78
methyl ethanoate	-4.34	-4.05	-3.78	-3.98	-3.83	-3.32
ethyl methanoate	-3.63	-2.85	-3.04	-3.80	-3.58	-2.65
methyl propanoate	-3.62	-3.61	-3.49	-3.63	-3.51	-2.93
ethyl ethanoate	-4.15	-3.89	-3.44	-3.87	-3.60	-3.10
methyl butanoate	-2.98	-3.21	-3.10	-3.14	-3.12	-2.83
propyl ethanoate	-3.73	-3.72	-3.19	-3.55	-3.34	-2.86
methyl pentanoate	-2.63	-3.02	-2.92	-2.97	-2.96	-2.57
butyl ethanoate	-3.34	-3.47	-2.97	-3.30	-3.13	-2.55
methyl hexanoate	-2.28	-2.81	-2.72	-2.78	-2.77	-2.49
pentyl ethanoate	-3.02	-3.29	-2.79	-3.12	-2.96	-2.45
methyl octanoate	-1.76	-2.56	-2.39	-2.50	-2.46	-2.04
mean unsigned error	0.65	0.59	0.39	0.68	0.55	
mean signed error	-0.57	-0.59	-0.39	-0.68	-0.55	
root mean square error	0.75	0.63	0.40	0.74	0.60	0.32^b

Bifunctional CHO

2-propen-1-ol	-5.35	-4.87	-4.62	-4.41	-4.52	-5.08
2-methoxyethanol	-7.51	-7.04	-7.23	-7.20	-7.25	-6.77
butenyne	-1.07	-0.47	-0.36	-0.36	-0.40	0.04
m-hydroxybenzaldehyde	-8.46	-8.58	-9.26	-9.03	-9.12	-9.51
p-hydroxybenzaldehyde	-8.77	-8.79	-9.45	-9.59	-9.69	-10.48
mean unsigned error	0.98	0.72	0.52	0.58	0.54	
mean signed error	0.13	0.41	0.18	0.24	0.17	
root mean square error	1.09	0.91	0.59	0.60	0.55	3.73^b

Inorganic H,O compounds

Water	-6.67	-10.15	-9.63	-10.26	-9.22	-6.31
H_2	0.00	3.15	3.16	3.06	3.07	2.33
mean unsigned error	1.35	2.33	2.07	2.34	1.82	
mean signed error	-1.35	-1.51	-1.25	-1.61	-1.08	

root mean square error	1.67	2.78	2.42	2.84	2.12	4.32[b]

Subtotal for compounds containing at most C, H, and/or O

mean unsigned error	0.43	0.43	0.43	0.49	0.44	
mean signed error	-0.09	-0.04	0.00	-0.03	-0.01	
root mean square error	0.63	0.66	0.61	0.71	0.61	3.06[b]

Compounds containing N
Aliphatic amines

methylamine	-4.61	-5.06	-5.18	-5.27	-5.01	-4.56
ethylamine	-3.80	-4.51	-4.47	-4.34	-4.24	-4.50
dimethylamine	-3.98	-3.95	-4.67	-4.69	-4.76	-4.29
azetidine	-3.54	-5.55	-6.21	-6.05	-6.38	-5.56
propylamine	-3.63	-4.40	-4.30	-4.13	-4.06	-4.39
trimethylamine	-3.35	-2.37	-3.06	-3.39	-3.10	-3.23
pyrrolidine	-3.73	-4.72	-5.02	-4.94	-5.18	-5.48
piperazine	-8.42	-9.09	-8.52	-9.33	-9.22	-7.40
butylamine	-3.41	-4.26	-4.15	-3.91	-3.85	-4.29
diethylamine	-2.71	-3.31	-3.14	-3.19	-3.19	-4.07
N-methylpiperazine	-8.03	-7.53	-6.46	-7.81	-7.05	-7.77
piperidine	-3.41	-4.63	-4.67	-4.24	-4.56	-5.11
pentylamine	-3.19	-4.11	-3.97	-3.74	-3.68	-4.10
N,N'-dimethylpiperazine	-7.58	-5.94	-4.38	-6.18	-4.72	-7.58
dipropylamine	-2.53	-3.54	-3.30	-3.37	-3.34	-3.66
mean unsigned error	0.86	0.50	0.67	0.59	0.72	
mean signed error	0.67	0.20	0.30	0.09	0.25	
root mean square error	1.07	0.73	1.02	0.77	1.00	1.39[b]

Aromatic amines

pyridine	-4.72	-4.86	-4.92	-5.04	-5.04	-4.70
2-methylpyrazine	-6.98	-7.49	-7.61	-6.48	-6.88	-5.57
aniline	-5.26	-6.02	-5.46	-5.35	-5.74	-5.49
2-methylpyridine	-3.88	-4.32	-3.91	-4.47	-4.18	-4.63
3-methylpyridine	-4.63	-4.95	-4.87	-4.67	-4.67	-4.77
4-methylpyridine	-4.71	-4.96	-4.85	-4.74	-4.68	-4.94
2-ethylpyrazine	-6.49	-6.56	-7.08	-5.58	-6.36	-5.51
2,4-dimethylpyridine	-3.82	-4.40	-3.86	-4.16	-3.81	-4.86
2,5-dimethylpyridine	-3.81	-4.54	-3.92	-4.36	-3.91	-4.72
2,6-dimethylpyridine	-3.06	-3.89	-2.83	-4.03	-3.29	-4.60
mean unsigned error	0.72	0.55	0.83	0.36	0.67	
mean signed error	0.24	-0.22	0.05	0.09	0.12	
root mean square error	0.89	0.77	1.09	0.45	0.80	0.37[b]

Nitriles

ethanonitrile	-4.51	-3.69	-3.67	-4.43	-4.65	-3.89
propanonitrile	-3.66	-3.31	-3.37	-3.85	-3.80	-3.85
butanonitrile	-3.11	-2.92	-2.94	-3.28	-3.15	-3.64
benzonitrile	-4.51	-4.27	-4.15	-3.56	-3.45	-4.10
mean unsigned error	0.44	0.41	0.36	0.36	0.49	
mean signed error	-0.08	0.32	0.34	0.09	0.10	
root mean square error	0.47	0.47	0.44	0.43	0.56	0.16[b]

Nitrohydrocarbons

nitroethane	-2.47	-4.10	-3.56	-4.21	-3.66	-3.71
1-nitropropane	-1.20	-3.12	-3.14	-3.27	-3.17	-3.34
2-nitropropane	-2.90	-4.01	-2.93	-3.71	-2.83	-3.14
1-nitrobutane	-0.54	-2.89	-2.89	-3.08	-2.95	-3.08
nitrobenzene	-2.65	-3.39	-4.25	-3.32	-4.35	-4.12
2-methyl-1-nitrobenzene	-2.58	-3.51	-4.25	-3.32	-4.11	-3.59
mean unsigned error	1.44	0.41	0.26	0.37	0.23	
mean signed error	1.44	-0.01	-0.01	0.01	-0.01	
root mean square error	1.62	0.50	0.32	0.46	0.28	0.36[b]

Amides

ethanamide	-7.47	-8.36	-7.35	-9.60	-9.92	-9.71
E-N-methylacetamide	-11.81	-9.88	-9.39	-7.88	-8.47	-10.00
Z-N-methylacetamide	-10.88	-7.95	-8.00	-7.60	-10.05	-10.00
mean unsigned error	1.65	1.17	1.66	1.54	0.60	
mean signed error	-0.15	1.17	1.66	1.54	0.42	
root mean square error	1.74	1.42	1.82	1.85	0.89	0.14[b]

Bifunctional

2-methoxyethanamine	-7.07	-7.15	-7.29	-6.58	-7.23	-6.55
morpholine	-6.73	-7.46	-8.51	-7.22	-8.59	-7.17
N-methylmorpholine	-6.24	-5.68	-6.51	-5.49	-6.32	-6.34
mean unsigned error	0.35	0.51	0.75	0.31	0.71	
mean signed error	0.01	-0.08	-0.75	0.26	-0.70	
root mean square error	0.40	0.54	0.89	0.49	0.91	0.36[b]

Ammonia, hydrazine

ammonia	-4.61	-6.50	-6.16	-4.78	-3.00	-4.29
hydrazine	-13.11	-8.08	c	-8.62	c	-9.30
mean unsigned error	2.06	1.71	1.87[c]	0.59	1.29[c]	
mean signed error	-2.06	-0.50	-1.87[c]	0.09	1.29[c]	
root mean square error	2.70	1.78	1.87[c]	0.59	1.29[c]	2.51[b]

Subtotal for compounds containing N

mean unsigned error	0.95	0.59	0.72[c]	0.53	0.62[c]	
mean signed error	0.38	0.10	0.17[c]	0.19	0.14[c]	

root mean square error	1.24	0.83	1.03[c]	0.76	0.84[c]	1.85[b]

Compounds containing P

phosphine	0.55	0.59	–	–	–	0.60
$OP(OCH_3)_3$	-7.25	-8.58	–	–	–	-8.70
$OP(OC_2H_5)_3$	-6.79	-7.23	–	–	–	-7.80
$OP(OC_3H_7)_3$	-6.05	-7.00	–	–	–	-6.10
mean unsigned error	0.64	0.40	–	–	–	
mean signed error	0.62	-0.05	–	–	–	
root mean square error	0.88	0.53	–	–	–	3.64[b]

Compounds containing S, but no halogen
Thiols

methanethiol	-0.91	-1.23	-1.29	-1.28	-1.22	-1.24
ethanethiol	-0.64	-0.67	-0.77	-0.94	-0.96	-1.30
1-propanethiol	-0.45	-0.53	-0.56	-0.58	-0.63	-1.05
thiophenol	-3.31	-3.75	-3.43	-3.21	-3.02	-2.55
mean unsigned error	0.59	0.59	0.49	0.38	0.31	
mean signed error	0.21	-0.01	0.02	0.03	0.08	
root mean square error	0.61	0.73	0.57	0.44	0.36	0.59[b]

Organic sulfides, H_2S

dimethyl sulfide	-1.66	-1.60	-1.79	-1.80	-1.67	-1.54
diethyl sulfide	-0.64	-0.46	-0.77	-1.06	-1.19	-1.43
dipropyl sulfide	-0.39	-0.26	-0.37	-0.36	-0.56	-1.27
hydrogen sulfide	-1.17	-0.84	-0.87	-0.82	-0.88	-0.70
thioanisole	-4.32	-3.80	-3.44	-3.18	-3.27	-2.73
mean unsigned error	0.77	0.65	0.54	0.42	0.36	
mean signed error	-0.10	0.14	0.09	0.09	0.02	
root mean square error	0.91	0.79	0.61	0.50	0.42	0.66[b]

Organic disulfides

dimethyl disulfide	-2.13	-2.19	-2.07	-2.13	-1.84	-1.83
diethyl disulfide	-1.34	-1.26	-1.37	-1.22	-1.61	-1.63
mean unsigned error	0.30	0.37	0.25	0.35	0.02	
mean signed error	-0.01	0.00	0.01	0.05	0.00	
root mean square error	0.30	0.37	0.25	0.36	0.02	0.10[b]

Subtotal for compounds containing S, H, and/or C

mean unsigned error	0.62	0.58	0.47	0.39	0.28	
mean signed error	0.03	0.06	0.05	0.06	0.04	
root mean square error	0.73	0.71	0.54	0.45	0.36	0.58[b]

Compounds containing halogens
Fluorinated hydrocarbons

fluoromethane	-0.26	-0.02	1.11	0.50	0.64	-0.22
1,1-difluoroethane	-0.55	-0.91	0.27	-0.28	-0.23	-0.11
fluorobenzene	-0.41	-0.72	-0.73	-0.59	-1.06	-0.78
mean unsigned error	0.29	0.35	0.59	0.36	0.42	
mean signed error	-0.04	-0.18	0.59	0.25	0.15	
root mean square error	0.34	0.48	0.80	0.44	0.53	0.29[b]

Chloroalkanes

chloromethane	-0.66	-0.30	0.18	-0.42	-0.41	-0.56
dichloromethane	-0.98	-0.99	-0.79	-1.57	-1.51	-1.36
trichloromethane	-0.80	-1.09	-1.25	-1.39	-1.32	-1.07
chloroethane	-0.33	-0.26	0.09	-0.62	-0.65	-0.63
1,1,1-trichloroethane	-0.77	-1.00	-0.74	-1.21	-0.83	-0.25
1,1,2-trichloroethane	-0.88	-1.02	-1.07	-1.71	-1.63	-1.95
1-chloropropane	0.11	0.04	0.40	-0.22	-0.26	-0.27
2-chloropropane	-0.12	-0.12	0.34	-0.37	-0.41	-0.25
mean unsigned error	0.39	0.39	0.61	0.26	0.20	
mean signed error	0.24	0.20	0.44	-0.15	-0.08	
root mean square error	0.49	0.49	0.64	0.38	0.27	0.58[b]

Chloroalkenes

chloroethene	-0.23	0.32	0.77	0.67	0.53	-0.59
3-chloropropene	-0.19	0.12	0.55	0.25	-0.02	-0.57
Z-1,2-dichloroethene	-0.80	-0.53	-0.08	-0.65	-0.74	-1.17
E-1,2-dichloroethene	-0.53	-0.15	-0.20	0.21	0.05	-0.76
trichloroethene	-0.79	-0.71	-0.97	-0.29	-0.32	-0.39
mean unsigned error	0.35	0.63	0.94	0.73	0.60	
mean signed error	0.19	0.50	0.71	0.73	0.60	
root mean square error	0.35	0.66	0.99	0.83	0.69	0.26[b]

Chloroarenes

chlorobenzene	-1.27	-1.33	-1.59	-1.24	-1.36	-1.12
o-dichlorobenzene	-1.46	-1.59	-2.16	-1.59	-1.68	-1.36
p-dichlorobenzene	-1.10	-1.09	-2.07	-0.95	-1.20	-1.01
mean unsigned error	0.11	0.17	0.77	0.14	0.25	
mean signed error	-0.11	-0.17	-0.77	-0.10	-0.25	
root mean square error	0.12	0.19	0.81	0.16	0.26	0.15[b]

Brominated hydrocarbons

bromomethane	-0.64	-0.41	-0.78	-0.98	-0.72	-0.82
dibromomethane	-1.63	-1.76	-1.78	-2.39	-1.95	-2.11
tribromomethane	-2.35	-2.65	-2.14	-2.30	-2.18	-1.98
bromoethane	-0.28	-0.28	-1.02	-0.99	-0.99	-0.70
1-bromopropane	0.11	0.03	-0.52	-0.52	-0.51	-0.56

2-bromopropane	0.01	0.03	-0.88	-0.54	-0.68	-0.48
1-bromobutane	0.40	0.21	-0.32	-0.32	-0.33	-0.41
1-bromopentane	0.68	0.38	-0.12	-0.15	-0.14	-0.08
bromobenzene	-2.15	-2.15	-1.72	-1.96	-1.99	-1.46
p-dibromobenzene	-2.97	-2.86	-2.31	-2.23	-2.43	-2.30
mean unsigned error	0.55	0.53	0.17	0.19	0.18	
mean signed error	0.21	0.14	-0.07	-0.15	-0.10	
root mean square error	0.58	0.54	0.22	0.24	0.23	0.76[b]

Iodinated hydrocarbons

diiodomethane	-2.51	-2.81	-2.17	-2.34	-2.45	-2.49
iodomethane	-0.54	-0.48	-0.94	-1.14	-0.73	-0.89
iodoethane	-0.32	-0.42	-1.04	-1.00	-0.84	-0.72
1-iodopropane	-0.16	-0.32	-0.56	-0.49	-0.48	-0.59
2-iodopropane	-0.06	-0.07	-0.76	-0.46	-0.47	-0.46
1-iodobutane	0.04	-0.22	-0.34	-0.28	-0.29	-0.25
1-iodopentane	0.29	-0.07	-0.14	-0.10	-0.10	-0.12
iodobenzene	-3.15	-3.07	-1.77	-1.59	-1.96	-1.73
mean unsigned error	0.46	0.39	0.15	0.12	0.09	
mean signed error	0.11	-0.03	-0.06	-0.02	-0.01	
root mean square error	0.60	0.55	0.20	0.15	0.11	0.76[b]

Other halo molecules

bromotrifluoromethane	0.42	0.26	1.36	1.70	1.40	1.79
chlorofluoromethane	-1.09	-0.88	-0.11	-1.16	-0.80	-0.77
chlorodifluoromethane	-0.72	-0.66	-0.09	-1.00	-0.47	-0.50
tetrafluoromethane	2.68	2.67	3.07	3.26	2.85	3.11
1-bromo-1-chloro-2,2,2-trifluoroethane	0.16	0.08	-0.27	0.19	-0.14	-0.13
1-bromo-2-chloroethane	-1.22	-1.25	-1.42	-1.77	-1.72	-1.95
1-bromo-1,2,2,2-tetrafluoroethane	0.32	0.15	0.37	0.41	0.26	0.52
tetrachloroethene	-0.91	-1.09	-1.72	0.03	0.01	0.05
1-chloro-2,2,2-trifluoroethane	0.02	0.04	0.21	-0.12	-0.20	0.06
1,1,2-trichloro-1,2,2-trifluoroethane	0.84	0.73	-0.20	1.86	1.73	1.77
2,2,2-trifluoroethanol	-5.26	-5.05	-5.03	-6.17	-5.69	-4.31
1-chloro-2,2,2-trifluoroethyl difluoromethyl ether	-0.26	-0.18	-0.71	-1.14	-0.42	0.11
1,1,1-trifluoropropan-2-ol	-3.95	-4.00	-4.00	-3.83	-3.76	-4.16
1,1,1,3,3,3-hexafluoropropan-2-ol	-1.42	-1.12	-2.59	-2.04	-2.12	-3.77
bis(2-chloroethyl)sulfide	-3.00	-2.96	-2.29	-2.89	-2.88	-3.92
2,2,2-trifluorethyl vinyl ether	-0.12	-0.16	0.12	0.26	0.24	-0.12
p-bromophenol	-6.71	-6.52	-6.45	-6.75	-6.70	-7.13
mean unsigned error	0.63	0.66	0.69	0.53	0.43	
mean signed error	-0.05	-0.04	-0.02	0.01	0.06	
root mean square error	0.85	0.93	0.91	0.77	0.64	2.60[b]

Subtotal for compounds containing halogens

mean unsigned error	0.48	0.51	0.52	0.35	0.31	
mean signed error	0.08	0.07	0.09	0.03	0.03	
root mean square error	0.63	0.68	0.70	0.55	0.46	1.56[b]

Entire Parameterization Set

mumber of solutes	219	219	214[c,d]	215[d]	214[c,d]	
mean unsigned error	0.56	0.49	0.51	0.46	0.44	
mean signed error	0.06	0.02	0.06	0.04	0.03	
root mean square error	0.80	0.70	0.73	0.67	0.62	2.89[b]

[a]See text, Section 2 of paper, for the sources of the experimental data.

[b]Dispersion of the experimental datafor the set of solutes in the group, i. e., root-mean-square deviation of the set from its mean.

[c]Hydrazine was not calculated using either of the SM5.4PD methods (see text).

[d]These models were not parameterized for phosphorus, thus those 4 compounds were dropped when calculating mean errors.