Supporting Information
Why computed protein folding landscapes are sensitive to the water model

Ramu Anandakrishnan, Saeed Izadi, and Alexey V. Onufriev

October 28, 2024

1 Choice of the implicit solvent model.

For the purpose of our analysis we chose the OBC variant of the generalized Born implicit solvent model (GB-OBC), which is arguably one of the most widely used, “general-purpose” GB models. To evaluate the potential effect of using an alternate implicit solvent model, we re-sampled 20000 snapshots from the GB-OBC simulation using GBNeck2 – a more recent variant of the generalized Born model specifically parametrized to improve protein simulations. Figure S1 shows that the distribution of total potential energies calculated with the two GB models are qualitatively similar.

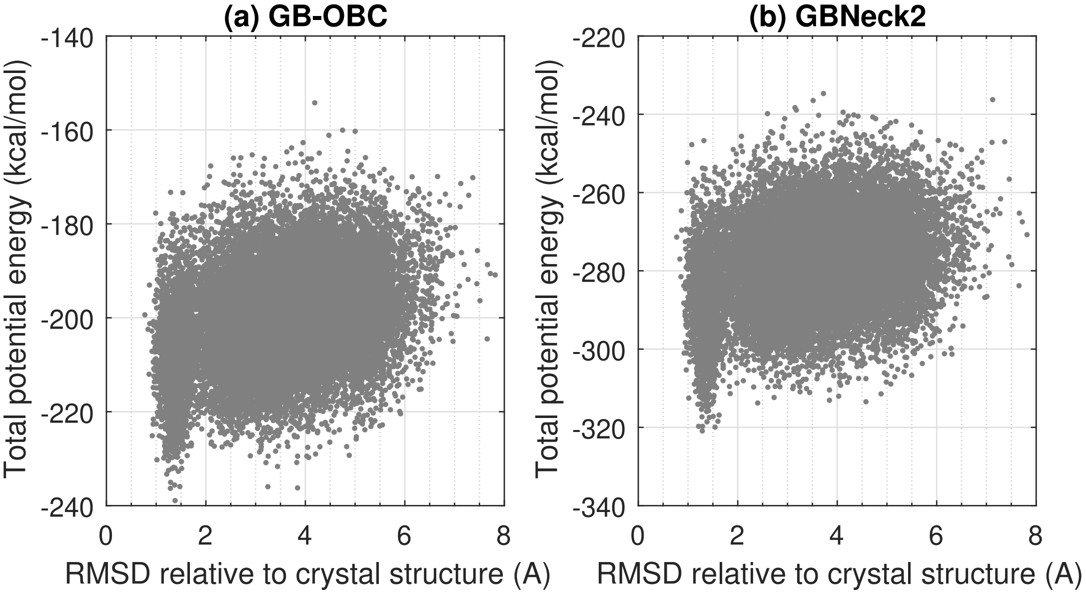

Figure S1: The potential energy landscape of CLN025. The total potential energy includes the solvation free energy estimated with either GB-OBC or GBNeck2 flavor of the generalized Born (GB) implicit solvent model. In both models a near native structure corresponds to the global minimum of the potential energy. In GB-neck2 structures with RMSD of about 4 Å from the native state are more strongly disfavored compared to GB-OBC. To obtain the GBNeck2 landscape, potential energy values of the 20000 snapshots from the original > 1 microsecond GB-OBC simulation were recomputed using GBNeck2 parameters.

2 Robustness of free energy landscape to the choice of gas-phase force-field

The free energy profiles from multiple (3) microsecond simulations of CLN025 using Amber ff12 force field (Figure S2) instead of ff10 (Figure 4 of the main text) shows the same clear differences between solvent models TIP3P, TIP4P/Ew and the GB.

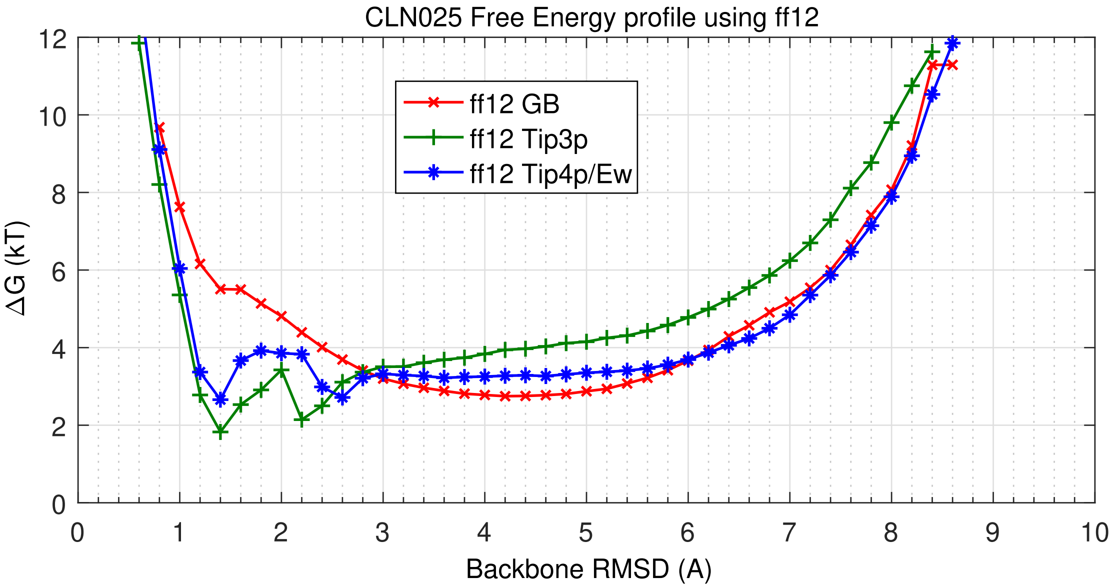

Figure S2: Computed free energy profile of CLN025 mini-protein at its experimental melting temperature, based on Amber ff12 force field. Connecting lines shown to guide the eye.

3 Random seed for Langevin dynamics ensures independent trajectories

We ran three separate one microsecond long simulations for each of the three water models. Each of these simulations follow distinct and independent trajectories due to the use of random seeds for Langevin dynamics (Fig. S3).

4 Identification of native-like and extended state structures

We have used RMSD relative to the native structure as the metric for differentiating between native-like and extended state structures. Figure S4 shows that for the RMSD values of 1.5 and 7.3 Å corresponding to the compact native-like and extended state structures, the corresponding ranges of radius-of-gyration (RG) values are relatively small (compared to the entire range of RG), more likely reflecting the natural flexibility of these structures, rather than representing significantly different structures.

5 Robustness of the energy decomposition to the choice of representative structures

Analysis of two additional randomly selected compact and extended states also show that water–water electrostatic interactions are the largest contributor to differences in folding ΔE (Figs. S6 and S7, and Table

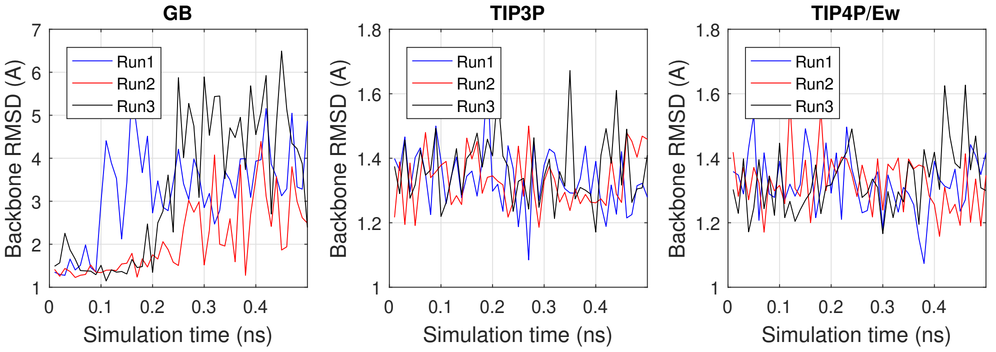

Figure S3: Simulation trajectories. Due to the use of Langevin dynamics with random seeds, each of the three simulations for each of the three water models follow distinct and independent trajectories. The first 0.5 nanoseconds of the 1 microsecond simulations are shown here. Connecting lines are shown to guide the eye.

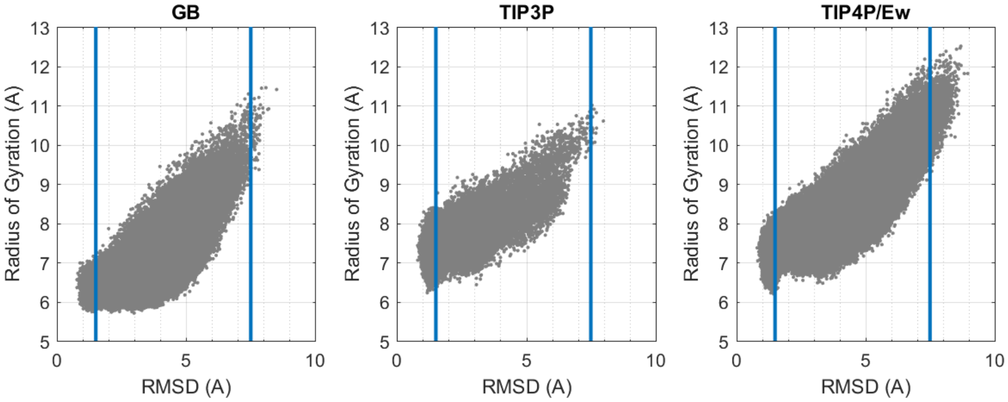

Figure S4: Comparison of RMSD and radius of gyration (RG). Vertical lines show the RMSD values corresponding to the compact native-like and extended state structures.

S1).

5.1 Alternative Estimation of the solvent free energy components

For the explicit solvent models, the solvent-solvent contribution $\Delta G_{solv-solv}$ to the total ΔG of folding can be estimated, at least qualitatively, as follows. First, we express $\Delta G_{solv-solv}$ via its enthalpy and entropy components:

$$\Delta G_{solv-solv} = \Delta H_{solv-solv} - T\Delta S_{solv} \tag{1}$$

We then approximate

$$T\Delta S_{solv} \approx T\Delta S - T\Delta S^{conf}_{prot} \tag{2}$$

where ΔS_{prot} is the configurational entropy contribution from the protein, $T\Delta S^{conf}_{prot} \approx -13$ kcal/mol based on a 4.2 cal/mol-K^{-1} per residue conformational entropy loss upon folding, used previously in a similar context, see main text. The decomposition, Eq. 2, of the total folding entropy ΔS into the solvent and protein parts should hold at least qualitatively, giving us an to order-of-magnitude estimate of $\Delta G_{solv-solv}$. Finally,

$$-T\Delta S = \Delta G - \Delta H \tag{3}$$

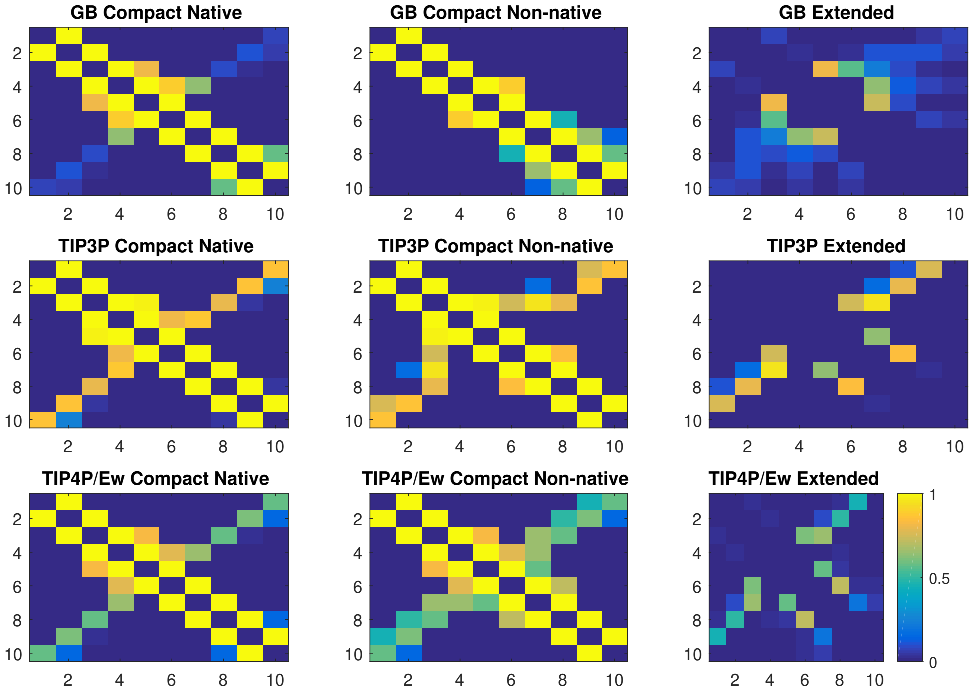

Figure S5: Residue-residue contact maps for the various structural states used in this study. Contact maps show clear differences between the compact native, compact non-native and extended state structures. The sampling of these states is qualitatively similar for the three water model, and is consistent with the free energy landscape shown in Fig. 1 of the main text. Contact is defined as the distance between alpha carbon atoms being < 6 Å. The color scale represents the frequency of each contact, averaged over the trajectory.

where the folding free energy ΔG is approximated from the computed free energy landscape, Figure 2 of the main text. The total enthalpy ΔH in the above equation, and $\Delta H_{solv-solv}$ in Eq. 1 can be approximated as $\Delta H \approx \Delta E$ and $\Delta H_{solv-solv} \approx \Delta E_{solv-solv}$ respectively, estimated from the potential energy components of the representative folded and unfolded structures as described in the main text.

$$\Delta G_{solv-solv} = \Delta E_{solv-solv} - \Delta G - \Delta E + T\Delta S_{prot}^{conf} \tag{4}$$

The results of using Eq. 4 for CLN025 are presented below, Table S2.

6 Energy decomposition data

The breakdown of the average energy from the restrained simulations of representative compact and extended states, as described in *Methods* in the main text, is shown in Table S3. This data is used to produce figure 4 of the main text. Here “vdw” refers to van der Waals interactions, “elec” to long-range electrostatic interactions, and “bond” to bonded interactions as defined by AMBER force field parameters. The terms not listed in the table can be calculated as described in *Methods* in the main text.

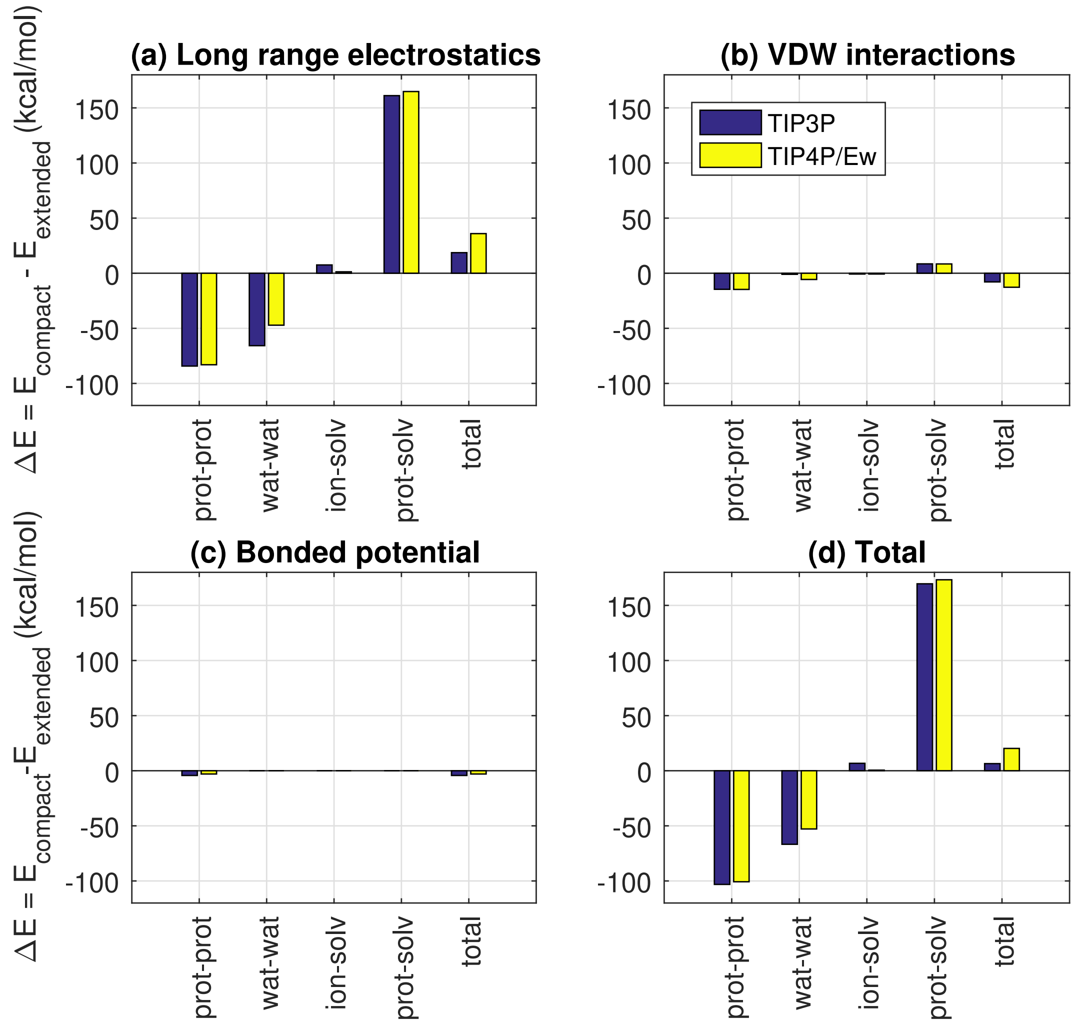

Figure S6: Decomposition of of average energy difference $\Delta E = E_{compact} - E_{extended}$ (kcal/mol) for an alternate set # 1 of randomly selected compact and extended state structures. Standard error of the mean for the total values are shown in Table S1

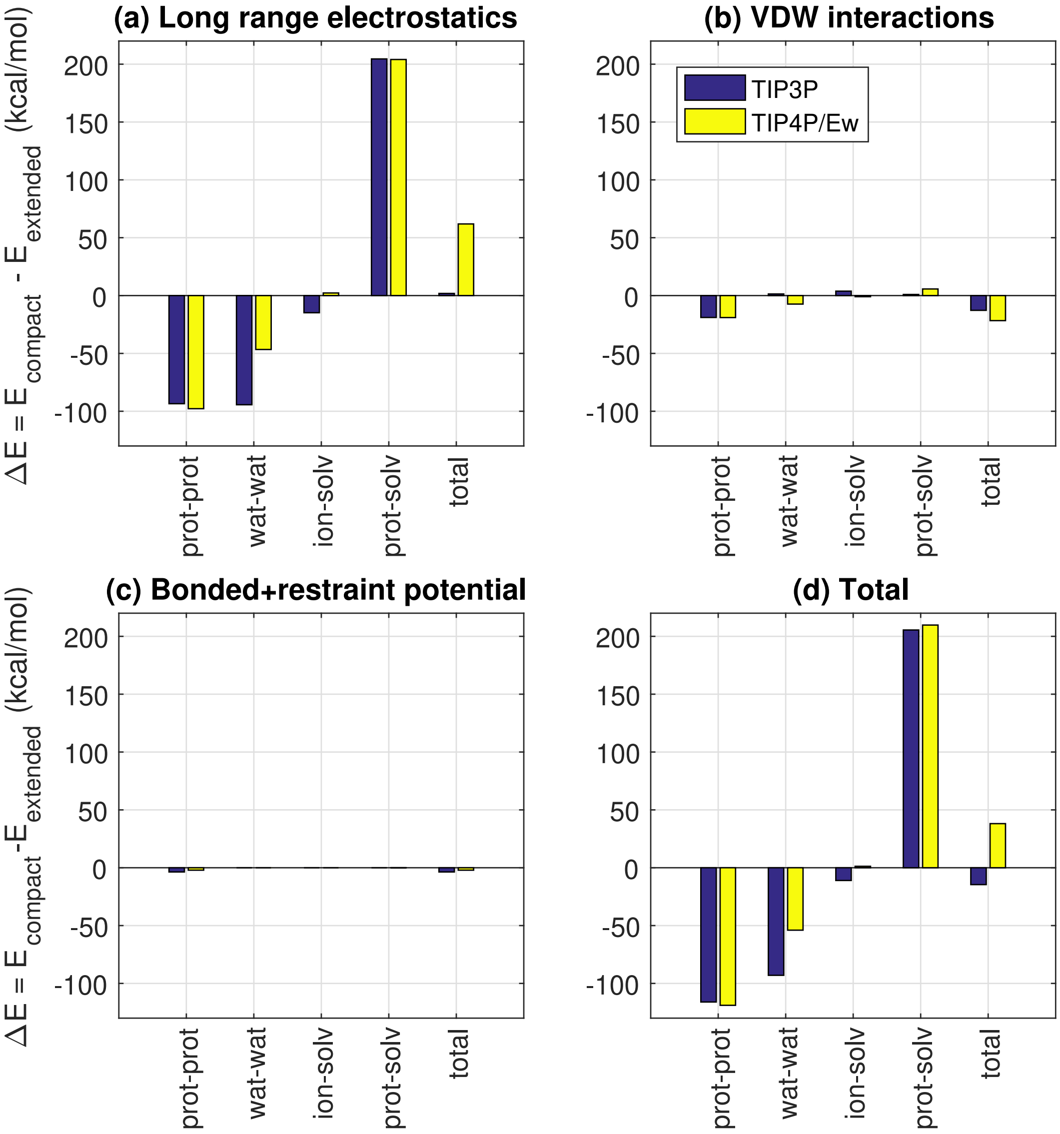

Figure S7: Decomposition of of average energy difference $\Delta E = E_{compact} - E_{extended}$ (kcal/mol) for an alternate set # 2 of randomly selected compact and extended state structures. Standard error of the mean for the total values are shown in Table S1

Table S1: Decomposition of average energy difference between two additional randomly selected alternate compact and extended structures. $\Delta E = E_{compact} - E_{extended}$ (kcal/mol). Standard error of the mean are shown next to each of the values.

Interaction	ΔE (kcal/mol)		
	TIP3P	TIP4P/Ew	$\Delta\Delta E$
Alternate set # 1 of compact and extended state structures			
Solvent-Solvent			
Water-Water	-66.74 ±1.29	-52.83 ±1.51	+13.91 ± 1.96
Ion-Solvent	-6.74 ±1.58	0.54 ±1.86	+7.28 ± 2.41
Protein-Solvent	+169.61 ±1.28	+173.26 ±1.54	+3.65 ± 1.98
Protein-Protein	-103.15 ±0.11	-100.73 ±0.13	+2.42 ± 0.16
Total	6.47 ±0.90	20.24 ±1.08	+13.77 ± 1.39
Alternate set # 2 of compact and extended state structures			
Solvent-Solvent			
Water-Water	-92.98 ±0.88	-53.96 ±0.78	+39.02 ±1.18
Ion-Solvent	-11.03 ±4.48	1.26 ±3.50	-9.77 ±5.68
Protein-Solvent	+205.48 ±6.42	+209.74 ±5.00	+4.27 ±8.14
Protein-Protein	-116.06 ±6.42	-118.92 ±5.02	-2.86 ±8.15
Total	-14.59 ±4.38	38.12 ±3.46	+52.71 ±5.58

Table S2: An alternative estimate of the contribution from solvent-solvent interactions to folding free energy ΔG (kcal/mol) of CLN025 mini-protein at its melting temperature. See *Methods* for details of the calculation. To be compared to Table 2 of the main text.

Free energy (kcal/mol)	TIP3P	TIP4P/Ew
ΔG (based on Table 1, main text)	-2.2	-0.67
$\Delta G_{solv-solv}$	-97.2	-109.1

Table S3: Average energy decomposition data (kcal/mol) used for Figure 4 of main text.

Component	Interaction	TIP3P		TIP4P/Ew		GB	
		Compact	Extended	Compact	Extended	Compact	Extended
Total	vdw	3248.71	3261.84	4299.57	4317.37		
Total	elec	-27742.13	-27757.31	-31664.01	-31702.60		
Total	bond	211.85	205.87	211.75	206.95		
prot-prot	vdw	-15.90	3.08	-16.16	2.90	-21.91	-1.11
prot-prot	elec	-250.70	-152.89	-249.64	-151.14	-199.29	-102.54
prot-prot	bond	211.85	205.87	211.75	206.95	219.14	219.77
solv-solv	vdw	3310.01	3308.49	4359.69	4363.44		
solv-solv	elec	-26936.43	-26847.25	-30817.89	-30747.61		
water-water	vdw	3169.68	3168.46	4216.83	4221.50		
water-water	elec	-24566.44	-24496.35	-28478.49	-28447.93		
prot-solv	vdw	-45.74	-49.74	-43.97	-48.98		
prot-solv	elec	-555.00	-757.18	-596.49	-808.00		
ion-water	vdw	140.32	140.04	142.86	141.94		
ion-water	elec	-2369.99	-2350.89	-2339.40	-2299.69		
solvation	total					-307.44	-402.74

Table S4: Standard error of the mean for the values shown in Figure 4 of the main text

interaction	prot-prot	wat-wat	ion-solv	prot-solv	total
TIP3P standard error (kcal/mol)					
Electrostatic	0.1	1.6	1.0	0.3	1.3
VDW	0.1	0.7	0.1	0.1	0.7
Bonded	0.1	0.0	0.0	0.0	0.1
Total	0.1	1.3	1.5	1.2	0.9
TIP4P/Ew standard error (kcal/mol)					
Electrostatic	0.1	2.0	1.1	0.3	1.7
VDW	0.1	1.0	0.1	0.1	1.0
Bonded	0.1	0.0	0.0	0.0	0.1
Total	0.1	1.5	1.8	1.5	1.1