Predicting binding free energies: Frontiers and benchmarks

David L. Mobley[1, *] and Michael K. Gilson[2, †]

[1] *Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine, CA, USA, 92697*
[2] *Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, CA, USA, 92092*

Binding free energy calculations based on molecular simulations provide predicted affinities for biomolecular complexes. These calculations begin with a detailed description of a system, including its chemical composition and the interactions between its components. Simulations of the system are then used to compute thermodynamic information, such as binding affinities. Because of their promise for guiding molecular design, these calculations have recently begun to see widespread applications in early stage drug discovery. However, many challenges remain to make them a robust and reliable tool. Here, we briefly explain how the calculations work, highlight key challenges, and argue for the development of accepted benchmark test systems that will help the research community generate and evaluate progress.

I. INTRODUCTION

Molecular simulations provide a powerful technique for predicting and understanding the structure, function, dynamics, and interactions of biomolecules. Often, these techniques are valued because they provide a movie of what might be going on at the atomic level. However, simulations also can be used to make quantitative predictions of thermodynamic and kinetic properties, with applications in fields including drug discovery, chemical engineering, and nanoengineering. A thermodynamic property of particular interest is the binding affinity between biomolecules and ligands such as inhibitors, modulators, or activators. With accurate and rapid affinity predictions, we could use simulations in varied health-related applications, such as the prediction of biomolecular interaction networks in support of systems biology, or rapid design of new medications with reduced side-effects and drug resistance.

A. Imagining a tool for drug discovery

A major aim in the development of molecular simulations is to create quantitative, accurate tools which will guide early stage drug discovery. Consider a medicinal chemist in the not-too-distant future who has just finished synthesizing several new derivatives of an existing inhibitor as potential drug leads targeting a particular biomolecule, and has obtained binding affinity or potency data against the desired biomolecular target. Before leaving work, he or she generates ideas for perhaps 100 new compounds which could be synthesized next, then sets a computer to work overnight prioritizing them. By morning, the compounds have all been prioritized based on reliable predictions of their affinity for the desired target, selectivity against alternative targets which should be avoided, solubility, and membrane permeability. The chemist then looks through the predicted properties for the top few compounds and selects the next ones for synthesis. If synthesizing and testing each compound takes several days, this workflow compresses roughly a year's work into a few days.

While this workflow is not yet a reality, huge strides have been made in this direction, with calculated binding affinity predictions now showing real promise [17, 18, 23, 25, 77, 102, 116, 121], solubility predictions beginning to come online [66, 92, 100], and predicted drug resistance/selectivity also apparently tractable [63], with some headway apparent on membrane permeability [21, 59]. A considerable amount of science and engineering still remains to make this vision a reality, but, given recent progress, the question now seems more one of *when* rather than *whether*.

B. Increasing accuracy will yield increasing payoffs

Recent progress in computational power, especially the widespread availability of graphics processing units (GPUs) and advances in automation [68] and sampling protocols, have helped simulation-based techniques reach the point where they now appear to have sufficient accuracy to be genuinely useful in guiding pharmaceutical drug discovery at least for a certain subset of problems [17, 23, 51, 73, 102, 116, 121]. Specifically, in some situations, free energy calculations appear to be capable of achieving RMS errors in the 1-2 kcal/mol range with current force fields, even in prospective applications. As a consequence, pharmaceutical companies are beginning to use these methods in discovery projects. The most immediate application of these techniques is to guide synthesis for lead optimization, but applications to scaffold hopping and in other areas also appear possible.

At the same time, it is clear that not all situations are so favorable, so it is worth asking what level of accuracy

* dmobley@mobleylab.org
† mgilson@ucsd.edu

is actually needed. It is often suggested that we need binding free energy predictions accurate to within ~ 1 kcal/mol, but we are not aware of a clear basis for this figure beyond the fact it is a pleasingly round number that is close to the thermal kinetic energy, RT. Instead of setting a single threshold requirement for accuracy, it is more informative to consider how accurate calculations must be to reduce the number of compounds synthesized and tested by some factor, relative to the number required without computational prioritization. If one targets a three-fold reduction, the answer appears to be that calculations with a 2 kcal/mol RMS error will suffice [77, 106]. Thus, one can gain substantial benefit from simulations that are good yet still quite imperfect.

More broadly, this analysis does not address the net value of computational affinity predictions in drug discovery. Costs include those of the software, computer time, and personnel required to incorporate calculations into the workflow; while benefits include the savings, revenue gains, and externalities attributable to reducing the number of low-affinity compounds synthesized and arriving earlier at a potent drug candidate. In addition, with sufficiently reliable predictions, chemists may choose to tackle difficult synthesis efforts they otherwise might have avoided, resulting in more novel and valuable chemical matter.

C. Overview of free energy calculations

The present review focuses on a class of methods in which free energy differences are computed with simulations that sample Boltzmann distributions of molecular configurations. These samples are usually generated by molecular dynamics (MD) simulations [57], with the system effectively coupled to a heat bath at constant temperature, but Monte Carlo methods may also be used [20, 70, 71]. In either case, the energy of a given configuration is provided by a potential function, or force field, which estimates the potential energy of a system of solute and solvent molecules as a function of the coordinates of all of its atoms. Such simulations may be used in several different ways to compute binding free energies or relative binding free energies, as detailed elsewhere [16, 19, 71, 105] and summarized below. In all cases, however, the calculations yield the free energy difference between two states of a molecular system, and they do so by computing the reversible work for changing the initial state to the final one. Two broad approaches deserve mention.

The first general approach directly computes the standard free energy of binding of two molecules by computing the reversible work of transferring the ligand from the binding site into solution. (This is sometimes called an absolute binding free energy calculation.) The pathway of this change may be one that is physically realizable, or one that is only realizable *in silico*, in which case it is sometimes called an "alchemical" pathway. Physical pathway methods provide the standard binding free energy by computing the reversible work of, in effect, pulling the ligand out of the binding site. Although, by definition, the pathway used must be a physical one that could occur in nature, it need not be probable, and improbable pathways, governed by an order parameter specifying how far the ligand is from the binding site, are often used [8, 48, 52, 115, 125, 130]. In addition, artificial restraints may be useful to avoid sampling problems in the face of often complex barriers along the pathway [8, 48, 52, 115, 125]. By contrast, alchemical pathway methods artificially decouple the ligand from the binding site and then recouple it to solution from the protein [9, 43, 49, 56, 74]. Although alchemical decoupling methods may avoid clashes of the ligand with the protein that might be problematic in pathway methods for a tight binding site, they still can pose some of the same sampling challenges. For example, sampling of the unbound receptor must be adequate after the ligand is removed, and water molecules must have time to equilibrate in the vacated binding site. Given that free energy is a state function, it is not surprising that alchemical and physical pathway approaches yield apparently comparable results when applied to the same systems [24, 47, 62, 128].

The second general approach computes the difference between the binding free energies of two different ligands for the same receptor, by computing the work of artificially converting one ligand into another, first in the bound state and then free in solution [16, 19, 71, 112]. Because these conversions are not physically realizable, such calculations are, again, called alchemical. These calculations can be quite efficient if the two ligands are very similar to each other, but they become more complicated and pose greater sampling problems if the two ligands are very different chemically or if there is a high barrier to interconversion between their most stable bound conformations [68]. In addition, there may be concerns about slow conformational relaxation of the protein in response to the change in ligand. Nonetheless, alchemical relative free energy calculations currently are the best automated and most widely used free energy methods [68, 77, 121].

Importantly, the accuracy and precision of all of these methods are controlled by the same considerations. First, many conformations typically need to be generated, or sampled, in order to obtain an adequate representation of the Boltzmann distribution. In the limit of infinite sampling, a correctly implemented method would yield the single value of the free energy difference dictated by the specification of the molecular system and the chosen force field. In reality, however, only finite sampling is possible, so the reported free energy will differ from the nominal value associated with infinite sampling. In addition, because sampling methods are typically stochastic and the dynamics of molecular systems are highly sensitive to initial conditions [3], repeated calculations, using different random number seeds or initial states, will yield different results. The problem of finite sampling is most acute for systems where low-energy (hence highly

occupied) conformational states are separated by high effective barriers, whether energetic or entropic. Second, even if adequate sampling is achievable, free energy differences may disagree substantially with experiment if the force field is not sufficiently accurate. Third, errors may also arise if the representation of the system in the simulation does not adequately represent the actual system, e.g. if protonation states are assigned incorrectly and held fixed.

D. Challenges and the domain of applicability

Thus, in order for a free energy calculation to be reliable, it must use an appropriate representation of the physical system and an accurate force field, and it must adequately sample the relevant molecular configurations. In the case of the more widely used alchemical relative free energy approach, this means that the best results are expected when:

- a high quality receptor structure is available, without missing loops or other major uncertainties
- the protonation state of the ligand and binding-site residues (as well as any other relevant residues) can reliably be inferred
- the ligand binding mode is defined by crystallographic studies and is not expected to change much on modification
- the receptor does not undergo substantial or slow conformational changes
- key interactions are expected to be well-described by underlying force fields

Beyond this domain of applicability—whose dimensions are, in fact, still somewhat vague — substantial challenges may be encountered. For example, binding free energy calculations for a cytochrome C peroxidase mutant suggest limitations of fixed-charge force fields. In this case, the strength of electrostatic interactions in a buried, relatively nonpolar binding site appears to be overestimated by a conventional fixed-charge force field, likely due to underestimation of polarization effects [96]. Sampling problems are also common, with slow sidechain rearrangements and ligand binding mode rearrangements in model binding sites in T4 lysozyme posing timescale problems unless enhanced or biased sampling methods are carefully applied [11, 35, 54, 75, 76, 120]; and larger-scale protein motions induced by some ligands also posing challenges [11, 64].

Although such problems need not prevent free energy calculations from being used, they can require specific adjustment of procedures and parameters based on experience and knowledge of the system at hand. Thus, a key challenge for the field is how to use insights from well-studied cases to enable automation and reduce the detailed knowledge of each system required to carry out high quality simulations.

Troubleshooting is also a major challenge. In most cases where calculations diverge substantially from experiment, the reason for the discrepancy is not apparent. Is the force field inaccurate? Would the results improve with more sampling? Were protonation states misassigned—or do they perhaps even change on binding? There might even be a software bug [28] or a human error in the use of the software. As a consequence, it is not clear what steps are most urgently needed to advance the field as a whole.

II. THE NEED FOR WELL-CHOSEN BENCHMARK SYSTEMS

Although tests of individual free energy methods are not uncommon today [17, 23, 73, 116, 121], the use of nonoverlapping molecular systems and computational protocols makes it difficult to compare methods on a rigorous basis. In addition, few studies are designed to identify key sources of error and thereby focus future research and development. A few molecular systems have now emerged as *de facto* standards for general study (Section III). These selections result in part from two series of blinded prediction challenges (SAMPL [85], and CSAR [27] followed by D3R [38]), which have helped focus the computational chemistry community on a succession of test cases and highlighted the need for methodological improvements. However, broader adoption of a larger and more persistent set of test cases is needed. By coalescing around a compact set of benchmarks, well chosen to challenge and probe free energy calculations, practitioners and developers will be able to better assess and drive progress in binding free energy calculations.

A. Benchmark types and applications

We envision two classes of benchmark cases: "hard" benchmarks, which are simple enough that well-converged results can readily be computed; and "soft" benchmarks, for which convincingly converged results cannot readily be generated, but which are still simple enough that concerted study by the community can delineate key issues that might not arise in the simpler "hard" cases. The following subsections provide examples of how hard and soft benchmark systems may be used to address important issues in free energy simulations.

1. *Hard benchmarks*

a. Systems to test software implementations and usage It is crucial yet nontrivial to validate that a simulation package correctly implements and applies the desired methods [104], and benchmark cases can help with this. First, all software packages could be tested for their ability to generate correct potential energies for a single

configuration of the specified molecular system and force field. These results should be correct to within rounding error and the precision of the physical constants used in the calculations [104]. Similarly, different methods and software packages should give consistent binding free energies when identical force fields are applied with identical simulation setups and compositions. The benchmark systems for such testing can be simple and easy to converge, and high precision free energies (e.g., uncertainty $\approx$ 0.1 kcal/mol) should serve as a reference. Test calculations should typically agree with reference results to within 95% confidence intervals, from established methods [33, 103], For this purpose, the correctly computed values need not agree with experiment; indeed, experimental results are unnecessary.

b. Systems to check sampling completeness and efficiency As discussed above, free energy calculations require thorough sampling of molecular configurations from the Boltzmann distribution dictated by the force field that is employed. This sampling is typically done by running molecular dynamics simulations, and for systems as large and complex as proteins, it is difficult to carry out long enough simulations. Calculations with inadequate sampling yield results that are imprecise, in the sense that multiple independent calculations with slightly different initial conditions will yield significantly different results, and these ill-converged results will in general be poor estimates of the ideal result obtained in the limit of infinite sampling. Advanced simulation methods have been developed to speed convergence [105, 111], but it is not always clear how various methods compare to one another. To effectively compare such enhanced sampling methods, we need benchmark molecular systems, parameterized with a force field that many software packages can use, that embody various sampling challenges, such as high dimensionality and energetic and entropic barriers between highly occupied states, but which are just tractable enough that reliable results are available via suitable reference calculations. Again, experimental data are not required, and the point of comparison may be, at least in part, sampling measures.

c. Systems to assess force field accuracy Some molecular systems are small and simple enough that current technology allows thorough conformational sampling, and hence well converged calculations of experimental observables. This has long been feasible for liquids [55]; for example, it is easy to precisely compute the heat of vaporization of liquid acetone with one of the standard force fields. More recently, advances in hardware and software have made it possible to compute binding thermodynamics to high precision for simple molecular recognition systems [48], as further discussed below. In such cases, absent complications like uncertain protonation states, the level of agreement with experiment reports directly on the accuracy of the force field. Thus, simple molecular recognition systems with reliable experimental binding data represent another valuable class of benchmarks. Here, of course, experimental data are needed. Ideally, the physical materials will be fairly easy to obtain so that measurements can be replicated or new experimental conditions (such as temperature and solvent composition) explored.

2. Soft benchmarks

a. Systems to challenge conformational sampling techniques Enhanced sampling techniques (Section II A 1 b), designed to speed convergence of free energy simulations, may not be adequately tested by any hard benchmark, because such systems are necessarily rather simple. Thus, despite the fact that reliable reference results are not available for soft benchmarks, they are still important for method comparisons. For example, it may become clear that some methods are better at sampling in systems with high energy barriers, and others in high-dimensional systems with rugged energy surfaces. Developers should test methods on a standard set of benchmark systems for informative comparisons.

b. Direct tests of protein-ligand binding calculations Although it is still very difficult to convincingly verify convergence of many protein-ligand binding calculations, it is still important to compare the performance of various methods in real-world challenges. Appropriate soft benchmarks are likely to be cases which are still relatively tractable, involving small proteins and simple binding sites. We need a series of benchmark protein-ligand systems that introduce various challenges in a well-understood manner. Systems should introduce none, one, two, or N of the following challenges in various combinations:

1. Sampling challenges
   (a) Sidechains in the binding site rearrange on binding different ligands
   (b) Modest receptor conformational changes, such as loop motion
   (c) Large scale conformational changes, such as domain motions and allostery
   (d) Ligand binding modes change unpredictably with small chemical modifications
   (e) High occupancy water sites rearrange depending on bound ligand
2. System challenges
   (a) Protonation state of ligand and/or protein changes on binding
   (b) Multiple protonation states of the ligand and/or receptor are relevant
   (c) Results are sensitive to buffer, salts or other environmental factors
3. Force field challenges
   (a) Strong electric fields suggest that omission of explicit electronic polarizability will limit accuracy
   (b) Ligands interact directly with metal ions
   (c) Ligands or co-factors challenge existing force fields

c. Progression of soft benchmarks We envision these more complex benchmark systems proceeding through stages, initially serving effectively as a playground where major challenges and issues are explored, documented, and become well-known. Eventually, some will become sufficiently well characterized and sampled that they become hard benchmarks.

B. Applications and limitations of benchmark systems

Standard benchmark systems along the lines sketched above will allow potential solutions to be tested in a straightforward, reproducible manner. For example, force fields may be assessed by swapping new parameters, or even a new functional form, into an existing workflow to see the impact on accuracy for a hard benchmark test. Sampling methods may be assessed by using various enhanced sampling methods for either hard or soft sampling benchmarks, here without focusing on accuracy relative to experiment. And system preparation tools could be varied to see how different approaches to assigning protonation states, modeling missing loops, or setting initial ligand poses, affect agreement with experiment—with the understanding that force field and sampling also play a role. Such studies will be greatly facilitated by well-characterized standard benchmarks.

At the same time, there is a possibility that that some methods will inadvertently end up tuned specifically to generate good results for the set of accepted benchmarks. In such cases, the results for systems outside the benchmark set might still be disappointing. This means the field will need to work together to develop a truly representative set of benchmarks. This potential problem can also be mitigated by sharing of methods to enable broader testing by non-developers, and by participation in blinded prediction challenges, such as SAMPL and D3R, which confront methods with entirely new challenge cases.

III. BENCHMARK SYSTEMS FOR BINDING PREDICTION

No molecular systems have been explicitly accepted by the field as benchmarks for free energy calculations, but certain host molecules (see below) and designed binding sites in the enzyme T4 lysozyme have emerged as particularly helpful and widely studied test cases. Here, we describe these artificial receptors and propose specific host-guest and T4 lysozyme-ligand combinations as initial benchmark systems for free energy calculations. We also point to several additional hosts and small proteins that also have potential to generate useful benchmarks in the future (Section VI). The present focus is on cases where experimental data are available and add value, rather than ones chosen specifically to test conformational sampling methods, where experimental data are not required (Section II A).

A. Host-guest benchmarks

Chemical hosts are small molecules, often comprising fewer than 100 non-hydrogen atoms, with a cavity or cleft that allows them to bind other compounds, called guests, with significant affinity. Hosts bind their guests via the same basic forces that proteins used to bind their ligands, so they can serve as simple test systems for computational models of noncovalent binding. Moreover, their small size, and, in many cases, their rigidity, can make it feasible to sample all relevant conformations, making for "hard" benchmarks as defined above (Section II A). Furthermore, experiments can often be run under conditions that make the protonation states of the host and guest unambiguous. Under these conditions, the level of agreement of correctly executed calculations with experiment effectively reports on the validity of the force field (Section II A 1 c. For a number of host-guest systems, the use of isothermal titration calorimetry (ITC) to characterize binding provides both binding free energies and binding enthalpies. Binding enthalpies can often also be computed to good numerical precision [48], so they provide an additional check of the validity of simulations.

Hosts fall into chemical families, such that all members of each family share a major chemical motif, but individuals vary in terms of localized chemical substitutions and, in some families, the number of characteristic monomers they comprise. For example, all members of the cyclodextrin family are chiral rings of glucose monomers; family members then differ in the number of monomers and in the presence or absence of various chemical substituents. For tests of computational methods ultimately aimed at predicting protein-ligand binding affinities in aqueous solution, water soluble hosts are, arguably, most relevant. On the other hand, host-guest systems in organic solvents may usefully test how well force fields work in the nonaqueous environment within a lipid membrane. Here, we focus on two host families, the cucurbiturils [34, 79]; and the octa-acids (more generally, Gibb deep cavity cavitands) [39, 50], which have already been the subject of concerted attention from the simulation community, due in part to their use in the SAMPL blinded prediction challenges [85, 87, 128].

1. *Cucurbiturils*

The cucurbiturils (**Figure 1**) are achiral rings of glycoluril monomers [34]. The first characterized family member, cucurbit[6]uril, has six glycoluril units, and subsequent synthetic efforts led to the five-, seven-, eight- and ten-monomer versions, cucurbit[n]uril (n=5,6,7,8,10) [67], which have been characterized to different extents. Of note, the n=6,7,8 variants accommo-

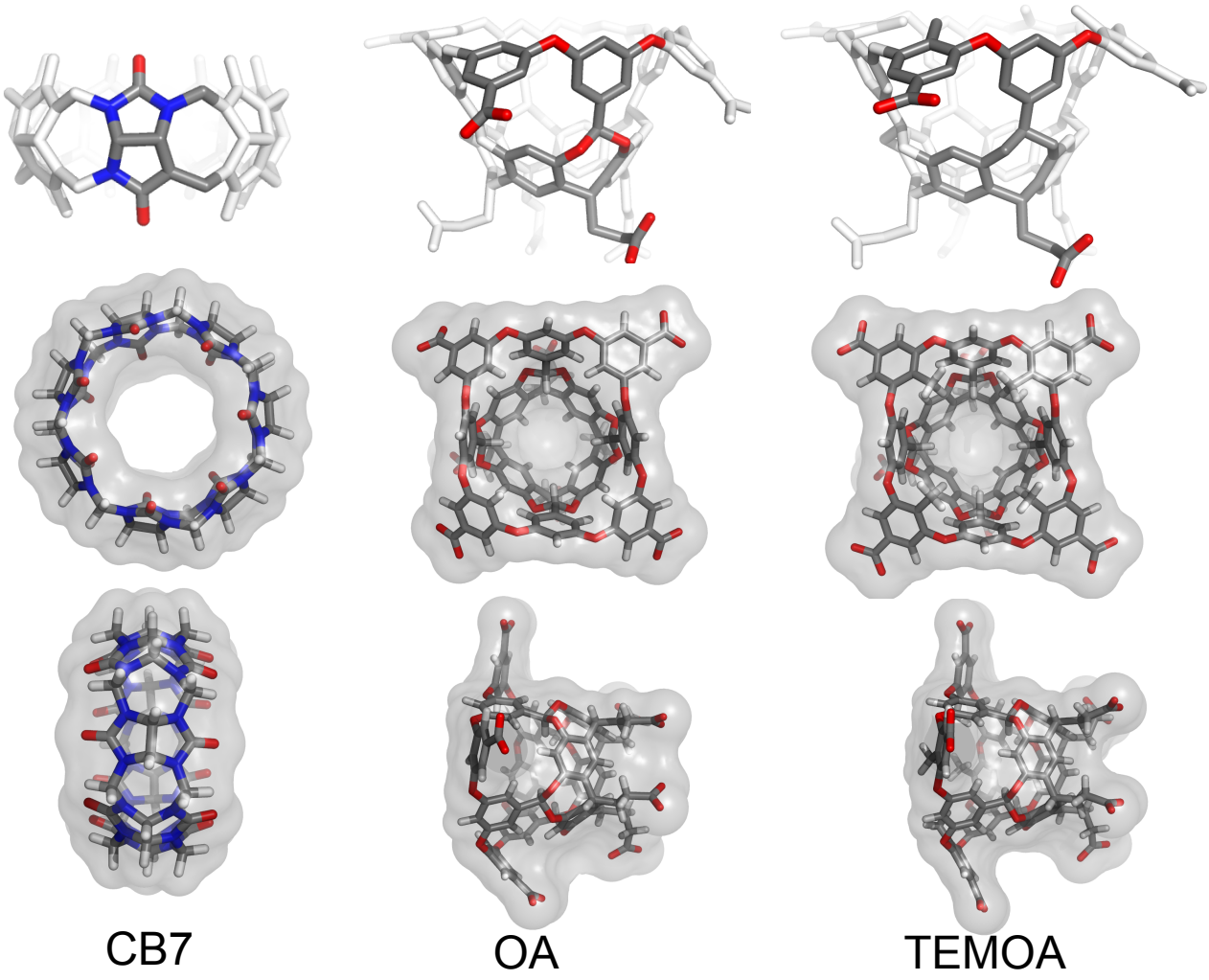


FIG. 1. OA, TEMOA, and CB7 hosts. Shown are the hosts which are the focus of our host-guest benchmark sets – two variants of the octa-acid GDCC, and CB7, a cucurbituril. Guest structures are available in the supplemental material.

date guests of progressively larger size, but are consistent in preferring to bind guests with a hydrophobic core sized to fit snugly into the relatively nonpolar binding cavity, along with at least one cationic moiety (though neutral compounds do bind [60, 126]) that forms stabilizing interactions with the oxygens of the carbonyl groups fringing both portals of the host [67]. Although derivatives of these parent compounds, have been made [4, 22, 61, 117], most of the binding data published for this class of hosts pertain to the non-derivatized forms.

We propose cucurbit[7]uril (CB7) as the basis of one series of host-guest benchmark systems (**Figure 1**, **Tables I and II**). This host is convenient experimentally, because it is reasonably soluble in water; and computationally, because it is quite rigid and lacks acidic or basic groups. In addition, it has attracted particular interest because of the high binding affinities of some guests, exceeding even the tightest-binding protein-ligand systems [14, 67, 80, 95]. Finally, CB7 is already familiar to a number of computational chemistry groups, as it figured in two of the three SAMPL challenges that included host-guest components [85, 87], and it is currently the focus of the "hydrophobe challenge" [101].

a. CB7 presents several challenges Despite the simplicity of CB7, calculations of its binding thermodynamics are still challenging, with several known complexities:

1. **Tight exit portal**: Guest molecules with bulky hydrophobic cores, such as adamantyl or [2.2.2]bicyclooctyl [80, 81] groups, do not fit easily through the constrictive portals [114]. As a consequence, free energy methods which compute the work of binding along a physical dissociation pathway may encounter a high barrier as the bulky core exits the cavity, and this can lead to subtle convergence problems [48, 115]. One way to solve this problem is to reversibly add restraints that open the portal, then remove the guest, and finally reversibly remove the restraints [48], including all of these contributions in the overall work of dissociation.

2. **Water binding and unbinding**: If one computes the work of removing the guest from the host by a nonphysical pathway, in which the bound guest is gradually decoupled from the host and surrounding water [43], large fluctuations in the number of water molecules within the host's cavity can occur when the guest is partly decoupled, and these fluctuations can slow convergence [98].
3. **Salt concentration and buffer conditions**: Binding thermodynamics are sensitive to the composition of dissolved salts, both experimentally [80, 81, 85] and computationally [52, 88]. As a consequence, to be valid, a comparison of calculation with experiment must adequately model the experimental salt conditions.
4. **Finite-size artifacts due to charge modification**: Because many guest molecules carry net charge, it should be ascertained that calculations in which guests are decoupled from the system do not generate artifacts related to the treatment of long-ranged Coulombic interactions [65, 93, 97, 107].

b. The proposed CB7 benchmark sets comprise two compound series For CB7, we have selected two sets of guests that were studied experimentally under uniform conditions (50 mM sodium acetate buffer, pH 4.74, 298K) by one research group [14, 67]. Each series is based on a common chemical scaffold, making it amenable to not only absolute but also alchemical relative free energy calculations (Section I C). One set is based on an adamantane core (**Table I**), and the other on an aromatic ring (**Table II**). These systems can be run to convergence to allow detailed comparisons among methods and with experiment. Their binding free energies range from -5.99 to -17.19 kcal/mol, with the adamantane series spanning a particularly large range of free energies.

c. Prior studies provide additional insight into CB7's challenges Sampling of the host appears relatively straightforward in CB7 as it is quite rigid and its symmetry provides for clever convergence checks [48, 82]. Due to its top-bottom symmetry, flips of guests from "head-in" to "head-out" configurations are not necessary to obtain convergence [31]. However, sampling of the guest geometry can be a challenge, with transitions between binding modes as slow as 0.07 flips/ns [82], and flexible guests also presenting challenges [82]. As noted above, water sampling can also be an issue, with wetting/dewetting transitions occurring on the 50 ns timescale [98].

Salt and buffer conditions are also key. In addition to the strong salt-dependence of binding [81], acetic acid (such as in a sodium acetate buffer) can compete with guests for the binding site [80]. This may partially explain systematic errors in some computational studies [52, 88]. Indeed, the difference between 50 mM sodium acetate buffer and 100 mM sodium phosphate buffer impacts measured binding free energies by 2.5-2.8 kcal/mol [85, 88]. Cationic guests could also have substantial and differing interactions with the counterions in solution as well, potentially lowering affinity relative to zero-salt conditions [85]. Thus, one group found a 6.4-6.8 kcal/mol dependence on salt concentration [52], possibly impacting other studies as well [82]

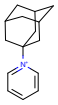
Despite these issues, CB7 appears to be at the point where careful studies can probe the true accuracy of our force fields [37, 48, 127], and the results can be sobering, with RMS errors in the binding free energies as high as 8 kcal/mol [48, 82]. More encouragingly, the values of R^2 values can be as high as 0.92 [48].) Some force fields appear relatively worse than others [52, 86]. Calculated values are in many cases quite sensitive to details of force field parameters [81, 82, 86]. For example, modest modification of some Lennard-Jones parameters yielded dramatic improvements in calculated values [127], and host-guest binding data has, accordingly, been suggested as an input for force field development [37, 48, 127]. Water structure around CB7 and calculated binding enthalpies also appear particularly sensitive to the choice of water model [31, 37, 98], and water is clearly important for modulating binding [90]. The water model also impacts the number of sodium ions which must be displaced (in sodium-based buffer) on binding [37, 48].

Despite its apparent simplicity, CB7 is still a challenging benchmark that can put important issues into high relief. For example, in SAMPL4, free energy methods yielded R^2 values from 0.1 to 0.8 and RMS errors of about 1.9 to 4.9 kcal/mol for the same set of CB7 cases. This spread of results across rather similar methods highlights the need for shared benchmarks. Potential explanations include convergence difficulties, subtle methodological differences, and details of how the methods were applied [85]. Until the origin of such discrepancies is clear, it is difficult to know how accurate our methods truly are.

2. Gibb Deep Cavity Cavitands (GDCC)

The octa-acids (OA) (**Figure 1**) are synthetic hosts with deep, basket-shaped, hydrophobic binding sites [39]. The eight carboxylic acidic groups for which they were originally named make these hosts water-soluble, but do not interact directly with bound hosts; instead, they project outward into solvent. Binding data have been reported for the original form of this host (OA) [39] and for a derivative with four added methyl groups at equivalent locations in the entryway, where they can contact a bound guest (TEMOA) [36, 109]. (Note that OA and TEMOA have also been called OAH and OAMe, respectively [128].) Additional family members with other substituents around the portal have been reported, as has a new series in which the eponymic carboxylic groups are replaced by various other groups, including a number of basic amines [50]. However, we are not aware of binding data for these derivatives. In view of these other hosts, however, we propose the more general name Gibb deep cavity cavitands (GDCCs) for this family of hosts. The binding cavities of the GDCCs are fairly rigid, though

TABLE I. Proposed CB7 Set 1 benchmark data

ID[a]	name	PC CID[b]	2D	SMILES	ΔG^{c} $(kcal/mol)$
1	Memantine	4054		CC12CC3CC(C1)(CC(C3)(C2)N)C	-5.99 ± 0.05 [d]
3	1,3-Bis(trimethylaminio)adamantane	101379195		C[N+](C)(C)C12CC3CC(C1)CC(C3)(C2)[N+](C)(C)C	-6.55 ± 0.05 [d]
5	N-(1-Adamantyl)ethylenediamine	303798		C1C2CC3CC1CC(C2)(C3)NCCN	-18.22 ± 0.09 [e]
17	Adamantane-1,3-diamine	213512		C1C2CC3(CC1CC(C2)(C3)N)N	-11.33 ± 0.05 [d]
18	1-Adamantanecarboxylic acid	13235		C1C2CC3CC1CC(C2)(C3)C(=O)O	-11.59 ± 0.06 [d]
22	1-Adamantyltrimethylaminium	3010127		C[N+](C)(C)C12CC3CC(C1)CC(C3)C2	-16.66 ± 0.08 [d]
23	amantadine	2130		C1C2CC3CC1CC(C2)(C3)N	-17.19 ± 0.08 [d]
24	N-(1-Adamantyl)pyridinium	3848257		C1C2CC3CC1CC(C2)(C3)[N+]4=CC=CC=C4	-16.75 ± 0.07 [d]

[a] Compound ID from original paper; [b] PubChem Compound ID; [c] Standard binding free energy, where all measurements were done via NMR in 50mM sodium acetate buffer in D_2O at pH 4.74 and 298 K. Uncertainties are obtained by taking the reported standard deviations across triplicate measurements [53] and dividing by $\sqrt{3}$; [d] drawn from [67]; [e] drawn from [14].

TABLE II. Proposed CB7 Set 2 benchmark data

ID[a]	name	PC CID[b]	2D	SMILES	$\Delta G^{c,d}$ $(kcal/mol)$
2	Dopamine	681		C1=CC(=C(C=C1CCN)O)O	-6.31 ± 0.05
4	O-phenylenediamine	7243		C1=CC=C(C(=C1)N)N	-6.68 ± 0.05
5	m-Phenylenediamine	7935		C1=CC(=CC(=C1)N)N	-6.69 ± 0.02
7	4-(Aminomethyl)pyridine	77317		C1=CN=CC=C1CN	-7.56 ± 0.06
8	p-Phenylenediamine	7814		C1=CC(=CC=C1N)N	-8.60 ± 0.06
9	P-toluidine	7813		CC1=CC=C(C=C1)N	-9.43 ± 0.05
20	P-Xylylenediamine	68315		C1=CC(=CC=C1CN)CN	-12.62 ± 0.06

[a] Compound ID from original paper; [b] PubChem Compound ID; [c] Standard binding free energy, where all measurements were done via NMR in 50mM sodium acetate buffer in D_2O at pH 4.74 and 298 K. Uncertainties are obtained by taking the reported standard deviations across triplicate measurements [53] and dividing by $\sqrt{3}$; [d] drawn from [67].

less so than the cucurbiturils. Some simulators report "breathing" motions that vary the diameter of the entry by up to 8 Å[72]; and, in some studies, the benzoic acid "flaps" around the entry occasionally flip upward and into contact with the guest [113, 129], though this motion has not been verified experimentally. Additionally, the four priopionate groups protruding into solution from the exterior base of the cavity are all flexible.

The octa-acids tend to bind guest molecules possessing a hydrophobic moiety that fits into the host's cavity and a hydrophilic moiety that projects into the aqueous solvent. Within these specifications, they bind a diversity of ligands, including both organic cations and anions, as well as neutral compounds with varying degrees

of polarity [40, 42]. Compounds with adamantane or noradamantane groups display perhaps the highest affinities observed so far, with binding free energies ranging to about -8 kcal/mol [110]. Much of the experimental binding data comes from ITC, so binding enthalpies are often available.

Two experimental aspects of binding are particularly intriguing and noteworthy. First, the binding thermodynamics of OA is sensitive to the type and concentration of anions in solution. Although NaCl produces relatively modest effects, 100mM sodium perchlorate, chlorate and isothiocyanate can shift binding enthalpies by up to about 10 kcal/mol and free energies by around 2 kcal/mol [41]. These effects are due in part to binding of anions by the host; indeed, trichloroacetate is reported to bind OA with a free energy of -5.2 kcal/mol [108], and competition of other guests with bound anions leads to entropy-enthalpy tradeoffs. Second, elongated guests can generate ternary complexes, in which two OA hosts encapsulate one guest, especially if both ends of the guest are not very polar [40].

a. The proposed GDCC benchmark sets are drawn from SAMPL As a core benchmark series for this family, we propose two sets which formed part of the SAMPL4 and SAMPL5 challenges, based on adamantane derivatives (Table III) and cyclic (aromatic and saturated) carboxylic acids (Table IV) binding to hosts OA and TEMOA with free energies of -3.7 to -7.6 kcal/mol. These cases offer aqueous binding data with a reasonably broad range of binding free energies, frequently along with binding enthalpies; the hosts and many or all of their guests are small and rigid enough to allow convincing convergence of binding thermodynamics with readily feasible simulations; and, like the cucurbiturils, they are already emerging as *de facto* computational benchmarks, due to their use in the SAMPL4 and SAMPL5 challenges [85, 128].

b. OA introduces new challenges beyond CB7 Issues deserving attention when interpreting the experimental data and calculating the binding thermodynamics of these systems include the following:

1. **Tight exit portal**: The methyl groups of the TEMOA variant narrow the entryway and can generate a barrier to the entry or exit of guest molecules with bulky hydrophobic cores, though the degree of constriction is not as marked as for CB7 (above). The TEMOA methyls groups can additionally hinder sampling of guest poses in the bound state, leading to convergence problems [128] specific to TEMOA.
2. **Host conformational sampling**: Although the flexible propionate groups are not proximal to the binding cavity, they are charged and so can have long-ranged interactions. As a consequence, it may be important to ensure their conformations are well sampled, though motions may be slow [72]. Similarly, benzoic acid flips [113, 129] could potentially be an important challenge in some force fields.
3. **Water binding and unbinding:** Water appears to undergo slow motions into and out of the OA host, on timescales upwards of 5ns [30]. This poses significant challenges for some approaches, such as metadynamics, where deliberately restraining water to stay out of the cavity when the host is not bound (and computing the free energy of doing so) can help convergence [8], and perhaps for other methods as well.
4. **Salt concentration and buffer conditions:** As in the case of CB7, binding to GDCCs is modulated by the composition of dissolved salts, both experimentally [41, 108] and computationally [91, 113]. As a consequence, to be valid, a comparison of calculation with experiment must adequately model the experimental salt conditions.
5. **Finite-size artifacts due to charge modification**: As for CB7, it should be ascertained that calculations in which charged guests are decoupled from the system do not generate artifacts related to long range Coulomb interactions. [65, 93, 97, 107].
6. **Protonation state effects**: Although experiments are typically run at pH values that lead to well-defined protonation states of the host and its guests, this may not always hold [30, 85, 113], particularly given experimental evidence for extreme binding-driven pKa shifts of 3-4 log units for some carboxylate compounds [108, 119]. Thus, attention should be given to ionization states and their modulation by binding.

c. Prior studies provide additional insight into the challenges of OA As noted, two different host conformational sampling issues have been observed, with dihedral transitions for the proprionate groups occurring on 1-2 ns timescales [72]); motions of the benzoic acid flaps were also relatively slow [113, 129] though perhaps thermodynamically unimportant. Guest sampling can also be an issue, at least in TEMOA [128], and this hosts's tight cavity may also have implications for binding entropy [129].

Salt concentration strongly modulates binding affinity, at least for anions, and the nature of the salt also plays an important role [15]. Co-solvent anions can also increase *or* decrease binding depending on their identity [41]. Some salts even bind to OA themselves, with perchlorate [41] and trichloroacetate [108] being particularly potent, and thus will compete with guests for binding. Computationally, including additional salt beyond that needed for system neutralization changed binding free energies by up to 4 kcal/mol [113].

Naively, protonation states of the guests might seem clear and unambiguous. But since OA can bind guests of diverse net charges, the protonation state may not always be clear. One study used absolute binding free energy calculations for different guest charge states, coupled with pKa calculations, and found that inclusion of pKa corrections and the possibility of alternate charge states of the guests affected calculated binding free energies by up to

TABLE III. Proposed GDCC Set 1 benchmark data

ID[a]	name	PC CID[b]	2D	SMILES	ΔG^c $(kcal/mol)$	ΔH^d $(kcal/mol)$
Octa Acid binders						
3 / OA-G1	5-Hexynoic acid	143036		C#CCCCC(=O)O	-5.40 ± 0.003	-7.71 ± 0.05
4 / OA-G6	3-nitrobenzoic acid	8497		C1=CC(=CC(=C1)[N+](=O)[O-])C(=O)O	-5.34 ± 0.005	-5.67 ± 0.01
5 / OA-G2	4-cyanobenzoic acid	12087		C1=CC(=CC=C1C#N)C(=O)O	-4.73 ± 0.01	-4.45 ± 0.08
6 / OA-G4	4-bromoadamantane-1-carboxylic acid	12598766		C1C2CC3CC(C2)(CC1C3Br)C(=O)O	-9.37 ± 0.01	-14.78 ± 0.02
7 / OA-G3	N,N,N-trimethylhexan-1-aminium	84774		CCCCCC[N+](C)(C)C	-4.49 ± 0.01	-5.91 ± 0.10
8 / OA-G5	trimethylphenethylaminium	14108		C[N+](C)(C)CCC1=CC=CC=C1	-3.72 ± 0.01	-9.96 ± 0.11
TEMOA/OAMe binders						
3 / OA-G1	5-Hexynoic acid	143036		C#CCCCC(=O)O	-5.476 ± 0.006	-9.961 ± 0.006
4 / OA-G6	3-nitrobenzoic acid	8497		C1=CC(=CC(=C1)[N+](=O)[O-])C(=O)O	-4.52 ± 0.02	-9.1 ± 0.1
5 / OA-G2	4-cyanobenzoic acid	12087		C1=CC(=CC=C1C#N)C(=O)O	-5.26 ± 0.01	-7.6 ± 0.1
6 / OA-G4	4-bromoadamantane-1-carboxylic acid	12598766		C1C2CC3CC(C2)(CC1C3Br)C(=O)O	ND[e]	ND[e]
7 / OA-G3	N,N,N-trimethylhexan-1-aminium	84774		CCCCCC[N+](C)(C)C	-5.73 ± 0.06	-6.62 ± 0.2
8 / OA-G5	trimethylphenethylaminium	14108		C[N+](C)(C)CCC1=CC=CC=C1	ND[e]	ND[e]

[a] Compound ID from [109] and SAMPL5 ID from [128]; [b] PubChem Compound ID; [c] Standard binding free energy from [109], where all measurements were done via ITC in 50 mM sodium phosphate buffer at pH 11.5 and 298 K. Uncertainties, drawn from the experimental paper, were computed from triplicate measurements taken with freshly made solutions of host and guest. However, based on personal communication with the authors, it may be advisable to regard the accuracy more conservatively, at ∼2% for ΔG and ∼6% for ΔH; [d] measured binding enthalpy [109], subject to the same conditions/caveats as [c]. [e] not done.

TABLE IV. Proposed GDCC Set 2 benchmark data

ID[a]	name	PC CID[b]	2D	SMILES	Method	ΔG^c $(kcal/mol)$
1	Benzoic acid	243		C1=CC=C(C=C1)C(=O)O	NMR	-3.72 ± 0.03
2	4-Methylbenzoic acid	7470		CC1=CC=C(C=C1)C(=O)O	NMR	-5.85 ± 0.06
3	4-ethylbenzoic acid	12086		CCC1=CC=C(C=C1)C(=O)O	ITC	-6.27 ± 0.01
4	4-Chlorobenzoic acid	6318		C1=CC(=CC=C1C(=O)O)Cl	ITC	-6.72 ± 0.01
5	3-chlorobenzoic acid	447		C1=CC(=CC(=C1)Cl)C(=O)O	NMR	-5.24 ± 0.02
6	cyclohexanecarboxylic acid	7413		C1CCC(CC1)C(=O)O	NMR	-5.62 ± 0.04
7	trans-4-Methylcyclohexanecarboxylic acid	20330		[C@@H]1(CC[C@@H](CC1)C(=O)O[H])C	ITC	-7.61 ± 0.04

[a] Compound ID from original paper [42]; [b] PubChem Compound ID; [c] Standard binding free energy from [42], where all measurements were done in 10 mM sodium tetraborate buffer at pH 9.2 and 298 K. A quirk is that for the NMR measurements, the guest was titrated in from 50 mM sodium tetraborate buffer, so the buffer concentration changed during the titration. Uncertainty is the standard error of the mean in free energy, computed from the reported standard deviations in K_a. Again, based on personal communication with the authors, uncertainties of perhaps 10% may be more appropriate.

2 kcal/mol [113]. As noted above, experimental evidence also indicates major pKa shifts on binding so that species such as acetate, formate and others would bind in neutral form at neutral pH [108, 119]. Even the host protonation state may be unclear; while OA is often assumed to have all eight carboxylic acids deprotonated at the basic pH of typical experiments, the four at the bottom are in close proximity, and these might make hydrogen bonds allowing retention of two protons [30]. Thus, there are uncertainties as to the host protonation state [30, 85], which perhaps also could be modulated by guest binding.

Several groups used different methods but the same force field and water model in SAMPL5, with rather varied levels of success because of discrepancies in calculated free energies [8, 10, 128]). However, some of these issues were resolved in follow-up work [8], bringing the methods into fairly good agreement for the majority of cases [10, 129].

B. Protein-ligand benchmarks: the T4 lysozyme model binding sites

Although we seek ultimately to predict binding in systems of direct pharmaceutical relevance, simpler protein-ligand systems can represent important stepping stones in this direction. Two model binding sites in T4 lysozyme have been particularly useful in this regard (**Figure 2**). These two binding sites, called L99A [83, 84] and L99A/M102Q [45, 122] for point mutations which create the cavities of interest, are created in artificial mutants of phage T4 lysozyme, and have been studied extensively experimentally and via modeling. As protein-ligand systems, they introduce additional complexities beyond those observed in host-guest systems, yet they share some of the same simplicity. The ligands are generally small, neutral, and relatively rigid, with clear protonation states. In many cases, substantial protein motions are absent, allowing calculated binding free energies to apparently converge relatively easily. However, like host-guest systems, these binding sites are still surprisingly challenging [11, 35, 54, 64, 74–76]. In addition, precise convergence is sometimes difficult to achieve, and it is in all cases essentially impossible to fully verify. As a consequence, these are "soft benchmarks" as defined above (Section II A). The importance of the lysozyme model sites is also driven by the relative wealth of experimental data. It is relatively easy to identify new ligands and obtain high quality crystal structures and affinity measurements, allowing two different rounds of blind predictions testing free energy calculations [11, 76].

1. *The apolar and polar cavities and their ligands*

The L99A site is also called the "apolar" cavity. It is relatively flat and elongated, and binds mostly nonpolar molecules such as benzene, toluene, p-xylene, and n-butylbenzene: basically, a fairly broad range of nonpola,r planar five- and six-membered rings and ring systems (such as indole). The polar version, L99A/M102Q, introduces an additional point mutation along one edge of the binding site, providing a glutamine that introduces polarity and the potential for hydrogen bonding. It still binds a variety of nonpolar ligands such as toluene (though not benzene). One small downside of these binding sites is that the range of affinities is relatively narrow: about -4.5 to -6.7 kcal/mol in the apolar site [76, 83], and about -4 to -5.5 kcal/mol in the polar site [11]). Thus, even the strongest binders are not particularly strong, and the weakest binders tend to run up against their solubility limits. Still, these sites offer immensely useful tests for free energy calculations.

For both sites, fixed charge force fields seem to yield reasonably accurate free energies, with RMS errors between 1-2 kcal/mol, and some level of correlation with experiment, despite limited dynamic range [11, 26, 35, 76, 118]. System composition/preparation issues also do not seem to be a huge factor. Instead, sampling issues predominate:

1. Ligand binding mode/orientational sampling: The binding sites are buried and roughly oblong, with ligands which are similar in shape. Ligands with axial symmetry typically have at least two reasonably likely binding modes, but broken symmetry can drive up the number of likely binding modes. For example, phenol has two plausible binding modes in the polar cavity [11, 46] but 3-chlorophenol has at least four, three of which appear to have some population in simulations [35], because the chlorine could point in either direction within the site. Timescales for binding mode interconversion are relatively slow, with in-plane transitions on the 1-10 nanosecond timescale, and out-of-plane transitions (e.g. between toluene's two symmetry-equivalent binding modes) taking hundreds of nanoseconds (Mobley group, unpublished data).
2. Sidechain rearrangements: Some sidechains are known to reorganize when binding certain ligands. The smallest ligands tend not to induce conformational changes, but larger ligands may induce sidechain rearrangements – often, rotamer flips – around the binding site region. These can be slow in the tightly packed binding site. This especially occurs for Val111 in the L99A site [54, 75, 84] and Leu118, Val11, and Val103 in L99A/M102Q [11, 46, 122, 123]. These sidechain motions typically present sampling problems for standard MD simulations [11, 54, 75, 76, 120].
3. Backbone sampling: Larger ligands induce shifts of the F helix, residues 107 or 108 to 115, adjacent to the binding site, allowing the site to enlarge. This occurs in both binding sites [11, 69, 123], but is best characterized for L99A [69]. There, addition of a series of methyl groups from benzene up to n-hexylbenzene causes a conformational transition in the protein from closed to intermediate to open conformations.

Tables V and VI introduce proposed benchmark sets for the apolar and polar cavities, giving ligands potentially amenable to both absolute and relative free energy calculations, and spanning the range of available affinities. Co-crystal structures are available in most cases, and the PDB IDs are provided in the tables. The selected ligands span a range of challenges and levels of difficulty, ranging from fairly simple to including most of the challenges noted above. Essentially all of them have been included in at least one prior computational study, and some have appeared in a variety of prior studies. Additional known ligands and non-binders are available, with binding affinities available for 19 compounds in the L99A site [29, 76, 83] and 16 in L99A/M102Q [11, 45, 122]. Because of the extent of the sampling challenges in lysozyme, binding of most ligands will currently constitute a soft benchmark, though long-timescale simulations

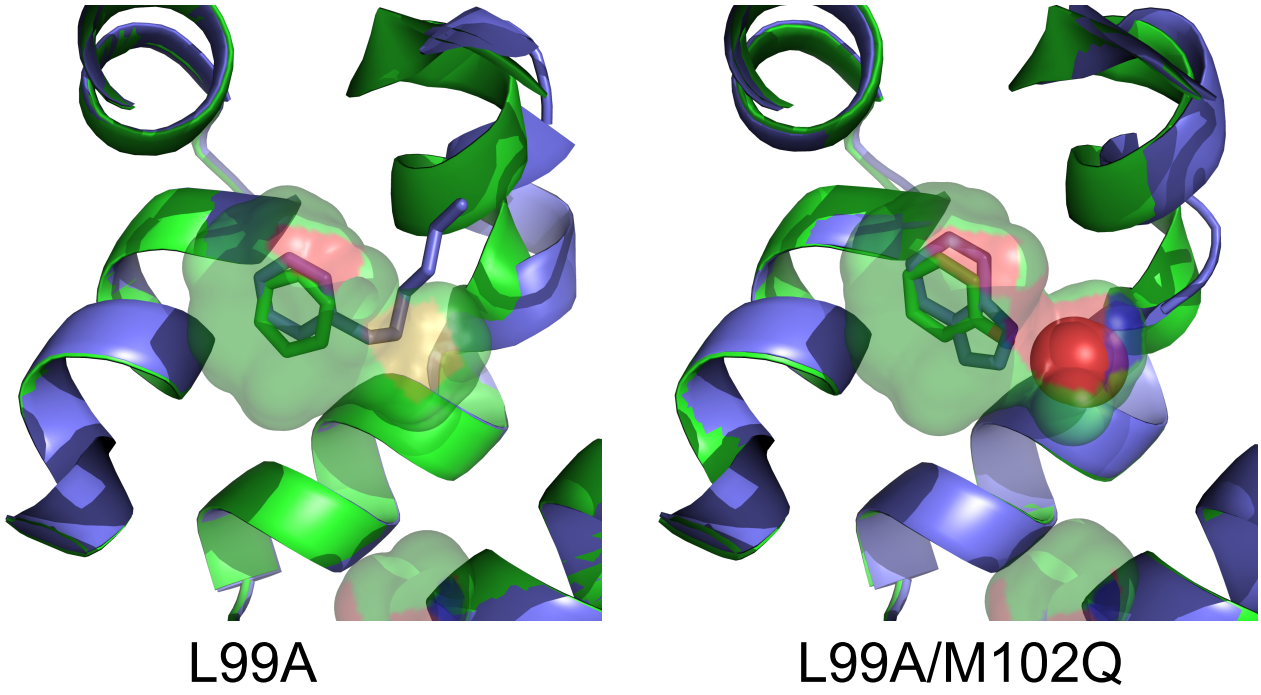


FIG. 2. Benzene and hexylbenzene in the lysozyme L99A site, and phenol and 4,5,6,7-tetrahydroindole in the L99A/M102Q site (PDBs 4W52, 4W59, 1LI2, and 3HUA, respectively). The binding site shape is shown as a semi-transparent surface, and the protein shown with cartoons. In both cases, the structure with the smaller ligand is shown in green and that with the larger ligand is shown in blue, and the larger ligand induces a motion of helix F bordering the binding site. Phenol and 4,5,6,7-tetrahydroindole both also bind with an ordered water, though this does not occur for all ligands in the polar L99A/M102Q site.

to turn these into hard benchmarks may already be feasible.

2. *Computational challenges posed by the T4 lysozyme benchmarks*

Early work on the lysozyme sites focused on the difficulty of predicting binding modes [11, 74, 76] because of the slow interconversions noted above. Docking methods often can generate reasonable poses spanning most of the important possibilities [11, 46, 74, 76] but do not accurately predict the binding mode of individual compounds [11, 46, 76]. Thus it appears necessary to consider the possibility of multiple binding modes; this is also important since some ligands actually populate multiple binding modes [11]. In a number of studies, candidate binding modes from docking are relaxed with MD simulations, then clustered to select binding modes for further study. It turns out an effective binding free energy for each distinct candidate binding mode can be computed separately [74] and combined to find the population of each binding mode and determine the overall binding free energy. However, this is costly since each candidate binding mode requires a full binding free energy calculation.

Relative binding free energy calculations do not dramatically simplify the situation. Introduction of a ligand modification can leave the binding mode uncertain (e.g., introducing a chlorine onto phenol leaves at least two possible binding modes even if the binding mode of phenol is known) [11]. A naïve solution is to consider multiple possible binding modes in relative free energy calculations [11], but this generates multiple results; determining the true relative binding free energy requires additional information [77]. Enhanced sampling approaches provide one possible solution to the binding mode problem. Particularly, with λ or Hamiltonian exchange techniques, ligands can easily switch between binding modes when they are non-interacting unless they are restrained, and then moves in λ space can allow transitions back to the interacting state. Thus, approaches employing this strategy can naturally sample multiple binding modes [35, 118].

While sidechain sampling has been a significant challenge, it is possible to use biased sampling techniques such as umbrella sampling to deliberately compute and include free energies of sampling slow sidechain rearrangements [75]. However, this is not a general solution, since it requires knowing what sidechains might rearrange on binding and then expending substantial computational power on sampling free energy landscapes for these rearrangements. An apparently better general strategy is including sidechains in enhanced sampling regions selected for Hamiltonian exchange [54, 58]

TABLE V. Proposed Lysozyme L99A Set benchmark data

name	PC CID[h]	2D	SMILES	ΔG^{a} $(kcal/mol)$	PDB code	reference
benzene[b]	241		c1ccccc1	-5.19 ± 0.16	181L [84], 4W52 [69]	[83]
toluene[b]	1140		Cc1ccccc1	-5.52 ± 0.04	4W53 [69]	[83]
ethylbenzene[b]	7500		CCc1ccccc1	-5.76 ± 0.07	1NHB [84], 4W54 [69]	[83]
propylbenzene[b]	7668		CCCc1ccccc1	-6.55 ± 0.02	4W55 [69]	[83]
butylbenzene[b]	7705		CCCCc1ccccc1	-6.70 ± 0.02	186L [84], 4W57 [69]	[83]
hexylbenzene[b]	14109		CCCCCCc1ccccc1	UNK[c]	4W59 [69]	[83]
p-xylene[d]	7809		Cc1ccc(cc1)C	-4.67 ± 0.06	187L [84]	[83]
benzofuran	9223		c1ccc2c(c1)cco2	-5.46 ± 0.03	182L [84]	[83]
thieno[2,3-c]pyridine	9224		c1cncc2c1ccs2	NB[e]	ND[f]	[76]
phenol[g]	996		c1ccc(cc1)O	NB[e]	ND[f]	[76, 83]

[a]T=302K, with compounds from [83] measured in 50mM sodium acetate at pH 5.5 and thieno[2,3-c]pyridine measured at pH 6.8 in 50 mM potassium chloride and 38% (v/v) ethylene glycol; [b] part of the series of [69], so larger ligands in the series induce conformational change; [c] unknown due to solubility limitations, but likely binds strongly; [d] L99A sidechain undergoes rotation; [e] nonbinder; [f] not done; [g] included since it is a binder in the polar cavity; [h] PubChem compound ID.

or REST [120], allowing sidechains to be alchemically softened or torsion barriers lowered (or both), to enhance sampling at alchemical intermediate states. With swaps between λ values, enhanced sidechain sampling at intermediate states can propagate to all states, improving convergence [54, 120].

Larger protein conformational changes in lysozyme have received less attention, partly because until very recently they seemed to be a peculiar oddity only rarely observed; i.e., for ligands 4,5,6,7-tetrahydroindole and benzyl acetate in the polar site [11]. However, recent work noted above highlighted how a helix in the apolar cavity can open to accommodate larger ligands [69]. Timescales for this motion appear to be on the order of 50 ns, so it can pose sampling challenges, even for relative free energy calculations [64]. Including part of the protein in the enhanced sampling region via REST2 provides some benefits, but sampling these motions will likely prove a valuable test for enhanced sampling methods.

IV. HOW TO USE BENCHMARK SYSTEMS

Benchmark systems will have several uses, as noted above **??**, but not all benchmark systems can cover all uses. Some will be particularly valuable for testing the accuracy of methods (and in this case, relatively large numbers of ligands or binders are needed to get good statistics) either against gold-standard results or against experimental values (for assessing force field accuracy). But other benchmark systems will be more useful for testing sampling techniques, and still others will initially serve primarily as a test bed to determine the sensitivity of the results to different factors.

In our view, benchmark systems will serve also to help design careful computational experiments. Researchers can test whether a particular method is sampling the motions which others have already shown to be important, or how the choice of starting conformation impacts the rate of convergence to a known, gold standard value for a particular force field and system composition. The availability of benchmark systems will also facilitate comparisons where only a single piece of a workflow is modified; for example, if a different protonation state assignment tool is applied to the receptor, how does that impact computed binding free energies without any other variations in the protocol? Of course these types of comparisons can already be done, but in the context of a benchmark system this will be more valuable information, as it provides insight for other researchers working on the same system into the relative importance of different protonation states, etc.

As noted above (Section **??**), the hope is that ultimately, gold standard binding free energy results will be made available for a set of benchmark systems. These would be from fully converged binding free energy calculations, and give correct results for a particular force field and system preparation, allowing quantitative comparison of the force field results with experiment. But such results will also facilitate a great deal of science on method efficiency; new methods which purport to be more efficient could easily and *automatically* be run on a standard set of systems to see how much more efficient they are than the (perhaps brute-force) method which yielded the gold standard results, and different enhanced sampling methods could easily be observed to

TABLE VI. Proposed Lysozyme L99A/M102Q Set benchmark data

ligand	PC CID[k]	2D	SMILES	ΔG^{a} $(kcal/mol)$	PDB code	reference
toluene[b]	1140		Cc1ccccc1	-4.93	ND[c]	[122]
phenol	996		c1ccc(cc1)O	-5.24	1LI2 [122]	[122]
catechol[h]	289		c1ccc(c(c1)O)O	-4.16 ± 0.03	1XEP [45]	[122]
2-ethoxyphenol[d]	66755		CCOc1ccccc1O	-4.02 ± 0.03	3HU8 [11]	[11]
benzyl acetate[e,f]	8785		CC(=O)OCc1ccccc1	-4.48 ± 0.16	3HUK [11]	[11]
4,5,6,7-tetrahydroindole[f]	57452536		c1c[nH]c2c1CCCC2	-4.61 ± 0.09	3HUA [11]	[11]
n-phenylglycinonitrile[g]	76372		c1ccccc1NCC#N	-5.52 ± 0.18	2RBO [11]	[11]
3-chlorophenol	7933		c1cc(cc(c1)Cl)O	-5.51	1LI3 [122]	[122]
2-methoxyphenol	460		COc1ccccc1O	NB[i]	ND[c]	[11]
4-vinylpyridine	7502		C=Cc1ccncc1	NB[i]	ND[c]	[122]

[a]T=283K, with measurements done at pH 6.8 in 50 mM potassium phosphate, 200 mM potassium chloride buffer in the case of [11]; [b] included for symmetry with the L99A site since this (unlike phenol and benzene) binds in both; [c] not determined; [d] fails to make crystallographic hydrogen bond [11]; [e] multiple binding modes; [f] induces helix F motion; [g] induces flip of Val111 sidechain; [h] induces flip of Leu118 sidechain; [j] nonbinder; [k] PubChem compound ID.

have strengths and weaknesses on known problem classes. Ultimately, these systems can allow automated testing of the efficiency of new methods on real-world problems.

V. WE NEED WORKFLOW SCIENCE

While the benchmark systems discussed here will already be useful, to fully realize their benefits a great deal of engineering needs to be done to facilitate workflow science. Currently, a wide variety of different tools are available for different stages of the free energy calculation process, from system preparation (protonation state assignment, building in missing residues and loops, adding counterions, etc.) to force field assignment, to planning and conducting the calculations themselves (choice of method, simulation package, and so on). Often, these tools live in their own ecosystems and are not typically designed to be easily interchangeable with tools from another ecosystem. What one would like to do is to easily interchange pieces of a particular workflow to assess how much difference each piece makes. For example, swapping different tools for assigning protein protonation states could yield valuable insights into the relative merits of these tools and the importance of protonation at specific residues, but currently, this is an arduous task.

A. Workflow automation is needed

At the most basic level, we need to allow calculations to be easily repeated on all of the benchmark systems via automated workflows. One should not have to become an expert in the systems being studied in order to be able to successfully apply calculations to them; inputs should be easily available and repeating calculations should become fully automated so that a new method can be tested by simply specifying the set of benchmarks to run on.

To achieve this, at least two major innovations are needed. First, we need automated workflows that can proceed from the specification of a system to target to yielding the desired results without human intervention. Second, we need a standard data structure for input to and output from these workflows so that people can easily obtain inputs for benchmark systems and only change the component they want to change (such as the force field or system preparation) and leave the other components unchanged so that, in an automated manner, they can focus their testing on only the components they want to test.

B. Analysis automation will also be needed

At the most basic level, we can simply check whether we are getting the expected answer for each calculation

performed, at least for systems where a gold standard result is available. However, this does not provide nearly enough insight. Are the relevant motions being sampled? We need ways to automatically check that we are sampling the right motions, identifying correct binding modes/conformations, and so on, without having to become experts on the specific systems examined. Probably we will need to define ways to automatically specify what order parameters should be monitored to assess for adequate sampling.

C. Modularization will be key

To achieve these goals, we will need to develop or package tools so that they take a set of well specified inputs and provide well specified outputs in an interchangeable way. This may involve containerizing key pieces of workflows such as in Docker [1] containers, and developing standards as to what inputs and outputs are provided to each component of the workflow. Again, the goal of this is to allow components (such as different tools for protein preparation, or different methods for free energy calculation) to be swapped without requiring changes elsewhere in the workflow. Currently, a change anywhere in the workflow often requires changes throughout the entire workflow.

Another key goal of modularization is to separate the *operator* from the *method*. Currently, binding calculations are most often done by a human expert who makes a variety of decisions along the way (though Schrödinger's workflow has gone a reasonable ways towards changing this [121]) and it is difficult to separate the importance of human expertise from the relative merits of the methods employed. Containerizing and modularization will be key for this, allowing methods to be employed only in a well-defined way which is reproducible. It is this type of science – coupled with benchmark tests – which is needed to advance the field.

VI. THE FUTURE OF BENCHMARKS AND OF THIS REVIEW

This work has so far presented a small set of benchmark systems for binding free energy calculations, and has highlighted some of the ways in which they have already proven their utility. However, the scope of these sets is still quite limited. More, increasingly diverse, host-guest systems will help probe the strengths and weaknesses of force fields, and to drive their improvement. At the other end of the spectrum, we need more complex and challenging benchmark sets for proteins including simple models, like T4 lysozyme as well as candidate drug targets. And there may be community interest in test systems specifically selected to challenge sampling algorithms, without reference to experimental data.

Several candidate hosts and proteins are worth mentioning in this regard. Among host-guest systems, there is a particularly extensive experimental literature on cyclodextrins [44, 94], and they are tractable computationally [48, 124]. As to artificial protein binding sites, the two variants of the CCP protein model binding site [5, 6, 32, 89, 96, 99] offer a modest increase in difficulty relative to the T4 lysozyme sites discussed above. And thrombin and the bromodomains appear to be promising examples of candidate drug targets for inclusion in a growing set of benchmark systems. Thrombin is a serine protease that has received prior attention from free energy studies [13, 120, 121]. Experimental data exhibits interesting trends [7] that can partly be explained by simulations [13]; but challenges remain [12]. Bromodomains may also be interesting, especially given that relatively high accuracies have been reported, relative to experiment. At the same time, binding modes may be non-obvious and the diversity of ligands could pose problems for relative free energy calculations [2]. Other systems will undoubtedly emerge as promising benchmarks as well, and we seek community input to help identify these.

In order to provide for updates of this material as new benchmark systems are defined, and to enable community input into the process of choosing them, we will make the LaTeX source for this article on GitHub at `http://www.github.com/mobleylab/benchmarksets`. We encourage use of the issue tracker for discussion, comments, and proposed updates. We plan to incorporate new material via GitHub as one would for a coding project, then make it available via a preprint server, likely bioRxiv. Given substantial changes to this initial version of the paper, it may ultimately be appropriate to make it available as a "perpetual review" [78] via another forum allowing versioned updates of publications.

VII. CONCLUSIONS AND OUTLOOK

Binding free energy calculations are a promising tool for predicting and understanding molecular interactions and appear to have enough accuracy to provide substantial benefits in a pharmaceutical drug discovery context. However, progress is needed to improve these tools so that they can achieve their potential. To achieve steady progress, and to avoid potentially damaging cycles of enthusiasm and disillusionment, we need to understand and be open and honest about key challenges. Benchmarks are vital for this, as they allow researchers in the field to rigorously test their methods, arrive at a shared understanding of problems, and measure progress on well-characterized yet challenging systems. It is also worth emphasizing the importance of sharing information about apparently well thought-out and even promising methods that do *not* work, rather than sharing only what does appear to work. Identifying and addressing failure cases and problems is critically important to advancing

this technology, but failures can be harder to publish, and may even go unpublished, even though they serve a unique role in advancing the field. We therefore strongly encourage that such results be shared and welcomed by the research community.

Here, we proposed several benchmark systems for binding free energy calculations. These embody a subset of the key challenges facing the field, and we plan to expand the set as consensus emerges. Hopefully, these systems will serve as challenging standard test cases for new methods, force fields, protocols, and workflows. Our desire is that these benchmarks will advance the science and technology of modeling and predicting molecular interactions, and that other researchers in the field will contribute to identifying new benchmark sets and updating the information provided about these informative systems.

DISCLOSURE STATEMENT

D.L.M. is a member of the Scientific Advisory Board for Schrödinger, LLC. M.K.G. is a cofounder and has equity interest in the company VeraChem LLC.

ACKNOWLEDGMENTS

DLM appreciates financial support from the National Institutes of Health (NIH; 1R01GM108889-01) and the National Science Foundation (NSF; CHE 1352608). MKG thanks the NIH for partial support of this work through grant R01GM061300. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or the NSF.

We also appreciate helpful discussions with a huge number of people in the field, including a wide variety of participants at recent meetings such as the 2016 Workshop on Free Energy Methods in Drug Discovery. Conversations with John Chodera (MSKCC), Chris Oostenbrink (BOKU), Julien Michel (Edinburgh), Robert Abel (Schrödinger), Bruce Gibb (Tulane), Matt Sullivan (Tulane), and Lyle Isaacs (Maryland) were particularly helpful.

[1] What is Docker? https://www.docker.com/what-docker, 2015-05-14T16:17:40-07:00.

[2] M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knapp, and P. C. Biggin. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.*, 7(1):207–218, 2016.

[3] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford Science Publications. Oxford University Press, New York, NY, June 1989.

[4] K. I. Assaf and W. M. Nau. Cucurbiturils: From synthesis to high-affinity binding and catalysis. *Chem Soc Rev*, 44(2):394–418, Jan. 2015.

[5] S. Banba and C. L. Brooks III. Free energy screening of small ligands binding to an artificial protein cavity. *The Journal of Chemical Physics*, 113(8):3423–3433, Aug. 2000.

[6] S. Banba, Z. Guo, and C. L. Brooks III. Efficient Sampling of Ligand Orientations and Conformations in Free Energy Calculations Using the λ-Dynamics Method. *J. Phys. Chem. B*, 104(29):6903–6910, July 2000.

[7] B. Baum, L. Muley, M. Smolinski, A. Heine, D. Hangauer, and G. Klebe. Non-additivity of Functional Group Contributions in Protein–Ligand Binding: A Comprehensive Study by Crystallography and Isothermal Titration Calorimetry. *Journal of Molecular Biology*, 397(4):1042–1054, Apr. 2010.

[8] S. Bhakat and P. Söderhjelm. 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*, 2016.

[9] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *The Journal of Physical Chemistry B*, 107(35):9535–9551, Sept. 2003.

[10] S. Bosisio, A. S. J. S. Mey, and J. Michel. Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *J Comput Aided Mol Des*, 2016.

[11] S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill, and B. K. Shoichet. Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.*, 394(4):747–763, Dec. 2009.

[12] G. Calabrò. Accelerating molecular simulations implication for rational drug design. Nov. 2015.

[13] G. Calabrò, C. J. Woods, F. Powlesland, A. S. J. S. Mey, A. J. Mulholland, and J. Michel. Elucidation of Nonadditive Effects in Protein–Ligand Binding Energies: Thrombin as a Case Study. *J. Phys. Chem. B*, June 2016.

[14] L. Cao, M. Šekutor, P. Y. Zavalij, K. Mlinarić-Majerski, R. Glaser, and L. Isaacs. Cucurbit[7]uril·Guest Pair with an Attomolar Dissociation Constant. *Angew. Chem. Int. Ed.*, 53(4):988–993, Jan. 2014.

[15] R. S. Carnegie, C. L. D. Gibb, and B. C. Gibb. Anion Complexation and The Hofmeister Effect. *Angew. Chem.*, 126(43):11682–11684, Oct. 2014.

[16] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande. Alchemical free energy methods for drug discovery: Progress and challenges. *Curr Opin Struct Biol*, 21(2):150–160, Feb. 2011.

[17] C. D. Christ. Binding affinity prediction from molecular simulations: A new standard method in structure-based drug design?, May 2016.

[18] C. D. Christ and T. Fox. Accuracy Assessment and Automation of Free Energy Calculations for Drug Design. *J. Chem. Inf. Model.*, 54(1):108–120, Jan. 2014.

[19] C. D. Christ, A. E. Mark, and W. F. van Gunsteren. Basic ingredients of free energy calculations: A review. *J. Comput. Chem.*, 31(8):1569–1582, June 2010.

[20] D. J. Cole, J. Tirado-Rives, and W. L. Jorgensen. Molecular dynamics and Monte Carlo simulations for protein–ligand binding and inhibitor design. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1850(5):966–971, May 2015.

[21] J. Comer, K. Schulten, and C. Chipot. Calculation of Lipid-Bilayer Permeabilities Using an Average Force. *J Chem. Theory Comput.*, 10(2):554–564, Feb. 2014.

[22] H. Cong, X. L. Ni, X. Xiao, Y. Huang, Q.-J. Zhu, S.-F. Xue, Z. Tao, L. F. Lindoy, and G. Wei. Synthesis and separation of cucurbit[n]urils and their derivatives. *Org. Biomol. Chem.*, 14(19):4335–4364, May 2016.

[23] G. Cui. Affinity Predictions with FEP+: A Different Perspective on Performance and Utility, May 2016.

[24] A. de Ruiter and C. Oostenbrink. Protein–Ligand Binding from Distancefield Distances and Hamiltonian Replica Exchange Simulations. *J. Chem. Theory Comput.*, 9(2):883–892, Feb. 2013.

[25] N. Deng, S. Forli, P. He, A. Perryman, L. Wickstrom, R. S. K. Vijayan, T. Tiefenbrunn, D. Stout, E. Gallicchio, A. J. Olson, and R. M. Levy. Distinguishing Binders from False Positives by Free Energy Calculations: Fragment Screening Against the Flap Site of HIV Protease. *J. Phys. Chem. B*, 119(3):976–988, Jan. 2015.

[26] Y. Deng and B. Roux. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *Journal of Chemical Theory and Computation*, 2(5):1255–1273, Sept. 2006.

[27] J. B. Dunbar, R. D. Smith, C.-Y. Yang, P. M.-U. Ung, K. W. Lexa, N. A. Khazanov, J. A. Stuckey, S. Wang, and H. A. Carlson. CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. *J. Chem. Inf. Model.*, 51(9):2036–2046, Sept. 2011.

[28] A. Eklund, T. E. Nichols, and H. Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.*, 113(28):7900–7905, July 2016.

[29] A. E. Eriksson, W. A. Baase, J. A. Wozniak, and B. W. Matthews. A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. *Nature*, 355(6358):371–373, Jan. 1992.

[30] J. Ewell, B. C. Gibb, and S. W. Rick. Water Inside a Hydrophobic Cavitand Molecule. *The Journal of Physical Chemistry B*, 112(33):10272–10279, Aug. 2008.

[31] A. T. Fenley, N. M. Henriksen, H. S. Muddana, and M. K. Gilson. Bridging Calorimetry and Simulation through Precise Calculations of Cucurbituril–Guest Binding Enthalpies. *Journal of Chemical Theory and Computation*, 10(9):4069–4078, Sept. 2014.

[32] M. M. Fitzgerald, R. A. Musah, D. E. McRee, and D. B. Goodin. A ligand-gated, hinged loop rearrangement opens a channel to a buried artificial protein cavity. *Nat Struct Mol Biol*, 3(7):626–631, July 1996.

[33] H. Flyvbjerg and H. G. Petersen. Error estimates on averages of correlated data. *The Journal of Chemical Physics*, 91(1):461, July 1989.

[34] W. A. Freeman, W. L. Mock, and N. Y. Shih. Cucurbituril. *J. Am. Chem. Soc.*, 103(24):7367–7368, Dec. 1981.

[35] E. Gallicchio, M. Lapelosa, and R. M. Levy. Binding Energy Distribution Analysis Method (BEDAM) for Estimation of Protein-Ligand Binding Affinities. *Journal of Chemical Theory and Computation*, 6(9):2961–2977, Sept. 2010.

[36] H. Gan, C. J. Benjamin, and B. C. Gibb. Nonmonotonic Assembly of a Deep-Cavity Cavitand. *Journal of the American Chemical Society*, 133(13):4770–4773, Apr. 2011.

[37] K. Gao, J. Yin, N. M. Henriksen, A. T. Fenley, and M. K. Gilson. Binding Enthalpy Calculations for a Neutral Host–Guest Pair Yield Widely Divergent Salt Effects across Water Models. *Journal of Chemical Theory and Computation*, 11(10):4555–4564, Oct. 2015.

[38] S. Gathiaka, S. Liu, M. Chiu, H. Yang, J. Stuckey, Y. Kang, J. Delproposto, G. Kubish, J. Dunbar, H. Carlson, S. Burley, W. Walters, R. Amaro, V. Feher, and M. Gilson. D3R Grand Challenge 2015: Evaluation of Protein-Ligand Pose and Affinity Predictions. *J. Comput. Aided Mol. Des.*, (In press), 2016.

[39] C. L. D. Gibb and B. C. Gibb. Well-Defined, Organic Nanoenvironments in Water: The Hydrophobic Effect Drives a Capsular Assembly. *J. Am. Chem. Soc.*, 126(37):11408–11409, Sept. 2004.

[40] C. L. D. Gibb and B. C. Gibb. Guests of differing polarities provide insight into structural requirements for templates of water-soluble nano-capsules. *Tetrahedron*, 65(35):7240–7248, Aug. 2009.

[41] C. L. D. Gibb and B. C. Gibb. 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, May 2011.

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

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

[44] L. A. Godínez, L. Schwartz, C. M. Criss, and A. E. Kaifer. Thermodynamic Studies on the Cyclodextrin Complexation of Aromatic and Aliphatic Guests in Water and Water-Urea Mixtures. Experimental Evidence for the Interaction of Urea with Arene Surfaces. *J. Phys. Chem. B*, 101(17):3376–3380, Apr. 1997.

[45] A. P. Graves, R. Brenk, and B. K. Shoichet. Decoys for Docking. *Journal of Medicinal Chemistry*, 48(11):3714–3728, June 2005.

[46] A. P. Graves, D. M. Shivakumar, S. E. Boyce, M. P. Jacobson, D. A. Case, and B. K. Shoichet. Rescoring Docking Hit Lists for Model Cavity Sites: Predictions and Experimental Testing. *Journal of Molecular Biology*, 377(3):914–934, Mar. 2008.

[47] J. C. Gumbart, B. Roux, and C. Chipot. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.*, 9(1):794–802, Jan. 2013.

[48] N. M. Henriksen, A. T. Fenley, and M. K. Gilson. Computational Calorimetry: High-Precision Calculation of Host–Guest Binding Thermodynamics. *Journal of Chemical Theory and Computation*, 11(9):4377–4394, Sept. 2015.

[49] J. Hermans and S. Subramaniam. The free energy of xenon binding to myoglobin from molecular dynamics simulation. *Isr. J. Chem.*, 27:225–227, Jan. 1986.

[50] M. B. Hillyer, C. L. D. Gibb, P. Sokkalingam, J. H. Jordan, S. E. Ioup, and B. C. Gibb. Synthesis of Water-Soluble Deep-Cavity Cavitands. *Org. Lett.*, 18(16):4048–4051, Aug. 2016.

[51] N. Homeyer, F. Stoll, A. Hillisch, and H. Gohlke. Binding Free Energy Calculations for Lead Optimization: Assessment of Their Accuracy in an Industrial Drug Design Context. *J. Chem. Theory Comput.*, 10(8):3331–3344, Aug. 2014.

[52] Y.-W. Hsiao and P. Söderhjelm. Prediction of SAMPL4 host–guest binding affinities using funnel metadynamics. *J Comput Aided Mol Des*, 28(4):443–454, Feb. 2014.

[53] L. Isaacs. Personal communication, Sept. 2016.

[54] W. Jiang and B. Roux. Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *Journal of Chemical Theory and Computation*, 6(9):2559–2565, Sept. 2010.

[55] W. L. Jorgensen. Quantum and statistical mechanical studies of liquids. 12. Simulation of liquid ethanol including internal rotation. *Journal of the American Chemical Society*, 103(2):345–350, Jan. 1981.

[56] W. L. Jorgensen, J. K. Buckner, S. Boudon, and J. Tirado-Rives. Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *The Journal of Chemical Physics*, 89(6):3742–3746, Sept. 1988.

[57] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat Struct Mol Biol*, 9(9):646–652, Sept. 2002.

[58] I. V. Khavrutskii and A. Wallqvist. Improved Binding Free Energy Predictions from Single-Reference Thermodynamic Integration Augmented with Hamiltonian Replica Exchange. *Journal of Chemical Theory and Computation*, 7(9):3001–3011, Sept. 2011.

[59] C. T. Lee, J. Comer, C. Herndon, N. Leung, A. Pavlova, R. V. Swift, C. Tung, C. N. Rowley, R. E. Amaro, C. Chipot, Y. Wang, and J. C. Gumbart. Simulation-Based Approaches for Determining Membrane Permeability of Small Compounds. *J. Chem. Inf. Model.*, 56(4):721–733, Apr. 2016.

[60] J. W. Lee, H. H. L. Lee, Y. H. Ko, K. Kim, and H. I. Kim. Deciphering the Specific High-Affinity Binding of Cucurbit[7]uril to Amino Acids in Water. *The Journal of Physical Chemistry B*, 119(13):4628–4636, Apr. 2015.

[61] J. W. Lee, S. Samal, N. Selvapalam, H.-J. Kim, and K. Kim. Cucurbituril homologues and derivatives: New opportunities in supramolecular chemistry. *Acc. Chem. Res.*, 36(8):621–630, Aug. 2003.

[62] M. S. Lee and M. A. Olson. Calculation of Absolute Protein-Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophysical Journal*, 90(3):864–877, Feb. 2006.

[63] G. Leonis, T. Steinbrecher, and M. G. Papadopoulos. A Contribution to the Drug Resistance Mechanism of Darunavir, Amprenavir, Indinavir, and Saquinavir Complexes with HIV-1 Protease Due to Flap Mutation I50V: A Systematic MM–PBSA and Thermodynamic Integration Study. *J. Chem. Inf. Model.*, 53(8):2141–2153, Aug. 2013.

[64] N. M. Lim, L. Wang, R. Abel, and D. L. Mobley. Sensitivity in binding free energies due to protein reorganization. *Journal of Chemical Theory and Computation*, July 2016.

[65] Y.-L. Lin, A. Aleksandrov, T. Simonson, and B. Roux. An Overview of Electrostatic Free Energy Computations for Solutions and Proteins. *J. Chem. Theory Comput.*, 10(7):2690–2709, July 2014.

[66] S. Liu, S. Cao, K. Hoang, K. L. Young, A. S. Paluch, and D. L. Mobley. Using MD Simulations To Calculate How Solvents Modulate Solubility. *Journal of Chemical Theory and Computation*, 12(4):1930–1941, Feb. 2016.

[67] S. Liu, C. Ruspic, P. Mukhopadhyay, S. Chakrabarti, P. Y. Zavalij, and L. Isaacs. The Cucurbit[n]uril Family: Prime Components for Self-Sorting Systems. *Journal of the American Chemical Society*, 127(45):15959–15967, Nov. 2005.

[68] S. Liu, Y. Wu, T. Lin, R. Abel, J. P. Redmann, C. M. Summa, V. R. Jaber, N. M. Lim, and D. L. Mobley. Lead optimization mapper: Automating free energy calculations for lead optimization. *J Comput Aided Mol Des*, 27(9):755–770, Sept. 2013.

[69] M. Merski, M. Fischer, T. E. Balius, O. Eidam, and B. K. Shoichet. Homologous ligands accommodated by discrete conformations of a buried cavity. *PNAS*, 112(16):5039–5044, Apr. 2015.

[70] J. Michel and J. W. Essex. Hit Identification and Binding Mode Predictions by Rigorous Free Energy Simulations. *J. Med. Chem.*, 51(21):6654–6664, Nov. 2008.

[71] J. Michel and J. W. Essex. Prediction of protein–ligand binding affinity by free energy simulations: Assumptions, pitfalls and expectations. *J Comput Aided Mol Des*, 24(8):639–658, May 2010.

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

[73] P. Mikulskis, S. Genheden, and U. Ryde. A Large-Scale Test of Free-Energy Simulation Estimates of Protein–Ligand Binding Affinities. *J. Chem. Inf. Model.*, 54(10):2794–2806, Oct. 2014.

[74] D. L. Mobley, J. D. Chodera, and K. A. Dill. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *J. Chem. Phys.*, 125:084902, Jan. 2006.

[75] D. L. Mobley, J. D. Chodera, and K. A. Dill. Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *Journal of Chemical Theory and Computation*, 3(4):1231–1235, July 2007.

[76] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet, and K. A. Dill. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, 371(4):1118–1134, Aug. 2007.

[77] D. L. Mobley and P. V. Klimovich. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.*, 137(23):230901, Jan. 2012.

[78] D. L. Mobley and D. M. Zuckerman. A proposal for regularly updated review/survey articles: "Perpetual Reviews". *arXiv:1502.01329 [cs]*, Feb. 2015.

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

[80] S. Moghaddam, Y. Inoue, and M. K. Gilson. Host-Guest Complexes with Protein-Ligand-like Affinities: Computational Analysis and Design. *Journal of the American Chemical Society*, 131(11):4012–4021, Mar. 2009.

[81] S. Moghaddam, C. Yang, M. Rekharsky, Y. H. Ko, K. Kim, Y. Inoue, and M. K. Gilson. 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. *Journal of the American Chemical Society*, 133(10):3570–3581, Mar. 2011.

[82] J. I. Monroe and M. R. Shirts. Converging free energies of binding in cucurbit[7]uril and octa-acid host–guest systems from SAMPL4 using expanded ensemble simulations. *J Comput Aided Mol Des*, 28(4):401–415, Mar. 2014.

[83] A. Morton, W. A. Baase, and B. W. Matthews. Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 lysozyme. *Biochemistry*, 34(27):8564–8575, July 1995.

[84] A. Morton and B. W. Matthews. Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: Linkage of dynamics and structural plasticity. *Biochemistry*, 34(27):8576–8588, July 1995.

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

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

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

[88] H. S. Muddana, J. Yin, N. V. Sapra, A. T. Fenley, and M. K. Gilson. Blind prediction of SAMPL4 cucurbit[7]uril binding affinities with the mining minima method. *J Comput Aided Mol Des*, 28(4):463–474, Feb. 2014.

[89] R. A. Musah, G. M. Jensen, S. W. Bunte, R. J. Rosenfeld, and D. B. Goodin. Artificial protein cavities as specific ligand-binding templates: Characterization of an engineered heterocyclic cation-binding site that preserves the evolved specificity of the parent protein1. *Journal of Molecular Biology*, 315(4):845–857, Jan. 2002.

[90] C. N. Nguyen, T. K. Young, and M. K. Gilson. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *The Journal of Chemical Physics*, 137(4):044101, July 2012.

[91] R. K. Pal, K. Haider, D. Kaur, W. Flynn, J. Xia, R. M. Levy, T. Taran, L. Wickstrom, T. Kurtzman, and E. Gallicchio. A combined treatment of hydration and dynamical effects for the modeling of host-guest binding thermodynamics: The SAMPL5 blinded challenge. *Journal of Computer-Aided Molecular Design*, 2016.

[92] J. Park, I. Nessler, B. McClain, D. Macikenas, J. Baltrusaitis, and M. J. Schnieders. Absolute Organic Crystal Thermodynamics: Growth of the Asymmetric Unit into a Crystal via Alchemy. *J. Chem. Theory Comput.*, 10(7):2781–2791, July 2014.

[93] M. M. Reif and C. Oostenbrink. Net charge changes in the calculation of relative ligand-binding free energies via classical atomistic molecular dynamics simulation. *J. Comput. Chem.*, 35(3):227–243, Jan. 2014.

[94] M. V. Rekharsky and Y. Inoue. Complexation Thermodynamics of Cyclodextrins. *Chem. Rev.*, 98(5):1875–1918, July 1998.

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

[96] G. J. Rocklin, S. E. Boyce, M. Fischer, I. Fish, D. L. Mobley, B. K. Shoichet, and K. A. Dill. Blind Prediction of Charged Ligand Binding Affinities in a Model Binding Site. *J. Mol. Biol.*, 425(22):4569–4583, Nov. 2013.

[97] G. J. Rocklin, D. L. Mobley, K. A. Dill, and P. H. Hünenberger. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *J. Chem. Phys.*, 139(18):184103, Jan. 2013.

[98] K. E. Rogers, J. M. Ortiz-Sánchez, R. Baron, M. Fajer, C. A. F. de Oliveira, and J. A. McCammon. On the Role of Dewetting Transitions in Host–Guest Binding Free Energy Calculations. *Journal of Chemical Theory and Computation*, 9(1):46–53, Jan. 2013.

[99] R. J. Rosenfeld, A.-M. A. Hays, R. A. Musah, and D. B. Goodin. Excision of a proposed electron transfer pathway in cytochrome c peroxidase and its replacement by a ligand-binding channel. *Protein Science*, 11(5):1251–1259, May 2002.

[100] M. J. Schnieders, J. Baltrusaitis, Y. Shi, G. Chattree, L. Zheng, W. Yang, and P. Ren. The Structure, Thermodynamics, and Solubility of Organic Crystals from Simulation with a Polarizable Force Field. *J. Chem. Theory Comput.*, 8(5):1721–1736, May 2012.

[101] P. Schreiner. Theoretical prediction of affinities to cucurbiturils –the blind prediction hydrophobe challenge. https://www.uni-giessen.de/fbz/fb08/dispersion/projects/HydrophobeChallenge, 2016.

[102] Sherborne, Bradley. Opening the lid on FEP. *J Comput Aided Mol Des*, 2016.

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

[104] M. R. Shirts, C. Klein, J. M. Swails, J. Yin, M. K. Gilson, D. L. Mobley, D. A. Case, and M. R. Shirts. Lessons learned from comparing molecular dynamics englines on the SAMPL5 dataset. *J Comput Aided Mol Des*, 2016.

[105] M. R. Shirts and D. L. Mobley. An Introduction to Best Practices in Free Energy Calculations. In *Biomolecular Simulations*, volume 924. Methods in Molecular Biology, Jan. 2013.

[106] M. R. Shirts, D. L. Mobley, and S. P. Brown. Free-energy calculations in structure-based drug design. In J. Merz, Kenneth M, D. Ringe, and C. H. Reynolds, editors, *Drug Design: Structure and Ligand-Based Approaches*. Cambridge University Press, Jan. 2010.

[107] T. Simonson and B. Roux. Concepts and protocols for electrostatic free energies. *Molecular Simulation*, 42(13):1090–1101, Sept. 2016.

[108] P. Sokkalingam, J. Shraberg, S. W. Rick, and B. C. Gibb. Binding Hydrated Anions with Hydrophobic Pockets. *Journal of the American Chemical Society*, 138(1):48–51, Jan. 2016.

[109] M. R. Sullivan, P. Sokkalingam, T. Nguyen, J. P. Donahue, and B. C. Gibb. Binding of carboxylate and trimethylammonium salts to octa-acid and TEMOA deep-cavity cavitands. *J Comput Aided Mol Des*, pages 1–8, July 2016.

[110] H. Sun, C. L. D. Gibb, and B. C. Gibb. Calorimetric Analysis of the 1:1 Complexes Formed between a Water-soluble Deep-cavity Cavitand, and Cyclic and Acyclic Carboxylic Acids. *Supramolecular Chemistry*, 20(1-2):141–147, Jan. 2008.

[111] K. Tai. Conformational sampling for the impatient. *Biophysical Chemistry*, 107(3):213–220, Feb. 2004.

[112] B. L. Tembe and J. A. McCammon. Ligand Receptor Interactions. *Comput Chem*, 8(4):281–283, Jan. 1984.

[113] F. Tofoleanu, J. Lee, F. C. Pickard IV., G. König, J. Huang, M. Baek, C. Seok, and B. R. Brooks. Absolute binding free energy calculations for octa-acids and guests. *J Comput Aided Mol Des*, 2016.

[114] C. Velez-Vega and M. K. Gilson. Force and Stress along Simulated Dissociation Pathways of Cucurbituril–Guest Systems. *J. Chem. Theory Comput.*, 8(3):966–976, Mar. 2012.

[115] C. Velez-Vega and M. K. Gilson. Overcoming dissipation in the calculation of standard binding free energies by ligand extraction. *J. Comput. Chem.*, 34(27):2360–2371, Oct. 2013.

[116] A. Verras. Free Energy Perturbation at Merck: Benchmarking against Faster Methods, May 2016.

[117] B. Vinciguerra, P. Y. Zavalij, and L. Isaacs. Synthesis and Recognition Properties of Cucurbit[8]uril Derivatives. *Org. Lett.*, 17(20):5068–5071, Oct. 2015.

[118] K. Wang, J. D. Chodera, Y. Yang, and M. R. Shirts. Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics. *J Comput Aided Mol Des*, 27(12):989–1007, Dec. 2013.

[119] K. Wang, P. Sokkalingam, and B. C. Gibb. ITC and NMR analysis of the encapsulation of fatty acids within a water-soluble cavitand and its dimeric capsule. *Supramolecular Chemistry*, 28(1-2):84–90, Feb. 2016.

[120] L. Wang, B. J. Berne, and R. A. Friesner. On achieving high accuracy and reliability in the calculation of relative protein–ligand binding affinities. *PNAS*, 109(6):1937–1942, July 2012.

[121] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner, and R. Abel. 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, Feb. 2015.

[122] B. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet. A Model Binding Site for Testing Scoring Functions in Molecular Docking. *Journal of Molecular Biology*, 322(2):339–355, Sept. 2002.

[123] B. Q. Wei, L. H. Weaver, A. M. Ferrari, B. W. Matthews, and B. K. Shoichet. Testing a Flexible-receptor Docking Algorithm in a Model Binding Site. *Journal of Molecular Biology*, 337(5):1161–1182, Apr. 2004.

[124] L. Wickstrom, N. Deng, P. He, A. Mentes, C. Nguyen, M. K. Gilson, T. Kurtzman, E. Gallicchio, and R. M. Levy. Parameterization of an effective potential for protein–ligand binding from host–guest affinity data. *J. Mol. Recognit.*, 29(1):10–21, Jan. 2016.

[125] H.-J. Woo and B. Roux. Calculation of absolute protein–ligand binding free energy from computer simulations. *PNAS*, 102(19):6825–6830, Oct. 2005.

[126] I. W. Wyman and D. H. Macartney. Cucurbit[7]uril host–guest complexes with small polar organic guests in aqueous solution. *Organic & Biomolecular Chemistry*, 6(10):1796, 2008.

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

[128] J. Yin, N. M. Henriksen, D. R. Slochower, M. W. Chiu, D. L. Mobley, and M. K. Gilson. Overview of the SAMPL5 Host-Guest Challenge: Are We Doing Better? *J Comput Aided Mol Des*, 2016.

[129] J. Yin, N. M. Henriksen, D. R. Slochower, and M. K. Gilson. The SAMPL5 Host-Guest Challenge: Binding Free Energies and Enthalpies from Explicit Solvent Simulations. *J Comput Aided Mol Des*, 2016.

[130] F. M. Ytreberg. Absolute FKBP binding affinities obtained via nonequilibrium unbinding simulations. *The Journal of Chemical Physics*, 130(16):164906, Apr. 2009.