13 Supplementary Information

Table S1. Error metrics for SAMPL7 methods (ranked and non-ranked) for datasets with optional systems. The root mean square error (RMSE), mean absolute error (MAE), signed mean error (ME), coefficient of correlation (R^2), slope (m), and Kendall's rank correlation coefficient (Tau) were computed via bootstrapping with replacement. Shown are results for individual host categories with optional systems, which includes the combined OA and exoOA dataset (**GDCC-OA and exoOA**) and Cyclodextrin derivatives. Statistics include optional host-guest systems OA-g1, OA-g2, OA-g3 OA-g4, OA-g5, OA-g6, bCD-g1, and bCD-g2. Optional GDCC systems were not included for reference calculations (*Docking/GAFF/YANK_REF*), thus only cyclodextrin statistics are included.

ID	sid	RMSE [kcal/mol]	MAE [kcal/mol]	ME [kcal/mol]	R^2	m	τ
GDCC-OA and exoOA							
AMOEBA/DDM/BAR	29	1.05 [0.78, 2.17]	0.79 [0.61, 1.76]	-0.30 [-1.19, 0.54]	0.83 [0.43, 0.93]	1.14 [0.70, 1.79]	0.78 [0.38, 0.93]
RESP/GAFF/MMPBSA-Cor	20	1.45 [1.05, 2.47]	1.16 [0.82, 2.13]	1.02 [0.15, 1.90]	0.70 [0.03, 0.87]	0.61 [0.13, 1.03]	0.57 [0.00, 0.84]
xtb-GNF/Machine Learning/CORINA MD	28	1.77 [1.15, 2.83]	1.27 [0.86, 2.36]	0.31 [-0.78, 1.45]	0.17 [0.00, 0.61]	0.27 [-0.22, 0.87]	0.34 [-0.24, 0.67]
AMOEBA/DDM/BAR_2	30	1.89 [1.22, 3.05]	1.41 [0.92, 2.51]	-0.99 [-2.10, 0.07]	0.43 [0.02, 0.78]	0.70 [0.12, 1.43]	0.50 [-0.02, 0.81]
AMOEBA/DDM/BAR_3	31	2.10 [1.48, 3.15]	1.73 [1.15, 2.74]	0.24 [-1.04, 1.54]	0.53 [0.08, 0.79]	1.18 [0.46, 1.91]	0.48 [0.02, 0.80]
B2PLYPD3/SMD_QZ-R	23	3.92 [2.53, 5.47]	3.00 [1.85, 4.57]	1.84 [-0.03, 3.77]	0.29 [0.02, 0.61]	1.17 [0.29, 2.23]	0.35 [-0.06, 0.66]
FSDAM/GAFF2/OPC3	14	4.57 [3.28, 7.62]	4.17 [2.63, 6.56]	-0.40 [-3.54, 2.55]	0.04 [0.00, 0.48]	-0.41 [-1.68, 1.70]	-0.05 [-0.56, 0.41]
RESP/GAFF/MMPBSA/Nmode	18	5.26 [4.26, 6.47]	4.96 [3.89, 6.12]	-4.96 [-6.12, -3.88]	0.68 [0.24, 0.88]	1.30 [0.70, 2.02]	0.61 [0.18, 0.87]
B2PLYPD3/SMD_TZ	22	6.70 [3.64, 9.78]	4.84 [2.74, 7.55]	3.09 [0.13, 6.31]	0.30 [0.04, 0.66]	2.00 [0.62, 3.74]	0.38 [-0.04, 0.71]
B2PLYPD3/SMD_QZ-NR	24	6.78 [3.43, 10.40]	4.71 [2.58, 7.69]	2.61 [-0.42, 6.08]	0.29 [0.03, 0.66]	2.04 [0.63, 4.12]	0.40 [-0.03, 0.72]
B2PLYPD3/SMD_DZ	21	7.12 [5.27, 8.96]	6.16 [4.32, 8.11]	5.44 [2.96, 7.79]	0.25 [0.01, 0.62]	1.41 [0.00, 2.49]	0.34 [-0.10, 0.63]
RESP/GAFF/MMPBSA	19	8.66 [7.54, 9.83]	8.48 [7.32, 9.62]	8.48 [7.32, 9.62]	0.70 [0.16, 0.91]	1.36 [0.70, 1.82]	0.57 [0.17, 0.88]
AM1-BCC/GAFF/MMPBSA	17	10.67 [9.13, 12.16]	10.29 [8.64, 11.89]	10.29 [8.64, 11.89]	0.63 [0.13, 0.90]	1.74 [0.88, 2.38]	0.57 [0.19, 0.88]
RESP/GAFF/MMGBSA	16	11.43 [10.11, 12.79]	11.19 [9.78, 12.56]	11.19 [9.78, 12.56]	0.51 [0.04, 0.87]	1.27 [0.37, 1.89]	0.52 [0.08, 0.84]
Cyclodextrin derivatives							
FSDAM/GAFF2/OPC3_ranked	12	1.23 [1.36, 3.39]	1.01 [1.06, 2.84]	0.47 [-0.90, 1.87]	0.04 [0.00, 0.46]	0.17 [-1.26, 1.66]	0.23 [-0.41, 0.55]
Noneq/Alchemy/CGENFF	26	1.55 [1.17, 2.33]	1.35 [0.93, 2.03]	0.99 [0.24, 1.74]	0.05 [0.00, 0.39]	0.24 [-0.45, 0.95]	0.10 [-0.41, 0.49]
Noneq/Alchemy/consensus	27	1.62 [1.21, 2.17]	1.38 [0.96, 1.90]	1.08 [0.43, 1.72]	0.03 [0.00, 0.30]	0.18 [-0.33, 0.74]	0.03 [-0.38, 0.45]
FSDAM/GAFF2/OPC3_JB	13	1.71 [1.55, 3.76]	1.48 [1.21, 3.19]	0.54 [-0.94, 2.04]	0.01 [0.00, 0.41]	-0.14 [-1.58, 1.47]	0.03 [-0.44, 0.48]
Noneq/Alchemy/GAFF	25	1.84 [1.35, 2.58]	1.54 [1.07, 2.24]	1.17 [0.37, 1.97]	0.01 [0.00, 0.28]	0.12 [-0.55, 0.83]	0.02 [-0.36, 0.43]
Docking/GAFF/YANK_REF	REF1	2.64 [1.87, 3.42]	2.19 [1.51, 2.94]	0.64 [-0.58, 1.84]	0.02 [0.00, 0.36]	-0.29 [-1.59, 0.87]	-0.10 [-0.44, 0.24]
AM1-BCC/MD/GAFF/TIP4PEW/QMMM	15	46.62 [22.85, 65.69]	32.00 [17.92, 49.22]	31.27 [16.89, 48.87]	0.04 [0.00, 0.33]	7.62 [-3.31, 30.72]	0.24 [-0.13, 0.52]

Table S2. Error metrics for ranked method submission of absolute binding free energy calculations of all host-guest systems. The root mean square error (RMSE), mean absolute error (MAE), signed mean error (ME), coefficient of correlation (R^2), slope (m), and Kendall's rank correlation coefficient (τ) were computed, with confidence intervals from bootstrapping with replacement. All three datasets (**TrimerTrip**, **GDCC-OA and exoOA**, **Cyclodextrin derivatives**), and an artificial sub-dataset of exo-OA ranked submissions (**GDCC-exoOA**) are included. Statistical values in this table do not include optional host-guest systems OA-g1, OA-g2, OA-g3, OA-g4, OA-g5, OA-g6, bCD-g1, and bCD-g2, for which values had been released previously. Each method has an assigned unique submission ID (sid).

ID	sid	RMSE [kcal/mol]	MAE [kcal/mol]	ME [kcal/mol]	R^2	m	τ
TrimerTrip							
AMOEBA/DDM/BAR	6	2.76 [1.83, 3.98]	2.12 [1.35, 3.33]	-1.69 [-2.98, -0.44]	0.50 [0.13, 0.77]	1.25 [0.53, 2.06]	0.47 [0.12, 0.74]
FSDAM/GAFF2/OPC3	4	2.97 [2.11, 5.13]	2.24 [1.62, 4.22]	0.43 [-1.59, 2.33]	0.12 [0.00, 0.56]	0.60 [-0.51, 1.60]	0.24 [-0.23, 0.61]
MD/DOCKING/GAFF/xtb-GNF/	5	5.65 [3.87, 7.36]	4.51 [3.01, 6.40]	-4.23 [-6.19, -2.23]	0.00 [0.00, 0.26]	-0.10 [-1.02, 0.80]	-0.05 [-0.41, 0.35]
GDCC - OA and exoOA							
RESP/GAFF/MMPBSA-Cor	20	1.24 [0.76, 2.46]	0.95 [0.59, 2.15]	0.94 [-0.13, 1.99]	0.94 [0.11, 0.97]	0.65 [0.18, 1.14]	0.83 [0.03, 1.00]
AMOEBA/DDM/BAR	29	1.25 [0.68, 2.53]	0.92 [0.54, 2.12]	-0.36 [-1.54, 0.83]	0.80 [0.34, 0.97]	1.11 [0.57, 1.97]	0.72 [0.18, 1.00]
xtb-GNF/Machine Learning/CORINA_MD	28	2.26 [1.39, 3.44]	1.91 [1.10, 3.12]	0.37 [-1.31, 2.06]	0.01 [0.00, 0.78]	0.04 [-0.58, 0.54]	0.06 [-0.64, 0.81]
B2PLYPD3/SMD_QZ-R	23	4.52 [2.52, 6.39]	3.70 [1.96, 5.67]	3.15 [0.84, 5.44]	0.49 [0.03, 0.93]	1.43 [-0.11, 2.98]	0.37 [-0.31, 0.87]
GDCC - exoOA							
AMOEBA/DDM/BAR	29	1.27 [0.56, 2.72]	0.91 [0.45, 2.31]	-0.66 [-1.98, 0.61]	0.81 [0.30, 0.99]	1.05 [0.45, 2.12]	0.71 [0.05, 1.00]
RESP/GAFF/MMPBSA-Cor	20	1.32 [0.68, 2.65]	1.03 [0.54, 2.34]	1.01 [-0.18, 2.20]	0.95 [0.04, 0.99]	0.61 [0.04, 1.20]	0.81 [-0.14, 1.00]
xtb-GNF/Machine Learning/CORINA MD	28	2.43 [1.40, 3.71]	2.11 [1.10, 3.42]	0.82 [-1.12, 2.77]	0.00 [0.00, 0.91]	0.01 [-0.81, 0.57]	0.05 [-0.78, 1.00]
B2PLYPD3/SMD_QZ-R	23	4.76 [2.26, 6.93]	3.90 [1.81, 6.26]	3.50 [0.91, 6.12]	0.72 [0.24, 0.99]	1.97 [0.88, 3.77]	0.59 [-0.06, 1.00]
Cyclodextrin derivatives							
FSDAM/GAFF2/OPC3_ranked	12	1.28 [1.32, 3.51]	1.04 [1.04, 2.95]	0.63 [-0.84, 2.10]	0.01 [0.00, 0.50]	0.12 [-1.62, 2.30]	0.21 [-0.46, 0.57]
Noneq/Alchemy/consensus	27	1.70 [1.27, 2.28]	1.48 [1.03, 2.04]	1.21 [0.52, 1.87]	0.02 [0.00, 0.29]	0.16 [-0.48, 0.93]	-0.02 [-0.43, 0.45]
AM1-BCC/MD/GAFF/TIP4PEW/QMMM	15	46.62 [22.85, 65.69]	32.00 [17.92, 49.22]	31.27 [16.89, 48.87]	0.04 [0.00, 0.33]	7.62 [-3.31, 30.72]	0.24 [-0.13, 0.52]

Table S3. Error metrics for methods used in reference binding free energy calculations of all host-guest systems. Please see section 6.1.1 for details on the submission methodology. Optional systems in the GDCC and cyclodextrin datasets (OA-g1, OA-g2, OA-g3, OA-g4, OA-g5, OA-g6, bCD-g1, and bCD-g2) are not part of this analysis. This table includes the method ID, method submission ID (sid), root mean squared error (RMSE), mean absolute error (MAE), mean signed error (ME), coefficient of determination (R^2), linear regression slope (m), and kendall rank correlation coefficient (τ) for cyclodextrin, TrimerTrip, and GDCC datasets (includes both OA and exoOA predictions). An artificial separation of GDCC was done to obtain a exoOA sub-dataset for analysis.

ID	sid	RMSE [kcal/mol]	MAE [kcal/mol]	ME [kcal/mol]	R^2	m	τ
Cyclodextrin derivatives							
Docking/GAFF/YANK_REF	REF1	2.64 [1.87, 3.42]	2.19 [1.51, 2.94]	0.64 [-0.58, 1.84]	0.02 [0.00, 0.36]	-0.29 [-1.59, 0.87]	-0.10 [-0.44, 0.24]
TrimerTrip							
Docking/GAFF/YANK_REF	REF2	7.18 [5.63, 8.71]	6.57 [5.16, 8.10]	-6.57 [-8.09, -5.16]	0.11 [0.00, 0.59]	0.57 [-0.56, 1.55]	0.12 [-0.35, 0.56]
Docking/GAFF/YANK_REF_2	REF3	7.21 [5.73, 8.75]	6.63 [5.26, 8.13]	-6.63 [-8.12, -5.26]	0.12 [0.00, 0.59]	0.57 [-0.55, 1.54]	0.12 [-0.34, 0.57]
GDCC - OA and exoOA							
Docking/GAFF/YANK_REF	REF4	4.05 [1.54, 5.88]	2.90 [1.21, 4.93]	2.40 [0.41, 4.67]	0.12 [0.00, 0.65]	-0.30 [-1.06, 0.53]	-0.11 [-0.70, 0.60]
GDCC - exoOA							
Docking/GAFF/YANK_REF	REF4	4.48 [1.56, 6.43]	3.25 [1.10, 5.65]	2.60 [0.06, 5.40]	0.37 [0.03, 0.95]	-0.58 [-1.56, 0.08]	-0.43 [-1.00, 0.33]

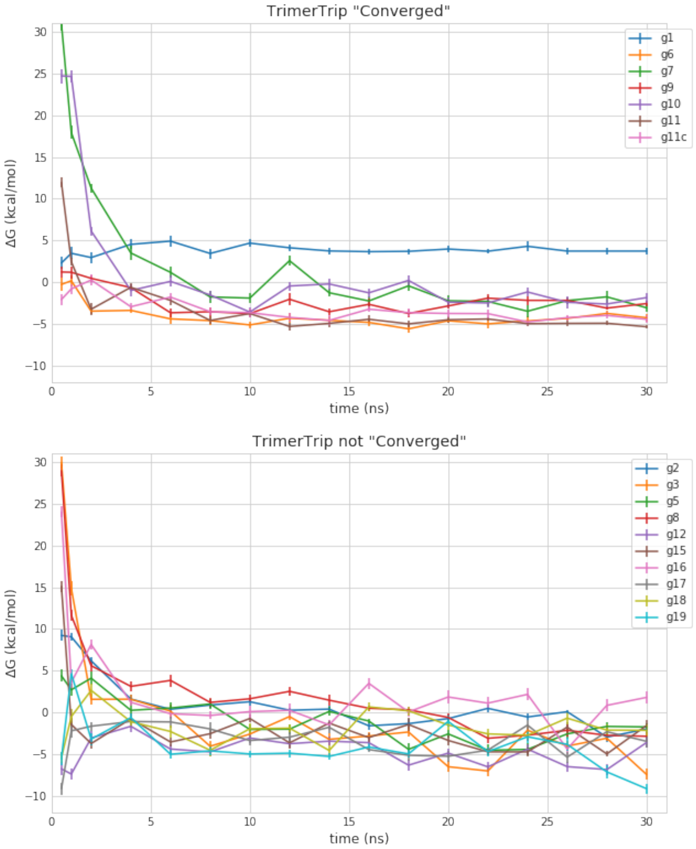

Figure S1. Reference calculations for TrimerTrip. Plots showing converging free energy estimates (top) or lack of convergence (bottom) for the TrimerTrip dataset. The calculation for clip-g11c is with g11 but run with an open TrimerTrip conformer extracted from one of our previous simulations.

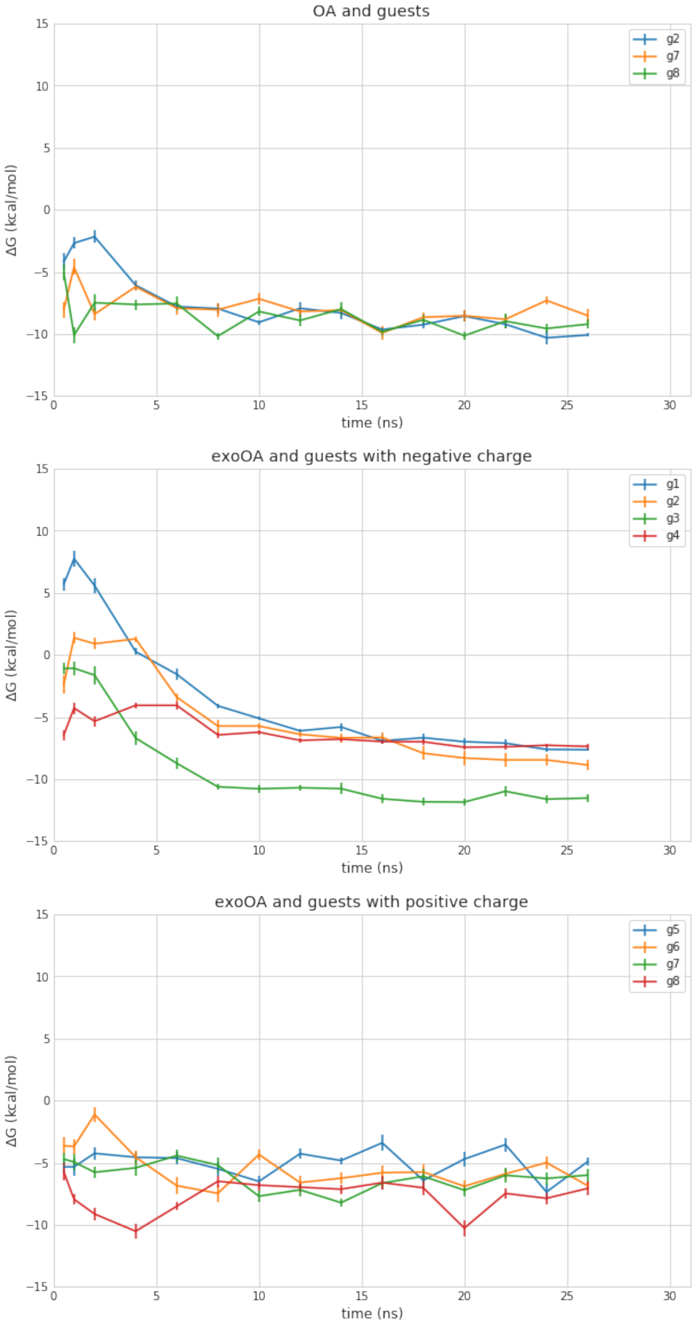

Figure S2. Reference calculations for Cavitands. Plots showing converging free energies for the GDCC dataset which includes the OA and exoOA hosts. (Top) Free energy estimate plotted as a function of time for the OA system with the required guests. (Middle) Free energies estimates plotted as a function of time for the exoOA host with negatively charged guests. For these systems the free energy is closely converged. (Bottom) The free energy estimates for exoOA with a postively charged guest are not readily converged, particularly in comparison to other systems in the GDCC dataset.

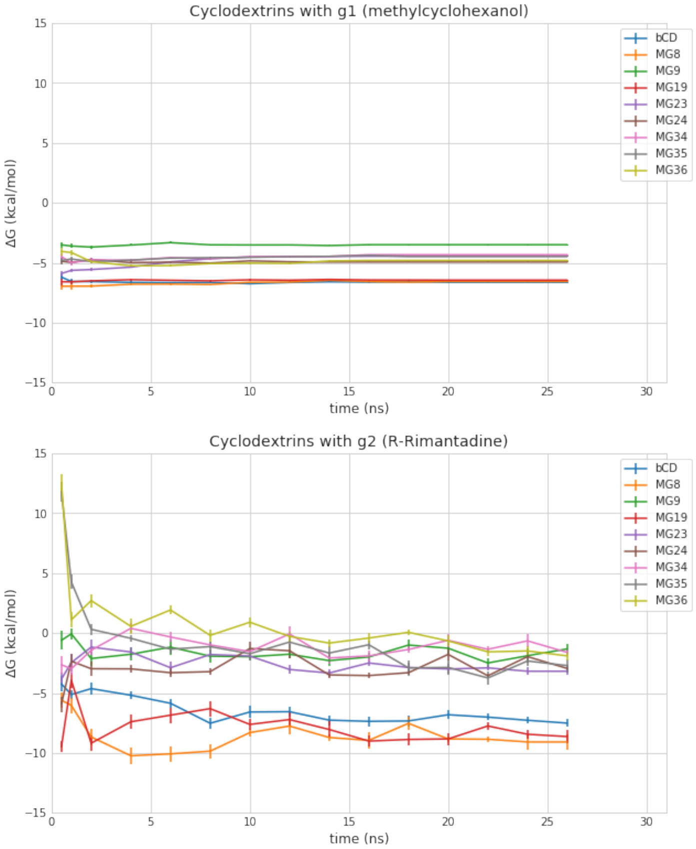

Figure S3. Reference calculations for Cyclodextrin derivatives. Plots showing the convergence of free energy estimates for cyclodextrins with g1 (top) or with g2 (bottom). Free energies are well converged for systems with g1, while not all systems with g2 are convincingly converged at the simulated timescale.

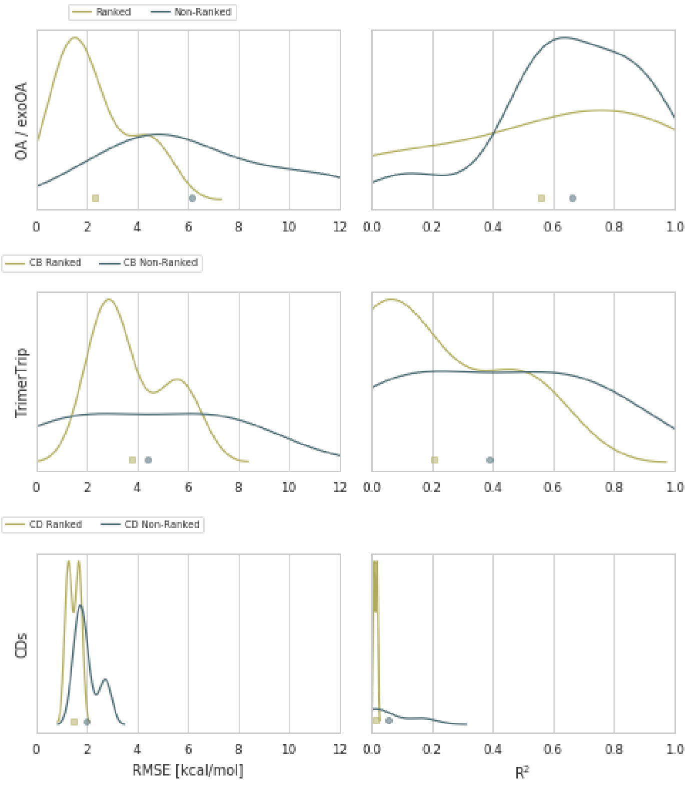

Figure S4. Comparing ranked and non-ranked methods based on RMSE and R^2. Plots compare the distribution of predictive and correlational statistics comparing ranked and non-ranked methods for each dataset (GDCCs (OA/exoOA), TrimerTrip, and Cyclodextrins (CDs)) in the SAMPL7 host-guest challenge. Ranked methods statistics are shown in yellow, and non-ranked are shown in blue. In addition, the mean is of the distributions are marked by a dot under the curves. On average the RMSE for ranked methods was better compared to non-ranked methods. However, on average non-ranked methods had a better R^2 for all datasets.