Supplementary Material for:

Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and their Complexes with Proteins

Hai Nguyen,[a,b] *Alberto Pérez,*[b] *Sherry Bermeo*[a], *Carlos Simmerling*[a,b*]
[a)] Department of Chemistry, Stony Brook University, Stony Brook, NY 11794, USA
[b)] Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA
*carlos.simmerling@stonybrook.edu

Table S1. Summary of training and test sets for GB-Neck2

	Name	#structures
Training set	dnadup $(CCAACGTTGG)_2$	370
	rnadup $(CCAACGUUGG)_2$	187
Type I test set	dnadup_plus150 $(CCAACGTTGG)_2$	520
	rnadup_plus200 $(CCAACGUUGG)_2$	387
Type II test set	DNA duplex $(CGCGAATTCGCG)_2$	650
	RNA duplex $(CGCGAAUUCGCG)_2$	600
	DNA/protein complex 1GCC	850

Table S2. Parameters for the lowest 10 out of 600 runs for the final optimization round, ranked according to objective function values for weight factors wr=2.5, w_rel=5.0. "Sx < 0" means one or more of the scaling factors is negative while "Sx > 0" means all the scaling factors are positive. The three last rows show the RMSD between GB and PB absolute energies (kcal/mol) for dnadup and rnadup and the effective radii RMSD (Å) for the dnadupRad training set. The relative energy RMSDs are shown in parentheses. The best run was picked if it satisfies: 1.) having low objective function. 2) having positive values for all scaling factor S_x. Based on this criteria, we picked runs #3 and #4 for final comparison (bot have similar objective functions but significantly different parameters). We finally chose parameter set #4 as our final candidate since it has slightly lower relative energy RMSD than set #3 for both training sets.

	#1	#2	#3	**#4**	#5	#6	#7	#8	#9	#10
Pars	Sx < 0	Sx < 0	Sx > 0	**Sx > 0**	Sx < 0	Sx > 0	Sx > 0	Sx < 0	Sx > 0	Sx > 0
S_H	-0.556	-0.546	1.175	**1.697**	0.536	1.225	1.127	1.201	1.211	1.229
S_C	0.920	0.887	0.669	**1.269**	0.817	0.644	0.611	0.178	0.175	0.095
S_N	1.118	1.083	1.066	**1.426**	1.065	1.075	0.997	1.011	1.016	1.032
S_O	-0.309	0.221	0.184	**0.184**	-0.333	0.405	0.183	-0.019	0.184	0.184
Sp	1.500	1.451	1.487	**1.545**	1.432	1.491	1.418	1.434	1.448	1.476
α_H	1.373	1.359	1.184	**0.537**	1.241	1.193	1.421	1.368	1.180	1.315
β_H	2.114	2.146	1.592	**0.363**	1.985	1.575	2.184	2.011	1.503	1.853
γ_H	1.338	1.453	1.067	**0.117**	1.494	1.022	1.543	1.527	1.189	1.375
α_C	0.750	1.165	-0.204	**0.332**	-0.402	0.789	0.036	0.452	1.198	1.794
β_C	-0.384	0.578	-1.198	**0.197**	-2.769	1.024	-0.826	0.039	1.901	3.236
γ_C	-0.337	0.223	-0.233	**0.093**	-1.473	0.934	0.039	0.572	1.673	2.357
α_N	2.361	2.773	1.503	**0.686**	0.364	2.104	1.944	2.096	2.565	2.905
β_N	2.648	3.843	1.953	**0.463**	-1.239	3.314	3.055	3.456	4.612	5.418
γ_N	1.013	1.772	1.208	**0.139**	-0.837	1.952	2.000	2.373	3.071	3.511

α_O	1.277	1.234	1.137	**0.606**	1.898	0.947	1.063	1.294	1.081	1.005
β_O	2.470	2.459	1.937	**0.463**	4.436	1.381	1.794	2.419	1.814	1.587
γ_O	1.918	2.075	1.396	**0.142**	3.514	0.990	1.464	1.928	1.477	1.277
α_P	1.222	0.812	1.077	**0.418**	0.612	1.104	1.109	1.234	1.143	1.102
β_P	3.150	2.028	2.321	**0.290**	0.963	1.782	2.164	1.911	1.776	1.525
γ_p	4.847	4.199	3.061	**0.106**	3.062	2.044	3.197	2.524	2.570	2.202
obj_funct	0.324	0.327	0.336	**0.338**	0.339	0.340	0.340	0.349	0.351	0.353
dnadup	11.6 (12.7)	11.6 (13.0)	14.8 (11.4)	**14.0 (10.3)**	13.6 (11.8)	15.6 (11.8)	15.8 (11.7)	15.5 (11.8)	15.8 (12.2)	16.1 (12.5)
rnadup	9.3 (9.8)	9.5 (10.4)	24.0 (12.4)	**25.4 (11.7)**	14.5 (10.4)	25.1 (12.8)	24.6 (12.7)	25.4 (12.5)	24.1 (12.5)	23.8 (12.7)
dnadupRad	0.046	0.044	0.035	**0.041**	0.048	0.033	0.037	0.037	0.038	0.037

Table S3. Summary of structures used in this study. "*GB vs. PB*" means the structure was used for comparing GB and PB calculations. "*MD simulation*" means the structures were used to carry out MD simulations. "x" denotes the structures were used for the corresponding calculations. Blank indicates the pair of structure and calculation was not carried out.

System	System size (Number of residues)	Net Charge	GB vs. PB	MD simulation	
				GB	TIP3P
DNA duplex $(CCAACGTTGG)_2$	20	-18	x	x	x
DNA duplex $(CGCGAATTCGCG)_2$	24	-22	x	x	x
RNA duplex $(CCAACGUUGG)_2$	20	-18	x	x	x
RNA duplex $(CGCGAAUUCGCG)_2$	24	-22	x	x	x
DNA duplex $(CTAGGTGGATGACTCATT)_2$	36	-24		x	x
DNA G-quadruplex $(GGGG)_4$[1]	16	-12		x	x
DNA G-quadruplex 2 (PDB ID: 1L1H)[2]	24	-22	x	x	x
DNA-protein complex (PDB ID: 1GCC)[3]	85	-13	x	x	x
DNA GCA hairpin loop (PDB: 1ZHU)[4]	7	-6		x	
RNA UUCG hairpin loop (PBD: 2KOC)[5]	14	-13		x	

Table S4. Comparison of energy and effective radii RMSD between GB and PB calculations for the different training and testing sets with different weighting factors (w_r = 1.5, 2.5, 5.0; w_{rel} = 5.0, 10.0). w_{abs} is 1.0. The best solution for each of these weighting factor combinations is shown. We performed 300-600 function minimization runs for each combination. The fitting parameters from (w_r = 2.5, w_{rel} = 5.0, w_{abs} = 1.0) were chosen as the final parameters since we felt that they have the best compromise between low energy RMSD and low effective radii RMSD with respect to the PB calculation. For the dnadup and rnadup training sets, the values are the absolute and (relative) energy RMSD values in kcal/mol. For dnaduprad, the values are the effective Born radii RMSD in Å.

Training set		GB-Neck2		
	w_r=1.5 w_{rel}=5.0	**w_r=2.5** **w_{rel}=5.0**	w_r=5.0 w_{rel}=5.0	w_r=2.5 w_{rel}=10.0
dnadup	18.7 (10.2)	**14.0 (10.3)**	17.1 (12.5)	14.7 (12.0)
rnadup	26.4 (11.3)	**25.4 (11.7)**	29.6 (12.9)	31.1 (11.2)

dnadupRad	0.051	**0.041**	0.030	0.050

Table S5. Atoms **(**in RNA B-form duplex CGCGAAUUCGCG) that have greater than 2.0 Å errors in GB-Neck model effective Born radii as compared to PB perfect radii. For this system, GB-Neck2 had no errors larger than 2.0 Å.

C5_1@H1' means: Atom H1' at residue 1 (C5) (and so on)

Atom	GB-Neck effective radii (Å)	PB perfect radii (Å)
C5_1@H1'	6.20	2.74
C5_1@O2'	5.07	2.91
C5_1@HO2'	5.07	2.21
G_2@H1'	5.66	3.15
G_2@C2'	5.43	3.11
G_2@O2'	5.69	3.16
G_2@HO2'	6.66	2.75
C_3@O5'	4.98	2.95
C_3@C1'	6.39	3.53
C_3@H1'	7.31	3.08
C_3@C2'	5.45	2.98
C_3@O2'	5.46	2.93
C_3@HO2'	5.33	2.22
G_4@H5'	4.08	1.94
G_4@H1'	5.52	3.10
G_4@C2'	5.15	2.96
G_4@O2'	5.18	2.92
G_4@HO2'	5.05	2.21
A_5@H1'	5.13	2.93
A_5@C2'	5.13	2.96
A_5@O2'	5.16	2.91
A_5@HO2'	4.98	2.21
A_6@C1'	5.43	3.33
A_6@H1'	5.16	2.56
A_6@C2'	5.52	3.09
A_6@O2'	5.76	3.15
A_6@HO2'	6.65	2.78
U_7@O5'	5.05	2.95
U_7@C1'	6.48	3.41
U_7@H1'	6.84	2.77
U_7@H6	4.21	2.19
U_7@C2'	5.95	3.11
U_7@O2'	6.14	3.17
U_7@HO2'	7.10	2.79

U_8@O5'	5.16	2.96
U_8@H5'	4.10	1.90
U_8@C1'	6.56	3.41
U_8@H1'	6.86	2.76
U_8@H6	4.37	2.19
U_8@C2'	5.99	3.10
U_8@O2'	6.12	3.14
U_8@HO2'	7.06	2.75
C_9@O5'	5.13	2.95
C_9@H5'	4.11	1.91
C_9@C1'	6.22	3.34
C_9@H1'	7.09	2.82
C_9@H6	4.25	2.18
C_9@C2'	5.52	2.96
C_9@O2'	5.51	2.91
C_9@HO2'	5.39	2.22
G_10@H5'	4.06	1.90
G_10@C1'	5.84	3.60
G_10@H1'	5.85	3.14
G_10@C2'	5.57	3.11
G_10@O2'	5.77	3.15
G_10@HO2'	6.70	2.75
C_11@O5'	5.03	2.95
C_11@H5'	3.96	1.94
C_11@C1'	6.33	3.53
C_11@H1'	7.19	3.08
C_11@C2'	5.38	2.98
C_11@O2'	5.28	2.93
C_11@HO2'	5.12	2.22
G3_12@H5'	3.99	1.93
C5_13@H1'	6.20	2.73
C5_13@O2'	5.07	2.89
C5_13@HO2'	5.07	2.20
G_14@H1'	5.66	3.16
G_14@C2'	5.43	3.11
G_14@O2'	5.69	3.16
G_14@HO2'	6.66	2.77
C_15@O5'	4.98	2.96
C_15@C1'	6.39	3.53
C_15@H1'	7.30	3.09
C_15@C2'	5.45	2.97
C_15@O2'	5.45	2.92
C_15@HO2'	5.33	2.22
G_16@H5'	4.07	1.93
G_16@H1'	5.52	3.12
G_16@C2'	5.15	2.97

G_16@O2'	5.19	2.93
G_16@HO2'	5.05	2.23
A_17@H1'	5.13	2.92
A_17@C2'	5.13	2.97
A_17@O2'	5.16	2.93
A_17@HO2'	4.98	2.23
A_18@C1'	5.43	3.34
A_18@H1'	5.17	2.58
A_18@C2'	5.52	3.09
A_18@O2'	5.76	3.14
A_18@HO2'	6.65	2.77
U_19@O5'	5.06	2.96
U_19@C1'	6.48	3.42
U_19@H1'	6.84	2.76
U_19@H6	4.22	2.20
U_19@C2'	5.95	3.11
U_19@O2'	6.13	3.16
U_19@HO2'	7.10	2.78
U_20@O5'	5.16	2.96
U_20@H5'	4.10	1.90
U_20@C1'	6.56	3.42
U_20@H1'	6.86	2.77
U_20@H6	4.37	2.20
U_20@C2'	5.99	3.11
U_20@O2'	6.12	3.14
U_20@HO2'	7.06	2.75
C_21@O5'	5.14	2.95
C_21@H5'	4.10	1.90
C_21@C1'	6.22	3.33
C_21@H1'	7.09	2.80
C_21@H6	4.25	2.19
C_21@C2'	5.52	2.96
C_21@O2'	5.51	2.92
C_21@HO2'	5.39	2.22
G_22@H5'	4.07	1.91
G_22@C1'	5.84	3.61
G_22@H1'	5.85	3.15
G_22@C2'	5.57	3.11
G_22@O2'	5.77	3.15
G_22@HO2'	6.70	2.75
C_23@O5'	5.03	2.95
C_23@H5'	3.96	1.94
C_23@C1'	6.33	3.53
C_23@H1'	7.20	3.09
C_23@C2'	5.38	2.98
C_23@O2'	5.28	2.93

C_23@HO2'	5.13	2.21
G3_24@H5'	4.00	1.93

Table S6. Average groove widths (Å) of a DNA duplex $(CCAACGTTGG)_2$ and a RNA duplex $(CCAACGUUGG)_2$ from GB-Neck2 and TIP3P MD simulations. There are two runs for each solvent model, starting from A and B-forms, with resulting uncertainties shown in parentheses. These DNA and RNA duplexes were used for training GB-Neck2 parameters.

Groove width (Å)	DNA $(CCAACGTTGG)_2$		RNA $(CCAACGUUGG)_2$	
	GB-Neck2	TIP3P	GB-Neck2	TIP3P
Major	18.6 (0.1)	18.1 (0.1)	15.2 (0.1)	19.0 (0.1)
Minor	13.0 (0.1)	12.4 (0.1)	15.9 (0.1)	15.4 (0.1)

Table S7. Average major and minor groove widths (Å) of DNA duplex $(CTAGGTGGATGACTCATT)_2$ from GB-Neck2 and TIP3P MD simulations. There are two runs (starting from A and B forms) for each solvent model, with resulting uncertainties shown in parentheses.

Groove width (Å)	GB-Neck2	TIP3P
Major	19.1 (0.1)	19.5 (0.1)
Minor	13.2 (0.1)	12.6 (0.1)

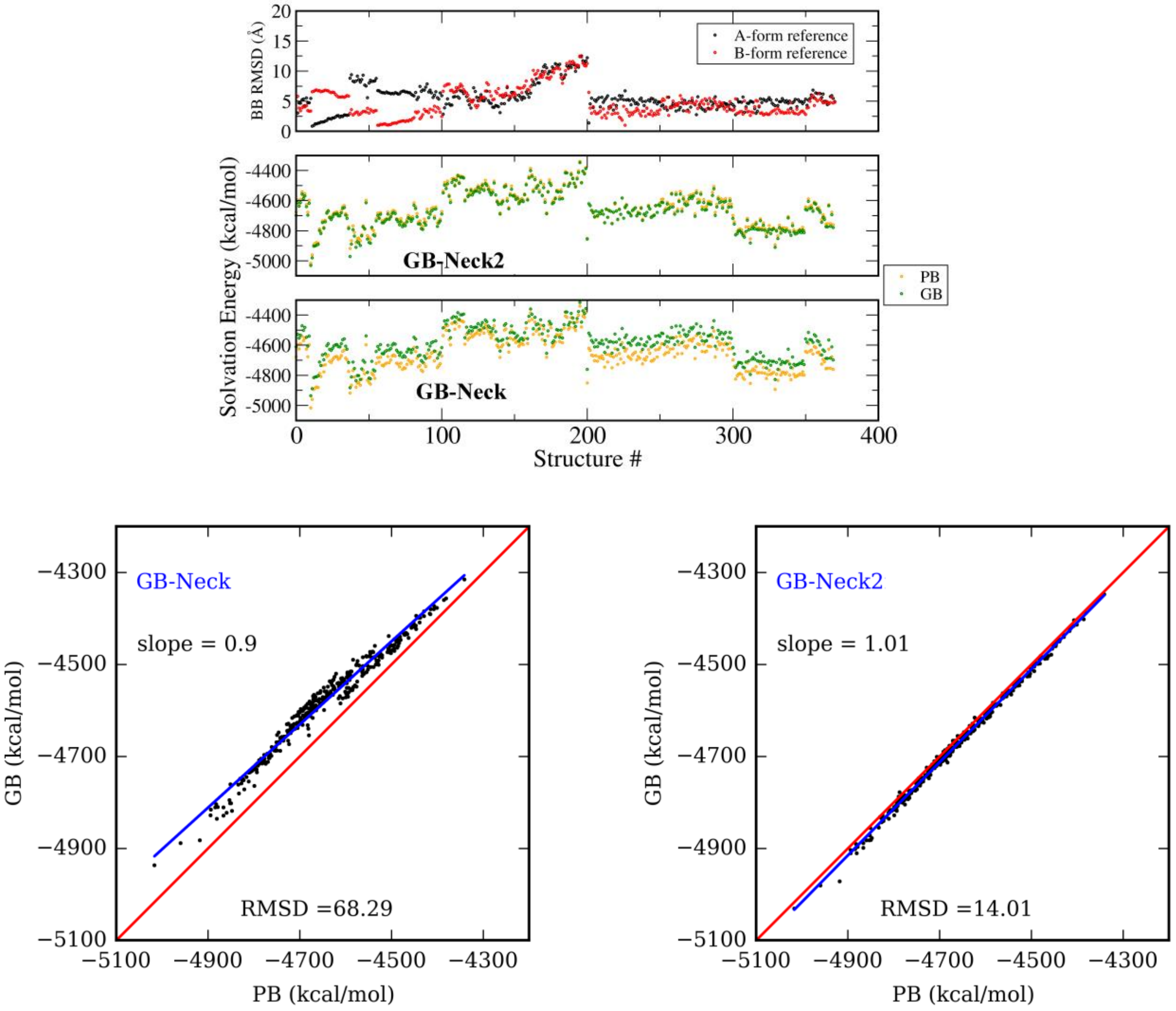

Figure S1. **(Upper figure)** Structural and energetic diversity of the **training set** for DNA (dnadup). The first panel shows the backbone RMSD (Å) of each structure to canonical A and B-forms of DNA. The bottom two panels compare GB and PB energies (kcal/mol) for individual structures in the DNA training set. **(Lower figure)** Comparison between GB (left: GB-Neck; right: GB-Neck2) and PB energies (kcal/mol) for the same data shown in the upper figure.

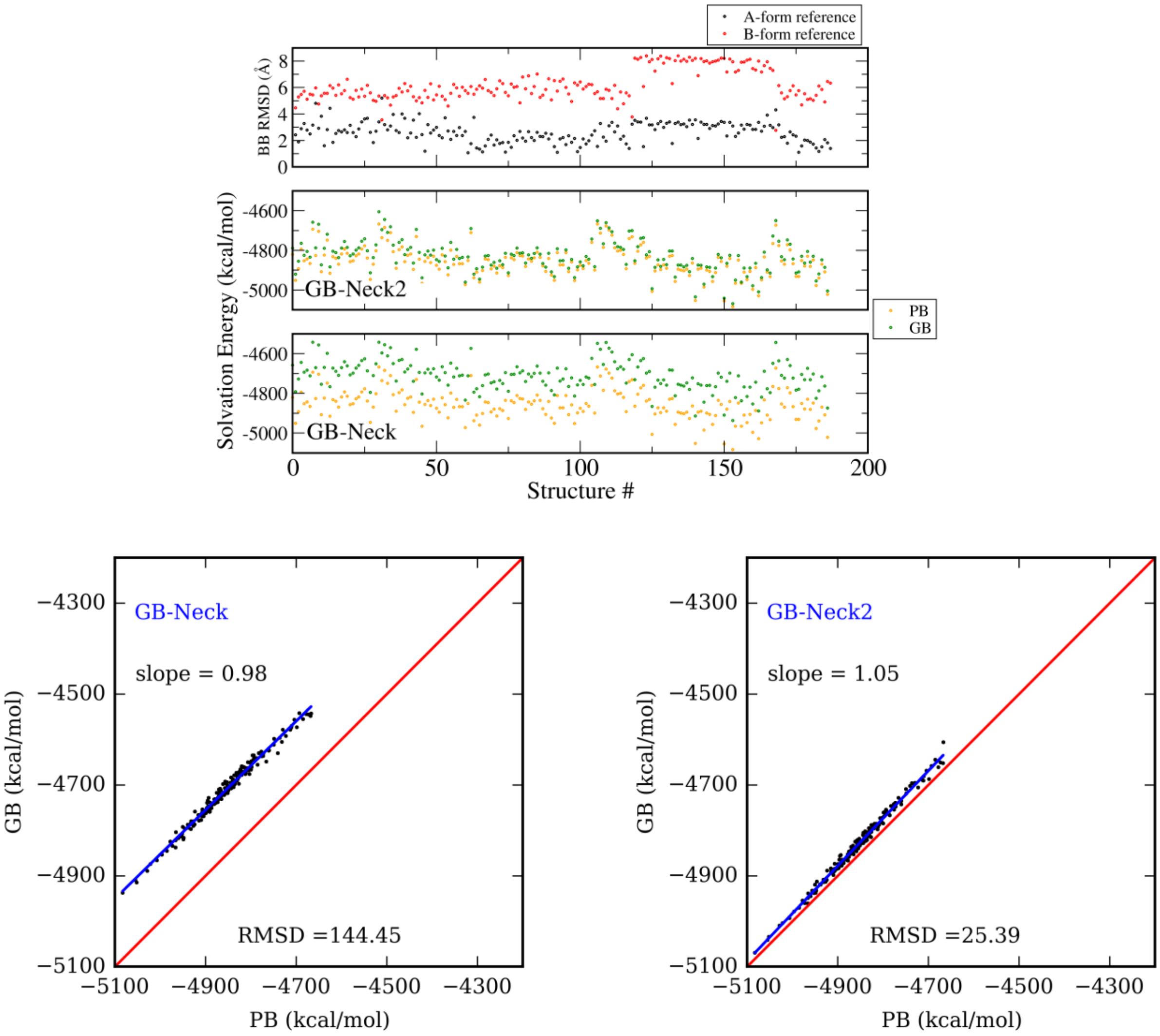

Figure S2. **(Upper figure)** Structural and energetic diversity of the **training set** for RNA (rnadup). The first panel shows the backbone RMSD of each structure to canonical A and B-forms of RNA. The bottom two panels compare GB and PB energies for individual structures in the RNA training set. **(Lower figure)** Comparison between GB (left: GB-Neck; right: GB-Neck2) and PB energies for the same data as in the upper figure.

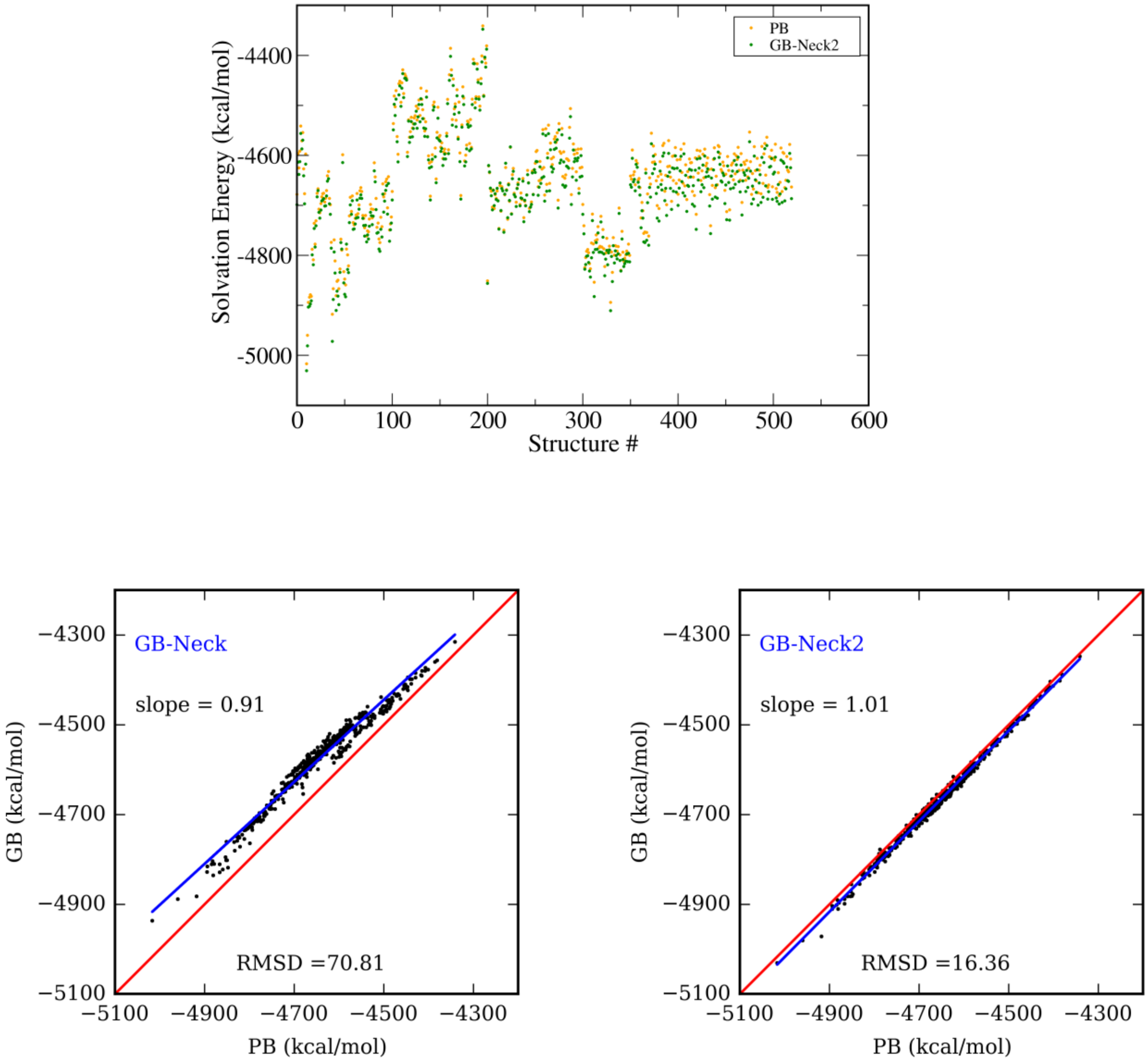

Figure S3. Comparison between GB and PB energies during the training of the DNA duplex, including training structures as shown in **Figure S1**, as well as additional MD structures obtained for the system after training. (**Upper**) The first 370 structures in the plot correspond to the individual structures used in the DNA training set (in the 5th round: dnadup), and the last 150 structures come from a 0.5μs MD simulation of the same structure, using GB parameters from the 5th round. Optimization of GB parameters was stopped after the 5th round since there is no strong energy bias for the new structures. (**Lower**) Comparison between GB (left: GB-Neck; right: GB-Neck2) and PB energies for all of the training set structures for the DNA duplex. The blue line and red line indicate the best fit and x=y respectively. Including the last 150 structures in the analysis did not significantly reduce the error between GB and PB energies (vs. **Figure S1**). This indicates that our training set was reasonably converged after five rounds.

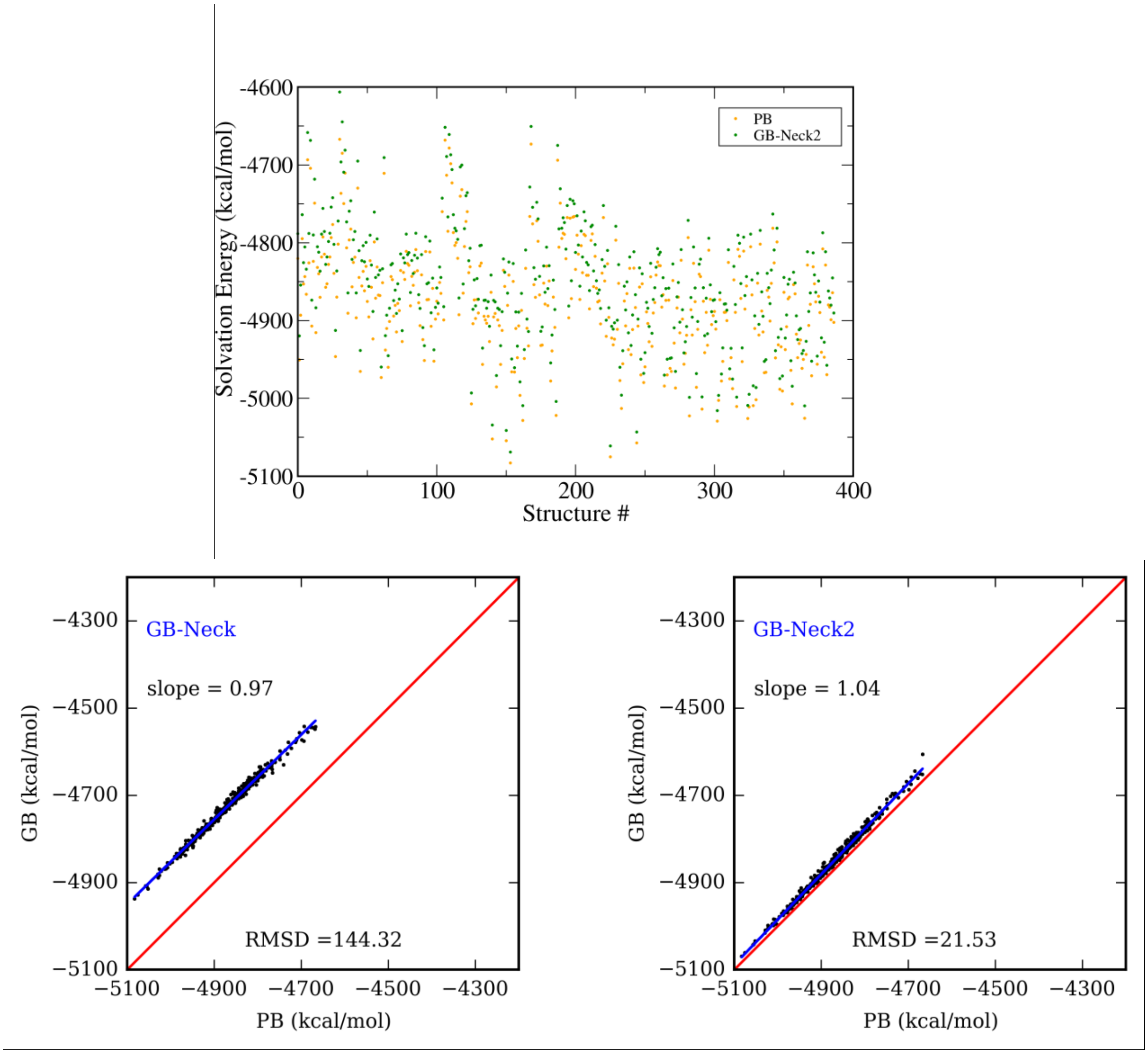

Figure S4. Comparison between GB and PB energies during the training of the RNA duplex, including training structures as shown in **Figure S2**, as well as additional MD structures obtained for the system after training. (**Upper**) The first 187 structures in the plot correspond to the individual structures used in the RNA training set (in the 5th round: rnadup), and the last 200 structures come from a 1.0 µs MD simulation of the same structure, using GB parameters from the 5th round. Optimization of GB parameters was stopped after the 5th round since there is no strong energy bias for the new structures. (**Lower**) Comparison between GB (left: GB-Neck; right: GB-Neck2) and PB energies for all of the training set structures for the RNA duplex. The blue line and red line indicate the best fit and x=y respectively. Including the last 200 structures in the analysis did not significantly reduce the error between GB and PB energies (vs. **Figure S2**). This indicates that our training set was reasonably converged after five rounds.

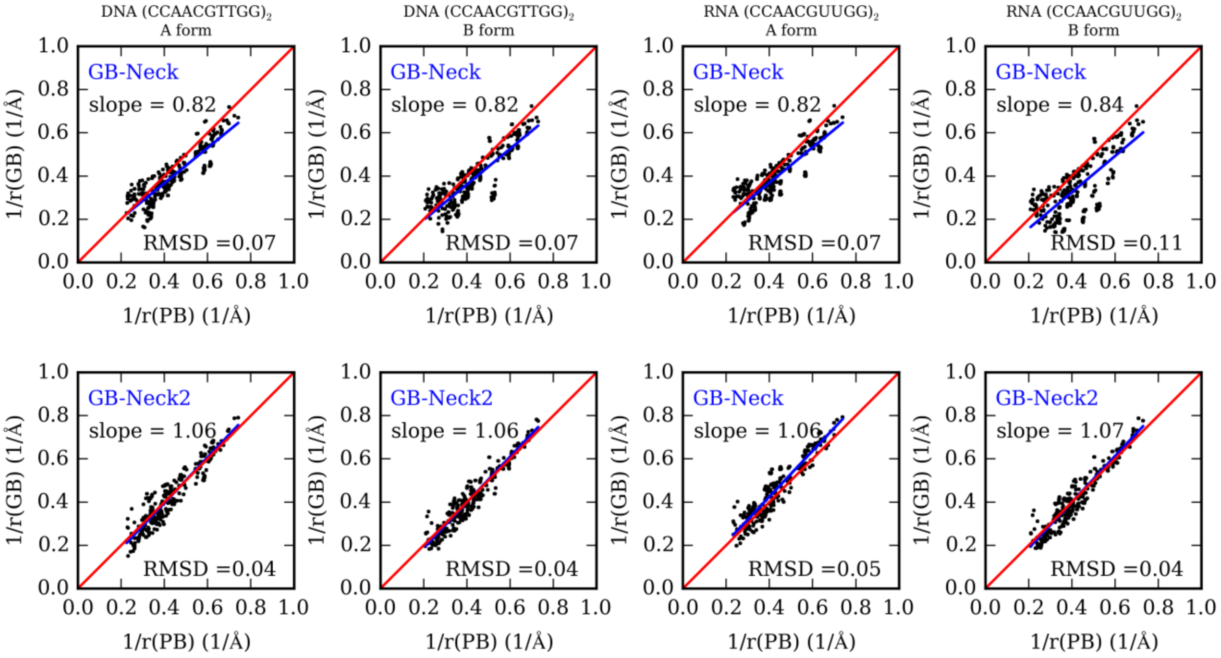

Figure S5. Correlations between GB and PB radii. The top row shows the inverse of the effective radii in the original GB-Neck vs the inverse of PB "perfect" radii for different systems used for radii training of GB-Neck2. The bottom row shows the results for GB-Neck2. The red line in each subplot indicates the ideal agreement between GB and PB effective radii. The blue line indicates the best fit line. Slope deviations from 1 indicate that the change in effective radius with degree of burial is not accurately reproduced, and will lead to bias in simulations.

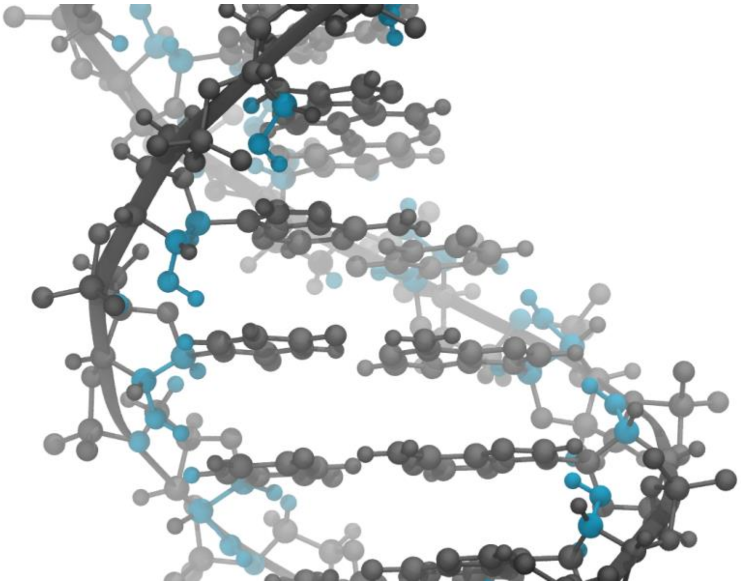

Figure S6. Effective radii overestimation in the original GB-Neck for B-form RNA (CGCGAAUUCGCG)$_2$. Atoms shown in cyan color are those having (GB effective radii – PB effective radii) > 2 Å. The raw data are shown in **Table S5**. For this system, GB-Neck2 does not have any atoms with radii overestimation greater than 2. Å.

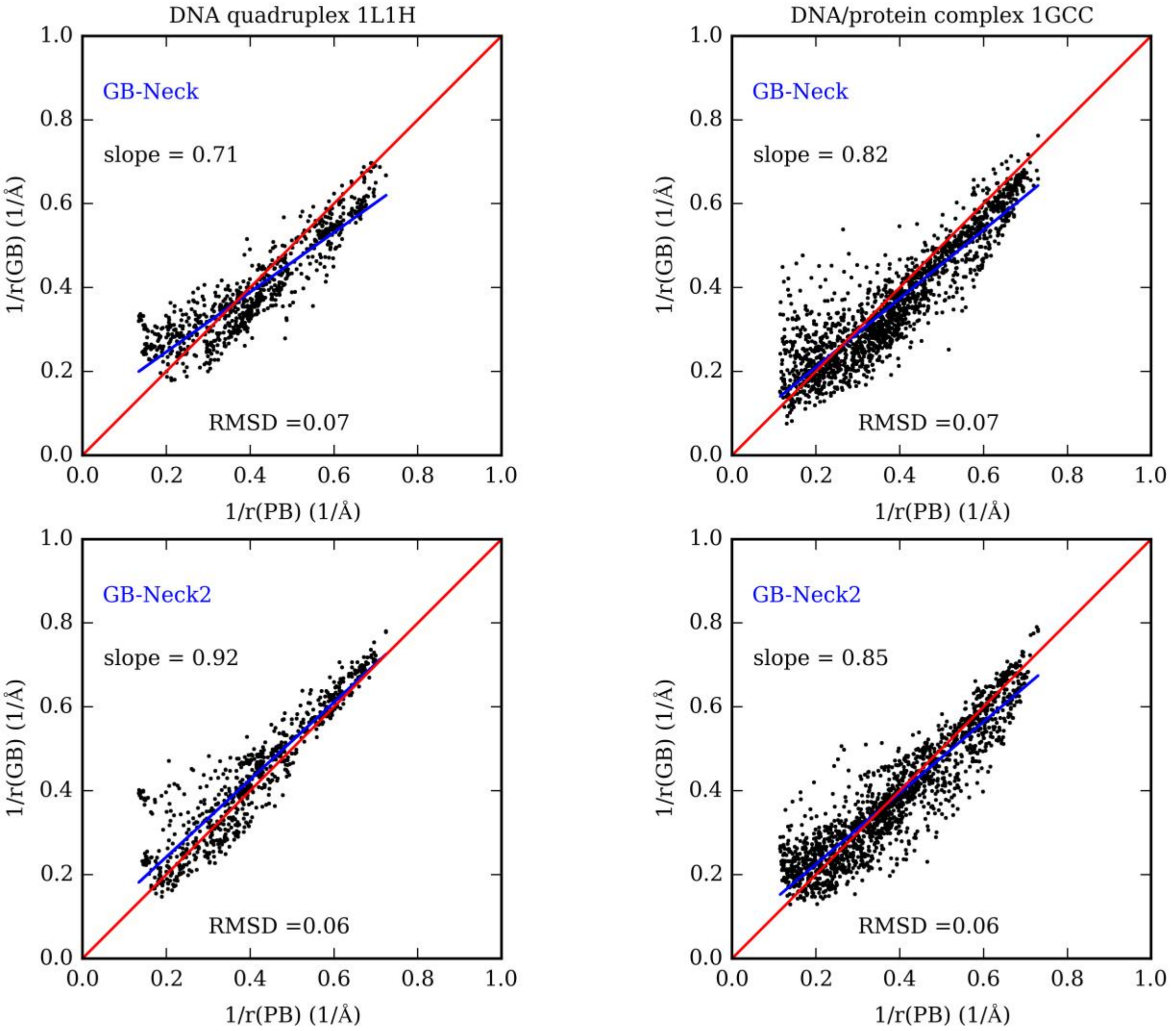

Figure S7. Comparison of the inverse of effective radii between GB-Neck (top), GB-Neck2 (bottom) and the inverse of PB "perfect" radii for DNA quadruplex (PDB ID 1L1H)[9] and DNA/protein complex (PDB ID 1GCC).[3]

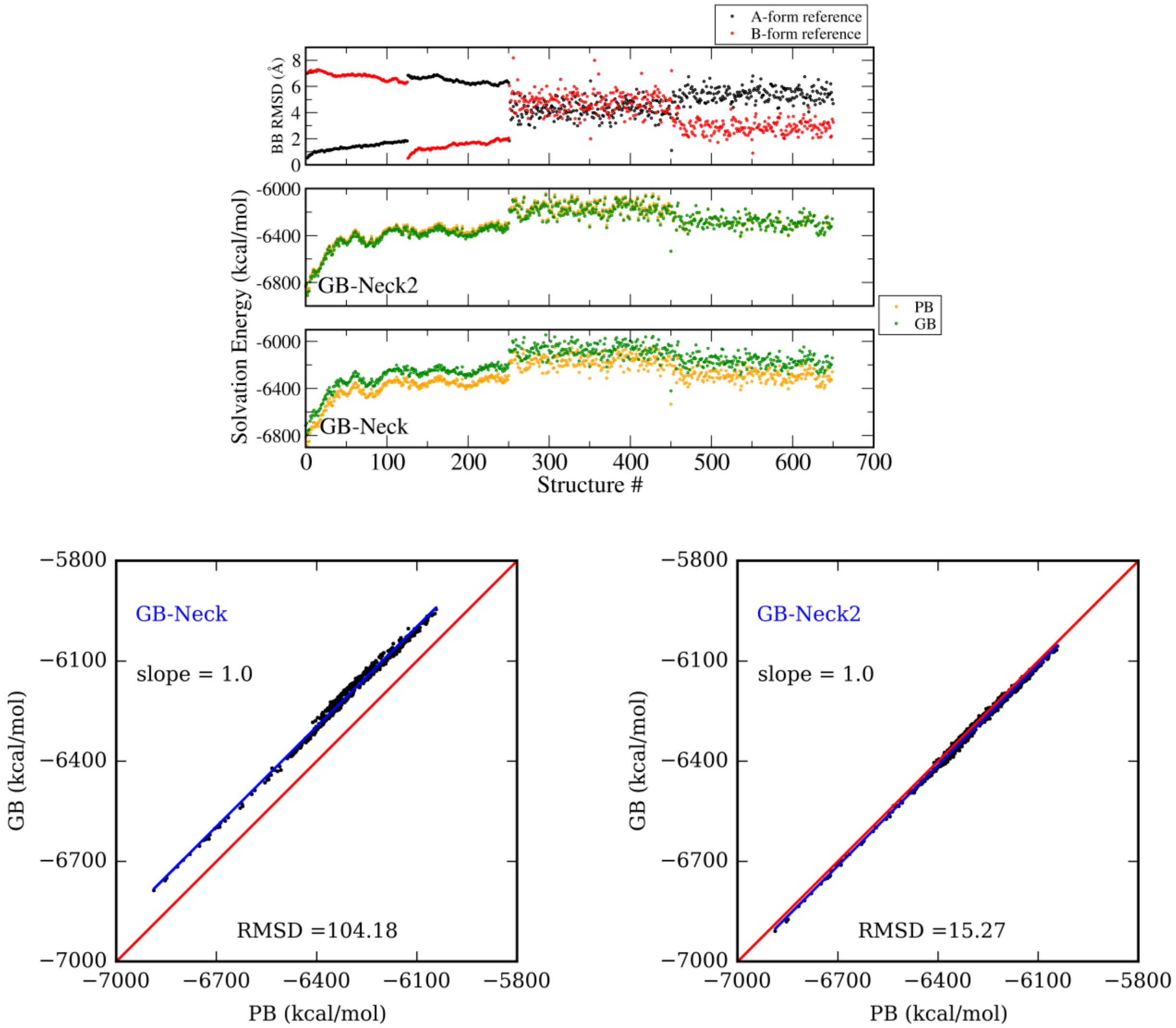

Figure S8. Comparison between GB and PB energies for DNA duplex (CGCGAATTCGCG)$_2$ **test set**. The first 450 structures come from GB-intermediate MD simulations and the last 200 structures were from a TIP3P MD simulation. **(Upper)**: Top panel shows the backbone RMSD of each structure to canonical A and B-form RNA. Middle and bottom panels compare GB and PB energies for individual structures in the training set: GB-Neck2 (middle) and GB-Neck (bottom). **(Lower):** The same data as in the upper figure, but with the GB and PB energies shown as a scatter plot.

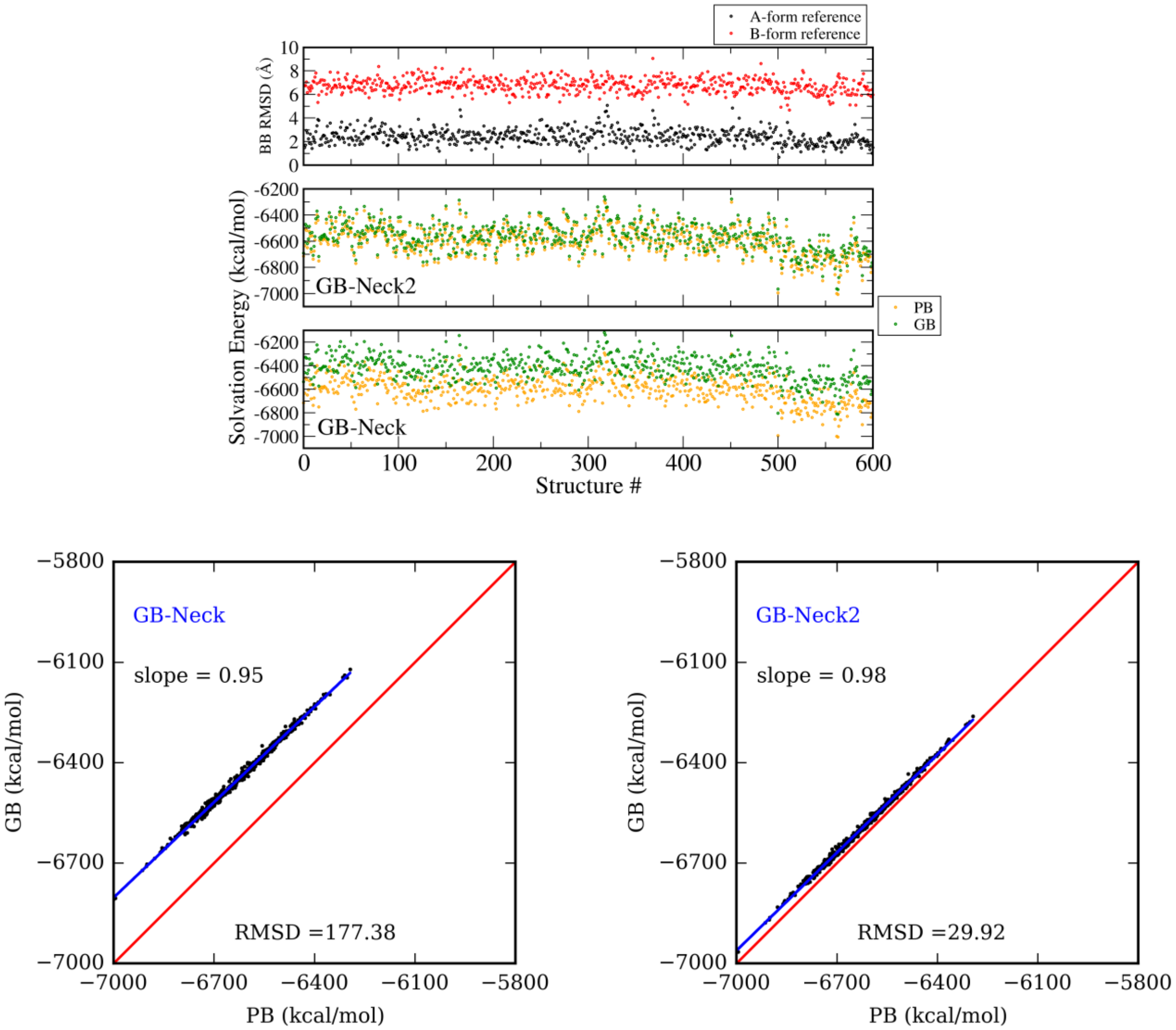

Figure S9. Comparison between GB and PB energies for RNA duplex (CGCGAAUUCGCG)$_2$ **test set**. The first 500 structures come from GB-intermediate MD simulations and the last 100 structures were from a TIP3P MD simulation. **(Upper)**: Top panel shows the backbone RMSD of each structure to canonical A and B-form RNA. Middle and bottom panels compare GB and PB energies for individual structures in the training set: GB-Neck2 (middle) and GB-Neck (bottom). **(Lower):** The same data as in the upper figure, but with the GB and PB energies shown as a scatter plot.

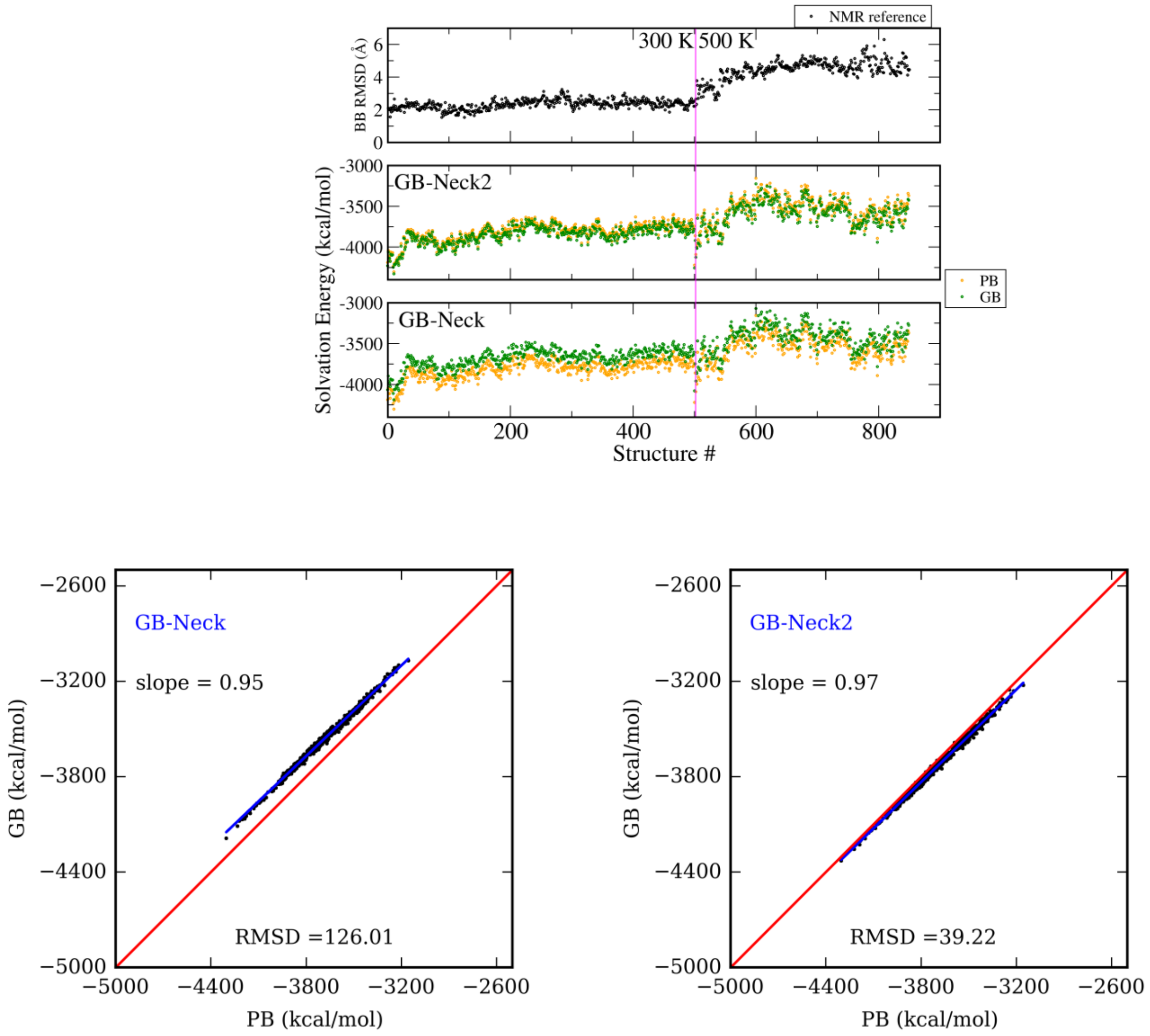

Figure S10. Structural and energetic analysis in the DNA/protein complex (PDB id: 1GCC)[3] **test set.** The first 500 structures come from a simulation in TIP3P explicit solvent at 300K; the last 350 structures were in TIP3P at 500K (high temperature MD was used to have more diverse structure conformations). **(Upper):** Top panel shows the backbone RMSD to the NMR structure. Flexible termini were excluded from the RMSD calculation (residues 23th to 26th and residues 75th to 85th in the complex). Middle and bottom panels compare GB and PB energies for individual structures in the training set: GB-Neck2 (middle) and GB-Neck (bottom). For the GB-Neck plot, GB-Neck parameters were applied for both protein and DNA. **(Lower):** The same data as the upper figure, but with the GB vs PB energies shown as a scatter plot.

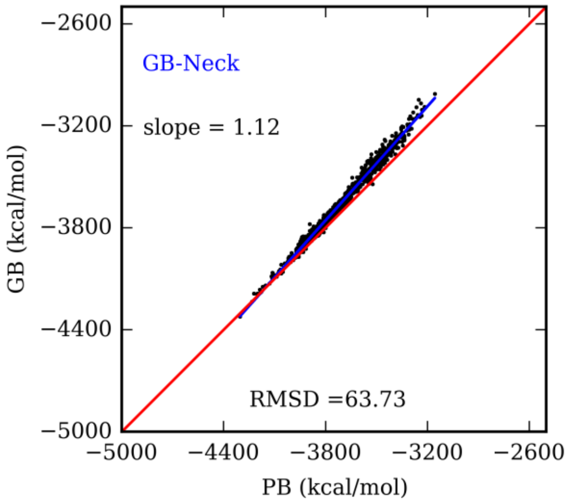

Figure S11. Comparison between GB and PB energies for DNA/protein complex (PDB id: 1GCC)[3] test set, using the same structures as shown in **Figure S10.** Here, for the "GB-Neck" plot, GB-Neck parameters were only applied for DNA while GB-Neck2 parameters[6] were applied to the protein.

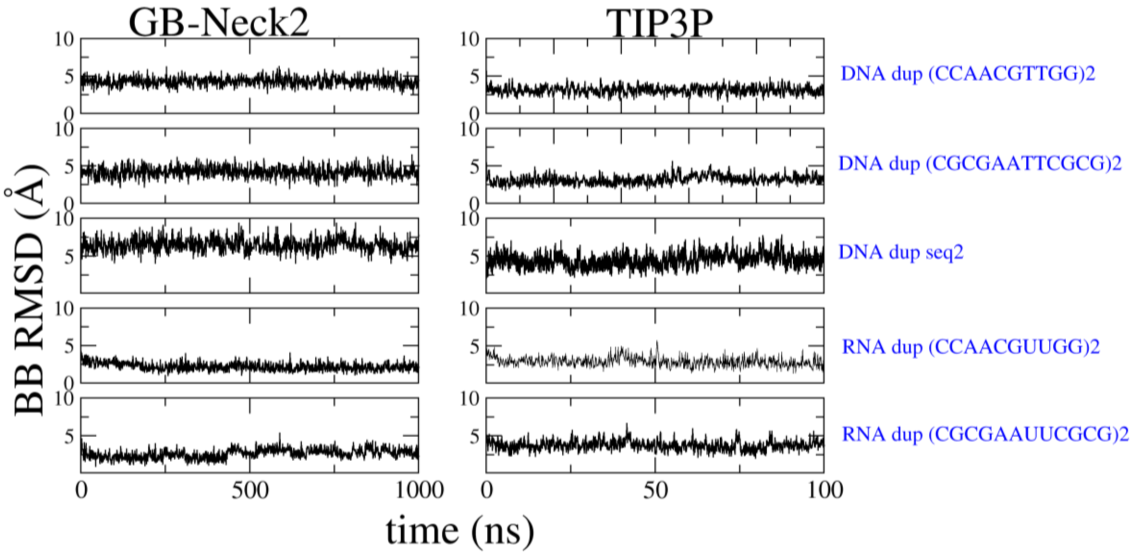

Figure S12. Backbone RMSD evolution of DNA and RNA duplexes for GB-Neck2 (left) and TIP3P (right) MD simulations. Stable structure in experiment was used as reference for RMSD calculation. For DNA duplexes, MD simulations started from A-form. For RNA duplexes, MD simulations started from B-form. Experimental structures were used as reference structure for RMSD calculation. GB-Neck2 MD simulations are 10-fold longer than TIP3P MD ones (1000 and 100 ns for GB-Neck2 and TIP3P, respectively). "DNA dup seq2" corresponds to DNA duplex $(CTAGGTGGATGACTCATT)_2$ and the TIP3P trajectory was taken from Pérez et al.[7]

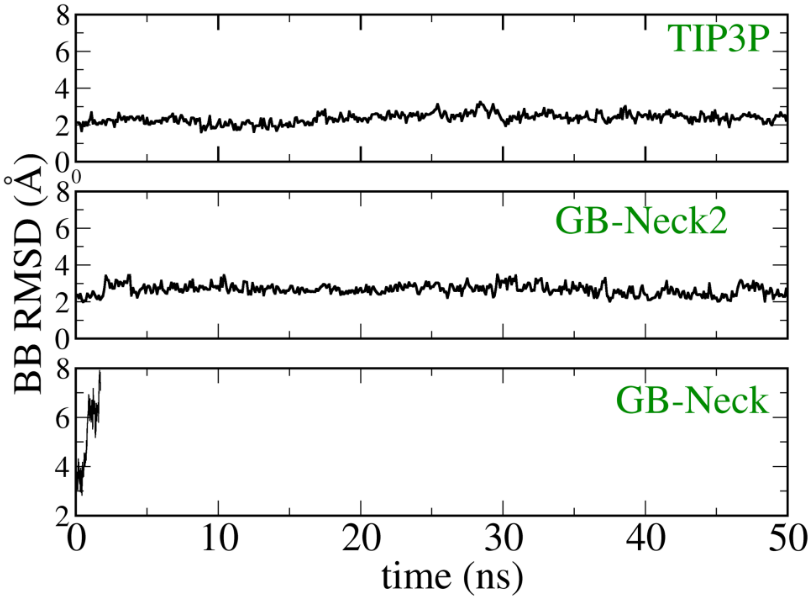

Figure S13. Backbone RMSD evolution of protein/DNA complex 1GCC.

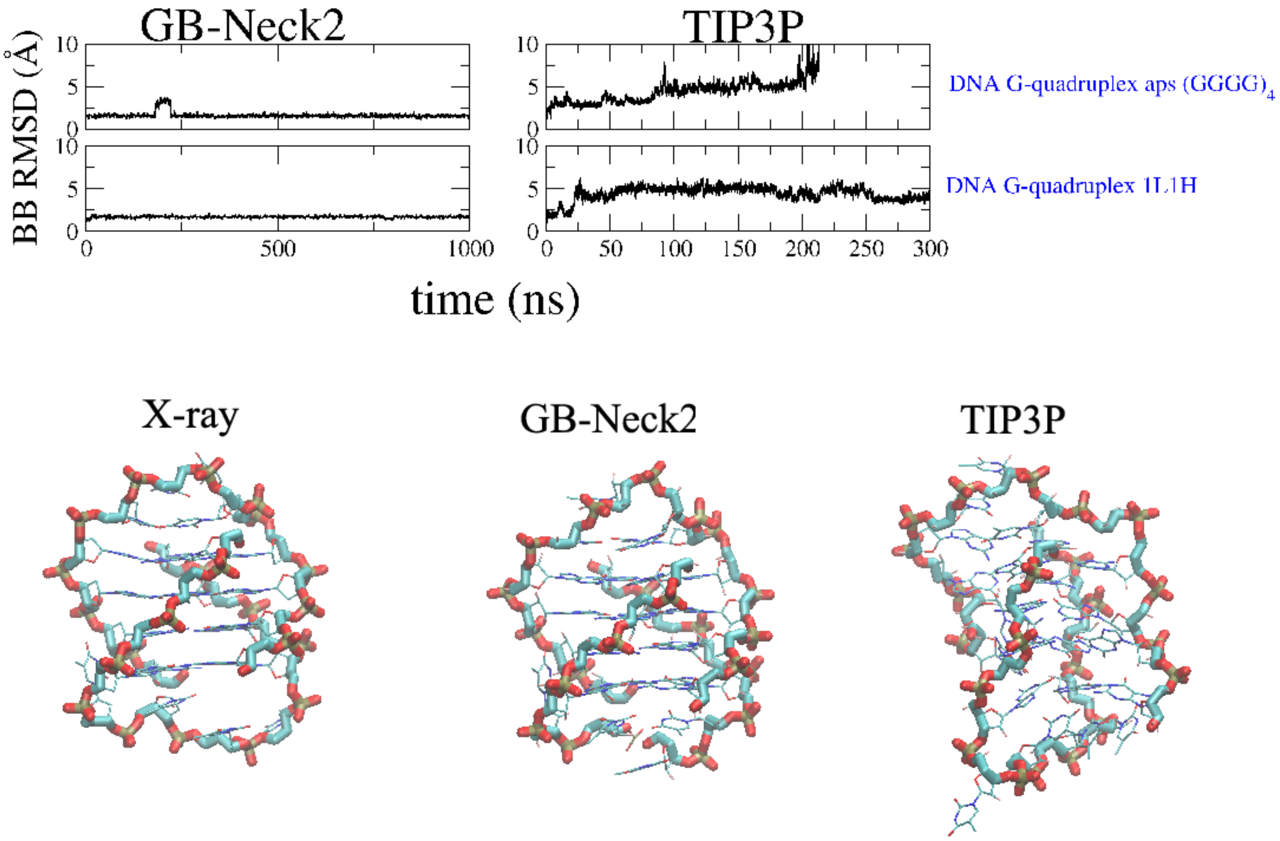

Figure S14. (Top) Backbone RMSD of two DNA quadruplexes for GB-Neck2 (left) and TIP3P (right) MD simulations. (Bottom) X-ray structure of DNA quadruplex 1L1H (PDB: 1L1H)[2] with the representative structure of the most populated cluster from GB-Neck2 (1000 ns) and TIP3P (300 ns without ions) MD simulations. Without salt, the TIP3P structure adopts a different, compacted structure; this is consistent with a previous study.[8]

References

(1) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham III, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817.

(2) Haider, S. M.; Parkinson, G. N.; Neidle, S. *J.Mol.Biol.* **2003**, *326*, 117.

(3) Allen, M. D.; Yamasaki, K.; Ohme Takagi, M.; Tateno, M.; Suzuki, M. *EMBO J.* **1998**, *17*, 5484.

(4) Zhu, L.; Chou, S.-H.; Xu, J.; Reid, B. R. *Nat Struct Biol.* **1995**, *2*, 1012.

(5) Nozinovic, S.; Fürtig, B.; Jonker, H. R. A.; Richter, C.; Schwalbe, H. *Nucl. Acids Res.* **2010**, *38*, 683.

(6) Nguyen, H.; Roe, D. R.; Simmerling, C. *J. Chem. Theory Comput.* **2013**, *9*, 2020.

(7) Pérez, A.; Lankas, F.; Luque, F. J.; Orozco, M. *Nucl. Acids Res.* **2008**, *36*, 2379.

(8) Stadlbauer, P.; Krepl, M.; Cheatham, T. E.; Koča, J.; Šponer, J. *Nucleic Acids Res* **2013**, *41*, 7128.

(9) Haider, S. M.; Parkinson, G. N.; Neidle, S. *J Mol Biol.* **2003**, *326*, 117.