Supporting Information for: Introducing Charge Hydration Asymmetry into the Generalized Born Model

Abhishek Mukhopadhyay,[†] Boris H Aguilar,[‡] Igor S Tolokh,[‡] and Alexey V Onufriev[*,‡,†]

Department of Physics, Virginia Tech, Blacksburg, VA 24061, USA, and Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

E-mail: alexey@cs.vt.edu

1 Molecule Sets

The four molecule sets used in this work are detailed as follows.

N- and P- bracelets. A set of artificial neutral molecules or "bracelets" (name, structure and parameters adopted from Ref. 1) with planar and regular polygonal geometry were constructed with the aromatic carbon ("ca" atom type in the generalized Amber force field (GAFF)[2]) with the Lennard Jones parameters, $\sigma = 3.39967$Å and $\varepsilon = 0.086$ kcal/mol. Two adjacent beads/atoms were connected by a bond of length 1.4 Å. We used the distributed charge scheme from Ref. 1, *e.g.*, for a n-beaded bracelet one bead was charged $+1$ (P-bracelet) or -1 (N-bracelet) whereas the other $n-1$ beads were assigned equal charge such that overall molecule was neutral. Six charge

[*]To whom correspondence should be addressed
[†]Department of Physics, Virginia Tech, Blacksburg, VA 24061, USA
[‡]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

inverted clones (pairs of P- and N-bracelets) were constructed with n ranging from 3 (triangle) to 8 (octagon). We performed standard thermodynamic integration (TI) calculations (detailed in the preceding section) in explicit TIP3P and TIP4P-Ew water to obtain the ΔG_{pol}. For TIP5P-E the polar solvation energy, ΔG_{pol} were obtained from the total solvation energy, ΔG_{solv}, as provided in Ref. 1. The non-polar part of the solvation energy is approximated as $\Delta G_{np} = \gamma SASA$ (the effective surface energy coefficient $\gamma = 5$ cal/mol/Å^2, *SASA* is the solvent accessible surface area obtained using the MSMS package,[3] for simplicity the dispersive van-der Waals term was ignored in these calculations) was subtracted from ΔG_{solv} values. The water model charge asymmetry parameters R^z_{OH} used for CHA-GB are provided in the Main Text for all three water models used in this work.

Neutral Small Molecules. A set of 504 neutral small molecules to study hydration models were compiled previously by Mobley *et. al.*[4] In the original work, these molecules were prepared using the GAFF[2] small molecule parameters as assigned by Antechamber. Merck-Frosst implementation of AM1-BCC[5,6] was used to assign the partial charges. The explicit (TIP3P) solvation free energies were computed using the Bennett acceptance ratio[7] (BAR) in standard TIP3P water without employing any restraint to avoid the conformational variability. To minimize possible uncertainties due to inadequate conformational sampling of flexible molecules, here we restrict ourselves to a smaller subset of 248 rigid molecules as discussed in the main text. Note that the time trajectories for explicit (TIP3P) simulation were not provided in the original work, Ref. 4, and were hence obtained from implicit molecular dynamics simulations.[8] These implicit simulations were performed on the same 504 small molecule set using a GB implementation (igb=5) of AMBER,[9] without the surface area term, which was, however, added by re-weighting, see Ref. 8 for details. Using these time trajectories we computed the root mean square deviation (RMSD) of atomic positions of these molecules in 10 ns time trajectories. 248 rigid molecules were chosen for which the RMSD's were below 0.3 Åwith respect to their initial conformations, see Figure 1. In order to compare CHA-GB with GB and 3D-RISM[10] we have used the rigid molecule subset, however the full set of 504 molecules was used to compare with the SEA model[11] and the experimental solvation free

energies.[4]

Amino Acid Analogs This set is comprised of 48 structures. The coordinates, atomic partial charges, and the explicit solvent (TIP3P) solvation free energies in TIP3P of 40 structures were obtained from Ref. 12 – two conformations for each of 20 nonzwiterionic single residue amino acid side chain dipeptides of the form N-acetyl-X-N-methylamide, where X refers to one of the twenty standard amino acids. Only the charged states of the titratable amino acids, ASP, LYS, GLU, and ARG were considered in Ref. 12. We therefore added 8 additional structures corresponding to the neutral states of these four amino acids. The same coordinates as that of the corresponding charged structures were used. The atomic partial charges of the neutral ASP, GLU and LYS were obtained from AMBER force field parameters, whereas partial charges of neutral ARG were obtained from Ref. 13. The polar part of the solvation free energies of these 8 additional structures were computed using standard TI in explicit (TIP3P) water, discussed later.

Protein set 19 small proteins were randomly selected from a larger data set of representative proteins structures from Feig et al.[14] with PDB IDs 1az6, 1bh4, 1bku, 1brv, 1byy, 1cmr, 1dfs, 1dmc, 1eds, 1fct, 1fmh, 1fwo, 1g26, 1ha9, 1hzn, 1paa, 1qfd, 1qk7, and 1scy. Chain "A" or "model 1" (as referred to in the original work) has been chosen when appropriate. We used the H++ server[15] to assign partial charges and the protonation states of ionizable amino acids. Using specific values of pH in H++ we transformed the structures such that overall molecule was neutral. The random selection resulted in a fairly representative sampling of various structural classes. The structural composition of the proteins is as follows: 6 mostly α helical, 4 mostly β sheet, 4 roughly equal mix of α/β, and 5 mostly disordered. The size of most of these proteins is about 30 amino acids.

2 Simulation Protocol

Standard thermodynamic integration(TI) protocol for neutral molecules adopted from Ref.[16] was used to obtain the explicit solvent (TIP3P, TIP4P-Ew) solvation free energies. Amber 12[17] simu-

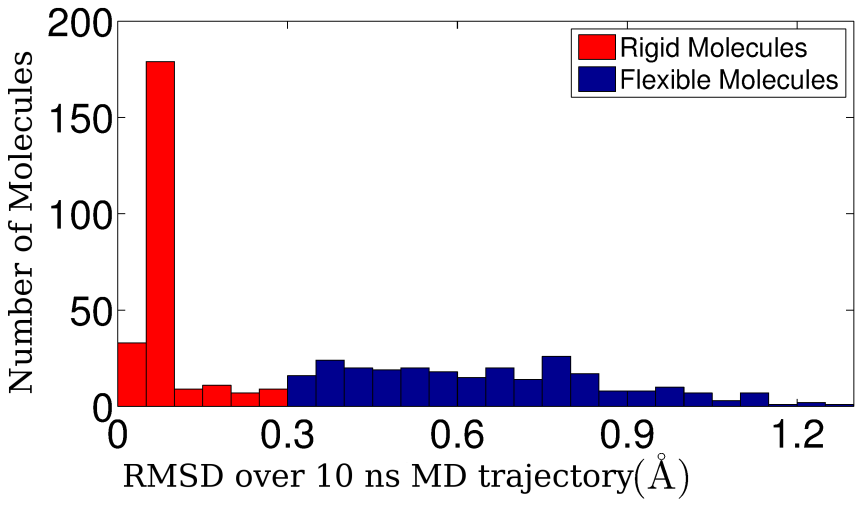


Figure 1: The distribution of *rmsd* of conformational change during 10 ns molecular dynamics simulation in TIP3P using initial conformation as reference. The trajectories were obtained from Ref. 4. Red bars correspond to the selected 248 rigid set of small molecule, *rmsd* < 0.3Å

lation package was employed. The details of the TIP5P-E solvation free energies are provided in the section on molecule sets, above. The polar contribution was computed as the difference of the charging energy of the molecular cavity in the aqueous phase and the gas phase.[18] The TI integrals were approximated using a five point Gaussian weighted sum. All simulations were performed using the Langevin thermostat with a collision frequency of 2 ps^{-1} and a time step of 2 fs. Hydrogen bonds were constrained with SHAKE[19] using a geometrical tolerance of 10^{-6} Å. For the aqueous phase, the molecules were placed in a truncated octahedral box such that the minimum distance between the solute atoms and the box edge was 12 Å. The non-bonded interaction cutoff was 10.0 Å, and long-range electrostatic interactions were calculated using periodic boundary conditions *via.* the particle mesh Ewald (PME) summation.[20,21] Positional restraints of 50 kcal/mol/ Å^2 on all atoms were employed to hold the solute in the desired conformation. The system was gradually heated at constant volume for 50 ps followed by a 1 ns equilibration at constant pressure of 1 atm and pressure relaxation time of 2 ps. The last 1 ns of a 2 ns constant volume simulation was used for the free energy calculations.

3 Parameter Optimization

3.1 The Training and Test Sets

The model parameters (9 intrinsic atomic radii and τ for CHA-GB, 9 intrinsic atomic radii for GB) were optimized using a training set designed using molecules from the rigid small molecule set and the set of amino acid analogs. This training set consisted of a total of 148 molecules, specifically, 124 molecules were chosen from the rigid set and 24 molecules from the amino acid analogs. The molecules in the training set were chosen such that the atom types and the polar solvation energy of each of the two molecule classes, the small molecules and the amino acid analogs, are equally represented as that of the rest of the molecules in the respective sets. A test set was designed using the rest of the molecules from the two sets. The training set and test set are provided at the end, in Table 8 and Table 9, respectively.

3.2 Optimization protocol

We optimize the model parameters using an objective function *rmse*(rigid molecules)+*rmse*(amino acid analogs) such that the two molecule classes are equally represented during optimization. We use a heuristic nonlinear optimization technique namely the Nelder-Mead simplex algorithm, that uses initial guess values of the parameters. With termination criteria of 10^{-3} for the parameter set and convergence criteria of 10^{-4} for the objective function, several parameter sets randomly selected within a physical range were used as initial guess, Table 1.

Robustness and Validation The optimum parameter set pertains to the converged set with the lowest objective function. 10 independent optimizations led to the converged objective function of 1.58 kcal/mol for CHA-GB, with the *rmse* of the full 248 small molecules set and the full set of 48 amino acid analogs, 0.90 kcal/mol and 0.93 kcal/mol respectively. However for GB the converged objective function value was 2.82 kcal/mol with the *rmse* of the two molecule sets being 1.35 and 1.40 kcal/mol, respectively. For more refined estimate of ΔG_{pol}, we obtained our

Table 1: Initial Parameters used for the optimization of parameters for CHA-GB and GB: random numbers from a uniform distribution were drawn from the lower bound (Min) and upper bound (Max) for various parameters. The intrinsic radii are in Å

Model		**Initial parameter values used in the optimization**									
		ρ(**C**)	ρ(**H**)	ρ(**N**)	ρ(**O**)	ρ(**S**)	ρ(**F**)	ρ(**Cl**)	ρ(**Br**)	ρ(**I**)	τ
CHA-GB	Min	1.2	0.3	1.2	1.2	1.7	1.2	1.5	1.7	2.0	1
	Max	1.7	0.8	1.7	1.7	2.2	1.7	2.0	2.2	2.5	2
GB	Min	1.5	1.0	1.2	1.2	1.7	1.0	1.1	1.3	1.5	N/A
	Max	2.0	1.5	1.7	1.7	2.2	1.5	1.6	1.8	2.0	N/A

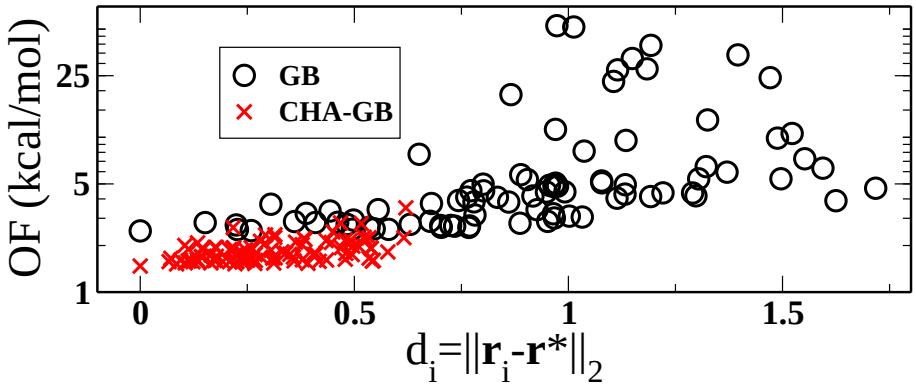


Figure 2: The converged value of the objective function (OF), *rmse*(small molecules)+*rmse*(amino acid analogs) for each of the 100 random optimization runs. The OF is plotted against the parameter distance metric $d_{i0} = ||\mathbf{r}_i - \mathbf{r}^*||_2$, where $\mathbf{r}_i$ is the converged parameter set for the ith optimization run and $\mathbf{r}^*$ is the optimum parameter set used in this work for CHA-GB (in red) and GB (in black).

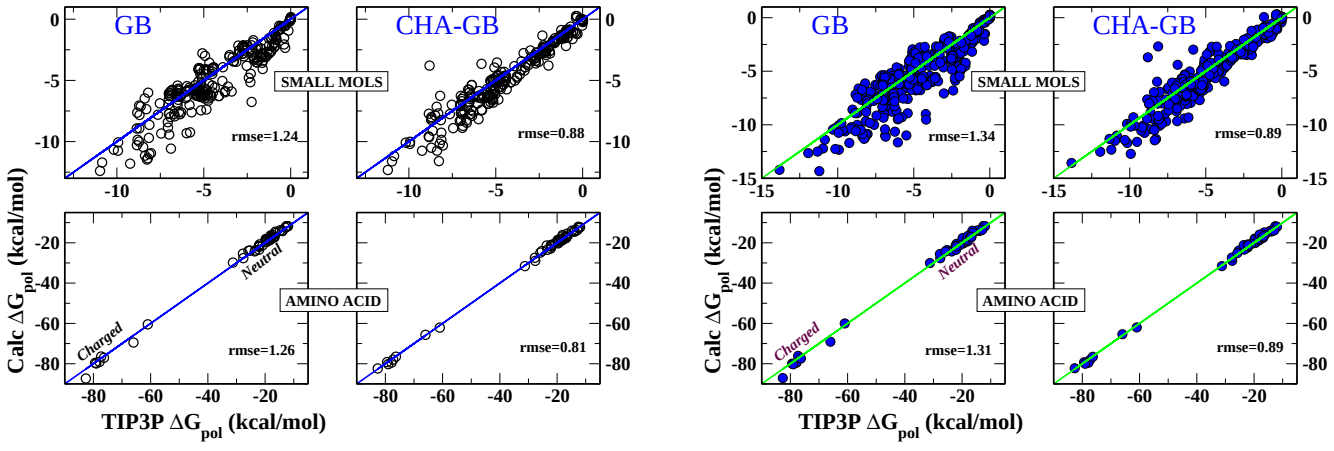
final set of parameters from 100 random optimizations. The final value of the objective function for CHA-GB was 1.47 kcal/mol whereas for GB it was 2.48. Note that the final outcome of the optimizations with 10 runs was similar to that obtained with 100 runs. Models' performance on the training set and the test set compares well, see Table 2. Comparison of the converged parameter sets of all 100 optimizations with the respective converged objective function, Figure 2 reveal that the parameter set for CHA-GB is more robust than that of the GB. The resulting parameter sets and the objective functions of these optimizations form a tight cluster *i.e.* close to the global optimum, whereas for GB the optimized parameters vary significantly from one optimization run to another. Note that the parameter sets were multi-dimensional and hence we used the distance

metric (a measure of disparity between these parameter sets) namely, $d_i = ||\mathbf{r}_i - \mathbf{r}^*||_2$, where $\mathbf{r}_i$ is a parameter set for the i^{th} optimization run and $\mathbf{r}^*$ is the optimum parameter set. In Figure 3a we compare the performance of GB and CHA-GB for all 248 small molecules and 48 amino acid analogs against the explicit (TIP3P) ΔG_{pol}. For the 248 rigid small molecules, we further analyzed the accuracy of ΔG_{pol} estimated via CHA-GB and GB for different degrees of molecular polarity, as quantified by explicit (TIP3P) ΔG_{pol}; small ($\Delta G_{pol} > -3.0$ kcal/mol), intermediate (-3.0 kcal/mol$> \Delta G_{pol} > -6.0$ kcal/mol) and large ($\Delta G_{pol} < -6.0$ kcal/mol), see Figure 4. We find that CHA-GB consistently provides a more accurate estimate over GB in each ΔG_{pol} range.

Table 2: Performance of the GB and CHA-GB in Training and Test sets

Method		Training Set		Test Set	
		Small Mols	Amino Acid Analogs	Small Mols	Amino Acid Analogs
GB	*rmse*	1.22	1.26	1.25	1.26
	$\langle error \rangle$	-0.50	0.13	-0.55	0.34
	r^2	0.85	0.997	0.87	0.998
CHA-GB	*rmse*	0.83	0.64	0.92	0.96
	$\langle error \rangle$	-0.39	-0.03	-0.36	0.22
	r^2	0.95	0.999	0.90	0.998

Improved radii transferability is due to introduced CHA. To further investigate the importance of CHA in the improvements offered by the CHA-GB model, we again performed a set of 100 optimizations for GB, but now with the new dielectric boundary definition, see Main Text, that we have so far used exclusively in CHA-GB. Namely, we used a probe of radius $\rho_w - R_s = 0.88$ Åto define the solute/solvent boundary over which the "R6" surface integration is performed to obtain the effective Born radii. The use of the new surface in GB makes the model formally equivalent to CHA-GB with the water model asymmetry "switched off", $R^z_{OH} = 0$. After the radii optimization against the same training set as before, the model yields 1.01 kcal/mol *rmse* error in ΔG_{pol} for the rigid molecule set and 1.27 kcal/mol error for the set of amino acid analogs. Recall that the corresponding CHA-GB errors are 0.88 and 0.81 kcal/mol, which means that the modified surface definition just by itself can not bring about the uniform *rmse* accuracy of better than 1 kcal/mol seen in CHA-GB.



(a) (b)

Figure 3: The polar solvation free energies, ΔG_{pol} of GB (left panel) and CHA-GB (right panel) against the reference TIP3P simulation[4] for the two molecule sets, neutral small molecules (top panels) and amino acid analogs (bottom panel) using the optimum radii sets obtained with Figure 3a just the 248 rigid molecules from the neutral small molecule set in the training set and Figure 3b the training set with flexible molecules included

Outliers Although CHA-GB shows a noticeable improvement in accuracy over the canonical GB, we find one prominent outlier in the rigid molecule set, and two in the full set of 504 molecules. Namely dimethyl-sulfate and methyl-methanesulfonate, each show about 5 kcal/mol deviation from the reference explicit ΔG_{pol}. A possible explanation could be that both molecules contain a highly charged Sulphur (S) atom (with partial atomic charge 1.6-1.8 e), which misrepresent the solvent polarization (the sign of CHA) of the atoms in the neighborhood. The CHA-scaling factor in the proposed CHA-GB model *i.e.* η, uses a simple exponential interpolation to account for the contribution of neighboring charges to determine the sign of effective solvent polarization for a particular atom. This rather simplistic approximation apparently fails to reproduce a proper sign dependence of this polarization for the immediate neighbors of these highly charged S-atoms which, in turn causes wrong CHA-contribution for their neighbors, finally leading to erroneous estimates of ΔG_{pol}.

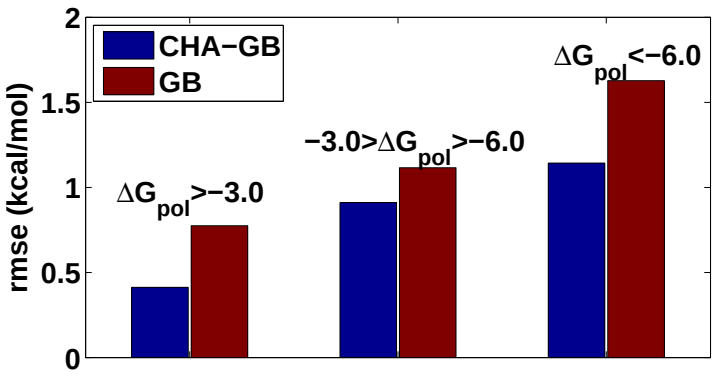


Figure 4: CHA-GB provides consistent improvement in ΔG_{pol} accuracy for molecules of different degrees of polarity. Shown is the root-mean-square error (*rmse*) of CHA-GB (blue bars) and GB (red bars) for the 248 rigid neutral small molecules against the TIP3P polar solvation energies (ΔG_{pol}). The three bars correspond to the three ranges of ΔG_{pol}; $\Delta G_{pol} > -3.0$ kcal/mol, -3.0 kcal/mol $> \Delta G_{pol} > -6.0$ kcal/mol and $\Delta G_{pol} < -6.0$ kcal/mol.

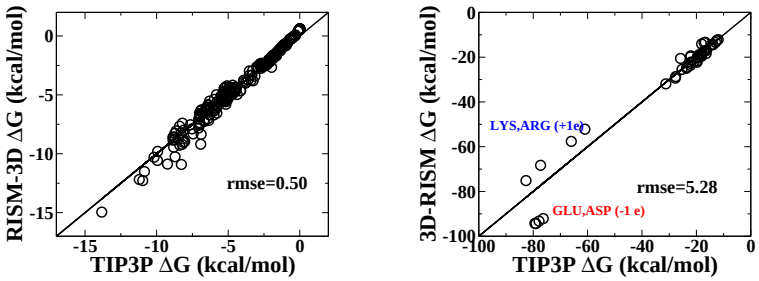
4 3D RISM: additional accuracy metrics

The single point 3D-RISM ΔG_{pol} (TIP3P) were computed using the 3D-RISM implementation[22] in AMBER[17] and corrected[10] using two parameters, a_1 and a_2, which was obtained by fitting against the explicit ΔG_{pol},

$$\Delta G_{pol}^{corr} = \Delta G_{pol}^{3DRISM/GF} + a_1 \rho V + a_2, \tag{1}$$

Here, $\Delta G_{pol}^{3DRISM/GF}$, is the computed 3D-RISM ΔG_{pol} with Kovalenko-Hirata closure[23] assuming Gaussian fluctuation of the solvent, V is the computed partial molar volume and $\rho = 0.0333A^{-3}$ is the solvent number density. The corrected polar solvation energy, ΔG_{pol}^{corr} were obtained using optimizations performed using Nelder-Mead simplex algorithm. The same training set that was used for the rigid molecules and the amino acid analogs, Table 8 and the same objective function *rmse*(small molecules) + *rmse*(amino acid analogs) was used. These optimizations led to $a_1 = -0.0118$ kcal/mol and $a_2 = 0.6419$ kcal/mol. The performance of 3D-RISM against the explicit (TIP3P) ΔG_{pol} is shown in Figure 5 and Table 3. For the charge inverted "bracelets" the optimum values of $a_1 = 0.759$ kcal/mol and $a_2 = 0.1991$ kcal/mol were obtained by fitting with

the corresponding explicit ΔG_{pol} values in TIP4P-Ew water.



(a) (b)

Figure 5: The polar solvation free energies, ΔG_{pol} using 3D-RISM (corrected by two fitting parameters $a_1 = 0.0118$ kcal/mol and $a_2 = 0.6419$ kcal/mol, see Main Text) against the explicit (TIP3P) ΔG_{pol} of the rigid neutral molecule set (left) and the amino acid analogs (right)

Table 3: Accuracy of ΔG_{pol} computed using 3D-RISM relative to the reference explicit (TIP3P) simulation[4,12] for the rigid neutral molecule set and the amino acid analogs

	Small Mols.	**Amino Acids**
rmse	0.50	5.28
$\langle$**error**$\rangle$	-0.05	0.65
$\langle$|**error**|$\rangle$	0.36	2.95
corr. coef. $(\mathbf{r^2})$	0.98	0.95
%$\|$**error**$\| > \mathbf{2k_BT}$	2.4%	43.8%
RMS of worst 5%	2.95	15.58

5 Parameter re-optimization for flexible molecules

To minimize possible uncertainties due to inadequate conformational sampling of flexible molecules, in the Main Text we have restricted ourselves to a subset of 248 rigid molecules. However, including the flexible molecules to train the model parameters does not affect our overall conclusions. To this end, we re-optimize the models' parameters by using a new, larger training set. It contains the same 24 molecules of the amino acid analogs in Table 8. To these, we now add 124 molecules including both rigid and flexible kind from the small molecule set were chosen while keeping equal

representation of solvation free energy and atom types between the training set and the test set. We note that the new set now has one extra atom type namely Phosphorus (P) which was missing among the 248 rigid molecules. The parameter optimizations were performed using the same protocol (same objective function and validation) as in the case of the rigid molecules detailed in the Main Text. The optimum radii set is provided in the Table 4. Note that the radii values of this set are similar to the one found earlier, see Main Text. The optimum value of $\tau = 1.3$. The performance of the GB and CHA-GB models in Figure 3b and Table 5, shows similar agreement as that of the earlier optimization both for GB and CHA-GB.

Table 4: Intrinsic radii sets simultaneously re-optimized for GB and CHA-GB for all 504 molecules from the neutral molecule set (including the flexible ones) and the same 48 amino acid analogs used in the Main Text.

	Radii Set(Å)									
	C	**H**	**N**	**O**	**S**	**P**	**F**	**Cl**	**Br**	**I**
CHA-GB	1.60	0.52	1.58	1.36	1.72	1.63	1.22	1.63	1.84	2.14
GB	1.85	1.30	1.40	1.49	1.46	1.20	0.82	1.87	1.47	1.31

Table 5: Accuracy in ΔG_{pol} computed using GB and CHA-GB against the reference explicit (TIP3P) simulation[4,12] for all 504 molecules from the neutral molecule set and 48 amino acid analogs

	Small Mols		**Amino Acid Analogs**	
	GB	CHA-GB	GB	CHA-GB
rmse	1.34	0.89	1.31	0.89
$\langle$**error**$\rangle$	-0.43	-0.34	0.15	0.20
$\langle\lvert$**error**$\rvert\rangle$	0.98	0.62	1.00	0.67
corr. coef. $(\mathbf{r^2})$	0.82	0.92	0.997	0.998
%$\lvert$**error**$\rvert > \mathbf{2k_BT}$	28.4%	12.7%	27.1%	18.8%
RMS of worst 5%	3.86	2.70	3.85	2.29

6 Optimizing the non-polar part of solvation energy and comparison with experiment

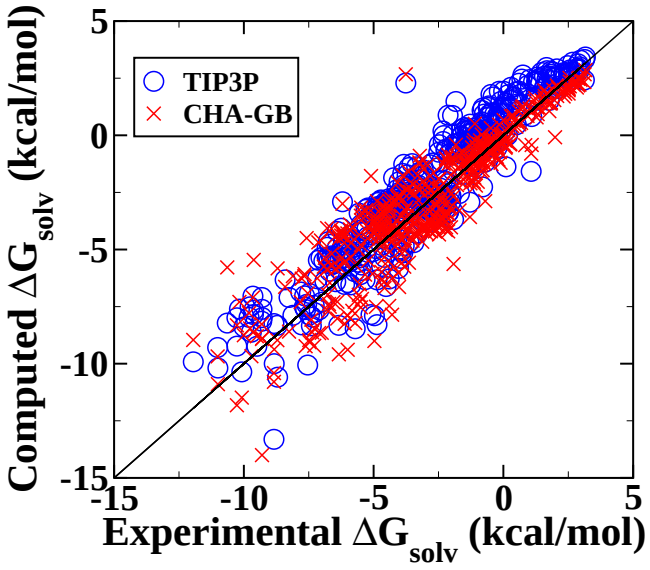


Figure 6: The solvation free energies, ΔG_{solv} using CHA-GB, red crosses and explicit (TIP3P) alchemical estimates, blue open circles[4] against the reference experimental values.[4]

We optimize ΔG_{np} against experimental solvation free energy under the approximation that the total solvation energy, $\Delta G_{solv} = \Delta G_{pol} + \Delta G_{np}$. The ΔG_{pol} values, Figure 3b, are taken from our previously optimized ΔG_{pol} using the optimum radii set from Table 4. The optimization protocol is adopted from Ref. 24. The non-polar component of the solvation energy can be decomposed into cavity (ΔG_{cav}) and van der Waals dispersion (ΔG_{vdw}) terms.[25]

$$\Delta G_{np} = \Delta G_{cav} + \Delta G_{vdw} \tag{2}$$

i.e.,

$$\Delta G_{np} = \gamma \cdot SASA - \sum_i \frac{16}{3}\pi d_w \varepsilon_{iw} \sigma_{iw}^6 \frac{\mu_i}{(R_i + \rho_w - R_s)^3} \tag{3}$$

Here γ is the effective surface tension coefficient and *SASA* is the solvent accessible surface area

Table 6: Lennard-Jones Parameters Used for Computation of ΔG_{np} for GB and CHA-GB; for GB optimum $\gamma = 0.0104$ kcal/mol/Å^2 and for CHA-GB $\gamma = 0.0178$ kcal/mol/Å^2

	σ_i(Å)	ε_i(kcal/mol)	μ_i	
			GB	CHA-GB
H	2.64953	0.0157	-0.0361	0.0192
C	3.39967	0.1094	0.1744	0.1296
O	2.9592	0.2100	-0.0460	0.1098
N	3.25	0.1700	0.1856	0.5433
S	3.56359	0.2500	-0.0172	0.3554
P	3.74177	0.2000	-0.2767	-0.2218
F	3.11815	0.061	-0.3751	0.0012
Cl	3.47094	0.265	0.1229	0.2464
Br	3.95559	0.320	0.0517	0.3007
I	4.18722	0.40	-0.0986	0.2722

Table 7: Accuracy in ΔG_{solv} computed using GB and CHA-GB compared to experimental values[4]

	TIP3P	GB	CHA-GB
rmse	1.26	1.45	1.22
$\langle$**error**$\rangle$	0.68	0.04	0.02
$\langle$|**error**|$\rangle$	1.03	1.09	0.91
corr. coef. ($\mathbf{r^2}$)	0.89	0.79	0.84
%|**error**| $>$ $\mathbf{2k_BT}$	40%	34%	30%
RMS of worst 5%	2.95	3.71	3.46

of a solute computed using the MSMS package[3] with the standard 1.4 Å water probe radius. In principle[25] γ can be different, specific to the atom type of a solute. However in this work we use a global(same for all atom types) value of γ similar to Ref. 24. $d_w = 0.033428$ Å^{-3} is the number density of water at standard conditions. ε_{iw} and σ_{iw} are computed using,

$$\begin{aligned} \sigma_{iw} &= \frac{1}{2}(\sigma_i + \sigma_w) \\ \varepsilon_{iw} &= \sqrt{\varepsilon_i \varepsilon_w} \end{aligned} \tag{4}$$

where $\sigma_w = 3.1507$ Å and $\varepsilon_w = 0.152$ kcal/mol are the Lennard-Jones (LJ) parameters for oxygen

in TIP3P water and σ_i, ε_i are the LJ parameters for the atom type i, standard GAFF[2] values were used in this work.

The values of global γ and μ_i were optimized using a training set comprised of the same 124 small molecules used earlier for the optimization of parameter sets provided in Table 4. Nelder-Mead simplex algorithm with 100 random initial seeds was used for this optimization. The final value of parameters(γ, μ_i's) were the ones pertaining to the lowest value of the objective function (*rmse* against the experimental ΔG_{solv}). The optimized parameters are provided in Table 6. and the performance of the two models are shown in Figure 6 and Table 7. Note that the μ_i's for certain atom types are negative, which leads to positive values of the dispersive terms in Eq. (3). This is unphysical because in Eq. (3) the dispersive terms are separated from the repulsive cavity term. The issue was discussed in Ref. 24; it is suggestive of inconsistencies involved in Eq. (3) itself.

Table 8: Training set

The part of training set with the rigid neutral small molecule set.			
111trichloroethane	thiophenol	35dimethylpyridine	cyclohexanol
112trichloro122trifluoroethane	trichloroethene	3acetylpyridine	dimethylamine
1234tetrachlorobenzene	Z12dichloroethene	3cyanophenol	dimethylether
1245tetrachlorobenzene	123trimethylbenzene	3methylbutanoicacid	dinbutylether
124trichlorobenzene	124trimethylbenzene	3methylpyridine	dinpropylether
135trichlorobenzene	12ethanediol	4acetylpyridine	Ebut2enal
14dichlorobenzene	135trimethylbenzene	4cyanophenol	ethanamide
2bromo2methylpropane	13dimethylnaphthalene	4methylacetophenone	ethane
2chloropyridine	14dioxane	4methylbenzaldehyde	methylcyanoacetate
2chlorotoluene	1methylnaphthalene	4methylpyridine	methylcyclohexane
2iodophenol	1methylpyrrole	acenaphthene	mxylene
2methylthiophene	1naphthol	acetaldehyde	Nacetylpyrrolidine
3chloroaniline	22dimethylpropane	aceticacid	naphthalene
4bromophenol	23dimethylnaphthalene	acetonitrile	nbutane
4chloroaniline	23dimethylphenol	acetophenone	nbutylacetate
4chlorophenol	26dimethylphenol	alphamethylstyrene	nitromethane
benzylbromide	26dimethylpyridine	ammonia	Nmethylacetamide
bromotrifluoromethane	2methoxyethanol	aniline	Nmethylmorpholine
chlorodifluoromethane	2methylbut2ene	anthracene	Nmethylpiperazine
chloroethane	2methylbut2ene	azetidine	NNdimethylformamide
chlorofluoromethane	2methylpropan2ol	benzaldehyde	NNdimethylpnitrobenzamide
diiodomethane	2methylpropane	benzamide	npentylacetate
dimethyldisulfide	2methylpropene	benzene	npropylbutyrate
dinpropylsulfide	2methylpyrazine	benzonitrile	piperidine
E12dichloroethene	2methylpyridine	but1yne	propan2ol
iodobenzene	2naphthol	buta13diene	pyrrole
methanethiol	2naphthylamine	butan2ol	pyrrolidine
methyltrifluoroacetate	33dimethylpentane	cis12dimethylcyclohexane	quinoline
pdibromobenzene	34dimethylphenol	cyanobenzene	styrene
tetrachloroethene	34dimethylpyridine	cyclohepta135triene	triacetylglycerol
tetrafluoromethane	35dimethylphenol	cyclohexane	trimethoxymethane
The part of training set from the set of amino acid analogs			
gly2-abt	phe2-abt	tyr2-abt	glh2-abt
ala2-abt	trp2-abt	asn2-abt	ash2-abt
val2-abt	met2-abt	gln2-abt	arg2-abt
leu2-abt	ser2-abt	hsd2-abt	lys2-abt
ile2-abt	thr2-abt	arn2-abt	asp2-abt
pro2-abt	cys2-abt	lyn2-abt	glu2-abt

Table 9: Test Set

Part of the test set with the rigid small neutral molecules			
11dichloroethane	3cyanopyridine	ethanol	methyltrimethylacetate
11dichloroethene	3formylpyridine	ethene	morpholine
11difluoroethane	3hydroxybenzaldehyde	ethylamine	npropylformate
1235tetrachlorobenzene	3methyl1hindole	ethylbutanoate	npropylpropanoate
123trichlorobenzene	4bromotoluene	ethylhexanoate	ocresol
12dichlorobenzene	4chloro3methylphenol	ethylpentanoate	otoluidine
13dichlorobenzene	4cyanopyridine	ethylpropanoate	oxylene
14dimethylnaphthalene	4fluorophenol	fluorene	pcresol
14dimethylpiperazine	4formylpyridine	fluorobenzene	pentanoicacid
1iodopropane	4hydroxybenzaldehyde	fluoromethane	phenanthrene
1methylcyclohexene	4methyl1himidazole	imidazole	phenol
1methylimidazole	bromobenzene	indane	piperazine
1naphthylamine	bromoethane	iodoethane	propane
22dimethylpentane	chlorobenzene	isobutylacetate	propanenitrile
23dimethylpyridine	chloroethylene	isopropylacetate	propanoicacid
24dimethylphenol	chloromethane	isopropylformate	propanone
24dimethylpyridine	cyclohexanone	mcresol	propene
25dimethylphenol	cyclohexylamine	methane	propionaldehyde
25dimethylpyridine	cyclopentane	methylamine	propyne
25dimethyltetrahydrofuran	cyclopentanol	methylbenzoate	ptoluidine
26dimethylaniline	cyclopentanone	methylbutanoate	pxylene
26dimethylnaphthalene	cyclopentene	methylchloroacetate	pyrene
2bromopropane	cyclopropane	methylcyclohexanecarboxylate	pyridine
2chloro2methylpropane	dibromomethane	methylcyclopentane	tetrachloromethane
2chloroaniline	dichloromethane	methylcyclopropanecarboxylate	tetrahydrofuran
2chlorophenol	diethylamine	methylcyclopropylketone	tetrahydropyran
2chloropropane	diethylmalonate	methylmethanesulfonate	thiophene
2fluorophenol	diethylsulfide	methylpentanoate	toluene
2iodopropane	diisopropylether	methylpmethoxybenzoate	trans14dimethylcyclohexane
3chlorophenol	dimethylsulfone	methylpropanoate	tribromomethane
3chloropyridine	ethanethiol	methyltbutylether	trichloromethane
Part of the test set from the amino acid analogs			
gly-abt	phe-abt	tyr-abt	glh-abt
ala-abt	trp-abt	asn-abt	ash-abt
val-abt	met-abt	gln-abt	arg-abt
leu-abt	ser-abt	hsd-abt	lys-abt
ile-abt	thr-abt	arn-abt	asp-abt
pro-abt	cys-abt	lyn-abt	glu-abt

References

(1) Mobley, D. L.; Ii, A. E.; Fennell, C. J.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 2405–2414.

(2) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. *J. Comput. Chem.* **2004**, *25*, 1157–74.

(3) Sanner, M. F.; Olson, A. J.; Spehner, J. C. *Biopolymers* **1996**, *38*, 305–320.

(4) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(5) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.

(6) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(7) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.

(8) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(9) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–94.

(10) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. *J. Phys. Cond. Matt.* **2010**, *22*, 492101+.

(11) Fennell, C. J.; Kehoe, C. W.; Dill, K. A. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3234–3239.

(12) Swanson, J. M. J.; Adcock, S. A.; McCammon, J. A. *J. Chem. Theory Comput.* **2005**, *1*, 484–493.

(13) Sigfridsson, E.; Ryde, U. *J. Comput. Chem.* **1998**, *19*, 377–395.

(14) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265–284.

(15) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. *Nucleic Acids Res.* **2012**, *40*, W537–W541.

(16) Shirts, M.; Mobley, D. *Biomolecular Simulations*; Methods in Molecular Biology; 2013; Vol. 924; pp 271–311.

(17) Case, D. et al. *University of California, San Francisco* **2012**,

(18) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.

(19) Ryckaert, J.; Ciccotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327–341.

(20) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(21) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(22) Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. *J. Chem. Theory Comput.* **2010**, *6*, 607–624.

(23) Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **1999**, *103*, 7942–7957.

(24) Aguilar, B.; Onufriev, A. V. *J. Chem. Theory Comput.* **2012**, *8*, 2404–2411.

(25) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.