# Overview of the SAMPL6 host-guest binding affinity prediction challenge

**Andrea Rizzi[1,2], Steven Murkli[3], John N. McNeill[3], Wei Yao[4], Matthew Sullivan[4], Michael K. Gilson[5], Michael W. Chiu[6], Lyle Isaacs[3], Bruce C. Gibb[4], David L. Mobley[7*], John D. Chodera[1*]**

[1]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; [2]Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY 10065, USA; [3]Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, USA; [4]Department of Chemistry, Tulane University, Louisiana, LA 70118, USA; [5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA; [6]Qualcomm Institute, University of California, San Diego, La Jolla, CA 92093, USA; [7]Department of Pharmaceutical Sciences and Department of Chemistry, University of California, Irvine, California 92697, USA

***For correspondence:**
dmobley@uci.edu (DLM); john.chodera@choderalab.org (JDC)

**Abstract** Accurately predicting the binding affinities of small organic molecules to biological macromolecules can greatly accelerate drug discovery by reducing the number of compounds that must be synthesized to realize desired potency and selectivity goals. Unfortunately, the process of assessing the accuracy of current computational approaches to affinity prediction against binding data to biological macromolecules is frustrated by several challenges, such as slow conformational dynamics, multiple titratable groups, and the lack of high-quality blinded datasets. Over the last several SAMPL blind challenge exercises, host-guest systems have emerged as a practical and effective way to circumvent these challenges in assessing the predictive performance of current-generation quantitative modeling tools, while still providing systems capable of possessing tight binding affinities. Here, we present an overview of the SAMPL6 host-guest binding affinity prediction challenge, which featured three supramolecular hosts: octa-acid (OA), the closely related tetra-endo-methyl-octa-acid (TEMOA), and cucurbit[8]uril (CB8), along with 21 small organic guest molecules. A total of 119 entries were received from 10 participating groups employing a variety of methods that spanned from electronic structure and movable type calculations in implicit solvent to alchemical and potential of mean force strategies using empirical force fields with explicit solvent models. While empirical models tended to obtain better performance than first-principle methods, it was not possible to identify a single approach that consistently provided superior results across all host-guest systems and statistical metrics. Moreover, the accuracy of the methodologies generally displayed a substantial dependence on the system considered, emphasizing the need for host diversity in blind evaluations. Several entries exploited previous experimental measurements of similar host-guest systems in an effort to improve their physical-based predictions via some manner of rudimentary machine learning; while this strategy succeeded in reducing systematic errors, it did not correspond to an improvement in statistical correlation. Comparison to previous rounds of the host-guest binding free energy challenge highlights an overall improvement in the correlation obtained by the affinity predictions for OA and TEMOA systems, but a surprising lack of improvement regarding root mean square error over the past several challenge rounds. The data suggests that further refinement of force field parameters, as well as improved treatment of chemical effects (e.g., buffer salt conditions, protonation states) may be required to further enhance predictive accuracy.

## Introduction

Quantitative physical and empirical modeling approaches have played a growing role in aiding and directing the design of small molecule biomolecular ligands for use as potential therapeutics or chemical probes [1–4, 23, 65]. The degree of inaccuracy of these predictions largely determines how effective they can be in prioritizing synthesis of small molecule ligands [105]. Retrospective estimates have suggested that current methodologies are capable of achieving about 1–2 kcal/mol inaccuracy for well-behaved protein-ligand systems [6, 122], but more work remains to be done to extend the applicability domain of these technologies.

Assessment of how much of this inaccuracy can be attributed to fundamental limitations of the *force field* in accurately modeling energetics is complicated by the presence of numerous additional factors [78]. Proteins are highly dynamic entities, and many common drug targets—such as kinases [115] and GPCRs [63]—possess slow dynamics with timescales of microseconds to milliseconds [62] that frustrate the computation of true equilibrium affinities. While there has been some attempt to curate benchmark sets of protein-ligand affinity data in well-behaved model protein-ligand systems that are believed to be mostly free of slow-timescale motions that would convolve convergence issues with forcefield inaccuracies [78], other effects can complicate assessment of the accuracy of physical modeling benchmarks. Ionizable residues, for example, comprise approximately 29% of all protein residues [56], and large-scale computational surveys suggest that 60% of all protein-ligand complexes undergo a change in ionization state upon binding [5], with several notable cases characterized experimentally [24, 25, 91, 111]. For physical or empirical modeling approaches that assume fixed protonation states throughout the complexation process, protonation state effects are hopelessly convolved with issues of force field inaccuracy.

### Host-guest systems are a tractable model for assessing force field inaccuracies

Over the last decade, supramolecular host-guest complexes have emerged as a practical and useful model system for the quantitative assessment of modeling errors for the interaction of druglike small molecules with receptors. Supramolecular hosts such as cucurbiturils, cavitands, and cyclodextrins can bind small druglike molecules with affinities similar to protein-ligand complexes [84, 85, 99]. The lack of slowly relaxing conformational degrees of freedom of these hosts eliminates the potential for slow microsecond-to-millisecond receptor relaxation timescales as a source of convergence issues [78], while the small size of these systems allows many methodologies to take advantage of faster simulation times to rapidly assess force field quality. The high solubilities of these systems permit high-quality biophysical characterization of their interactions via gold-standard methods such as isothermal titration calorimetry (ITC) and nuclear magnetic resonance (NMR) [22, 38, 114]. Additionally, the stability of supramolecular hosts at extreme pH allows for strict control of protonation states in a manner not possible with protein-ligand systems, allowing confounding protonation state effects to be eliminated from consideration if desired [114]. Collectively, these properties have made host-guest systems a productive route for revealing deficiencies in modern force fields through blind community challenge exercises we have organized as part of the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) series of blind prediction challenge [86, 88, 107, 127].

### SAMPL host-guest challenges have driven advances in our understanding of sources of error

The SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) challenges are a recurring series of blind prediction challenges for the computational chemistry community [76, Drug Design Data Resource]. Through these challenges, SAMPL aims to evaluate and advance computational tools for rational drug design: By focusing the community on specific phenomena relevant to drug discovery—such as the contribution of force field inaccuracy to binding affinity prediction failures—isolating these phenomena from other confounding factors in well-designed test systems, evaluating tools prospectively, enforcing data sharing to learn from failures, and releasing the resulting high-quality datasets into the community as benchmark sets, SAMPL has driven progress in a number of areas over five previous rounds of challenge cycles [9, 34, 35, 43, 44, 81, 82, 86, 88, 93, 107, 107, 108, 127].

More specifically, SAMPL host-guest challenges have provided key tests for modeling of binding interactions [78], motivating increased attention to how co-solvents and ions modulate binding (resulting in errors

of up to 5 kcal/mol when these effects are neglected) and the importance of adequately sampling water rearrangements [17, 78, 86, 127]. In turn, this detailed examination has resulted in clear improvements in subsequent SAMPL challenges [127], though host-guest binding remains difficult to model accurately [47], in part due to force field limitations (spawning new efforts to remedy major force field deficiencies [125]).

## SAMPL6 host-guest systems

Three hosts were selected for the SAMPL6 host-guest binding challenge from the Gibb Deep Cavity Cavitand (GDCC) [36, 48, 79, 80] and the cucurbituril (CB) [31, 69, 83] families (***Figure 1***). The guest ligand sets were purposefully selected for the SAMPL6 challenge. The utility of these particular host systems for evaluating free energy calculations has been reviewed in detail elsewhere [79, 80].

The two GDCCs, octa-acid (OA) [36] and tetra-endo-methyl-octa-acid (TEMOA) [33], are low-symmetry hosts with a basket-shaped binding site accessible through the larger entryway located at the top. These hosts also appeared in two previous SAMPL host-guest challenges— SAMPL4 [86] and SAMPL5 [127]—with the names of OAH and OAMe respectively with different sets of guests. OA and TEMOA differ by four methyl groups that reduce the size of the binding site entryway (***Figure 1***). Both hosts expose eight carboxyl groups that increase their solubility. The molecular structures of the eight guests selected for the SAMPL6 challenge for characterization against both OA and TEMOA are shown in ***Figure 1*** (denoted OA-G0 through OA-G7). These guests feature a single polar group situated at one end of the molecule that tends to be exposed to solvent when complexed, while the rest of the compound remains buried in the hydrophobic binding site.

A second set of guest ligands were developed for the host cucurbit[8]uril (CB8). This host previously appeared in the SAMPL3 host-guest binding challenge [87], but members of the same family or analogs such as cucurbit[7]uril (CB7) and CBClip [128] were featured in SAMPL4 and SAMPL5 challenges as well. CB8 is a symmetric ($D_{8h}$), ring-shaped host comprising eight identical glycoluril monomers linked by pairs of methylene bridges. Its top-bottom symmetry means that asymmetric guests have at least two symmetry-equivalent binding modes that can be kinetically separated by timescales not easily achievable by standard molecular dynamics (MD) or Monte Carlo simulations and may require special considerations, in particular in alchemical absolute binding free energy calculations [75]. The CB8 guest set (compounds CB8-G0 to CB8-G13 in ***Figure 1***) includes both fragment-like and bulkier drug-like compounds.

Some of the general modeling challenges posed by both families of host-guest systems have been characterized in previous studies. While their relatively rigid structure minimizes convergence difficulties associated with slow receptor conformational dynamics, both families have been shown to bind guest ligands via a dewetting processes—in which waters must be removed from the binding site to accommodate guests—in a manner that can frustrate convergence for strategies based on molecular simulation. In the absence of tight-binding guest ligands, the octa-acid host experiences fluctuations in the number of bound waters on timescales of several nanoseconds [30]; a similar phenomenon was observed in alchemical absolute binding free energy calculations of CB7 at intermediate alchemical states with partially decoupled Lennard-Jones interactions [101]. In addition, hosts in both families have been shown to bind ions that can compete with and lower the binding affinity of other guests in solution [37, 98, 109]. Depending on differences in concentration and composition, the effect on the binding free energy can be between 1–2 kcal/mol [85, 94, 98]. Sensitivity of the guest affinity to ion concentration has been observed also with computational methods [50, 89, 95], which suggests that careful modeling of the buffer conditions is in principle necessary for a meaningful comparison to experiments.

## Experimental host-guest affinity measurements

A detailed description of the experimental methodology used to collect binding affinity data for OA, TEMOA, and CB8 host-guest systems is described elsewhere [90]. Briefly, all host-guest binding affinities were determined via direct or competitive isothermal titration calorimetry (ITC) at 298 K. OA and TEMOA measurements were performed in 10 mM sodium phosphate buffer at pH 11.7±0.1 whereas CB8 guests binding affinities were measured in a 25 mM sodium phosphate buffer at pH 7.4. Phosphate buffer is a common choice of buffer for its relevance to biology, and can be prepared over a wide pH range for exerting control over

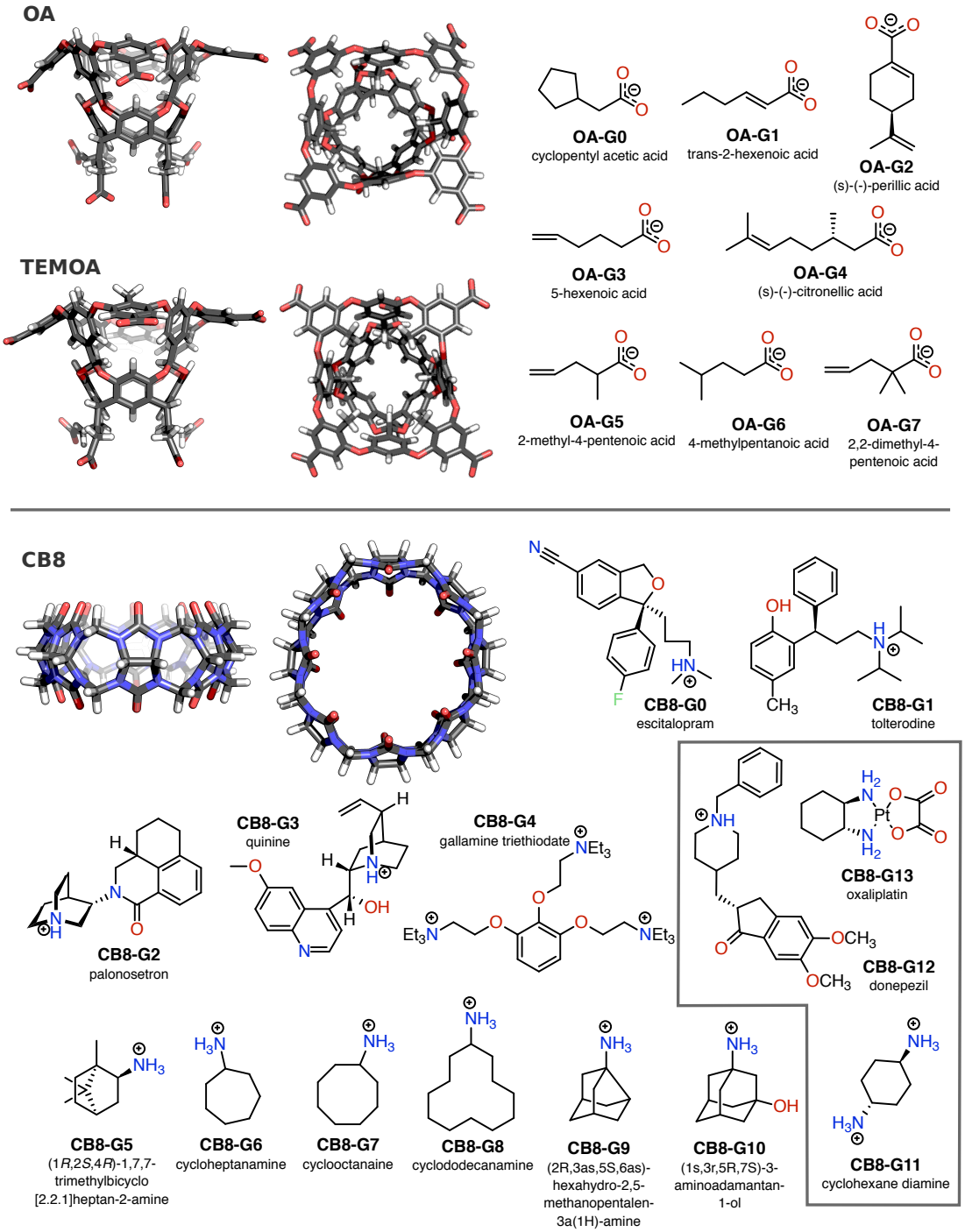



**Figure 1. Hosts and guests featured in the SAMPL6 host-guest blind challenge dataset.** Three-dimensional structures of the three hosts featured in the SAMPL6 challenge dataset (OA, TEMOA, and CB8) are shown in stick view from top and side perspective views. Carbon atoms are represented in gray, hydrogens in white, nitrogens in blue, and oxygens in red. Guest ligands for each complex are shown as two-dimensional chemical structures annotated by hyphenated host and guest names. Protonation states of the guest structures correspond to the predicted dominant microstate at the experimental pH at which binding affinities were collected, and matches those provided in the `mol2` and `sdf` input files shared with the participants when the challenge was announced. The same set of guests OA-G0 through OA-G7 was used for both OA and TEMOA hosts. The gray frame (lower right) contains the three CB8 guests that constitute the bonus challenge.

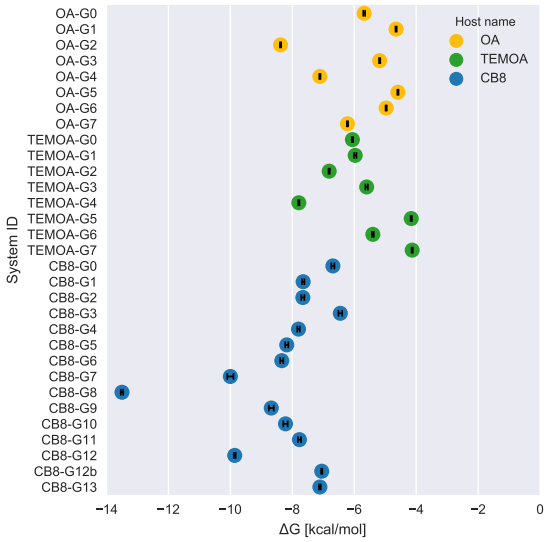



**Figure 2. Overview of experimental binding affinities for all host-guest complexes in the SAMPL6 challenge set.** Binding free energies ($\Delta G$) measured via isothermal titration calorimetry (ITC) are shown (filled circles), along with experimental uncertainties denoting standard error of the mean (black error bars), for OA (yellow), TEMOA (green), and CB8 (blue) complexes.

protonation states. Binding stoichiometries were determined by $^1$H NMR spectral integration and/or by ITC. The ITC titration curves were fitted to a single-site model or a competition model for all guests, except for CB8-G12 (donepezil), for which a sequential binding model was used. The stoichiometry coefficient was either fitted simultaneously with the other parameters or fixed to the value verified by the NMR titrations, which is the case for the CB8 guest set, as well as for OA-G5, TEMOA-G5, and TEMOA-G7.

To determine experimental uncertainties, we added the relative error in the nonlinear fit-derived association constant ($K_a$) or binding enthalpy ($\Delta H$) with the relative error in the titrant concentration in quadrature [20]. We decided to arbitrarily assume a relative error in the titrant concentration of 3% after personal communication with Professor Lyle Isaacs who suggested a value inferior to 5% based on his experience. The minimum relative nonlinear fit-derived uncertainty permitted was 1%, since the fit uncertainty was reported by the ITC software as smaller than this in some cases. It should be noted that the error propagation strategy adopted here assumes that the stoichiometry coefficient is fitted to the ITC data in order to absorb errors in cell volume and titrand concentration; this approach is exact only for the OA/TEMOA sets with the exclusion of OA-G5, TEMOA-G5, and TEMOA-G7, and an underestimate of the true error for the remaining cases. The error was then further propagated to the binding free energies and entropies that were calculated from $K_a$ and $\Delta H$. The final estimated experimental uncertainties are relatively small, never exceeding 0.1 kcal/mol.

The resulting experimental measurements with their uncertainties are reported in ***Table 1*** and ***Figure 2***. The dynamic range of the binding free energy $\Delta G$ spans 4.25 kcal/mol for the merged OA and TEMOA guest set, and 7.05 kcal/mol for CB8. The relatively wide cavity of CB8 enables binding stoichiometries different than 1:1. This is the case for three of the CB8 guests, specifically CB8-G1 (tolterodine), CB8-G4 (gallamine triethiodate), and CB8-G12 (donepezil). Curiously, while CB8-G12 was found to bind in 2:1 complexes (two guests bound to the same host), the NMR experiments determined stoichiometries of 1:2 and 1:3 for CB8-G1 and CB8-G4 respectively (one guest bound to multiple hosts). For the last two guests, the ITC titration curves fit well to a single set of sites binding model which indicates that the each of the binding events are equivalent. In ***Table 1*** and ***Figure 2*** we report the binding affinity of both the 1:1 and the 2:1 complex for CB8-G12, which are identified by CB8-G12a and CB8-G12b respectively, and the free energy of the 1:1 complex for CB8-G1 and CB8-G4.

# Methods

## Challenge design and logistics

### Challenge timeline

On August 24th, 2017, we released in a publicly accessible GitHub repository [51] a brief description of the host-guest systems and the experimental methodology, together with the challenge directions, and input files in `mol2` and `sdf` formats for the three hosts and their guests. The instructions shared online included information about buffer concentrations, temperature, and pH used for the experiments. The participants were asked to submit their predicted absolute binding free energies and, optionally, binding enthalpies, along with a detailed description of the methodology and the software employed through the Drug Design Data Resource (D3R) website (https://drugdesigndata.org/about/sampl6) by January 19th, 2018. We also encouraged the inclusion of uncertainties and/or standard error of the mean (SEM) of the predictions when available. The results of the experimental assays were released on January 26th in the same GitHub repository. The challenge culminated in a conference held on February 22–23, 2018 in La Jolla, CA where the participants shared lessons learned from participating in the challenge after performing retrospective analysis of their data.

### Bonus challenge

Three molecules in the CB8 guest sets, namely CB8-G11, CB8-G12, and CB8-G13, were proposed to participants as an optional bonus challenge since they were identified in advance to present some atypical difficulties for molecular modeling. In particular, the initial experimental data suggested both CB8-G11 and CB8-G12 to bind with 2:1 binding stoichiometry while CB8-G13 was deemed to be an especially challenging case for modeling due to the presence of a coordinated platinum atom, which is commonly not readily handled by classical force fields and usually requires larger basis sets for quantum mechanics (QM) calculations than those commonly employed with simple organic molecules. Further investigation after the start date of the challenge revealed an error in the calibration of a CB8 solution which affected the measurement of CB8-G11. After correcting the error, a 1:1 stoichiometry was recovered, and the experiment was repeated to validate the result. Unfortunately, the new data was obtained too late to send out a correction to all participants, so only six entries included predictions for this guest.

### Preparation of standard input files

Standard input files for the three hosts were generated for the previous rounds of the SAMPL host-guest binding challenge and uploaded to the repository unchanged, while the guests' atomic coordinates were generated from their SMILES string representation through the OMEGA library [46] in the OpenEye Toolkit (version 2017.Oct.b5) except for oxaliplatin (CB8-G13), which was generated with OpenBabel to handle the platinum atom. The compounds were then docked into their hosts with OpenEye's FRED docking facility [72, 73]. Stereochemistry of the 3D structures recapitulated the stereochemistry of compounds assayed experimentally; experimental assays for chiral compounds were enantiopure except OA-G5, which was measured as a racemic mixture. For this molecule, we picked at random one of the two enantiomers under the assumption that the guest chirality (for this guest with a single chiral center) would not affect the binding free energy to an achiral host such as OA and TEMOA since the system otherwise contains no chiral centers. This information was included in the instructions when the challenge was released. Guest `mol2` files also included AM1-BCC point charges generated with the AM1-BCC charge engine in the Quacpac tool from the OpenEye toolkit [54, 55]. ***Figure 1*** shows the protonation state of the molecules as provided in the input files, which reflects the most likely protonation state as predicted by Epik [41, 102] from the Schrödinger Suite 2017-2 (Schrödinger) at experimental buffer pH (11.7 for OA and 7.4 for CB8). This resulted in all molecules possessing a net charge, with the exception of oxaliplatin and the CB8 host, which have no acidic or basic groups. Specifically, the eight carboxyl groups of OA and TEMOA were modeled as deprotonated and charged. The instructions stated clearly that the protonation and tautomeric states provided were not guaranteed to be optimal. In particular, participants in the bonus challenge were advised to treat CB8-G12 with care as, in its protonated state, the nitrogen proton could be placed so that the substituent was axial or equatorial. The latter solution was arbitrarily adopted by the tools used to generate the input files for

CB8-G12.

## Statistical analysis of challenge entries

### Performance statistics

We computed root mean square error (RMSE), mean signed error (ME), coefficient of determination ($R^2$), and Kendall rank correlation coefficient ($\tau$) comparing experimentally determined binding free energies with blinded participant free energy predictions.

The mean signed error (ME), which quantifies the bias in predictions, was computed as

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta G_i^{(exp)} - \Delta G_i^{(calc)}\right) \tag{1}$$

where $\Delta G_i^{(exp)}$ and $\Delta G_i^{(calc)}$ are the experimental measurement of the binding free energy and its computational prediction respectively for the $i$-th molecule, and $N$ is the total number of molecules in the dataset. A positive ME reflects an overestimated binding free energy $\Delta G$ (or underestimated affinity $K_d = e^{-\beta\Delta G} \times (1\,\mathrm{M})$.

Some of the methods appearing in SAMPL6 were also used in previous rounds of the same challenge to predict relative binding free energies of similar host-guest systems. In order to comment on the performance of these methods over sequential challenges, for which statistics on absolute free energies are not readily available, we computed a separate set of statistics defined as *offset statistics*, as opposed to the *absolute statistics* defined above, in the same way they were reported in previous challenge overview papers. These statistics are computed identically to absolute statistics but by substituting $\Delta G_i^{(calc)}$ with

$$\Delta G_{i,o}^{(calc)} = \Delta G_i^{(calc)} - \mathrm{ME} \tag{2}$$

in the estimator expressions. The offset root mean square error computed from the $\Delta G_{i,o}^{(calc)}$ data points is termed $\mathrm{RMSE}_o$. It should be noted, however, that $R^2$ and $\tau$ are invariant under a constant shift of the data points. For this reason, we will use the symbols $R^2$ and $\tau$ both for the absolute and the offset correlation statistics.

Given the similarities of the two octa-acid hosts the set of their guest molecules, and that the large majority of the submitted methodologies were applied to both sets, we decided to report here the statistics computed using all the 16 predictions performed for OA and TEMOA (i.e., 8 predictions for each host). This merged set will be referred to as OA/TEMOA set in the rest of the work. The only method used to predict the binding free energies of the TEMOA set but not of the OA set was US-CGenFF (see ***Table 2*** for a schematic description of the methodology). We also decided to calculate separate statistics for the CB8 to highlight the general difference in performance between the predictions of the two host families. Statistics calculated on the two separate OA and TEMOA sets, as well as on the full dataset including CB8, OA, and TEMOA, are available on the GitHub repository (https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis/Accuracy/).

We generated bootstrap distributions of the statistics and computed 95-percentile bootstrap confidence intervals of the point estimates by generating 100 000 bootstrap samples through random sampling of the set of host-guest pairs with replacement. When the submission included SEMs for each prediction, we accounted for the statistical uncertainty in predictions by adding, for each bootstrap replicate, an additional Gaussian perturbation to the prediction with a standard deviation indicated by the SEM for that prediction.

### Null model

In order to compare the results obtained by the participants to a simple model that can be evaluated with minimal effort, we computed the binding free energy predicted by MM-GBSA rescoring [40] using Prime [52, 53] with the OPLS3 forcefield [45] in the Schrödinger Suite 2018-1 (Schrödinger). We used the same docked poses provided in the input files that were shared with all the participants as the initial coordinates for all the calculations. All docked positions were minimized before being rescored with the OPLS3 force field and the VSGB2.1 solvent model. The only exception to this was CB8-G4, which was manually re-docked into the host, as the initial structure contained steric clashes that could not be relaxed by minimization, causing the predicted binding free energy to spike to an unreasonable value of +2443 kcal/mol.

## Results

We received 42 submissions for the OA guest set, 43 for TEMOA, and 34 for CB8, for a total of 119 submissions, from 10 different participants, 5 of whom uploaded predictions for the three compounds in the bonus challenge as well. Only two groups submitted enthalpy predictions, which makes it impractical to draw general conclusions about the state of the field regarding the reliability of enthalpy predictions. Moreover, the predictive performance was generally poor (see Supplementary ***Figure 9***). The results of the enthalpy calculations are thus not discussed in detail here, but they are nevertheless available on the GitHub repository.

### Overview of the methodologies

Including the null model, 41 different methodologies were applied to one or more of the three datasets. In particular, the submissions included a total of 25 different variations of the movable type method exploring the effect of the input structures, the force field, the presence of conformational changes upon binding, and the introduction of previous experimental information on the free energy estimates. In order to facilitate the comparison among methods, we focus in this analysis on a representative subset of 7 different variations of the methodology. Supplementary ***Figure 7*** and Supplementary ***Figure 8*** show statistic bootstrap distributions and correlation plots for all the movable type free energy calculations submitted. As many of the methodologies are reported in detail elsewhere, in this section, we give a brief overview of the different strategies employed for the challenge to model the host-guest systems and estimate the binding free energies, and we leave the detailed descriptions of the various methodologies to the articles referenced in ***Table 2***.

#### Modeling

The majority of the participants either used the docked poses provided in the input files or ran a separate docking program to generate the initial complex conformation for the calculations. In few cases, the starting configuration was found by manually placing the guest inside the host. Surprisingly, the most common solvent model used in classical simulations was still TIP3P [57], a water model parameterized by Jorgensen 35 years ago for use with a fixed-cutoff Monte Carlo code neglecting long-range dispersion interactions and omitting long-range electrostatics. The only other explicit water models used in this round of the challenge were the significantly more modern AMOEBA [96] and TIP4P-Ew [49] water models, which was used to sample conformations to evaluate at the QM level. Implicit solvent models were adopted only in MMPBSA and for the movable type and QM calculations. We observed more variability in the treatment of buffer salt concentrations despite the known importance of this element in affecting the binding predictions, which may reflect a lack of standard practices in the field. Some entries modeled the buffer ionic strength explicitly with Na+ and Cl- ions while others included only the neutralizing counterions or used a uniform neutralizing charge. One of the participating groups submitted multiple variants of the SOMD method either utilizing only neutralizing counterions or including additional ions simulating the ionic strength at experimental conditions, which makes it possible to directly assess the effect of this modeling decision on the selected host-guest systems.

Most methods employing classical force fields used GAFF [121] or GAFF2 (still under active development) with AM1-BCC [54, 55] or RESP [12] charges, which were usually derived at the Hartree-Fock or MP2 level of theory. Other approaches made use of the AMOEBA polarizable model [96], CGenFF [28] or force matching [120] starting from CGenFF parameters. The movable type calculations utilized either the KECSA [129] scoring algorithm or the more recently developed GARF [11]. Several submissions employed QM potentials at the semi-empirical PM6-DH+ [64, 100] or DFT level of theory either modeling the full host-guest system or in hybrid QM/MM approaches that treated quantum mechanically the guest only. DFT calculations employed B3LYP [13], B3PW91 [13], or TPSS [116] functionals and often the DFT-D3 dispersion correction [42].

#### Sampling and free energy prediction

All the challenge entries used MD to sample host-guest conformations; uses of docking were limited to preparation of initial bound geometries for subsequent simulations. This was also the case also for QM

and movable type calculations, where samples generated from MD were in some cases clustered prior to quantum chemical energy evaluations. In a few cases, enhanced sampling techniques were used; in particular, the entries identified by DDM-FM and DDM-FM-QMM used Hamiltonian Replica Exchange (HREX) [113] as part of their double decoupling method (DDM) calculation [39] while Replica Exchange with Solute torsional Tempering (REST) [68, 71] was employed in FSDAM to generate from equilibrium the starting configurations for the fast switching protocol. Many groups used the double decoupling or the double annihilation method with purely classical force fields or with hybrid QM/MM potentials and either Bennett acceptance ratio (BAR) [15, 103] or the multistate Bennett acceptance ratio (MBAR) [104] to estimate free energies for the aggregated simulation data. Other classes of methodologies applied to this dataset include umbrella sampling (US) [119], movable type [130], MMPBSA [110], and free energy predictions based on QM calculations.

The repeat appearance of hosts chosen from the octa-acid and cucurbituril families as test systems for the SAMPL binding challenge, which reflects the continuous contribution of experimental data from the Gibb and Isaacs laboratories, led some groups to take advantage of previously available experimental data to improve their computational predictions. Several entries (e.g., SOMD-D, US-GAFF-C, and MovTyp-GE3L) were submitted with a linear[1] correction of the form

$$\Delta G^{(corrected)} \quad = \quad \mathrm{a} \cdot \Delta G^{(calc)} + \mathrm{b} \tag{3}$$

where the slope and offset coefficients (i.e., a and b respectively) were trained on data generated for previous rounds of the challenge. In some of the movable type calculations (e.g., MovTyp-GE3O), the coefficient $a$ was fixed to unity and the training data used to determine a purely additive bias correction. Relatedly, RFEC-GAFF2 and RFEC-QMMM, which included predictions for the OA and TEMOA guest sets, calculated the relative binding free energy between the compound and determined the offsets necessary to obtain absolute free energy using binding measurements of similar OA and TEMOA guests.

## Submission performance statistics

As mentioned above, we present here the statistics obtained by the challenge entries on the CB8 dataset and the merged OA and TEMOA dataset with the exception of US-CGenFF, for which we received a submission for the TEMOA set only. Moreover, since only a minority of entries had predictions for the bonus challenge, we excluded CB8-G11, CB8-G12, and CB8-G13 when computing the statistics of all the methodologies in order to compare them on the same set of compounds. ***Table 3*** reports such statistics with 95-percentile confidence intervals and ***Figure 4*** show the statistics bootstrap distributions. Some of the methods were used to estimate the binding free energy of only one between the OA/TEMOA and the CB8 sets, and, as a consequence, some of the table entries are missing. For the methodologies that made predictions of the bonus compounds, we report the statistics obtained including them separately in ***Table 4***. While it is difficult to isolate methods and models that performed very well across datasets and statistics, a few patterns emerged from comparing the different entries.

[1] Technically, this is an affine transformation in the general case since $b \neq 0$ for some of the corrections employed by participants, but we will refer to it as *linear* here.

**Table 1. Summary of ITC and NMR measurements for the SAMPL6 host-guest dataset.** Guest identifiers (ID), association constants ($K_a$), binding free energies ($\Delta G$), enthalpies ($\Delta H$), entropies at room temperature ($T\Delta S$) and stoichiometric ratios ($n$) as determined by ITC and NMR assays are reported for all compounds featured in the challenge. All quantities are reported as point estimates ± statistical error obtained by error propagation. For $K_a$ and $\Delta H$, the reported uncertainties incorporate both the uncertainty in the ITC enthalpogram least-squares fit and an assumed 3% uncertainty in titrant concentration. A minimum least-squares fit uncertainty of $1\%$ was assumed for fit errors reported by instrumentation as $< 1\%$. $\Delta G$ and $T\Delta S$ and their uncertainties were obtained from the first two quantities. Some of the compounds in the CB8 guest set can be bound by their hosts with stoichiometries different than 1:1. For CB8-G1 and CB8-G4, which can form 1:2 (two hosts bound to the same guest) and 1:3 complexes with CB8, respectively, we report the thermodynamic quantities of only one of the equivalent binding events—the value used to calculate the statistics for challenge entries. For CB8-G12, we report the measurements of both the 1:1 (CB8-G12a) and the 2:1 (CB8-G12b) bound complexes. The original data can be found at https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis/ExperimentalMeasurements/experimental_measurements.csv. Eventual updates or corrections to the data will be made available at the same URL, and anyone wishing to reuse the data should refer there.
[a] Point estimate and uncertainties computed from the $K_a$ measurements by error propagation.
[b] All experiments were performed at 298 K.
[c] The thermodynamic quantities given here represent the binding free energy and enthalpy of one of the $1/n$ equivalent binding events.
[d] Units of $M^{-2}$.

| ID | $K_a$ ($M^{-1}$) | $\Delta G$ (kcal/mol) [a] | $\Delta H$ (kcal/mol) | $T\Delta S$ (kcal/mol) [b] | $n$ |
|---|---|---|---|---|---|
| OA-G0 | $(147 \pm 7) \times 10^2$ | -5.68 ± 0.03 | -4.8 ± 0.2 | 0.8 ± 0.2 | 1 |
| OA-G1 | $(26 \pm 1) \times 10^2$ | -4.65 ± 0.02 | -5.5 ± 0.2 | -0.9 ± 0.2 | 1 |
| OA-G2 | $(140 \pm 6) \times 10^4$ | -8.38 ± 0.02 | -12.1 ± 0.5 | -3.7 ± 0.5 | 1 |
| OA-G3 | $(62 \pm 2) \times 10^2$ | -5.18 ± 0.02 | -7.5 ± 0.3 | -2.4 ± 0.3 | 1 |
| OA-G4 | $(164 \pm 7) \times 10^3$ | -7.11 ± 0.02 | -6.9 ± 0.3 | 0.2 ± 0.3 | 1 |
| OA-G5 | $(233 \pm 9) \times 10$ | -4.59 ± 0.02 | -5.3 ± 0.2 | -0.7 ± 0.2 | 1 |
| OA-G6 | $(44 \pm 2) \times 10^2$ | -4.97 ± 0.02 | -5.3 ± 0.2 | -0.3 ± 0.2 | 1 |
| OA-G7 | $(36 \pm 1) \times 10^3$ | -6.22 ± 0.02 | -7.4 ± 0.3 | -1.2 ± 0.3 | 1 |
| TEMOA-G0 | $(28 \pm 1) \times 10^3$ | -6.06 ± 0.02 | -7.8 ± 0.4 | -1.8 ± 0.4 | 1 |
| TEMOA-G1 | $(24 \pm 2) \times 10^3$ | -5.97 ± 0.04 | -8.2 ± 0.6 | -2.3 ± 0.6 | 1 |
| TEMOA-G2 | $(98 \pm 4) \times 10^3$ | -6.81 ± 0.02 | -9.3 ± 0.4 | -2.5 ± 0.4 | 1 |
| TEMOA-G3 | $(128 \pm 9) \times 10^2$ | -5.60 ± 0.04 | -8.9 ± 0.4 | -3.2 ± 0.4 | 1 |
| TEMOA-G4 | $(51 \pm 2) \times 10^4$ | -7.79 ± 0.02 | -8.9 ± 0.4 | -1.1 ± 0.4 | 1 |
| TEMOA-G5 | $(113 \pm 5) \times 10$ | -4.16 ± 0.02 | -8.0 ± 0.3 | -3.8 ± 0.3 | 1 |
| TEMOA-G6 | $(91 \pm 5) \times 10^2$ | -5.40 ± 0.03 | -6.2 ± 0.2 | -0.8 ± 0.2 | 1 |
| TEMOA-G7 | $(107 \pm 4) \times 10$ | -4.13 ± 0.02 | -8.3 ± 0.3 | -4.2 ± 0.3 | 1 |
| CB8-G0 | $(81 \pm 6) \times 10^3$ | -6.69 ± 0.05 | -4.2 ± 0.2 | 2.5 ± 0.2 | 1 |
| CB8-G1[c] | $(40 \pm 3) \times 10^4$ | -7.65 ± 0.04 | -5.0 ± 0.2 | 2.6 ± 0.2 | 0.5 |
| CB8-G2 | $(41 \pm 4) \times 10^4$ | -7.66 ± 0.05 | -6.5 ± 0.3 | 1.2 ± 0.3 | 1 |
| CB8-G3 | $(53 \pm 5) \times 10^3$ | -6.45 ± 0.06 | -2.5 ± 0.1 | 4.0 ± 0.2 | 1 |
| CB8-G4[c] | $(51 \pm 4) \times 10^4$ | -7.80 ± 0.04 | -9.8 ± 0.4 | -2.0 ± 0.4 | 0.33 |
| CB8-G5 | $(99 \pm 9) \times 10^4$ | -8.18 ± 0.05 | -3.2 ± 0.1 | 5.0 ± 0.1 | 1 |
| CB8-G6 | $(13 \pm 1) \times 10^5$ | -8.34 ± 0.05 | -5.7 ± 0.2 | 2.6 ± 0.2 | 1 |
| CB8-G7 | $(21 \pm 4) \times 10^6$ | -10.0 ± 0.1 | -6.5 ± 0.3 | 3.5 ± 0.3 | 1 |
| CB8-G8 | $(83 \pm 6) \times 10^8$ | -13.50 ± 0.04 | -14.4 ± 0.6 | -0.9 ± 0.6 | 1 |
| CB8-G9 | $(23 \pm 3) \times 10^5$ | -8.68 ± 0.08 | -4.6 ± 0.2 | 4.0 ± 0.2 | 1 |
| CB8-G10 | $(10 \pm 1) \times 10^5$ | -8.22 ± 0.07 | -2.00 ± 0.08 | 6.2 ± 0.1 | 1 |
| CB8-G11 | $(50 \pm 4) \times 10^4$ | -7.77 ± 0.05 | -2.11 ± 0.08 | 5.7 ± 0.1 | 1 |
| CB8-G12a | $(167 \pm 9) \times 10^5$ | -9.86 ± 0.03 | -9.2 ± 0.4 | 0.7 ± 0.4 | 1 |
| CB8-G12b | $(146 \pm 6) \times 10^3$ [d] | -7.05 ± 0.02 | -4.8 ± 0.2 | 2.2 ± 0.2 | 2 |
| CB8-G13 | $(161 \pm 8) \times 10^3$ | -7.11 ± 0.03 | -6.8 ± 0.3 | 0.3 ± 0.3 | 1 |

**Table 2. Summary of methodologies used by the participants in the SAMPL6 host-guest challenge.**
When a method uses multiple models (e.g., MM is used to generate the conformations to evaluate at the QM level in DFT(TPSS)-D3), only the energy and solvation models used for the final free energy prediction are listed. COSMO-RS: conductor-like screening model for real solvents [61]; DDM: double decoupling method [39]; FM: Force Matching [28]; FSDAM: Fast switching double annihilation method [92, 97] KMTISM: KECSA-Movable Type Implicit Solvation Model [131]; MD: molecular dynamics; MovTyp Movable Type method [130]; PBSA: Poisson-Boltzmann surface area [106]; REST: replica exchange with solute torsional tempering [68, 71]; RFEC: relative free energy calculation; QM/MM: mixed quantum mechanics and molecular mechanics; SOMD: double annihilation or decoupling method performed with Sire/OpenMM6.3 software [27, Woods et al.]; SQM: semi-empirical quantum mechanics; US: umbrella sampling [119]; VSGB2.1: VSGB2.0 solvation model refit to OPLS2.1/3/3e [67];
(a) Alchemical calculations are flagged by (A). All of these are absolute free energy calculations except for the RFEC entries.
(b) (E) and (I) denote explicit and implicit solvation models respectively.
(c) The corrections based on previous experimental data either apply only an additive term (offset) or both an additive term and a multiplicative factor (linear).
(d) Only a subset of the 25 movable type variations are included here. The four-letter suffix of each movable type submission is to be interpreted as following: first letter indicates the force field (G: GARF; K: KECSA), the second letter input structures (E: ensemble of structures from MD sampling; T: lowest energy structure during movable type scoring), the third letter is the number of states (1: only the complex is considered, 3: includes also the energy scores of host and guest in solution), and the fourth letter the type of experimental correction (L: linear; O: offset; N: no correction).
(e) Both RFEC-GAFF2 and RFEC-QMMM report the results of relative free energy calculations. The offsets were determined from experimental data for similar OA or TEMOA guests.
(f) SOMD submissions denoted with the *nobuffer* suffix include only the neutralizing counterions while the others add extra ions to model the buffer salt concentration. SOMD-A has no corrections. SOMD-B adds corrections for missing long-range dispersion interactions and for the flat-bottomed restraint to bring the ligand to standard state concentration. SOMD-D includes a linear correction fit to previously-available experimental data.

| Method ID(a) | Sampling | Energy model | Solvation model(b) | Experimental fit correction(c) | SAMPL6 reference |
|---|---|---|---|---|---|
| DDM-AMOEBA (A) | MD | AMOEBA | AMOEBA (E) | no | |
| DDM-FM (A) | HREX; MD | Force-Matching/RESP | TIP3P (E) | no | |
| DDM-FM-QMMM (A) | HREX; MD | Force-Matching/RESP; DFT(B3LYP) | TIP3P (E) | no | |
| DDM-GAFF (A) | MD | GAFF/AM1-BCC | TIP3P (E) | no | |
| DFT(B3PW91) | MD; clustering | DFT(B3PW91) | SMD (I) | no | |
| DFT(B3PW91)-D3 | MD; clustering | DFT(B3PW91)-D3 | SMD (I) | no | |
| DFT(TPSS)-D3 | MD | DFT(TPSS)-D3 | COSMO-RS (I) | no | |
| FSDAM (A) | REST; MD | GAFF2/AM1-BCC | TIP3P (E) | no | |
| NULL | docking | OPLS3 | VSGB2.1 (I) | no | |
| MMPBSA-GAFF | MD; clustering | GAFF/RESP | PBSA (I) | no | |
| MovTyp-GE3N(d) | MD; clustering | GARF | KMTISM (I) | no | |
| MovTyp-GE3O | MD; clustering | GARF | KMTISM (I) | offset | |
| MovTyp-GE3L | MD; clustering | GARF | KMTISM (I) | linear | |
| MovTyp-GT1N | MD; clustering | GARF | KMTISM (I) | no | |
| MovTyp-GT1L | MD; clustering | GARF | KMTISM (I) | linear | |
| MovTyp-KT1N | MD; clustering | KECSA | KMTISM (I) | no | |
| MovTyp-KT1L | MD; clustering | KECSA | KMTISM (I) | linear | |
| RFEC-GAFF2 (A)(e) | MD | GAFF2/RESP | TIP3P (E) | offset | |
| RFEC-QMMM (A) | MD | GAFF2/RESP; PM6-DH+ | TIP3P (E) | offset | |
| SQM(PM6-DH+) | MD | PM6-DH+ | COSMO-RS (I) | no | |
| SOMD-A (A)(f) | MD | GAFF/AM1-BCC | TIP3P (E) | no | |
| SOMD-A-nobuffer (A) | MD | GAFF/AM1-BCC | TIP3P (E) | no | |
| SOMD-C (A) | MD | GAFF/AM1-BCC | TIP3P (E) | no | |
| SOMD-C-nobuffer (A) | MD | GAFF/AM1-BCC | TIP3P (E) | no | |
| SOMD-D (A) | MD | GAFF/AM1-BCC | TIP3P (E) | linear | |
| SOMD-D-nobuffer (A) | MD | GAFF/AM1-BCC | TIP3P (E) | linear | |
| US-CGenFF | MD | CGenFF | TIP3P (E) | no | |
| US-GAFF | MD | GAFF/AM1-BCC | TIP3P (E) | no | |
| US-GAFF-C | MD | GAFF/AM1-BCC | TIP3P (E) | linear | |

Source: https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis/Submissions

**Table 3. Method performance statistics and bootstrap confidence intervals on OA/TEMOA and CB8 datasets.** Root mean square error (RMSE), mean signed error (ME), coefficient of determination ($R^2$), and Kendall correlation coefficient ($\tau$) obtained by each methodology on the merged OA/TEMOA and the CB8 datasets. The only exception is US-CGenFF whose OA/TEMOA statistics were computed using only the TEMOA set since no submission was received for OA. Table entries are left blank for those methods that were applied to only one of the guest sets. The predictions performed for the bonus challenge guests were excluded when computing the statistics for the CB8 dataset. Each statistic is reported with bootstrap distribution mean (between parentheses) and 95-percentile bootstrap confidence interval (square brackets) obtained through 100 000 cycles of resampling with replacement. The standard errors of the mean of the predictions reported in the submissions are included in the confidence intervals. The original data for the combined OA/TEMOA and CB8 datasets can be found respectively at https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis/Accuracy/OA-TEMOA/StatisticsTables/statistics.csv and https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis/Accuracy/CB8-NOBONUS/StatisticsTables/statistics.csv. Eventual updates or corrections to the data will be made available at the same URL, and anyone wishing to reuse the data should refer there.

| Method | OA/TEMOA dataset | | | | CB8 dataset (no bonus challenge) | | | |
|---|---|---|---|---|---|---|---|---|
| | RMSE | ME | $R^2$ | $\tau$ | RMSE | ME | $R^2$ | $\tau$ |
| DDM-AMOEBA | 2.2 (2.1) [1.3, 2.9] | -0.8 (-0.8) [-1.9, 0.1] | 0.1 (0.2) [0.0, 0.5] | -0.2 (-0.2) [-0.6, 0.2] | 3.9 (3.8) [1.5, 5.7] | 2.3 (2.3) [0.5, 4.3] | 0.1 (0.3) [0.0, 0.8] | 0.1 (0.1) [-0.5, 0.6] |
| DDM-FM | | | | | 4.7 (4.6) [3.4, 6.0] | 2.7 (2.7) [0.2, 4.8] | 0.4 (0.4) [0.0, 0.9] | 0.5 (0.5) [0.1, 0.9] |
| DDM-FM-QMMM | | | | | 5.5 (5.5) [4.0, 7.3] | 2.3 (2.3) [-1.1, 5.0] | 0.4 (0.5) [0.1, 0.8] | 0.6 (0.6) [0.1, 0.9] |
| DDM-GAFF | 3.4 (3.3) [2.1, 4.6] | 1.0 (1.0) [-0.7, 2.5] | 0.1 (0.3) [0.0, 0.8] | 0.4 (0.4) [-0.1, 0.8] | 7.2 (7.2) [5.6, 8.5] | 6.4 (6.4) [4.4, 8.2] | 0.3 (0.3) [0.0, 0.7] | 0.3 (0.3) [-0.2, 0.8] |
| DFT(B3PW91) | 16.7 (15.9) [8.7, 25.2] | 5.4 (5.4) [-3.4, 11.1] | 0.3 (0.3) [0.0, 0.7] | -0.2 (-0.2) [-0.7, 0.2] | 17.7 (17.5) [11.9, 22.9] | -14.8 (-14.8) [-20.6, -9.1] | 0.0 (0.1) [0.0, 0.7] | 0.3 (0.3) [-0.3, 0.8] |
| DFT(B3PW91)-D3 | 37.7 (37.7) [34.8, 40.1] | 35.3 (35.3) [27.8, 39.9] | 0.1 (0.3) [0.0, 0.8] | 0.4 (0.4) [-0.1, 0.9] | 36.6 (36.4) [28.7, 44.2] | 33.9 (33.9) [25.7, 42.0] | 0.0 (0.2) [0.0, 0.8] | -0.3 (-0.3) [-0.7, 0.2] |
| DFT(TPSS)-D3 | 3.1 (3.0) [2.3, 3.7] | -1.6 (-1.6) [-2.8, -0.2] | 0.5 (0.5) [0.1, 0.8] | 0.3 (0.4) [-0.1, 0.7] | | | | |
| FSDAM | 2.5 (2.5) [1.5, 3.3] | 0.8 (0.8) [-0.5, 1.9] | 0.5 (0.5) [0.0, 0.9] | 0.5 (0.5) [0.0, 0.9] | | | | |
| MMPBSA-GAFF | 7.0 (7.0) [5.4, 8.5] | 6.4 (6.4) [4.9, 7.9] | 0.8 (0.8) [0.6, 0.9] | 0.7 (0.6) [0.4, 0.8] | 17.9 (17.8) [13.9, 21.5] | 16.7 (16.7) [13.0, 20.6] | 0.0 (0.3) [0.0, 0.8] | -0.4 (-0.4) [-0.9, 0.1] |
| MovTyp-GE3L | 2.3 (2.3) [1.5, 3.1] | 1.9 (1.9) [1.2, 2.6] | 0.3 (0.4) [0.0, 0.8] | 0.3 (0.3) [-0.2, 0.7] | 5.8 (5.8) [4.5, 7.0] | 5.4 (5.4) [4.1, 6.7] | 0.3 (0.4) [0.1, 0.8] | 0.5 (0.5) [0.1, 0.8] |
| MovTyp-GE3N | 1.8 (1.8) [1.0, 2.6] | 1.2 (1.2) [0.5, 1.9] | 0.3 (0.4) [0.0, 0.8] | 0.3 (0.3) [-0.2, 0.7] | 4.7 (4.7) [3.5, 5.8] | -4.2 (-4.2) [-5.4, -2.8] | 0.3 (0.4) [0.1, 0.8] | 0.5 (0.5) [0.1, 0.8] |
| MovTyp-GE3O | 1.3 (1.3) [0.8, 1.8] | 0.8 (0.8) [0.3, 1.3] | 0.3 (0.4) [0.0, 0.8] | 0.3 (0.3) [-0.1, 0.7] | 5.3 (5.3) [4.2, 6.3] | 5.0 (5.0) [3.9, 6.1] | 0.3 (0.4) [0.1, 0.8] | 0.5 (0.5) [0.1, 0.8] |
| MovTyp-GT1L | 3.3 (3.3) [2.8, 3.8] | 3.2 (3.2) [2.7, 3.6] | 0.5 (0.5) [0.1, 0.8] | 0.4 (0.4) [-0.1, 0.8] | 5.4 (5.4) [4.1, 6.6] | -5.0 (-5.0) [-6.3, -3.7] | 0.4 (0.4) [0.1, 0.7] | 0.5 (0.5) [0.1, 0.9] |
| MovTyp-GT1N | 4.4 (4.4) [3.9, 4.9] | 4.3 (4.3) [3.8, 4.8] | 0.5 (0.5) [0.1, 0.8] | 0.4 (0.4) [-0.1, 0.8] | 2.0 (2.0) [1.3, 2.6] | -0.4 (-0.4) [-1.5, 0.8] | 0.4 (0.4) [0.1, 0.7] | 0.5 (0.5) [0.1, 0.9] |
| MovTyp-KT1L | 1.0 (0.9) [0.7, 1.2] | -0.5 (-0.5) [-0.9, -0.0] | 0.6 (0.6) [0.2, 0.8] | 0.4 (0.4) [-0.1, 0.7] | 2.9 (2.8) [1.7, 3.8] | 1.6 (1.6) [0.2, 3.0] | 0.1 (0.1) [0.0, 0.5] | 0.1 (0.1) [-0.4, 0.6] |
| MovTyp-KT1N | 2.9 (2.9) [2.4, 3.3] | 2.7 (2.7) [2.3, 3.2] | 0.5 (0.5) [0.1, 0.8] | 0.3 (0.3) [-0.1, 0.7] | 4.8 (4.8) [3.9, 5.6] | 4.4 (4.4) [3.3, 5.4] | 0.1 (0.1) [0.0, 0.5] | 0.1 (0.1) [-0.4, 0.6] |
| **NULL** | 26.3 (26.2) [23.3, 29.2] | 25.6 (25.6) [22.8, 28.6] | 0.6 (0.6) [0.2, 0.8] | 0.5 (0.6) [0.2, 0.8] | 17.6 (17.6) [14.2, 21.2] | 14.9 (14.9) [8.7, 19.9] | 0.0 (0.1) [0.0, 0.5] | -0.1 (-0.1) [-0.6, 0.5] |
| RFEC-GAFF2 | 1.5 (1.5) [1.2, 1.8] | -1.2 (-1.2) [-1.6, -0.7] | 0.7 (0.7) [0.3, 0.9] | 0.6 (0.6) [0.2, 0.9] | | | | |
| RFEC-QMMM | 1.6 (1.6) [1.3, 2.0] | -1.0 (-1.0) [-1.6, -0.3] | 0.8 (0.8) [0.6, 0.9] | 0.8 (0.7) [0.5, 0.9] | | | | |
| SOMD-A | 5.7 (5.7) [4.7, 6.6] | 5.4 (5.4) [4.5, 6.3] | 0.8 (0.8) [0.6, 0.9] | 0.8 (0.7) [0.4, 0.9] | 5.1 (5.2) [3.8, 6.8] | 4.4 (4.4) [2.6, 6.1] | 0.1 (0.2) [0.0, 0.8] | 0.1 (0.1) [-0.6, 0.7] |
| SOMD-A-nobuffer | 4.9 (4.9) [3.9, 5.9] | 4.5 (4.5) [3.6, 5.5] | 0.8 (0.8) [0.6, 0.9] | 0.7 (0.7) [0.4, 0.9] | 7.9 (7.9) [6.1, 9.7] | 7.3 (7.3) [5.5, 9.2] | 0.1 (0.2) [0.0, 0.7] | 0.1 (0.0) [-0.6, 0.6] |
| SOMD-C | 3.7 (3.7) [2.7, 4.5] | 3.2 (3.2) [2.4, 4.1] | 0.8 (0.8) [0.6, 0.9] | 0.7 (0.7) [0.4, 0.9] | 3.8 (3.9) [2.7, 5.5] | 2.8 (2.8) [1.0, 4.6] | 0.1 (0.2) [0.0, 0.8] | 0.1 (0.1) [-0.6, 0.7] |
| SOMD-C-nobuffer | 3.0 (3.0) [2.1, 3.9] | 2.4 (2.4) [1.4, 3.3] | 0.8 (0.8) [0.5, 0.9] | 0.7 (0.7) [0.4, 0.9] | 6.6 (6.6) [4.9, 8.4] | 6.0 (6.0) [4.1, 7.8] | 0.1 (0.2) [0.0, 0.7] | 0.1 (0.1) [-0.6, 0.6] |
| SOMD-D | 1.8 (1.7) [1.1, 2.4] | 1.0 (1.0) [0.4, 1.8] | 0.8 (0.8) [0.6, 0.9] | 0.7 (0.7) [0.4, 0.9] | 2.6 (2.6) [1.4, 3.8] | -1.8 (-1.8) [-3.1, -0.7] | 0.1 (0.2) [0.0, 0.8] | 0.1 (0.1) [-0.6, 0.7] |
| SOMD-D-nobuffer | 1.6 (1.6) [1.0, 2.2] | 0.3 (0.3) [-0.5, 1.1] | 0.8 (0.8) [0.5, 0.9] | 0.7 (0.7) [0.4, 0.9] | 1.9 (1.9) [1.2, 2.7] | -0.2 (-0.2) [-1.4, 1.0] | 0.1 (0.2) [0.0, 0.7] | 0.1 (0.0) [-0.6, 0.6] |
| SQM(PM6-DH+) | 2.7 (2.6) [1.8, 3.4] | 1.1 (1.1) [-0.0, 2.3] | 0.1 (0.2) [0.0, 0.7] | 0.3 (0.3) [-0.2, 0.8] | | | | |
| US-CGenFF | 1.3 (1.4) [0.7, 2.1] | -0.1 (-0.1) [-1.1, 1.0] | 0.5 (0.5) [0.0, 1.0] | 0.4 (0.4) [-0.4, 1.0] | | | | |
| US-GAFF | 2.9 (2.9) [2.2, 3.5] | 2.5 (2.5) [1.8, 3.2] | 0.9 (0.8) [0.6, 1.0] | 0.7 (0.7) [0.4, 0.9] | 8.0 (7.9) [4.9, 11.0] | 6.7 (6.7) [4.3, 9.5] | 0.0 (0.2) [0.0, 0.6] | -0.1 (-0.1) [-0.6, 0.5] |
| US-GAFF-C | 1.0 (0.9) [0.6, 1.2] | -0.5 (-0.5) [-0.9, -0.1] | 0.9 (0.8) [0.6, 1.0] | 0.7 (0.7) [0.4, 0.9] | 3.5 (3.4) [1.4, 5.2] | 1.6 (1.6) [-0.1, 3.6] | 0.0 (0.2) [0.0, 0.6] | -0.1 (-0.1) [-0.6, 0.5] |

## Challenge entries generally performed better on OA/TEMOA than CB8

In general, the CB8 guest set proved to be more challenging than the OA/TEMOA set both in terms of error and correlation statistics. It is rarely the case that the same method scored better statistics on the former set, and only MovTyp-GT1N does so with statistical significance while the opposite can be observed relatively often. ***Figure 5***-A shows the root mean square error (RMSE) and mean signed error (ME) with 95-percentile bootstrap confidence interval computed for each molecule using the ten methods that scored best in RMSE statistics in the merged OA/TEMOA set or the CB8 set (excluding the bonus challenge), which formed a set of 14 different techniques employing GAFF and GAFF2 [121], CGenFF [120], force matching [28], AMOEBA [96], and QM/MM potentials using DFT(B3LYP) [13] or PM6-DH+ [64, 100]. These top ten methods performed poorly on eight out of the eleven CB8 compounds, and while confidence intervals for all the statistics are generally large, they also performed significantly worse on several CB8 guests than the OA/TEMOA ligands they accurately predicted affinities for. This loss of accuracy seems to be fairly consistent across models and methodologies, but the data is not sufficient to determine the exact cause of this behavior (e.g., force field parameters, the generally larger dimensions of the CB8 guests, protonation states). However, the results of the related SAMPL6 SAMPLing challenge does suggest that properly accounting for slow conformational dynamics for some of the CB8 guests may require longer simulation times than for the OA compounds [51], which may have contributed to poorer performance over the OA set. Moreover, explicitly modeling the buffer salt concentration in SOMD significantly reduced the difference in error on the two guest sets (compare SOMD-C with SOMD-C-nobuffer), albeit without a commensurate improvement in correlation statistics, so the issue of missing chemical effects may also have role.

The same trend appears when examining the performance of methods in correctly predicting the tightest binder of the three guest sets ***Figure 5***-B. About 61% and 66% of the methods correctly ranked OA-G2 and TEMOA-G4 as the tightest complexes in their respective sets, while CB8-G8 was correctly classified in only about 43% of the cases. In particular, the latter observation is interesting when considering that the binding free energy of G8 to CB8 is 3.5 kcal/mol greater than the second tightest binder (G7), despite the structural similarities between both guests. It is also worth mentioning that SOMD method was the only methodology that correctly ranked the tightest binder of the three separate guest sets, although the prediction that G8 was the highest affinity guest for CB8 did not hold when buffer salt conditions were modeled explicitly.

## Linear corrections fit to prior experimental data can reduce error without improving correlation

Nine of the entries represented in ***Figure 4*** incorporate fits to prior experimental data with the goal of either improving the computationally-predicted affinities or determining the offset necessary to convert relative free energy estimates into absolute binding affinities. It should be noted that a constant offset or multiplicative factor modifying all data points *cannot* alter the $R^2$ statistic besides correcting an inverse correlation, and they can change $\tau$ only if the transformation is such that the ranking of at least two data points is switched, which a single linear transformation with positive slope cannot do. However, since some of the entries fit distinct correction terms for OA and TEMOA guests, correlation statistics for the combined OA/TEMOA set were affected (see, e.g., Supplementary ***Figure 8*** results for SOMD-C and SOMD-D, MovTyp-GE3N and MovTyp-GE3S). We can thus observe the effects of the linear transformations trained on experimental data on both the error and correlation statistics.

The corrections were generally successful in reducing RMSE. Among the top 10 methods scoring the lowest RMSE on the OA/TEMOA set, seven employ a correction. Moreover, when considering multiple submissions of the same technique that differ only in whether a fit to prior experimental data was included, the entry with the lowest RMSE incorporates experimental data in every case. However, the results are less consistent when considering the CB8 guest set. The trend is the same for the SOMD, US-GAFF, and MovTyp submissions that used the KECSA potential, but it is reversed for the majority of the MovTyp submissions employing the GARF energy model (see also Supplementary ***Figure 8***). It should be noted that many of the MovTyp corrections were trained on a dataset that pooled binding measurements of OA, TEMOA, and CB8 guests, so it is possible that the approach failed to generalize when the methodology was affected by a systematic error of opposite sign on the OA/TEMOA and CB8 sets (see ***Figure 3***). The methods that scored best (in terms of lowest RMSE) are US-GAFF-C for OA/TEMOA, and SOMD-D-nobuffer for CB8; excluding

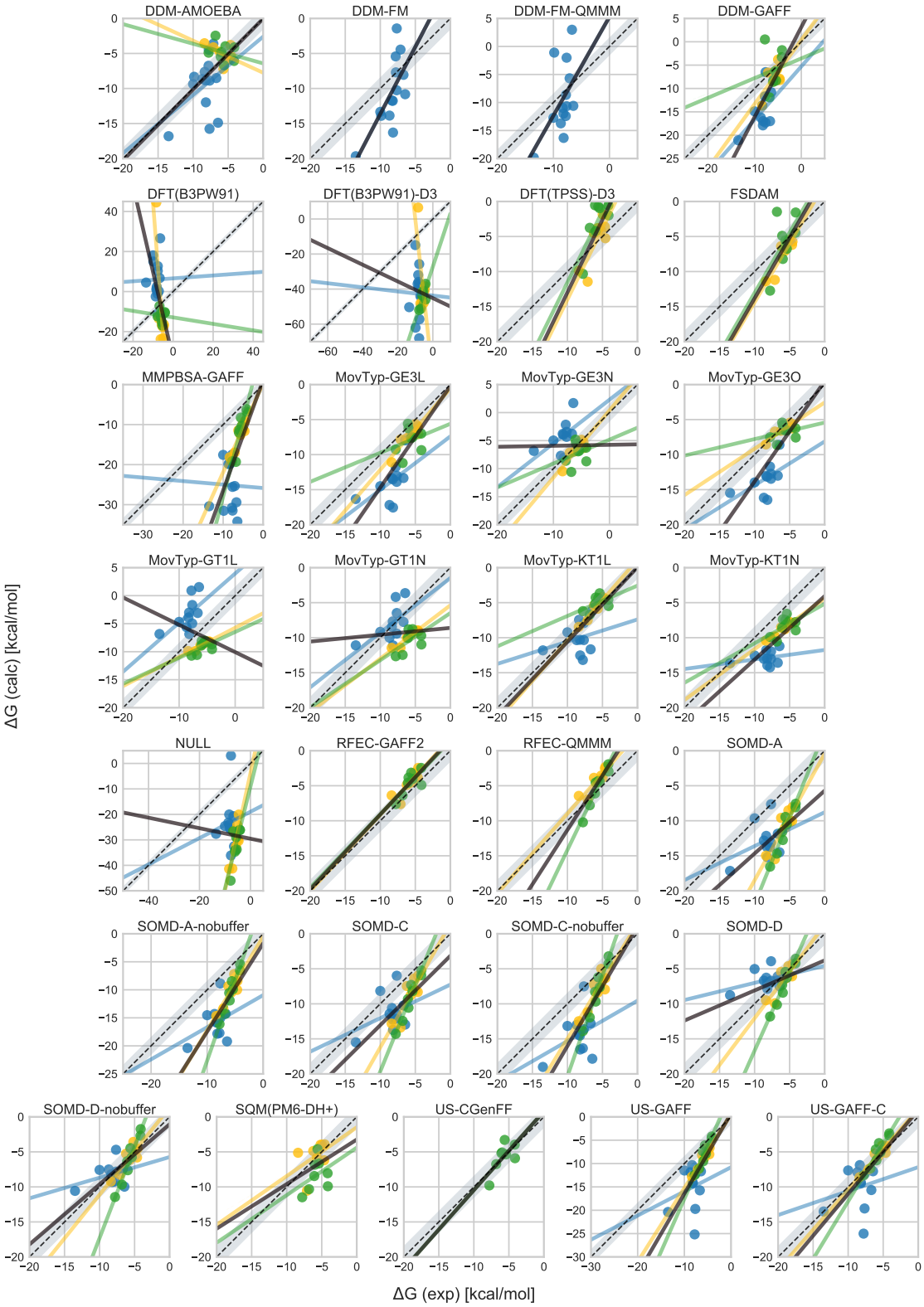



**Figure 3. Free energy correlation plots obtained by the methods on the three host-guest sets.**
Scatter plots showing the experimental measurements of the host-guest binding free energies (horizontal axis) against the methods' predictions on the OA (yellow), TEMOA (green), and CB8 (blue) guest sets with the respective regression lines of the same color. The solid black line is the regression line obtained by using all the data points. The gray shaded area represent the points within 1.5 kcal/mol from the diagonal (dashed black line). Only a representative subset of the movable type calculations results are shown. See Supplementary ***Figure 7*** for the free energy correlation plots of all the movable type predictions.

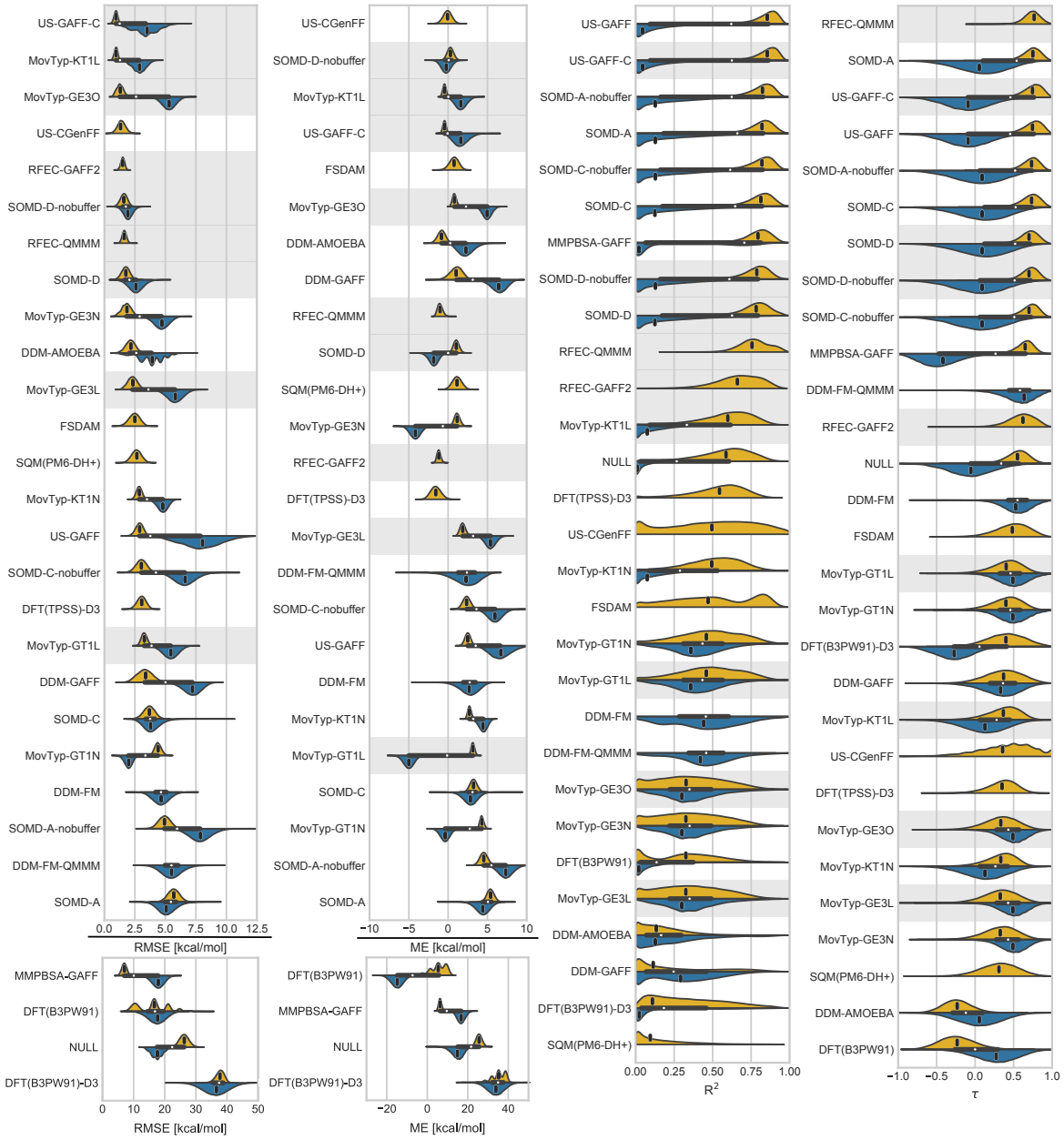



**Figure 4. Bootstrap distribution of the methods performance statistics.**
Bootstrap distributions of root mean square error (RMSE), mean signed error (ME), coefficient of determination ($R^2$) and Kendall rank correlation coefficient ($\tau$). For each methodology and statistic, two distributions are shown for the merged OA/TEMOA set (yellow, pointing upwards) and the CB8 set excluding the bonus challenge compounds (blue, downwards). The black horizontal box between the two distributions of each method shows the median (white circle) and interquartile range (box extremes) of the overall distribution of statistics (i.e., pooling together the OA/TEMOA and CB8 statistic distributions). The short vertical segment in each distribution is the statistic computed using all the data. The distributions of the methods that incorporate previous experimental data into the computational prediction are highlighted in gray. Methodologies are ordered using the statistics computed on the OA/TEMOA set, unless only data for the CB8 set was submitted (e.g., DDM-FM), in which case the CB8 set statistic was used to determine the order. Only a representative subset of the movable type calculations results are shown. See Supplementary ***Figure 8*** for the bootstrap distributions including all the movable type submissions.

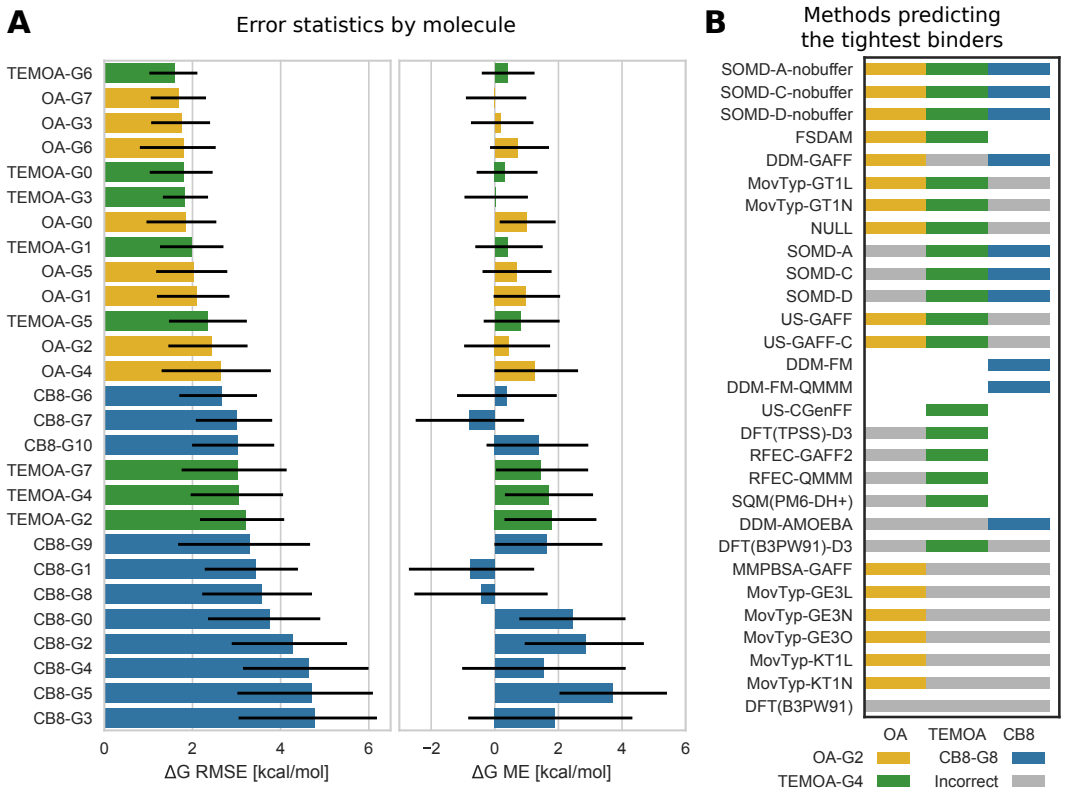



**Figure 5. Free energy error statistics by molecule and tightest binders ranking.**
(**A**) Root mean square error (RMSE) and mean signed error (ME) computed using the ten methodologies with the lowest RMSE on the merged OA/TEMOA and CB8 datasets (excluding bonus challenge compounds) for all guests binding to OA (yellow), TEMOA (green), and CB8 (blue). Error bars represent 95-percentile bootstrap confidence intervals. (**B**) Ranking of the tightest binder of each host-guest dataset for all methods. Methods that correctly predicted OA-G2, TEMOA-G4, and CB8-G8 to be the tightest binders of the OA (yellow), TEMOA (green), and CB8 (blue) guest sets respectively are marked by a colored cell. A gray cell is shown when the method incorrectly predicted the tightest binder, and a white space is left if no submissions were received for that method and guest set. The methods are ordered by the number of correctly ranked tightest binders in the three guest sets.

methods utilizing fits to experimental data, US-CGenFF and MovTyp-GT1N have the lowest RMSE on the OA/TEMOA and CB8 sets, respectively.

On the other hand, integrating prior experimental data did not appreciably impact correlation statistics, and the same methods with or without experimental correction show very similar $R^2$ and $\tau$ bootstrap distributions. It is true that the initial performance of these methods without the experiment-based correction on the separated OA and TEMOA sets was relatively similar, thus leaving a small margin of improvement for this type of correction to reduce the data variance around the regression line and increasing $R^2$. However, comparing the statistics computed pooling together the OA/TEMOA and CB8 predictions, which displayed very different correlation statistics, did not show any significant improvement (data not shown). In fact, $R^2$ for the SOMD-C calculations decreased from 0.47 [0.09,0.78] to 0.18 [0.01,0.48] when incorporating the experimental correction in SOMD-D, despite the expected drop in RMSE, and a similar observation can be made for SOMD-D-nobuffer and the $\tau$ statistic.

### GAFF/AM1-BCC and TIP3P consistently overestimated the host-guest binding affinities

Several entries used GAFF to parameterize the host-guest systems with AM1-BCC charges and TIP3P water molecules (i.e., SOMD, US-GAFF, DDM-GAFF) so it is possible to make relatively general observations about the performance of this model. Firstly, if we ignore the submissions that employ an experiment-based correction, every single method in this group predicted tighter binding than what supported by experiments with both the OA/TEMOA and the CB8 sets. This observation extends to MMPBSA-GAFF as well, which still used GAFF but with RESP charges and the implicit PBSA solvent model, but many of the methodologies that entered the challenge display a similar systematic error (see also ME in ***Figure 5***), although GAFF is the only force field that was independently adopted by multiple groups and used with various classes of techniques.

Secondly, while error statistics vary substantially among GAFF entries, the correlation statistics are quite similar. Most of these are among the best-performing methods for the OA/TEMOA set, with $\tau$ ranging between 0.7–0.8, despite showing poor correlations on the CB8 set. The main exception to this pattern is given by DDM-GAFF, which shows moderate correlations for both datasets. The reason for this is not entirely clear, as the methodology adopted for DDM-GAFF entry is very similar to SOMD-C-nobuffer. Their main difference appears to lie in their treatment of long-range electrostatics, with SOMD using reaction field electrostatics [117] and DDM-GAFF using PME [29], as well as the use of restraints, with SOMD employing a single flat-bottom restraint to keep the guest in the host's cavity and DDM-GAFF restraining the relative orientation of the guest by means of harmonic restraining potentials applied to one distance, two angles, and three torsions.

### Models accounting for polarization did not perform significantly better than point charge models

Several of the entries adopted explicit model of electrostatic polarization through either QM potentials or the AMOEBA force field. Two groups submitted predictions obtained both a point-charge force field that were corrected with the free energy of moving to a QM/MM potential. This is the case of RFEC-QMMM and DDM-FM-QMMM, both of which included only the guest in the QM region using PM6-DH+ and DFT(B3LYP) respectively. In both cases, when compared to the pure MM model, the correlation slightly increased, although this difference was not statistically significant. Notably, RFEC-QMMM and DDM-FM-QMMM scored the top $\tau$ for the OA/TEMOA and CB8 set respectively.

On the other hand, calculations based on the polarizable AMOEBA force field or pure QM potentials were generally outperformed by point charge force-fields and QM/MM models in terms of correlation with experimental data. However, when limiting the comparison to methods that did not include a linear correction fitted on previous experimental data, SQM(PM6-DH+), DFT(TPSS)-D3, and in particular DDM-AMOEBA obtained a relatively low RMSE in spite of their poor correlation with experimental data. It is of interest to note that SQM(PM6-DH+) and DFT(TPSS)-D3 performed similarly. Indeed, the two methodologies were submitted by the same group and differ only by the potential function used to compute the energy of the complex on a set of configurations sampled with MD. SQM(PM6-DH+) scored a slightly lower RMSE and DFT(TPSS)-D3 obtained slightly higher correlation statistics, but the difference is not statistically significant in either case. The data, however, seems to suggest opposite tendencies of the two models in regard to the

bias, with SQM(PM6-DH+) and DFT(TPSS)-D3 overestimating and underestimating the binding affinity of the OA/TEMOA guest set respectively. Similarly, DFT(B3PW91) and DFT(B3PW91)-D3 differ exclusively by the addition of the dispersion correction, which, surprisingly, significantly worsen the error for both guest sets.

## Comparison to null model

The vast majority of the entries statistically outperformed the MMGBSA calculation we used as a null model. Surprisingly, while the null model correlation on the CB8 set was objectively poor ($R^2$ = 0.0 [0.0, 0.5], $\tau$ = -0.1 [-0.6, 0.5]), the $R^2$ and $\tau$ statistics obtained by the MMGBSA null model on the OA/TEMOA set was comparable to more expensive methods and, in fact, surpassed many of the challenge entries (***Table 3***). Nevertheless, the MMGBSA null model was in general poorly accurate in terms of RMSE. We note the difference of our null model with the MMPBSA-GAFF, which generally performed better than MMGBSA on the OA/TEMOA guest set but similarly or slightly worse on the CB8 set. Besides differences in solvent model (i.e., Generalized Born and Poisson-Boltzmann respectively), the former used OPLS3 to rescore a single docked pose, while the second one used GAFF and molecular dynamics to collect samples that were subsequently clustered for the purpose of rescoring.

**Table 4. Performance statistics including the bonus challenge molecules.** Root mean square error (RMSE), mean signed error (ME), coefficient of determination ($R^2$), and Kendall correlation coefficient ($\tau$) obtained by all methods applied to the bonus challenge on the full CB8 set (left super column), including the three bonus molecules. Statistics computed excluding the bonus molecules are reported again here (right super column) for easy comparison. Bootstrap distribution mean and 95-percentile confidence intervals are reported between parentheses and square brackets respectively.

| Method | CB8 dataset (with bonus challenge) | | | | CB8 dataset (no bonus challenge) | | | |
|---|---|---|---|---|---|---|---|---|
| | RMSE | ME | $R^2$ | $\tau$ | RMSE | ME | $R^2$ | $\tau$ |
| DDM-AMOEBA | 3.7 (3.6) [1.9, 5.2] | 1.2 (1.2) [-0.6, 3.1] | 0.1 (0.2) [0.0, 0.7] | 0.1 (0.1) [-0.3, 0.6] | 3.9 (3.8) [1.5, 5.7] | 2.3 (2.3) [0.5, 4.3] | 0.1 (0.3) [0.0, 0.8] | 0.1 (0.1) [-0.5, 0.6] |
| DDM-FM | 4.3 (4.4) [3.1, 5.6] | 2.2 (2.2) [0.0, 4.2] | 0.5 (0.5) [0.1, 0.8] | 0.5 (0.5) [0.2, 0.8] | 4.7 (4.6) [3.4, 6.0] | 2.7 (2.7) [0.2, 4.8] | 0.4 (0.4) [0.0, 0.9] | 0.5 (0.5) [0.1, 0.9] |
| DDM-FM-QMMM | 5.4 (5.5) [3.8, 7.3] | 1.1 (1.1) [-2.0, 3.8] | 0.3 (0.3) [0.0, 0.7] | 0.5 (0.4) [-0.0, 0.8] | 5.5 (5.5) [4.0, 7.3] | 2.3 (2.3) [-1.1, 5.0] | 0.4 (0.5) [0.1, 0.8] | 0.6 (0.6) [0.1, 0.9] |
| DFT(B3PW91) | 17.3 (17.1) [12.1, 21.8] | -14.4 (-14.4) [-19.5, -9.5] | 0.0 (0.1) [0.0, 0.5] | 0.1 (0.1) [-0.4, 0.5] | 17.7 (17.5) [11.9, 22.9] | -14.8 (-14.8) [-20.6, -9.1] | 0.0 (0.1) [0.0, 0.7] | 0.3 (0.3) [-0.3, 0.8] |
| DFT(B3PW91)-D3 | 37.0 (36.8) [29.9, 43.5] | 34.3 (34.3) [27.0, 41.4] | 0.0 (0.1) [0.0, 0.5] | -0.1 (-0.1) [-0.6, 0.2] | 36.6 (36.4) [28.7, 44.2] | 33.9 (33.9) [25.7, 42.0] | 0.0 (0.2) [0.0, 0.8] | -0.3 (-0.3) [-0.7, 0.2] |
| MMPBSA-GAFF | 17.8 (17.7) [14.5, 20.8] | 16.7 (16.7) [13.5, 19.9] | 0.0 (0.1) [0.0, 0.7] | -0.2 (-0.2) [-0.7, 0.2] | 17.9 (17.8) [13.9, 21.5] | 16.7 (16.7) [13.0, 20.6] | 0.0 (0.3) [0.0, 0.8] | -0.4 (-0.4) [-0.9, 0.1] |

Data Source: https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis

**Table 5. Offset statistics of the methods appearing in previous rounds of the SAMPL host-guest binding challenge.**
Root mean square error (RMSE), coefficient of determination ($R^2$), Kendall correlation coefficient ($\tau$), and offset root mean square error ($RMSE_o$) computed by subtracting the mean signed error from the free energy predictions. Absolute and offset statistics for $R^2$ and $\tau$ are identical, and they are thus reported only once. Absolute statistics are identical to those presented before, but, consistently with the format adopted in the SAMPL5 host-guest binding challenge overview paper, the are reported as mean $\pm$ standard deviation of the bootstrap distribution (between parentheses) instead of the 95-percentile confidence interval.

| Method | dataset | RMSE | $RMSE_o$ | $R^2$ | $\tau$ |
|---|---|---|---|---|---|
| DDM-AMOEBA | CB8 | 3.9 (3.8 $\pm$ 1.0) | 3.2 (3.0 $\pm$ 0.7) | 0.1 (0.3 $\pm$ 0.2) | 0.1 (0.1 $\pm$ 0.3) |
| DFT(TPSS)-D3 | OA/TEMOA | 3.1 (3.0 $\pm$ 0.4) | 2.6 (2.5 $\pm$ 0.4) | 0.5 (0.5 $\pm$ 0.2) | 0.3 (0.4 $\pm$ 0.2) |
| MovTyp-GE3N | OA/TEMOA | 1.8 (1.8 $\pm$ 0.4) | 1.4 (1.4 $\pm$ 0.3) | 0.3 (0.4 $\pm$ 0.2) | 0.3 (0.3 $\pm$ 0.2) |
| MovTyp-KT1N | OA/TEMOA | 2.9 (2.9 $\pm$ 0.2) | 0.9 (0.9 $\pm$ 0.2) | 0.5 (0.5 $\pm$ 0.2) | 0.3 (0.3 $\pm$ 0.2) |
| MovTyp-KT1L | OA/TEMOA | 1.0 (0.9 $\pm$ 0.1) | 0.8 (0.8 $\pm$ 0.1) | 0.6 (0.6 $\pm$ 0.2) | 0.4 (0.4 $\pm$ 0.2) |
| SOMD-A-nobuffer | OA/TEMOA | 4.9 (4.9 $\pm$ 0.5) | 1.9 (1.9 $\pm$ 0.3) | 0.8 (0.8 $\pm$ 0.1) | 0.7 (0.7 $\pm$ 0.1) |
| SOMD-C-nobuffer | OA/TEMOA | 3.0 (3.0 $\pm$ 0.4) | 1.9 (1.9 $\pm$ 0.3) | 0.8 (0.8 $\pm$ 0.1) | 0.7 (0.7 $\pm$ 0.1) |
| SOMD-D-nobuffer | OA/TEMOA | 1.6 (1.6 $\pm$ 0.3) | 1.6 (1.5 $\pm$ 0.3) | 0.8 (0.8 $\pm$ 0.1) | 0.7 (0.7 $\pm$ 0.1) |

## Bonus challenge

The platinum atom in CB8-G13 required particular attention during parameterization as this atom is not customarily handled by general small molecule force fields. Even in the case of DFT(B3PW91) and DFT(B3PW91)-D3, the configurations used for the QM calculations were generated by classical molecular dynamics requiring empirical parameters. In general, all the participants to the bonus challenge relied on DFT-level quantum mechanics calculation to address the problem. In MMPBSA-GAFF, DFT(B3PW91), and DFT(B3PW91)-D3, Mulliken charges were generated from DFT(B3LYP), which were subsequently used to determine AM1-BCC charges. A different approach was adopted in DDM-FM-QMMM in which the platinum was substituted by palladium, and the conformations necessary to the force matching parameterization procedure were obtained by MNDO(d) dynamics.

All groups participating to the bonus challenge submitted 1:1 complex predictions also for CB8-G11 and CB8-G12, for which the initial experimental data suggested the possibility of 2:1 complexes (two guests simultaneously bound to one host). This later turned out to be correct only for CB8-G12, and several groups reported to have computationally tested the hypothesis for CB8-G11 with the correct outcome. DDM-AMOEBA was used to estimate affinity of both the 1:1 and 2:1 complexes, but in the end the first one was used in the submission as the two predicted binding free energies differed by only 0.1 kcal/mol. Accordingly, we used the experimental measurement determined for the first binding event to compute the statistics (CB8-G12a in ***Table 1***).

Summary statistics incorporating bonus challenge compounds are reported in ***Table 4***. Although the RMSE generally improves in most cases, it should be noted that this effect varies greatly across the three molecules, and this improvement is mainly due to CB8-G11, whose predictions are regularly much closer to the experimental measurement than the estimates provided for the other two compounds.

## Comparison to previous rounds of the SAMPL host-guest binding challenge

Since previous rounds of the host-guest binding challenge featured identical or similar hosts to those tested in SAMPL6, it is possible to compare earlier results and observe the evolution of methodological performance.

### Accuracy improvements over SAMPL5 for OA/TEMOA were driven by fits to prior experimental data

SAMPL5 featured a set of compounds binding to both OA and TEMOA, which will be referred in the following as the OA/TEMOA-5 set to differentiate it from the combined OA/TEMOA set used in this round of the challenge. In the top row of ***Figure 6***-A, we show median and fitted distributions of the RMSE and $R^2$ statistics taken from the SAMPL5 overview paper [127] together with the results from SAMPL6. OA was used as a test system in SAMPL4 as well, but in this case, only relative free energy predictions were submitted so we cannot draw a direct comparison. Prediction accuracy displays a slight improvement of the median RMSE from the previous round from 3.00 [2.70, 3.60] kcal/mol to 2.76 [1.85, 3.28] kcal/mol (95-percentile bootstrap confidence intervals of the medians not shown in ***Figure 6***-A). However, this change seems to be entirely driven by the methods employing experiment-based fit corrections since removing them results in a median RMSE that is essentially identical to SAMPL5. The data raises the question of whether the field is hitting the accuracy limit of current general force fields.

On the other hand, the median $R^2$ improved with respect to the last round from 0.0 [0.0,0.8] to 0.5 [0.4,0.8]. Even in this case, we observe a slightly lower SAMPL6 median $R^2$ when ignoring methods incorporating experimental data, but this is likely due not to the correction itself but to the fact that the top performing methods were generally submitted with and without correction, thus reducing the number of data points with high $R^2$. Indeed, as already discussed, no positive effect on correlation was evident from the inclusion of a trained linear correction. The improvement is particularly evident when considering only free energy-based methodologies (e.g., alchemical and potential of mean force calculations). It should be pointed out out that the higher median $R^2$ observed in SAMPL6 can, in principle, be explained not only by recent methodological advancements and the composition of the methods entering the challenge, but also by the particular set of assayed guests. While the first explanation is obviously the most desirable, the latter is a confounding factor

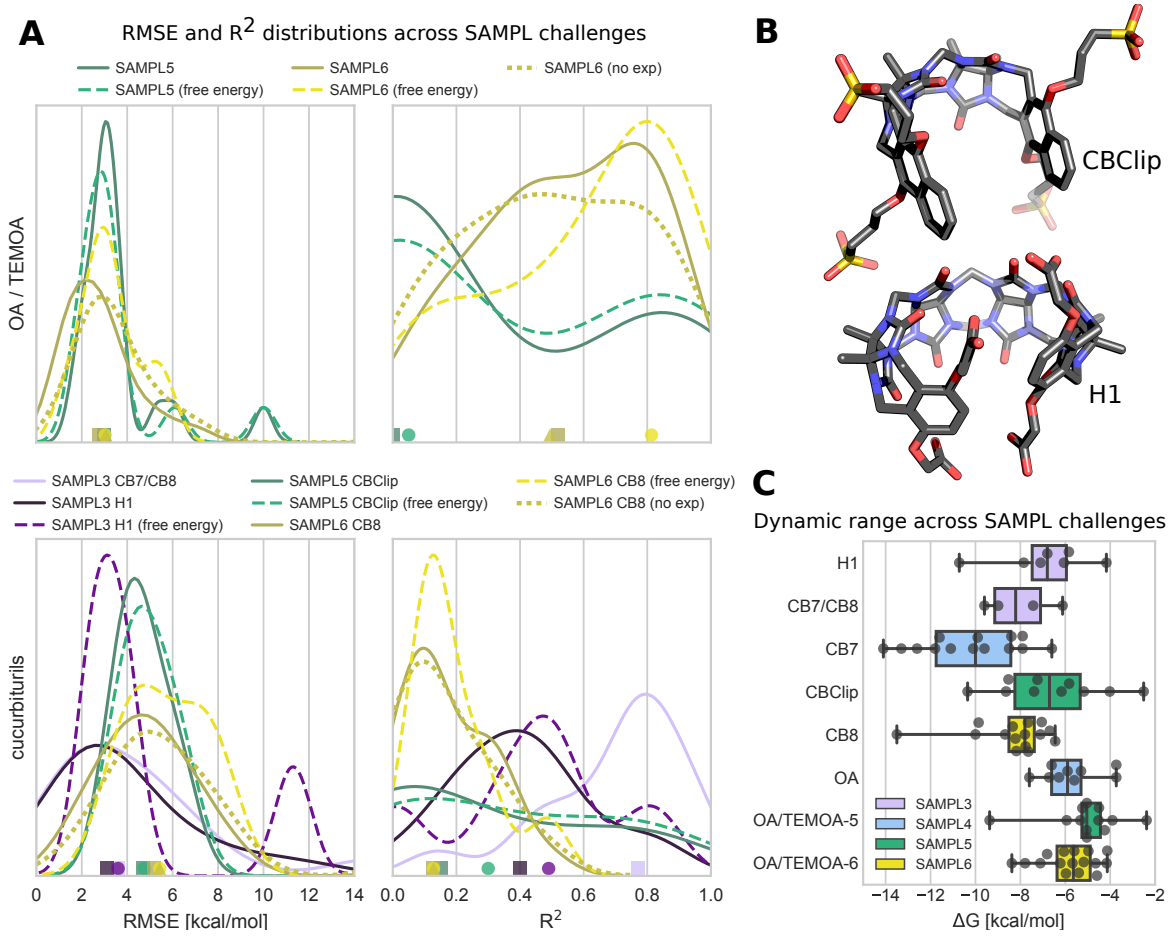



**Figure 6. CB analogues and distribution of RMSE and $R^2$ achieved by methods in SAMPL3 and SAMPL5.** (**A**) Probability distribution fitting of root mean square error (RMSE, left column) and coefficient of determination ($R^2$, right column) achieved by all the methods entering the SAMPL6 (yellow), SAMPL5 (green), and SAMPL3 (purple) challenge. Statistics for SAMPL4 are not shown in the panel because the subject of the challenge was confined to relative binding affinity predictions. The markers on the x-axis indicate the medians of the distributions. Distributions are shown for all the methods entering the challenge (solid line, square marker), excluding the SAMPL6 entries that used previous experimental data (dotted line, triangle marker), or isolating alchemical and potential of mean force methodologies that did not use an experiment-based correction (dashed line, circle marker). The RMSE axis is truncated to 14 kcal/mol, and a few outlier submissions are not shown. The data shows an essentially identical median RMSE and an increased median correlation on the combined OA/TEMOA guest sets (top row) with respect to the previous round of the challenge. The comparison of the results to different sets of guests binding few cucurbit[n]uril and cucurbit[n]uril-like hosts appearing in SAMPL3 and SAMPL5 (bottom row) shows instead a deteriorated performance in the most recent round of the challenge, which is likely explained by the major complexity of the SAMPL6 C8 guest set. (**B**) Three-dimensional structures in stick view of the CBClip (top) and H1 (bottom) hosts featuring in SAMPL5 and SAMPL3 respectively. Carbon atoms are represented in gray, nitrogens in blue, oxygens in red, and sulfur atoms in yellow. Hydrogen atoms are not shown. (**C**) Box plot comparing the range of the binding affinity experimental measurements used as references for the host-guest systems entering the SAMPL3 (purple), SAMPL4 (light blue), SAMPL5 (green), or SAMPL6 (yellow) challenges. The gray data points represent the measurements for the single host-guest entries. The the inter-quartile range and the median represented by the rectangular box were obtained by linear interpolation. The whiskers span the entire dynamic range of reported experiemntal measurements.

when attempting to associate the results of the challenge to the progress of the community.

Since SOMD calculations entered the SAMPL5 challenge as well [19], we can compare directly the same statistics obtained by the method on the two guest sets to form an idea about the relative complexity of the two sets for free energy methods. To this end, we report in ***Table 5*** the uncertainties of the absolute statistics in terms of the mean and standard deviations of the bootstrap distributions instead of their 95-percentile confidence intervals to allow a direct comparison to those published in the SAMPL5 overview paper. The results of the SOMD methods applied to the OA/TEMOA-5 were submitted with a restraint and long-range dispersion correction, similarly to SOMD-C-nobuffer here, and without it, similarly to SOMD-A-nobuffer here. The two methods were referred as SOMD-3 and SOMD-1 respectively in the SAMPL5 overview. In both cases, the calculations used GAFF with AM1-BCC charges and TIP3P water molecules as well as a single flat-bottom restraint. The RMSE obtained by SOMD-C-nobuffer increased with respect to the statistic computed for SOMD-3 on OA/TEMOA-5 from 2.1 (2.1 $\pm$ 0.3) kcal/mol to 3.0 (3.0 $\pm$ 0.4) kcal/mol, where the number outside the parentheses is the statistic computed using all the data, and the numbers between parentheses are the mean and standard deviation of the bootstrap distribution. Incorporating experimental data into the prediction improved the error as SOMD-D-nobuffer obtained a RMSE of 1.6 (1.6 $\pm$ 0.3) kcal/mol. On the other hand, the Kendall correlation coefficient slightly increased on the SAMPL6 dataset from 0.4 (0.4 $\pm$ 0.2) to 0.7 (0.7 $\pm$ 0.4) while $R^2$ remained more or less stationary from the already high value of 0.9 (0.7 $\pm$ 0.2) obtained on OA/TEMOA-5. Very similar observations can be made for SOMD-A-nobuffer and SOMD-1. While the improved $\tau$ correlation does not rule out the possibility of system-dependent effects on $R^2$, it is unlikely for the difference between the median $R^2$ of SAMPL5 and SAMPL6 (amounting to 0.76) to be entirely explained by the different set of guests, and the improvement is likely due, at least in part, to the different methodologies entering the challenge. In particular, SAMPL5 featured several free energy methods that scored near-zero $R^2$ on the OA/TEMOA-5 set, affecting considerably the SAMPL5 median statistic. One of these methods is BEDAM, which used the OPLS-2005 [8, 58] force field and the implicit solvent model AGBNP2 [32], none of which entered the latest round of the challenge. However, the rest of these methods consist of double decoupling calculations carried out either with thermodynamic integration (TI) [60, 112] or HREX and BAR that employed CGenFF and TIP3P, which performed relatively well in SAMPL6 on OA/TEMOA. It should be noted that the TI and HREX/BAR methodologies in SAMPL5 made use of a Boresch-style restraint [18] harmonically constraining one distance, two angles, and three dihedrals. This is similar to the solution adopted in DDM-GAFF in SAMPL6, which also showed a relatively low $R^2$ compared to the other free energy submissions in the same round of the challenge so it is natural to suspect that it may be particularly challenging to treat this class of host-guest systems with this type of restraint in alchemical calculations.

An improvement can also be observed for the movable type method, which was applied to the OA/TEMOA-5 set as well [10] using the KECSA 1 and KECSA 2 potentials. These two submissions, identified with MovTyp-1 and MovTyp-2 respectively in the SAMPL5 overview paper, obtained similar statistics so we will use MovTyp-2 for the comparison. The SAMPL6 entry MovTyp-KT1N, which uses the KECSA energy model too, obtained a comparable RMSE of 2.9 (2.9 $\pm$ 0.2) kcal/mol against the 3.1 (2.9 $\pm$ 1.1) kcal/mol achieved by MovTyp-2 on OA/TEMOA-5, but, even in this case, the error becomes statistically distinguishable once the experimental-based correction is included (i.e., in MovTyp-KT1L), which decreases the RMSE to 1.0 kcal/mol. The correlation statistics generally compare favorably with respect to SAMPL5 with $R^2$ moving from 0.0 (0.3 $\pm$ 0.3) to 0.5 (0.5 $\pm$ 0.2) and $\tau$ going from 0.1 (0.1 $\pm$ 0.3) to 0.3 (0.3 $\pm$ 0.2), although the uncertainties are too large to achieve statistical significance. Moreover, MovTyp-GE3N, which employs the more recently developed GARF energy model, obtained a better RMSE (1.8 (1.8 $\pm$ 0.4) kcal/mol) and comparable correlation statistics to MovTyp-KT1N.

Finally, it seems appropriate to compare the performance of DFT(TPSS)-D3 on OA/TEMOA to DFT/TPSS-c [21] in SAMPL5 and RRHO-551 [74] in SAMPL4 [86]. DFT(TPSS)-D3 an DFT/TPSS-c are very similar in that they both use the DFT-D3 approach to include dispersion correction, but while DFT(TPSS)-D3 generated an ensemble of configurations with MD, DFT/TPSS-c estimated the binding free energy from a single minimized structure. On the other hand, RRHO-551 does use MD for conformational sampling, but it employs DTF-D to correct for dispersion interactions, which was developed earlier than DFT-D3. As already mentioned, SAMPL4

featured a set of 9 OA guests [86], but only relative free energy predictions were submitted so absolute statistics are not available. Thus, in order to facilitate the comparison, we decided to report offset statistics for the subset of the SAMPL6 methods analyzed in this section in the same way they were computed in the previous two rounds of the challenge. The results are given in ***Table 5***. The RMSE of the two models was relatively similar in SAMPL4 and SAMPL5: 5.8 $\pm$ 2.6 kcal/mol for RRHO-551 and 5.3 (5.2 $\pm$ 0.8) kcal/mol for DFT/TPSS-c, where the estimate for RRHO-551 does not include the mean of the statistic bootstrap distribution, which was not reported in the SAMPL4 overview paper. However, the SAMPL6 DFT(TPSS)-D3 calculations attained a lower error (2.6 (2.5 $\pm$ 0.4) kcal/mol) while maintaining a similar coefficient of determination of 0.5 (0.5 $\pm$ 0.2) against the 0.3 (0.4 $\pm$ 0.2) and 0.5 $\pm$ 0.2 of DFT/TPSS-c and RRHO-551 respectively.

### The SAMPL6 CB8 system presents significant challenges to modern methodologies

A different perspective is offered by the history of the binding free energy predictions involving cucurbituril hosts. CB8 and the closely related CB7 appeared previously in SAMPL3 [87] together with an acyclic cucurbit[n]uril-type molecular container referred to as H1 [70]. Moreover, SAMPL5 featured another acyclic CB analogue called CBClip [128]. The 3D structures of the last two hosts are shown in ***Figure 6***-B, while in ***Figure 6***-A (bottom row), we show the distribution of RMSE and $R^2$ computed from the binding free energy predictions submitted for SAMPL3 and SAMPL5 against these four hosts.

In general, both statistics appear to have deteriorated from SAMPL3 to SAMPL5. Even though H1 and CBClip are sufficiently different for system-dependent effects to reasonably dominate the overall performance, the most marked difference appears from the comparison of the SAMPL6 predictions to those submitted for CB7 and CB8 in SAMPL3, which achieved a much greater $R^2$ in spite of the smaller dynamic range of the binding affinity measurements and none of which involved simulation-based methods. The explanation for this inequality is likely to be found in the complexity of the guest sets rather than a methodological regression as SAMPL3 featured only two relatively simple fragment-like binders while the latest round of the challenge included compounds of moderate size and/or complex stereochemistry (e.g., gallamine triethiodate, quinine).

That the CB8 guests in SAMPL6 were particularly challenging is corroborated by the comparison between the performance of DDM-AMOEBA and the results obtained by BAR-560, which also uses the double decoupling method and the AMOEBA polarizable force field, on the CB7 guests in SAMPL4 [14]. In this case as well, only offset statistics are available for comparison as SAMPL4 accepted exclusively relative free energy predictions. DDM-AMOEBA generally performed worse on the CB8 guest set featured in SAMPL6 with $R^2$ decreasing from 0.6 $\pm$ 0.1 to 0.1 (0.3 $\pm$ 0.2) and RMSE increasing from 2.2 $\pm$ 0.4 to 3.2 (3.0 $\pm$ 0.7). While the CB8 guest set featured in SAMPL6 highlights the limits of current free energy methodologies, it also uncovers new learning opportunities that can be exploited to push the boundaries of the domain of applicability of these technologies.

Similarly to the OA/TEMOA guest set, simulation-based free energy methods display a higher median $R^2$ than the global $R^2$ computed from considering all the methods in the challenge, albeit a slightly higher RMSE as well. The pattern is consistent across the three rounds of the challenge, but the distributions of the statistics are too wide to draw statistically significant conclusions without collecting more data.

## Discussion

As in previous years, the SAMPL host-guest binding challenge has provided an opportunity for the computational chemistry community to focus on a common set of systems to assess the state-of-the-art practices and performance of current binding free energy calculation methodologies. The value of the blind challenge does not lie exclusively in the comparison and benchmarking of different methods, but also in its ability to highlight general areas of weakness in the field as a whole on which the community can focus. The latter aspect, in particular, risks to become of secondary importance in retrospective studies. Moreover, the consistent use of octa-acid and cucurbiturils since SAMPL3, which took place in 2011, give us the opportunity to make general observations over a longer time span.

### The variability in difficulty highlights the need to evaluate methodologies on the same systems

Several recurring themes have emerged from this and previous rounds of the challenge. Firstly, even for systems relatively simple as supramolecular host-guests, the performance of free energy methodologies and models can be heavily system-dependent. This is evident not only from the results of the same method applied to different guest sets, but also from the relative performance of the methods against different molecules. For example, most of the predictions employing GAFF obtained among the highest correlation statistics on the OA/TEMOA set while ranking among the lowest positions on the CB8 set. This stresses the importance of using the same set of systems when comparing multiple methodologies, which, without any coordination between groups, is a difficult task to carry out on a medium-large scale given the amount of expertise and resources necessary to perform this type of studies.

A useful dataset should be large enough to have the statistical power to resolve differences in performance, and diverse enough for the distribution of the binding affinity to approximate the distribution of the population of interest and reflect how the method would perform on new data. At the same time, however, correlation statistics tend to increase with the dynamic range spanned by the data, and some methods, such as relative free energy calculations, often impose practical limits to the structural differences between compounds. For example, RFEC-GAFF and RFEC-QMMM submitted predictions only for the OA/TEMOA set, where the similarities between the guests are more prominent. These contrasting requirements, together with practical problems connected to the availability of experimental data and resources, make crafting an appropriate dataset a very challenging task.

### Force field accuracy is a dominant limiting factor for modeling affinity

A second consideration surfacing from previous SAMPL rounds as well is the tendency of classical methods to overestimate the binding affinities. Since the results of the related SAMPLing challenge support the claim that convergence for this class of systems is achievable [51], and considering that the RMSE has not improved significantly across rounds of the challenge, this seem to suggest that an investment of resources into improving the empirical parameters of force fields and solvent models could have a dramatic impact. It should be noted that, while these systems do not put to the test protein parameters, they rely on general force fields that are routinely used in drug and small molecule design.

### Other missing chemical details may also be major limiting factors

However, the problem of missing details of the chemical environment such as salts and alternative protomers cannot be ruled out as a major determinant of predictive accuracy. Explicitly modeling the buffer salt concentrations in the SOMD-C predictions reduced the RMSE from 7.9 to 5.1 kcal/mol for two sets of simulations otherwise identical, and, curiously, it had the opposite effect of increasing the error statistics on the OA/TEMOA set. Despite the sensitivity of the free energy prediction to the presence of ions, a lack of standard best practices emerges from the challenge entries. Many participants decided to add only neutralizing counterions or use a uniform neutralizing charge, and others did not include information about how the buffer was modeled in the submitted method sections, which possibly reflects a generally minor role currently played by this particular aspect of the decision-making process during the modeling step in comparison to other elements (e.g., charges, force field parameters, water model).

### Even at extreme pH, protonation state effects may still contribute

The possible influence of multiple accessible protonation states of the guest compounds on the binding free energy was left unexplored during the challenge, mirroring the widespread tendency in the free energy literature to neglect its effect, and participants largely used the most likely protonation states predicted by Epik that were provided in the input `mol2` and `sdf` files. However, the p$K_a$ free energy penalties estimated by Epik for the second most probable protonation state of the CB8 guests in water at experimental pH (***Table 6***), which is obtained in all cases by the deprotonation of the charged nitrogen atoms as given in ***Figure 1***, suggest that for several guests, and in particular for CB8-G3 and CB8-G11, the deprotonated state is accessible by paying a cost of a few $k_B T$ (where $k_B$ is Boltzmann's constant and $T$ is the absolute temperature), and a change in relative populations between the end states driven by the hydrophobic binding cavity may have a non-negligible effect on the binding affinity. Furthermore, even if the probability of having the carboxyl

**Table 6. p$K_a$ free energy penalties predicted by Epik for the second most likely protonation state of the CB8 guests.**
In all cases, the second most probable protonation state predicted by Epik can be obtained by removing the nitrogen proton of the dominant state. The estimated free energy penalties to access the deprotonated state are reported in kcal/mol and units of $k_BT$, where $k_B$ is the Boltzmann's constant and $T$ is the absolute temperature, taken to be 298 K (i.e., the temperature at experimental conditions). For all the other compounds, including the octa-acid guests, Epik was not able to find a second protonation state within a tolerance of 3 pH units.

| Complex | p$K_a$ penalty | |
|---|---|---|
| | [kcal/mol] | [$k_BT$] |
| CB8-G0 | 2.86 | 4.82 |
| CB8-G1 | 2.67 | 4.50 |
| CB8-G2 | 3.20 | 5.40 |
| CB8-G3 | 1.41 | 2.37 |
| CB8-G11 | 2.76 | 4.65 |
| CB8-G12 | 1.58 | 2.66 |

group of the octa-acid guests protonated at pH 11.7 is usually neglected, a previous study performed for SAMPL5 showed that modeling changes in protonation state populations upon binding resulted in improved predictive performance for a set of OA and TEMOA guests that, similarly to the latest round of the challenge, included several carboxylic acids and was measured at a similar buffer pH [118]. Experimentally, net proton gain or loss during complexation could straightforwardly be assessed for highly soluble host-guest systems via isothermal titration calorimetry (ITC) in buffers with the same pH but different ionization heats for proton loss from solvent [7], a technique that has been used for protein-ligand systems [24, 25, 91, 111]. Similarly to buffer salts, there are few established practices in the community to treat multiple protonation states in free energy calculations [66], and further development and testing of force fields and solvent models with the goal of improving accuracy to experiments should consider these issues as ignoring them during the fitting procedure could push the error caused by missing essential chemicals (e.g., ions, protonation and tautomeric states) to other force field parameters with the risk of decreasing the transferability of the model.

### Linear corrections fit to prior experimental measurements do not improve predictive utility

The experimental-based correction adopted by several groups introduces a new theme in the challenge which pertains to strategies that can be used to inject previous knowledge into molecular simulations. Force field parameters are in principle capable of incorporating experimental data, but an update of the model driven by binding free energy measurements or other ensemble observables is doubtlessly challenging and may involve calculations as expensive as the production calculations so this is normally not routinely viable, although previous studies indicated the validity and feasibility of such an approach [125, 126]. Other schemes that emerged in particular from the field of crystallographic structural refinement avoid modifying the force field parameters and instead add one or more biasing terms to the simulation to replicate experimental measurements that the underlying force field cannot reproduce [16, 123]. The simple linear corrections used independently by various participants in this round of the challenge had a positive impact on the error, but a very small effect in terms of correlation, which is often of central importance in the context of molecular design. However, the simplicity of its application, which is confined entirely to the post-processing step, was such that the participants were able to submit multiple entries with and without the correction.

### Outlook for future SAMPL host-guest challenges

The SAMPL roadmap [77] outlines a proposal for subsequent host-guest challenges for SAMPL7–10. While the future of these blind exercises is uncertain given the absence of a sustainable funding source, we briefly review the likely future design of these host-guest challenges below.

In one line of exploration ([77], section 2.2), SAMPL7 proposes to explore variants of Gibb deep cavity cavitands (related to OA/TEMOA) in which carboxylate substitutent locations are modified, comparing multiple host variants against a set of guests to explore how well affinities and selectivities could be predicted.

SAMPL8 would provide a second iteration of this experiment with novel guests and a trimethylammonium-substituted host variant to assess how algorithmic improvements from the first round could lead to improved performance. SAMPL9–10 would consider the effect of common biologically relevant salts, comparing the effects of NaCl and NaI on various host variants, while SAMPL11 would consider the effects of cosolvents that might compete for the binding site or modulate the strength of the hydrophobic effect.

In another line of exploration ([77], section 2.1), SAMPL7-11 are also proposed to feature cucubituril variants, including methylated forms of CB8, glcoyuracil hexamer, and acyclic forms of CB[n]-type receptors. By comparing the constrained cyclic and less constrained acyclic forms of CB[n] hosts, the accuracy with which participants can model the energetics of receptor flexibility and receptor desolvation can be probed. SAMPL8–9 also plans to feature small molecule guests with pKa values between 3.8–7.4, which brings the possibility that host binding can induce substantial shifts in protonation state.

Finally, recent work by one of the authors has demonstrated how a library of monosubstituted $\beta$-cyclodextrin analogues can be generated via a simple chemical route [59]. This strategy could ultimately lead to the attachment of chemical groups that resemble biopolymer residues, such as amino or nucleic acids, allowing interactions between small druglike molecules and biopolymer-like functional groups to be probed without the multifold challenges that protein-ligand interactions present. While development of this system is still ongoing, it is likely to make an appearance in upcoming SAMPL host-guest challenges.

## Code and data availability

- Input files and setup scripts: https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/
- Analysis scripts: https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis/Scripts/
- Analysis results: https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis/Accuracy
- Participants' submissions: https://github.com/MobleyLab/SAMPL6/tree/master/host_guest/Analysis/Submissions

## Author Contributions

Conceptualization, AR, JDC, DLM; Methodology, AR, JDC, DLM; Software, AR; Formal Analysis, AR, JDC; Investigation, AR, QY, SM, MS, JNM; Resources, JDC, BCG, LI, MWC, MKG, DLM; Data Curation, AR, MWC; Writing-Original Draft, AR, JDC; Writing - Review and Editing, AR, JDC, DLM, MKG, LI, BCG, SM; Visualization, AR, SM; Supervision, JDC, DLM; Project Administration, AR, JDC, DLM; Funding Acquisition, JDC, DLM, MKG, BCG, LI.

## Acknowledgments

AR and JDC acknowledge support from the Sloan Kettering Institute. JDC acknowledges support from NIH grant P30CA008748. JDC, AR, and DLM gratefully acknowledge support from NIH grant R01GM124270 supporting SAMPL blind challenges. AR acknowledges partial support from the Tri-Institutional Program in Computational Biology and Medicine. LI thanks the National Science Foundation for supporting (CHE-1404911) the participation in SAMPL6. DLM appreciates financial support from the National Institutes of Health (1R01GM108889-01), the National Science Foundation (CHE 1352608). AR and JDC are grateful to OpenEye Scientific for providing a free academic software license for use in this work. We thank four anonymous reviewers, whose comments helped us improve the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Disclosures

JDC was a member of the Scientific Advisory Board for Schrödinger, LLC during part of this study. JDC and DLM are current members of the Scientific Advisory Board of OpenEye Scientific Software. The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health, the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Entasis Therapeutics, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, the Molecular Sciences

Software Institute, the Starr Cancer Consortium, Cycle for Survival, a Louis V. Gerstner Young Investigator Award, and the Sloan Kettering Institute. A complete funding history for the Chodera lab can be found at http://choderalab.org/funding.

## References

[1] Abel, R. and Bhat, S. (2017). Free Energy Calculation Guided Virtual Screening of Synthetically Feasible Ligand R-Group and Scaffold Modifications: An Emerging Paradigm for Lead Optimization. In *Annual Reports in Medicinal Chemistry*, volume 50, pages 237–262. Elsevier.

[2] Abel, R., Mondal, S., Masse, C., Greenwood, J., Harriman, G., Ashwell, M. A., Bhat, S., Wester, R., Frye, L., Kapeller, R., and Friesner, R. A. (2017a). Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. *Curr. Opin. Struct. Biol.*, 43:38–44.

[3] Abel, R., Wang, L., Harder, E. D., Berne, B. J., and Friesner, R. A. (2017b). Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc. Chem. Res.*, 50(7):1625–1632.

[4] Abel, R., Wang, L., Mobley, D. L., and Friesner, R. A. (2017c). A Critical Review of Validation, Blind Testing, and Real-World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. *Curr. Top. Med. Chem.*, 17(23).

[5] Aguilar, B., Anandakrishnan, R., Ruscio, J. Z., and Onufriev, A. V. (2010). Statistics and Physical Origins of pK and Ionization State Changes upon Protein-Ligand Binding. *Biophys. J.*, 98(5):872–880.

[6] Aldeghi, M., Heifetz, A., Bodkin, M. J., Knapp, S., and Biggin, P. C. (2017). Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. *J. Am. Chem. Soc.*, 139(2):946–957.

[7] Baker, B. M. and Murphy, K. P. (1996). Evaluation of linked protonation effects in protein binding reactions using isothermal titration calorimetry. *Biophysical journal*, 71(4):2049–2055.

[8] Banks, J. L., Beard, H. S., Cao, Y., Cho, A. E., Damm, W., Farid, R., Felts, A. K., Halgren, T. A., Mainz, D. T., Maple, J. R., et al. (2005). Integrated modeling program, applied chemical theory (impact). *Journal of computational chemistry*, 26(16):1752–1780.

[9] Bannan, C. C., Burley, K. H., Chiu, M., Shirts, M. R., Gilson, M. K., and Mobley, D. L. (2016). Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *J Comput Aided Mol Des*, 30(11):1–18.

[10] Bansal, N., Zheng, Z., Cerutti, D. S., and Merz, K. M. (2017). On the fly estimation of host–guest binding free energies using the movable type method: participation in the sampl5 blind challenge. *Journal of computer-aided molecular design*, 31(1):47–60.

[11] Bansal, N., Zheng, Z., Song, L. F., Pei, J., and Merz Jr, K. M. (2018). The role of the active site flap in streptavidin/biotin complex formation. *Journal of the American Chemical Society*, 140(16):5434–5446.

[12] Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry*, 97(40):10269–10280.

[13] Becke, A. D. (1993). Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of chemical physics*, 98(7):5648–5652.

[14] Bell, D. R., Qi, R., Jing, Z., Xiang, J. Y., Mejias, C., Schnieders, M. J., Ponder, J. W., and Ren, P. (2016). Calculating binding free energies of host–guest systems using the amoeba polarizable force field. *Physical Chemistry Chemical Physics*, 18(44):30261–30269.

[15] Bennett, C. H. (1976). Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268.

[16] Best, R. B. and Vendruscolo, M. (2004). Determination of protein structures consistent with nmr order parameters. *Journal of the American Chemical Society*, 126(26):8090–8091.

[17] Bhakat, S. and Söderhjelm, P. (2017). Resolving the problem of trapped water in binding cavities: Prediction of host-guest binding free energies in the SAMPL5 challenge by funnel metadynamics. *J Comput Aided Mol Des*, 31(1):119–132.

[18] Boresch, S., Tettinger, F., Leitgeb, M., and Karplus, M. (2003). Absolute binding free energies: a quantitative approach for their calculation. *The Journal of Physical Chemistry B*, 107(35):9535–9551.

[19] Bosisio, S., Mey, A. S. J. S., and Michel, J. (2017). Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *J Comput Aided Mol Des*, 31(1):61–70.

[20] Boyce, S. E., Tellinghuisen, J., and Chodera, J. D. (2015). Avoiding accuracy-limiting pitfalls in the study of protein-ligand interactions with isothermal titration calorimetry. *bioRxiv*, page 023796.

[21] Caldararu, O., Olsson, M. A., Riplinger, C., Neese, F., and Ryde, U. (2017). Binding free energies in the SAMPL5 octa-acid host–guest challenge calculated with DFT-D3 and CCSD(T). *J Comput Aided Mol Des*, 31(1):87–106.

[22] Cao, L. and Isaacs, L. (2014). Absolute and relative binding affinity of cucurbit[7]uril towards a series of cationic guests. *Supramol. Chem.*, 26(3-4):251–258.

[23] Cournia, Z., Allen, B., and Sherman, W. (2017). Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.*, 57(12):2911–2937.

[24] Czodrowski, P. (2012). Who cares for the protons? *Bioorganic & medicinal chemistry*, 20(18):5453–5460.

[25] Czodrowski, P., Sotriffer, C. A., and Klebe, G. (2007). Protonation changes upon ligand binding to trypsin and thrombin: structural interpretation based on pka calculations and itc experiments. *Journal molecular biology*, 367(5):1347–1356.

[Drug Design Data Resource] Drug Design Data Resource. Sampl.

[27] Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., et al. (2017). Openmm 7: rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659.

[28] Ercolessi, F. and Adams, J. B. (1994). Interatomic potentials from first-principles calculations: the force-matching method. *EPL (Europhysics Letters)*, 26(8):583.

[29] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh ewald method. *The Journal of chemical physics*, 103(19):8577–8593.

[30] Ewell, J., Gibb, B. C., and Rick, S. W. (2008). Water inside a hydrophobic cavitand molecule. *The Journal of Physical Chemistry B*, 112(33):10272–10279.

[31] Freeman, W., Mock, W., and Shih, N. (1981). Cucurbituril. *Journal of the American Chemical Society*, 103(24):7367–7368.

[32] Gallicchio, E., Paris, K., and Levy, R. M. (2009). The agbnp2 implicit solvation model. *Journal of chemical theory and computation*, 5(9):2544–2564.

[33] Gan, H., Benjamin, C. J., and Gibb, B. C. (2011). Nonmonotonic assembly of a deep-cavity cavitand. *Journal of the American Chemical Society*, 133(13):4770–4773.

[34] Geballe, M. T. and Guthrie, J. P. (2012). The SAMPL3 blind prediction challenge: Transfer energy overview. *J Comput Aided Mol Des*, 26(5):489–496.

[35] Geballe, M. T., Skillman, A. G., Nicholls, A., Guthrie, J. P., and Taylor, P. J. (2010). The SAMPL2 blind prediction challenge: Introduction and overview. *J Comput Aided Mol Des*, 24(4):259–279.

[36] Gibb, C. L. and Gibb, B. C. (2004). Well-defined, organic nanoenvironments in water: the hydrophobic effect drives a capsular assembly. *Journal of the American Chemical Society*, 126(37):11408–11409.

[37] Gibb, C. L. and Gibb, B. C. (2011). Anion binding to hydrophobic concavity is central to the salting-in effects of hofmeister chaotropes. *Journal of the American Chemical Society*, 133(19):7344–7347.

[38] Gibb, C. L. D. and Gibb, B. C. (2013). Binding of cyclic carboxylates to octa-acid deep-cavity cavitand. *J Comput Aided Mol Des*, 28(4):319–325.

[39] Gilson, M. K., Given, J. A., Bush, B. L., and McCammon, J. A. (1997). The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical journal*, 72(3):1047–1069.

[40] Graves, A. P., Shivakumar, D. M., Boyce, S. E., Jacobson, M. P., Case, D. A., and Shoichet, B. K. (2008). Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *Journal of molecular biology*, 377(3):914–934.

[41] Greenwood, J. R., Calkins, D., Sullivan, A. P., and Shelley, J. C. (2010). Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of computer-aided molecular design*, 24(6-7):591–604.

[42] Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. (2010). A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15):154104.

[43] Guthrie, J. P. (2009). A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J Phys Chem B*, 113(14):4501–4507.

[44] Guthrie, J. P. (2014). SAMPL4, a blind challenge for computational solvation free energies: The compounds considered. *J Comput Aided Mol Des*, 28(3):151–168.

[45] Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., Wang, L., Lupyan, D., Dahlgren, M. K., Knight, J. L., et al. (2015). Opls3: a force field providing broad coverage of drug-like small molecules and proteins. *Journal of chemical theory and computation*, 12(1):281–296.

[46] Hawkins, P. C., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. (2010). Conformer generation with omega: algorithm and validation using high quality structures from the protein databank and cambridge structural database. *Journal of chemical information and modeling*, 50(4):572–584.

[47] Henriksen, N. M., Fenley, A. T., and Gilson, M. K. (2015). Computational Calorimetry: High-Precision Calculation of Host–Guest Binding Thermodynamics. *J. Chem. Theory Comput.*, 11(9):4377–4394.

[48] Hillyer, M. B., Gibb, C. L., Sokkalingam, P., Jordan, J. H., Ioup, S. E., and Gibb, B. C. (2016). Synthesis of water-soluble deep-cavity cavitands. *Organic letters*, 18(16):4048–4051.

[49] Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., and Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *The Journal of chemical physics*, 120(20):9665–9678.

[50] Hsiao, Y.-W. and Söderhjelm, P. (2014). Prediction of sampl4 host–guest binding affinities using funnel metadynamics. *Journal of computer-aided molecular design*, 28(4):443–454.

[51] Isik, M., Rizzi, A., Mobley, D. L., and Shirts, M. (2018). MobleyLab/SAMPL6: Version 1.12: Update preliminary SAMPLing analysis.

[52] Jacobson, M. P., Friesner, R. A., Xiang, Z., and Honig, B. (2002). On the role of the crystal environment in determining protein side-chain conformations. *Journal of molecular biology*, 320(3):597–608.

[53] Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., and Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367.

[54] Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. (2000). Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *Journal of computational chemistry*, 21(2):132–146.

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

[56] Jordan, I. K., Kondrashov, F. A., Adzhubei, I. A., Wolf, Y. I., Koonin, E. V., Kondrashov, A. S., and Sunyaev, S. (2005). A universal trend of amino acid gain and loss in protein evolution. *Nature*, 433(7026):633–638.

[57] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935.

[58] Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001). Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28):6474–6487.

[59] Kellett, K., Duggan, B. M., and Gilson, M. K. (2018). Facile synthesis of a diverse library of mono-3-substituted $\beta$-cyclodextrin analogues. *chemRxiv*.

[60] Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313.

[61] Klamt, A. (1995). Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *The Journal of Physical Chemistry*, 99(7):2224–2235.

[62] Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.*, 19(2):120–127.

[63] Kohlhoff, K. J., Shukla, D., Lawrenz, M., Bowman, G. R., Konerding, D. E., Belov, D., Altman, R. B., and Pande, V. S. (2014). Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.*, 6(1):15–21.

[64] Korth, M. (2010). Third-generation hydrogen-bonding corrections for semiempirical qm methods and force fields. *Journal of Chemical Theory and Computation*, 6(12):3808–3816.

[65] Kuhn, B., Tichý, M., Wang, L., Robinson, S., Martin, R. E., Kuglstatter, A., Benz, J., Giroud, M., Schirmeister, T., Abel, R., Diederich, F., and Hert, J. (2017). Prospective Evaluation of Free Energy Calculations for the Prioritization of Cathepsin L Inhibitors. *J. Med. Chem.*, 60(6):2485–2497.

[66] Lee, J., Miller, B. T., and Brooks, B. R. (2016). Computational scheme for ph-dependent binding free energy calculation with explicit solvent. *Protein Science*, 25(1):231–243.

[67] Li, J., Abel, R., Zhu, K., Cao, Y., Zhao, S., and Friesner, R. A. (2011). The vsgb 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 79(10):2794–2812.

[68] Liu, P., Kim, B., Friesner, R. A., and Berne, B. (2005a). Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13749–13754.

[69] Liu, S., Ruspic, C., Mukhopadhyay, P., Chakrabarti, S., Zavalij, P. Y., and Isaacs, L. (2005b). The cucurbit[n]uril family: prime components for self-sorting systems. *Journal of the American Chemical Society*, 127(45):15959–15967.

[70] Ma, D., Zavalij, P. Y., and Isaacs, L. (2010). Acyclic cucurbit[n]uril congeners are high affinity hosts. *The Journal of organic chemistry*, 75(14):4786–4795.

[71] Marsili, S., Signorini, G. F., Chelli, R., Marchi, M., and Procacci, P. (2010). Orac: A molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level. *Journal of computational chemistry*, 31(5):1106–1116.

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

[73] McGann, M. (2012). Fred and hybrid docking performance on standardized datasets. *Journal of computer-aided molecular design*, 26(8):897–906.

[74] Mikulskis, P., Cioloboc, D., Andrejić, M., Khare, S., Brorsson, J., Genheden, S., Mata, R. A., Söderhjelm, P., and Ryde, U. (2014). Free-energy perturbation and quantum mechanical study of SAMPL4 octa-acid host–guest binding energies. *J Comput Aided Mol Des*, 28(4):375–400.

[75] Mobley, D. L., Chodera, J. D., and Dill, K. A. (2006). On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *The Journal of chemical physics*, 125(8):084902.

[76] Mobley, D. L., Chodera, J. D., Isaacs, L., and Gibb, B. C. (2016a). Advancing predictive modeling through focused development of model systems to drive new modeling innovations. *UC Irvine: Department of Pharmaceutical Sciences, UCI*.

[77] Mobley, D. L., Chodera, J. D., Isaacs, L., and Gibb, B. C. (2016b). Advancing predictive modeling through focused development of model systems to drive new modeling innovations. Retrieved from https://escholarship.org/uc/item/7cf8c6cr.

[78] Mobley, D. L. and Gilson, M. K. (2016). Predicting binding free energies: Frontiers and benchmarks.

[79] Mobley, D. L. and Gilson, M. K. (2017). Predicting binding free energies: Frontiers and benchmarks. *Annual review of biophysics*, 46:531–558.

[80] Mobley, D. L., Heinzelmann, G., Henriksen, N. M., and Gilson, M. K. (2017). Predicting binding free energies: Frontiers and benchmarks (a perpetual review). *UC Irvine: Department of Pharmaceutical Sciences, UCI*.

[81] Mobley, D. L., Liu, S., Lim, N. M., Wymer, K. L., Perryman, A. L., Forli, S., Deng, N., Su, J., Branson, K., and Olson, A. J. (2014a). Blind prediction of HIV integrase binding from the SAMPL4 challenge. *J Comput Aided Mol Des*, 28(4):327–345.

[82] Mobley, D. L., Wymer, K. L., Lim, N. M., and Guthrie, J. P. (2014b). Blind prediction of solvation free energies from the SAMPL4 challenge. *J Comput Aided Mol Des*, 28(3):135–150.

[83] Mock, W. and Shih, N. (1983). Host-guest binding capacity of cucurbituril. *The Journal of Organic Chemistry*, 48(20):3618–3619.

[84] Moghaddam, S., Inoue, Y., and Gilson, M. K. (2009). Host-Guest Complexes with Protein-Ligand-like Affinities: Computational Analysis and Design. *J. Am. Chem. Soc.*, 131(11):4012–4021.

[85] Moghaddam, S., Yang, C., Rekharsky, M., Ko, Y. H., Kim, K., Inoue, Y., and Gilson, M. K. (2011). New Ultrahigh Affinity Host-Guest Complexes of Cucurbit[7]uril with Bicyclo[2.2.2]octane and Adamantane Guests: Thermodynamic Analysis and Evaluation of M2 Affinity Calculations. *J. Am. Chem. Soc.*, 133(10):3570–3581.

[86] Muddana, H. S., Fenley, A. T., Mobley, D. L., and Gilson, M. K. (2014a). The SAMPL4 host–guest blind prediction challenge: An overview. *J Comput Aided Mol Des*, 28(4):305–317.

[87] Muddana, H. S. and Gilson, M. K. (2012). Prediction of SAMPL3 host–guest binding affinities: Evaluating the accuracy of generalized force-fields. *J Comput Aided Mol Des*, 26(5):517–525.

[88] Muddana, H. S., Varnado, C. D., Bielawski, C. W., Urbach, A. R., Isaacs, L., Geballe, M. T., and Gilson, M. K. (2012). Blind prediction of host–guest binding affinities: A new SAMPL3 challenge. *J Comput Aided Mol Des*, 26(5):475–487.

[89] Muddana, H. S., Yin, J., Sapra, N. V., Fenley, A. T., and Gilson, M. K. (2014b). Blind prediction of sampl4 cucurbit[7]uril binding affinities with the mining minima method. *Journal of computer-aided molecular design*, 28(4):463–474.

[90] Murkli, S., McNeill, J. N., and Isaacs, L. (2018). Cucurbit[8]uril guest complexes: Blinded dataset for the SAMPL6 challenge. *Supramolecular Chemistry*, accepted.

[91] Neeb, M., Czodrowski, P., Heine, A., Barandun, L. J., Hohn, C., Diederich, Fran c., and Klebe, G. (2014). Chasing protons: How isothermal titration calorimetry, mutagenesis, and p$k_a$ calculations trace the locus of charge in ligand binding to a tRNA-binding enzyme. *Journal medicinal chemistry*, 57(13):5554–5565.

[92] Nerattini, F., Chelli, R., and Procacci, P. (2016). Ii. dissociation free energies in drug–receptor systems via nonequilibrium alchemical simulations: application to the fk506-related immunophilin ligands. *Physical Chemistry Chemical Physics*, 18(22):15005–15018.

[93] Nicholls, A., Mobley, D. L., Guthrie, J. P., Chodera, J. D., Bayly, C. I., Cooper, M. D., and Pande, V. S. (2008). Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.*, 51(4):769–779.

[94] Ong, W. and Kaifer, A. E. (2004). Salt effects on the apparent stability of the cucurbit [7] uril- methyl viologen inclusion complex. *The Journal of organic chemistry*, 69(4):1383–1385.

[95] Pal, R. K., Haider, K., Kaur, D., Flynn, W., Xia, J., Levy, R. M., Taran, T., Wickstrom, L., Kurtzman, T., and Gallicchio, E. (2017). A combined treatment of hydration and dynamical effects for the modeling of host-guest binding thermodynamics: The SAMPL5 blinded challenge. *J. Comput. Aided Mol. Des.*, 31(1):29–44.

[96] Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., Haque, I., Mobley, D. L., Lambrecht, D. S., DiStasio Jr, R. A., et al. (2010). Current status of the amoeba polarizable force field. *The journal of physical chemistry B*, 114(8):2549–2564.

[97] Procacci, P. (2016). I. dissociation free energies of drug–receptor systems via non-equilibrium alchemical simulations: a theoretical framework. *Physical Chemistry Chemical Physics*, 18(22):14991–15004.

[98] Rekharsky, M. V., Ko, Y. H., Selvapalam, N., Kim, K., and Inoue, Y. (2007a). Complexation thermodynamics of cucurbit[6]uril with aliphatic alcohols, amines, and diamines. *Supramolecular Chemistry*, 19(1-2):39–46.

[99] Rekharsky, M. V., Mori, T., Yang, C., Ko, Y. H., Selvapalam, N., Kim, H., Sobransingh, D., Kaifer, A. E., Liu, S., Isaacs, L., Chen, W., Moghaddam, S., Gilson, M. K., Kim, K., and Inoue, Y. (2007b). A synthetic host-guest system achieves avidin-biotin affinity by overcoming enthalpy–entropy compensation. *PNAS*, 104(52):20737–20742.

[100] Řezáč, J., Fanfrlík, J., Salahub, D., and Hobza, P. (2009). Semiempirical quantum chemical pm6 method augmented by dispersion and h-bonding correction terms reliably describes various types of noncovalent complexes. *Journal of Chemical Theory and Computation*, 5(7):1749–1760.

[101] Rogers, K. E., Ortiz-Sánchez, J. M., Baron, R., Fajer, M., de Oliveira, C. A. F., and McCammon, J. A. (2012). On the role of dewetting transitions in host–guest binding free energy calculations. *Journal of chemical theory and computation*, 9(1):46–53.

[102] Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. (2007). Epik: a software program for pk a prediction and protonation state generation for drug-like molecules. *Journal of computer-aided molecular design*, 21(12):681–691.

[103] Shirts, M. R., Bair, E., Hooker, G., and Pande, V. S. (2003). Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Physical review letters*, 91(14):140601.

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

[105] Shirts, M. R., Mobley, D. L., and Brown, S. P. (2010). Free energy calculations in structure-based drug design. *Drug design: structure-and ligand-based approaches*, pages 61–86.

[106] Sitkoff, D., Sharp, K. A., and Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *The Journal of Physical Chemistry*, 98(7):1978–1988.

[107] Skillman, A. G. (2012). SAMPL3: Blinded prediction of host–guest binding affinities, hydration free energies, and trypsin inhibitors. *J Comput Aided Mol Des*, 26(5):473–474.

[108] Skillman, A. G., Geballe, M. T., and Nicholls, A. (2010). SAMPL2 challenge: Prediction of solvation energies and tautomer ratios. *J Comput Aided Mol Des*, 24(4):257–258.

[109] Sokkalingam, P., Shraberg, J., Rick, S. W., and Gibb, B. C. (2015). Binding hydrated anions with hydrophobic pockets. *Journal of the American Chemical Society*, 138(1):48–51.

[110] Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., and Case, D. A. (1998). Continuum solvent studies of the stability of dna, rna, and phosphoramidate- dna helices. *Journal of the American Chemical Society*, 120(37):9401–9409.

[111] Steuber, H., Czodrowski, P., Sotriffer, C. A., and Klebe, G. (2007). Tracing changes in protonation: a prerequisite to factorize thermodynamic data of inhibitor binding to aldose reductase. *Journal molecular biology*, 373(5):1305–1320.

[112] Straatsma, T. and McCammon, J. (1991). Multiconfiguration thermodynamic integration. *The Journal of chemical physics*, 95(2):1175–1188.

[113] Sugita, Y., Kitao, A., and Okamoto, Y. (2000). Multidimensional replica-exchange method for free-energy calculations. *The Journal of Chemical Physics*, 113(15):6042–6051.

[114] Sullivan, M. R., Sokkalingam, P., Nguyen, T., Donahue, J. P., and Gibb, B. C. (2017). Binding of carboxylate and trimethylammonium salts to octa-acid and TEMOA deep-cavity cavitands. *J Comput Aided Mol Des*, 31(1):1–8.

[115] Sultan, M. M., Denny, R. A., Unwalla, R., Lovering, F., and Pande, V. S. (2017). Millisecond dynamics of BTK reveal kinome-wide conformational plasticity within the apo kinase domain. *Sci. Rep.*, 7(1).

[116] Tao, J., Perdew, J. P., Staroverov, V. N., and Scuseria, G. E. (2003). Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Physical Review Letters*, 91(14):146401.

[117] Tironi, I. G., Sperb, R., Smith, P. E., and van Gunsteren, W. F. (1995). A generalized reaction field method for molecular dynamics simulations. *The Journal of chemical physics*, 102(13):5451–5459.

[118] Tofoleanu, F., Lee, J., Pickard IV, F. C., König, G., Huang, J., Baek, M., Seok, C., and Brooks, B. R. (2017). Absolute binding free energies for octa-acids and guests in sampl5. *Journal of computer-aided molecular design*, 31(1):107–118.

[119] Torrie, G. M. and Valleau, J. P. (1974). Monte carlo free energy estimates using non-boltzmann sampling: Application to the sub-critical lennard-jones fluid. *Chemical Physics Letters*, 28(4):578–581.

[120] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., et al. (2010). Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of computational chemistry*, 31(4):671–690.

[121] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174.

[122] Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M. K., Greenwood, J., Romero, D. L., Masse, C., Knight, J. L., Steinbrecher, T., Beuming, T., Damm, W., Harder, E., Sherman, W., Brewer, M., Wester, R., Murcko, M., Frye, L., Farid, R., Lin, T., Mobley, D. L., Jorgensen, W. L., Berne, B. J., Friesner, R. A., and Abel, R. (2015). Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.*, 137(7):2695–2703.

[123] White, A. D. and Voth, G. A. (2014). Efficient and minimal method to bias molecular simulations with experimental data. *Journal of chemical theory and computation*, 10(8):3023–3030.

[Woods et al.] Woods, C. J., Mey, A. S., Calabro, G., and Julien, M. Sire molecular simulation framework.

[125] Yin, J., Fenley, A. T., Henriksen, N. M., and Gilson, M. K. (2015). Toward Improved Force-Field Accuracy through Sensitivity Analysis of Host-Guest Binding Thermodynamics. *J. Phys. Chem. B*, 119(32):10145–10155.

[126] Yin, J., Henriksen, N. M., Muddana, H. S., and Gilson, M. K. (2018). Bind3p: Optimization of a water model based on host-guest binding data. *Journal of chemical theory and computation*.

[127] Yin, J., Henriksen, N. M., Slochower, D. R., Shirts, M. R., Chiu, M. W., Mobley, D. L., and Gilson, M. K. (2017). Overview of the SAMPL5 host–guest challenge: Are we doing better? *J Comput Aided Mol Des*, 31(1):1–19.

[128] Zhang, B. and Isaacs, L. (2014). Acyclic cucurbit[n]uril-type molecular containers: influence of aromatic walls on their function as solubilizing excipients for insoluble drugs. *Journal of medicinal chemistry*, 57(22):9554–9563.

[129] Zheng, Z. and Merz Jr, K. M. (2013). Development of the knowledge-based and empirical combined scoring algorithm (kecsa) to score protein–ligand interactions. *Journal of chemical information and modeling*, 53(5):1073–1083.

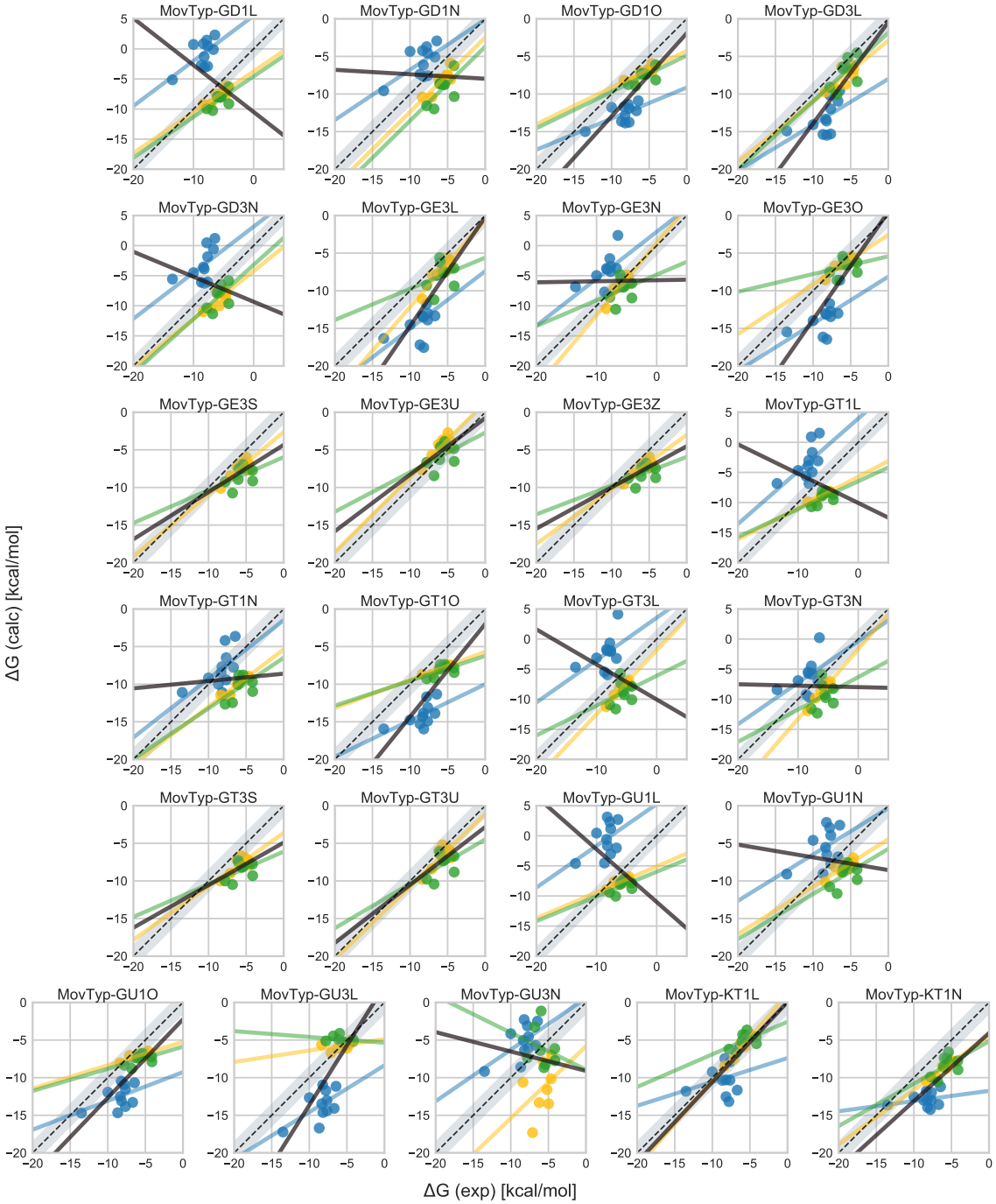
[130] Zheng, Z., Ucisik, M. N., and Merz, K. M. (2013). The movable type method applied to protein–ligand binding. *Journal of chemical theory and computation*, 9(12):5526–5538.

[131] Zheng, Z., Wang, T., Li, P., and Merz Jr, K. M. (2015). Kecsa-movable type implicit solvation model (kmtism). *Journal of chemical theory and computation*, 11(2):667–682.

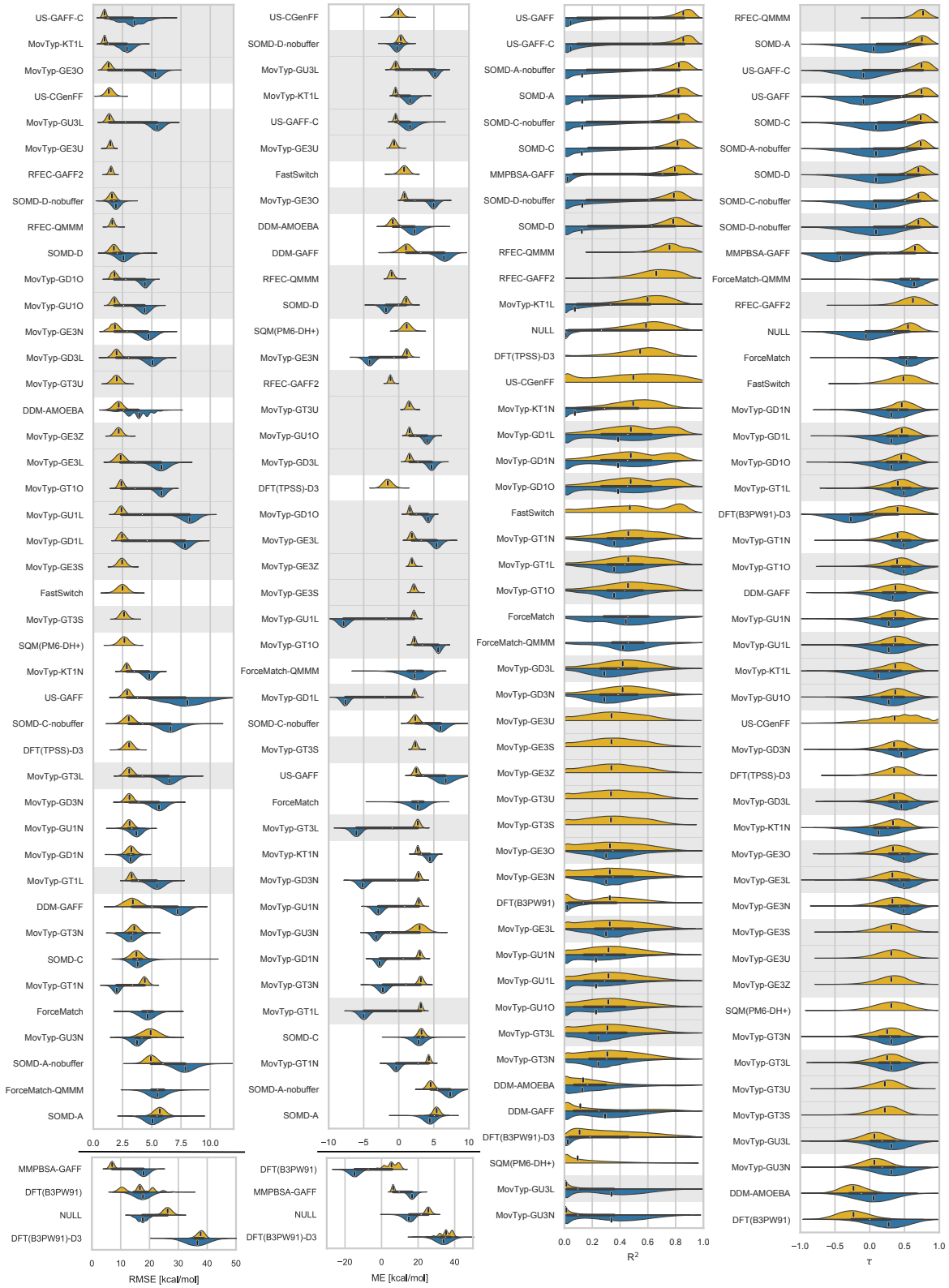
## List of abbreviations

- AM1-BCC: Austin model 1 bond charge correction [54, 55]
- AMOEBA: atomic multipole optimized energetics for biomolecular simulation [96]
- B3LYP: Becke 3-parameter Lee-Yang-Parr exchange-correlation functional [13]
- B3PW91: Becke 3-parameter Perdew-Wang 91 exchange-correlation functional [13]
- CGenFF: CHARMM generalized force field [120]
- COSMO-RS: conductor-like screening model for real solvents [61]
- DDM: double decoupling method [39]
- DFT-D3: density functional theory with the D3 dispersion corrections [42]
- FM: Force Matching [28]
- FSDAM: Fast switching double annihilation method [92, 97]
- GAFF: generalized AMBER force field [121]
- HREX: Hamiltonian replica exchange [113]
- KECSA: knowledge-based and empirical combined scoring algorithm [129]
- KMTISM: KECSA-Movable Type Implicit Solvation Model [131]
- MD: molecular dynamics
- MMPBSA: molecular mechanics Poisson Boltzmann/solvent accessible surface area [110]
- MovTyp Movable Type method [130]
- OPLS3: optimized potential for liquid simulations [45]
- PBSA: Poisson-Boltzmann surface area [106]
- PM6-DH+: PM6 semiempirical method with dispersion and hydrogen bonding corrections [64, 100]
- RESP: restrained electrostatic potential [12]
- REST: replica exchange with solute torsional tempering [68, 71]
- RFEC: relative free energy calculation
- QM/MM: mixed quantum mechanics and molecular mechanics
- SOMD: double annihilation or decoupling method performed with Sire/OpenMM6.3 software [27, Woods et al.]
- SQM: semi-empirical quantum mechanics
- TIP3P: transferable interaction potential three-point [57]
- TPSS: Tao, Perdew, Staroverov, and Scuseria exchange functional [116]
- US: umbrella sampling [119]
- VSGB2.1: VSGB2.0 solvation model refit to OPLS2.1/3/3e [67]

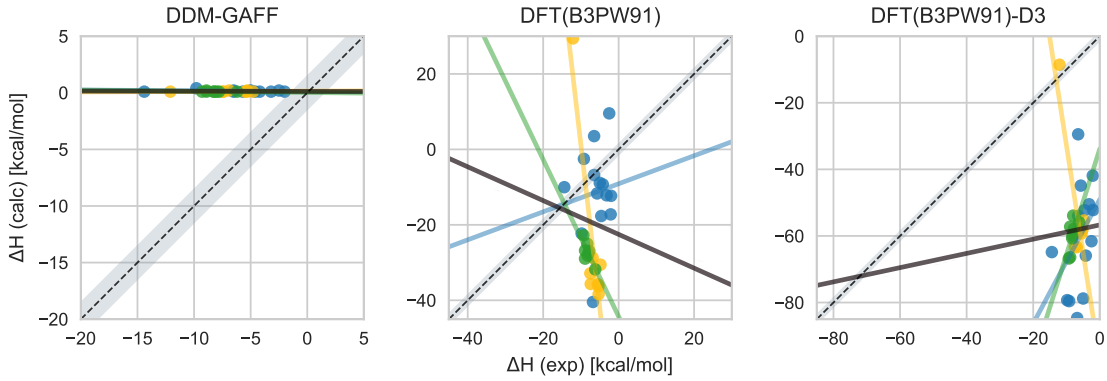
## Supplementary figures




**Appendix 0 Figure 7. Free energy correlation plots of all movable type submissions.**
The four-letter suffix of each movable type submission is to be interpreted as following: first letter indicates the force field (G: GARF; K: KECSA), the third letter input structures (D: final frame of MD sampling; E: ensemble of structures from MD sampling; T: lowest energy structure during movable type scoring; U: lowest energy structure obtained during the sampling in US-GAFF), the third letter is the number of states (1: only the complex is considered, 3: includes also host and guest in solution), and the fourth letter the type of experimental correction (N: no correction; L: linear correction trained a single dataset including OA, TEMOA, and CB8; O: offset correction trained a single dataset including OA, TEMOA, and CB8; U: linear correction trained on a set excluding CB8 guests; S: two different linear corrections for OA and TEMOA predictions trained on two separated sets including either OA or TEMOA measurements; Z: same as S but with only offset term).




**Appendix 0 Figure 8. Bootstrap distributions including all the movable type submissions.**




**Appendix 0 Figure 9. Enthalpy correlation plots obtained by the methods on the three host-guest sets.**
Scatter plots showing the experimental measurements of the host-guest binding enthalpies (horizontal axis) against the methods' predictions on the OA (yellow), TEMOA (green), and CB8 (blue) guest sets with the respective regression lines of the same color. The solid black line is the regression line obtained by using all the data points. The gray shaded area represent the points within 1.5 kcal/mol from the diagonal (dashed black line).