PCCP

Expanded ensemble predictions of absolute binding free energies in the SAMPL9 host–guest challenge

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-05-2023-002197.R1
Article Type:	Paper
Date Submitted by the Author:	30-Oct-2023
Complete List of Authors:	Hurley, Matthew; Temple University, Department of Chemistry Raddi, Robert; Temple University, Department of Chemistry Pattis, Jason; Temple University, Department of Chemistry Voelz, Vincent; Temple University, Department of Chemistry

PCCP

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Expanded ensemble predictions of absolute binding free energies in the SAMPL9 host–guest challenge

Matthew F. D. Hurley,[‡a] Robert M. Raddi,[‡a] Jason G. Pattis,[a] and Vincent A. Voelz[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

As part of the SAMPL9 community-wide blind host–guest challenge, we implemented an expanded ensemble workflow to predict absolute binding free energies for 13 small molecules against pillar[6]arene. Notable features of our protocol include consideration of a variety of protonation and enantiomeric states for both host and guests, optimization of alchemical intermediates, and analysis of free energy estimates and their uncertainty using large numbers of simulation replicates performed using distributed computing. Our predictions of absolute binding free energies resulted in a mean absolute error of 2.29 kcal mol^{-1} and an R^2 of 0.54. Overall, results show that expanded ensemble calculations using all-atom molecular dynamics simulations are a valuable and efficient computational tool in predicting absolute binding free energies.

1 Introduction

The Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) host–guest challenges provides a unique opportunity to benchmark the accuracy and performance of computational methods for binding free energy prediction.[1–6] Like other blind challenges,[7–9] the SAMPL host–guest challenges ensure unbiased assessment of various methods by curating experimental measurements to be released only after predictions are made.

1.1 Molecular simulation approaches in previous host–guest challenges

In the most recent SAMPL host–guest challenges, molecular simulation approaches using classical fixed-charge molecule mechanics (MM) force fields were the most widely used, although polarizable force field models, MM/PBSA, quantum mechanical (QM), and empirical machine learning approaches have seen increasing use.[6] Of the MM-based methods, alchemical free energy calculations[10,11] using double-decoupling schemes remain popular, with a variety of sampling strategies employed, ranging from Hamiltonian replica exchange (HRE),[12] non-equilibrium switching,[13,14] and expanded ensemble methods.[15,16] Other methods have included SILCS,[17] attach-pull-release (APR),[18], weighted ensemble approaches,[19] umbrella sampling with HRE,[20] and Gaussian accelerated MD (LiGaMD).[21]

The SAMPL6 host guest challenge focused on three hosts: two octa-acid, and one cucurbit[8]uril. An evaluation of the results,[4] and careful comparison of reliability and efficiency of various methods,[22] identified several problems in simulation methodologies that continue to pose a challenge in predicting absolute binding free energies by molecular simulation, including: sensitivity to simulation parameters and preparation protocols, proper estimation of prediction uncertainties, and long correlation times that may need to overcome due the rearrangements of water molecules in the binding cavity. Nevertheless, alchemical methods for absolute binding free energy calculations were found to give reasonably accurate estimates, with enhanced-sampling strategies generally leading to increased convergence.

The SAMPL7 host–guest challenge focused on the binding affinity of several small molecules to cucurbituril derivatives CB[n], CB-Clip and TrimerTrip, as well as octa-acid (OA) and exo-octa-acid.[5] The results showed that polarizable force fields like AMOEBA[23] can outperform non-polarizable force fields in these systems. SAMPL8 examined the host cucurbit[8]uril with guests that can be categorized as "drugs of abuse": methamphetamine, fentanyl, morphine, hydromorphone, ketamine, phencyclidine, and cocaine.[24] SAMPL8 also examined tetramethyl octa-acid (TEMOA) and tetraethyl octa-acid as host molecules.[25]

The latest challenge, SAMPL9, is focused on a water-soluble pillar[n]arene host called WP6 and 13 guest molecules (Figure 1). Because WP6 is highly carboxylated and expected to be anionic in solution neutral pH, and most of the guests are highly cationic salts, a careful treatment of electrostatics needs to be considered for accurate prediction of binding affinity. WP6 has many applications: it can been used as a chiral switch or for chiral sensing due to its planar chirality.[26] When WP6 is combined with a organic pyridinium salt guest, a change in pH can induce organization into nanotubes and vesicles.[27]

[a] Department of Chemistry, Temple University, Philadelphia, PA, USA. Fax: +1 215 204 1532; Tel: +1 215 204 7118; E-mail: voelz@temple.edu
‡ These authors contributed equally to this work.

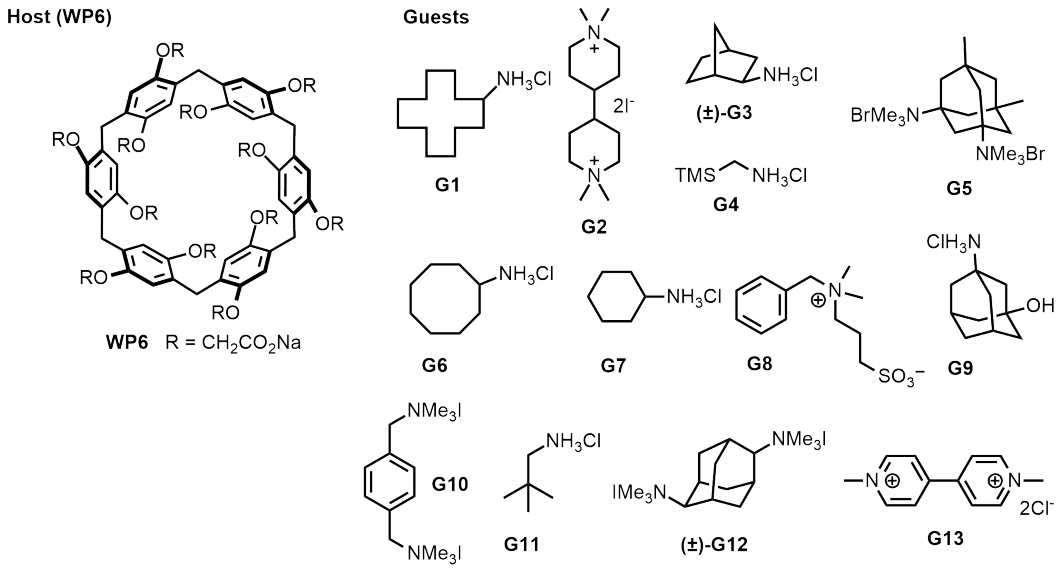

Fig. 1 Molecular structures of the pillar[6]arene host (WP6) and guests (G1–G13).

1.2 A test of expanded ensemble methods for absolute binding free energy

1.2.1 The expanded ensemble approach

Here, we use the SAMPL9 challenge to test an expanded ensemble (EE) approach to calculating absolute binding free energies. In the EE approach, double-decoupling alchemical free energy calculations are performed in which non-bonded interactions with the guest are turned off, in both host-bound and host-unbound states. Each of these calculations is performed by defining a series of N alchemical intermediates, parameterized by a coupling parameter $\lambda = 0 \rightarrow 1$, and performing Markov chain Monte Carlo (MCMC) to periodically accept or reject proposed transitions from one intermediate thermodynamic ensemble parameterized by λ_k to another parameterized by λ_l, in the presence of constant bias potentials g_i applied to each thermodynamic ensemble i. If the g_i are equal to the negative free energies $-\tilde{f}_i$ of each ensemble, then the probabilities of forward ($k \rightarrow l$) and backward ($l \rightarrow k$) transitions will be equal.[15] Therefore, the EE approach can be thought of as a procedure to adaptively *learn* the values of the biases that will lead to a uniform random walk in λ-space. (In this sense, EE is similar in spirit to sampling methods like Metadynamics,[28] which seeks to adaptively learn a bias potential along a collective variable (CV) surface to achieve uniform sampling. Indeed, recent work by Hsu et al. uses PLUMED to perform joint sampling over both CVs and alchemical λ-space.[29])

In our EE approach, we use the Wang-Landau flat-histogram method[30] to adaptively learn the (reduced) free energy surface $-\tilde{f}(\lambda)$. This method (which has long been available in GROMACS[31]) works by storing a histogram of counts h_i tracking the number of visits to each thermodynamic ensemble i. At each iteration t of the MCMC algorithm, the histogram for the current ensemble i is incremented, and the bias potential $g_i = -\tilde{f}_i$ is updated, according to

$$\begin{aligned} h_i^{(t+1)} &\leftarrow h_i^{(t)} + 1 \\ \tilde{f}_i^{(t+1)} &\leftarrow \tilde{f}_i^{(t)} - \delta \end{aligned} \tag{1}$$

where δ is positive value called the Wang-Landau (WL) increment. This has the effect of *penalizing* repeated visits to the thermodynamic state i, and making it more likely for MCMC moves to other thermodynamic states to be accepted. Once the histogram of visited states is sufficiently "flat", then the WL increment is scaled by a factor $\alpha < 1$ (in this study, $\alpha = 0.8$), the histogram counts are reset to zero, and the process continues. The histogram is deemed sufficiently flat when the ratio of all histogram counts h_i to the mean $\bar{h} = (\sum_i h_i)/N$ satisfy the criterion $\eta < h_i/\bar{h} < \eta^{-1}$, where η is called the Wang-Landau (WL) ratio (in this study, $\eta = 0.7$).

In a typical EE simulation, this process continues until the WL increment dips below a preset tolerance (in this study, $\delta < 10^{-5}$); after this point, the biases $\tilde{f}_i$ are held constant, while the simulation can continue. In practice, however, this stopping criterion may not be reached within a reasonable simulation time. In our approach, we collect samples of $\tilde{f}_i$ after some convergence criterion is reached ($\delta < 0.01$, for example) and estimate the (reduced) free energy of the alchemical transformation as the average value of $\Delta\tilde{f} = \tilde{f}_N - \tilde{f}_1$ for samples taken after this convergence criterion is reached. Since free energies are only defined up to some additive constant, throughout the EE simulation the value of $\tilde{f}_1$ is subtracted from all $\tilde{f}_i$ as an offset, making $\Delta\tilde{f} = \tilde{f}_N$.

A distinct advantage EE approaches have over alternative methods is the ability to sample all alchemical intermediates in a *single* simulation. Other methods such as λ-dynamics[32,33] also have this ability, but EE is conceptually simpler to implement (EE is essentially an MCMC wrapper *outside* of the molecular dynamics integration, whereas λ-dynamics requires a specialized integrator). Given this simplicity and self-containment in

a single simulation, EE approaches are ideal for large-scale virtual screening on distributed computing platforms such as Folding@home,[34] where simulation instances are necessarily asynchronous and only loosely uncoupled to other instances. Indeed, EE was recently deployed on Folding@home[35] to screen potential inhibitors of the SARS-CoV-2 main protease as part of the COVID Moonshot initiative.[36])

1.2.2 Progress and challenges in expanded ensemble approaches for binding free energies

Expanded ensemble (EE) methods have been used previously in SAMPL challenges, mostly through the capabilities coded into GROMACS by the Shirts group.[31] Debuting EE in the SAMPL4 challenge, Monroe et al. found that despite problems in force field parameterization for host–guest interactions, GROMACS/EE was able to produce well-converged free energy estimates.[16] While GROMACS/EE methods were not entered in the SAMPL6 blind challenge, the Shirts group tested the reliability and efficiency of GROMACS/EE on SAMPL6 targets against comparable methods and found favorable accuracy and convergence.[22]

Despite the promise of EE approaches, they remain relatively underutilized, arguably due to a few technical complications that have impeded their wider adoption. One problem is that EE methods are sensitive to the initial choice of the λ_k; poor choices of alchemical intermediates lead to poor MCMC acceptance probabilities, and extremely long convergence times. To ameliorate this problem, here we have developed a simple scheme to optimize the schedule of λ_k values, by equally spacing ensembles in thermodynamic length,[37,38] similar to the "thermodynamic trailblazing" method of Rizzi et al.[39] (see Methods).

Another issue with EE is uncertainty due to the run-to-run variation of the predictions.[22] This is a feature of any stochastic sampling algorithm, but exacerbated in EE by the saturation of the error that is known to occur with Wang-Landau flat-histogram sampling.[40] Because the histogram increment scales as a power law ($\sim \alpha^{-t}$ where t is the simulation time and α is some positive constant), the free energy estimate will converge to a fixed value, but this value may still contain error. This problem can be alleviated by scaling the histogram increment as $\sim t^{-1}$,[41] but this method is not yet implemented in GROMACS, nor has its performance and efficiency been thoroughly characterized for alchemical free energy calculations. In the meantime, we have dealt with this issue by performing many parallel replicates of a given alchemical transformation using distributed computing, and compute estimates as averages of individual trials, as further described below. Using this improved strategy, Zhang et al. recently showed that EE methods can predict relative binding free energies of Tyk2 inhibitors to within a mean unsigned error (MUE) of 0.75 $\pm$ 0.12 kcal mol^{-1}.[42] The current SAMPL9 host–guest challenge represents the first time these improved protocols have been tested for predicting absolute binding free energies.

2 Methods

Absolute binding free energies for host–guest interactions were calculated using a double-decoupling method in which the alchemical free energies of decoupling the guest in the presence and absence of the host were computed using expanded-ensemble molecular simulations performed on the Folding@home distributed computing platform[34,35]. A three-part workflow was implemented to (1) prepare systems, (2) perform expanded ensemble simulations on Folding@home and Temple University high-performance computing (HPC) clusters, and (3) analyze the results.

2.1 System preparation

2.1.1 Microstate enumeration

To estimate the ionization state the WP6 host at pH 7.4, we considered the fluorescence emission spectra vs. pH published in Yu et al.[27] A titration curve fit to this data suggest a pKa of 6.997 and a Hill coefficient of 3.519, for a model where approximately 4 protons cooperatively dissociate upon varying the pH from 2 to 11. Based on this result, and the absence of other information, we assumed that the most populated microstate of WP6 at pH 7.4 has a -12 net charge, and that titration to lower pH cooperatively adds 4 protons to form a -8 net charge state. Therefore, we considered three different protonation states of the WP6 host: -12, -10, -8 net charge, each with equal numbers of deprotonated groups above and below the pillarene ring (Supporting Figure S1), as the host microstates likely contributing most in the binding reaction.

Reference ionization states for each guest molecule were determined by OpenEye's Quacpac module[43], which selected the most energetically favorable ionization state at pH 7.4. While the reference state is likely to have the greatest population, we additionally considered a larger ensemble of enumerated microstates that may be populated near pH 7.4. This resulted in between 1 and 4 microstates per guest molecule. We also considered each enantiomer of chiral guest molecules as separate microstates (Supporting Figure S2).

2.1.2 Simulation preparation

System preparation was performed semi-automatically using a series of in-house Python scripts. Force field parameters for nearly all hosts and guests used OpenFF-2.0.0.[44,45] The only exception to this was for the guest G4. This molecule contained a silane group for which OpenFF parameters were unavailable. We instead used GAFF-2.11[46] for G4. Partial charges for all molecules were assigned using AM1-BCC[47].

Initial poses for receptor-ligand systems were prepared by docking guests to the host via OpenEye's OEDocking[48] module using the FRED score function[49] and saving the minimum-energy structure. Systems were solvated with TIP3P water and neutralizing counterions at 137 mM NaCl. Ligand-only simulations used a 3.5 nm cubic box, while receptor-ligand simulations used a 4.5 nm cubic box. Ligand-only simulations were minimized and equilibrated at 298.15 K using GPU-accelerated OpenMM version 7.5.0[50]; subsequent production runs were performed in GROMACS (see below). Receptor-ligand were minimized and equilibrated at 298.15 K in GROMACS version 2020.3[31] using position restraints with a force constant of 800 kJ mol^{-1} nm^{-2} all heavy atoms of the host, and all atoms of the guest. Equilibration was performed in the isobaric-isothermal ensemble.

2.2 Expanded Ensemble simulation

Absolute binding free energies computed using expanded-ensemble (EE) methods have been used previously in SAMPL challenges,[16,22] and our methods closely follow these efforts, with some innovations inspired by recent work[42].

The free energy ΔG_L of decoupling the guest from solvent in a ligand-only (L) simulation was calculated using 101 alchemical intermediates in which Coulomb (coul) interactions are turned off, and then van der Waals (vdW) interactions. The free energy ΔG_{RL} of decoupling the guest from a receptor-ligand (RL) simulation was calculated using 101 alchemical intermediates in which a restraint potential is turned *on*, then Coulomb interactions are turned off, and then vdW interactions. The restraint potential was a harmonic distance restraint between the center of mass of the six benzene rings of the WP6 host (6 rings $\times$ 6 carbons = 36 atoms), and all non-hydrogen guest atoms, with a force constant of 800 kJ mol^{-1} nm^{-2}, and an equilibrium distance of 0 nm. The absolute free energy of binding ΔG is estimated as

$$\Delta G = \Delta G_{\text{rest}} + \Delta G_L - \Delta G_{RL}, \tag{2}$$

where ΔG_{rest} is the free energy cost of restraining the guest from standard volume to a restricted volume, determined by the force constant 800 kJ mol^{-1} nm^{-2}, which we compute to be $\Delta G_{\text{rest}} = +6.42\,RT$. The $-\Delta G_{RL}$ term includes the free energy of removing this restraint.

2.2.0.1 Optimization of alchemical intermediates To avoid sampling bottlenecks in the EE algorithm that would impede the efficient exploration of all alchemical intermediates, we implemented a custom optimization algorithm called `pylambdaopt` (Zhang et al., in preparation). This algorithm works in two steps:

First, a trial EE simulation is performed with initial guesses for the Coulomb and vdW λ_i values. From this trial simulation, estimates of the thermodynamic length $|\ell(\lambda_{k+1}) - \ell(\lambda_k)|$ between each pair of intermediates are made.[37,38] The thermodynamic length is estimated as the variance in the distributions $P(\Delta u_{k,k+1})$, where $\Delta u_{k,k+1} = u_{k+1} - u_k$ is the change in (reduced) energy incurred by bringing a sample from thermodynamic ensemble k to thermodynamic ensemble l.[39]

Second, cubic spline fitting is used to find a continuous and differentiable function $\ell(\lambda)$ that interpolates the $\ell(\lambda_i)$. Steepest-descent minimization is then used to find new values λ_i^* that minimize the loss function $\mathscr{L} = \sum_k |\ell(\lambda_{k+1}) - \ell(\lambda_k)|^2$. This results in a series of λ_i^* values that are equidistant from each other in thermodynamic length. This procedure equalizes the EE acceptance probabilities, ameliorating MCMC sampling bottlenecks with poor transitions. The optimized λ_i^* values are then used for production runs. An example of lambda values before and after optimization is shown in Figure 2.

For this study, simulations of each guest-only (L) and host-guest (RL) system (using the -12 charge state of the host) were run for 24 hours in order to sample Δu_{kl} distributions over the course of an EE simulation, using an initial guess for the schedule of λ-values that control the alchemical transformation. From this information, optimized λ-values were obtained and used for production-run EE simulations on the Folding@home distributed computing platform[34,35].

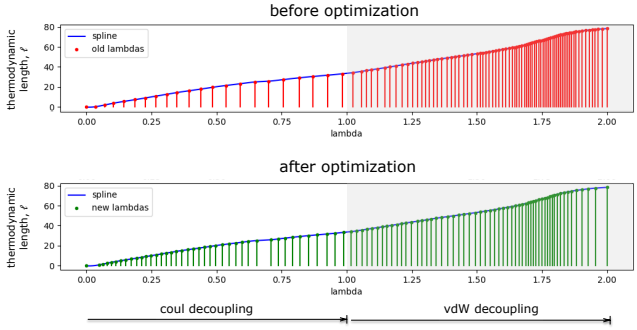

Fig. 2 An example of λ-value optimization using pylambdaopt. The top panel shows λ_i values before optimization. The bottom panel shows the optimized λ_i^* values. Since the vdW decoupling transformation occurs independently and subsequent to the Coulomb decoupling transformation, `pylambdaopt` treats the transformation using a single $\lambda = 0 \to 2$, where $\lambda = 0 \to 1$ represents Coulomb decoupling and $\lambda = 1 \to 2$ represents vdW decoupling.

For each set of host–guest microstate pairs, fifty parallel production-run EE simulations were performed in GROMACS 2020.3. Simulations used a timestep of 2 fs, 0.9 nm cutoffs for long-range electrostatics, LINCS constraints on H-bonds,[51] with frames were saved every 50 ps. The Wang-Landau method and Metropolized-Gibbs move set was used for EE simulations. The initial Wang-Landau (WL) bias increment was set to 10 k_BT, and was scaled by a factor of 0.8 every time the histogram of sampled intermediates was sufficiently flat.

2.3 Analysis of free energies and uncertainties

The convergence of the EE predictions was monitored by the progressive decrease of the Wang-Landau (WL) increment. We considered the EE simulations to be sufficiently converged if the WL increment went below 0.01 and 0.02 for the L and RL simulations, respectively. Free energies were computed as the average of all free energy estimates reported after the convergence threshold was reached, across all converged trajectories. In the case that less than five trajectories reached convergence according to our criteria, the five (or more) trajectories with the smallest WL increments were used to compute the average free energy. These instances were few, and only included affected calculations for guests G5 and G8.

Uncertainties in our computed binding free energies ΔG (in units of RT) come from the standard deviations from the sample mean of computed ΔG_L and ΔG_{RL} values across multiple parallel simulations.

2.3.1 Binding free energy predictions consider the full set of host and guest microstates.

Our final ranked predictions of the absolute binding free energy ΔG for each host–guest interaction (in units RT) are computed as

$$\Delta G = -\ln \frac{\sum_{i \in \text{bound}} e^{-\Delta G_i}}{\sum_{i \in \text{unbound}} e^{-\Delta G_i}}, \tag{3}$$

where each ΔG_i are either host-bound or host-unbound guest microstate free energies. Free energy differences relating bound and unbound microstates are provided by the double decoupling EE free energy simulations.[52,53] Free energy differences relating protonation states of the WP6 host were given by our model of cooperative titration of 4 protons at pH 6.997. At pH 7.4, this model rewards the removal of two protons by -1.856 RT. Free energy differences between the protonation microstates of the guests are provided by microstate pKa estimates obtained using the `luoszgroup` pK_a predictor from Qi Yang et al.[54].

Model uncertainties σ_{model} were calculated as

$$\sigma_{\text{model}} = \left(\sigma_{\Delta G}^2 + \sigma_{\text{sys}}^2\right)^{1/2} \quad (4)$$

where $\sigma_{\text{sys}} = 0.6857\ RT$ is assumed to be independent systematic error arising from the reported 1.7 kJ mol^{-1} accuracy of OpenFF 2.0.0[44].

2.4 Overview of our SAMPL9 submissions

In addition to our ranked SAMPL9 submission ("Voelz rnkd"), we also submitted two unranked SAMPL9 submissions of binding free energies (in units kcal mol^{-1}). One included all samples of the free energy estimates throughout the simulations, regardless of the WL convergence ("Voelz all"). The other used only the -8 net charge microstate of the host ("Voelz RL8"). Results for these submissions can be found in Table 1.

One minor complication was that we were unable to fix errors in our G8 simulations with WP6 in charge states of -8 and -10 before submitted results were due. Therefore, the $\Delta G_{\text{binding}}$ for G8 is simply our $\Delta G_{\text{binding}}$ prediction for G8 with WP6 in the -12 charge state.

An interactive webpage of all EE simulations (WL increment over time and estimated free energy over time, for all alchemical transformations) and our computed binding free estimates are available at `https://vvoelz.github.io/sampl9-voelzlab/`. Below we discuss the results for our ranked submission.

3 Results and Discussion

3.1 Performance of expanded ensemble approach for absolute binding free energy prediction

Here, we present absolute binding free energies for each microstate, describe the overall performance statistics for our submissions, and review some of the interesting cases found during our analysis. Furthermore, we will discuss a pitfall in our restraint protocol and how corrections were made to counter this mistake.

Absolute binding free energy predictions were calculated for the full set of host and guest microstates (Figure 3), as described in Methods. The various charge states of the host and guest are given in Supporting Table 1). For the majority of guest molecules, binding free energies calculated for simulations of the host in a charge state of -8 give lower values of $\Delta G_{\text{binding}}$ than hosts in -10 and -12 charge states. Typically, absolute binding free energies across microstates are dominated by the -8 host state, with the exception of G4. Binding free energy predictions for G4 across microstates span ~5 kcal mol^{-1} and yield relatively large uncertainty (~ 1 kcal mol^{-1}). As mentioned in Methods, G4 was parameterized using GAFF-2.11, which may be an influencing factor.

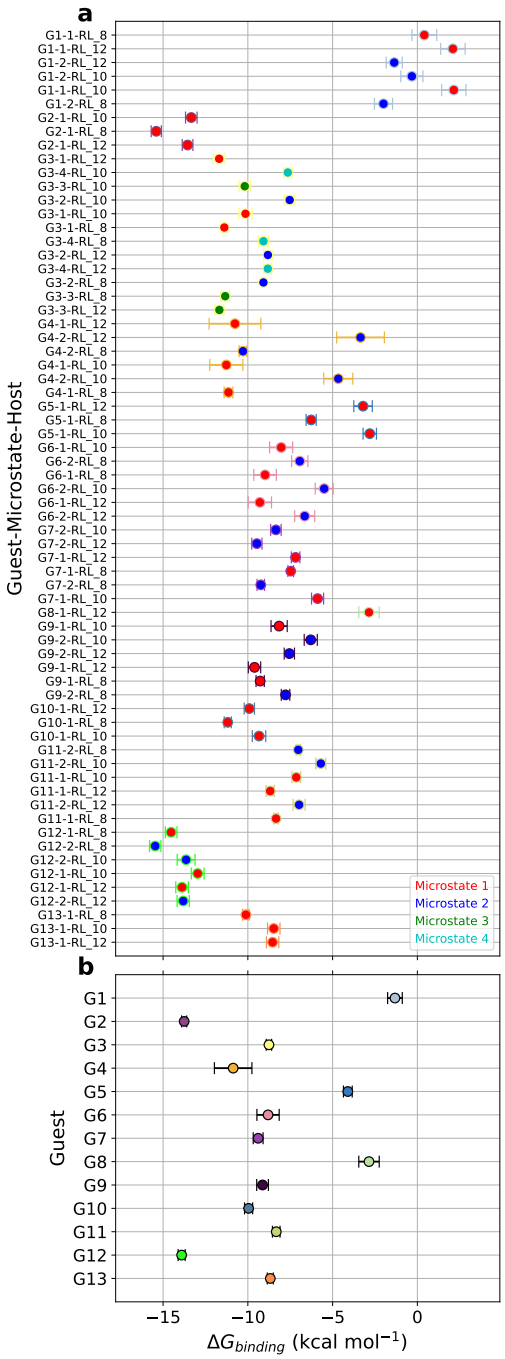

Fig. 3 (a) Predicted free energies of host–guest binding using single microstates i ($\Delta G_{\text{binding}} = \Delta G_{\text{rest}} + \Delta G_{L,i} - \Delta G_{RL,i}$) between host microstates and guest microstates. Error bars are colored by guest (see panel b) and data markers are colored by microstate index (see legend). (b) $\Delta G_{\text{binding}}$ in kcal mol^{-1} for each guest to WP6 (colored by guest) calculated according to Equation (3)

.

Overall performance of our ranked submission gave an overall mean absolute error (MAE) of 2.29 kcal mol^{-1}, root mean

Table 1 Absolute binding free energy predictions and uncertainties for our three submissions.

Group	Exp	Exp std	Voelz rnkd	Voelz rnkd std	Voelz RL8	Voelz RL8 std	Voelz all	Voelz all std
WP6-G1	-6.44	0.01	-1.32	0.43	-2.00	0.53	-1.63	0.53
WP6-G2	-10.44	0.05	-13.76	0.15	-15.40	0.21	-13.74	0.21
WP6-G3	-7.92	0.02	-8.76	0.16	-9.08	0.16	-8.75	0.16
WP6-G4	-6.41	0.01	-10.87	1.11	-11.14	0.26	-10.78	1.10
WP6-G5	-5.39	0.02	-4.10	0.26	-6.27	0.30	-4.00	0.32
WP6-G6	-7.98	0.04	-8.81	0.66	-8.59	0.65	-9.20	0.61
WP6-G7	-6.98	0.02	-9.39	0.29	-9.24	0.22	-9.38	0.33
WP6-G8	-5.96	0.01	-2.85	0.60	-2.85	0.60	-2.25	0.63
WP6-G9	-6.24	0.05	-9.13	0.34	-8.91	0.23	-9.14	0.36
WP6-G10	-9.82	0.03	-9.96	0.24	-11.19	0.21	-9.88	0.29
WP6-G11	-6.17	0.01	-8.32	0.22	-8.08	0.13	-8.32	0.27
WP6-G12	-10.87	0.02	-13.90	0.21	-15.16	0.28	-13.91	0.22
WP6-G13	-8.47	0.04	-8.68	0.18	-10.11	0.14	-8.68	0.18

RL8 submission uses our RL12 prediction for G8 due to the issue mentioned in the methods section.

squared error (RMSE) of 2.74, and an R^2 of 0.54 (Figure 4), which falls just under the median of all SAMPL9 submissions (Table 2). As we discuss below, our "Voelz all" predictions were able to best rank the binding free energies, and our "Voelz RL8" predictions had the most correlation with experimental observations across all SAMPL9 submissions.

Table 2 Summary of participant performance in free energy predictions over all host–guest systems. Statistics include the correlation coefficient (R^2), mean absolute error (MAE), mean standard error (MSE), root-mean squared error (RMSE). Ponder and Voelz submissions had the highest R^2 values, while the U-Barcelona submission had the lowest absolute error.

	R^2	MAE	MSE	RMSE
group				
Voelz rnkd	0.535	2.292	7.527	2.743
Voelz RL8	0.618	2.618	9.049	3.008
Voelz all	0.541	2.338	7.619	2.760
Ponder	0.582	1.930	7.165	2.677
U-Pittburgh	0.396	1.933	6.176	2.485
U-Barcelona	0.141	1.600	4.159	2.039
Procacci-DSSB	0.162	3.752	19.632	4.431
Procacci-VINARDO	0.003	2.007	8.461	2.909

3.2 Reliability and convergence of EE predictions

In previous SAMPL host–guest challenges, it was noted that seemingly small differences in free energy protocols could non-trivially affect predictions.[22] One sign that our expanded ensemble protocol is relatively robust in this aspect is the consistency of predictions across our ranked and unranked submissions (Table 1).

Our approach was able to predict the affinity some guests better than others (Table 1 and Figure 4) Our most accurate predictions were made for G10, G13, G6, G3, G5; moderately accurate predictions were made for G11, G7 and G9, and poor predictions (greater than 3.0 kcal mol^{-1} error) for G12, G8, G2, G4, and G1.

Our predictions varied the most from other groups' for guests G1 and G13. Our expanded-ensemble approach was able to make very accurate predictions for G13, where most groups were unable to do so (Figure 5). Traces of expanded ensemble free energy estimates ΔG_L for guest-only decoupling display excellent convergence (Figure 6). Traces of free energy estimates ΔG_{RL} for host–guest decoupling show more variance across parallel expanded ensemble simulations, but with a robust sample mean (Figure 7).

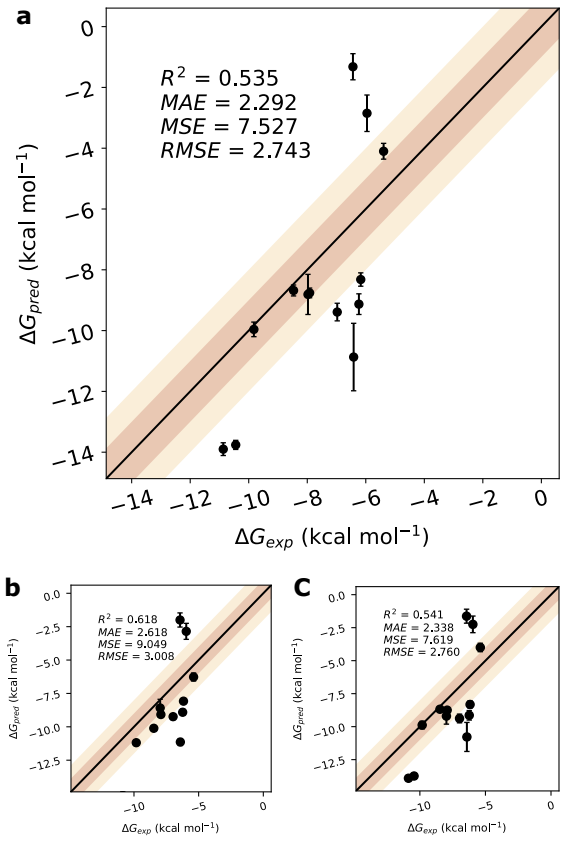

Fig. 4 (a) Comparisons of predicted (ranked submission) vs. experimental binding free energies for all host–guest systems. Annotations report the correlation coefficient (R^2), mean absolute error (MAE), mean standard error (MSE), root-mean squared error (RMSE). (b) Comparisons for the "Voelz RL8" submission. (c) Comparisons for the "Voelz all" submission.

Our most inaccurate predictions were made for G1, with more error than all other groups. In our ranked submission, we incorrectly reported the ΔG of binding for G1 as -0.78 ± 0.43 kcal mol^{-1}, due to a error converting from RT to kcal mol^{-1} (this er-

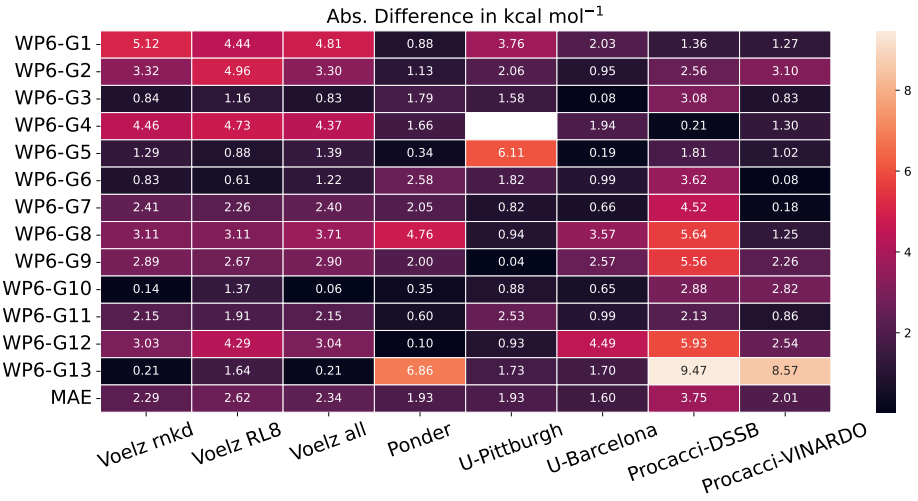

Fig. 5 Predictions of host–guest binding free energies submitted by all SAMPL9 participants. The color map and numerical values inside the cells are the absolute difference in predictions against experiment (in kcal mol^{-1}). The last row is the mean absolute error (MAE) over all host–guest predictions for each group.

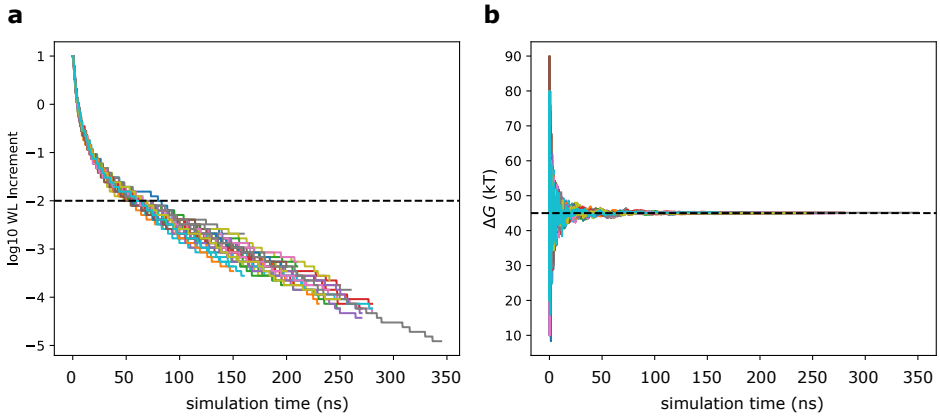

Fig. 6 Convergence of expanded ensemble (EE) estimates of guest-only decoupling free energies ΔG_L for G13. Superimposed are (up to) fifty independent trajectories (distinguished by color). (a) The Wang-Landau (WL) increment vs. simulation time, with a dotted line denoting our 0.02 convergence threshold. (b) EE estimates of ΔG_L vs. simulation time, with a dotted line denoting the final estimate.

ror was only made for G1). The correct value is -1.32 kcal mol^{-1}, which we will continue to use in the analysis reported here.

G4 was our next-most inaccurate prediction. One source of error might be inaccurate force field parameters for the trimethylsilyl group (we were forced to parameterize this molecule using GAFF due to the absence of silane parameters in OpenFF). Another source of error for G4 may arise from poor sampling. We inspected the simulation trajectories for this guest and found slow binding events of sodium cations to the WP6 binding pocket while nonbonded interactions for guest were decoupled, which hindered re-coupling of the guest due to steric clashes with ions inside the host. Because of the long timescale needed for coupling and decoupling, cycling between the two endpoints impeded, ultimately hindering convergence and causing additional variance in the predicted free energies.

While binding free energies for G6 were accurately predicted, the convergence behavior of the expanded ensemble method for this guest highlights the importance of adequate conformational sampling. Inspection of simulated trajectories reveals slow transitions between two metastable states for G6, corresponding to boat and chair conformations. The timescale of these transitions are slow enough that the expanded ensemble approach exhibits hysteresis as it tries to learn the free energy profile for one conformation, then the other. Traces of ΔG over the course of the expanded ensemble trajectory show large variability ($\pm$ 10 RT), giving rise to a large uncertainty in ΔG. This behavior can be seen most clearly in the ligand-only (L) trajectories (Figure 8) but can also be seen in the receptor-ligand (RL) trajectories (Figure 9). This finding presents a strong argument for the use of multiple independent trajectories in estimate binding free energies, since the average between these states over all trajectories allows us to predict the converged free energy very early into sampling.

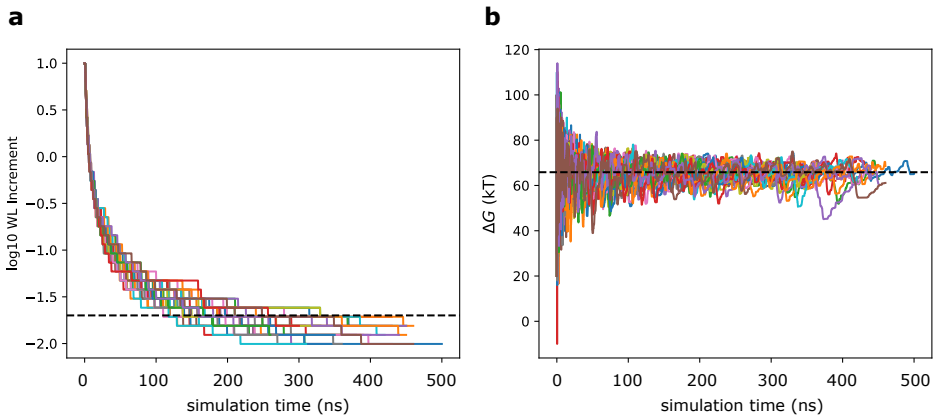

Fig. 7 Convergence of expanded ensemble (EE) estimates of host–guest decoupling free energies ΔG_{RL} for G13. Superimposed are (up to) fifty independent trajectories (distinguished by color). (a) The Wang-Landau (WL) increment vs. simulation time, with a dotted line denoting our 0.02 convergence threshold. (b) EE estimates of ΔG_{RL} vs. simulation time, with a dotted line denoting the final estimate.

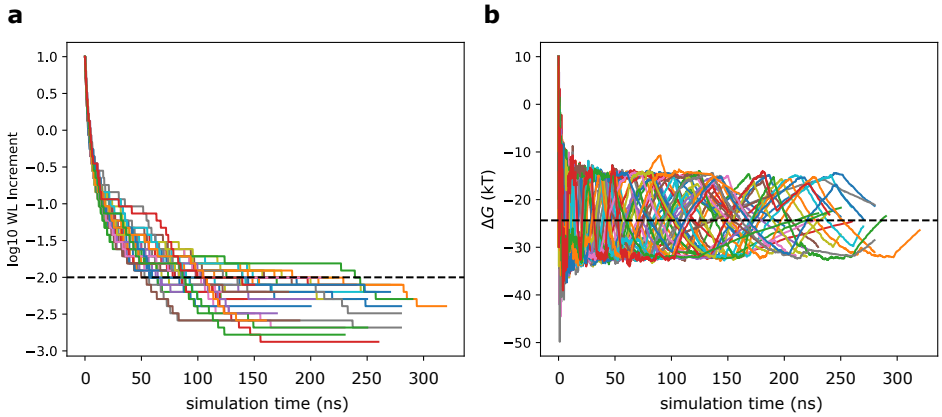

Fig. 8 Convergence of expanded ensemble (EE) estimates of guest-only decoupling free energies ΔG_{L} for G6. Superimposed are (up to) fifty independent trajectories (distinguished by color). (a) The Wang-Landau (WL) increment vs. simulation time, with a dotted line denoting our 0.02 convergence threshold. (b) EE estimates of ΔG_{L} vs. simulation time, with a dotted line denoting the final estimate.

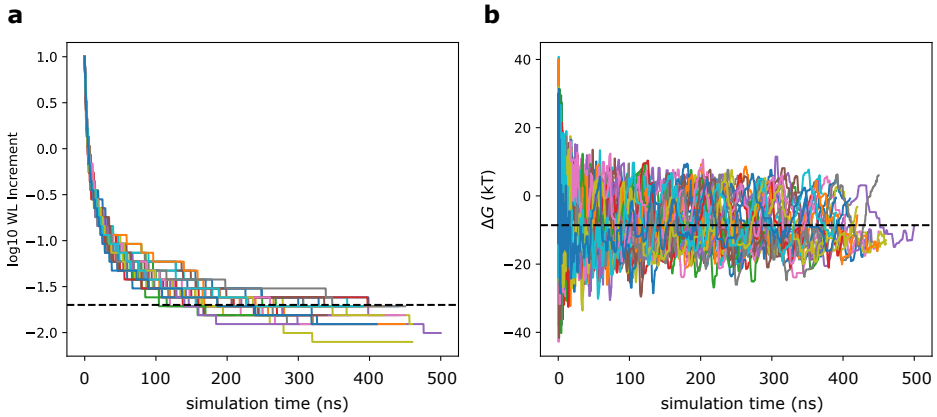

Fig. 9 Convergence of expanded ensemble (EE) estimates of host–guest decoupling free energies ΔG_{RL} for G6. Superimposed are (up to) fifty independent trajectories (distinguished by color). (a) The Wang-Landau (WL) increment vs. simulation time, with a dotted line denoting our 0.02 convergence threshold. (b) EE estimates of ΔG_{RL} vs. simulation time, with a dotted line denoting the final estimate.

3.3 Correcting for the free energy bias of restrained ligands

In preparing this manuscript, we realized there was an error in the way harmonic restraints were implemented during the expanded ensemble simulations. Our computed value of ΔG_{RL} was supposed to include the free energy of adding a harmonic restraint, but the simulations were performed with the restraint always on. This means that our submitted estimates contain a systematic *positive* bias, as they do not include the favorable reward of turning off the restraint from $-\Delta G_{RL}$. This reward should be small for well-chosen restraint potentials, but large in situations where the restraint doesn't match the equilibrium pose(s) of the guest.

To correct for this bias, we performed two additional simulations for each ligand in which we altered the restraint potential to 400 kJ mol^{-1} nm^{-2}, and then 0 kJ mol^{-1} nm^{-2}. These, combined with our original simulations with 800 kJ mol^{-1} nm^{-2} restraints, allowed us to use the Multi-state Bennett Acceptance Ratio (MBAR)[55] to estimate the (negative) free energy reward of removing the restraint. The corrected ΔG values (Figure 10) show a shift in our estimates towards higher binding affinities that are relatively minor (between -0.45 and -1.80 RT, i.e. between -0.26 and -1.1 kcal mol^{-1}), with two exceptions: G1 and G5.

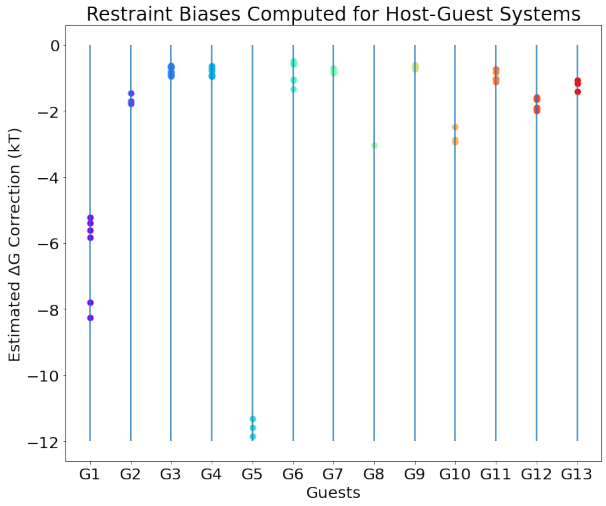

Fig. 10 Estimated correction terms to our submitted free energy predictions for each guest, to correctly account for restraint biases.

G1 and G5 show restraint bias corrections of 3.76 and 6.86 kcal mol^{-1} respectively. In analyzing the updated free energy prediction of G1, we observe that our predicted free energy estimate is improved considerably, resulting in an absolute error of only 1.36 RT, or 0.8 kcal mol^{-1}. In contrast, the large restraint bias correction for G5 increases the absolute error in our prediction of binding free energy from 0.76 to 3.3 kcal mol^{-1}. Further analysis of the trajectory data for G5 reveals a bimodal distribution of displacements of the guest from the host center-of-mass, and time courses showing slow transitions between guest poses near the top and bottom rims of the host (Figure 11). This suggests that the harmonic potential used, which restrains the guest at the center of the host, is a poor choice for this guest. Sampling issues thus may be a reason for the large prediction error for this guest.

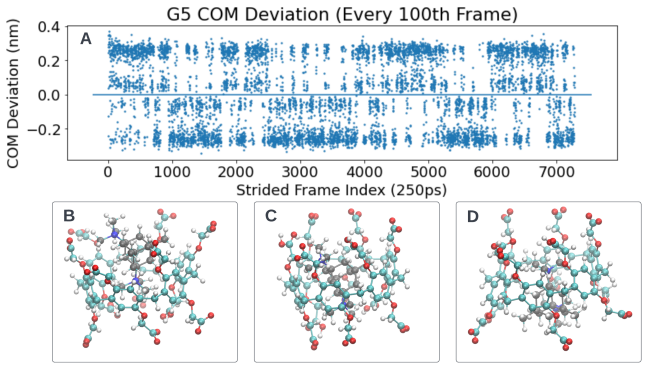

Fig. 11 Transitions between a number of host-bound states are shown for G5. (a) A trace of the z-axis displacement of the guest center-of-mass (COM) with respect to the host COM shows free energy minima at both +/- edges of the host, in addition to those near the center of mass. (b,c, and d) depict Conformations representative of binding for G5 are shown for poses near the (b) top, (c) middle, and (d) bottom of the host.

Overall, when our predictions of host–guest binding free energies are updated with the restraint bias corrections, agreement with experiment worsens by all metrics (R^2 coefficient, RMSE and MAE, Figure 12). The uncorrected predictions tend to underestimate ΔG (i.e. predict tighter binding), and the correction exacerbates this trend.

In our study, we did not attempt Boresch-style restraints.[56] While this restraint scheme gives an exact expression for the restraint free energy, we deemed it very difficult to choose the anchor points *a priori*, especially given the our imprecise knowledge of the binding pose (which is likely to be highly dynamic). For an excellent recent comparison of restraint schemes for absolute binding free energies, we refer the reader to Clark et al.[57]

In SAMPL6, which also featured highly carboxylated hosts—octa-acid (OA) and tetramethyl octa-acid (TEMOA)—participants using GAFF/AM1-BCC with TIP3P consistently underestimated ΔG (i.e. predicted tighter binding),[4] similar to our results. These results point to the general need for better treatment of electrostatics. In this particular, perhaps consideration of other host protonation states (-9, -7, -6, etc.) may have resulted in improved predictions. Despite having the least amount of net charge, host protonation states of -8 tended to predict the greatest decoupling free energies, suggesting subtle preferences of guests for ionic host sidechains and their arrangements with counterions may be important.

3.4 Comparison of the convergence of EE biases versus MBAR free energy estimates

In their initial work exploring the performance of EE in the SAMPL4 host–guest challenge, Monroe et al.[16] waited until the biases were no longer updated, and then used the subsequent "production-run" simulation to collect samples at each lambda value as input for the Multistate Bennett Acceptance Ratio (MBAR) free energy estimator.[55] This method assumes that EE biases have converged to the true free energies. Only then

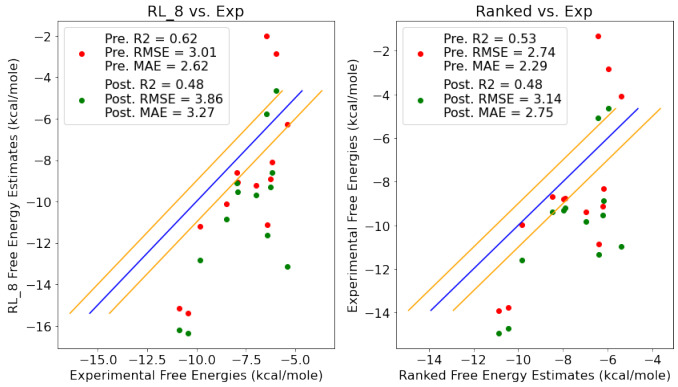

Fig. 12 Changes to our submitted predicted before (red) and after (green) corrections are applied to correctly account for restraint biases. Left: "Voelz RL_8" submissions vs. experiment. Right: "Voelz ranked" submissions vs. experiment.

will the MCMC acceptance criteria produce samples that are correctly drawn from their equilibrium distribution at each lambda value.

In contrast, the protocol used in this study is to wait until the Wang-Landau flat-histogram increment reaches a critically small threshold, and then collect values of the still-updating biases, which vary about some sample mean. An interesting question to consider: is it possible to use these *nearly*-equilibrated samples as input to the MBAR estimator? Technically, the samples are out of equilibrium, but perhaps the MBAR estimator might have some unforeseen advantages. MBAR estimates free energies using computed values of Δu_{kl} for every sample in the simulation (the change in the reduced energy for a configuration sampled in ensemble k brought to ensemble l) for all ensembles l. The multi-ensemble sampling approach of MBAR tends to give smaller estimated uncertainties than other methods,[58] which is highly desirable considering the large fluctuations in our free energy estimates that sometimes persist beyond 100 ns (e.g. see Figs. 7 and 8).

To compare the performance of EE vs. MBAR estimates made from expanded ensemble sampling, we use as an example the G13 host–guest decoupling EE simulations (see Fig. 7). Note that in this case, MBAR analysis is possible because we used simulation settings that periodically saved all values of Δu_{kl} to the GROMACS `dhdl.xvg` file. The EE approach does not strictly require this, but it is often very useful despite the computational expense.

Just as with our EE protocol, estimates are made for each trajectory, using only samples taken after the WL increment reaches a value below 0.02 (our criterion for the RL simulations). Once this threshold is reached, some additional time is needed before the EE simulations sample all every lambda value. Once this is achieved, the sampled values of Δu_{kl} are sorted by ensemble k and used as input to the MBAR estimator. In this way, an MBAR estimate of the free energy $(\Delta G)_m$ and its uncertainty $(\sigma_{\Delta G})_m$ is obtained for each independent trajectory m.

To compare EE and MBAR estimates, we compute as a function of simulation time: the free energy estimate ΔG, its standard deviation across trajectories $\sigma_{\Delta G}$, and the standard error of the mean (SEM). If there are M viable trajectories, the free energy estimate is calculated as

$$\Delta G = \frac{1}{M}\sum_{m=1}^{M}(\Delta G)_m,$$

For EE, the standard deviation across trajectories is calculated as the standard deviation from the sample mean:

$$\sigma_{\Delta G} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\Delta G_m - \Delta G)^2}.$$

For MBAR, we use the uncertainty of each MBAR estimate $(\sigma_{\Delta G})_m$ to calculate the standard deviation across trajectories:

$$\sigma_{\Delta G} = \sqrt{\sum_{m=1}^{M}(\sigma_{\Delta G})_m^2}.$$

For both EE and MBAR, the standard error of the mean (SEM) is calculated as

$$\mathrm{SEM} = \frac{\sigma_{\Delta G}}{\sqrt{M}}.$$

Note that for both EE and MBAR estimates, we do not correct for time-correlated samples by subsampling the input data, as is often recommended to avoid artificially low uncertainty estimates. Based on our experience, correlation times would likely correspond to slow conformational rearrangements that can occur around the 100-ns timescale, resulting in few samples after subsampling. Instead, we are using the variation across independent simulation trajectories to provide an estimate of this uncertainty.

A comparison of the EE and MBAR estimates as a function of simulation time are shown for 10 independent trajectories in Figure 13. Interestingly, while EE estimates for each trajectory continue to stochastically fluctuate past 200 ns (Figure 13a), MBAR estimates for each trajectory are smooth and robust once sufficient input data is achieved (Figure 13b). Run-to-run variation of MBAR estimates, however, still vary considerably across trajectories. This suggests that regardless of the estimator used to calculate ΔG_{RL}, each trajectory has not reached a global convergence, likely due to slow conformational motions that limit sampling.

Comparisons of the average ΔG_{RL} across trajectories computed from the EE biases (Figure 13c) versus the MBAR estimator (Figure 13d) show no clear advantage of using the MBAR estimator over the EE biases for free energy estimation. The standard deviation across trajectories $\sigma_{\Delta G}$, and the standard errors in the mean (SEM) are highly similar over the length of the trajectories (Figures 13e–f).

Our interpretation of these results is that, despite the favorable properties of the MBAR estimator, the out-of-equilibrium sampling present in the EE simulations violate the sampling conditions of the estimator, limiting its performance. A related question we do not address here is how a non-equilibrium work (NEW) estimator might perform given input data from from EE trajectories, that is a subject for future work.

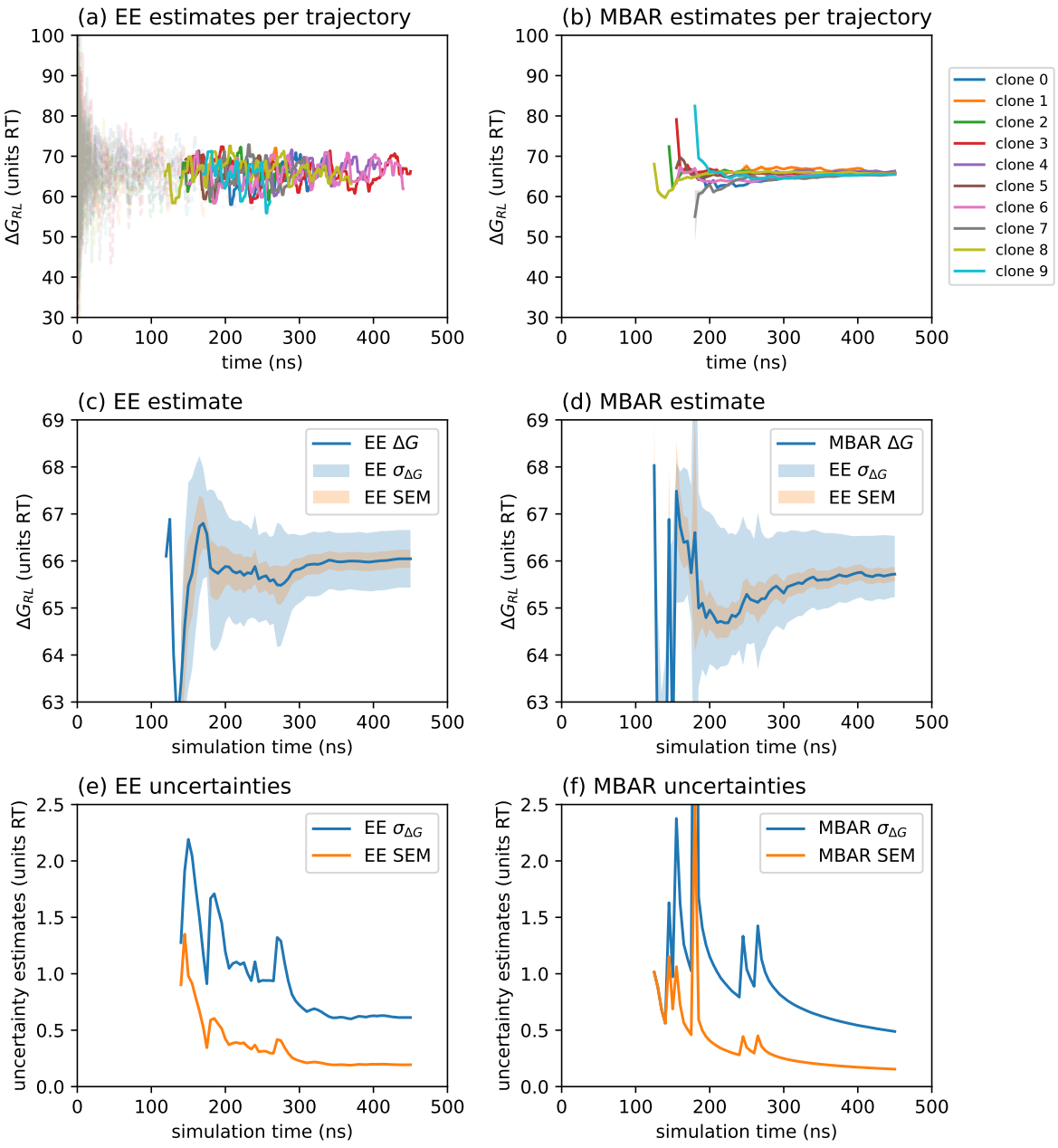

Fig. 13 Comparison of the convergence of free energies from expanded ensemble simulations, calculated using the EE biases versus the MBAR free energy estimator. Ten independent EE trajectories for the decoupling of guest G13 from the WP6 (-12) host (see Figure 7) were used for these tests. (a) Free energies ΔG_{RL} over time for each trajectory estimated from the EE biases. Traces change from transparent to solid lines as each trajectory reached the $\delta < 0.02$ convergence criterion. (b) Free energies ΔG_{RL} over time for each trajectory estimated from MBAR. The appearance of each trace over time occurs when both the convergence criterion is met, and there are samples from each thermodynamic ensemble. Mean free energy estimates averaged over each trajectory, standard deviations $\sigma_{\Delta G}$ across trajectories, and standard errors of the mean, shown over time for the EE bias estimator (c,e) and MBAR estimator (d,f).

Table 3 Spearman rank correlation coefficients of SAMPL9 host–guest participant rankings.

Group	r_s	p-value
Voelz rnkd	0.385	0.094
Voelz RL8	-0.335	0.869
Voelz all	0.451	0.059
Ponder	0.418	0.076
U-Pittburgh	-0.176	0.717
U-Barcelona	0.236	0.212
Procacci-DSSB	0.247	0.202
Procacci-VINARDO	-0.407	0.915

3.5 Comparison of rank ordering with other challenge participants

Virtual screening in drug discovery often relies on the ability of computational models to correctly rank order predicted binding affinities of ligands. To evaluate the extent to which submitted predictions correctly ranked the binding affinity of guests G1 through G13 compared to experiment, we used the Spearman rank correlation coefficient,[59]

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}, \tag{5}$$

where d_i are differences in (integer) ranks for each guest, and $n = 13$ is the number of ranked items (Figure 14 and Table 3). According to this metric, our "Voelz all" submissions give the most correctly ranked predictions compared to experiment, with a r_s value of 0.45. To gauge the statistical significance of this result, we computed one-sided p-values by using 100,000 random rank perturbations to non-parametrically construct the null distribution of r_s. With a p-value of 0.059, the "Voelz all" ranking is not quite significant enough to reject the null hypothesis that the measured value of r_s is due to random chance.

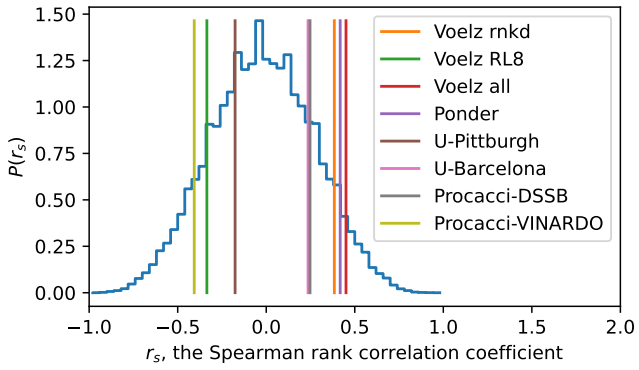

Fig. 14 Spearman rank correlation coefficients r_s for participants in the SAMPL9 host–guest challenge, in comparison with the null distribution $P(r_s)$ constructed from the nonparametric sampling of randomly permuted ranks.

4 Conclusion

From the results of our participation in the SAMPL9 host–guest challenge, we conclude that expanded ensemble (EE) simulations are up to the task of accurately and efficiently predicting absolute binding free energies, achieving a mean absolute error of 2.29 kcal mol^{-1} for 13 small molecules against pillar[6]arene. Our improved EE protocol includes pre-optimization of the schedule of alchemical intermediates, and collecting statistical distributions of predicted free energies from parallel distributed computing. We expect further sampling improvements may still be possible, through more judicious choices of harmonic bias, and better anticipation of configurational transitions that occur on timescales similar to EE convergence times. Better treatment of electrostatics for acidic hosts, and general force field improvements also bode well for future implementations of EE methods for absolute binding free energy estimation.

Code, Data, and Submissions

SAMPL9 host–guest challenge instructions, experimental data, submissions and analysis are available at `https://github.com/samplchallenges/SAMPL9`. An interactive web-page containing of all the raw data for computed binding free estimates as well as additional figures for all three submissions is available at `https://vvoelz.github.io/sampl9-voelzlab/`. All of the code involving microstate enumeration, system preparation, optimization of alchemical intermediates, and analysis can be found in our GitHub repository: `https://github.com/vvoelz/sampl9-voelzlab`.

Author Contributions

Matthew F. D. Hurley: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – Original Draft **Robert M. Raddi**: Conceptualization, Data curation, Formal Analysis, Investigation, Writing - Original draft preparation. **Jason G. Pattis**: Investigation, Methodology, Resources, Software, Visualization, Writing - Original draft preparation. **Vincent A. Voelz**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Writing - Original draft preparation, Writing – Review Editing

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work is supported by NIH R01GM123296. This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189. We thank the participants of Folding@home, who made this work possible. We appreciate the National Institutes of Health for its support of the SAMPL project via R01GM124270 to David L. Mobley (UC Irvine).

Notes and references

1 H. S. Muddana, C. Daniel Varnado, C. W. Bielawski, A. R. Urbach, L. Isaacs, M. T. Geballe and M. K. Gilson, *Journal of computer-aided molecular design*, 2012, **26**, 475–487.

2 H. S. Muddana, A. T. Fenley, D. L. Mobley and M. K. Gilson, *Journal of computer-aided molecular design*, 2014, **28**, 305–317.

3 J. Yin, N. M. Henriksen, D. R. Slochower, M. R. Shirts, M. W. Chiu, D. L. Mobley and M. K. Gilson, *Journal of computer-aided molecular design*, 2017, **31**, 1–19.

4 A. Rizzi, S. Murkli, J. N. McNeill, W. Yao, M. Sullivan, M. K. Gilson, M. W. Chiu, L. Isaacs, B. C. Gibb, D. L. Mobley *et al.*, *Journal of computer-aided molecular design*, 2018, **32**, 937–963.

5 M. Amezcua, L. El Khoury and D. L. Mobley, *Journal of computer-aided molecular design*, 2021, **35**, 1–35.

6 M. Amezcua, J. Setiadi, Y. Ge and D. L. Mobley, *Journal of Computer-Aided Molecular Design*, 2022, **36**, 707–734.

7 C. D. Parks, Z. Gaieb, M. Chiu, H. Yang, C. Shao, W. P. Walters, J. M. Jansen, G. McGaughey, R. A. Lewis, S. D. Bembenek *et al.*, *Journal of computer-aided molecular design*, 2020, **34**, 99–119.

8 A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis and J. Moult, *Proteins: Structure, Function, and Bioinformatics*, 2021, **89**, 1607–1617.

9 S. Ackloo, R. Al-Awar, R. E. Amaro, C. H. Arrowsmith, H. Azevedo, R. A. Batey, Y. Bengio, U. A. Betz, C. G. Bologa, J. D. Chodera *et al.*, *Nature Reviews Chemistry*, 2022, **6**, 287–295.

10 M. R. Shirts, D. L. Mobley and J. D. Chodera, *Annual reports in computational chemistry*, 2007, **3**, 41–59.

11 A. S. Mey, B. Allen, H. E. B. Macdonald, J. D. Chodera, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi *et al.*, *arXiv preprint arXiv:2008.03067*, 2020.

12 K. Han, P. S. Hudson, M. R. Jones, N. Nishikawa, F. Tofoleanu and B. R. Brooks, *Journal of computer-aided molecular design*, 2018, **32**, 1059–1073.

13 Y. Khalak, G. Tresadern, B. L. de Groot and V. Gapsys, *Journal of computer-aided molecular design*, 2021, **35**, 49–61.

14 P. Procacci and G. Guarnieri, *The Journal of Chemical Physics*, 2022, **156**, 164104.

15 A. Lyubartsev, A. Martsinovski, S. Shevkunov and P. Vorontsov-Velyaminov, *The Journal of chemical physics*, 1992, **96**, 1776–1783.

16 J. I. Monroe and M. R. Shirts, *Journal of computer-aided molecular design*, 2014, **28**, 401–415.

17 H. Goel, A. Hazel, W. Yu, S. Jo and A. D. MacKerell, *New Journal of Chemistry*, 2022, **46**, 919–932.

18 D. R. Slochower, N. M. Henriksen, L.-P. Wang, J. D. Chodera, D. L. Mobley and M. K. Gilson, *Journal of chemical theory and computation*, 2019, **15**, 6225–6242.

19 T. Dixon, S. D. Lotz and A. Dickson, *Journal of computer-aided molecular design*, 2018, **32**, 1001–1012.

20 D. Markthaler, H. Kraus and N. Hansen, *Journal of Computer-Aided Molecular Design*, 2022, 1–9.

21 Y. Miao, A. Bhattarai and J. Wang, *Journal of chemical theory and computation*, 2020, **16**, 5526–5547.

22 A. Rizzi, T. Jensen, D. R. Slochower, M. Aldeghi, V. Gapsys, D. Ntekoumes, S. Bosisio, M. Papadourakis, N. M. Henriksen, B. L. De Groot *et al.*, *Journal of computer-aided molecular design*, 2020, **34**, 601–633.

23 J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr *et al.*, *The journal of physical chemistry B*, 2010, **114**, 2549–2564.

24 M. Ghorbani, P. S. Hudson, M. R. Jones, F. Aviat, R. Meana-Pañeda, J. B. Klauda and B. R. Brooks, *Journal of computer-aided molecular design*, 2021, **35**, 667–677.

25 S. Azimi, J. Z. Wu, S. Khuttan, T. Kurtzman, N. Deng and E. Gallicchio, *arXiv preprint arXiv:2107.05155*, 2021.

26 J.-F. Chen, J.-D. Ding and T.-B. Wei, *Chemical Communications*, 2021, **57**, 9029–9039.

27 G. Yu, M. Xue, Z. Zhang, J. Li, C. Han and F. Huang, *Journal of the American Chemical Society*, 2012, **134**, 13248–13251.

28 A. Laio and M. Parrinello, *Proceedings of the national academy of sciences*, 2002, **99**, 12562–12566.

29 W.-T. Hsu, V. Piomponi, P. T. Merz, G. Bussi and M. R. Shirts, *Journal of Chemical Theory and Computation*, 2023, **19**, 1805–1817.

30 F. Wang and D. P. Landau, *Physical review letters*, 2001, **86**, 2050.

31 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1**, 19–25.

32 X. Kong and C. L. Brooks III, *The Journal of chemical physics*, 1996, **105**, 2414–2423.

33 R. L. Hayes, J. Buckner and C. L. Brooks III, *Journal of chemical theory and computation*, 2021, **17**, 6799–6807.

34 V. A. Voelz, V. S. Pande and G. R. Bowman, *Biophysical Journal*, 2023.

35 M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora *et al.*, *Nature chemistry*, 2021, **13**, 651–659.

36 H. Achdout, A. Aimon, E. Bar-David and G. Morris, *BioRxiv*, 2020.

37 D. A. Sivak and G. E. Crooks, *Physical review letters*, 2012, **108**, 190602.

38 D. K. Shenfeld, H. Xu, M. P. Eastwood, R. O. Dror and D. E. Shaw, *Phys. Rev. E*, 2009, **80**, 46705.

39 A. Rizzi, *PhD thesis*, Weill Medical College of Cornell University, 2020.

40 R. Belardinelli and V. Pereyra, *The Journal of chemical physics*, 2007, **127**, 184105.

41 R. Belardinelli, S. Manzi and V. Pereyra, *Physical Review E*, 2008, **78**, 067701.

42 S. Zhang, D. F. Hahn, M. R. Shirts and V. A. Voelz, *Journal of Chemical Theory and Computation*, 2021, **17**, 6536–6547.

43 OpenEye Scientific Software Inc., *QUACPAC*, `http://www.eyesopen.com`.

44 J. Wagner, M. Thompson, D. Dotson, hyejang, Simon-Boothroyd and J. Rodríguez-Guerra, *openforcefield/openff-forcefields: Version 2.0.0 "Sage"*, 2021, `https://doi.org/10.`

5281/zenodo.5214478.

45 V. T. Lim, D. F. Hahn, G. Tresadern, C. I. Bayly and D. L. Mobley, *F1000Research*, 2020, **9**, year.

46 J. Wang, W. Wang, P. A. Kollman and D. A. Case, *Journal of molecular graphics and modelling*, 2006, **25**, 247–260.

47 A. Jakalian, D. B. Jack and C. I. Bayly, *Journal of computational chemistry*, 2002, **23**, 1623–1641.

48 OpenEye Scientific Software Inc., *OEDOCKING*, `http://www.eyesopen.com`.

49 M. McGann, *Journal of chemical information and modeling*, 2011, **51**, 578–596.

50 P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern *et al.*, *PLoS computational biology*, 2017, **13**, e1005659.

51 B. Hess, H. Bekker, H. J. Berendsen and J. G. Fraaije, *Journal of computational chemistry*, 1997, **18**, 1463–1472.

52 C. De Oliveira, H. S. Yu, W. Chen, R. Abel and L. Wang, *Journal of chemical theory and computation*, 2018, **15**, 424–435.

53 M. R. Gunner, T. Murakami, A. S. Rustenburg, M. Işık and J. D. Chodera, *Journal of Computer-Aided Molecular Design*, 2020, **34**, 561–573.

54 Q. Yang, Y. Li, J.-D. Yang, Y. Liu, L. Zhang, S. Luo and J.-P. Cheng, *Angewandte Chemie*, 2020, **132**, 19444–19453.

55 M. R. Shirts and J. D. Chodera, *The Journal of chemical physics*, 2008, **129**, 124105.

56 S. Boresch, F. Tettinger, M. Leitgeb and M. Karplus, *The Journal of Physical Chemistry B*, 2003, **107**, 9535–9551.

57 F. Clark, G. Robb, D. J. Cole and J. Michel, *Journal of Chemical Theory and Computation*, 2023.

58 H. Paliwal and M. R. Shirts, *Journal of chemical theory and computation*, 2011, **7**, 4115–4134.

59 E. C. Fieller, H. O. Hartley and E. S. Pearson, *Biometrika*, 1957, **44**, 470–481.