Relative binding enthalpies from molecular dynamics simulations using a direct method

Amitava Roy,[‡1] *Duy P. Hua,*[‡1] *Joshua M. Ward,*[§1] *and Carol Beth Post*[*‡]

[‡]Department of Medicinal Chemistry, Markey Center for Structural Biology, and Purdue Center for Cancer Research, Purdue University, West Lafayette, IN 47907

[§]*current address: Department of Chemistry, University of Oulu, PO Box 3000, FIN-90014 Oulu, Finland*

[1]**These authors contributed equally to this work.**

Supporting Information

Local Mean Potential Energy, the cumulative average and limiting slopes

The averaged value as a function of the production time per simulation used for averaging is one indicator for the convergence of the estimated value. This cumulative average for the local potential energy $\bar{E}_k^T$ from individual trajectories of the complexes and peptides is shown in figure S1 as a function of the time for averaging. Results for the forty simulations are plotted in one panel. Limiting slopes determined from the time period of 8 to 10 ns are listed in Table S1. Initial fluctuations in the cumulative average of each simulation largely subside after 2 to 4 ns of simulation time, yet there is a range in the final $\bar{E}_k^T$ and the limiting slopes are considerably larger than that of the ensemble-averaged plot (figure 2 and table 2 in main text).

Convergence of the statistical uncertainty in the ensemble-averaged mean estimates for the various energy terms

The dependence of the statistical certainty of the energy estimates on simulation time was determined using the bootstrap method[40] (see Methods in main text). The decrease of the 95% CI for $\left\langle E^T \right\rangle$ is shown in Figure S2 for Src SH2-cpYEEI and SH2-fpYEEI, and S3 for the three unbound peptide ligands. Figure S4 and S5 are 95% CI for the component energy terms for solute-solute interactions, E^{UU}, solute-solvent interactions, E^{UV}, and solvent-solvent interactions, E^{VV}, for the complexes and unbound peptides, respectively. These figures are 2-dimensional plots with 95% CI as a function of the length of the individual trajectories and as a function of the number of trajectories, for which a subset of the forty trajectories is used in the bootstrap analysis. Black contour lines in the figure denote a constant computer simulation time according to the combined number and length of the individual trajectories.

Overlap of trajectories from individual simulations

A motivation for using multiple trajectories is to efficiently sample a free-energy basin in the conformational space by launching individual trajectories to explore in parallel different regions within the same basin of the conformational space. The free-energy basin is rugged with transitions occurring on a nanosecond timescale between wells separated by low energy barriers. Transitions of individual trajectories occur frequently between the wells observed as densely populated regions in the energy-rmsd distributions (main text figure 5). To query whether trajectories are likely sampling the same free-energy basin, we determined the nearness of trajectories in the conformational space and overlap of the sampling from individual simulations by comparing the distribution of pairwise root-mean-square difference (rmsd) in backbone coordinates for snapshots from the same trajectory with that for snapshots taken from two different trajectories. We evaluated an all-against-all pairwise rmsd, as opposed to the rmsd against a single reference structure, with the notion that a distribution is a better indicator of the similarity of conformations and thus the nearness of one trajectory to another in the conformational space. All-against-all rmsd distributions for pairs of snapshots from the same trajectory versus pairs from two independent trajectories are shown in figure S6 using three out of the forty trajectories for each complex. The right column results are for two trajectories starting from different initial coordinates (DIC), and the left column corresponds to two trajectories with the same initial coordinates (SIC) but different initial velocities. Each panel shows the all-against-all rmsd distribution for two pairs from individual trajectories and that for pairs between the two trajectories. The within-trajectory distributions are solid lines (blue and red), and the between-trajectory distributions are dashed (green). Examination of the combined panels finds that while the 'between rmsd' distributions fall at larger rmsd values, there is always overlap in the distributions, suggesting that the trajectories sample similar regions of conformational space. The DIC comparison has less overlap than that for SIC, indicating a broader sampling of conformational space, but the between-rmsd distribution still overlaps the within-rmsd distribution, consistent with the premise that the trajectories sample the same free-energy basin.

Correlative behavior of the component energy terms

To elucidate the source of the covariance in the solute-solute and solute-solvent components for $\Delta\Delta E$ (figure 4 main text), we plot the deviation from the ensemble-mean total potential energy of the local mean energy values $\bar{E}_k$ for forty MD trajectories of the complex Src SH2-pYEEI in Figure S7. The inverse correlation between the relative $\Delta\Delta E^{UV}$ with $\Delta\Delta E^{UU}$ arises from the strong covariance of the individual trajectory values $\bar{E}_k^{UU}$ and $\bar{E}_k^{UV}$. It should be noted that the apparent correlation of $\bar{E}_k^{VV}$ derives largely from solvation water molecules near the protein and is less diminished as the number of water molecules relative to solute is increased. The strong anticorrelation of $\bar{E}_k^{UU}$ and $\bar{E}_k^{UV}$ is reminiscent of the cancellation of coulombic interaction and reaction field energy and therefore suggests that the differences among multiple trajectories is dominated by solvation effects rather than protein intramolecular configurational effects.

Comparison of confidence interval calculated from the bootstrap error and standard error

In our analysis of relative binding enthalpies, a bootstrap analysis was used to estimate uncertainty in the mean energy values for the total energy and the component energies. We satisfy the bootstrap requirement of independent samples by using independent trajectories, rather than subdivision of a trajectory into small time blocks that can have correlated effects. The uncertainty estimated from bootstrap analysis was compared to the uncertainty obtained using the standard error of local energy mean values. Analogous to figure 3 in the main text, we show the uncertainty from the standard error plotted as a function of the simulation time for averaging and the number of trajectories taken from the forty 10-ns SH2-pYEEI simulations. The 95% CI (1.96 $\delta_{N,K}$) for E^T (Figure S8) and for the component energy terms E^{UU}, E^{UV} and E^{VV}, (Figure S9) were calculated from the local mean energy values, $\bar{E}_k^N$, over N snapshots for K trajectories:

$$\delta_{\langle E_K^N\rangle} = \sqrt{\frac{\sum_{k=1}^{K}\left(\bar{E}_k^N - \langle \bar{E}^N\rangle\right)^2}{K(K-1)}}$$

$$\delta_{N,K} = \sqrt{\frac{\sum_{i=1}^{50}\left(\delta^i_{\langle E_K^N\rangle}\right)^2}{50}} \qquad \text{S(1)}$$

N varies from 50 ps to 10 ns and K varies from 2 to 40 trajectories. For K trajectories, 50 sets of K unique trajectories were chosen randomly from the forty computed trajectories.

For a small sample number, the uncertainty obtained by the bootstrap method is a better estimate of the width of the underlying Gaussian distribution of mean values than the standard error. Nevertheless, the estimates from the two procedures approach each other as the number of simulations increases so that the bootstrap error and the standard error are similar for 40 trajectories. For example, for the SH2-pYEEI trajectories, the standard error of E^T is ±1.11 kcal/mol (±2.2 kcal/mol 95% CI) and bootstrap error is ±1.07 kcal/mol (±2.1 kcal/mol 95% CI).

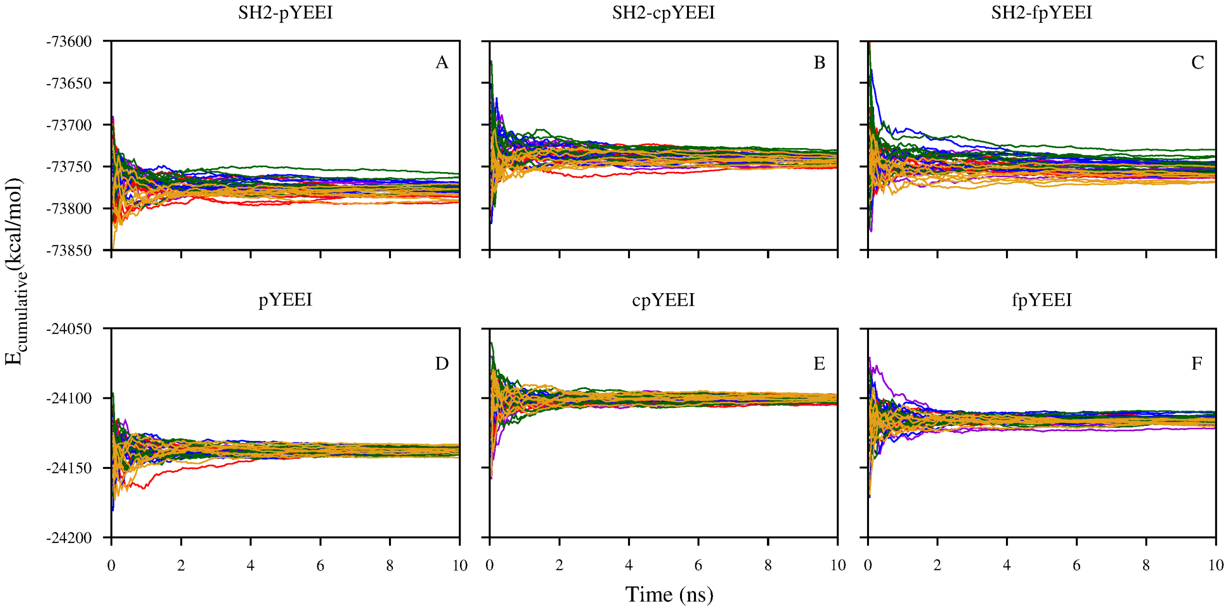

Figure S. 1 Cumulative average for the local potential energy $\bar{E}_k^T$ from the forty individual trajectories for the indicated Src SH2 complexes (top row) and unbound peptide ligands (bottom row).

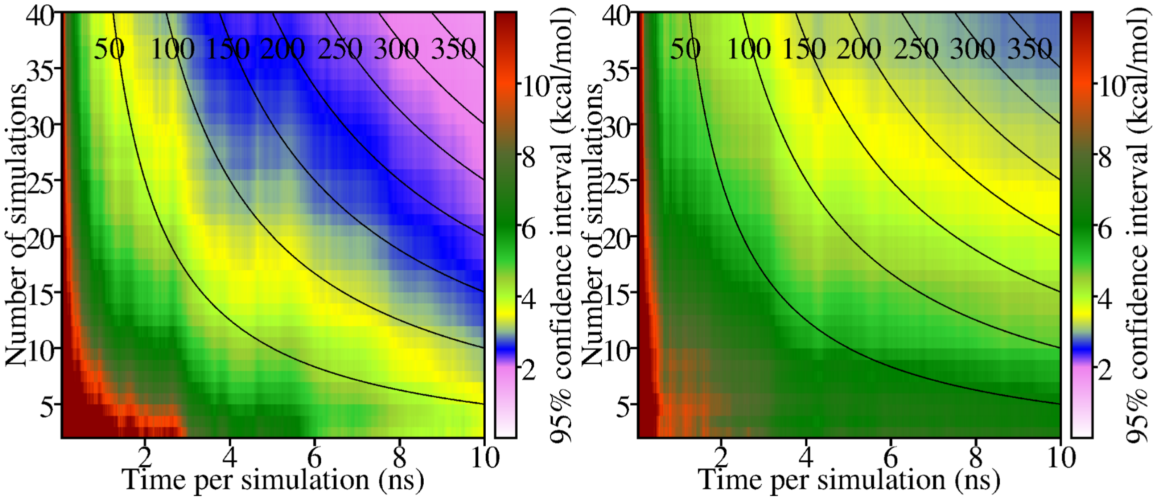

Figure S. 2 Convergence of the statistical uncertainty from bootstrap analysis in the estimate of E^T for SH2-cpYEEI (left) and SH2-fpYEEI (right). The 95% CI (1.96 $\delta_{<E>}$) for E^T narrows with increasing number of simulations and time per simulation. Uncertainties were determined for subsets of the forty 10-ns simulations and over varying time of averaging as detailed in main text. Solid black curves indicate equal total simulation time (ns) spread over the number of simulations and the time per simulation.

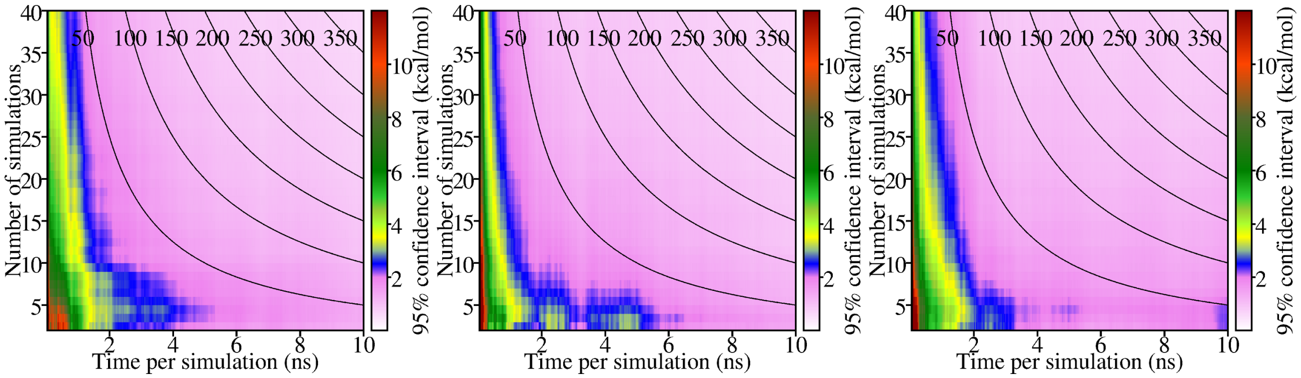

Figure S. 3 Convergence of the statistical uncertainty from bootstrap analysis in the estimate of E^T for the peptides: pYEEI, cpYEEI, and fpYEEI. See figure S2 caption for details.

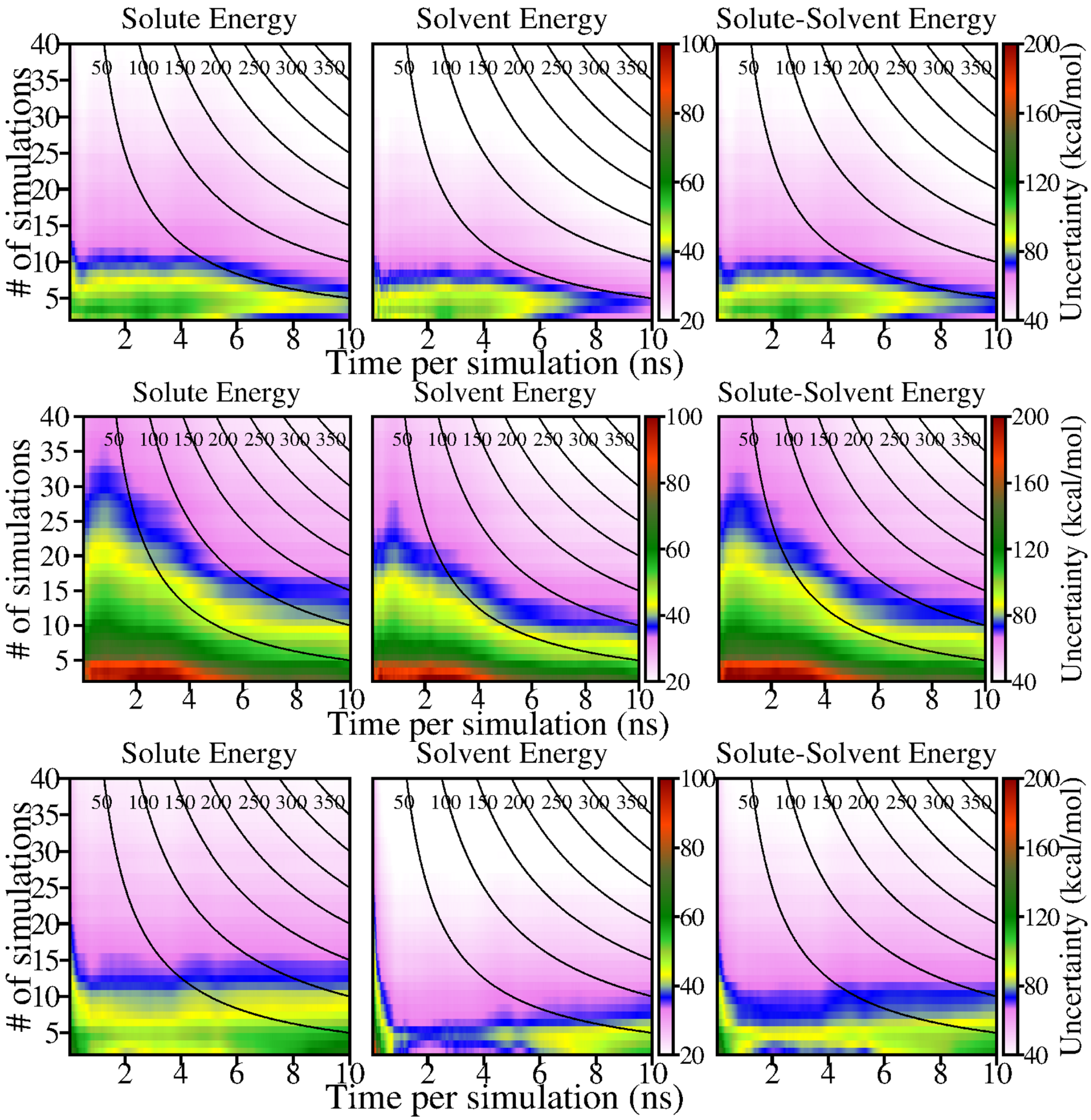

Figure S. 4 Convergence of the statistical uncertainty from bootstrap analysis in the estimate of component energy terms (E^{UU}, E^{VV}, E^{UV}) for complexes SH2-pYEEI (top), SH2-cpYEEI (middle) and SH2-fpYEEI (bottom). See figure S2 caption for details.

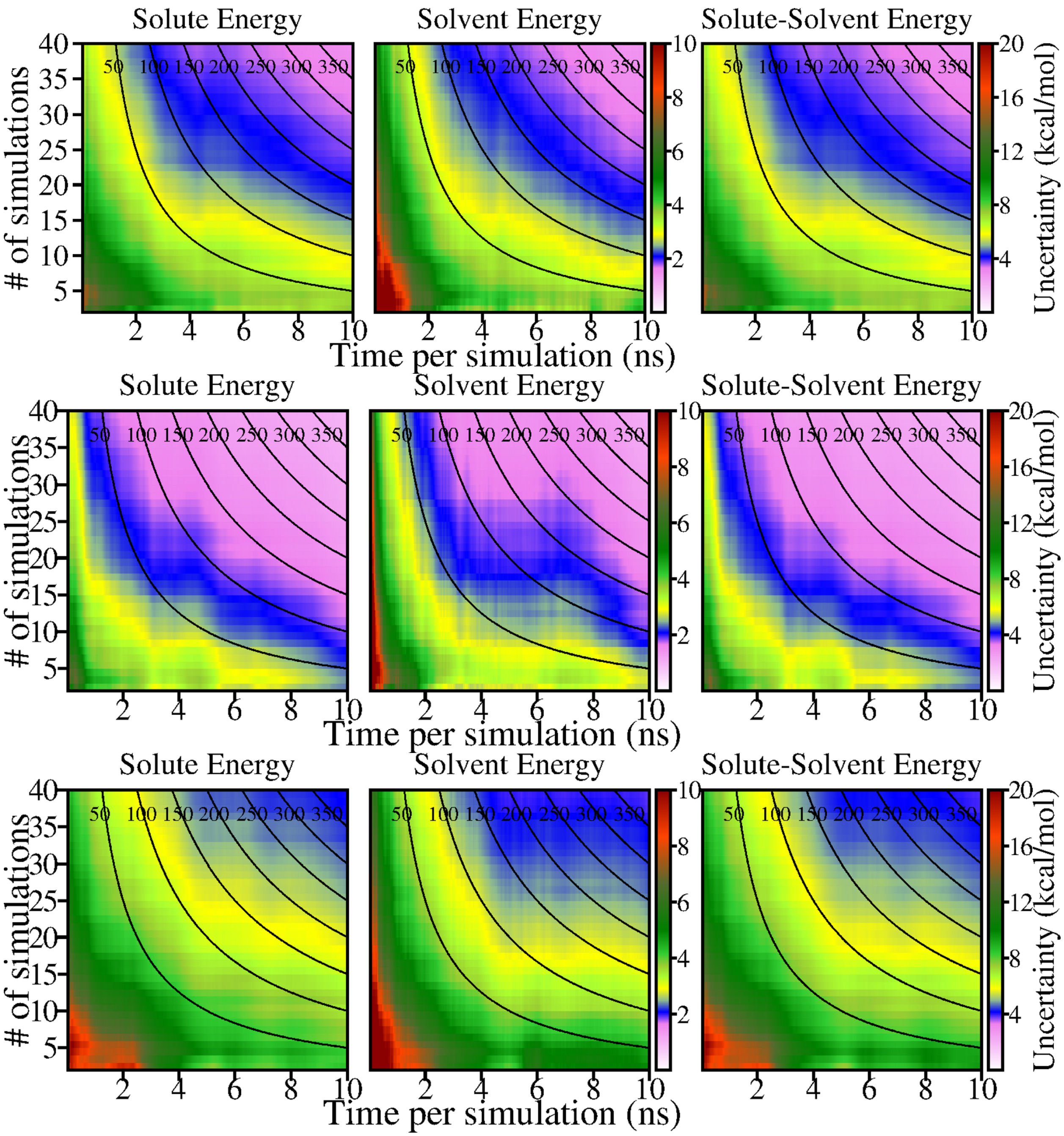

Figure S. 5 Convergence of the statistical uncertainty from bootstrap analysis in the estimate of component energy terms (E^{UU}, E^{VV}, E^{UV}) for peptides pYEEI (top), cpYEEI (middle) and fpYEEI (bottom). See figure S2 caption for details.

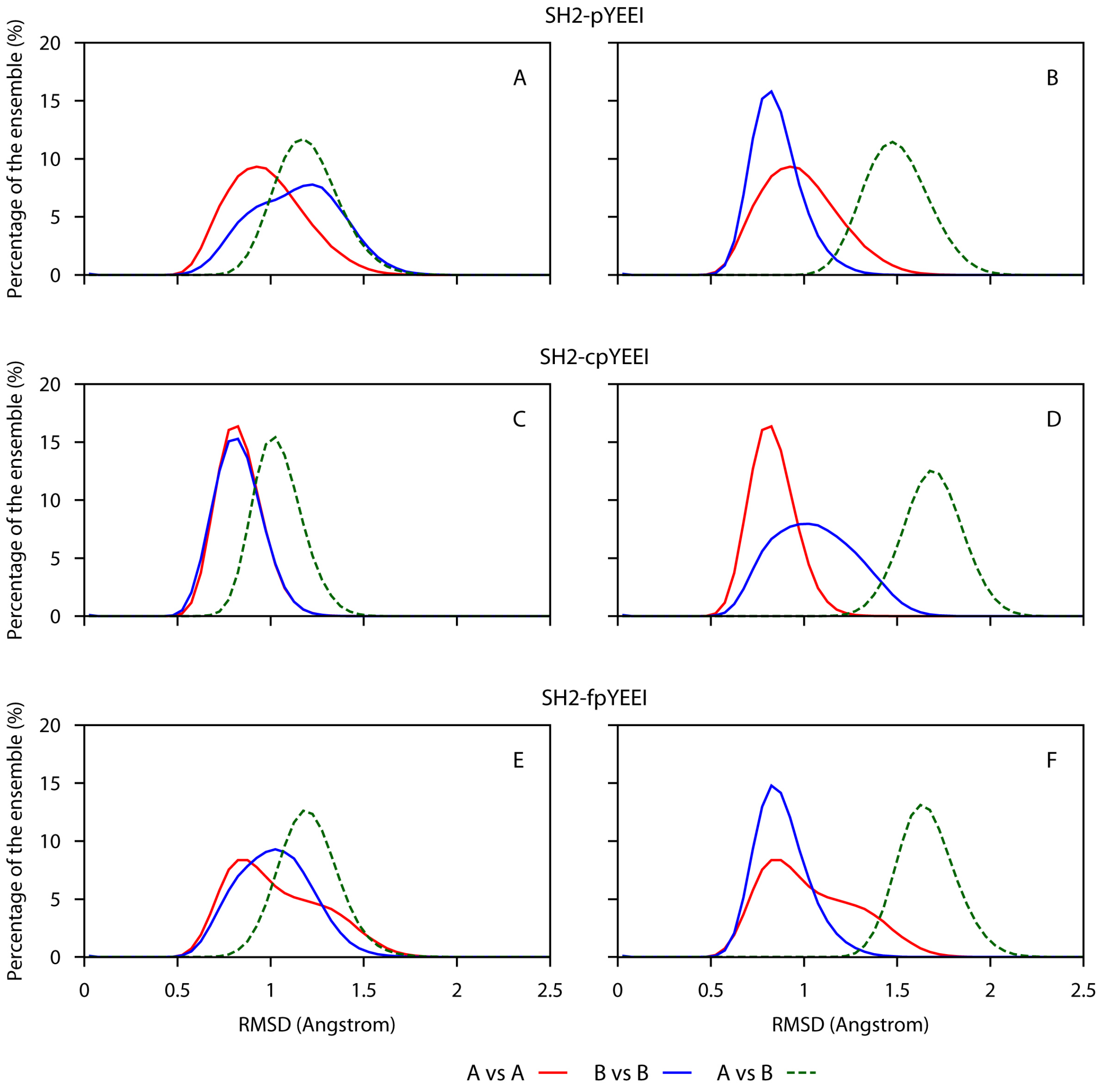

Figure S. 6 Figure S. 6 Distribution of pairwise rmsd values calculated between all pairs of snapshots either within one trajectory or between two trajectories. The rmsd was calculated over all backbone N, Cα, C atoms after superposition with respect to these same atoms. Three trajectories were selected out of the 40 computed for each complex, with two being initiated from different initial coordinates (DIC) and two with different initial velocities and the same initial coordinates (SIC). The within-rmsd distributions (solid red and blue curves) and between-rmsd distributions (dashed line, green) for SIC (left column) or DIC (right column) pairs are shown. The distributions overlap consistent with the premise that the individual trajectories sample the same free-energy basin.

.

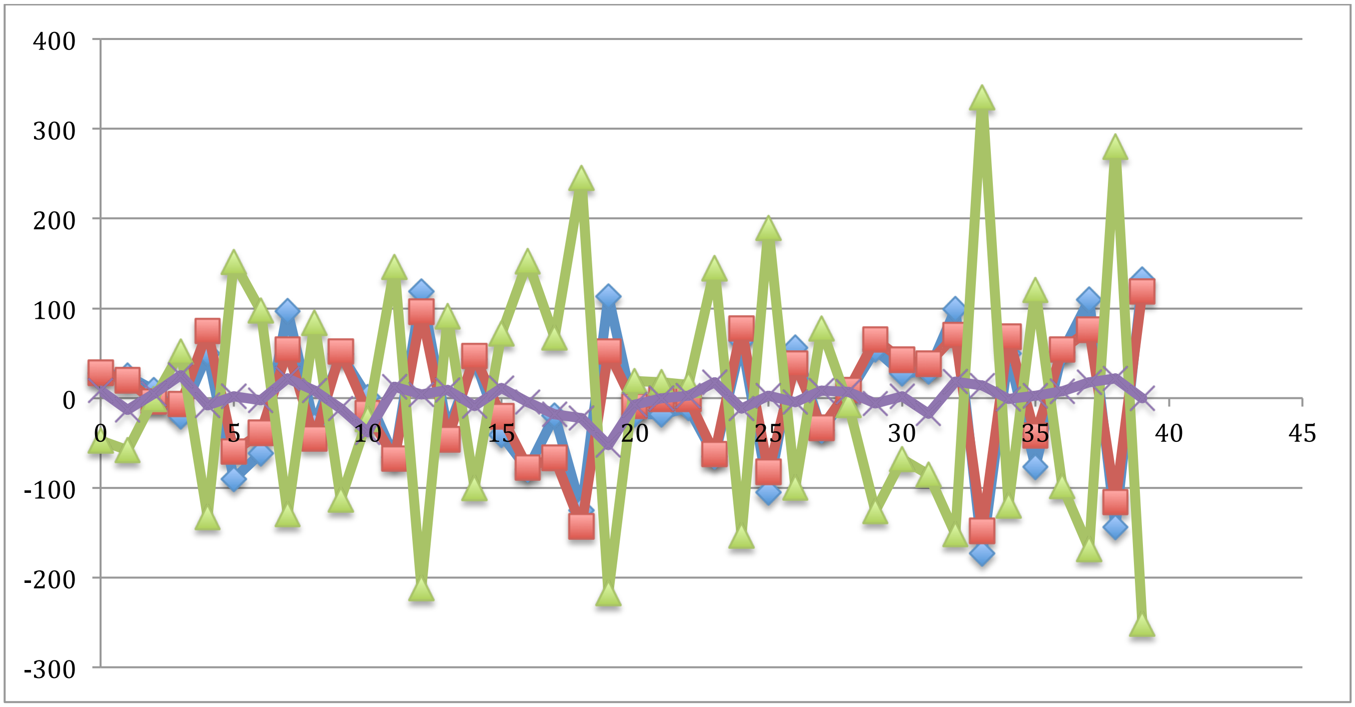

Figure S.7 Deviation from the ensemble mean values $\langle E \rangle$ of the local mean values $\overline{E}_k$ for forty MD trajectories of the complex Src SH2-pYEEI for the total $\overline{E}_k^T$ (purple), solute-solute $\overline{E}_k^{UU}$ (blue), solvent-solvent $\overline{E}_k^{VV}$ (red), and solute-solvent $\overline{E}_k^{UV}$ (green) interactions.

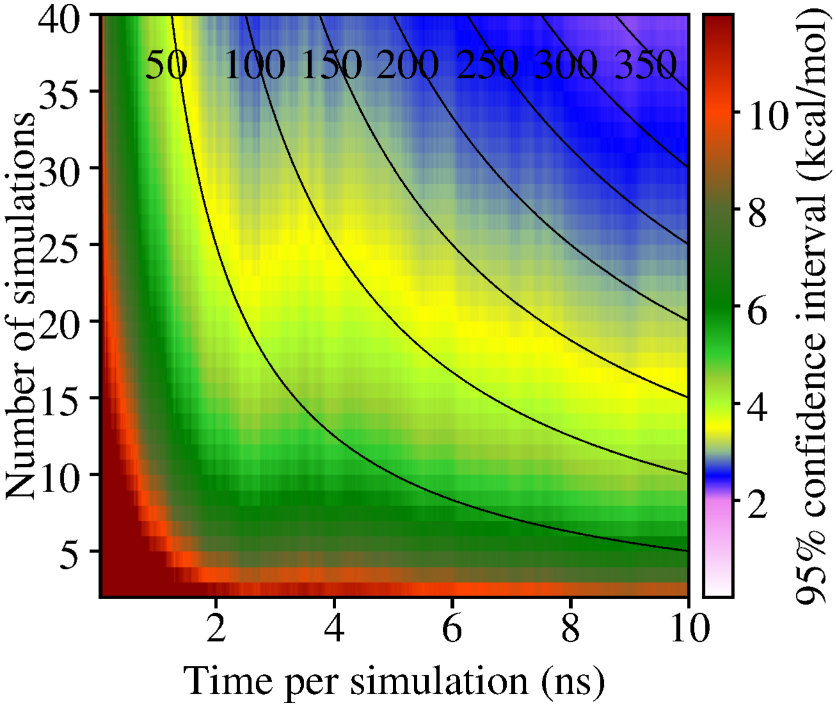

Figure S.8 Convergence of the statistical uncertainty from standard error of E^T for SH2-pYEEI analogous to figure 3 in main text. Uncertainties were calculated from equation S(1).

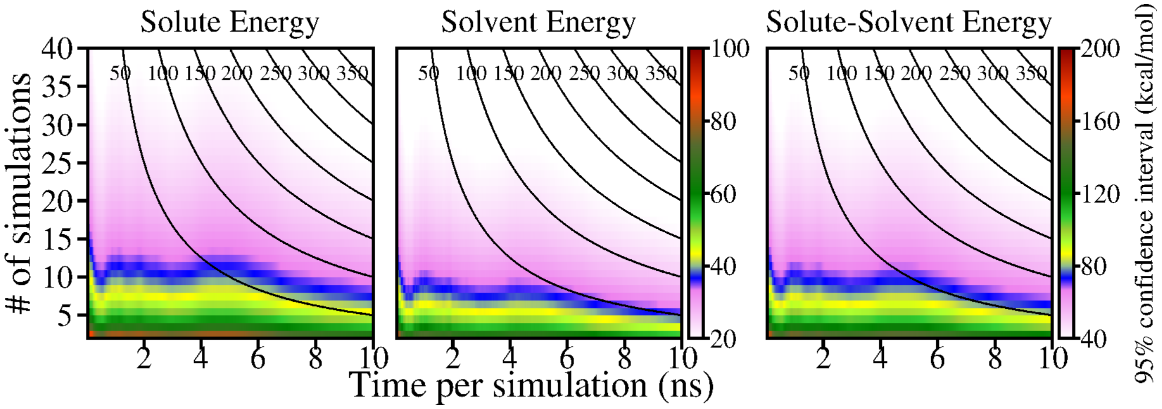

Figure S.9 Convergence of the statistical uncertainty from standard error of component energy terms (E^{UU}, E^{VV}, E^{UV}) for SH2-pYEEI, analogous to figure S4 (top).

Table S. 1 Least-squares fitted slopes at long production time (from 8 to 10 ns) of the cumulative local average of total potential energy (E^T) of 40 individual trajectories.

	Slope $(kcal/(mol * ns))$		
	SH2-pYEEI	SH2-cpYEEI	SH2-fpYEEI
1	2.14	-1.35	-2.26
2	-0.39	-2.01	0.17
3	-1.93	-0.66	-1.19
4	1.05	-1.28	-0.01
5	0.11	-1.12	1.14
6	-0.75	1.36	0.66
7	0.19	0.29	0.99
8	1.19	-0.95	-0.07
9	-0.34	-0.67	1.07
10	-1.39	-0.74	0.28
11	0.02	0.60	-1.19
12	1.25	-1.79	0.01
13	-2.15	0.77	-0.97
14	-1.09	-1.16	-0.61
15	0.90	0.37	-0.98
16	-1.40	-0.21	0.90
17	-1.55	-0.89	-0.73
18	-1.54	-0.47	-1.79
19	-2.05	-1.49	-0.63
20	1.57	-0.76	0.02
21	-0.55	-0.32	0.71
22	0.25	-0.48	-0.40
23	-1.06	0.86	-0.91
24	-0.72	-0.15	0.25
25	-1.69	0.17	-0.37
26	-2.28	-0.02	0.25
27	1.96	-0.49	-0.72
28	0.42	-1.89	-0.06
29	0.31	0.85	-0.05
30	0.99	1.43	-2.22
31	0.67	0.76	0.65
32	2.07	-1.66	-0.93
33	0.20	-1.33	-0.24
34	1.45	-1.65	-0.60
35	-1.05	-2.25	-1.09
36	-0.45	0.82	-0.71
37	-1.87	-0.23	-1.53
38	0.76	-1.04	0.61
39	1.06	-1.44	-0.85
40	-0.90	-1.93	-0.72