Supporting Information

A Fast Pairwise Approximation of Solvent Accessible Surface Area for Implicit Solvent Simulations of Proteins on CPUs and GPUs

He Huang[1,2] and Carlos Simmerling*[,1,2]

1. Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York, 11794, United States

2. Department of Chemistry, Stony Brook University, Stony Brook, New York, 11794, United States

Telephone: (631)-632-5424

Email: carlos.simmerling@stonybrook.edu

The derivation of *shielded_SASA* calculation formula

When m=12, n=6, the van der Waals Lennard-Jones potential is expressed as below:

$$formula(vdw) = \varepsilon_{i,j}\left((\frac{\sigma_{i,j}}{R_{i,j}})^{12} - 2(\frac{\sigma_{i,j}}{R_{i,j}})^{6}\right) \qquad \text{(S1)}$$

But a more general form is:

$$formula(vdw\_like) = \varepsilon_{i,j}\left(\frac{\frac{n}{m-n}{\sigma_{i,j}}^{m}}{\left(R_{i,j}\right)^{m}} - \frac{\frac{m}{m-n}{\sigma_{i,j}}^{n}}{\left(R_{i,j}\right)^{n}}\right) \qquad \text{(S2)}$$

A, B, and C steps correspond to the transformations described in **Figure S1**:

$$formula(after\ A) = \ \varepsilon_{i,j}\left(\frac{\frac{n}{m-n}{\sigma_{i,j}}^{m}}{\left(-R_{i,j}\right)^{m}} - \frac{\frac{m}{m-n}{\sigma_{i,j}}^{n}}{\left(-R_{i,j}\right)^{n}}\right) \qquad \text{(S3)}$$

$$formula(after\ B) = \ \varepsilon_{i,j}\left(\frac{\frac{n}{m-n}{\sigma_{i,j}}^{m}}{\left(Cutoff_{i,j} + \ \sigma_{i,j} - R_{i,j}\right)^{m}} - \frac{\frac{m}{m-n}{\sigma_{i,j}}^{n}}{\left(Cutoff_{i,j} + \ \sigma_{i,j} - R_{i,j}\right)^{n}}\right) \qquad \text{(S4)}$$

$$formula(after\ C) = \ \varepsilon_{i,j}\left(\frac{\frac{n}{m-n}{\sigma_{i,j}}^{m}}{\left(Cutoff_{i,j} + \ \sigma_{i,j} - R_{i,j}\right)^{m}} - \frac{\frac{m}{m-n}{\sigma_{i,j}}^{n}}{\left(Cutoff_{i,j} + \ \sigma_{i,j} - R_{i,j}\right)^{n}}\right) + \ \varepsilon_{i,j} \qquad \text{(S5)}$$

Our pairwise formula transformed from vdw formula (same as **Equation 5** in the main text):

$$shielded\_SASA_{i,j} = \begin{cases} \varepsilon_{i,j}\left(\dfrac{\frac{n}{m-n}}{\left(1 + \frac{Cutoff_{i,j} - R_{i,j}}{\sigma_{i,j}}\right)^{m}} - \dfrac{\frac{m}{m-n}}{\left(1 + \frac{Cutoff_{i,j} - R_{i,j}}{\sigma_{i,j}}\right)^{n}}\right) + \ \varepsilon_{i,j}, & R_{i,j} < Cutoff_{i,j} \\ 0, & R_{i,j} \geq Cutoff_{i,j} \end{cases} \qquad \text{(S6)}$$

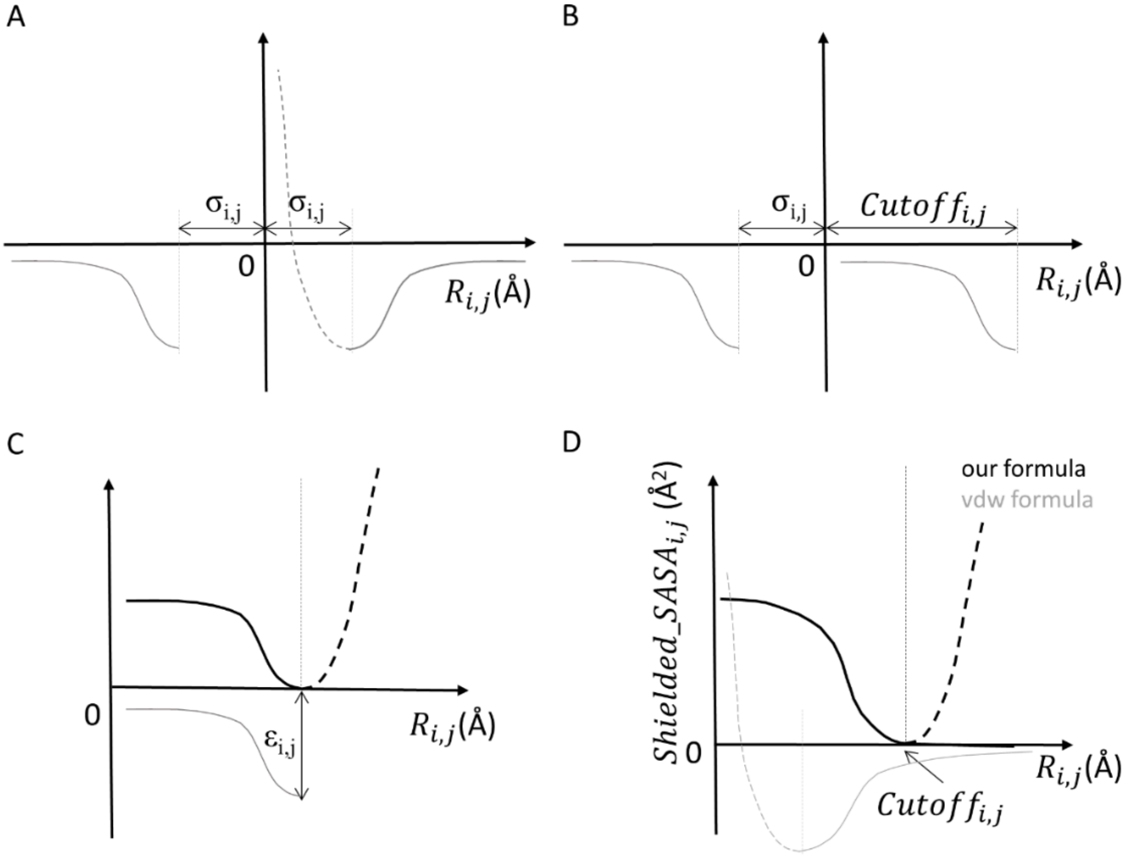


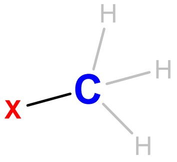
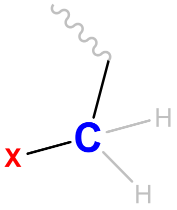
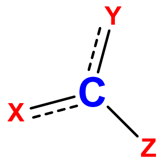
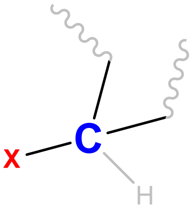
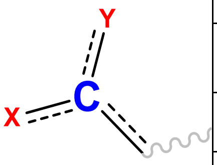
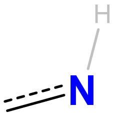
Figure S1. Transformation of our formula (**Equation S6**) from vdw function (**Equation S1**, more general form **Equation S2**) in schematic representations. (A) starting from vdw function (only the beyond vdw radius part is kept, shown in solid gray line on the right side of the y axis), to reflect it by y-axis results in **Equation S3**; (B) right shift it by $\sigma_{i,j}$ + Cutoff$_{i,j}$ results in **Equation S4**; (C) up shift the curve by $\varepsilon_{i,j}$ results in **Equation S5**, which is the (0, Cutoff$_{i,j}$) ; (D) a comparison of our final formula and vdw formula. Dashed lines represent the repulsive portion of the original function that was discarded in our transformation.

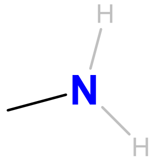
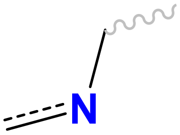
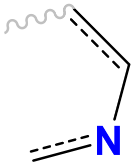
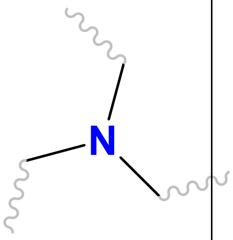
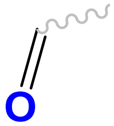
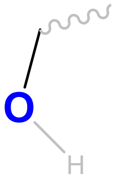
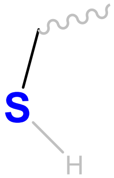
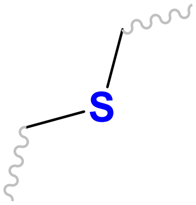
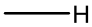
Definition of SASA types and parameters

We defined 30 atom types (just for SASA estimation, so we term them SASA types) based on one atom's bonded heavy atoms and hybridization state. The nomenclature system of SASA type is 1 digit followed by several capital letters. The digit indicates the category the central atom falls into, depending on how many heavy atoms are bonded to the central atom or the just the group index. For example, for Carbon atoms, we categorize their bonding environments into 5 groups, group 1 means the central atom is single bonded to one heavy atom, group 2 means it is bonded to two heavy atoms. As for group 3, it also contains two heavy atoms bonded central carbons, but instead of a sp3 hybridization as in group 2, the central carbon is double-bonded (or conjugated to) one or two heavy atoms, the change in bond angles changes the accessibility of the central carbons in group 3 compared to those in group 2. Furthermore, in group 4, three heavy atoms are bonded with the central carbon and in group 5, there are conjugated double bonds so that the central atoms in group 4 and 5 are categorized respectively.

Depending on the element type of the heavy atoms, the 5 groups are further divided into sub-groups. All the detailed division and definitions are included in Table S1 below.

Table S1. Definition of 30 SASA types and their occurrences in the training and test sets.

Element	Hybri-dizatio n	Generic formula	SASA type	Locations	# in training set	# in test set	Atom radius
Carbon	sp3	 1C-*X*	**1C**C	Ala side chain	80	444	1.7
			1CN	NME	10	0	
			1CS	Met side chain	10	18	
		 2CC-*X*	**2CC**C	Arg, Lys, Pro, Trp, Tyr, Phe, His side chain	220	914	
			2CCN	Arg, Lys, Gly, Pro side chain	40	220	
			2CCO	Ser side chain	10	36	
			2CCS	Cys, Met side chain	20	18	
	sp2	 3-*XYZ*	3CC	Tyr, Phe, Trp side chain	130	404	
			3CCC	Thr, Phe, Trp side chain	40	101	
			3CCN	His, Trp side chain	40	24	
			3CCO	Tyr side chain	10	29	
			3CNN	His side chain	30	12	
	sp3	 4CC-*X*	**4CC**C	Ile, Leu, Val side chain	30	155	
			4CCN	all backbone Cα except Gly	210	787	
			4CCO	Thr side chain	10	46	
	sp2	 5C-*XY*	**5C**CN 1	Trp side chain	10	12	
			5CCN 2	His side chain	30	12	
			5CNN	Arg side chain	10	44	
			5CNO	Backbone and Asn, Gln side chain carbonyl	240	915	
			5COO	Terminal carbonyl	20	133	
Nitrogen	sp3		**1NC**1	Arg, Asn, Gln side chain	40	165	1.55

			**1NC**2	Terminal amide, Lys	20	100	
	sp2, aliphatic		2NCC	Arg side chain, backbone amide	220	858	
	sp2, aromatic		3NCC	His side chain	70	36	
	sp3		4NCC	Pro backbone amide	10	24	
Oxygen	sp3		1OC1	Backbone and deprotonated carbonyl	280	1181	1.5
			1OC2	Ser, Thr side chain hydroxyl	20	82	
			1OC3	Tyr side chain hydroxyl	10	29	
Sulfur	sp3		1SC	reduced Cys	10	0	1.8
			2SCC	Met side chain, Cys in disulfide bonds	10	18	
Hydrogen	N/A		1H	all hydrogens	1780	6813	0*

*Zero radii are set for the Hydrogen atoms only for SASA calculations.

Table S2. Optimized (sigma and epsilon) and calculated (cutoff and max_SASA) parameters

SASA type	Cutoff (Å)	σ (Sigma)	ε(Epsilon)	Max SASA
1CC	3.1	4.370116	19.592480	89.5418522949
1CN		0.317054	7.148281	93.3032786658
1CS		1.208319	14.405034	101.172789338
2CCC		7.249659	18.793198	67.848137034
2CCN		5.492838	19.214204	78.9232674787
2CCO		1.568006	4.738056	62.3149039685
2CCS		3.856010	13.792617	72.0652500364
3CC		5.523807	13.179907	72.2402965204
3CCC		7.925637	0.700405	14.8528474421
3CCN		1.137401	4.914482	70.6469194565
3CCO		2.852471	0.690433	20.2921312847
3CNN		4.793021	17.786058	85.8858602953
4CCC		2.463345	1.586756	29.2869985239
4CCN		0.100000	0.328277	22.5608806495
4CCO		1.842881	1.600013	33.1334096093
5CCN1		3.532759	0.371097	16.5093987597
5CCN2		0.902828	0.021241	6.90712731848
5CNN		6.516442	3.681840	32.0850765017
5CNO		5.997082	0.739936	16.1244401843
5COO		9.776595	1.438099	16.6436516422
1NC1	2.95	3.008485	23.511977	94.0970695867
1NC2		4.290955	34.274575	95.1108847805
2NCC		3.296031	0.919202	22.8610751485
3NCC		5.589998	16.263937	64.7488801386
4NCC		1.0^{β}	0.000000^{Υ}	0.180783832809
1OC1	2.9	6.764858	12.670634	58.3692979586

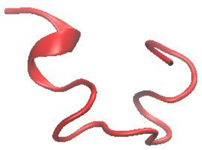
1OC2		2.827230	11.236117	69.8105657286
1OC3		2.827230	11.236117	80.1047057149
1SC	3.2	2.520362	16.788985	105.113824567
2SCC		1.133725	5.828670	75.8598197695
1H	1.4[α]	1.0 [β]	0. [ϒ]	0.

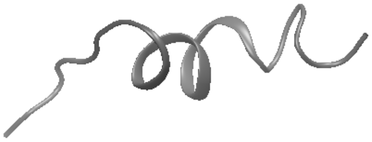
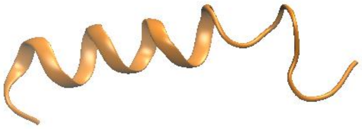
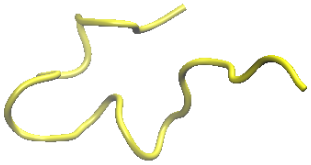
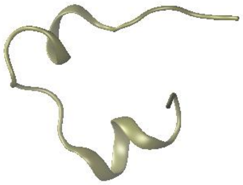
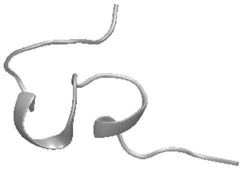
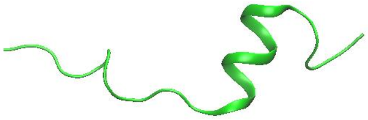
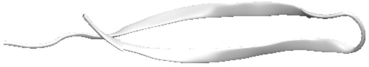
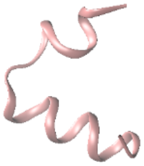
[α] Zero radii are set for the Hydrogen atoms, so the cutoff is always the probe of water radius 1.4 Å.
[β] Sigma = 1.0 for hydrogen and 4NCC are to make sure the denominator is not 0 in Equation 6.
[ϒ] Epsilon = 0 enforces zero contribution in shield SASA for all hydrogen and 4NCC involved atom pairs.

Description of parameter fitting and optimization

The initial parameters for σ were randomized in the range of 2.7-3.6 Å, and the range for ε was 1.5-1.9 Å^2. When we attempted the optimizations by varying functional forms, m and n values, the best performing parameter set was always saved and used as input for the next round of minimization; the ranges of initial parameters are not related to the final parameters. There were four rounds of optimization, each searching for best option for one thing. In the first two rounds, the functional form used was hyperbolic function, as was for the vdw dispersion energy. The objective function was molecular SASA and residual SASA. In the first round we used the hyperbolic functional form to optimize molecular SASA, starting at n=6 (same order as Lennard-Jones dispersion term); with the converged parameter set outcome, the second round of optimization was made of 6 runs optimizing molecular SASA by varying the n value from 1 to 6. We found n=3 with a cutoff at 12 Å or n=4 with no cutoff performed better than the others; thus we kept the two parameter sets, applied to MD simulations but found that we could not reproduce LCPO results. Then we changed to the current functional form, aiming to minimize the atomic SASA differences for another round. With the two sets from previous fitting, one of the resulting parameter sets assigned 0 to Cα atoms (4CCN SASA type), the simulation results were not as effective either. In the other set, Cα atoms contribute to pairwise *shield_SASA*, and LCPO could be reproduced effectively, so we chose this parameter set.

Table S3. Sequences and conformational features in scrambled peptide training set

scrambled sequence index	Sequence	Secondary Structure	Representative structure of largest cluster (percentage population)
1	RAH$^{\delta\varepsilon}$TH$^{\delta}$GYKMDNPEQIH$^{\varepsilon}$LFWCVS-NME	antiparallel, α-helix, coil	(17.2%)
2	RWMCDVAGIH$^{\varepsilon}$ENLTPH$^{\delta\varepsilon}$SKH$^{\delta}$QYF-NME	α-helix, coil	 (13.8%)

3	ENLVAFPITWYQ$H^{\delta}H^{\varepsilon}$RMCKDGS$H^{\delta\varepsilon}$-NME	α-helix, coil	 (45.4%)
4	NVWPEC$H^{\delta\varepsilon}$LQYDTIH^{ε}FH^{δ}ASKRGM-NME	α-helix, coil	 (38.4%)
5	FMIH^{δ}SE$H^{\delta\varepsilon}$CLWH^{ε}QANRKGTVDYP-NME	antiparallel, turn	 (11.7%)
6	FKH^{δ}A$H^{\delta\varepsilon}$ECQH^{ε}RGLIVPSMYNTDW-NME	α-helix, coil	 (19.1%)
7	YIKQPSDFVWLGTH^{ε}NAH^{δ}EMCR$H^{\delta\varepsilon}$-NME	α-helix, coil	 (23.2%)
8	LDKH^{ε}AG$H^{\delta\varepsilon}$VSREFIH^{δ}TWNQCMYP-NME	α-helix, coil	 (23.8%)
9	FH^{δ}RLQMDKEYNPSGAWI$H^{\delta\varepsilon}$TCVH^{ε}-NME	antiparallel, turn	 (82.3%)
10	EDKLH^{ε}ASRPH^{δ}WYV$H^{\delta\varepsilon}$CFMTQNGI-NME	α-helix, turn, coil	 (15.8%)

Note: H^{δ}, H^{ε}, $H^{\delta\varepsilon}$ are Histidine that is protonated at N^{δ}, N^{ε}, or both N^{δ} and N^{ε}, respectively. This training set has been developed in the experimental state of pairwise SASA algorithm, at that time hydrogens were considered in the SASA calculations; as we decided to exclude hydrogens in SASA estimation, the protonation states do not make any difference in the SASA fitting (but do impact the conformations sampled).

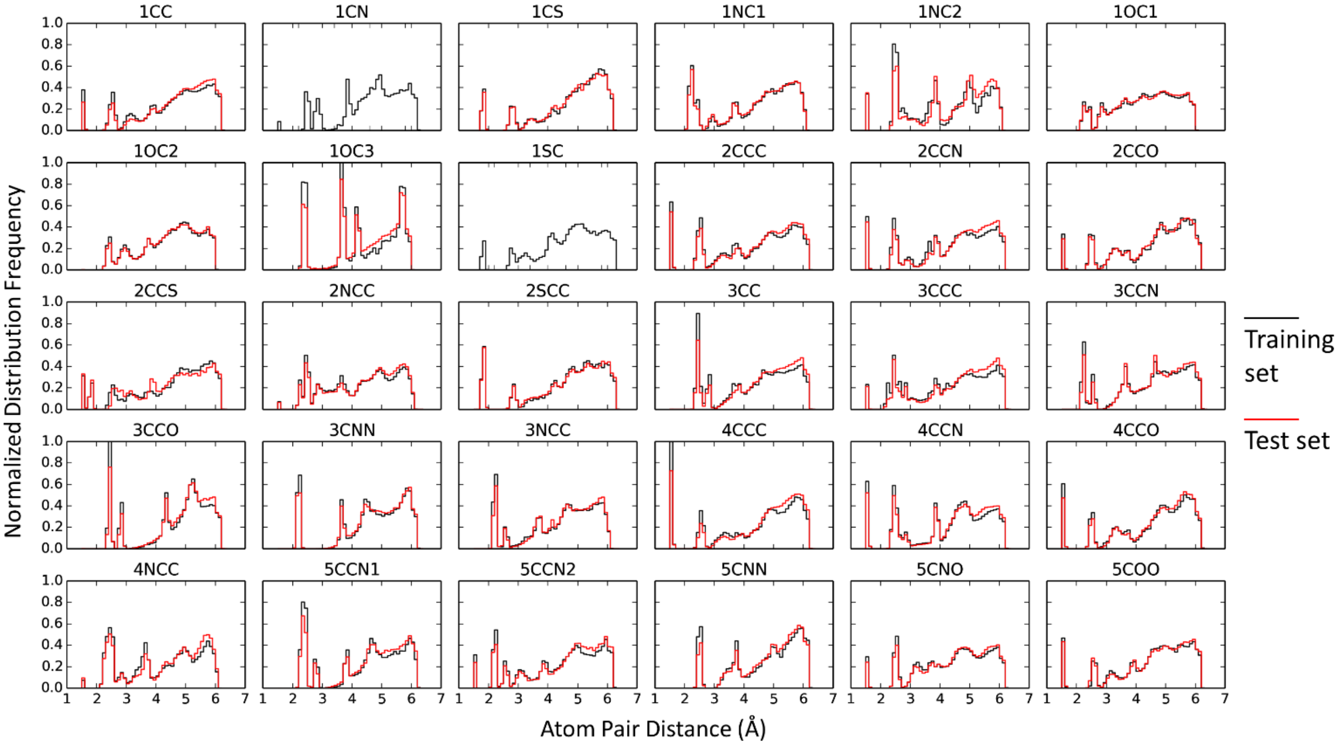


Figure S2. Distribution of pairwise atom distances within corresponding cutoffs for each SASA type for training set peptides and test set proteins

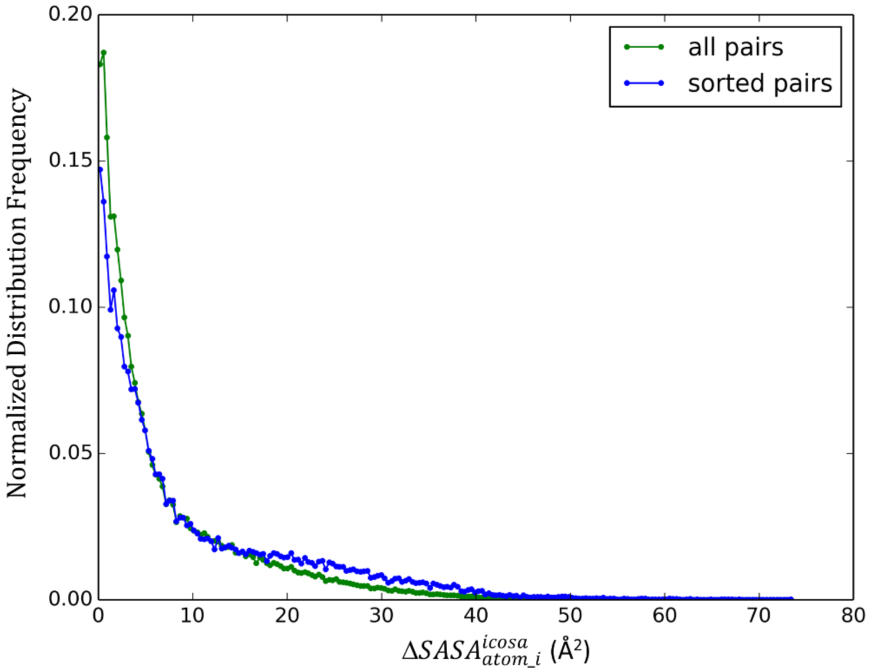


Figure S3. Normalized distribution of $\Delta SASA_{atom\_i}^{icosa}$ including all frame pairs (green) or sorted frame pairs (blue) for training set peptide atoms

Evaluation of energy conservation in MD

We carried out ~ 3.5 ns of constant energy MD simulations using HP36 at 300K starting from the equilibrated NMR structure. The time step was 0.5 fs and SHAKE was not used. Mixed precision[1] GPU version (pmemd.cuda_SPFP) of Amber 18 and our modified code were used for GB and pwSASA GB/SA simulations, respectively. For GB/SA simulations, surface tensions varied from 5, 10, 20 cal/mol/Å^2. The total energy deviations from the respective starting point were plotted to evaluate the energy conservation and force stability of the GPU code with and without the pwSASA calculation. As shown in **Figure S4** and **Table S4**, introduction of pwSASA results in no increase in the energy drift, not does it increase the standard deviation in the energy. As expected due to the continuous derivatives, we conclude that adding pwSASA introduces no additional significant force instability.

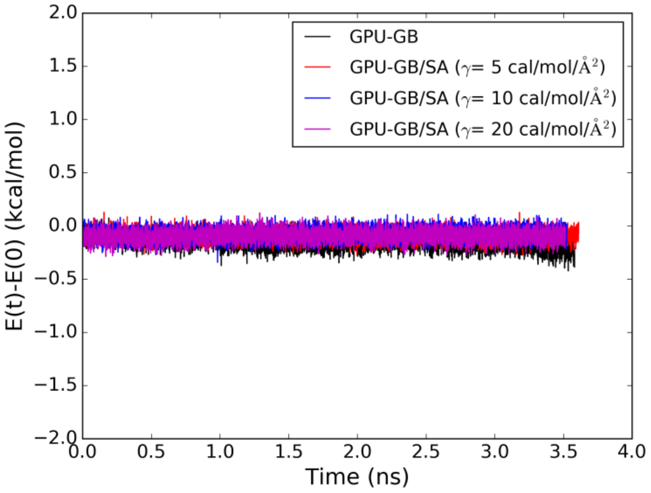


Figure S4. Energy deviation from initial energy, shown as a function of time during four simulations. Data are shown for standard GB calculations as well as pwSASA GB/SA with three different surface tension values.

Table S4. The averages and standard deviations of the total energies for constant energy simulations for HP36.

Surface Tension (cal/mol/Å^2)	Etot_avg (kcal/mol)	E(t)-E(0)_avg (kcal/mol)	Std (kcal/mol)
0 (GB)	-358.61056	-0.1543588	0.06529644
5	-341.85268	-0.0844846	0.05576449
10	-325.15303	-0.0728288	0.05765302
20	-291.78761	-0.0906127	0.05853439

Table S5. Temperature ladders for all REMD simulations.

System	Solvent model	REMD temperatures (K)
HC16	TIP3P	266.7, 270.2, 273.8, 277.4, 281.0, 284.7, 288.5, 292.3, 296.1, 300.0, 303.9, 307.9, 312.0, 316.1, 320.3, 324.5,

		328.8, 333.1, 337.5, 341.9, 346.4, 351.0, 355.6, 360.3, 365.0, 369.8, 374.7, 379.6, 384.6, 389.7, 394.8, 400.0
	GB/SA (pwSASA, LCPO)	279.5, 300.0, 321.9, 345.5, 370.8, 397.9
CLN025	GBNeck2, GB/SA (pwSASA, LCPO)	252.3, 275.1, 300.0, 327.2, 356.8, 389.1
Trp-cage		247.7, 264.0, 281.4, 300.0, 319.8, 340.9, 363.3, 387.3
HP36		250.0, 262.2, 275.0, 288.4, 300.0, 317.3, 332.8, 349.0
Homeodomain	GBNeck2, GB/SA (pwSASA)	288.7, 300.0, 311.7, 323.9, 336.6, 349.8, 363.5, 377.7, 392.4, 407.8, 423.8, 440.3

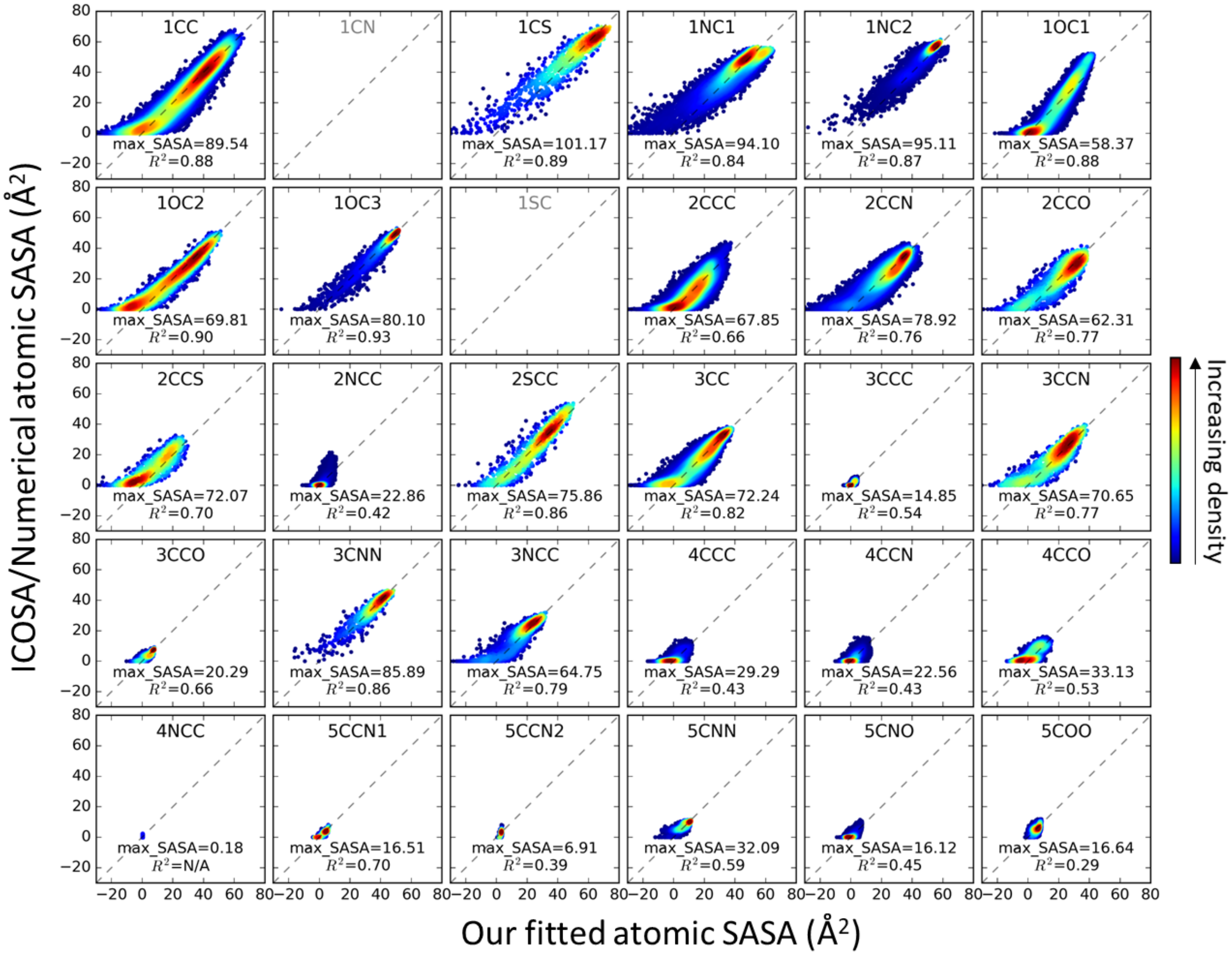


Figure S5. 2D histograms of pwSASA fitted atomic SASA of each SASA type versus ICOSA numerical values for the test set. Perfect agreement is shown by the diagonal dashed lines. The color indicates the kernel density estimated using scipy gaussian_kde.

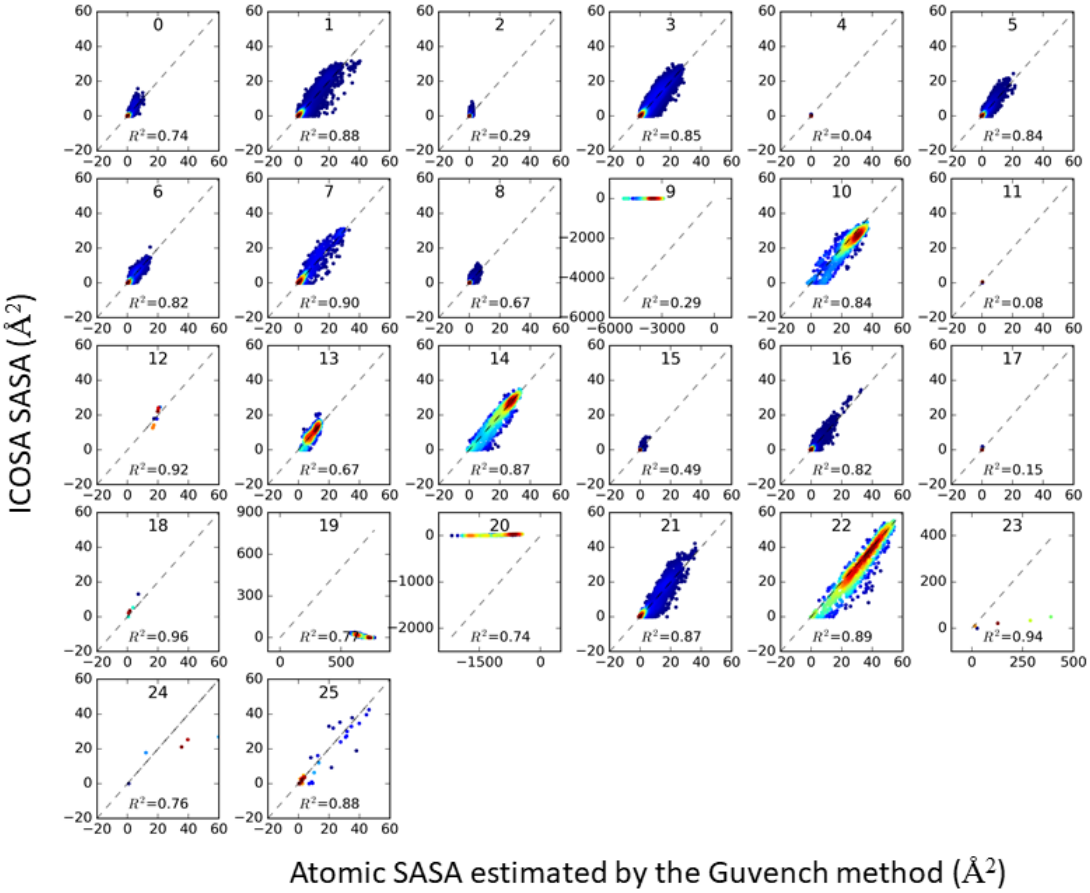


Figure S6. 2D histograms of fitted atomic SASA of each SASA type using Guvench et al.'s method, parameters and atom types[2] versus ICOSA numerical values for the test set. Perfect agreement is shown by the diagonal dashed lines. Significant deviation from the diagonal is seen for 3 of the atom types (Type 9: N sp3 bonded to 1 heavy and 3 hydrogens, Type 19: O sp3 bonded to 1 heavy and 1 H, Type 20: H connected with type 19 O). The color indicates the kernel density estimated using scipy gaussian_kde.

Re-optimization of parameters for the Guvench et al. SASA estimator[2]

The training set was composed of four proteins (PDBID: 1APS, 1COA, 1LMB and 1CSP) as described in Guvench et al's Parameterization section[2]. The 20 conformations were generated following their preparation protocol except we used the same simulation methods as described in the Methods section of this article, including the Amber software, ff14SBonlysc force field and GBSA (GBNeck2, LCPO with a surface tension of 5 cal/mol/Å^2). For each protein, 5 conformations were included. One of the five conformations was the last frame of 600 ps of backbone positional restrained MD simulation at 298K (force constant of 5 kcal/mol/ Å^2). The other four conformations were the last frames of 600 ps unrestrained MD simulations at 298, 400, 600, and 800 K, respectively. The van der Waals radii in their Table I and radii of 1.6 Angstrom for solvent probe were used as described in their report[2] to calculate A_i values for each atom. We then used non-linear least squares (curve_fit in scipy) to determine the set of c values. The boundaries of c0 values were set between -123 and 131, c2 values were bounded to be positive and c3 to be negative, to match to the reported[2] parameters. Our newly fit parameters are shown in **Table S6**. We also used the atomic A_i values to generate the 4th-order polynomials for each atom type, shown in **Figure S7**. Consistent with the comparison of estimated and exact SASA values in **Figure S6**, the same three atom types (9,19,20) show a significant deviation between the actual atomic SASA values and the polynomial curve. When our refit c_i parameters (**Table S6**) are used, the polynomial curves show much better fit to the data. The deviation of the two polynomials outside the data points for nearly all atom types suggests the possibility for training set sensitivity of the c_i parameters.

Table S6. Atom types and newly refit parameters c_k for the training set of Guvench et al. The "# in set (reported) " column in red is taken from their publication[2], while the "# in set" is the number of atoms of that type that we calculated for the four proteins. The c_i parameters in the table are obtained from our refitting procedure.

Atom	Atom Type Index	# in set (reported)	# in set	c_0	c_1	c_2	c_3	c_4
C sp^3	0	920	1200	-123	-41.4634	66.4298	-17.2423	1.32261
H	1	2760	3600	-123	-38.2899	105.021	-34.4741	3.2005
C sp^3	2	2165	2820	-53.6689	26.5466	0.0000	-1.3691	0.148054
H	3	4330	5640	-123	48.1305	48.1953	-21.877	2.26138
C sp^2	4	1865	2405	1.51508	-0.467078	0.0261628	0.0000	0.00028649
H	5	1865	2405	131	10.3734	0.0000	-5.98703	0.906069
C sp^2	6	640	770	131	-180.416	91.6942	-18.7412	1.31759
H	7	640	770	-123	96.2129	10.4313	-12.5749	1.52784
C sp^2	8	2240	2870	-24.2353	21.1903	0.0000	-1.54218	0.179456
N sp^3	9	150	205	-52.3656	26.6936	0.0000	-1.52838	0.175802
H	10	450	615	-123	-100.595	169.269	-54.4652	5.14834
N sp^3	11	5	5	-20.95	-14.8988	4.90531	0.0000	-0.0229411
H	12	10	10	-122.751	-209.62	115.064	0.0000	-3.31968
N sp^2	13	285	360	96.9616	-22.7412	0.0000	0.0000	0.0277059
H	14	570	720	-123	-91.3066	153.69	-48.9391	4.58758
N sp^2	15	1615	2075	131	-52.5365	5.46968	0.0000	-0.0079142
H	16	1615	2075	131	-8.64628	0.0000	-3.41924	0.543425
N sp^2	17	45	55	38.9238	-10.1705	0.0491014	0.0000	0.0172401
N sp^2	18	5	5	131	-12.1068	0.0000	-2.4721	0.383666
O sp^3	19	225	295	131	-226.634	138.874	-31.8983	2.43034
H	20	225	295	-123	17.3263	63.3762	-24.3485	2.39696
O sp^2	21	1695	2015	-123	93.135	24.9148	-15.9794	1.65768
O carboxylate	22	490	800	130.997	-150.453	107.684	-28.6736	2.4228
S	23	5	5	131	33.3752	0.0000	-6.90118	0.912177
H	24	5	5	131	-77.5976	17.6689	0.0000	-0.335062
S	25	40	55	-123	-299.889	261.799	-63.2109	4.77654

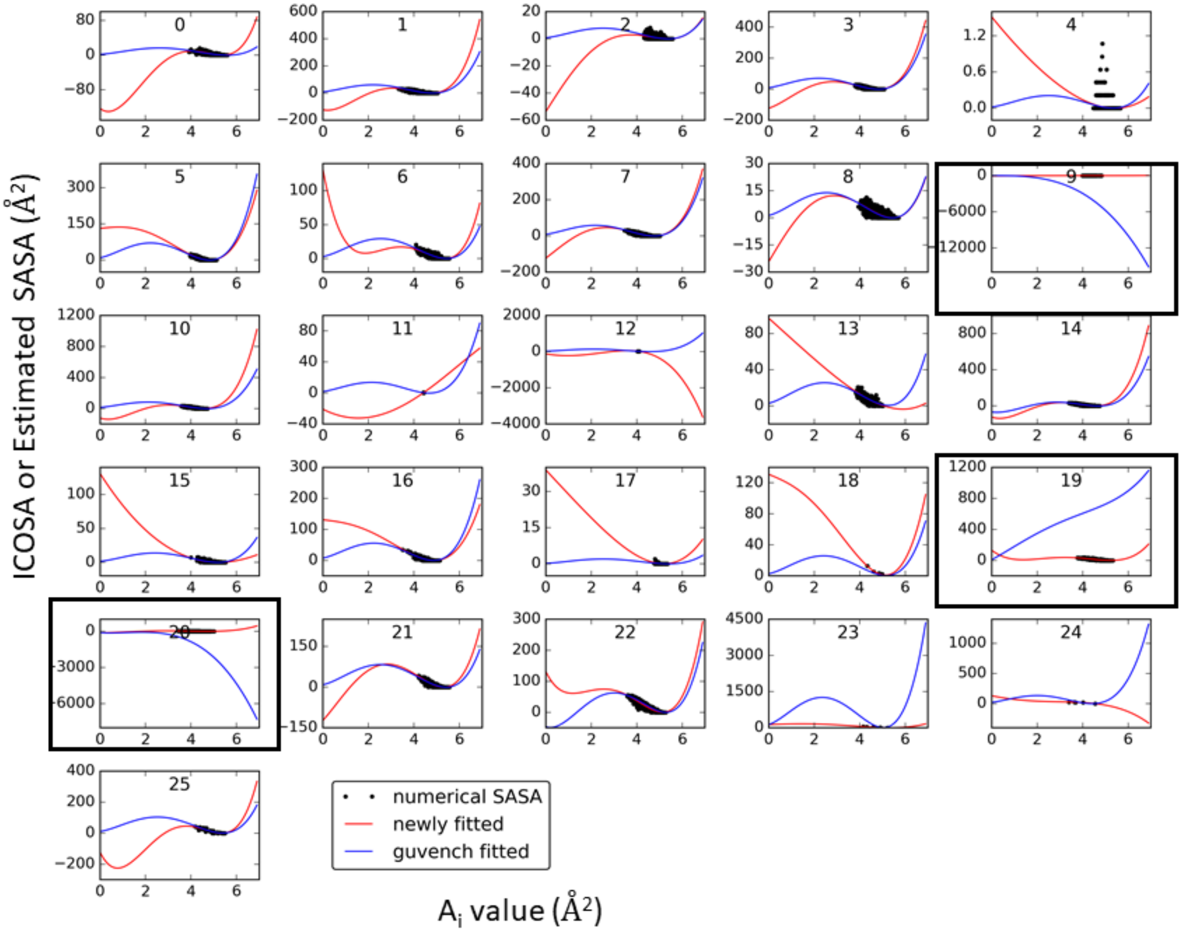


Figure S7. Estimated atomic SASA values from three sources, with 1 plot shown for each atom typeas defined by Guvench et al.[2] Data are shown for the four protein systems described above. For the black dots, the X axis is the A_i value for each atom calculated using the Guvench et al. approach[2] and the Y value is the ICOSA-calculated atomic SASA. For the red and blue curves, the X axis represents input A_i values for the Guvench polynomial SASA estimator, with the estimated SASA on the Y axis. Blue curves correspond to the estimated atomic SASA using the published parameters[2], while the red curve represents the estimation using our refit parameters (**Table S6**). Black boxes surround the three atom types where the polynomial curve poorly estimates the actual SASA values for atoms with these A_i values (black dots); all three are significantly improved using our refit parameters.

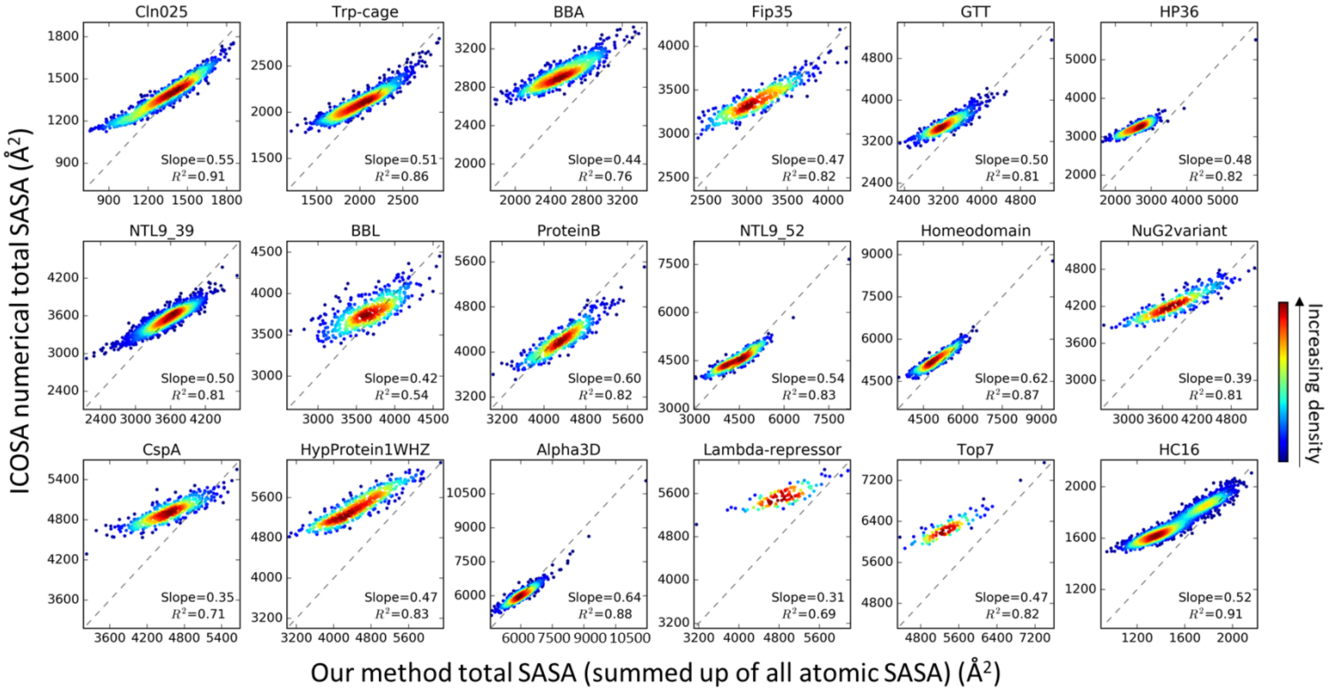


Figure S8. Deviation of sum of atomic SASA from the numerical SASA, represented in 2D histograms of total SASA versus ICOSA numerical values for the test set.

The transformation to get molecular SASA from atomic SASA

We observed that the sums of estimated atomic SASA values systematically deviate from the numerical molecular values (**Figure S8)** due to the occurrence of negative SASAs and other inaccuracies in compact conformations compared to the extended ones. We see the larger molecular SASA (more extended conformation) values are closer to the diagonal dashed line indicating perfect agreement, while the smaller SASA (more compact conformation) data points are underestimated. These systematic deviations result in overestimation in the molecular SASA changes as conformation changes, indicated by a universal decrease in the slope of estimated vs. numerical SASA. We further adjusted it by a scaling factor that reduces the estimated SASA changes, followed by a linear regression that finds a best fit total *adjusted_max_SASA* for each system through a common formula (**Figure S9)**. As we have found that a common scaling factor of 0.6 multiplied to directly summed atomic SASA would improve the slopes in **Figure S8**, we then used the following transformation:

$$SASA_{molecule}^{icosa} = offset\ - 0.6 \times SASA_{molecule}^{estimated} \tag{S7}$$

$$SASA_{molecule}^{estimated}\ =\ \sum_{i}^{natoms} (max\_SASA_i - shielded\_SASA_i) \tag{S8}$$

where an offset is needed for the scaled molecular $SASA_{atom\_i}^{estimated}$ to match the ICOSA molecular SASA values ($SASA_{molecule}^{icosa}$) for each protein system in the test set, respectively.

One last transformation is done for an analytical fitting formula to calculate this offset, which compensates the difference of $SASA_{molecule}^{icosa}$ and the scaled *shielded_SASA$_i$* sums, termed *adjusted_max_SASA* **(Equation S9)**.

$$SASA_{molecule}^{icosa} = adjusted\_\mathrm{max}\_SASA + 0.6 \sum_{i}^{natoms} shield\_SASA_i \tag{S9}$$

For each of the 18 proteins, we calculated the required *adjusted_max_SASA* from the numerical molecular SASA and scaled molecular *shielded_SASA* using similar fashion as the calculations for the SASA type specific *max_SASA*.

$$adjusted\_max\_SASA_{protein\_i} = \frac{1}{N} \sum_{conformations}^{N} (SASA_{protein\_i}^{icosa} + 0.6 * shielded\_SASA_{protein\_i}^{estimated}) \tag{S10}$$

The last step was to plot the molecular *max_SASA* vs. *adjusted_max_SASA* for a linear regression, which results in a nearly perfect linear correlation of R^2=0.9995 (**Figure S9**). Therefore, it is reasonable to calculate *adjusted_max_SASA* from *max_SASA* using **Equation S7**.

$$adjusted\_\mathrm{max}\_SASA = 0.6814 \sum_{i}^{natoms} max\_SASA_i + 361.1 \tag{S11}$$

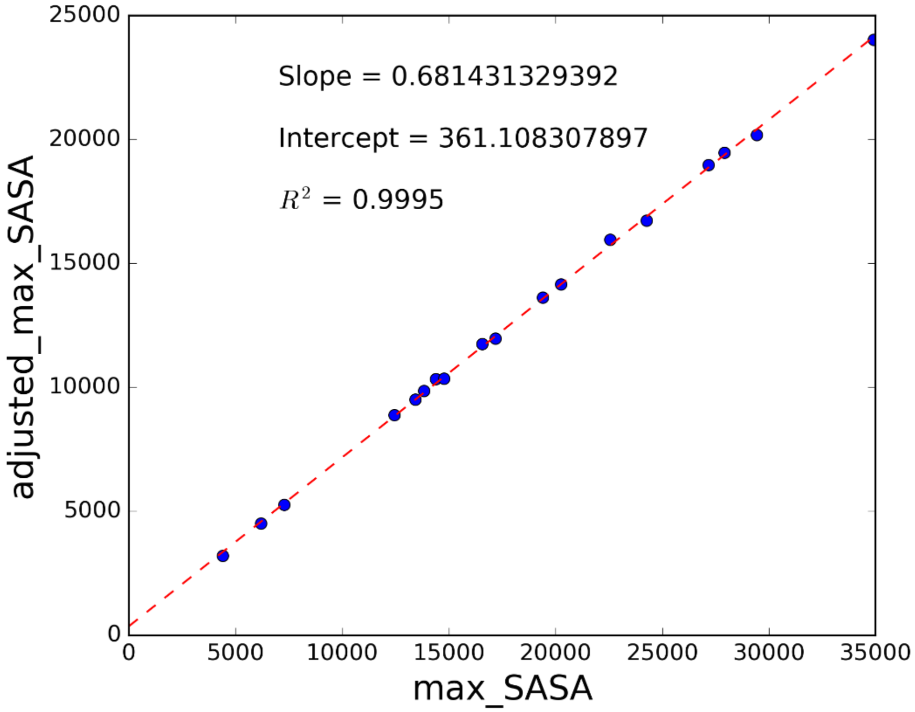


Figure S9. Transformation of *max_SASA* to *adjusted_max_SASA* by linear regression. Each data point corresponds to a protein in the test set.

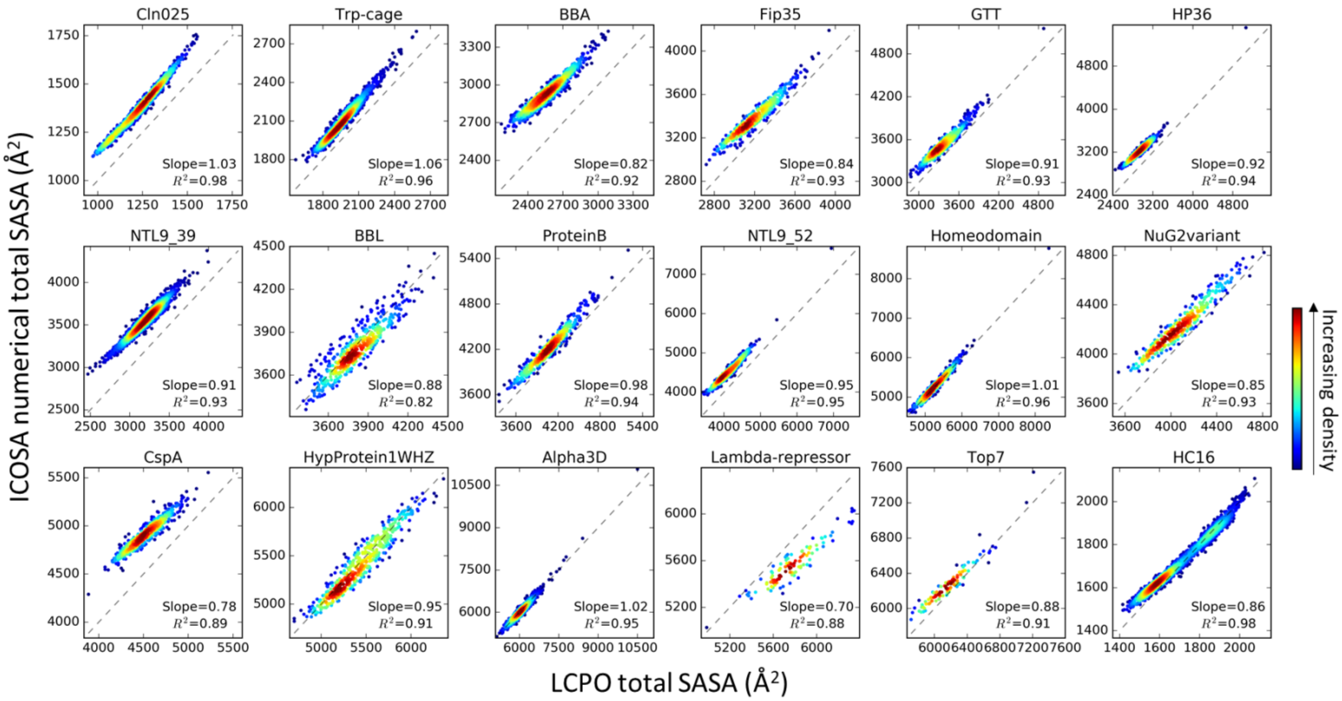


Figure S10. 2D histograms of LCPO fitted molecular SASA of each SASA type versus ICOSA numerical values for the test set.

Table S7. The properties of the top 3 cluster representative structures from cluster analysis $^{\alpha}$ on HC16 GB simulations (300K) and their occurrences in other simulations (300K trajectories from GB/SA and TIP3P)

Clustering and occurrence Analysis		Cluster 1	Cluster 2	Cluster 3
Representative structure Cα-RMSD (Å)		1.00	4.12	5.37
Representative structure ICOSA SASA (Å^2)		1686.4	1911.7	1813.9
Representative structure SASA-based (γ=7cal/mol· Å^2) nonpolar energy (kcal/mol)		11.8	13.4	12.7
Cluster population in GB (%)		57.0 (0.1)	14.9 (0.1)	5.3 (0.1)
Occurrence$^{\beta}$ in each trajectory at 300K (%)	GB	57.9 (1.6)	15.0 (0.9)	4.9 (0.2)
	GB/SA: pwSASA	91.4 (0.1)	1.8 (0.1)	3.0 (0.1)
	GB/SA: LCPO	93.2 (1.2)	1.4 (0.1)	2.3 (0.2)
	TIP3P	95.7 (1.0)	0.1 (0.1)	0.2 (0.1)

[α] For clustering analysis done on GB trajectories, 16000 frames in total were evenly obtained from the last halves of the two MD simulations starting from different initial structures. The clustering criterion is pairwise RMSDs based on all Cα atoms, using bottom-up aggregating average linkage algorithm, with centroid distances < 2.0 Å.

[β] The occurrence of certain cluster in GB/SA (γ=7cal/mol· Å^{2}) or TIP3P trajectories was measured by the number of conformations that are < 2.0 Å (Cα-RMSD) from the representative structure of this cluster, divided by the total frame number of the whole simulated trajectory at 300K.

Table S8. Cluster analysis [α] for HP36 combined trajectory at 250K and occurrences of the top 7 cluster representative structures in the four 300K trajectories, respectively

Clustering and occurrence Analysis		Cluster 1	Cluster 2	Cluster 3	c4	c5	c6	c7
Representative structure Cα-RMSD (Å)		2.40	7.68	3.03	4.70	6.17	5.47	7.21
Representative structure Cα-RMSD on structured region 3-32 (Å)		1.57	6.87	2.57	3.83	5.42	4.55	6.23
Average Cα-RMSD on structured region 3-32 (Å)		1.96 (0.55)	6.79 (0.14)	2.71 (0.32)	3.74 (0.27)	5.56 (0.23)	4.57 (0.15)	6.24 (0.16)
Average ICOSA SASA (Å^{2})		3155.7 (105.8)	2914.9 (132.4)	3195.8 (113.3)	3365.7 (85.4)	3320.0 (132.1)	3376.7 (124.0)	3230.2 (110.6)
SASA-based (γ=7cal/mol· Å^{2}) nonpolar energy (kcal/mol)		**22.1 (0.7)**	**20.4 (0.9)**	22.4 (0.8)	23.6 (0.6)	23.2 (0.9)	23.6 (0.9)	22.6 (0.8)
Fraction (cluster population) in each trajectory at 250K (%)	GB	30.4 (10.4)	11.6 (5.9)	18.3 (0.2)	6.0 (6.0)	5.3 (5.3)	0.0 (0.0)	0.0 (0.0)
	GB/SA: pwSASA	8.7 (1.3)	75.9 (0.3)	2.8 (0.1)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.1)
	GB (14sb)	35.7 (1.8)	4.3 (2.5)	4.4 (0.8)	4.7 (4.8)	3.4 (3.4)	8.9 (7.1)	1.7 (1.1)
	GB/SA: pwSASA (14sb)	76.3 (2.6)	2.1 (1.8)	6.3 (0.2)	0.2 (0.2)	1.2 (1.2)	0.1 (0.1)	2.6 (0.1)
Occurrence [β] in each trajectory at 300K (%)	GB	1.4 (0.2)	0.7 (0.5)	1.7 (0.3)	0.6 (0.2)	0.2 (0.2)	0.1 (0.1)	0.0 (0.0)
	GB/SA: pwSASA	18.1 (0.4)	26.9 (1.6)	14.1 (0.6)	0.4 (0.1)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
	GB (14sb)	2.8 (0.4)	0.4 (0.3)	2.2 (0.4)	0.7 (0.7)	0.2 (0.2)	1.6 (0.3)	0.3 (0.2)
	GB/SA: pwSASA (14sb)	47.4 (0.3)	1.6 (1.4)	30.6 (0.2)	0.5 (0.5)	1.4 (1.4)	0.5 (0.2)	3.1 (0.1)

[α] Clustering analysis was done on combined trajectory of the four (GB, GB/SA: pwSASA, GB(14sb), GB/SA: pwSASA (14sb)) methods. In total 40,000 frames (10,000 frames from each trajectory) are evenly obtained from

250K trajectories, clustering criterion is pairwise RMSDs based on structured region (residue 3 to 32 Cα atoms), using bottom-up aggregating average linkage algorithm, with centroid distances < 2.0 Å.

[β] The occurrences are measured in the similar fashion as in Table S7, i.e. all frames < 2.0 Å (Cα-RMSD in region 3-32) from the representative structure of this cluster, divided by the total frame number of the whole simulated trajectory at 300K.

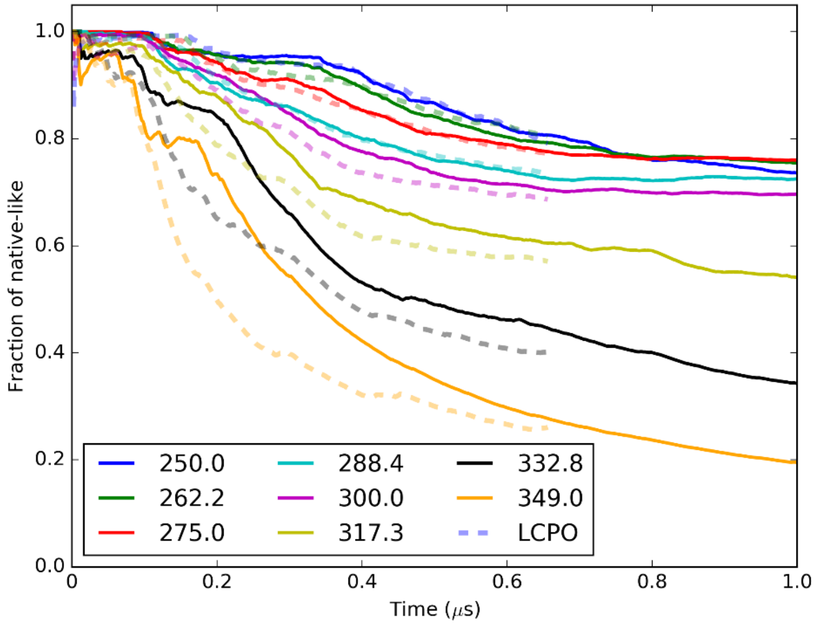


Figure S11. Reduction of HP36 NMR-like structures in all temperature trajectories observed in two GB/SA solvent simulations using ff14SBonlysc: pwSASA (solid lines) and LCPO (dashed lines). Both REMD simulations started from NMR structure. The fraction of folded is calculated on conformations < 3.5 Å Cα-RMSD excluding flexible termini. Only the first 1 μs of data is shown.

1. Le Grand, S.; Gotz, A. W.; Walker, R. C., SPFP: Speed without compromise-A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications* **2013,** *184* (2), 374-380.

2. Guvench, O.; Brooks, C. L., Efficient approximate all-atom solvent accessible surface area method parameterized for folded and denatured protein conformations. *J Comput Chem* **2004,** *25* (8), 1005-1014.