HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2022 December 22.

Published in final edited form as:

J Chem Inf Model. 2022 December 12; 62(23): 6084–6093. doi:10.1021/acs.jcim.2c01115.

Practical Guidance for Consensus Scoring and Force Field Selection in Protein-Ligand Binding Free Energy Simulations

Han Zhang[1], **Seonghoon Kim**[2], **Wonpil Im**[1,*]

[1]Departments of Biological Sciences, Chemistry, Bioengineering, and Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, USA

[2]School of Computational Sciences, Korea Institute for Advanced Study, Seoul 02455, Republic of Korea

Abstract

The advances in ligand binding affinity prediction have been fostered by system generation tools and improved force fields (FFs). CHARMM-GUI *Free Energy Calculator* provides input and post-processing scripts for AMBER-TI free energy calculations with various FFs. In this study, we used 12 different FF combinations (ff14SB and ff19SB for protein, GAFF2.2 and OpenFF for ligand, and TIP3P, TIP4PEW, and OPC for water) to calculate relative binding free energies ($\Delta\Delta G_{bind}$) for 80 alchemical transformations (among the JACS benchmark set) with different numbers of λ windows (12 or 21) and simulation times (1, 5, or 10 ns). Our results show that 12 λ windows and 5 ns simulation time for each window are sufficient to obtain reliable $\Delta\Delta G_{bind}$ with 4 independent runs for the current benchmark set. While there is no statistically noticeable performance difference among 12 different FF combinations compared to the experimental values, a combination of ff14SB + GAFF2.2 + TIP3P FFs appears to be best with a mean unsigned error of 0.87 [0.69, 1.07] kcal/mol, a root-mean-square error of 1.22 [0.94, 1.50] kcal/mol, a Pearson's correlation of 0.64 [0.52, 0.76], a Spearman's correlation of 0.73 [0.58, 0.83], and a Kendell's correlation of 0.54 [0.42, 0.64]. This large-scale $\Delta\Delta G_{bind}$ calculation study provides useful information about $\Delta\Delta G_{bind}$ prediction with different AMBER FF combinations and presents valuable suggestions for FF selection in AMBER-TI $\Delta\Delta G_{bind}$ calculations.

Graphical Abstract

[*]**Corresponding Author:** wonpil@lehigh.edu.

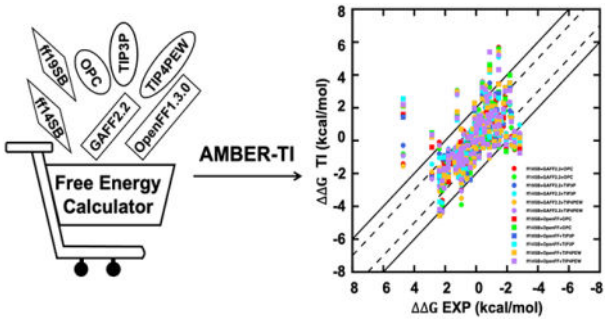
SUPPORTING INFORMATION

The Supporting Information is available free of charge.

The 10 transformations for each protein system (Table S1), the structures of 10 pairs in BACE1 (Table S2), the overall workflow of system preparation and simulation procedure (Figure S1), the correlation between predicted and experimental $\Delta\Delta G_{bind}$ values for 8 protein systems using ff19SB + GAF2.2 + OPC and 12 λ or 21 λ (Figure S2), and the correlation between predicted and experimental $\Delta\Delta G_{bind}$ values for 8 protein systems with 12 FF combinations (Figure S3).

Conflict of Interest

W.I. is the co-founder and CEO of MolCube INC.



INTRODUCTION

Ligand binding interactions often alter target protein conformations,[1–3] and thus protein functions and biological activities can be regulated by ligand binding.[4–6] A ligand with a high binding affinity to a target protein can occupy the ligand binding site, trigger a physiological response, and achieve an expected therapeutic effect with a relatively low ligand concentration.[7] This is the reason why a greater affinity of compounds to their target protein is one of the crucial features of new drug candidates. However, experimental high-throughput ligand screening is costly and time-consuming. Thus, computational protein-ligand binding affinity prediction methods, such as molecular docking,[7, 8] molecular mechanics/generalized Born surface area (MM/GBSA),[9] molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA),[10] and alchemical free energy calculations,[11–14] have become popular since they require less resources and time, but yield reasonable accuracy. Among them, alchemical free energy calculations have shown a high accuracy in ligand binding affinity prediction and have been used as an important tool in computer-aided drug discovery and drug design.

Ligand binding free energy calculations can be classified into absolute and relative ones, which are determined by the thermodynamic end states.[15] The protein-ligand relative binding free energy ($\Delta\Delta G_{bind}$) calculations require less computational time than their absolute counterparts,[16–19] and can be estimated using thermodynamic integration (TI),[20] Bennett's acceptance ratio (BAR),[21] multistate-BAR (MBAR),[22] or unbinned weighted histogram analysis method (UWHAM).[23]

The advances in ligand binding free energy calculations have been fostered by enhanced sampling algorithms, improved force fields (FFs), state-of-the-art high-performance computing architectures, and the tools to generate molecular systems reliably. The FF selection for protein, ligand, and water is vital to obtain accurate ligand binding free energies with alchemical free energy calculations. However, while there are many protein and ligand AMBER FFs and water models available, there are few systematic evaluation studies of these FFs that can inform the best combination for ligand binding free energy calculations.[24, 25] Huggins assessed 3 protein FFs (ff15SB, ff15ipq, AMBER-FB15), 5 water models (SPC/E, TIP3P, TIP3P-FB, TIP4PEW, and TIP4PFB), and 1 ligand FF (GAFF2.1) with two partial charge models (AM1-BCC[26] or RESP[27]) using OpenMM.[24, 28, 29] It was reported that the ff15ipq + TIP3P + GAFF2.1 (AM1-BCC) FF combination presents the best

results. The author showed that there is no clear improvement using RESP for ligand charge in their ligand binding free energy calculations. Gapsys et al. tested AMBER99SB*ILDN and CHARMM36m for proteins, GAFF2.1, OpenFF1.2.0, and CGenFF3.0.1 for ligands using GROMACS.[25, 30] AMBER99SB*ILDN + GAFF2.1 and AMBER99SB*ILDN + OpenFF1.2.0 performed better than CHARMM36m + CGenFF3.0.1. However, to the best of our knowledge, there is no study that comprehensively tests different FFs for AMBER-TI free energy calculations.

In this study, we systematically tested AMBER-TI[31] with large-scale $\Delta\Delta G_{bind}$ calculations. 8 protein systems (BACE1, TYK2, CDK2, MCL1, JNK1, p38, Thrombin, and PTP1B) of the JACS benchmark set[32] were used to examine which FF combination would give better $\Delta\Delta G_{bind}$ prediction. 12 different FF combinations were tested for protein (ff14SB[33] and ff19SB[34]), ligand (GAFF2.2[35] and OpenFF1.3.0[36]), and water models (TIP3P[37], TIP4PEW[38], and OPC[39]). In addition, the consensus FF approach, averaging out the predicted $\Delta\Delta G_{bind}$ values from multiple FFs, is also used to further test the accuracy in $\Delta\Delta G_{bind}$ estimation.[40] This large-scale $\Delta\Delta G_{bind}$ calculation study provides useful information about $\Delta\Delta G_{bind}$ prediction using various AMBER FF combinations and presents a valuable guidance for consensus scoring and FF selection for AMBER-TI $\Delta\Delta G_{bind}$ calculations.

METHODS

System and force field selection

To have a better comparison with the previous studies, we used the same 8 systems as in Wang et al (the so-called JACS benchmark set).[32] The 8 systems and their PDB codes are BACE1 (PDB: 4DJW), TYK2 (PDB: 4GIH), CDK2 (PDB: 1H1Q), MCL1 (PDB: 4HW3), JNK1 (PDB: 4GMX), p38 (PDB: 3FLY), Thrombin (PDB: 2ZFF), and PTP1B (PDB: 2QBS). To systematically and efficiently perform large-scale AMBER-TI $\Delta\Delta G_{bind}$ calculations, we selected 10 transformations from each protein system (Table S1) whose $\Delta\Delta G_{bind}$ values cover a relatively large range based on the experimental results. All 80 transformation systems and inputs were prepared with the 12 different FF combinations using *Free Energy Calculator* in CHARMM-GUI.[41–44] Both complex and ligand systems were prepared for a total of 960 transformations (i.e., 10 transformations × 8 proteins × 12 FFs). The ligand charges were calculated using AM1-Mulliken and AM1-BCC for OpenFF1.3.0 and GAFF2.2, respectively.[26, 45]

AMBER-TI

The following methods were used for all AMBER-TI simulations in this work. $\Delta\Delta G_{bind}$ between two ligands (L0 to L1) is calculated as

$$\Delta\Delta G_{bind}^{L0\rightarrow L1} = \Delta G_{complex}^{L0\rightarrow L1} - \Delta G_{ligand}^{L0\rightarrow L1} \quad (1)$$

where $\Delta G_{complex}^{L0\rightarrow L1}$ and $\Delta G_{ligand}^{L0\rightarrow L1}$ are the alchemical transformations of L0 to L1 in the complex and solution, respectively. In this study, the reported TI free energy values used the trapezoidal rule for the numerical integration to obtain all necessary ΔG values.

Long-range electrostatics was treated with the particle mesh Ewald (PME) method, and the van der Waals interactions were calculated with a cutoff distance of 10 Å.[46, 47] The second-order smoothstep softcore potential, SSC(2), was applied.[31] The values of 0.2 and 50 Å^2 were used for the parameters α and β of SSC(2), respectively. Equilibration was performed for 5 ps employing the NPT (constant particle number, pressure, and temperature) ensemble after minimization in each λ window. AMBER-TI simulations were performed in the NPT ensemble at 300 K and 1 atm (1.0135 bar) with the pmemd.cuda module of AMBER20.[31] All simulations were performed sequentially from λ=0 to λ=1. The alchemical transformations in this work were done using the unified protocol with a 4 fs timestep and the hydrogen mass repartitioning scheme.[48, 49] The last 4 ns of the simulation results of each λ was utilized for the final $\Delta\Delta G_{bind}$ values for 5 ns AMBER-TI simulations, while the last 5 ns was used for 10 ns AMBER-TI simulations. For statistical analysis, 4 independent runs were performed for each pair and the mean value was recorded as $\Delta\Delta G_{bind}$ for all calculations throughout this work. The overall workflow of system preparation and simulation procedure is shown in Figure S1.

We first selected 10 pairs (Table S2) from the BACE1 benchmark set to test three different 12 λ sets and different simulation lengths (1, 5, or 10 ns) per each λ. The same FF combination, ff19SB + GAFF2.2 + OPC, was used to avoid any FF influence on $\Delta\Delta G_{bind}$. The three different 12 λ sets are based on a linear scheme (0.000, 0.091, 0.182, 0.273, 0.364, 0.455, 0.545, 0.636, 0.727, 0.818, 0.909 and 1.000), the scheme used by Song et al. (0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, 0.99078),[50] and the scheme used by Lee et al. (0.0000, 0.0479, 0.1151, 0.2063, 0.3161, 0.4374, 0.5626, 0.6839, 0.7937, 0.8850, 0.9521, and 1.0000).[51] The convergence analysis was applied to ensure that the obtained 12 λ $\Delta\Delta G_{bind}$ values were from the equilibrated simulations. For this, the trajectories used for the final $\Delta\Delta G_{bind}$ were divided into 12 blocks to estimate the cumulative average $\Delta\Delta G_{bind}$ in the forward direction. And, following Yang et al,[52] the cumulative $\Delta\Delta G_{bind}$ values were also obtained from the time-reversed data starting from the end of trajectories but using the same amount of simulation time.

Then, we tested different numbers of λ windows (12 or 21) and different simulation lengths (5 ns or 10 ns) per each λ with ff19SB + GAFF2.2 + OPC for 8 protein systems. Only the third scheme of 12 λ (used by Lee et al.) was used here. For 21 λ windows, λ values are from 0 to 1 with $\Delta\lambda$ = 0.05. The $\Delta\Delta G_{bind}$ values with 12 λ or 21 λ windows and 5 ns or 10 ns simulation times were almost the same (see Results and Discussion).

Therefore, we chose the third 12 λ scheme and 5 ns simulation time per each λ to further test which FF combination is the best. For a consistent comparison, the same ligand-only systems were used if they contained the same FFs for the ligand and water model. For example, the complex systems with ff19SB + GAFF2.2 + OPC and ff14SB + GAFF2.2 + OPC share one ligand system with GAFF2.2 and OPC. Therefore, we built 960 alchemical transformation systems and ran a total of 5,760 (3,840 complex for $\Delta G_{complex}$ and 1,920 ligand systems for ΔG_{ligand}) AMBER-TI simulations. The agreement between the predicted $\Delta\Delta G_{bind}$ values and the experiment was quantified by the mean unsigned error (MUE) and root-mean-square error (RMSE), and Pearson's (r_P), Spearman's (ρ), and Kendell's (τ)

correlations. 95% confidence intervals of MUE, RMSE, r_P, ρ, and τ were obtained with bootstrapping by resampling $\Delta\Delta G_{bind}$ values 5,000 times with replacement.

A consensus FF approach, the average of $\Delta\Delta G_{bind}$ predicted by multiple FFs, was used to further check the accuracy of $\Delta\Delta G_{bind}$ estimation. Gapsys et al. have pioneered the application of the consensus FF approach, which averages the estimated $\Delta\Delta G_{bind}$ values from different FFs and multiple independent runs. For example, they calculated consensus $\Delta\Delta G_{bind}$ by averaging the $\Delta\Delta G_{bind}$ values from two FFs (Amber99SB*ILDN + GAFF2.1 + TIP3P and CHARMM36m+ CGenFF + TIP3P) and showed that the consensus approach yielded higher accuracy than the individual FF.[53] In 2022, Gapsys et al mentioned that they did not include OpenFF into the consensus because of the high similarity between the early OpenFF version and GAFF2.1.[25] However, it has been reported that small variations in FFs may cause significant differences in optimized geometries,[54–56] indicating that minor differences might be enough to alter the geometries of ligands and present different $\Delta\Delta G_{bind}$ values. In addition, both GAFF2.2 and OpenFF1.3.0 had substantial improvement compared to the previous versions. Therefore, we decided to apply the consensus FF approach by averaging $\Delta\Delta G_{bind}$ values from 12 FFs and 4 independent runs of each FF to obtain the consensus results. In addition, to examine the performance of the consensus approach, we calculated the consensus results by averaging $\Delta\Delta G_{bind}$ values from different FF combinations (see Results and Discussion).

RESULTS AND DISCUSSION

Number of λ windows and simulation length for AMBER-TI calculations

For the selected 10 pairs of the BACE1 benchmark set (Table S2), we first tested three different sets of 12 λ windows (see Methods) and different simulation lengths (1, 5, or 10 ns) per each λ window. Each $\Delta\Delta G$ value was obtained from 4 independent runs of 1, 5, or 10 ns per windows simulations. Table 1 shows the results across three tested λ sets and three simulation lengths. Not surprisingly, longer simulations show better agreement with experiments because of the better convergence. In addition, the MUE, RMSE, r_P, ρ, and τ show reasonable accuracy for the predicted $\Delta\Delta G_{bind}$ from 1 ns and 5 ns simulations. To further investigate the accuracy of $\Delta\Delta G_{bind}$ estimation from different simulation lengths, we plotted the predicted $\Delta\Delta G_{bind}$ compared to experimental data (Figures 1A–1C). Clearly, the $\Delta\Delta G_{bind}$ values from more extended simulations tend to be closer to the experiments since more results from 10 ns simulations (blue dots in Figure 1A–1C) are within 1 kcal/mol (e.g., dots distribute within the dashed lines), followed by 5 ns simulations (yellow dots), and then 1 ns simulations (red dots). We observed the opposite signs between predicted $\Delta\Delta G_{bind}$ and experimental ones from the results of 1 ns length of simulations for the transformation from CAT-13d to CAT-13h. For example, the experimental $\Delta\Delta G_{bind}$ value for this pair is 0.84 kcal/mol, but all three λ sets of 1 ns simulations showed the $\Delta\Delta G_{bind}$ values (green in Figure 1A–1C) are all negative (−0.69 ± 0.18, −0.32 ± 0.31, and −0.26 ± 0.20 kcal/mol. Therefore, to obtain higher accuracy of $\Delta\Delta G_{bind}$, we decided to use a longer simulation time (5 ns or 10 ns) for the following study. For the performance comparison of different λ sets, three λ sets show similar results in terms of MUE, RMSE, r_P, ρ, and τ and their 95% confidence

intervals (Table 1). Therefore, only the third scheme (used by Lee et al.)[51] was used for the following study.

In addition, to ensure our results from the 12 λ and 5 ns simulation length are from the equilibrated simulations, the convergence analysis was performed. The convergence of $\Delta\Delta G_{bind}$ from forward and time-reversed data and the overlap matrix for the transformation pair from CAT-13j to CAT-4o are shown in Figure 2A. As reported by Klimovich et al.,[57] the forward and time-reversed free energy estimations agree within errors, indicating that $\Delta\Delta G_{bind}$ is converged well using 12 λ windows. Also, previous studies reported that the overlap matrix should be at least tridiagonal (i.e., green in Figure 2A) and the values in these diagonals should be higher than 0.03 to obtain trustworthy results.[57] Therefore, both the forward and time-reversed free energy estimations and the overlap matrix show a good convergence and reliability of our simulations. The forward and time-reversed free energy estimations for the other 9 pairs are in Figure 2B, also indicating that our simulations with 12 λ windows are well converged.

We also used TI-3 (cubic spline interpolation), BAR, and MBAR to obtain $\Delta\Delta G_{bind}$ for the same pairs and the same trajectories to validate our analysis from TI (based on the trapezoidal rule). TI-3 performs the integration from a natural cubic spline interpolation to obtain the free energy values. Figure 1D is a spider plot comparing the results from TI and other estimation methods using the results from the third 12 λ scheme (e.g., used by Lee et al.) and 5 ns simulations. The range of estimation difference between TI and BAR (blue in Figure 1D) is from 0.00 ± 0.07 (pair CAT-4m → CAT-4p) to 0.07 ± 0.38 (pair CAT-17b → CAT-17e) kcal/mol. For the comparison between TI and MBAR (orange in Figure 1D), the largest difference is 0.13 ± 0.21 kcal/mol (pair CAT-13d → CAT-13h) and the smallest one is 0.01 ± 0.05 kcal/mol (pair CAT-4m → CAT-4p). Between TI and TI3 (purple in Figure 1D), the difference is from 0.00 to 0.08 kcal/mol. Overall, the estimated results obtained from TI, TI-3, BAR, and MBAR are almost indistinguishable. Therefore, our calculation of TI (e.g., trapezoidal rule) appears to be appropriate and the trapezoidal numerical integration is used to do the TI calculation for the following study.

Next, we further tested two simulation lengths (5 or 10 ns per window) using 8 protein sets. In addition, two different numbers of λ windows (12 or 21 λ windows) were tested. Our results in Table 2 (MUE, RMSE, r_P, ρ, τ, and their 95% confidence intervals) and Figure S2 show that the $\Delta\Delta G_{bind}$ values with 12 λ or 21 λ windows and 5 ns or 10 ns simulation time are very similar, indicating that 12 λ windows and 5 ns simulation time are sufficient to obtain reliable $\Delta\Delta G_{bind}$ with 4 independent AMBER-TI runs.

All AMBER-TI simulations for the BACE1 complex systems (~86,000 atoms) were conducted using one GTX 1080 Ti or one RTX 2080 Ti GPU with a speed of approximately 71 ns/day or 169 ns/day, respectively. Therefore, 5 ns of 12 λ window simulations take ~20 (one GTX 1080 Ti GPU) or ~8.5 (one RTX 2080 Ti GPU) hours with a 4 fs timestep, while 10 ns of 24 λ window simulations take ~81 (one GTX 1080 Ti GPU) or ~34 (one RTX 2080 Ti GPU) hours. Using 12 λ windows and 5 ns simulations is ~4 times faster and yields one $\Delta\Delta G_{bind}$ value (with 4 independent runs) in ~1.5 days (one RTX 2080 Ti GPU) for the BACE1 complex system.

Overall performance of 12 FF combinations for $\Delta\Delta G_{bind}$ prediction

Using 12 λ windows, 5 ns simulation time in each window, and 4 independent AMBER-TI runs for each pair, we examined the FF influence on $\Delta\Delta G_{bind}$. A statistical analysis of the $\Delta\Delta G_{bind}$ prediction performance of each FF combination is shown in Table 3. All FF combinations present a good agreement with experimental data based on MUE, RMSE, r_P, ρ, and τ. Using ff14SB + GAFF2.2 + TIP3P, we can obtain the best rank correlations (r_P = 0.64 [0.52, 0.76], ρ = 0.73 [0.58, 0.83], and τ =0.54 [0.42, 0.64]) for the 80 transformations (bold in Table 3), which is similar to Wang et al. using FEP+.[32] The highest r_P, ρ, and τ inform that this FF combination provides the most correct trend compared to experimental data. In addition, the lowest MUE (0.87 [0.69, 1.07]) and RMSE (1.22 [0.94, 1.50]) indicate that $\Delta\Delta G_{bind}$ calculations are closest to experimental results, i.e., the highest accuracy of $\Delta\Delta G_{bind}$ prediction, which is slightly better than Wang et al.[32] In comparison, ff19SB + OpenFF + TIP4PEW and ff14SB + OpenFF + TIP4PEW show the highest MUE (1.01 [0.80, 1.25]) and RMSE (1.51 [1.04, 1.92]), respectively, while ff14SB + OpenFF + OPC presents the lowest r_P (0.55 [0.33, 0.76]), ρ (0.65[0.47, 0.79]), and τ (0.48[0.35, 0.61]). Overall, the 95% confidence intervals are in the acceptable ranges for all the tested FF combinations, although 95% confidence intervals for some FF combinations are slightly broader.

Figure S3 shows the correlation between predicted and experimental $\Delta\Delta G_{bind}$ for 12 FF combinations, showing 960 $\Delta\Delta G_{bind}$ values (8 systems × 10 pairs × 12 FF combinations) and each $\Delta\Delta G_{bind}$ is the mean of 4 independent runs. Among these 960 $\Delta\Delta G_{bind}$ values, 69% pairs are within 1.0 kcal/mol from their experimental value, and 88% predictions show less than 2 kcal/mol deviation. The correlation between predicted and experimental $\Delta\Delta G_{bind}$ for each individual FF combination is shown in Figure 3. With ff14SB + GAFF2.2 + TIP3P, only 6 out of 80 pairs (7.5%) deviate more than 2 kcal/mol from their experimental value, while there are more pairs showing more than 2 kcal/mol deviation in other FF combinations (Table 3). Note that there are 9 pairs showing more than 2 kcal/mol deviation in Wang et al. using FEP+.[32] The performance of ff14SB + GAFF2.2 + TIP3P, therefore, is the best, while other FF combinations also present a good agreement with experimental $\Delta\Delta G_{bind}$.

Impact of protein, ligand, or water FFs on $\Delta\Delta G_{bind}$ prediction

Next, we investigated the influence of protein, ligand, and water FFs separately on the accuracy of $\Delta\Delta G_{bind}$ prediction. ff14SB showed improved accuracy of protein side chain and backbone parameters compared to ff99SB,[60] while ff19SB showed further improvement of backbone conformational profiles for all 20 amino acid residues. To explore the effect of protein FFs on $\Delta\Delta G_{bind}$ prediction, we divided 12 FF combinations into 2 groups based on the protein FFs. As shown in Table 4, the average values of MUE and RMSE inform that the usage of ff19SB or ff14SB has no statistical difference on $\Delta\Delta G_{bind}$. In particular, all FF combinations with either protein FF show the same rank correlation (i.e., the similar r_P, ρ, and τ).

Two ligand FFs are tested in this study: the updated second version of GAFF and the SMIRKS-based FFs from the Open FF Initiative (OpenFF1.3.0). GAFF2.2, compared to GAFF2.1, added more than 50% of bond and angle parameters and more than 35% of torsional angle parameters. OpenFF 1.0.0 showed an accuracy similar to that of GAFF2.1,

OPLS3e, and CGenFF when estimating $\Delta\Delta G_{bind}$ in the study of Qiu et al.[36] On the premise to keep the improvement of OpenFF 1.1.0 and 1.2.0, OpenFF 1.3.0 improved the performance in reproducing amide torsional energy profiles and added the appropriate torsion parameters for dialkyl amides. To explore the impact of GAFF2.2 and OpenFF 1.3.0 on $\Delta\Delta G_{bind}$ prediction, we split 12 FF combinations into 2 groups based on the ligand FF. Table 5 shows the average values of MUE, RMSE, r_P, ρ, and τ and their 95% confidence intervals. Overall, there is no statistically significant difference in performance between OpenFF1.3.0 and GAFF2.2 in this study.

OPC, TIP3P, and TIP4PEW are common water models that are frequently used in all-atom simulations. In particular, a water model is very crucial for calculations of both $\Delta G_{complex}$ and ΔG_{ligand} in $\Delta\Delta G_{bind}$. To explore the effects of the water model on $\Delta\Delta G_{bind}$ prediction, we divided 12 FF combinations into 3 groups based on the water model. As shown in Table 6, the average values of MUE, RMSE, r_P, ρ, τ, and their 95% confidence intervals inform that the usage of different water models has a statistically unnoticeable effect on $\Delta\Delta G_{bind}$ prediction.

$\Delta\Delta G_{bind}$ prediction using a consensus force field approach

A consensus approach using the average of the predicted $\Delta\Delta G_{bind}$ values from multiple FFs could help minimize a bias from the parametrization of individual FFs and provide higher accuracy of $\Delta\Delta G_{bind}$ prediction.[40] The $\Delta\Delta G_{bind}$ estimation accuracy with the consensus approach (case 1 with all 12 FFs in Table 7) is improved or similar compared to those from individual FFs (Table 3). For instance, using the consensus FF, 59 out of 80 pairs within 1.0 kcal/mol from their experimental data and 72 pairs within the deviation of 2.0 kcal/mol (case 1), while the individual FF ff19SB + GAFF2.2 + OPC (case 1 in Table 3) shows 57 and 70 pairs have less than 1 and 2 kcal/mol deviation from experiments, respectively. Although we could see some improvements in accuracy from the consensus approach, each $\Delta\Delta G_{bind}$ value was averaged from 12 FFs and 4 independent runs. It is not economical to obtain one $\Delta\Delta G_{bind}$ value from 12 FFs, considering the time for system preparation with various FFs and the computing resources. Therefore, we tested a consensus FF comprising GAFF2.2 and OpenFF1.3.0. For example, the results of ff19SB + OPC consensus FF were obtained by averaging $\Delta\Delta G_{bind}$ values from ff19SB + GAFF2.2 + OPC and ff19SB + OpenFF + OPC. It is worth noting that, using consensus FFs (Table 7), the number of pairs within 1 kcal/mol deviation from experiments is always more than their individual FFs (Table 3). For example, the ff19SB + OPC consensus FF contain 61 pairs (case 2 in Table 7) deviate from their experiments within 1 kcal/mol, while ff19SB + GAFF2.2 + OPC and ff19SB + OpenFF + OPC are 57 and 58 pairs (cases 1 and 7 in Table 3), respectively. In addition, the consensus FF approach shows a better $\Delta\Delta G_{bind}$ estimation than OpenFF1.3.0 in terms of MUE, RMSE, r_P, ρ, and τ. Most consensus FF combinations also show a similar or better $\Delta\Delta G_{bind}$ prediction than GAFF2.2, except for the cases 5 and 7 (bold) in Table 7. Our result indicates that the consensus approach is an applicable strategy to obtain similar or relatively higher accuracy and could be more trustworthy values in $\Delta\Delta G_{bind}$ estimation. Overall, the tested consensus FFs (FF combinations 2 to 7 in Table 7) show a similar performance including the 95% confidence intervals.

CONCLUSIONS

Nowadays, owing to both economic and efficient perspectives, virtual high throughput screening using alchemical free energy calculations gains more attention in drug discovery and drug design. To calculate reliable $\Delta\Delta G_{bind}$ values using AMBER-TI, we previously showed that at least 4 independent runs with 12 λ windows and 5 ns simulation time per window are sufficient for soluble protein.[44] In this study, we first confirmed this by examining the effects of using different numbers of λ windows and simulation times per window on $\Delta\Delta G_{bind}$. Then, we investigated the performance of 12 FF combinations for AMBER-TI $\Delta\Delta G_{bind}$ prediction. An AMBER-TI run (with 12 λ windows and 5 ns simulation time) for ~86,000 atoms takes ~8.5 hours using one RTX 2080 Ti GPU, meaning that one $\Delta\Delta G_{bind}$ value (with 4 independent runs) can be calculated in ~1.5 days. The overall real computational time can be further reduced by performing the simulations from different windows in parallel across multiple GPUs. With this speed, we believe that screening hundreds of lead compounds can be achieved in a timely manner with highly accurate $\Delta\Delta G_{bind}$ estimation. One may want to use larger λ windows or longer simulation time if their systems are more complex than the current 8 protein systems or when $\Delta\Delta G_{bind}$ values do not converge within 5 ns.

Our large-scale $\Delta\Delta G_{bind}$ calculation study provides good practices of FF selection for AMBER-TI $\Delta\Delta G_{bind}$ prediction. The accuracy of ff14SB + TIP3P + GAFF2.2 in terms of MUE, RMSE, r_P, ρ, and τ is the best among the 12 FF combinations tested in this study. This FF combination shows higher accuracy than Wang et al[32] using FEP+ in terms of MUE and RMSE. The ff14SB protein model being trained using TIP3P water model could explain why this FF combination presents great $\Delta\Delta G_{bind}$ prediction. Both ff19SB and ff14SB show a similar accuracy, indicating that the usage of ff19SB or ff14SB has a statistically negligible effect on $\Delta\Delta G_{bind}$ estimation and both protein FFs are suitable for $\Delta\Delta G_{bind}$ calculations. Also, there is no statistically significant difference in performance between OpenFF1.3.0 and GAFF2.2 and both ligand FFs present good performance. Note that the latest OpenFF2.0.0 was recently released, so it will be interesting to check the accuracy of OpenFF2.0.0 in the future. In addition, the water models used in this study also show comparable accuracy on $\Delta\Delta G_{bind}$ prediction.

The application of a consensus FF approach can yield similar or relatively higher accuracy in $\Delta\Delta G_{bind}$ estimation and thus is a good strategy to obtain $\Delta\Delta G_{bind}$ values. However, it is worthy of note that the preparation of AMBER-TI systems and implementation of simulations with different FFs require additional time and resources, especially for large numbers of ligands. Generally, due to the limited test cases and systems, any FF recommendation from a benchmark study is biased. A more significant number of transformation pairs and systems are always required to draw a more solid conclusion of the FF recommendation for $\Delta\Delta G_{bind}$ prediction. Nonetheless, for the individual FF combinations, our benchmark testing recommends ff14SB + TIP3P + GAFF2.2 for virtual high throughput screening and lead optimization projects. In addition, since all tested individual FF combinations in this study present comparable accuracy for $\Delta\Delta G_{bind}$ prediction, one could consider their FF combination based on other desired properties such as improved protein behaviors (e.g., no protein aggregation).[61] We hope that this benchmark

study provides some useful suggestions for $\Delta\Delta G_{bind}$ estimation with AMBER FFs and AMBER-TI.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors thank to Junmei Wang for sharing GAFF2.2 prior to its publication. This work was supported in part by the grants from NIH GM126140 and GM138472 (to W.I.). S.K. was supported by an individual grant of Korea Institute for Advanced Study (CG080501).

DATA AND SOFTWARE AVAILABILITY

All the structures (e.g., complex and solvent systems) and simulation input files for the 8 benchmark cases and 12 FF combinations are freely available on GitHub: https://github.com/haz519/AMBER-TI. All the calculated data and the analysis script examples are also freely available on the same GitHub. All the initial complex and ligand systems were prepared using *Free Energy Calculator* module in CHARMM-GUI. All the necessary analysis scripts were also provided by CHARMM-GUI. *Free Energy Calculator* module can be accessed through the following link: https://www.charmm-gui.org/input/fec.

References:

1. Weikl TR; Paul F, Conformational selection in protein binding and function. Protein Sci 2014, 23 (11), 1508–18. [PubMed: 25155241]
2. Amaral M; Kokh DB; Bomke J; Wegener A; Buchstaller HP; Eggenweiler HM; Matias P; Sirrenberg C; Wade RC; Frech M, Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. Nat Commun 2017, 8 (1), 2276. [PubMed: 29273709]
3. Kumar A; Jernigan RL, Ligand Binding Introduces Significant Allosteric Shifts in the Locations of Protein Fluctuations. Front Mol Biosci 2021, 8, 733148. [PubMed: 34540902]
4. Popov VM; Yee WA; Anderson AC, Towards in silico lead optimization: scores from ensembles of protein/ligand conformations reliably correlate with biological activity. Proteins 2007, 66 (2), 375–87. [PubMed: 17078091]
5. Schena A; Griss R; Johnsson K, Modulating protein activity using tethered ligands with mutually exclusive binding sites. Nat Commun 2015, 6, 7830. [PubMed: 26198003]
6. Shortridge MD; Bokemper M; Copeland JC; Stark JL; Powers R, Correlation between protein function and ligand binding profiles. J Proteome Res 2011, 10 (5), 2538–45. [PubMed: 21366353]
7. Salahudeen MS; Nishtala PS, An overview of pharmacodynamic modelling, ligand-binding approach and its application in clinical practice. Saudi Pharm J 2017, 25 (2), 165–175. [PubMed: 28344466]
8. Kitchen DB; Decornez H; Furr JR; Bajorath J, Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 2004, 3 (11), 935–49. [PubMed: 15520816]
9. Ylilauri M; Pentikainen OT, MMGBSA as a tool to understand the binding affinities of filamin-peptide interactions. J Chem Inf Model 2013, 53 (10), 2626–33. [PubMed: 23988151]
10. Miller BR 3rd; McGee TD Jr.; Swails JM; Homeyer N; Gohlke H; Roitberg AE, MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. J Chem Theory Comput 2012, 8 (9), 3314–21. [PubMed: 26605738]
11. Mobley DL; Klimovich PV, Perspective: Alchemical free energy calculations for drug discovery. J Chem Phys 2012, 137 (23), 230901. [PubMed: 23267463]

12. Wade AD; Bhati AP; Wan S; Coveney PV, Alchemical Free Energy Estimators and Molecular Dynamics Engines: Accuracy, Precision, and Reproducibility. J Chem Theory Comput 2022, 18 (6), 3972–3987. [PubMed: 35609233]
13. Kuhn M; Firth-Clark S; Tosco P; Mey A; Mackey M; Michel J, Assessment of Binding Affinity via Alchemical Free-Energy Calculations. J Chem Inf Model 2020, 60 (6), 3120–3130. [PubMed: 32437145]
14. Schindler CEM; Baumann H; Blum A; Bose D; Buchstaller HP; Burgdorf L; Cappel D; Chekler E; Czodrowski P; Dorsch D; Eguida MKI; Follows B; Fuchss T; Gradler U; Gunera J; Johnson T; Jorand Lebrun C; Karra S; Klein M; Knehans T; Koetzner L; Krier M; Leiendecker M; Leuthner B; Li L; Mochalkin I; Musil D; Neagu C; Rippmann F; Schiemann K; Schulz R; Steinbrecher T; Tanzer EM; Unzue Lopez A; Viacava Follis A; Wegener A; Kuhn D, Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. J Chem Inf Model 2020, 60 (11), 5457–5474. [PubMed: 32813975]
15. Cournia Z; Allen BK; Beuming T; Pearlman DA; Radak BK; Sherman W, Rigorous Free Energy Simulations in Virtual Screening. J Chem Inf Model 2020, 60 (9), 4153–4169. [PubMed: 32539386]
16. Jiang W; Chipot C; Roux B, Computing Relative Binding Affinity of Ligands to Receptor: An Effective Hybrid Single-Dual-Topology Free-Energy Perturbation Approach in NAMD. J Chem Inf Model 2019, 59 (9), 3794–3802. [PubMed: 31411473]
17. Cournia Z; Allen B; Sherman W, Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. J Chem Inf Model 2017, 57 (12), 2911–2937. [PubMed: 29243483]
18. Cappel D; Hall ML; Lenselink EB; Beuming T; Qi J; Bradner J; Sherman W, Relative Binding Free Energy Calculations Applied to Protein Homology Models. J Chem Inf Model 2016, 56 (12), 2388–2400. [PubMed: 28024402]
19. Raman EP; Paul TJ; Hayes RL; Brooks CL, Automated, Accurate, and Scalable Relative Protein-Ligand Binding Free-Energy Calculations Using Lambda Dynamics. Journal of Chemical Theory and Computation 2020, 16 (12), 7895–7914. [PubMed: 33201701]
20. Kumar J; Dey TK; Sinha SK, Semiclassical statistical mechanics of hard-body fluid mixtures. J Chem Phys 2005, 122 (22), 224504. [PubMed: 15974688]
21. Bennett CH, Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. J Comput Phys 1976, 22 (2), 245–268.
22. Shirts MR; Chodera JD, Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys 2008, 129 (12), 124105. [PubMed: 19045004]
23. Tan Z; Gallicchio E; Lapelosa M; Levy RM, Theory of binless multi-state free energy estimation with applications to protein-ligand binding. J Chem Phys 2012, 136 (14), 144102. [PubMed: 22502496]
24. Huggins DJ, Comparing the Performance of Different AMBER Protein Forcefields, Partial Charge Assignments, and Water Models for Absolute Binding Free Energy Calculations. J Chem Theory Comput 2022, 18 (4), 2616–2630. [PubMed: 35266690]
25. Gapsys V; Hahn DF; Tresadern G; Mobley DL; Rampp M; de Groot BL, Pre-Exascale Computing of Protein-Ligand Binding Free Energies with Open Source Software for Drug Design. J Chem Inf Model 2022, 62 (5), 1172–1177. [PubMed: 35191702]
26. Jakalian A; Jack DB; Bayly CI, Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem 2002, 23 (16), 1623–41. [PubMed: 12395429]
27. Bayly CI C. P; Cornell W; Kollman PA, A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J. Phys. Chem 1993.
28. Friedrichs MS; Eastman P; Vaidyanathan V; Houston M; Legrand S; Beberg AL; Ensign DL; Bruns CM; Pande VS, Accelerating molecular dynamic simulation on graphics processing units. J Comput Chem 2009, 30 (6), 864–72. [PubMed: 19191337]
29. Eastman P; Pande VS, OpenMM: A Hardware Independent Framework for Molecular Simulations. Comput Sci Eng 2015, 12 (4), 34–39. [PubMed: 26146490]

30. Hess B; Kutzner C; van der Spoel D; Lindahl E, GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput 2008, 4 (3), 435–47. [PubMed: 26620784]

31. Lee TS; Allen BK; Giese TJ; Guo Z; Li P; Lin C; McGee TD Jr.; Pearlman DA; Radak BK; Tao Y; Tsai HC; Xu H; Sherman W; York DM, Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. J Chem Inf Model 2020, 60 (11), 5595–5623. [PubMed: 32936637]

32. Wang L; Wu Y; Deng Y; Kim B; Pierce L; Krilov G; Lupyan D; Robinson S; Dahlgren MK; Greenwood J; Romero DL; Masse C; Knight JL; Steinbrecher T; Beuming T; Damm W; Harder E; Sherman W; Brewer M; Wester R; Murcko M; Frye L; Farid R; Lin T; Mobley DL; Jorgensen WL; Berne BJ; Friesner RA; Abel R, Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. J Am Chem Soc 2015, 137 (7), 2695–703. [PubMed: 25625324]

33. Maier JA; Martinez C; Kasavajhala K; Wickstrom L; Hauser KE; Simmerling C, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 2015, 11 (8), 3696–713. [PubMed: 26574453]

34. Tian C; Kasavajhala K; Belfon KAA; Raguette L; Huang H; Migues AN; Bickel J; Wang Y; Pincay J; Wu Q; Simmerling C, ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. J Chem Theory Comput 2020, 16 (1), 528–552. [PubMed: 31714766]

35. He X; Man VH; Yang W; Lee TS; Wang J, A fast and high-quality charge model for the next generation general AMBER force field. J Chem Phys 2020, 153 (11), 114502. [PubMed: 32962378]

36. Qiu Y; Smith DGA; Boothroyd S; Jang H; Hahn DF; Wagner J; Bannan CC; Gokey T; Lim VT; Stern CD; Rizzi A; Tjanaka B; Tresadern G; Lucas X; Shirts MR; Gilson MK; Chodera JD; Bayly CI; Mobley DL; Wang LP, Development and Benchmarking of Open Force Field v1.0.0-the Parsley Small-Molecule Force Field. J Chem Theory Comput 2021, 17 (10), 6262–6280. [PubMed: 34551262]

37. Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML, Comparison of Simple Potential Functions for Simulating Liquid Water. J Chem Phys 1983, 79 (2), 926–935.

38. Horn HW; Swope WC; Pitera JW; Madura JD; Dick TJ; Hura GL; Head-Gordon T, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. J Chem Phys 2004, 120 (20), 9665–9678. [PubMed: 15267980]

39. Izadi S; Anandakrishnan R; Onufriev A, Building water models, a different approach. Abstr Pap Am Chem S 2015, 250.

40. Khalak Y; Tresadern G; Aldeghi M; Baumann HM; Mobley DL; de Groot BL; Gapsys V, Alchemical absolute protein-ligand binding free energies for drug design. Chem Sci 2021, 12 (41), 13958–13971. [PubMed: 34760182]

41. Jo S; Kim T; Iyer VG; Im W, CHARMM-GUI: a web-based graphical user interface for CHARMM. J Comput Chem 2008, 29 (11), 1859–65. [PubMed: 18351591]

42. Lee J; Hitzenberger M; Rieger M; Kern NR; Zacharias M; Im W, CHARMM-GUI supports the Amber force fields. J Chem Phys 2020, 153 (3).

43. Kim S; Oshima H; Zhang H; Kern NR; Re S; Lee J; Roux B; Sugita Y; Jiang W; Im W, CHARMM-GUI Free Energy Calculator for Absolute and Relative Ligand Solvation and Binding Free Energy Simulations. Journal of Chemical Theory and Computation 2020, 16 (11), 7207–7218. [PubMed: 33112150]

44. Zhang H; Kim S; Giese TJ; Lee TS; Lee J; York DM; Im W, CHARMM-GUI Free Energy Calculator for Practical Ligand Binding Free Energy Simulations with AMBER. J Chem Inf Model 2021, 61 (9), 4145–4151. [PubMed: 34521199]

45. Mulliken SR, Electronic Population Analysis on LCAOMO Molecular Wave Functions. I. J. Chem. Phys 1955, 23 (10), 1833–1840.

46. Darden T; York D; Pedersen L, Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. J Chem Phys 1993, 98 (12), 10089–10092.

47. Essmann U; Perera L; Berkowitz ML; Darden T; Lee H; Pedersen LG, A Smooth Particle Mesh Ewald Method. J Chem Phys 1995, 103 (19), 8577–8593.

48. Hopkins CW; Le Grand S; Walker RC; Roitberg AE, Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. Journal of Chemical Theory and Computation 2015, 11 (4), 1864–1874. [PubMed: 26574392]

49. Gao Y; Lee J; Smith IPS; Lee H; Kim S; Qi YF; Klauda JB; Widmalm G; Khalid S; Im W, CHARMM-GUI Supports Hydrogen Mass Repartitioning and Different Protonation States of Phosphates in Lipopolysaccharides. Journal of Chemical Information and Modeling 2021, 61 (2), 831–839. [PubMed: 33442985]

50. Song LF; Lee TS; Zhu C; York DM; Merz KM, Using AMBER18 for Relative Free Energy Calculations. Journal of Chemical Information and Modeling 2019, 59 (7), 3128–3135. [PubMed: 31244091]

51. Lee T T. H G. A; Giese T; and York D, Robust, Efficient and Automated Methods for Accurate Prediction of Protein-Ligand Binding Affinities in AMBER Drug Discovery Boost. ChemRxiv 2021.

52. Yang W; Bitetti-Putzer R; Karplus M, Free energy simulations: Use of reverse cumulative averaging to determine the equilibrated region and the time required for convergence. J Chem Phys 2004, 120 (6), 2618–2628. [PubMed: 15268405]

53. Gapsys V; Perez-Benito L; Aldeghi M; Seeliger D; Van Vlijmen H; Tresadern G; de Groot BL, Large scale relative protein ligand binding affinities using non-equilibrium alchemy. Chem Sci 2020, 11 (4), 1140–1152.

54. Ehrman JN; Lim VT; Bannan CC; Thi N; Kyu DY; Mobley DL, Improving small molecule force fields by identifying and characterizing small molecules with inconsistent parameters. J Comput Aided Mol Des 2021, 35 (3), 271–284. [PubMed: 33506360]

55. Fennell CJ; Wymer KL; Mobley DL, A fixed-charge model for alcohol polarization in the condensed phase, and its role in small molecule hydration. J Phys Chem B 2014, 118 (24), 6438–46. [PubMed: 24702668]

56. Sellers BD; James NC; Gobbi A, A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. J Chem Inf Model 2017, 57 (6), 1265–1275. [PubMed: 28485585]

57. Klimovich PV; Shirts MR; Mobley DL, Guidelines for the analysis of free energy calculations. J Comput Aided Mol Des 2015, 29 (5), 397–411. [PubMed: 25808134]

58. Mey A; Allen BK; Macdonald HEB; Chodera JD; Hahn DF; Kuhn M; Michel J; Mobley DL; Naden LN; Prasad S; Rizzi A; Scheen J; Shirts MR; Tresadern G; Xu H, Best Practices for Alchemical Free Energy Calculations [Article v1.0]. Living J Comput Mol Sci 2020, 2 (1).

59. Hahn DF; Bayly CI; Macdonald HEB; Chodera JD; Gapsys V; Mey A; Mobley DL; Benito LP; Schindler CEM; Tresadern G; Warren GL, Best practices for constructing, preparing, and evaluating protein-ligand binding affinity benchmarks. arXiv 2021.

60. Hornak V; Abel R; Okur A; Strockbine B; Roitberg A; Simmerling C, Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 2006, 65 (3), 712–25. [PubMed: 16981200]

61. Abriata AL; Peraro DM, Assessment of transferable forcefields for protein simulations attests improved description of disordered states and secondary structure propensities, and hints at multi-protein systems as the next challenge for optimization. Comput Struct Biotechnol J 2021, 19, 2626–2636. [PubMed: 34025949]

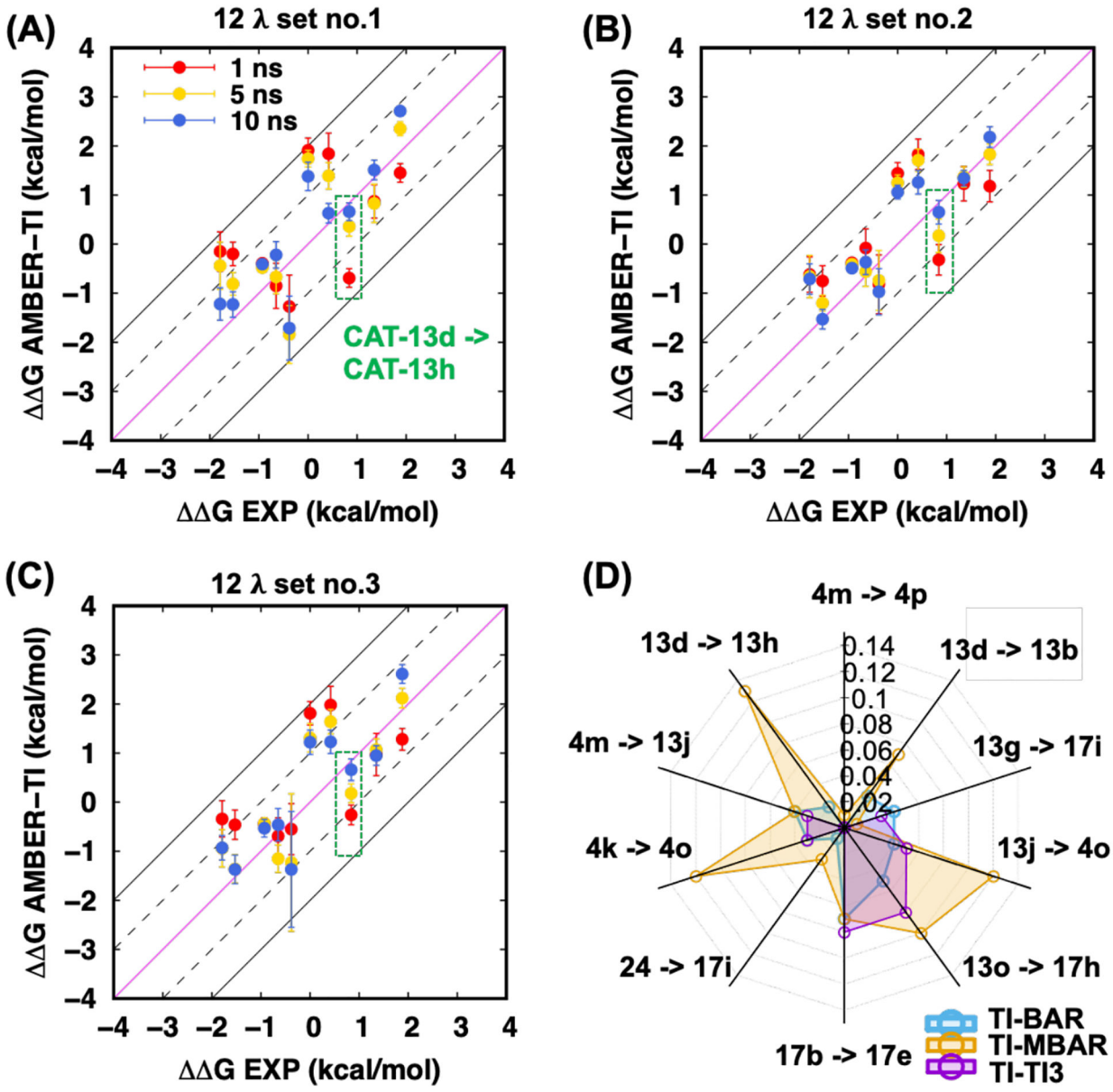


Figure 1.

Correlation between predicted and experimental $\Delta\Delta G_{bind}$ for BACE systems using ff19SB + GAFF2.2 +OPC FF combination and **(A)** the linear λ scheme, **(B)** the scheme used by Song et al,[50] or **(C)** the scheme used by Lee et al.[51] Different simulation time lengths used different colors. The error bars are the standard errors of 4 independent AMBER-TI runs. **(D)** The comparison of $\Delta\Delta G_{bind}$ estimation using TI (based on the trapezoidal rule) and other analysis methods (BAR, MBAR, or TI-3 (cubic spline interpolation)). The X-axis is the transformation pairs, and the Y-axis is the differences in $\Delta\Delta G_{bind}$ values between corresponding calculation methods. The difference values become higher as the point reaches the edge of the spike. Comparisons between the methodologies TI and BAR, TI and MBAR, and TI and TI-3 are shown in blue, orange, and purple, respectively.

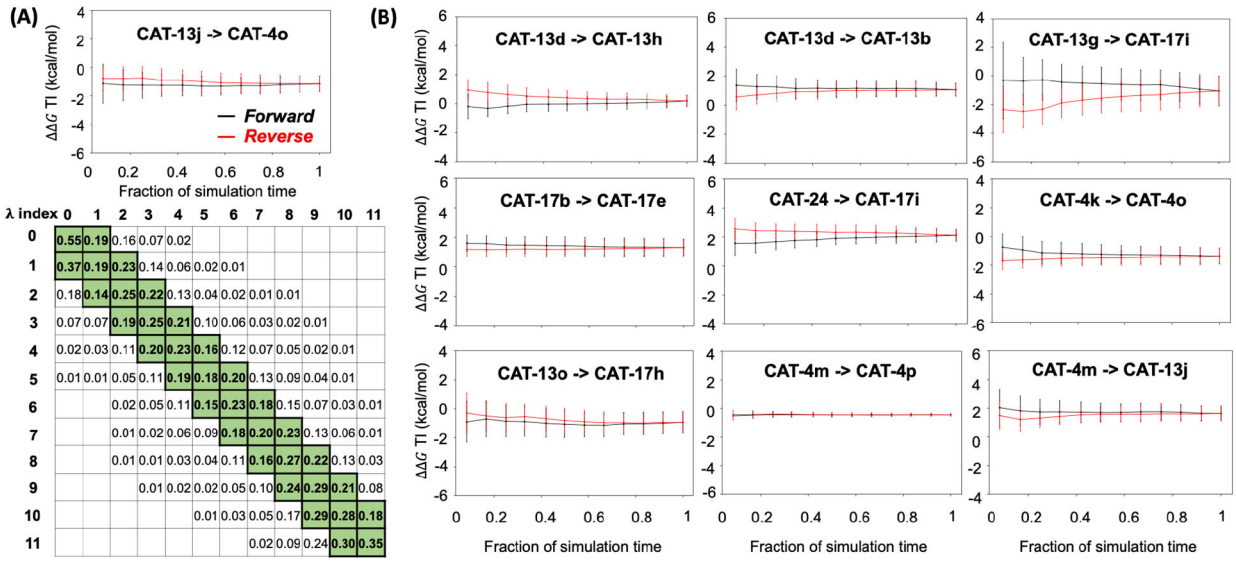


Figure 2.

(A) Top: free energy convergence as a function of time for pair CAT-13j to CAT-4o with a 12 λ window scheme (used by Lee et al.),[51] 5 ns simulation in each window, and ff19SB + GAFF2.2 + OPC. The $\Delta\Delta G_{bind}$ convergence was estimated from the forward (black) and time-reversed (red) data of the last 4 ns trajectory. Bottom: overlap matrix **O** for complex system averaged from 4 independent runs. The elements $\mathbf{O}_{ij}$ are the probabilities of observing a sample in state i (i^{th} row) from state j (j^{th} column). For example, a sample collected in state 1 from state 2 simulation is 0.23. **(B)** Free energy convergence as a function of time for the other 9 pairs in the BACE1 benchmark set.

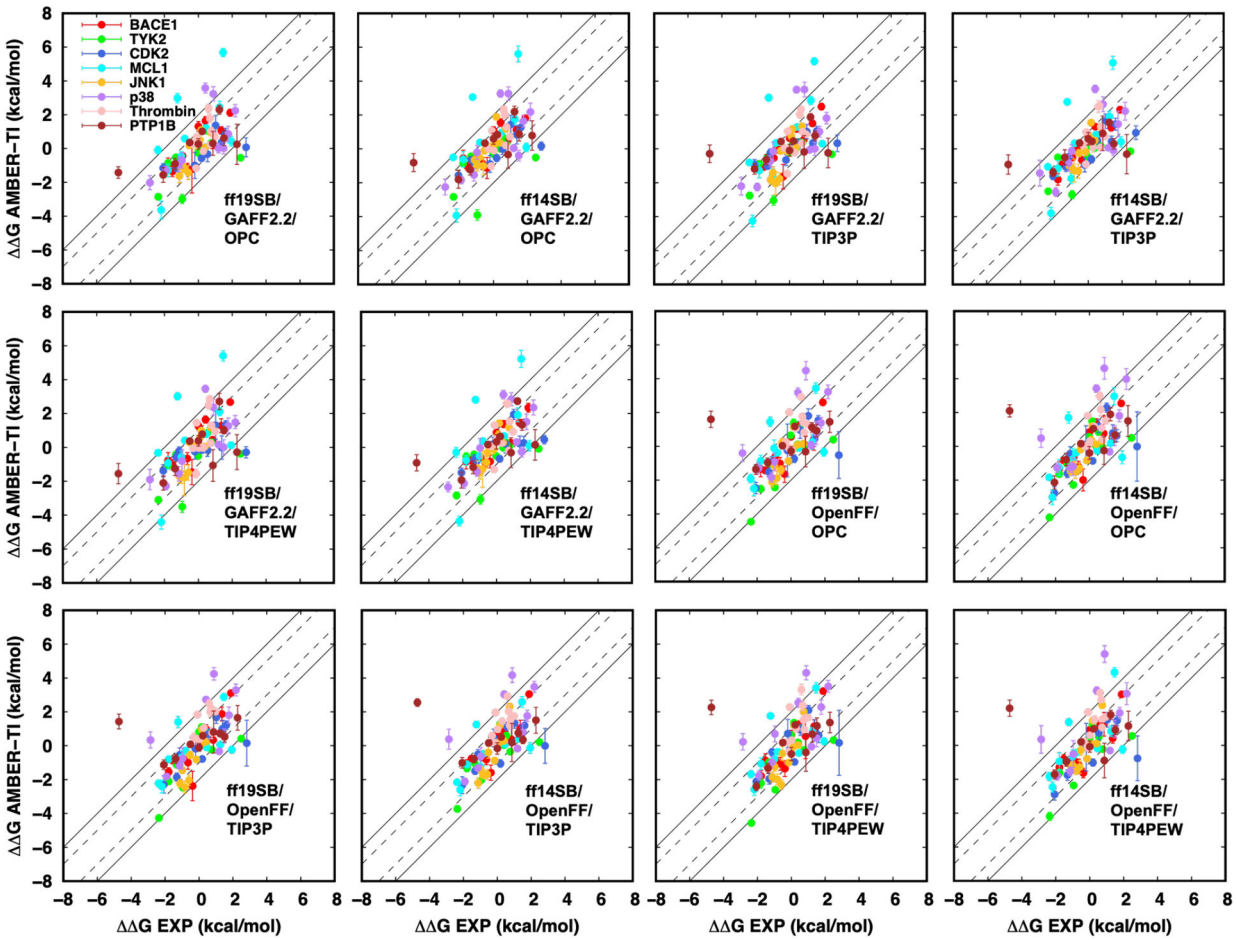


Figure 3.

Correlation between predicted and experimental $\Delta\Delta G_{bind}$ for 8 protein systems (10 pairs for each protein) with 12 different FF combinations. Different systems used different colors. The error bars are the standard errors of 4 independent AMBER-TI runs.

Table 1.

MUE, RMSE, r_P, ρ, and τ between calculated and experimental $\Delta\Delta G_{bind}$ for the BACE1 benchmark set with three different 12 λ schemes and simulation lengths using ff19SB + GAFF2.2 + OPC.

λ sets	Simulation length (ns)	MUE (kcal/mol)	RMSE (kcal/mol)	r_P	ρ	τ
no.1[1]	1	1.04 [0.63, 1.34]	1.18 [0.76, 1.44]	0.62 [0.36, 0.85]	0.52 [−0.39, 0.69]	0.17 [−0.49, 0.50]
	5	0.72 [0.53, 1.18]	0.94 [0.66, 1.29]	0.73 [0.38, 0.94]	0.65 [0.01, 0.91]	0.39 [−0.18, 0.79]
	10	0.59 [0.39, 0.92]	0.73 [0.45, 1.03]	0.87 [0.59, 0.99]	0.80 [0.22, 1.00]	0.67 [0.17, 1.00]
no.2[2]	1	0.83 [0.54, 1.06]	0.93 [0.64, 1.14]	0.78 [0.59, 0.96]	0.67 [0.12, 0.95]	0.44 [0.05, 0.85]
	5	0.57 [0.25, 0.88]	0.74 [0.41, 1.00]	0.87 [0.74, 0.97]	0.83 [0.37, 0.99]	0.67 [0.25, 0.95]
	10	0.48 [0.28, 0.76]	0.61 [0.40, 0.84]	0.90 [0.74, 0.98]	0.87 [0.41, 1.00]	0.78 [0.35, 1.00]
no.3[3]	1	0.87 [0.48, 1.24]	1.05 [0.65, 1.36]	0.70 [0.51, 0.93]	0.57 [−0.04, 0.79]	0.33 [−0.10, 0.65]
	5	0.66 [0.42, 0.90]	0.76 [0.50, 0.99]	0.84 [0.64, 0.95]	0.75 [0.15, 0.96]	0.56 [0.08, 0.90]
	10	0.48 [0.43, 0.85]	0.54 [0.51, 0.91]	0.87 [0.65, 0.98]	0.79 [0.34, 1.00]	0.65 [0.26, 1.00]

[1] λ = 0.000, 0.091, 0.182, 0.273, 0.364, 0.455, 0.545, 0.636, 0.727, 0.818, 0.909 and 1.000.

[2] λ = 0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, 0.99078.

[3] λ = 0.0000, 0.0479, 0.1151, 0.2063, 0.3161, 0.4374, 0.5626, 0.6839, 0.7937, 0.8850, 0.9521, and 1.0000. 95% confidence intervals of MUE, RMSE, r_P, ρ, and τ are provided in square brackets, which were calculated with bootstrapping.

Table 2.

MUE, RMSE, r_P, ρ, and τ between calculated and experimental $\Delta\Delta G_{bind}$ for 8 protein systems with different numbers of λ windows and simulation lengths using ff19SB + GAFF2.2 + OPC.

# of λ windows	Time[a]	MUE (kcal/mol)	RMSE (kcal/mol)	r_P	ρ	τ
12	5	0.95 [0.77, 1.15]	1.29 [1.00, 1.58]	0.60 [0.47, 0.73]	0.70 [0.54, 0.80]	0.50 [0.38, 0.60]
	10	0.94 [0.77, 1.13]	1.24 [0.98, 1.50]	0.61 [0.47, 0.73]	0.71 [0.56, 0.81]	0.52 [0.40, 0.62]
21	5	0.95 [0.76, 1.15]	1.28 [1.04, 1.58]	0.58 [0.44, 0.71]	0.70 [0.53, 0.81]	0.51 [0.38, 0.62]
	10	0.89 [0.73, 1.09]	1.21 [0.96, 1.47]	0.62 [0.47, 0.74]	0.71[0.55, 0.82]	0.53 [0.40, 0.64]

[a]Simulation length (ns) of each λ window. 95% confidence intervals of MUE, RMSE, r_P, ρ, and τ are provided in square brackets, which were calculated with bootstrapping.[58, 59]

Table 3.

MUE, RMSE, r_P, ρ, and τ between calculated and experimental $\Delta\Delta G_{bind}$ for 8 protein systems with12 different FF combinations.

FF[a]	MUE (kcal/mol)	RMSE (kcal/mol)	r_P	ρ	τ	# pairs[b]	# pairs[c]
1	0.95 [0.77, 1.15]	1.29 [1.00, 1.58]	0.60 [0.47, 0.73]	0.70 [0.54, 0.80]	0.50 [0.38, 0.60]	57	70
2	0.91 [0.72, 1.13]	1.31 [1.00,1.60]	0.60 [0.47, 0.73]	0.71 [0.56, 0.82]	0.52 [0.40, 0.62]	56	72
3	0.96 [0.77, 1.17]	1.33 [1.02, 1.62]	0.58 [0.45, 0.73]	0.70 [0.55, 0.81]	0.51 [0.40, 0.62]	53	70
4	**0.87 [0.69, 1.07]**	**1.22 [0.94, 1.50]**	**0.64 [0.52, 0.76]**	**0.73 [0.58, 0.83]**	**0.54 [0.42, 0.64]**	**58**	**74**
5	1.00 [0.80, 1.20]	1.35 [1.07, 1.61]	0.59 [0.46, 0.72]	0.67 [0.51, 0.79]	0.49 [0.36, 0.60]	54	69
6	0.94 [0.76, 1.13]	1.28 [1.02, 1.54]	0.62 [0.49, 0.74]	0.70 [0.56, 0.81]	0.51 [0.39, 0.61]	55	70
7	0.94 [0.75, 1.17]	1.35 [0.98, 1.74]	0.58 [0.37, 0.77]	0.68 [0.49, 0.80]	0.50 [0.36, 0.61]	58	69
8	0.98 [0.77, 1.22]	1.42 [1.01, 1.85]	0.55 [0.33, 0.76]	0.65 [0.47, 0.79]	0.48 [0.35, 0.61]	57	72
9	0.93 [0.73, 1.14]	1.33 [0.98, 1.72]	0.60 [0.41, 0.78]	0.69 [0.53, 0.80]	0.50 [0.38, 0.61]	54	71
10	0.94 [0.74, 1.18]	1.39 [0.96, 1.88]	0.55 [0.29, 0.77]	0.66 [0.47, 0.79]	0.49 [0.35, 0.61]	52	70
11	1.01 [0.80, 1.25]	1.43 [1.04, 1.87]	0.58 [0.33, 0.77]	0.66 [0.48, 0.79]	0.48 [0.35, 0.60]	51	70
12	0.99 [0.76, 1.25]	1.51 [1.04, 1.92]	0.55 [0.32, 0.76]	0.66 [0.49, 0.80]	0.50 [0.37, 0.63]	56	69
13	1.05	1.46	0.56	0.61	0.46	49	66
14	0.97	1.28	0.65	0.70	0.52	50	70

[a] FF combinations from number 1 to 12 are: (1) ff19SB/GAFF2.2/OPC, (2) ff14SB/GAFF2.2/OPC, (3) ff19SB/GAFF2.2/TIP3P, (4) ff14SB/GAFF2.2/TIP3P, (5) ff19SB/GAFF2.2/TIP4PEW, (6) ff14SB/GAFF2.2/TIP4PEW, (7) ff19SB/OpenFF/OPC, (8) ff14SB/OpenFF/OPC, (9) ff19SB/OpenFF/TIP3P, (10) ff14SB/OpenFF/TIP3P, (11) ff19SB/OpenFF/TIP4PEW, and (12) ff14SB/OpenFF/TIP4PEW. 13 and 14 are the results from Song et al.[50] and Wang et al.[32], respectively.

[b] Number of pairs with $|\Delta\Delta\Delta G| < 1$ kcal/mol.

[c] Number of pairs with $|\Delta\Delta\Delta G| < 2$ kcal/mol. 95% confidence intervals of MUE, RMSE, r_P, ρ, and τ are provided in square brackets, which were calculated with bootstrapping.

Table 4.

Averaged MUE, RMSE, r_P, ρ, and τ between calculated and experimental $\Delta\Delta G_{bind}$ from all FF combinations with ff19SB or ff14SB for 8 protein systems.

FF[a]	MUE (kcal/mol)	RMSE (kcal/mol)	r_P	ρ	τ
1	0.97 [0.88, 1.05]	1.35 [1.21, 1.50]	0.59 [0.51, 0.66]	0.68 [0.62, 0.74]	0.50 [0.45, 0.54]
2	0.94 [0.85, 1.02]	1.36 [1.20, 1.51]	0.58 [0.50, 0.66]	0.69 [0.62, 0.74]	0.50 [0.46, 0.55]

[a]FF combinations for number 1 and 2 are 6 FFs with ff19SB and 6 FFs with ff14SB, respectively. 95% confidence intervals of MUE, RMSE, r_P, ρ, and τ are provided in square brackets, which were calculated with bootstrapping.

Table 5.

Averaged MUE, RMSE, r_P, ρ, and τ between calculated and experimental $\Delta\Delta G_{bind}$ from all FF combinations with GAFF2.2 or OpenFF for 8 protein systems.

FF[a]	MUE (kcal/mol)	RMSE (kcal/mol)	r_P	ρ	τ
1	0.94 [0.86, 1.02]	1.30 [1.18, 1.41]	0.60 [0.55, 0.66]	0.70 [0.65, 0.75]	0.51 [0.46, 0.55]
2	0.96 [0.88, 1.06]	1.41 [1.23, 1.58]	0.57 [0.48, 0.66]	0.67 [0.60, 0.73]	0.49 [0.44, 0.54]

[a]FF combinations for number 1 and 2 are 6 FFs with GAFF and 6 FFs with OpenFF, respectively. 95% confidence intervals of MUE, RMSE, r_P, ρ, and τ are provided in square brackets, which were calculated with bootstrapping.

Table 6.

Average MUE, RMSE, r_P, ρ, and τ for computational and experimental $\Delta\Delta G_{bind}$ from all FF combinations with OPC, TIP3P, or TIP4PEW for 8 protein systems.

FF[a]	MUE (kcal/mol)	RMSE (kcal/mol)	r_P	ρ	τ
1	0.95 [0.85, 1.05]	1.34 [1.17, 1.52]	0.58 [0.49, 0.67]	0.69 [0.61, 0.75]	0.50 [0.44, 0.56]
2	0.93 [0.83, 1.03]	1.32 [1.14, 1.52]	0.59 [0.50, 0.68]	0.69 [0.62, 0.76]	0.51 [0.45, 0.56]
3	0.99 [0.88, 1.10]	1.39 [1.21, 1.58]	0.59 [0.48, 0.67]	0.68 [0.60, 0.74]	0.49 [0.43, 0.55]

[a] FF combinations for number 1, 2 and 3 are 4 FFs with OPC, 4 FFs with TIP3P, and 4 FFs with TIP4PEW, respectively. 95% confidence intervals of MUE, RMSE, r_P, ρ, and τ are provided in square brackets, which were calculated with bootstrapping.

Table 7.

MUE, RMSE, r_P, ρ, and τ of consensus FF for 8 protein systems.

FF[a]	MUE (kcal/mol)	RMSE (kcal/mol)	r_P	ρ	τ	# pairs[b]	# pairs[c]
1	0.85 [0.67, 1.06]	1.25 [0.93, 1.57]	0.62 [0.47, 0.77]	0.72 [0.58, 0.83]	0.54 [0.42, 0.64]	59	72
2	0.86 [0.68, 1.07]	1.24 [0.93, 1.55]	0.62 [0.48, 0.76]	0.71 [0.56, 0.82]	0.52 [0.40, 0.63]	61	72
3	0.87 [0.68, 1.08]	1.28 [0.97, 1.62]	0.60 [0.45, 0.76]	0.71 [0.56, 0.83]	0.53 [0.41, 0.65]	57	70
4	0.86 [0.69, 1.08]	**1.24 [0.94, 1.57]**	**0.62 [0.48, 0.77]**	**0.74 [0.59, 0.83]**	0.54 [0.43, 0.65]	59	73
5	0.84 [0.67, 1.06]	1.23 [0.92, 1.58]	0.62 [0.46, 0.78]	0.72 [0.56, 0.83]	0.53 [0.41, 0.65]	59	71
6	0.92 [0.74, 1.13]	1.29 [1.00, 1.60]	0.62 [0.47, 0.76]	0.71 [0.56, 0.81]	0.52 [0.40, 0.62]	55	70
7	0.92 [0.73, 1.15]	**1.31 [1.00, 1.65]**	**0.61 [0.44, 0.76]**	**0.71 [0.56, 0.82]**	0.52 [0.40, 0.63]	56	71

[a] FF combinations from number 1 to 7 are: (1) 12 FFs, (2) ff19SB/OPC/consensus FF, (3) ff14SB/OPC/consensus FF, (4) ff19SB/TIP3P/consensus FF, (5) ff14SB/TIP3P/consensus FF, (6) ff19SB/TIP4PEW/consensus FF, and (7) ff14SB/TIP4PEW/consensus FF, respectively.

[b] Number of pairs with $|\Delta\Delta G| < 1$ kcal/mol.

[c] Number of pairs with $|\Delta\Delta G| < 2$ kcal/mol. 95% confidence intervals of MUE, RMSE, r_P, ρ, and τ are provided in square brackets, which were calculated with bootstrapping.