EXPANDED ENSEMBLE PREDICTIONS OF ABSOLUTE BINDING FREE ENERGIES IN THE SAMPL9 HOST–GUEST CHALLENGE

Matthew F. D. Hurley *
Department of Chemistry
Temple University
Philadelphia, PA 19122
matthew.hurley@temple.edu

Robert M. Raddi *
Department of Chemistry
Temple University
Philadelphia, PA 19122
rraddi@temple.edu

Jason G. Pattis
Department of Chemistry
Temple University
Philadelphia, PA 19122
jason.pattis@temple.edu

Vincent A. Voelz
Department of Chemistry
Temple University
Philadelphia, PA 19122
voelz@temple.edu

June 15, 2023

ABSTRACT

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]

*These authors contributed equally to this work.

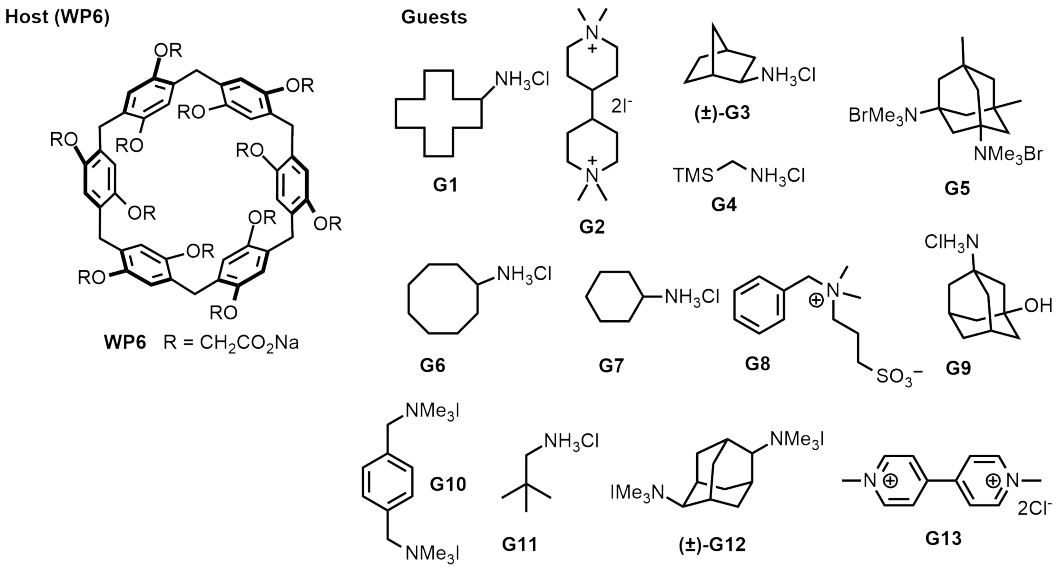

Figure 1: Molecular structures of the pillar[6]arene host (WP6) and guests (G1–G13).

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 of 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 be used as a chiral switch or for chiral sensing due to its planar chirality.[?] When WP6 is combined with an organic pyridinium salt guest, a change in pH can induce organization into nanotubes and vesicles.[26]

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

Here, we use the SAMPL9 challenge to test an expanded ensemble (EE) approach for 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 $\lambda = 0 \to 1$, and performing Markov chain Monte Carlo (MCMC) to propose moves from one intermediate thermodynamic ensemble parameterized by λ_k to another parameterized by λ_l. The Wang-Landau flat-histogram method[27] is used to adaptively learn a (reduced) bias potential $-\tilde{f}(\lambda)$ such that all intermediates are uniformly sampled. If a converged estimate of $\tilde{f}$ can be achieved, then $\tilde{f}(\lambda)$ is

equivalent to $f(\lambda)$, the (reduced) free energy versus λ, and the (reduced) free energy of the transformation can be estimated as $\Delta\tilde{f} = \tilde{f}_N - \tilde{f}_1$.[15]

A distinct advantage EE approaches have over alternative methods is the ability to sample all alchemical intermediates in a *single* simulation. This makes EE approaches ideal for large-scale virtual screening on distributed computing platforms such as Folding@home,[28] where simulation instances are necessarily asynchronous and only loosely uncoupled to other instances. Indeed, EE was recently deployed on Folding@home[29] to screen potential inhibitors of the SARS-CoV-2 main protease as part of the COVID Moonshot initiative.[30]

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, we have developed a simple scheme to optimize the schedule of λ_k values, by equally spacing ensembles in thermodynamic length,[32,33] similar to the "thermodynamic trailblazing" method of Rizzi et al.[34] (unpublished).

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.[35] 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 $\sim t^{-1}$,[36] 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}.[37] 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.[28,29] 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 of the WP6 host at pH 7.4, we considered the fluorescence emission spectra vs. pH published in Yu et al.[26] A titration curve fit to this data suggests 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, 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,[38] 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.

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.[39,40] 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[41] for G4. Partial charges for all molecules were assigned using AM1-BCC.[42]

Initial poses for receptor-ligand systems were prepared by docking guests to the host via OpenEye's OEDocking[43] module using the FRED score function[44] 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.[45] 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} to 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.[37]

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 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

$$\Delta G = \Delta G_{\text{rest}} + \Delta G_L - \Delta G_{RL}, \tag{1}$$

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 $\Delta G_1 = +6.42\,RT$. The $-\Delta G_{RL}$ term includes the free energy of removing this restraint.

Optimization of alchemical intermediates In order 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) that uses a steepest-descent algorithm to minimize the variance in the distributions $P(\Delta u_{kl})$, where Δu_{kl} is the change in (reduced) energy in going from thermodynamic ensemble k to a neighboring ensemble l. This has the effect of maximizing the EE transition probabilities.

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.[28,29]

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,[46] with frames 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.

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.

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.

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

$$\Delta G = -\ln \frac{\sum_{i \in \text{bound}} e^{-\Delta G_i}}{\sum_{i \in \text{unbound}} e^{-\Delta G_i}}, \tag{2}$$

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. 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.[47]

Model uncertainties σ_{model} were calculated as

$$\sigma_{\text{model}} = \left(\sigma_{\Delta G}^2 + \sigma_{\text{sys}}^2\right)^{1/2} \tag{3}$$

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.[39]

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.

Table 2: SMILES strings for all host (WP6) and guest microstates

Index	Guest	Microstate	Isomeric SMILES
0	G1	1	C1CCCCCC(CCCCC1)N
1	G1	2	C1CCCCCC(CCCCC1)[NH3+]
2	G2	1	C[N+]1(CCC(CC1)C2CC[N+](CC2)(C)C)C
3	G3	1	[NH3+][C@@H]1C[C@H]2CC[C@@H]1C2
4	G3	2	N[C@@H]1C[C@H]2CC[C@@H]1C2
5	G3	3	[NH2]C1CC2CCC1C2
6	G3	4	[NH3+]C1CC2CCC1C2
7	G4	1	C[Si](C)(C)C[NH3+]
8	G4	2	C[Si](C)(C)CN
9	G5	1	CC12CC3(CC(C1)(CC(C2)(C3)[N+](C)(C)C)[N+](C)(C)C)C
10	G6	1	C1CCCC(CCC1)[NH3+]
11	G6	2	C1CCCC(CCC1)N
12	G7	1	C1CCC(CC1)N
13	G7	2	C1CCC(CC1)[NH3+]
14	G8	1	C[N+](C)(CCCS(=O)(=O)[O-])Cc1ccccc1
15	G9	1	[NH3+]C12CC3CC(C1)CC(O)(C3)C2
16	G9	2	[NH2]C12CC3CC(C1)CC(O)(C3)C2
17	G10	1	C[N+](C)(C)Cc1ccc(cc1)C[N+](C)(C)C
18	G11	1	CC(C)(C)C[NH3+]
19	G11	2	CC(C)(C)CN
20	G12	1	C[N+](C)(C)[C@H]1[C@@H]2C[C@@H]3C[C@H]1C[C@H](C2)[C@H]3[N+](C)(C)C
21	G12	2	C[N+](C)(C)[C@H]1[C@H]2C[C@H]3C[C@@H]1C[C@@H](C2)[C@H]3[N+](C)(C)C
22	G13	1	C[n+]1ccc(cc1)c2cc[n+](cc2)C
1	WP6	-8	O=C([O-])COc1cc2c(OCC(=O)[O-])cc1Cc1cc(OCC(=O)O)c(cc1OCC(=O)[O-])Cc1cc... (OCC(=O)[O-])c(cc1OCC(=O)O)Cc1cc(OCC(=O)[O-])c(cc1OCC(=O)[O-])Cc1cc... (OCC(=O)[O-])c(cc1OCC(=O)O)Cc1cc(OCC(=O)O)c(cc1OCC(=O)[O-])C2
2	WP6	-10	O=C([O-])COc1cc2c(OCC(=O)[O-])cc1Cc1cc(OCC(=O)[O-])c(cc1OCC(=O)[O-])Cc1cc... (OCC(=O)[O-])c(cc1OCC(=O)O)Cc1cc(OCC(=O)[O-])c(cc1OCC(=O)[O-])Cc1cc... (OCC(=O)[O-])c(cc1OCC(=O)[O-])Cc1cc(OCC(=O)O)c(cc1OCC(=O)O)C2
3	WP6	-12	O=C([O-])COc1cc2c(OCC(=O)[O-])cc1Cc1cc(OCC(=O)[O-])c(cc1OCC(=O)[O-])Cc1cc... (OCC(=O)[O-])c(cc1OCC(=O)[O-])Cc1cc(OCC(=O)[O-])c(cc1OCC(=O)[O-])Cc1cc... (OCC(=O)[O-])c(cc1OCC(=O)[O-])Cc1cc(OCC(=O)[O-])c(cc1OCC(=O)[O-])C2

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 2), as described in Methods. The various charge states of the host and guest are given in Table 2). 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 $\sim$5 kcal mol^{-1} and yield relatively large uncertainty ($\sim$ 1 kcal mol^{-1}). As mentioned in Methods, G4 was parameterized using GAFF-2.11, which may be an influencing factor.

Overall performance of our ranked submission gave an overall mean absolute error (MAE) of 2.29 kcal mol^{-1}, root mean squared error (RMSE) of 2.74, and an R^2 of 0.54 (Figure 3), which falls just under the median of all SAMPL9 submissions (Table 3). 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.

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).

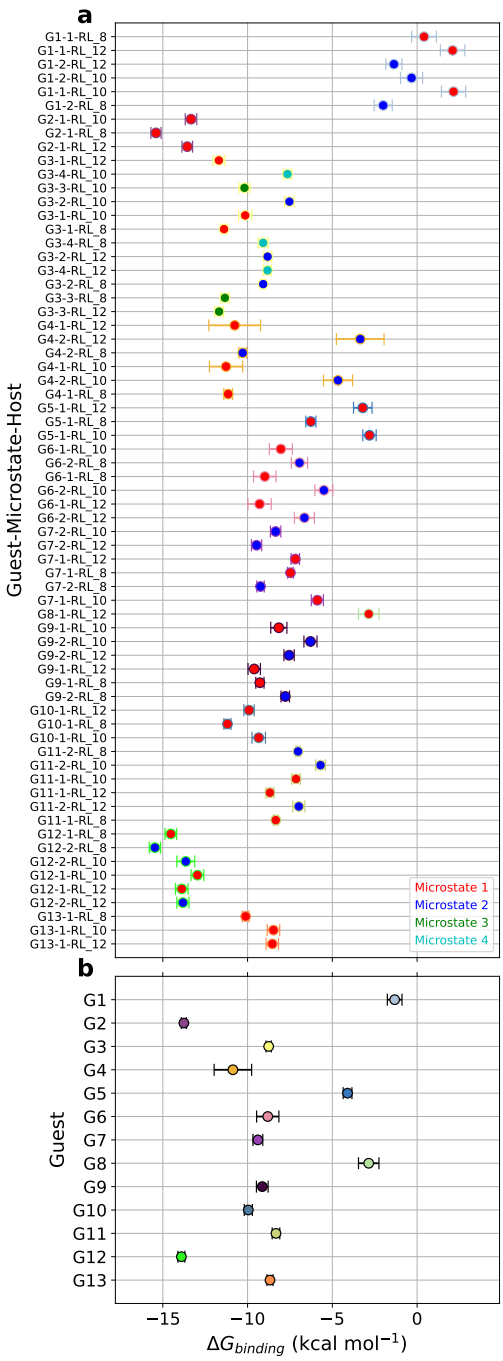

Figure 2: (a) Predicted free energies of host–guest binding using single microstates i ($\Delta G_{\text{binding}} = \Delta G_{\text{rest}} + \Delta G_{L,i} - \Delta G_{RL,i}$) between host microstates and guest microstates. Error bars are colored by guest (see panel b) and data markers are colored by 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 (2)

.

Our approach was able to predict the binding affinity for some guests better than others (Table 1 and Figure 3) 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.

Table 3: 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.

group	R^2	MAE	MSE	RMSE
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

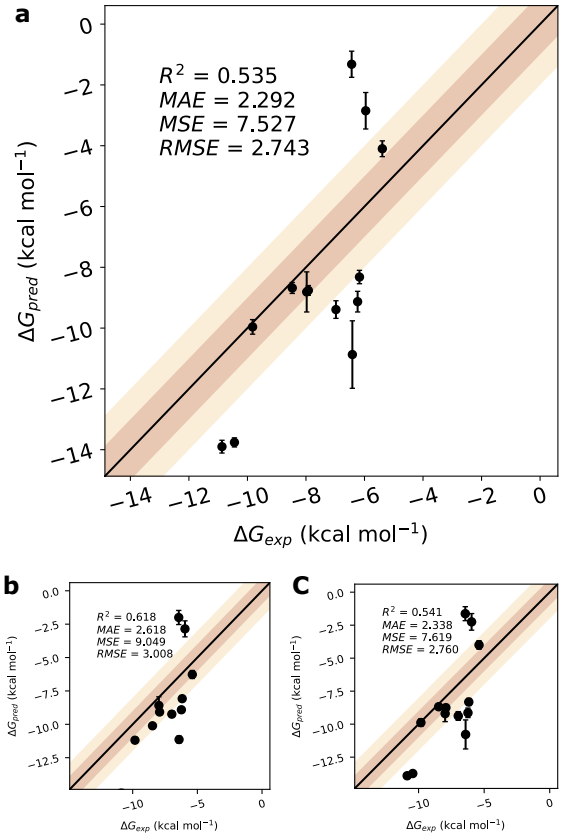

Figure 3: (a) Comparisons of predicted (ranked submission) vs. experimental binding free energies for all host–guest systems. Annotations report the correlation coefficient (R^2), mean absolute error (MAE), mean standard error (MSE), root-mean squared error (RMSE). (b) Comparisons for the "Voelz RL8" submission. (c) Comparisons for the "Voelz all" submission.

Our 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 4). Traces of expanded ensemble free energy estimates ΔG_L for guest-only decoupling display excellent convergence (Figure 5). 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 6).

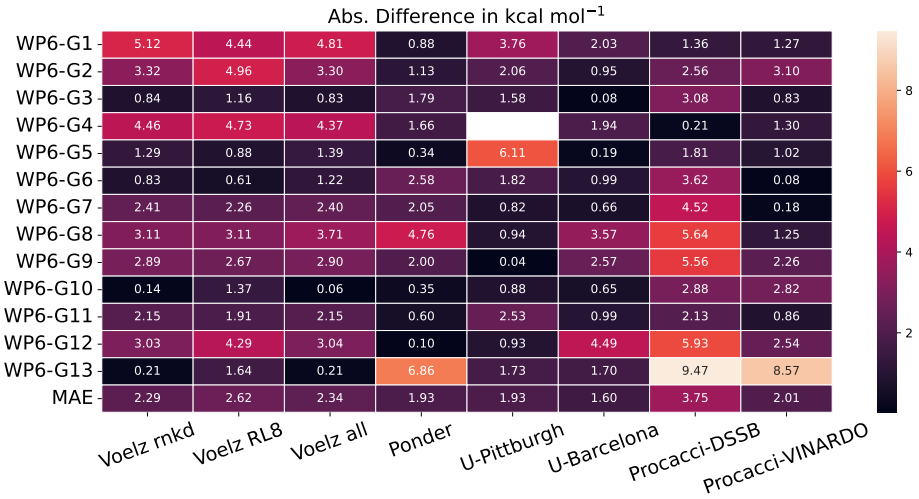

Figure 4: Predictions of host–guest binding free energies submitted by all SAMPL9 participants. The color map and numerical values inside the cells are the absolute difference in predictions against experiment (in kcal mol^{-1}). The last row is the mean absolute error (MAE) over all host–guest predictions for each group.

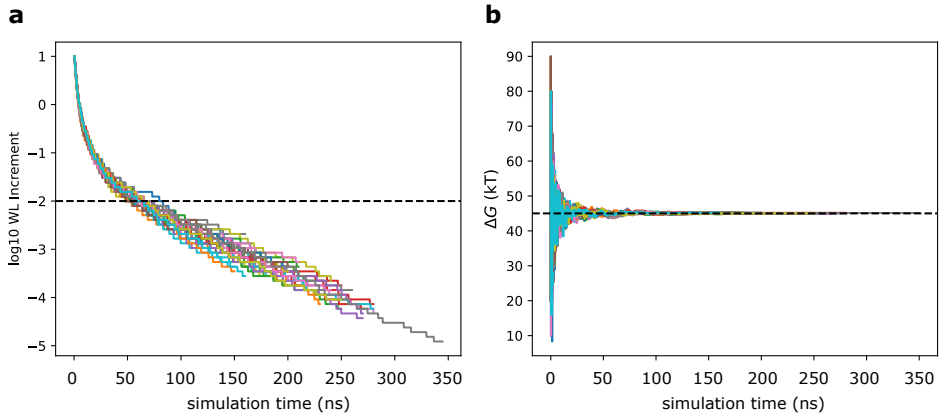

Figure 5: Convergence of expanded ensemble (EE) estimates of guest-only decoupling free energies ΔG_L for G13. Superimposed are (up to) fifty independent trajectories (distinguished by color). (a) The Wang-Landau (WL) increment vs. simulation time, with a dotted line denoting our 0.02 convergence threshold. (b) EE estimates of ΔG_L vs. simulation time, with a dotted line denoting the final estimate.

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 error 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

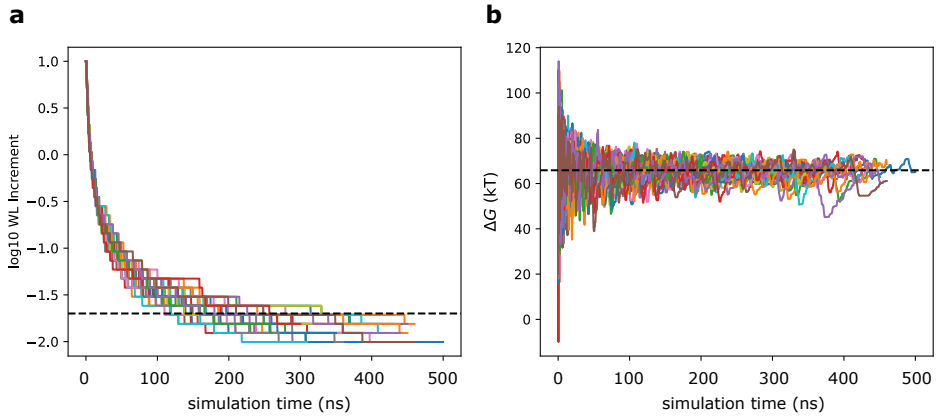

Figure 6: Convergence of expanded ensemble (EE) estimates of host–guest decoupling free energies ΔG_{RL} for G13. Superimposed are (up to) fifty independent trajectories (distinguished by color). (a) The Wang-Landau (WL) increment vs. simulation time, with a dotted line denoting our 0.02 convergence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

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 7) but can also be seen in the receptor-ligand (RL) trajectories (Figure 8). 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.

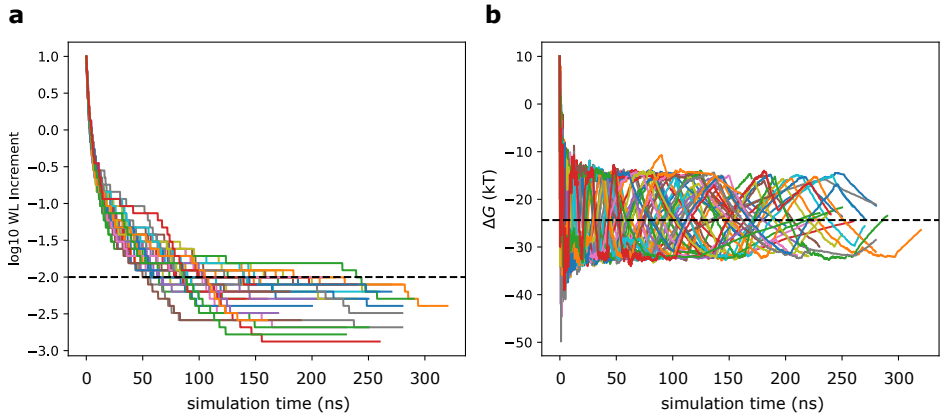

Figure 7: Convergence of expanded ensemble (EE) estimates of guest-only decoupling free energies ΔG_{L} for G6. Superimposed are (up to) fifty independent trajectories (distinguished by color). (a) The Wang-Landau (WL) increment vs. simulation time, with a dotted line denoting our 0.02 convergence threshold. (b) EE estimates of ΔG_{L} vs. simulation time, with a dotted line denoting the final estimate.

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

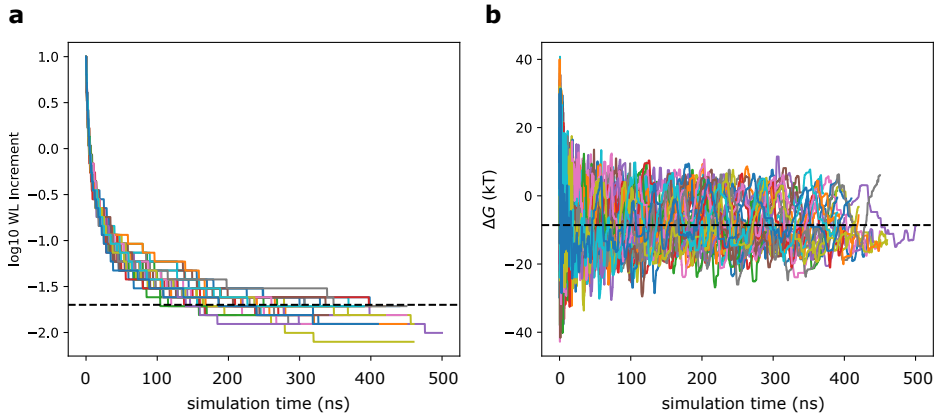

Figure 8: Convergence of expanded ensemble (EE) estimates of host–guest decoupling free energies ΔG_{RL} for G6. Superimposed are (up to) fifty independent trajectories (distinguished by color). (a) The Wang-Landau (WL) increment vs. simulation time, with a dotted line denoting our 0.02 convergence threshold.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

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)[48] to estimate the (negative) free energy reward of removing the restraint. The corrected ΔG values (Figure 9) 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.

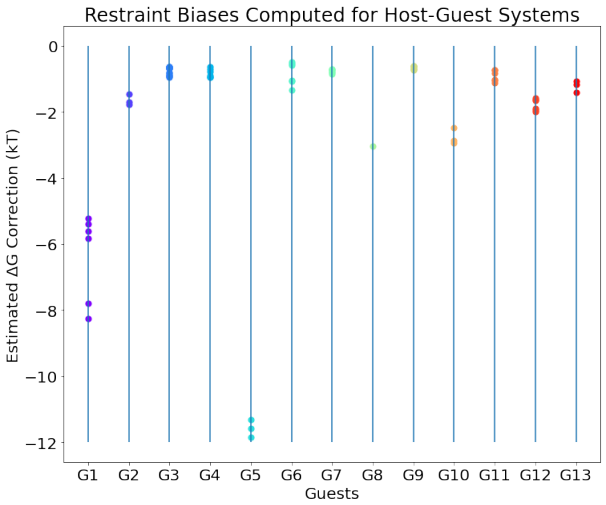

Figure 9: Estimated correction terms to our submitted free energy predictions for each guest, to correctly account for restraint biases.

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 predicted 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 the host, is a poor choice for this guest. Sampling issues thus may be a reason for the large prediction error for this guest.

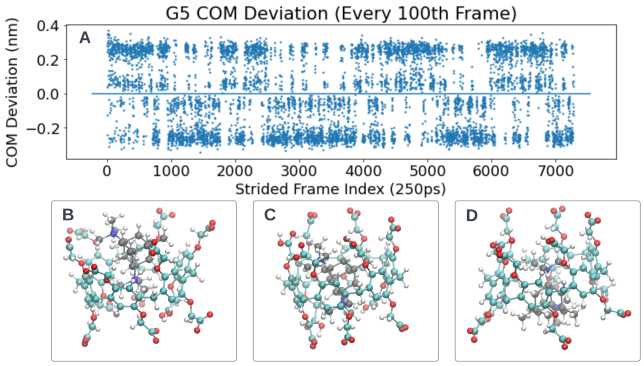

Figure 10: Transitions between a number of host-bound states are shown for G5. (a) A trace of the z-axis displacement of the guest center-of-mass (COM) with respect to the host COM shows free energy minima at both +/- edges of the host, in addition to those near the center of mass. (b,c, and d) depict 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 11). The uncorrected predictions tend to underestimate ΔG (i.e. predict tighter binding), and the correction exacerbates this trend.

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 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.

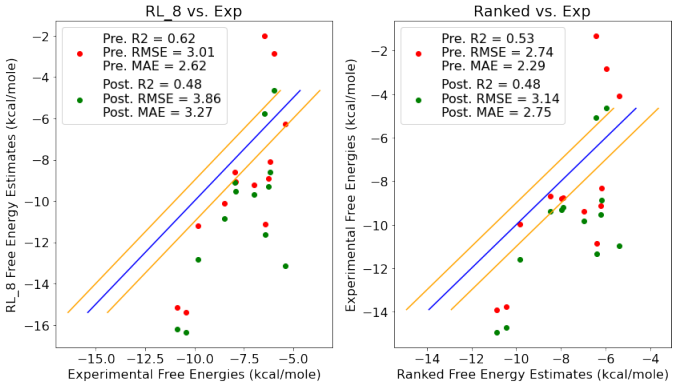

Figure 11: Changes to our submitted predicted before (red) and after (green) corrections are applied to correctly account for restraint biases. Left: "Voelz RL_8" submissions vs. experiment. Right: "Voelz ranked" submissions vs. experiment.

3.4 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

Table 4: 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

correlation coefficient,[49]

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{4}$$

where d_i are differences in (integer) ranks for each guest, and $n = 13$ is the number of ranked items (Figure 12 and Table 4). According to this metric, our "Voelz all" submissions give the most correctly ranked predictions compared to experiment, with an 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.

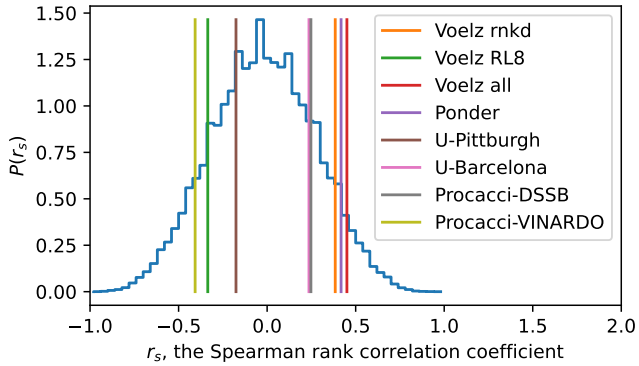

Figure 12: Spearman rank correlation coefficients r_s for participants in the SAMPL9 host–guest challenge, in comparison with the null distribution $P(r_s)$ constructed from the nonparametric sampling of randomly permuted ranks.

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 preparation**Robert M. Raddi**: Conceptualization, Data curation, Formal Analysis, Investigation, Visualization, 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).

References

[1] Hari S Muddana, C Daniel Varnado, Christopher W Bielawski, Adam R Urbach, Lyle Isaacs, Matthew T Geballe, and Michael K Gilson. Blind prediction of host–guest binding affinities: a new sampl3 challenge. *Journal of computer-aided molecular design*, 26:475–487, 2012.

[2] Hari S Muddana, Andrew T Fenley, David L Mobley, and Michael K Gilson. The sampl4 host–guest blind prediction challenge: an overview. *Journal of computer-aided molecular design*, 28:305–317, 2014.

[3] Jian Yin, Niel M Henriksen, David R Slochower, Michael R Shirts, Michael W Chiu, David L Mobley, and Michael K Gilson. Overview of the sampl5 host–guest challenge: Are we doing better? *Journal of computer-aided molecular design*, 31:1–19, 2017.

[4] Andrea Rizzi, Steven Murkli, John N McNeill, Wei Yao, Matthew Sullivan, Michael K Gilson, Michael W Chiu, Lyle Isaacs, Bruce C Gibb, David L Mobley, et al. Overview of the sampl6 host–guest binding affinity prediction challenge. *Journal of computer-aided molecular design*, 32(10):937–963, 2018.

[5] Martin Amezcua, Léa El Khoury, and David L Mobley. Sampl7 host–guest challenge overview: Assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *Journal of computer-aided molecular design*, 35(1):1–35, 2021.

[6] Martin Amezcua, Jeffry Setiadi, Yunhui Ge, and David L Mobley. An overview of the sampl8 host–guest binding challenge. *Journal of Computer-Aided Molecular Design*, 36(10):707–734, 2022.

[7] Conor D Parks, Zied Gaieb, Michael Chiu, Huanwang Yang, Chenghua Shao, W Patrick Walters, Johanna M Jansen, Georgia McGaughey, Richard A Lewis, Scott D Bembenek, et al. D3r grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of computer-aided molecular design*, 34:99–119, 2020.

[8] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiv. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021.

[9] Suzanne Ackloo, Rima Al-Awar, Rommie E Amaro, Cheryl H Arrowsmith, Hatylas Azevedo, Robert A Batey, Yoshua Bengio, Ulrich AK Betz, Cristian G Bologa, John D Chodera, et al. Cache (critical assessment of

computational hit-finding experiments): A public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nature Reviews Chemistry*, 6(4):287–295, 2022.

[10] Michael R Shirts, David L Mobley, and John D Chodera. Alchemical free energy calculations: ready for prime time? *Annual reports in computational chemistry*, 3:41–59, 2007.

[11] Antonia SJS Mey, Bryce Allen, Hannah E Bruce Macdonald, John D Chodera, Maximilian Kuhn, Julien Michel, David L Mobley, Levi N Naden, Samarjeet Prasad, Andrea Rizzi, et al. Best practices for alchemical free energy calculations. *arXiv preprint arXiv:2008.03067*, 2020.

[12] Kyungreem Han, Phillip S Hudson, Michael R Jones, Naohiro Nishikawa, Florentina Tofoleanu, and Bernard R Brooks. Prediction of cb [8] host–guest binding free energies in sampl6 using the double-decoupling method. *Journal of computer-aided molecular design*, 32(10):1059–1073, 2018.

[13] Yuriy Khalak, Gary Tresadern, Bert L de Groot, and Vytautas Gapsys. Non-equilibrium approach for binding free energies in cyclodextrins in sampl7: force fields and software. *Journal of computer-aided molecular design*, 35:49–61, 2021.

[14] Piero Procacci and Guido Guarnieri. Sampl9 blind predictions using nonequilibrium alchemical approaches. *The Journal of Chemical Physics*, 156(16):164104, 2022.

[15] AP Lyubartsev, AA Martsinovski, SV Shevkunov, and PN Vorontsov-Velyaminov. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *The Journal of chemical physics*, 96(3):1776–1783, 1992.

[16] Jacob I Monroe and Michael R Shirts. Converging free energies of binding in cucurbit [7] uril and octa-acid host–guest systems from sampl4 using expanded ensemble simulations. *Journal of computer-aided molecular design*, 28(4):401–415, 2014.

[17] Himanshu Goel, Anthony Hazel, Wenbo Yu, Sunhwan Jo, and Alexander D MacKerell. Application of site-identification by ligand competitive saturation in computer-aided drug design. *New Journal of Chemistry*, 46(3):919–932, 2022.

[18] David R Slochower, Niel M Henriksen, Lee-Ping Wang, John D Chodera, David L Mobley, and Michael K Gilson. Binding thermodynamics of host–guest systems with smirnoff99frosst 1.0. 5 from the open force field initiative. *Journal of chemical theory and computation*, 15(11):6225–6242, 2019.

[19] Tom Dixon, Samuel D Lotz, and Alex Dickson. Predicting ligand binding affinity using on-and off-rates for the sampl6 sampling challenge. *Journal of computer-aided molecular design*, 32(10):1001–1012, 2018.

[20] Daniel Markthaler, Hamzeh Kraus, and Niels Hansen. Binding free energies for the sampl8 cb8 "drugs of abuse" challenge from umbrella sampling combined with hamiltonian replica exchange. *Journal of Computer-Aided Molecular Design*, pages 1–9, 2022.

[21] Yinglong Miao, Apurba Bhattarai, and Jinan Wang. Ligand gaussian accelerated molecular dynamics (ligamd): Characterization of ligand binding thermodynamics and kinetics. *Journal of chemical theory and computation*, 16(9):5526–5547, 2020.

[22] Andrea Rizzi, Travis Jensen, David R Slochower, Matteo Aldeghi, Vytautas Gapsys, Dimitris Ntekoumes, Stefano Bosisio, Michail Papadourakis, Niel M Henriksen, Bert L De Groot, et al. The sampl6 sampling challenge: assessing the reliability and efficiency of binding free energy calculations. *Journal of computer-aided molecular design*, 34(5):601–633, 2020.

[23] Jay W Ponder, Chuanjie Wu, Pengyu Ren, Vijay S Pande, John D Chodera, Michael J Schnieders, Imran Haque, David L Mobley, Daniel S Lambrecht, Robert A DiStasio Jr, et al. Current status of the amoeba polarizable force field. *The journal of physical chemistry B*, 114(8):2549–2564, 2010.

[24] Mahdi Ghorbani, Phillip S Hudson, Michael R Jones, Félix Aviat, Rubén Meana-Pañeda, Jeffery B Klauda, and Bernard R Brooks. A replica exchange umbrella sampling (reus) approach to predict host–guest binding free energies in sampl8 challenge. *Journal of computer-aided molecular design*, 35(5):667–677, 2021.

[25] Solmaz Azimi, Joe Z Wu, Sheenam Khuttan, Tom Kurtzman, Nanjie Deng, and Emilio Gallicchio. Application of the alchemical transfer and potential of mean force methods to the sampl8 host-guest blinded challenge. *arXiv preprint arXiv:2107.05155*, 2021.

[26] Guocan Yu, Min Xue, Zibin Zhang, Jinying Li, Chengyou Han, and Feihe Huang. A water-soluble pillar [6] arene: synthesis, host–guest chemistry, and its application in dispersion of multiwalled carbon nanotubes in water. *Journal of the American Chemical Society*, 134(32):13248–13251, 2012.

[27] Fugao Wang and David P Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050, 2001.

[28] Vincent A Voelz, Vijay S Pande, and Gregory R Bowman. Folding@ home: achievements from over twenty years of citizen science herald the exascale era. *Biophysical Journal*, 2023.

[29] Maxwell I Zimmerman, Justin R Porter, Michael D Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L Mallimadugula, Catherine E Kuhn, Jonathan H Borowsky, Rafal P Wiewiora, et al. Sars-cov-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature chemistry*, 13(7):651–659, 2021.

[30] Hagit Achdout, Anthony Aimon, Elad Bar-David, and GM Morris. Covid moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *BioRxiv*, 2020.

[31] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[32] David A Sivak and Gavin E Crooks. Thermodynamic metrics and optimal paths. *Physical review letters*, 108(19):190602, 2012.

[33] Daniel K Shenfeld, Huafeng Xu, Michael P Eastwood, Ron O Dror, and David E Shaw. Minimizing Thermodynamic Length to Select Intermediate States for Free-Energy Calculations and Replica-Exchange Simulations. *Phys. Rev. E*, 80(4):46705, 2009.

[34] Andrea Rizzi. *Improving Efficiency and Scalability of Free Energy Calculations through Automatic Protocol Optimization*. PhD thesis, Weill Medical College of Cornell University, 2020.

[35] RE Belardinelli and VD Pereyra. Wang-landau algorithm: A theoretical analysis of the saturation of the error. *The Journal of chemical physics*, 127(18):184105, 2007.

[36] RE Belardinelli, S Manzi, and VD Pereyra. Analysis of the convergence of the 1/ t and wang-landau algorithms in the calculation of multidimensional integrals. *Physical Review E*, 78(6):067701, 2008.

[37] Si Zhang, David F. Hahn, Michael R. Shirts, and Vincent A. Voelz. Expanded ensemble methods can be used to accurately predict protein-ligand relative binding free energies. *Journal of Chemical Theory and Computation*, 17(10):6536–6547, 2021.

[38] OpenEye Scientific Software Inc. Quacpac.

[39] Jeff Wagner, Matt Thompson, David Dotson, hyejang, SimonBoothroyd, and Jaime Rodríguez-Guerra. openforcefield/openff-forcefields: Version 2.0.0 "Sage", August 2021.

[40] Victoria T Lim, David F Hahn, Gary Tresadern, Christopher I Bayly, and David L Mobley. Benchmark assessment of molecular geometries and energies from small molecule force fields. *F1000Research*, 9, 2020.

[41] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling*, 25(2):247–260, 2006.

[42] Araz Jakalian, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation. *Journal of computational chemistry*, 23(16):1623–1641, 2002.

[43] OpenEye Scientific Software Inc. Oedocking.

[44] Mark McGann. Fred pose prediction and virtual screening accuracy. *Journal of chemical information and modeling*, 51(3):578–596, 2011.

[45] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

[46] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[47] Qi Yang, Yao Li, Jin-Dong Yang, Yidi Liu, Long Zhang, Sanzhong Luo, and Jin-Pei Cheng. Holistic prediction of the pka in diverse solvents based on a machine-learning approach. *Angewandte Chemie*, 132(43):19444–19453, 2020.

[48] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12):124105, 2008.

[49] Edgar C Fieller, Herman O Hartley, and Egon S Pearson. Tests for rank correlation coefficients. i. *Biometrika*, 44(3/4):470–481, 1957.