THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

pubs.acs.org/JPCB

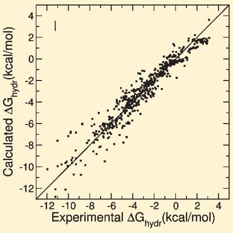
Hydration Free Energies Using Semiempirical Quantum Mechanical Hamiltonians and a Continuum Solvent Model with Multiple Atomic-Type Parameters

Victor M. Anisimov and Claudio N. Cavasotto*

School of Biomedical Informatics, University of Texas Health Science Center at Houston, 7000 Fannin Street, Houston, Texas 77030, United States

S Supporting Information

ABSTRACT: To build the foundation for accurate quantum mechanical (QM) simulation of biomacromolecules in an aqueous environment, we undertook the optimization of the COnductor-like Screening MOdel (COSMO) atomic radii and atomic surface tension coefficients for different semiempirical Hamiltonians adhering to the same computational conditions recently followed in the simulation of biomolecular systems. This optimization was achieved by reproducing experimental hydration free energies of a set consisting of 507 neutral and 99 ionic molecules. The calculated hydration free energies were significantly improved by introducing a multiple atomic-type scheme that reflects different chemical environments. The nonpolar contribution was treated according to the scaled particle Claverie—Pierotti formalism. Separate radii and surface tension coefficient sets have been developed for AM1, PM3, PM5, and RM1 semiempirical Hamiltonians, with an average unsigned error for neutral molecules of 0.64, 0.66, 0.73, and 0.71 kcal/mol, respectively. Free energy calculation of each molecule took on average 0.5 s on a single processor. The new sets of parameters will enhance the quality of semiempirical QM calculations using COSMO in biomolecular systems. Overall, these results further extend the utility of QM methods to chemical and biological systems in the condensed phase.



$$\Delta G_{hydr} = \Delta G_{elec} + \Delta G_{cav} + \Delta G_{disp} + \Delta G_{conc}$$

$$\Delta G_{elec} = \left\langle \hat{\mathrm{H}}_{\mathrm{o}} + \frac{1}{2}\sum_{i,e}\frac{q_i}{|\vec{r}_e - \vec{r}_i|} \right\rangle + \frac{1}{2}\sum_{i,a}\frac{q_i Z_a}{|\vec{R}_a - \vec{r}_i|} - \left\langle \hat{\mathrm{H}}_o \right\rangle_{\mathrm{o}}$$

1. INTRODUCTION

The accurate calculation of binding free energy in biomolecular systems is of the utmost importance in studying ligand—protein interaction and in computer-aided drug design. Since chemical processes occur in an aqueous environment, the theoretical and computational modeling of such processes has to account for solvent effects to match experimental conditions. A sound approach to model these effects in free energy calculations is to explicitly include a large number of water molecules around the solute, such as in free energy perturbation (FEP),[1] thermodynamic integration (TI),[2] and potential of mean force (PMF) approaches.[3,4] However, these simulations bring a significant computational overhead due to the increased number of degrees of freedom to account for and usually omit mutual electronic polarization since they are based on fixed-charge molecular mechanics (MM) calculations. Computationally more affordable and still quantitatively accurate are continuum solvent models, where the solvent is represented by a polarizable dielectric continuum, and its response to the presence of the solute is accounted for through the reaction field. This type of solvent representation has been of tremendous value in modeling molecular processes occurring in the condensed phase.[5−8]

End-point methods represent an appealing approach to calculate binding free energy. They account for adequate conformational flexibility and, since they use only the initial and final states of the system, are computationally more efficient than perturbative approaches. For example, the MM-Poisson—Boltzmann Surface Area (MM/PB-SA),[9,10] and the MM-Generalized Born Surface Area (MM/GB-SA)[11,12] methods rescore molecular dynamics (MD) trajectories using a continuum solvent model, PB or GB, respectively.

Recently, we introduced the MM/QM-COSMO method,[13] a semiempirical quantum mechanical (QM)-based end-point approach for calculating binding free energy in large biomolecular systems using the COnductor-like Screening MOdel (COSMO) continuum solvent model.[14] The MM/QM-COSMO correctly predicted the binding free energy of a series of phosphopeptides to the SH2 domain of human LCK, in contrast to the strong overestimation exhibited by the MM/PB-SA and MM/GB-SA methods. The use of QM also allowed both solvent and solute to exhibit mutual electronic polarization,[6] while the semiempirical level was dictated by the size of the biomolecular system. In fact, the progress in computer hardware and QM methodology has already made feasible several large-scale applications of QM methods to study protein—ligand interactions.[13,15−21]

Received: April 26, 2011
Revised: May 18, 2011
Published: May 18, 2011

Continuum solvent models used with semiempirical methods include PB,[22] SMx,[23,24] the Polarizable Continuum Model (PCM),[6] the Miertus−Scrocco−Tommasi (MST),[25,26] and COSMO. COSMO is particularly appealing due to its simplicity, robustness, and the ease with which it can be integrated in the QM framework (for a thorough overview of QM continuum solvent models, the reader is referred to the review by J. Tomasi et al.[6]). However, the accuracy of QM calculations in extended sytems using continuum solvent models depends on the availability of optimized atomic radii and surface tension coefficients. Due to the lack of hydration free energy data for macromolecules, the best approach is to optimize those parameters on a set of small molecules for which experimental data are available. The use of a large and diverse set of training molecules would guarantee parameter transferability from small to macromolecules. Thus, in the present work, we aim to derive optimized sets of COSMO radii and surface tension coefficients for different semiempirical Hamiltonians, with the long-term goal of using them for calculations in biomolecular systems. It has been remarked[27] that the optimal performance of continuum solvent models is achieved when they are used under the same computational conditions used in their parametrization. Thus, in this work, parameters were optimized to reproduce hydration free energies of a large set of neutral and charged molecules, following the same computational conditions recently used in macromolecular systems[13,21] concerning the level of theory, temperature, number of tesserae, solvent probe radius, and the use of a switching function for smooth analytical gradients, among other factors. This will clearly enhance the quality of future semiempirical QM calculations using COSMO in biomolecular systems.

As it pertains to all parametric methods, the quality of the model strongly depends on the amount and diversity of the training data set used in the parameter optimization. While large-scale training and extensive validation of implicit solvent models on hundreds of small molecules are common with classical force fields (for example, cfr. refs 28−31), fewer comparable efforts have been seen pertaining to semiempirical methods. Li et al. parametrized the SM5.42R solvent model using 275 neutral solutes for AM1, PM3, ab initio, and DFT methods.[32] However, it should be remarked that the latest member of the SM*x* family of solvent models, the SM8, was developed for nonempirical methods only.[33] Extending the use of the MST method to the RM1 Hamiltonian,[34] Forti et al.[35] used 81 neutral and 97 ionic molecules in their training set. They used a single set of atomic radii per chemical element in electrostatic calculations but scaling them in ionic molecules by a factor depending on the functional group and the type of Hamiltonian. In essence, this approach is equivalent to the use of multiple atom types. They found the scaling factor to be Hamiltonian-dependent, which supports the logical notion that different sets of atomic radii are necessary for different Hamiltonians.

Prompted by our long-term goal of routine semiempirical QM calculations with the COSMO continuum solvent model in macromolecular systems, we present in this work the derivation of optimal COSMO parameters (atomic radii and surface tension coefficients) for AM1,[36] PM3,[37] PM5,[38] and RM1[34] Hamiltonians using a set of 606 small molecules (507 neutral and 99 ionic), adhering to the computational conditions used in biomolecular simulations.[13,21] As a primary goal, we developed a fast and accurate protocol to calculate small-molecule hydration free energies. Parameters were optimized to reproduce experimental hydration free energies, using multiple atomic types to account for different chemical environments. The nonpolar contribution to the solvation free energy was accounted for via the scaled particle Claverie−Pierotti formalism.

The Methods section includes a survey of continuum solvent model theory, details of the QM calculation, and parametrization procedure, with special emphasis on cavity construction, the issue of solvent probe determination, and the use of a switching function. In the Results section, parameter optimization and hydration free energy calculations using single atomic-type and multiple atomic-type parameters are presented, with a detailed discussion about the justification for the multiple atomic-type approach and the performance of both approaches on different chemical classes; cross-validation and comparison with other models follow. A summary of the results and the implications of this work are presented in the Conclusions.

2. METHODS

2.1. Theory. The hydration free energy ΔG_{hydr} of transferring a molecule from the standard state in gas to the standard state in water can be calculated as[26]

$$\Delta G_{\text{hydr}} = \Delta G_{\text{elec}} + \Delta G_{\text{cav}} + \Delta G_{\text{disp}} + \Delta G_{\text{conc}} \qquad (1)$$

where the electrostatic term (ΔG_{elec}) represents the change in electrostatic energy upon solvation, including the mutual electronic polarization of the solute and solvent; the cavity term (ΔG_{cav}) represents the work necessary to create a cavity in bulk solvent; the dispersion term (ΔG_{disp}) represents the attractive solute−solvent dispersion interaction; and ΔG_{conc} accounts for the change in concentration between the gas phase and the aqueous phase. Since 1 mol/L was used for both phases, $\Delta G_{\text{conc}} = 0$ throughout this work. The second and third terms account for the nonpolar contribution to the free energy. A single-conformation approach is implicitly assumed in eq 1, in agreement with our and others' previous work.[28,29,32,33,35,39]

To determine the electrostatic component, we assume that the solvent is represented by an infinite isotropic continuum dielectric, whose polarization in the presence of the solute generates a reaction field represented by the perturbative operator $\hat{V}_{\text{R}}$. Thus, the electronic wave function of the system is determined by solving the Schrödinger equation

$$(\hat{H}_{\text{o}} + \hat{V}_{\text{R}})|\Psi\rangle = E|\Psi\rangle \qquad (2)$$

where $\hat{H}_{\text{o}}$ is the Hamiltonian of the solute in the gas phase and $\hat{V}_{\text{R}}$ is the reaction field operator. Assuming that the cavity surface is discretized in a large number N of surface elements (tesserae), such that the induced charge (q_i) on surface element i can be considered as a constant to be determined iteratively, the reaction field perturbative operator can be expressed as

$$\hat{V}_{\text{R}} = \sum_{i=1}^{N} \frac{q_i}{|\vec{r} - \vec{r}_i|} \qquad (3)$$

Thus, the electrostatic component of the hydration free energy takes the form[40]

$$\Delta G_{\text{elec}} = \left\langle \hat{H}_{\text{o}} + \frac{1}{2}\sum_{i,e} \frac{q_i}{|\vec{r}_e - \vec{r}_i|} \right\rangle + \frac{1}{2}\sum_{i,a} \frac{q_i Z_a}{|\vec{R}_a - \vec{r}_i|} - \langle \hat{H}_{\text{o}} \rangle_{\text{o}} \qquad (4)$$

where the Z_a are the solute atomic nuclear charges; the last term corresponds to the gas-phase energy; the sums over "*e*", "*a*", and

"i" refer to the electrons, nuclei, and surface charges, respectively; and the average value of the third term is calculated using the gas-phase wave function; eq 4 accounts for the work performed to polarize the solvent, which is half of the solute−solvent interaction energy.

The surface charges were calculated using the COSMO continuum solvent model,[14] where the solvent is represented by an ideal conductor for which the total electrostatic potential due to solute and solvent cancels out on the solute boundary. Assuming that a set of solute charges $\mathbf{Q}$ induces charges $\mathbf{q}'$ on the cavity surface, the vector of potentials Φ_i on the segments can be expressed as $\mathbf{\Phi} = \mathbf{BQ} + \mathbf{Aq}'$, where $\mathbf{A}$ and $\mathbf{B}$ are the surface−surface and surface−solute blocks of the Green function,[14] which depend only on the geometry of the system and the dielectric constant distribution. The condition of $\mathbf{\Phi} = 0$ in every surface element means

$$\mathbf{q}' = -\mathbf{A}^{-1}\mathbf{BQ} \tag{5}$$

The effects of the finite dielectric constant ε are recovered by scaling $\mathbf{q}'$ by a factor f

$$\mathbf{q} = f\mathbf{q}' \tag{6}$$

where $f = (\varepsilon - 1)/(\varepsilon + 0.5)$, with the relative error of using this approach being less than $(2\varepsilon)^{-1}$.[14] Thus, COSMO is very well suited for high dielectric constant solvents like water.

Equation 4 is solved iteratively, where both the wave function Ψ and the surface charges q_i are mutually dependent. In each self-consistent cycle step, q_i charges are updated using eqs 5 and 6, where $\mathbf{Q}$ represents the nuclei and Mulliken charges. Equation 5 was solved via the linear scaling conjugate gradient minimization technique proposed by York et al.[39]

The cavitation contribution to eq 1 was calculated based on the scaled particle theory developed by Reiss et al.[41] and adapted by Pierotti.[42] In the Claverie−Pierotti formalism,[43] which extends the theory to the treatment of nonspherical solutes, the cavitation contribution is expressed as a sum of atomic cavitation energies $\Delta G_{\text{cav}}(R_i)$ scaled by the ratio of the atomic solvent exposed area A_i to the total molecular solvent accessible surface area, A_{total}

$$\Delta G_{\text{cav}} = \sum_i \frac{A_i}{A_{\text{total}}} \Delta G_{\text{cav}}(R_i) \tag{7}$$

The atomic cavitation terms were computed using an expansion in series of powers of the effective radius $R_{\text{ms}} = R_{\text{m}} + R_{\text{solv}}$ (where R_{m} and R_{solv} are atomic and solvent radii, respectively) as described by Tomasi et al.[44]

The dispersion contribution to the free energy in eq 1 was evaluated in a similar manner from individual atomic contributions[40]

$$\Delta G_{\text{disp}} = \sum_i A_i \gamma_i \tag{8}$$

where γ_i is the atomic surface tension coefficient. For other approaches to calculate the dispersion term and general discussion on the subject, the reader is referred to specialized publications.[6,40,45,46]

2.2. Cavity Construction. We used two different cavity surfaces to determine polar and nonpolar free energy contributions (dual-cavity approach), similar to the MST model.[35] The electrostatic boundary surface was constructed directly from the atomic van der Waals radii. For the nonpolar contributions, we used the solvent accessible surface area, constructed by augmenting the atomic van der Waals radii by the solvent probe radius, R_{solv}. It should be remarked that the frequently used value of 1.4 Å for the solvent probe radius is not a universal standard. In the SMx model, the solvent probe radius is optimized to 0.4 Å for the nonpolar term.[47] In the MOPAC program, R_{solv} is reduced to 1.3 Å.[48] This suggests certain benefits from optimization of the solvent probe radius. Therefore, we checked for an optimal value of R_{solv} on alkanes where the nonpolar term is the dominant part of the hydration free energy. In this work, the solvent probe radius was set to 1.1 Å. Details of this choice are presented in the Results and Discussions.

For the purpose of placing induced charges on the boundary surface, each atom in the solute molecule was modeled by a sphere discretized in equal-area triangles.[48] The induced charge was placed at the geometric center of each triangle. The number of surface elements per atom determines the degree of discretization of cavity surface and, thus, the memory requirement to store the data. When dealing with macromolecules, the required memory allocation for surface elements quickly becomes prohibitively large, thus placing an upper limit on the possible surface discretization level. For this reason, we chose 32 triangles per non-hydrogen atom and 12 per hydrogen atom, which provided the best balance between accuracy in hydration-free energy for small molecules and computational demand for calculation in large biomacromolecular systems.[13,21] Performing parameter optimization using larger segmentation would perhaps improve the accuracy of the model but will lead to a dramatic rise in memory demand and computational cost when large systems are studied, which we intend to avoid.

2.3. Scaling of Induced Charges Located along Seams of Intersecting Atomic Spheres. Since an important utility of semiempirical QM calculations with COSMO is the feasibility of geometry optimization (and eventually, molecular dynamics), there is an actual need for smooth analytical derivatives of the induced charges with respect to the atomic coordinates. As the atoms move during geometry optimization, or molecular dynamics, new surface elements may emerge, while the previously exposed surface elements may disappear, which may cause gradient discontinuity.

To make the position of induced charges to be a smooth function of the nuclear coordinates, York et al.[49] proposed to weight the atomic surface charges, q_i, by a switching function Sw_i, which is constructed as a product of atomic switching functions $\text{Sw}_{i,\text{B}}$,

$$\text{Sw}_i = \prod_{\text{B} \neq \text{A}} \text{Sw}_{i,\text{B}} \tag{9}$$

The atomic switching function used in the present work was the one given by Senn et al.[50]

$$\text{Sw}_{i,\text{B}} = 1 - \exp[-(d_{i\text{A},j\text{B}} + 0.9)^{48}] \tag{10}$$

In eqs 9 and 10, the indices i and j of surface charges are associated with atoms A and B, respectively; index B runs over atoms located in the vicinity of the atom A; and $d_{i\text{A},j\text{B}}$ is the distance between surface element i located on atom A and the nearest surface element j located on atom B. In the case $d_{i\text{A},j\text{B}} \leq 0$, when charge $q_{i\text{A}}$ gets inside the sphere of atom B, $\text{Sw}_{i,\text{B}}$ is set to zero. All surface charges are weighted by the value of their switching function, thus leading to eq 6 being rewritten in the

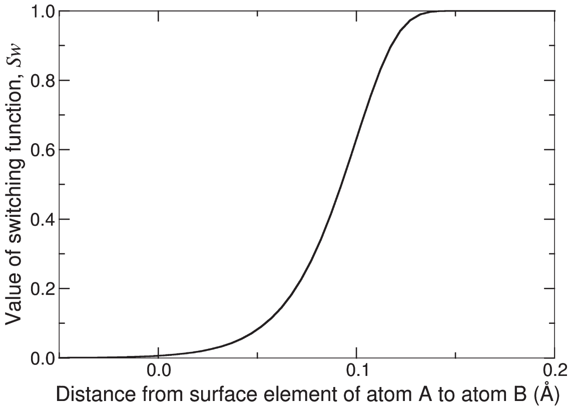


Figure 1. Profile of switching function defined by eq 10.

following form

$$q_i = \mathrm{Sw}_i \cdot f \cdot q_i' \qquad (11)$$

This effectively means that eq 11 replaces eq 6 in the section of COSMO theory. In this approach, none of the surface charges is discarded from the energy equation, so the number of surface charges (surface elements) remains constant during the entire simulation.

The numerical profile of switching function given by eq 10 is shown in Figure 1. The switching function nullifies the surface charge q_i on atom A in the energy expression if the charge enters the sphere of another atom. The surface charges located along the seams, where the spheres of different atoms intersect, are scaled with the weight in the range from 0 to 1. The surface charges away from the seams effectively retain their full unscaled values. When atoms move during geometry optimization or molecular dynamics, the new surface charges smoothly appear or disappear on the energy surface ensuring continuity of energy and gradients.

2.4. Parametrization Procedure. The optimal set of atomic radii and surface tension coefficients was determined by minimizing the squared difference between calculated and experimental hydration free energies, according to

$$F(\mathbf{x}) = \sum_i [\Delta G^i_{\mathrm{exp}} - \Delta G^i_{\mathrm{hydr}}(\mathbf{x})]^2 + \alpha \|\mathbf{x} - \mathbf{x_o}\|^2 \qquad (12)$$

where F is the error function to be minimized, $\mathbf{x} \equiv (\mathbf{R}, \gamma)$ with $\mathbf{R}$ and γ being the vectors of the radii and surface tension coefficients, respectively; ΔG_{hydr} is the calculated hydration free energy according to eq 1; and the last term corresponds to a mild parabolic restraint ($\alpha = 0.1$) introduced to improve stability, where $\mathbf{x_o}$ are the initial values of $\mathbf{x}$; the sum is performed over all available experimental data ΔG_{exp}. Using a large α would reduce convergence speed since more iterations will be needed to reach the minimum. On the other hand, having no restraint or a too small one might allow nonphysical values of the parameters. Thus, the chosen value of α provided a good balance. It should be remarked, as explained below, that the optimization was performed in cycles, using the results of the last optimization as an initial guess for the following. Thus, the impact of the restraint is essentially negligible in the last cycles.

In this work, we considered a limited number of chemical environment-dependent atom types,[35,39] which improved the accuracy of the model. Since the choice of specific multiple atom types is influenced by parameter correlation, different physical combinations of types were tried and only the best performing ones reported. In the end, the decision on the number of introduced atom types was guided by the sense to keep their number to a minimum since an excessive number of parameters can lead to overtraining.

The optimization procedure was designed as follows. The initial guess for atomic radii was taken from Bondi radii.[51] The starting values of atomic surface tension coefficients were set to zero. To prevent the uncertainties in experimental hydration free energies of ionic molecules from negatively affecting the optimization of atomic parameters for neutral ones, for which accurate experimental hydration free energies are available, for each Hamiltonian, the optimization was performed in three steps. In the first step, atomic radii and surface tension coefficients were optimized for neutral molecules only. In the second step, the parameters corresponding to the neutral molecules were held fixed, and the parameters unique to ions were optimized. On the third and final step, the optimization was performed on the entire set of neutral and ionic molecules to minimize the impact of the hierarchical procedure (steps one and two). In the last step, the experimental data for neutral molecules carried a weight of 1.0, whereas the experimental data for ionic molecules were assigned a weight of 0.1. In the first two steps dealing with separate optimization of neutral and ionic molecules, a uniform weight of 1.0 was applied. Each step was repeated three times before going to the next one, using the optimized parameters of one round as an initial guess for the next round, thus to minimize the influence of restraints on the quality of the fit. The iterations were terminated when convergence tolerance of 0.0001 was reached for the minimized least-squares function. The entire optimization procedure was independently performed for AM1, PM3, PM5, and RM1 Hamiltonians.

2.5. Levenberg−Marquardt Minimization Algorithm. Since $\Delta G_{\mathrm{hydr}}(\mathbf{x})$ in eq 12 depends nonlinearly on $\mathbf{x} \equiv (\mathbf{R}, \gamma)$, we performed the minimization of F using the Levenberg−Marquardt algorithm.[52] The solution is based on a Taylor expansion of the hydration free energy as

$$\Delta G_{\mathrm{hydr}}(\mathbf{x} + \boldsymbol{\delta}) \approx \Delta G_{\mathrm{hydr}}(\mathbf{x}) + \nabla[\Delta G_{\mathrm{hydr}}(\mathbf{x})] \cdot \boldsymbol{\delta} \qquad (13)$$

where $\boldsymbol{\delta}$ is the vector of variable increments, $\boldsymbol{\delta} = \mathbf{x}^{\mathrm{new}} - \mathbf{x}^{\mathrm{old}}$. The convenience of eq 13 is that it is linear in respect to unknown variables $\boldsymbol{\delta}$. The first derivatives of ΔG_{hydr} with respect to $\mathbf{x}$ define the Jacobian matrix, $\mathbf{J}$, as

$$\mathrm{J}_{ij} = \frac{\partial[\Delta G^i_{\mathrm{hydr}}(\mathbf{x})]}{\partial x_j} \qquad (14)$$

Substituting eqs 13 and 14 into eq 12, one obtains an expression, which is linear in $\boldsymbol{\delta}$

$$F(\mathbf{x} + \boldsymbol{\delta}) \approx \|\Delta G_{\mathrm{exp}} - \Delta G_{\mathrm{hydr}}(\mathbf{x}) - \mathbf{J}\boldsymbol{\delta}\|^2 + \alpha\|\mathbf{x} + \boldsymbol{\delta} - \mathbf{x_o}\|^2 \qquad (15)$$

Differentiating eq 15 with respect to $\boldsymbol{\delta}$, setting the result to zero, and adding a dumping term $\lambda \cdot \mathrm{diag}(\mathbf{J}^{+}\mathbf{J})$ leads to the final form of the Levenberg−Marquardt equation, namely

$$\begin{aligned}[\mathbf{J}^{+}\mathbf{J} + \alpha\mathbf{I} + \lambda \cdot \mathrm{diag}(\mathbf{J}^{+}\mathbf{J})] \cdot \boldsymbol{\delta} \\ = \mathbf{J}^{+}[\Delta G_{\mathrm{exp}} - \Delta G_{\mathrm{hydr}}] - \alpha(\mathbf{x} - \mathbf{x_o})\end{aligned} \qquad (16)$$

where $\mathbf{I}$ is the identity matrix. The iterative solution of eq 16 with respect to $\boldsymbol{\delta}$ leads to the determination of the optimal $\mathbf{x}$. In this equation, λ is an adjustable, dimensionless, and positive-value dumping factor controlling the stability of convergence of the optimization procedure. Choosing a smaller value for λ causes the minimization procedure to make a larger leap in the direction of the search. If the iteration was unsuccessful, its result was discarded and the minimization attempted from the previous successful step with increased λ by a factor of 10. After a successful iteration, its value was reduced by a factor of 10. Energy gradients with respect to atomic radii were computed numerically. The gradients with respect to atomic surface tension coefficients were determined analytically from eq 8. The starting value of the dumping factor λ in the Levenberg−Marquardt minimization procedure was set to 0.5.

2.6. Training Data Set. Our training set consisted of 606 small molecules (507 neutral and 99 ionic) for which experimental data are available. Experimental and calculated hydration free energies are given for 298 K and correspond to the transfer at 1 mol/L gas to 1 mol/L in solvent. Experimental data for molecules with codes 0001−0752 were obtained from the compilation of Mobley et al.[53] (molecule codes are explained in Table S1 of the Supporting Information), codes 0760−0761 from Vorobyov et al.,[54] and codes 0762−0950 from the compilation of Kelly et al.[47] For ionic molecules (codes 0801−0950), the hydration free energies were corrected by a factor of 1.9 kcal/mol as suggested by Forti et al.[35] to account for the most accurate experimental hydration free energy of the proton. The error in the experimental data of hydration free energy for neutral molecules is in the range from ±0.10 to ±1.93 kcal/mol as estimated by Guthrie on a set of biologically relevant compounds.[55] The error in experimental hydration free energies of ionic molecules is ±3 kcal/mol as estimated by Kelly et al.[47] In the set obtained from Mobley et al., butanone turned out to be butyraldehyde, so the molecular structure of the former was corrected. To make the analysis tractable we grouped the molecules in 38 chemical classes. The complete list of training set molecules along with their calculated and experimental hydration free energies at 298 K is provided in the Supporting Information (Table S1). The geometry of each molecule was gas-phase optimized using the corresponding semiempirical Hamiltonian, assuming that the difference in hydration free energy using a gas-phase geometry and an aqueous-phase geometry is less than the mean error of the model.[47]

2.7. Semiempirical Quantum Mechanical Calculations. Semiempirical quantum mechanical calculations were performed using LocalSCF[56] using its capability to read atomic radii and surface tension coefficients from an external file. Empirical corrections for nitrogen planarity in the peptide group (keyword MMOK) and for geometry of five- and six-membered rings (keyword MMCCROK)[48] were used. Tight SCF convergence criterion was enforced by the keyword "PRECISE". The water dielectric constant was set to 78 in COSMO calculations. Solvent (background) charges were placed on the analytical cavity surface constructed as the union of atomic van der Waals radii according to the algorithm described by Senn et al.[50] assigning 32 surface elements to non-hydrogen atoms and 12 surface elements to hydrogen atoms.

3. RESULTS AND DISCUSSION

In this work, we optimized the atomic radii and surface tension parameters for the continuum solvent model COSMO within the semiempirical QM framework and tested the model on its ability to accurately reproduce hydration free energy of 606 small molecules (507 neutral and 99 ionic molecules).

Table 1. Average Signed Error between Calculated and Experimental Hydration Free Energy ($\Delta G_{\mathrm{hydr}} - \Delta G_{\mathrm{exp}}$) Using Optimized Single Atomic-Type Radii and Surface Tension Coefficients[a]

chemical class	AM1	PM3	PM5	RM1
alkanes	−0.96	−1.69	−0.23	−0.94
alkenes	−0.75	−1.32	−0.06	−0.94
alkynes	−0.60	−1.25	0.13	−0.93
aromatics	−0.55	−0.81	−0.17	−0.51
alcohols	2.57	3.15	2.48	2.71
ethers	−0.14	0.23	−1.30	−0.02
aldehydes	0.31	0.69	−0.17	0.18
ketones	0.40	0.52	−0.55	0.35
carboxylic acids	2.03	2.98	2.19	2.03
esters	−1.45	−0.92	−1.68	−1.39
amines	0.90	1.71	0.82	1.26
heterocycles	−0.20	−0.10	−0.66	−0.84
amides	2.62	3.47	0.70	2.57
nitriles	−0.66	−2.03	−0.32	−1.24
nitrates	−1.69	−4.03	−0.34	−0.79
thiols	−0.53	0.14	−0.63	0.40
thioethers	0.28	−0.15	0.13	0.34
sulfoxides	−6.74	−7.50	−6.32	−10.65
phosphates	−7.55	−5.88	−0.06	−6.85
fluoro-derivatives	0.53	0.58	0.23	0.36
chloro-derivatives	0.22	0.28	−0.05	−0.03
bromo-derivatives	−0.06	−0.19	−0.17	−0.24
iodo-derivatives	0.00	−0.05	−0.01	−0.06
F−Cl−Br derivatives	−0.04	0.00	−0.12	−0.16
deprotonated carbon	−4.95	−5.09	−3.47	−2.35
deprotonated amines	−2.06	−2.39	−4.17	−1.78
deprotonated carbonyl	5.47	5.04	1.42	5.49
deprotonated carboxyl	2.39	1.68	1.82	3.01
deprotonated alcohols	11.69	9.54	6.31	11.89
deprotonated phenol	8.39	7.84	4.03	8.61
deprotonated thiols	−3.02	−2.79	−5.44	−3.72
protonated heterocycles	4.49	1.89	5.63	4.32
protonated amines	0.45	−1.03	2.69	1.64
protonated amides	5.78	4.78	7.61	6.12
protonated carbonyl	11.90	11.75	14.12	11.86
protonated O=X	2.58	3.49	4.85	3.01
protonated ether	10.70	8.93	13.84	10.69
protonated alcohol	14.88	15.02	18.67	16.08
AUE				
neutral	1.22	1.57	1.11	1.39
anions	6.36	5.62	4.39	6.56
cations	3.61	3.76	5.42	4.08
rmsd				
neutral	1.80	2.19	1.63	2.04
anions	7.66	6.89	5.41	7.84
cations	5.52	5.55	7.16	5.83

[a] Values in kcal/mol.

3.1. Solvent Probe Radius Determination. As described above, the solvent probe radius was used in this work for calculation of the nonpolar term only. We found in the present work that using the common value of 1.4 Å for the solvent probe radius leads to incorrect hydration free energy trend in linear alkanes. Energies should become less favorable as the number of carbons increases. However, COSMO calculations showed exactly the opposite unphysical trend for all considered Hamiltonians, namely, AM1, PM3, PM5, and RM1. Since no optimization of atomic radii and surface tension coefficients was able to restore the right trend, we performed several parameter optimization runs trying different values of solvent probe radius, noting that the trend for hydration free energy of alkanes strongly improves upon reducing the solvent probe radius. Thus, we settled on the solvent probe radius of 1.1 Å which is relatively close to the conventional value of 1.4 Å. Further reduction of the radius below 1.1 Å provided additional improvement in the free energy trend for alkanes, but we considered it necessary to restrain the solvent probe radius to a physically justified value. Throughout this work, the optimized solvent probe radius of 1.1 Å was used in all calculations of the nonpolar hydration free energy term.

3.2. Hydration Free Energy Calculation Using Single Atomic-Type Parameters. The results of hydration free energy calculation for the whole training set using single-type atomic radii are presented in Table 1, averaged for 38 different chemical classes of molecules. Detailed data for individual molecules are provided in the Supporting Information (Table S1). The average calculation time per molecule was 0.5 s.

Considering the 507 neutral molecules for which accurate experimental hydration free energies are available, the smallest average unsigned error (AUE) of 1.11 kcal/mol was obtained for the PM5 Hamiltonian. This result was closely followed by AM1 and RM1 Hamiltonians showing an AUE of 1.22 and 1.39 kcal/mol, respectively. The largest AUE of 1.57 kcal/mol was exhibited by PM3. At the level of individual classes of molecules, all Hamiltonians failed on sulfoxides with the signed error in the range from −6.32 to −10.65 kcal/mol, indicating that the computations predict a too favorable hydration free energy. The second most troublesome class was phosphates. Except for PM5, having the error near to zero, the signed error for other Hamiltonians ranged from −5.88 to −7.55 kcal/mol. Other problematic classes were amides, carboxylic acids, and alcohols with the errors in the range from 2.03 to 3.47 kcal/mol, with PM5 again showing the smallest error. PM3 revealed significant problems with nitrates as characterized by the signed error of −4.03 kcal/mol.

Ionic molecules showed the largest errors, which is not surprising considering the large absolute values of hydration free energy of these molecules. All Hamiltonians showed similar trends. For anionic molecules, the smallest AUE of 4.39 kcal/mol was obtained with PM5 and the largest AUE of 6.56 kcal/mol with RM1. For cationic molecules, the smallest AUE of 3.61 kcal/mol was obtained with AM1, whereas the largest AUE of 5.42 kcal/mol was with PM5. Among the individual classes, the worst agreement with experimental data was observed for protonated alcohols with the signed error being in the range from 14.88 to 18.67 kcal/mol. Deprotonated alcohols also turned to be problematic, having the signed error in the range from 6.31 to 11.89 kcal/mol. Other highly problematic ionic classes were protonated carbonyls (ketones and aldehydes) and protonated ethers. Although increasing the weight of ionic molecules in the parameter optimization procedure could reduce these errors, it was avoided in view of the limited accuracy of the experimental hydration free energies of ionic molecules. The reduction in error between computed and experimental values for ionic molecules would likely be achieved at the cost of worsening the overall agreement for neutral molecules, which is undesirable.

Table 2. Optimized Single Atomic-Type Radii, *R*, and Atomic Surface Tension Coefficients, γ[a]

	AM1	PM3	PM5	RM1
R(H)	1.2414	1.2685	1.2686	1.2592
R(C)	2.0065	2.0204	2.0023	2.0064
R(N)	1.8124	1.8668	1.8205	1.8222
R(O)	1.7699	1.7417	1.7665	1.7756
R(F)	1.7212	1.7217	1.7239	1.7223
R(P)	2.1674	2.1282	2.1088	2.1347
R(S)	2.1859	2.1803	2.1735	2.1884
R(Cl)	2.0495	2.0500	2.0514	2.0546
R(Br)	2.1601	2.1606	2.1608	2.1614
R(I)	2.3203	2.3204	2.3204	2.3204
γ(H)	0.0085	0.0065	0.0079	0.0131
γ(C)	−0.0122	−0.0201	−0.0012	−0.0235
γ(N)	−0.0337	−0.0370	−0.0092	−0.0177
γ(O)	0.0075	0.0267	0.0251	0.0059
γ(F)	0.0205	0.0183	0.0348	0.0249
γ(P)	0.0170	−0.0032	−0.0050	−0.0028
γ(S)	−0.0055	0.0068	−0.0049	−0.0054
γ(Cl)	−0.0035	−0.0067	−0.0011	0.0069
γ(Br)	−0.0069	−0.0037	−0.0043	0.0019
γ(I)	−0.0109	−0.0074	−0.0033	−0.0067

[a] Values of R in Å and γ in kcal/mol Å^2

The optimized values of single-type atomic parameters are provided in Table 2. The differences in values of atomic radii for different Hamiltonians are relatively small but sufficient to make an impact of several kilocalories/mole. Moreover, there is no physical reason to impose a single set of atomic radii to all Hamiltonians since the latter are unrelated. It is seen that H, N, and O atomic radii are the most Hamiltonian dependent.

These optimized single-type atomic radii and surface tension parameters are presumably the most transferable parameters when considering their application to untested molecules. The average error of ∼1 kcal/mol for neutral molecules is perhaps the best that can be achieved within this scheme, which is very respectable. However, there are persisting difficulties with phosphates, sulfoxides, some nitrogen derivatives, and ionic molecules. These errors are indicative of present limitations in the electronic part of the semiempirical Hamiltonians. It is well-known that accuracy in electronic calculations improves upon adding d-orbitals to phosphorus and sulfur as it is implemented in the PM6 Hamiltonian.[57] The Hamiltonians considered in this work use a minimum sp basis set, which is a contributing factor to the observed large discrepancies with experimental data. To address the present needs, we introduce chemical environment-depending atomic parameters, which result in the assignment of multiple chemical types to certain atoms.

3.3. Hydration Free Energy Calculation Using Multiple Atomic-Type Parameters. In an attempt to reduce the error of single atomic-type parameters, we introduced multiple atomic types for some chemical elements. Two separate types were

Table 3. Average Signed Error between Calculated and Experimental Hydration Free Energy ($\Delta G_{hydr} - \Delta G_{exp}$) Using Optimized Multiple Atomic-Type Radii and Surface Tension Coefficients[a]

chemical class	AM1	PM3	PM5	RM1
alkanes	−0.43	−0.69	0.02	−0.33
alkenes	−0.28	−0.49	0.26	−0.34
alkynes	−0.11	−0.48	0.59	−0.36
aromatics	0.15	0.20	0.44	0.28
alcohols	0.07	0.08	−0.05	0.14
ethers	0.00	0.10	−0.12	0.19
aldehydes	−0.89	−0.88	−0.34	−0.83
ketones	−0.06	−0.04	−0.22	0.10
carboxylic acids	0.65	0.18	1.06	0.48
esters	0.01	0.01	−0.17	0.00
amines	0.18	0.39	−0.06	0.03
heterocycles	0.33	0.27	0.21	0.28
amides	2.07	2.05	0.75	1.59
nitriles	0.00	−0.02	−0.04	−0.02
nitrates	0.41	0.04	0.22	0.87
thiols	−1.10	−0.44	−1.11	−0.66
thioethers	0.07	−0.07	0.04	0.08
sulfoxides	−0.34	−0.02	−1.73	−2.37
phosphates	−1.90	0.00	1.72	−1.29
fluoro-derivatives	0.35	0.30	0.12	0.17
chloro-derivatives	0.05	0.07	−0.21	−0.24
bromo-derivatives	0.01	−0.09	−0.14	−0.16
iodo-derivatives	−0.02	−0.04	−0.01	−0.05
F−Cl−Br derivatives	−0.04	0.08	−0.11	−0.15
deprotonated carbon	−4.62	−4.35	−2.83	−1.31
deprotonated amines	−0.02	0.00	−0.82	1.33
deprotonated carbonyl	5.21	3.90	0.31	5.27
deprotonated carboxyl	2.21	0.19	2.42	2.69
deprotonated alcohols	−0.37	−0.73	−0.33	−0.08
deprotonated phenol	0.91	1.57	−0.40	0.89
deprotonated thiols	−0.98	−0.25	−2.02	−0.77
protonated heterocycles	4.39	2.03	3.11	3.11
protonated amines	0.74	0.36	1.08	0.86
protonated amides	1.63	0.60	0.20	−0.19
protonated carbonyl	5.24	6.36	5.89	5.18
protonated O═X	1.34	2.94	3.22	2.98
protonated ether	7.65	4.31	6.69	4.19
protonated alcohol	−0.97	1.33	2.74	−1.05
AUE				
neutral	0.64	0.66	0.73	0.71
anions	2.42	2.26	2.48	2.32
cations	2.78	2.36	2.97	2.33
rmsd				
neutral	0.87	0.89	1.02	0.97
anions	2.95	2.73	3.10	2.87
cations	3.40	2.92	3.55	2.78

[a] Values in kcal/mol.

Table 4. Optimized Multiple Atomic-Type Radii, *R*, and Atomic Surface Tension Coefficients γ[a]

type	AM1	PM3	PM5	RM1
R(Hpol)	1.1132	1.1053	1.0410	1.0483
R(Hnon)	1.3810	1.3761	1.3228	1.3770
R(C)	1.9240	1.9288	1.9183	1.9246
R(Ngen)	1.8307	1.8472	1.8490	1.8447
R(NH)	1.5865	1.6050	1.6280	1.6209
R(N═O)	2.1133	2.1000	2.1061	2.1423
R(N≡C)	2.0863	2.0865	2.0853	2.0857
R(Npr)	1.7675	1.8669	1.8167	1.8546
R(Ogen)	1.6923	1.6698	1.6895	1.6924
R(OCC)	1.6317	1.6065	1.7325	1.6390
R(O═C)	1.7493	1.7551	1.7594	1.7419
R(O═X)	1.9251	1.9021	1.9173	1.9695
R(OH)	1.5634	1.5725	1.6110	1.5645
R(OHa)	1.5312	1.5761	1.6729	1.5581
R(Opr)	1.2891	1.4016	1.3700	1.3998
R(F)	1.7235	1.7216	1.7267	1.7230
R(P)	2.4107	2.3978	2.3984	2.4172
R(S)	2.2472	2.2504	2.2606	2.2782
R(Cl)	2.0308	2.0360	2.0313	2.0413
R(Br)	2.1592	2.1594	2.1601	2.1606
R(I)	2.3168	2.3165	2.3134	2.3151
γ(Hpol)	0.0632	0.0437	0.0895	0.0652
γ(Hnon)	0.0046	0.0039	0.0058	0.0059
γ(C)	−0.0115	−0.0196	0.0071	−0.0229
γ(Ngen)	−0.0414	−0.0401	−0.0102	−0.0098
γ(NH)	−0.0826	−0.1258	−0.0198	−0.0554
γ(N═O)	−0.1233	−0.0966	−0.1059	−0.1675
γ(N≡C)	−0.0295	−0.0162	−0.0198	−0.0168
γ(Npr)	−0.0379	−0.0409	−0.0750	−0.0486
γ(Ogen)	0.0371	0.0419	0.0659	0.0353
γ(OCC)	0.0277	0.0054	0.0763	0.0289
γ(O═C)	−0.0217	−0.0196	0.0215	−0.0162
γ(O═X)	0.0272	0.0494	0.0270	0.0265
γ(OH)	−0.0685	−0.0685	−0.0212	−0.0480
γ(OHa)	−0.0261	−0.0358	−0.0156	−0.0555
γ(Opr)	−0.0766	−0.0751	−0.0828	−0.0690
γ(F)	0.0218	0.0216	0.0358	0.0269
γ(P)	0.0082	−0.0119	−0.0129	−0.0086
γ(S)	−0.0215	−0.0115	−0.0169	−0.0242
γ(Cl)	−0.0051	−0.0079	−0.0019	0.0053
γ(Br)	−0.0092	−0.0068	−0.0051	−0.0003
γ(I)	−0.0120	−0.0093	−0.0023	−0.0073

[a] Values of *R* in Å and γ in kcal/mol Å^2

assigned to the hydrogen atom. The polar hydrogen (Hpol) was defined as a hydrogen atom connected to a noncarbon atom. The nonpolar hydrogen (Hnon) was the one connected to the carbon atom. The initially attempted introduction of sp^1, sp^2, and sp^3 carbon types did not lead to any sensible improvement in the quality of the fit; therefore, carbon was retained at a single type. Five distinct types were introduced for the nitrogen atom. These are amine nitrogen (NH) having at least one hydrogen atom connected to it, nitrogen in nitro-group (N═O), nitrogen in cyano-group (N≡C), protonated nitrogen (Npr), and generic nitrogen (Ngen), which covers all other nitrogen types not found within the specific first four groups. The introduction of a specific

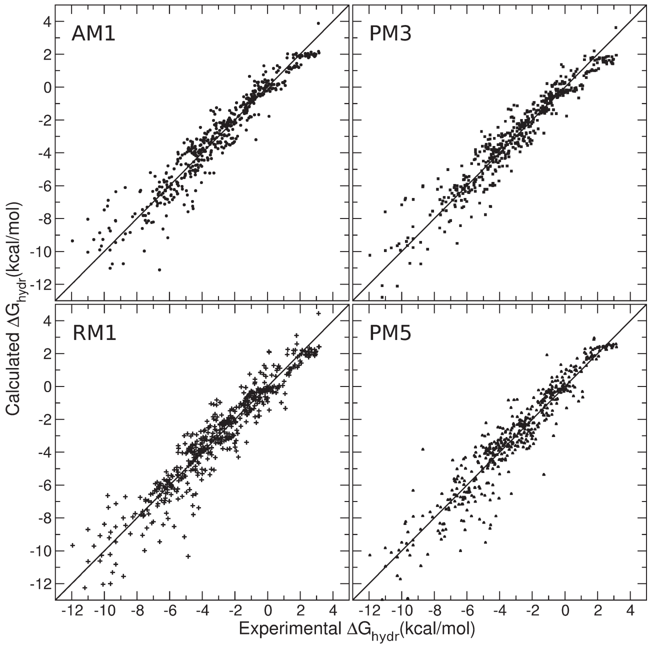


Figure 2. Calculated vs experimental hydration free energies for the set of 507 neutral molecules. The correlation coefficient was 0.96 for AM1 and PM3 and 0.95 for PM5 and RM1.

type for nitrogen in nitro- and cyano-groups was necessary to correct for a systematic difference between the calculated and experimental hydration free energies (see Table 1). These systematic differences persisted in all studied Hamiltonians, thus additionally supporting our decision to introduce the specific nitrogen types for nitro- and cyano-groups. Seven specific types were introduced for the oxygen atom. These are hydroxyl oxygen (OH), oxygen in ethers (OCC), carbonyl oxygen for aldehydes and ketones (O═C), sp^2 oxygen connected to a noncarbon atom (O═X), deprotonated hydroxyl oxygen (OHa), protonated oxygen (Opr), and generic oxygen (Ogen). All other chemical elements (C, F, P, S, Cl, Br, I) were treated at single type. Overall, 21 unique atom types were considered, three of which occur in ionic molecules only. Classification of the training set based on this atom typing scheme is provided in the Supporting Information (Table S2). Hydration free energies computed using optimized multiple atomic-type parameters are presented in Table 3.

Following the introduction of multiple atomic types, a significantly better agreement between the computed and experimental hydrations free energies was obtained for all classes of molecules (Table 3), relative to the case utilizing single atomic types (Table 1). Indeed, the AUE dropped from 1.22, 1.57, 1.11, and 1.39 kcal/mol for AM1, PM3, PM5, and RM1 Hamiltonians, respectively (Table 1), to 0.64, 0.66, 0.73, and 0.71 kcal/mol (Table 3). The improvement concerns alkanes, alcohols, carboxylic acids, and esters, among others, where previously the degree of agreement between the computed and experimental hydration free energies was unremarkable. Introduction of multiple atomic types solved the notorious problem with sulfoxides and phosphates. A noticeable improvement was obtained on ionic molecules where the error dropped roughly by a factor of 3 over the single atomic-type case. Also the difference for anions and cations became closer. The AUE for anionic molecules dropped to 2.42, 2.26, 2.48, and 2.32 kcal/mol for AM1, PM3, PM5, and RM1 Hamiltonians, respectively, from the corresponding larger values reported in Table 1. Cationic molecules were characterized with an AUE of 2.78, 2.36, 2.97, and 2.33 kcal/mol for the same list of Hamiltonians. The signed error for deprotonated alcohols dropped from 12 kcal/mol as reported in the case of single-type radii (Table 1) to less than 1 kcal/mol due to multiple atomic-type parameters (Table 3). The signed error for protonated alcohols dropped from 19 kcal/mol to less than 3 kcal/mol. Less but also significant improvement was observed with protonated ethers, where the signed error went down from 14 to 8 kcal/mol. Overall, these improvements justify the transition from 10 single atomic types (Table 2) to 21 multiple atomic-type parameters (Table 4). This remarkable agreement between the computed and hydration free energies, which is obtained on a very large set of 606 molecules, confirms the validity of the parametrization approach.

In Figure 2 we present the plots of calculated vs experimental hydration free energy for AM1, PM3, PM5, and RM1 Hamiltonians. Dense arrangement of the points on the plots along the diagonal line suggests good agreement between the computed and experimental data. Linear fit of data showed slopes of 0.95, 0.93, 1.02, and 0.97 and intercepts of −0.12, −0.20, 0.06, and 0.10 for AM1, PM3, PM5, and RM1 Hamiltonians, respectively. The slopes and intercepts are very close to the value of one and zero, respectively, thus demonstrating good correlation of computed results with experimental data. The correlation coefficients were 0.96 for AM1 and PM3 and 0.95 for PM5 and RM1.

In the free energy range from −4 to +4 kcal/mol, all considered Hamiltonians showed similar high accuracy. The errors increased for larger hydration free energies (lower left corner of the plots). In this higher free energy area, the PM3 Hamiltonian shows somewhat better accuracy over more recent PM5 and RM1 Hamiltonians.

Our attempts to use mutliple atomic types for phosphorus and sulfur were not successful. This is partially related to the small number of corresponding molecules in the training set. A complicating factor is the buried state of phosphorus and sulfur atoms, typically shielded from direct interaction with the solvent. In this light, it is not surprising that the introduction of the O═X type for the solvent-exposed oxygen atom perfomed well for sulfoxides, phosphates, and nitrates. Since this type worked well for a diverse set of molecules involving nitrogen, phosphorus, and sulfur, this may suggest that the O═X type is chemically justified.

The use of multiple atomic-type parameters for the continuum solvent model COSMO corrects the limitations of AM1, PM3, RM1, and PM5 Hamiltonians to describe the electronic structures of some molecules in the training set. For those atoms where the Hamiltonian underestimates the magnitude of the partial atomic charge, a smaller COSMO atomic radius effectively compensates for the reduced solute−solvent electrostatic interaction. An analogous argument applies when the Hamiltonian overestimates the partial electronic atomic charge. The success of this approach on continuum solvent models may suggest the next step to integrate the multiple atomic-type approach in the development of semiempirical Hamiltonians.

3.4. Validation of Multiple Atomic-Type Parameters. The final set of COSMO parameters listed in Table 4 was derived including all 606 molecules from the training set. We assessed the quality of our optimization by a 5-fold cross-validation. The training set was divided in five nearly equal-size groups by approximately assigning equal number of molecules from each chemical class to each of five groups. Such division guarantees an

Table 5. Average Unsigned Error for the 5-Fold Cross-Validation[a]

	AM1		PM3		PM5		RM1	
molecules	opt[b]	test[c]	opt[b]	test[c]	opt[b]	test[c]	opt[b]	test[c]
all	0.96	0.97	0.93	0.94	1.05	1.07	0.98	0.98
neutral	0.64	0.65	0.66	0.67	0.73	0.74	0.71	0.72
anions	2.42	2.41	2.26	2.29	2.48	2.47	2.32	2.31
cations	2.78	2.80	2.36	2.38	2.97	3.00	2.33	2.35

[a] Values in kcal/mol. [b] AUE corresponding to parameter optimization using the complete training set (cfr. Table 3). [c] Average AUE over the 5-fold cross-validation rounds using a ratio 1/4 between the test and the training set.

equal presence of each class in all the five subsets. Chemical classes consisting of fewer than five molecules could not be equally distributed in five groups for obvious reasons. Thus, some classes may be missing in some of the subsets. However, there are only a few such classes, so their unequal division is not going to significantly impact the equal distribution of classes over the five subsets. Each subset served as a test set once and four times was part of the training set. These calculations were separately repeated for AM1, PM3, PM5, and RM1 Hamiltonians. In each round, a full-parameter optimization was performed using the parameters from Table 4 as an initial guess.

The averaged AUE over the five rounds is presented in Table 5 (the values per round, Hamiltonian, and molecule group are displayed in Table S3 of the Supporting Information). It is observed that the values from the cross-validation are very close to those derived using the complete training set (cfr. Table 3). All the considered Hamiltonians in Table 5 exhibit comparable accuracy with an average unsigned error of 1 kcal/mol for the entire set of 606 small molecules representing 507 neutral, 48 negatively, and 51 positively charged molecules. Neutral molecules are characterized by an AUE of about 0.7 kcal/mol. Anionic molecules show an AUE of 2.41, 2.29, 2.47, and 2.31 kcal/mol for AM1, PM3, PM5, and RM1, respectively. Positively charged molecules exhibit an AUE of 2.80, 2.38, 3.00, and 2.35 kcal/mol, respectively.

Reported in Table S4 (Supporting Information) are the root-mean-square fluctuations of the optimized parameters as a result of the 5-fold cross-validation procedure. It can be readily seen that the majority of the optimized parameters are converged up to the fourth digit after the decimal point. Interestingly, the largest fluctuations are observed for polar and nonpolar hydrogen radii followed by a spike on the generic oxygen radius. These are likely the hot spots on the parameter space which point to the direction of potential improvement of the model and which can be addressed in a follow up study.

3.5. Comparison with Other Models.

The accuracy of our developed implicit solvent model can be compared to that of other models developed in similar studies. Among these is the work of Forti et al.[35] reporting the errors of 0.6 kcal/mol for neutral molecules, 4.3 kcal/mol for cations, and 4.4 kcal/mol for anions on a training set of 81 neutral, 47 cationic, and 51 anionic molecules. Our results are also comparable to the SM8 solvation model of Cramer and Trular,[33] which is based on the mPW1PW/6-31G(d) level of theory and characterized by an AUE of 0.55 kcal/mol for 274 neutral molecules, using single atomic types. They obtained an AUE of 2.7 and 3.7 kcal/mol for 52 cations and 60 anions, respectively. It should be mentioned that the SM8 can also predict solvation free energies in a variety of solvents (cf. also ref 58). Errors in this range are also found using other QM continuum solvent models.[27]

Among similar studies using a MM approach, we can quote the work of Bordner et al.[29] who optimized implicit solvent atomic parameters for the MMFF94 force field on 410 neutral small molecules, reporting a rmsd of 0.72 kcal/mol. Shivakumar et al.[30] reported an AUE of 1.10 kcal/mol for 239 neutral small molecules using MD FEP and the OPLS 2005 force field. Rizzo et al.[28] reported an AUE of 0.99 kcal/mol for 460 neutral molecules and an AUE of 4.46 kcal/mol for 42 charged molecules when using ChelpG charges in the PBSA solvation model and the general Amber force field. Dill and co-workers recently reported hydration free energy calculations using the semiexplicit assembly, with a rmsd of ∼1.3 kcal/mol on a set of 504 neutral molecules.[31]

This short overview confirms that our developed implicit solvent model is basically in line with the results obtained by other groups in terms of accuracy and also computing time. The advantage of our developed model is presumably its better transferability vs the implicit solvent models based on classical force fields due to accounting for solute polarization.

4. CONCLUSIONS

Toward the long-term goal of routinely using semiempirical QM calculations with a continuum solvent representation in macromolecular systems, we undertook the optimization of atomic radii and surface tension coefficients for the continuum solvent model COSMO integrated with semiempirical AM1, PM3, PM5, and RM1 Hamiltonians adhering to the same computational conditions recently used in biomolecular simulations.[13,21] The optimization showed the need to introduce a multiple atomic-type approach, using two specific types for the hydrogen atom (polar and nonpolar), five types for the nitrogen atom (amine, nitro-group, cyano-group, protonated nitrogen, and generic nitrogen), and seven types for the oxygen atom (hydroxyl, ether, carbonyl, noncarbon connected sp^2, deprotonated hydroxyl, protonated oxygen, and generic oxygen), whereas C, F, P, S, Cl, Br, and I chemical elements were reasonably well represented with single types.

Although the introduction of multiple atomic types might look to be deviating from the standard practice of QM methods to keep single type for chemical elements, it was fully justified by the significant improvement in the agreement with experimental data when compared to the single atomic-type approach. Besides, the number of introduced types is significantly smaller than that used in classical force fields. This is a favorable aspect of comparison of our model, which explicitly accounts for solute polarizability.

Confirmed by cross-validation, the developed implicit solvent model is characterized by an average unsigned error of ∼1 kcal/mol in reproducing hydration free energy for the entire set of 606 small molecules including 507 neutral and 99 ionic molecules for which experimental data are available. For the set of neutral molecules, the average unsigned error is 0.64, 0.66, 0.73, and 0.71 kcal/mol for AM1, PM3, PM5, and RM1 Hamiltonians, respectively. The AUE for anionic molecules is 2.42, 2.26, 2.48, and 2.32 kcal/mol, and the AUE for cationic molecules is 2.78, 2.36, 2.97, and 2.33 kcal/mol for AM1, PM3, PM5, and RM1 Hamiltonians, respectively.

The obtained level of accuracy of the implicit model COSMO at semiempirical quantum mechanical level of theory is comparable to other implicit solvent models at similar or higher QM levels of theory. This enables a consistent application of a quantum mechanical level of theory from small molecules to large biological macromolecules, which expands the range of chemical and biological applications in the condensed phase.

■ ASSOCIATED CONTENT

S Supporting Information. Calculated and experimental hydration free energies at 298 K for the complete molecule set (606 molecules); classification of the training set based on the atom typing scheme; results of the 5-fold cross-validation per round, Hamiltonian and molecule type; and root-mean-square fluctuations of the optimized parameters within the 5-fold cross-validation procedure. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: claudio.n.cavasotto@uth.tmc.edu.

■ ACKNOWLEDGMENT

The authors are grateful to the Texas Advanced Computing Center (TACC) (http://www.tacc.utexas.edu) for providing high-performance computing resources for this project. This work was partially supported by the R. A. Welch Foundation Chemistry and Biology Collaborative Grant from the John S. Dunn Gulf Coast Consortium for Chemical Genomics.

■ REFERENCES

(1) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431.

(2) Straatsma, T. P.; McCammon, J. A. *Methods Enzymol.* **1991**, *202*, 497.

(3) Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72*, 1568.

(4) Woo, H. J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825.

(5) Fogolari, F.; Brigo, A.; Molinari, H. *J. Mol. Recognit.* **2002**, *15*, 377.

(6) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(7) Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187.

(8) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.

(9) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401.

(10) Vorobjev, Y. N.; Hermans, J. *Biophys. Chem.* **1999**, *78*, 195.

(11) Feig, M.; Brooks, C. L., 3rd *Curr. Opin. Struct. Biol.* **2004**, *14*, 217.

(12) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005.

(13) Anisimov, V. M.; Cavasotto, C. N. *J. Comput. Chem.* **2011**, *32*, 2254.

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

(15) Fanfrlik, J.; Bronowska, A. K.; Rezac, J.; Prenosil, O.; Konvalinka, J.; Hobza, P. *J. Phys. Chem. B* **2010**, *114*, 12666.

(16) Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. *J. Mol. Biol.* **2003**, *327*, 549.

(17) Gräter, F.; Schwarzl, S. M.; Dejaegere, A.; Fischer, S.; Smith, J. C. *J. Phys. Chem. B* **2005**, *109*, 10474.

(18) Raha, K.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2004**, *126*, 1020.

(19) Soderhjelm, P.; Aquilante, F.; Ryde, U. *J. Phys. Chem. B* **2009**, *113*, 11085.

(20) Illingworth, C. J. R.; Morris, G. M.; Parkes, K. E. B.; Snell, C. R.; Reynolds, C. A. *J. Phys. Chem. A* **2008**, *112*, 12157.

(21) Anisimov, V. M.; Bugaenko, V. L.; Cavasotto, C. N. *ChemPhysChem* **2009**, *10*, 3194.

(22) Gogonea, V.; Merz, K. M. *J. Phys. Chem. A* **1999**, *103*, 5171.

(23) Winget, P.; Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 5160.

(24) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 4538.

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

(26) Luque, F. J.; Barril, X.; Orozco, M. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 139.

(27) Klamt, A.; Mennucci, B.; Tomasi, J.; Barone, V.; Curutchet, C.; Orozco, M.; Luque, F. J. *Acc. Chem. Res.* **2009**, *42*, 489.

(28) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J. Chem. Theory Comput.* **2005**, *2*, 128.

(29) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A. *J. Phys. Chem. B* **2002**, *106*, 11009.

(30) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. *J. Chem. Theory Comput.* **2010**, *6*, 1509.

(31) Fennell, C. J.; Kehoe, C. W.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 3234.

(32) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1999**, *103*, 9.

(33) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011.

(34) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.

(35) Forti, F.; Barril, X.; Luque, F. J.; Orozco, M. *J. Comput. Chem.* **2008**, *29*, 578.

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

(37) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

(38) Stewart, J. J. P. *MOPAC 2002*; Fujitsu Ltd: Tokyo, Japan, 2002.

(39) York, D. M.; Tai-Sung, L.; Weitao, Y. *Chem. Phys. Lett.* **1996**, *263*, 297.

(40) Luque, F. J.; Curutchet, C.; Muñoz-Muriedas, J.; Bidon-Chanal, A.; Soteras, I.; Morreale, A.; Gelpi, J. L.; Orozco, M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3827.

(41) Reiss, H.; Frisch, H. L.; Helfand, E.; Lebowitz, J. L. *J. Chem. Phys.* **1960**, *32*, 119.

(42) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717.

(43) Langlet, J.; Claverie, P.; Caillet, J.; Pullman, A. *J. Phys. Chem.* **1988**, *92*, 1617.

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

(45) Kar, P.; Seel, M.; Hansmann, U. H. E.; Hoefinger, S. *J. Phys. Chem. B* **2007**, *111*, 8910.

(46) Colominas, C.; Luque, F. J.; Teixidó, J.; Orozco, M. *Chem. Phys.* **1999**, *240*, 253.

(47) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

(48) Stewart, J. J. P. *MOPAC2000*, 1.0 ed.; Fujitsu Limited: Tokyo, Japan, 2000.

(49) York, D. M.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 11060.

(50) Senn, H. M.; Margl, P. M.; Schmid, R.; Ziegler, T.; Blochl, P. E. *J. Chem. Phys.* **2003**, *118*, 1089.

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

(52) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: New York, 1992; Vol. 1.

(53) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938.

(54) Vorobyov, I.; Li, L.; Allen, T. W. *J. Phys. Chem. B* **2008**, *112*, 9588.

(55) Guthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501.

(56) Bugaenko, V. L.; Bobrikov, V. V.; Andreyev, A. M.; Anikin, N. A.; Anisimov, V. M. *LocalSCF*, 2.1 ed.; Fujitsu Ltd.: Tokyo, Japan, 2005.

(57) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.

(58) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 760.