Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics

Kai Wang · John D. Chodera · Yanzhi Yang ·
Michael R. Shirts

Abstract We present a method to identify small molecule ligand binding sites and poses within a given protein crystal structure using GPU-accelerated Hamiltonian replica exchange molecular dynamics simulations. The Hamiltonians used vary from the physical end state of protein interacting with the ligand to an unphysical end state where the ligand does not interact with the protein. As replicas explore the space of Hamiltonians interpolating between these states, the ligand can rapidly escape local minima and explore potential binding sites. Geometric restraints keep the ligands from leaving the vicinity of the protein and an alchemical pathway designed to increase phase space overlap between intermediates ensures good mixing. Because of the rigorous statistical mechanical nature of the Hamiltonian exchange framework, we can also extract binding free energy estimates for all putative binding sites. We present results of this methodology applied to the T4 lysozyme L99A model system for three known ligands and one non-binder as a control, using an implicit solvent. We find that our methodology identifies known crystallographic binding sites consistently and accurately for the small number of ligands considered here and gives free energies consistent with experiment. We are also able to analyze the contribution of individual binding sites to the overall binding affinity. Our methodology points to near term potential applications in early-stage structure-guided drug discovery.

K. Wang · Y. Yang · M. R. Shirts (✉)
Department of Chemical Engineering, University of Virginia, Charlottesville, VA, USA
e-mail: michael.shirts@virginia.edu

J. D. Chodera
Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

Introduction

Determining small molecule binding sites and bound poses is an important part of the drug discovery process. When the co-crystal structure of a lead compound is unavailable, rationalizing affinity changes in a lead compound series and designing molecules with improved affinity can prove challenging. Even when the binding site is known, additional sites with varying druggability may exist, and targeting these alternative sites may produce desirable biological responses and hence provide new opportunities for drug discovery.

With rapid development in processing power and molecular simulation algorithms, computational methods are now playing an important role in predicting protein-ligand binding properties, especially in early-stage drug discovery. Docking methods, the most widely used class of structure-based drug design methods, aim to rapidly generate a comprehensive set of conformations of the protein-ligand complex and rank them using scoring functions of varying complexity and accuracy. Though docking methods can quickly rank and often identify binding sites and poses, the accuracy of docking is limited by a number of factors, including the effectiveness of semi-empirical scoring functions, the difficulty of including solvation

effects, and the problem of representing a statistical mechanical ensemble with one or a few configurations. Docking is therefore problematic in projects requiring detailed and reliable knowledge of location of ligand binding in the binding pocket and its interactions with the target in the binding site [1].

A number of studies have worked to fix many of these issues. Some studies have successfully improved docking methodologies by introducing receptor flexibility [2], explicit water molecules [3], or even using post-docking methods to rescore the ensemble of docked structures [4, 5]. Nevertheless, as shown by studies evaluating and comparing different docking programs, their intrinsic limitations, such as a low level of physical detail and lack of statistical mechanical considerations, make them unable to consistently identify ligand binding sites and poses [6–8]. Other structure-based drug design methods that are specifically designed for identifying binding sites based on geometric properties [9–11] or that are knowledge-based [12–14] have also been used with varying success, but these methods are only useful when the binding sites are well-defined pockets. Moreover, extensive usage of fitted models and parameters makes them less generalizable to systems for which they were not parameterized.

In contrast with cheap but approximate docking methods are more rigorous, physics-based techniques such as molecular dynamics (MD) and Monte Carlo (MC) simulations, which historically have found much less use in the drug design process because of their expense. With an all-atom representation of the protein and explicit or implicit representations of solvent, MD simulations can provide microscopic information about protein-ligand interactions, predict and calculate properties based on statistical averages of an ensemble of conformations, and have been shown to be capable of accurately predicting binding affinities in model systems [15–17]. In theory, MD simulations of a protein with a known ligand will eventually converge to the true distribution of bound structures if run sufficiently (though impractically) long with an accurate force field. Free energy calculation methods [18] can then in principle be used to either decide between the predicted poses or compare the results with experimental data.

In reality, optimizing these simulation tools individually and assembling them together to produce useful predictions on a timeline consistent with a realistic drug discovery project is still an unsolved problem. The rapid development of computer power and techniques such as GPU-accelerated simulations [19, 20], increasingly accurate biomolecular force fields [21–23], implicit solvent models [24–26], and simulation machines designed specifically for MD simulations [27, 28] have made these problems much more amenable to computation, but many issues must still be

addressed to enable simulations of sufficient accuracy to be useful in drug design or discovery.

Among these issues, poor or insufficient sampling is undoubtedly the most stubborn one [29]. A ligand in an MD simulation can easily become kinetically trapped for long periods of time, effectively preventing it from visiting the relevant parts of conformational space. This leads to incorrect sampling of the ensemble and results in computed binding affinities or observed binding modes that are sensitive to the initial configuration. In fact, without adequate sampling, even a perfect force field would be of limited use. As argued by Mobley [29] in a recent review, we are still running unconverged simulations with important unsampled configurations on a daily basis, hoping that the unsampled configurations are not essential to ligand binding or other events of interest. Overcoming this sampling problem could lead to direct use of more physical methods to understand and predict small molecule binding.

Because of these computational limitations, knowledge of the binding site is usually a prerequisite in standard ligand binding free energy calculation methods. A crystal structure of a related small molecule or, alternatively, a putative initial structure generated by docking tools is often used as the starting configuration to increase the likelihood that the free energy calculations can at least converge within the binding site in the simulation time available. But could these methods ever practically be used to identify binding sites and poses both rapidly and accurately without prior knowledge of the binding site? A number of docking-based tools and structure-guided drug design methods can sample putative binding sites to generate a putative ensemble of bound conformations [30], but in many cases the emphasis on making the process fast discards the physics required to obtain a properly weighted ensemble that would provide critical information about which sites are populated to which degree.

In this study, we investigate whether sufficiently optimized accelerated MD simulations in implicit solvent can discover binding sites and poses without prior knowledge of the binding site, even in a highly buried binding pocket. Many studies have investigated enhanced sampling methods for accelerating the rate at which MD can sample relevant conformations [31–37], and we focus specifically on Hamiltonian replica exchange molecular dynamics (HREMD) in this paper. In HREMD methods, individual replicas can visit a range of predefined Hamiltonians during the course of a simulation, with exchanges between pairs of replicas accepted according to a modified Metropolis criterion to ensure the equilibrium distribution is preserved for each Hamiltonian. Because kinetic barriers can vary drastically among Hamiltonians, correlation times can be reduced as replicas perform a random walk in Hamiltonian space.

HREMD has been shown to improve sampling in free energy calculations over the use of independent simulations at fixed Hamiltonians [32]. However, because of the large gap between the time scale that current computers can achieve and the time scale of most relevant biomolecular motions, we must further optimize HREMD [38] or combine it with other enhanced sampling methods to fully explore the biophysical configurations of interest in protein-ligand binding. In this study, we accelerate sampling beyond that which can typically be achieved by HREMD, without sacrificing thermodynamic accuracy, using a number of methods. Specifically, we employ flat-bottom restraints to keep the ligand near the protein, make use of multiple coupled and uncoupled states, incorporate Monte Carlo simulation techniques, and use GPU-accelerated molecular dynamics with the OpenMM toolkit [20, 39]. A number of other less conceptually central sampling enhancements are also incorporated as discussed below. Because of the rigorous statistical mechanical nature of the Hamiltonian replica exchange framework, we can also extract binding free energy estimates at all putative binding sites using the multistate Bennett acceptance ratio (MBAR) algorithm [40].

We note that the methodology presented here has many similarities to the Binding Energy Distribution Analysis Method (BEDAM) of Gallicchio et al. [41], in which Hamiltonian replica exchange in an implicit solvent system is used to enhance sampling. However, in our case no binding site is assumed, the Hamiltonian is designed to explicitly maximize phase space overlap between replicas, and no restraints are placed on the protein.

To test the methodology presented in this paper, we examine a model protein-ligand binding system consisting of the engineered L99A mutant of T4 lysozyme and a series of small aromatic ligands. This model system has been widely used by a number of researchers to test the accuracy of free energy methods [15, 16, 42]. T4 lysozyme L99A has a small, buried, hydrophobic internal pocket that has proven to be a difficult target for a number of docking methods [43–46]. Importantly, the crystallographic binding structures and binding free energies are well characterized for this system, allowing us to directly validate our methodology against experiment.

Theory and computational methods

System preparation

Protein parameterization: The T4 lysozyme L99A benzene-bound structure (PDB accession code 181L) was used for this study. The protein was parameterized with the AMBER parm96 forcefield [23] using LEaP from AmberTools 11 [47] (with the force field chosen to be consistent with previous studies of this system [15]).

Ligand parameterization: Ligand structures were created from IUPAC names using the OpenEye OEChem toolkit (version 2.3.2). Mobley et al. [48] have shown that the bond-corrected semi-empirical quantum mechanical AM1-BCC charge model [49, 50] for small molecules works almost as well as ab initio methods in calculating binding free energies for implicit systems. This treatment was used to derive charges for the ligand, and the other parameters were assigned from the AMBER GAFF force field [23, 51] using Antechamber [52].

Docking

To compare the performance of traditional docking methods and our methodology, AutoDock 4.2 was used to dock the same four ligands to the protein [53, 54]. Each ligand was docked twice, once with an entirely rigid protein and once with the same rigid protein except for three flexible residues, Val111, Val103 and Leu118. The three flexible residues were chosen based on their reorientation observed in X-ray structures in response to ligand binding previously reported [15]. All docking was performed to the same PDB structure 181L, the co-crystal structure of the L99A mutant with benzene. The protein for rigid and flexible docking was prepared according to standard AutoDockTools procedures, hydrogens were added to the original files and Gasteiger partial charges were assigned. The AutoDock default grid spacing was used, with the grid box sizes for all docking set to be the box size, which effectively covers the entire protein volume. The number of genetic algorithm runs was set at 50, resulting in 50 final poses.

We note that this docking setup is only partially blind, as the bound structure used is the actual crystal structure for one of the four ligands, so there is some degree of preorganization of the docking site for a bound ligand. Additionally, in the case of flexible docking, only residues which are known to potentially move in alternate crystal structures were made flexible. This therefore represents in many ways a best case scenario for docking.

Simulation methodology

The HREMD-based simulations utilized a modified version of the open-source Python alchemical free energy code YANK (https://github.com/choderalab/yank), which is built on the OpenMM GPU-accelerated molecular simulation library [20, 39]. We performed our simulations using a generalized Born (GB) implicit representation of water [25]. A Langevin dynamics integrator with a 2 fs time step and a 0.5 ps^{-1} collision frequency was used, with a bath temperature of 298 K, and bond lengths to hydrogen were

Fig. 1 Protein system and small molecule ligands used in this study. The T4 lysozyme L99A mutant (PDB 181L) and four small-molecule ligands (of which is one a non-binder) were examined. The ligand atoms closest to the molecular centroids, *circled in red*, were used to define the location of the ligand in subsequent analysis
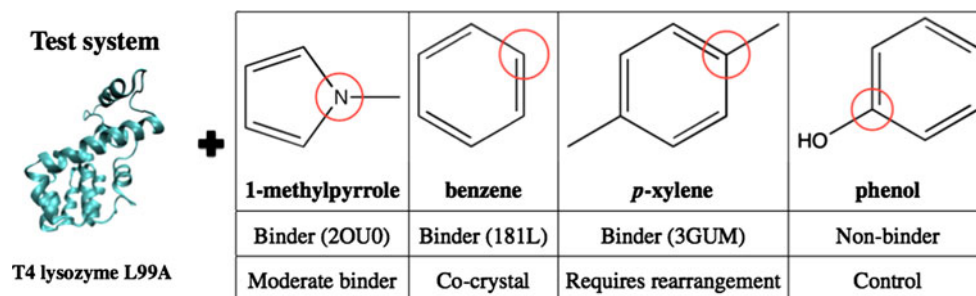
constrained by the CCMA method [55]. A flat-bottom restraint was implemented to keep the ligand in the vicinity of the protein while allowing it to sample in an unbiased way all spatially available and physically reasonable conformational space consistent with binding. The specific choices made for this potential are described below. Hamiltonian replica exchange [32] was used to improve sampling, along with a number of improvements described below. Simulations were run on GPU computing resources provided by XSEDE, including the NCSA Forge and Lincoln clusters.

All preliminary tests of simulation parameters and the tenfold replicate test of simulation consistency were performed with 1-methylpyrrole, a known binder. The ability of our approach to differentiate binders from non-binders was then examined by introducing another three ligands: benzene, a small binder; *p*-xylene, a larger binder which requires conformational change in Val111 upon binding; and phenol, a nonbinder, as a control [15]. By using *p*-xylene, the ability of the method to sample relevant biomolecular motions of the protein can be examined.

The system used in our simulations is shown in Fig. 1. With sufficient sampling of all relevant binding conformations, the simulations here can also be used to estimate protein-ligand free energy of binding. For this purpose, we additionally performed HREMD simulations of the ligand alone, in implicit solvent, with the same parameters as described above.

Flat-bottomed restraint: It is common in free energy calculations to employ restraints to keep the ligand close to the putative binding site, especially in alchemical states where the ligand has weakened interactions with the protein [56, 57]. In our case, we use the tendency of the uncoupled ligand to wander to our advantage in order to identify new binding sites. A restraint to a single binding site would defeat this objective. However, we still wish to keep the ligand near the protein, as the time the ligand spends in the solvent is not of interest, and without periodic boundary conditions the ligand could drift away indefinitely. We therefore used a flat-bottomed restraint to keep the ligand close to the protein. The restraint potential is zero inside a cutoff radius (r_0) with harmonic restraining walls outside of this radius, using the equation:

$$U(r) = \begin{cases} 0 & \text{if } r \leq r_0 \\ \frac{1}{2}k(r - r_0)^2 & \text{if } r > r_0 \end{cases} \quad (1)$$

where $U(r)$ is the restraining potential, k is the spring constant, r is the distance between the protein and ligand centers of geometry, and r_0 is the cutoff radius.

We set r_0 at half the maximum distance between protein atoms plus a 5 Å buffer so that the entire protein with a buffer zone for surface binding sites was within the cutoff. We set the spring constant $k = 5.92$ kcal/mol/Å^2, such that at 1 Å away from the cutoff, the potential energy rises to $5k_BT$. This minimizes the amount of time the ligand spends away from the protein. In this case, we obtain a cutoff r_0 of 35.34 Å from the center of the protein for this system. This restraint is present regardless of the degree the ligand is coupled to the protein. We validated our flat-bottom restraint and integration scheme for physical consistency as described in the Supplementary Material (section A). In the case of a less spherical protein, the amount of time spent sampling configurations away from the protein surface could be minimized using a more complicated shape such as an ellipsoid with major axes constrained to be oriented along the protein's corresponding major axes.

Hamiltonian replica exchange molecular dynamics (HREMD): In MD simulations of protein-ligand complexes, ligands are highly likely to become kinetically trapped in local minima in the free energy surface, potentially for tens of microseconds [58, 59]. These trapping events prevent the ligands from visiting other potential binding sites. Our proposed solution to this problem is to use Hamiltonian replica exchange molecular dynamics (HREMD) between coupled and uncoupled ligand states along an optimized path of alchemical intermediate states. Typically in HREMD, K replicas of simulations at different intermediates along the coupling pathway are run in parallel, with Monte Carlo exchanges attempted periodically between neighboring replicas. This process can lower correlation times for a particular Hamiltonian state of interest by allowing replicas to visit other Hamiltonian states with shorter correlation times. In our particular implementation, the states simulated are defined as follows, starting with the fully interacting state: charges are first scaled to zero, followed by removing the Lennard–Jones

interactions between ligand and protein through soft-core potentials [60–63], leaving an uncharged molecule decoupled from the protein at the other end state. The state of physical interest is fully coupled state, in which all protein–ligand interactions are turned on. However, by including partially and fully uncoupled states in our simulation we allow the ligand to escape from kinetically trapped states, such as nonspecific binding minima, on the time scale of tens or hundreds of picoseconds rather than microseconds. Here, we use a Langevin integrator, but many other sampling methods that preserve the canonical distribution are possible.

In order to efficiently discover putative ligand binding sites and geometries when such information is unavailable, we made a number of modifications to the standard Hamiltonian replica exchange algorithm and Langevin dynamics [32]. These included the use of Gibbs sampling for replica exchanges, the addition of Monte Carlo translation and rotation moves for the ligand, the initial seeding of replicas with distinct configurations, and the use of multiple coupled and uncoupled states to aid statistics.

Gibbs sampling for replica exchange: Recently, it was shown that replica exchange algorithms can be considered a form of Gibbs sampling, with approaches that speed mixing in the permutation of thermodynamic state indices associated with each replica also speeding overall mixing of the whole simulation Markov chain [38]. We make use of this scheme here by attempting many swaps of randomly selected replica pairs (i, j), using the acceptance criteria described in Eq. 24 of Ref. [38]. We attempt a total of K^5 swaps each iteration, where K is the total number of replicas, to ensure thorough mixing. Thus, instead of only jumping to the nearest neighbors, a given replica can jump to any Hamiltonian, though potentially with low probability. The stationary probability is correctly reproduced. In previous test cases, this increased the rate of sampling between 2 and 100 times, depending on the observables and systems examined, with negligible increase in computational cost [38]. The potential energy matrix of each configuration calculated at all alchemical states is calculated and stored for later MBAR analysis.

Monte Carlo ligand translational/rotational moves: To further enhance conformational sampling, we introduced Monte Carlo translational and rotational moves, carried out immediately prior to dynamics with each iteration of Hamiltonian exchange. For these moves, a random displacement of the ligand atoms is attempted, with the trial displacement in each dimension drawn from a normal distribution with 1 nm standard deviation, and acceptance or rejection determined by the Metropolis criterion. A uniform rotational move is chosen by drawing by generating a uniform quaternion [a uniform element of SO(3)] and computing the corresponding rotation matrix, with

rotations accepted or rejected separately from translation by the Metropolis criterion.

Seeding replicas with independent starting configurations: To eliminate biasing from the starting configuration, we initialized the simulations with random starting configurations in the allowed simulation space at all replicas. We applied random rotational and translational moves to the initial bound configurations of all replicas using the scheme described in the previous section without Metropolization. Translational moves were proposed by generating three random numbers from 0 to 2 nm corresponding to (x, y, z) translation from the initial bound configurations, followed by a rotational move as described above. This starting location was rejected if any atom was less than 3 Å from any protein atom.

Using multiple fully coupled and fully uncoupled states: Standard HREMD uses only one fully coupled state and one fully uncoupled state. We can increase the amount of physically meaningful sampling by using multiple fully coupled states. By also using multiple fully uncoupled states, we increase the chance of a ligand being exchanged into a fully uncoupled state, gaining the ability to move freely around the accessible volume.

In our HREMD simulations, the potential energy can be expressed in terms of two coupling parameters:

$$U(x, \lambda_{elec}, \lambda_{LJ}) = U_0(x) + U_{elec}(x; \lambda_{elec}) + U_{LJ}(x; \lambda_{LJ}) \quad (2)$$

where U_0 is the potential of the system with the noninteracting ligand. U_{elec} and U_{LJ} are the Lennard–Jones and electrostatic potentials. λ_{elec} and $\lambda_{LJ} \in [0, 1]$ are the corresponding coupling parameters. Note that the flat-bottom restraint and the ligand torsion, angle, and bond potentials are fully turned on in all states and therefore part of U_0.

For simulations of the ligand in complex, we use 24 total states, as this number is easily portable between configurations of 6 or 8 GPUs per CPU on the computing clusters simulations were run on. In this study, one iteration is defined as the period in MD time steps between replica exchanges. The MD time step was 2 fs, with 500 time steps between exchanges, making each iteration 1 ps long. Velocities were reassigned from the Maxwell–Boltzmann distribution at the beginning of each iteration to ensure the simulation is maintained in the canonical ensemble. Fewer time steps per iteration allows for more exchanges in state space in a given unit time, and thus for faster transitions of ligands in and out of putative binding sites [64]. However, at some point as exchanges become more frequent there is a tradeoff between the computational overhead required to perform state exchanges and the acceleration of binding transitions due to the exchanges. We ran a series of 1 ns simulations with different numbers of time steps per iteration (250, 500, 1,000, 2,500). We chose 500 steps for our

performance runs, because with 250 MD steps per swap the percentage of time spent performing exchanges was about twice as large as that for 500 and began to be a non-negligible fraction of the total simulation time. The total time taken was independent of whether Gibbs sampling or standard Metropolis neighbor exchange was performed. The particular tradeoffs involved in choosing this exchange frequency are highly sensitive to the particular CUDA implementation and the networking details of the computers on which simulations are run, and should not be taken as definitive for all hardware or software configurations.

We performed a series of runs using a beta version of the code to examine the sensitivity of the simulation efficiency on simulation parameters, including the number and spacing of intermediate states, the number of additional fully coupled and fully uncoupled states, and the size of the Monte Carlo displacements. The results showed that other than having sufficiently close spacing of intermediate states in λ space, sampling was not very sensitive to these simulation parameters, and thus no attempt at extensive optimization was made. A table of simulation parameters tested is included as Supplementary Material (section B).

The ligand was alchemically decoupled from the rest of the system through a series of discharging intermediates in which the ligand charges were scaled by the alchemical parameter λ_{elec} (charge annihilation). This was followed by a series of intermediates in which the ligand Lennard–Jones interactions were removed using the soft-core pathway in Pham et al. [63] with parameters $a = 1, b = 1$ and $c = 1$ using the alchemical parameter λ_{LJ} (Lennard–Jones decoupling). Specifically, we utilized the alchemical schedule (λ_{elec}: 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.85, 0.65, 0.35, 0; λ_{LJ}: 1.0, 0.95, 0.90, 0.85, 0.80, 0.70, 0.60, 0.50, 0.40, 0.30, 0.20, 0.10, 0.0, 0.0, 0.0), which were chosen to ensure that replica exchange success probabilities between neighboring states were approximately equal across the entire transformation. Here, $\lambda = 1$ represents the fully interacting potential term, while $\lambda = 0$ represents the noninteracting potential. Note that six fully coupled and three fully uncoupled states were used, for a total of 24 states. One equilibrium iteration was followed by production runs performed for 15,000 iterations (15 ns/replica).

For the ligand in solvent HREMD simulations (decoupling the ligand in implicit solvent), we used only three states—a fully coupled state ($\lambda_{elec} = \lambda_{LJ} = 1$), a fully discharged ($\lambda_{elec} = 0$; $\lambda_{LJ} = 1$), and a noninteracting state ($\lambda_{elec} = \lambda_{LJ} = 0$). This spacing was found to be sufficient to guarantee full mixing between states the solvent alone. All other simulation parameters were the same as in ligand–protein complex simulations.

Production runs

To test simulation consistency and repeatability, we performed ten independent runs of the 1-methylpyrrole/T4 lysozyme L99A system starting from random initial configurations for 15 ns per replica. After clustering the sampled ligand conformations from the fully interacting states, we then compared clustering patterns between these ten independent runs. Simulations starting from different configurations, if run sufficiently long, should converge to the same clusters, within some statistical noise.

We also performed simulations with two other binders and one other non-binder to examine whether this methodology was able to differentiate binders from non-binders. For the *p*-xylene case, a conformational change in Val111 is required for the ligand to bind [15], which provides a good opportunity to test the ability of our method to sample relevant biomolecular motions and ligand motions.

Binding site identification: The configurations sampled at all of the fully coupled (i.e., fully interacting) states were analyzed together to give final predictions of putative binding sites. In the analysis, the location of the ligand for any given configuration was determined by the ligand atom closest to the center of geometry of the ligand, circled in red in Fig. 1.

Protein alignment: Both the protein and ligand were flexible during our simulations. To be able to cluster all ligand binding sites, all protein conformations from all complexes had to be aligned to provide information on ligand locations relative to the protein. Alignments used the Kabsch algorithm [65, 66] as implemented by Ho [67]. All configurations were aligned to the alpha carbons of the crystal structure.

Clustering analysis: After alignment, the samples from all fully coupled states were clustered using the Density-Based Scan Algorithm with Noise (DBSCAN) [68]. The rationale behind this choice of clustering algorithm lies in the nature of the data. We do not know ahead of time how many alternative binding sites are possible, though we expect that the density in binding sites will be moderately localized, because the exponential nature of the Boltzmann distribution means that binding sites with free energy of binding several kcal/mol lower will have significantly higher density compared to other locations. However, there is also likely to be nonspecific binding density. We therefore expect distinct clusters, with moderate noise, but with the number of clusters unknown *a priori*. These requirements make K-means and hierarchical clustering algorithms less useful. Density-based clustering methods that cluster results based on the density of data points appear more applicable.

To simplify the clustering, we began the clustering process with a grid-based density analysis. Starting from

atomic coordinates of the protein, a three-dimensional cube with 36 Å edge length, just large enough to fit the observed data sampled during the flat-bottom restrained simulation, was centered on the center of geometry of the system and filled with a 2 Å resolution grid defining 46,656 cells of 8 Å^3 volume each. A 2 Å edge length was chosen based on the standard tolerance for the approximate maximum allowable fluctuations from crystal structure. The uniform density over all nonempty cells was calculated, and all cells with fewer than 8 times the background density were discarded. The factor of 8 was chosen for this model system because clusters that appeared visually distinct could not be separated by the clustering algorithms with a density cutoff factor less than 8. This choice of density threshold to exclude from the clustering introduces a small amount of bias, the effect of which we address later.

After this filtering, the DBSCAN algorithm was used to cluster the results [68]. We used a minimum threshold population of 1 % of the total number of samples remaining after low density filtering as the criteria for defining a cluster. Without this filtering removing the low density volumes, the DBSCAN algorithm tended to give large amorphous clusters. This initial density filtering resulted in well-defined clusters in all cases examined. The most populated cluster was then identified as the most probable binding site, with the the centroids of the clusters used to define the locations of the binding sites.

Binding pose identification: The bound configuration of the ligand is determined not only by the location of its center of geometry, but also by the orientation and conformation of the ligand within the binding site. It is therefore important to further analyze these clusters to find the most probable binding orientations and poses.

In order to identify poses, we ran LIGPLOT for each observed pose in the predicted binding sites [69]. The LIGPLOT program generates both lists of observed interactions (such as hydrogen bonding, π–π stacking, and hydrophobic contact interactions) and schematic 2-D representation of protein-ligand complexes in terms of these interactions. We first examined the hydrophobic interaction patterns of all the poses at each site by enumerating the interactions predicted by LIGPLOT. We then identified interactions that were frequently formed for low-RMSD structures and classified the poses based on possession of sets of these predicted interactions.

However, because of the small size of these ligands and the partial freedom the bound states have to reorient in the binding site, it was impossible to uniquely specify low RMSD configurations based solely on lists of observed contacts. We therefore default to classifying clusters based on the average RMSD values of all the poses in the most populated cluster from the ligand in the co-crystal structure after alpha carbon alignment in order demonstrate the performance of the methodology. This procedure requires having a crystal structure with the ligand of interest, but we anticipate that pose identification based on specific protein-ligand contacts in a crystal structure-agnostic method should work much more effectively than it worked here for other more complicated binding sites with larger, more chemically diverse ligands.

Computing binding free energies

Because the simulation algorithm presented here generates samples from all the intermediate states connecting the coupled and uncoupled states, we can use free energy perturbation and reweighting techniques to calculate binding free energies. In this case, we use the multistate Bennett acceptance ratio (MBAR) method to calculate free energies [40], as implemented as the `pymbar` Python code [70]. Because the Gibbs replica exchange scheme we employ requires the potential energies of all replicas be computed for all alchemical states anyway, no additional information is required to analyze the simulation using MBAR if these energies are written to disk during the simulation, as we do here.

The thermodynamic cycle used in this calculation is show in Fig. 2, and involves alchemical decoupling of the ligand both from a system containing a protein, and from a system without a protein. These two free energies are calculated using HREMD simulations as described in the "Theory and computational methods" section.

The free energy of then transferring the ideal gas ligand out of the simulation volume ($\Delta G_{transfer}$) is equal to $k_B T$ times the natural log of the ratio of the volumes the ideal gas ligand is sampling. We will then have for the overall binding free energy:

$$\Delta G_{binding} = \Delta G_{solvent} - \Delta G_{complex} + k_B T \ln \frac{V^{\circ}}{V_{binding}} \qquad (3)$$

where $\Delta G_{solvent}$ and $\Delta G_{complex}$ are the free energy of decoupling the ligand in solvent and complex and V° and V_{sphere} are the standard-state volumes for a single molecule in a box of size $1/N_A$ L, where N_A is Avogadro's number. and $V_{binding}$ is the volume of the binding site, which may change depending on the most appropriate definition of binding site. k_B and T are the Boltzmann constant and temperature in Kelvin, respectively.

$\Delta G_{complex}$ can be calculated by:

$$\Delta G_{complex} = -k_B T \ln Q/V^{\circ} \qquad (4)$$

where Q is canonical partition function, which is given by:

$$Q = \int_V e^{-U(\vec{x})/k_B T} d\vec{x} \qquad (5)$$

where U is the potential energy as a function of the coordinates $\vec{x}$ and V is the phase space volume of $\vec{x}$ over which we sample. There is also a intramolecular partition function
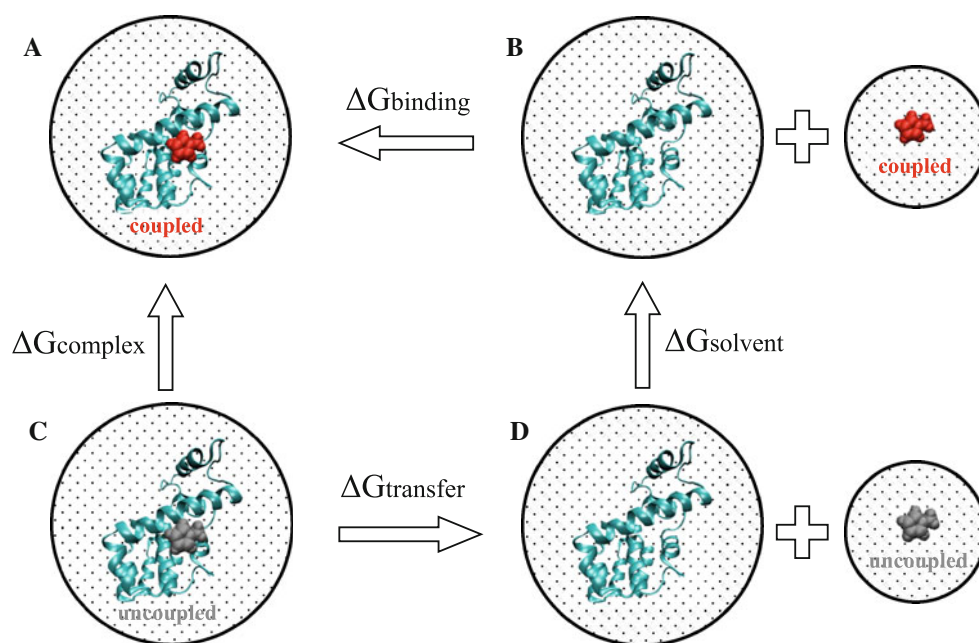
Fig. 2 Thermodynamic cycle for calculating binding free energy. To calculate the binding free energy (**B–A**), the ligand is first decoupled from the solvent (**B–D**), transferred into the protein binding site (**D–C**), and coupled with the protein (**C–A**), closing the cycle. The *dotted box* represents the implicit solvent environment. *Grey and red ligands* represent decoupled and coupled ligands, respectively. $\Delta G_{solvent}$ and $\Delta G_{complex}$ are the free energies of decoupling the ligand in solvent and complex, respectively

that must be included if the ideal gas ligand has intramolecular interactions. However, this contribution cancels exactly from the solvent and complex parts of the calculation, and thus is omitted here for clarity.

In our study, because we spatially restrict the ligand to the vicinity of the protein, we can calculate not only the overall free energy of the ligand binding to the protein, but also the binding free energy with respect to all potential binding sites considered jointly and the binding free energies of ligand binding to individual binding sites. The physical reason for the difference between the three binding free energies is that for each of them, we integrate over a different configurational volume.

Overall binding free energies calculations: The overall binding free energy is the free energy of the ligand considering the entire simulation volume, with partition function given by:

$$Q_{overall} = \int\limits_{V_{overall}} e^{-U(\vec{\mathbf{x}})/k_B T} d\vec{\mathbf{x}} \tag{6}$$

where $V_{overall}$ is total volume inside the flat-bottom sphere. In the limit of box that does not extend far beyond the edge of the protein, and with a sufficiently tight binding affinity, this would be the free energy consistent with an experimental measurement of protein association.

Binding free energies of individual sites: We can also calculate the binding free energies of the ligand to

individual binding sites. Using the grid constructed during the grid-based density analysis, we define a site as the volume made up of the smallest number of cells that include all the samples from that cluster. The partition function for the site is given by:

$$Q_{site} = \int\limits_{V_{site}} e^{-U(\vec{\mathbf{x}})/k_B T} d\vec{\mathbf{x}} \tag{7}$$

where the only difference is that V_{site} is volume within an individual binding site. This free energy will be equivalent to the binding free energy calculated for a method that requires binding at a specific site of a protein, such as fluorescence polarization competition assays. MBAR is applied to all samples that occur in that defined binding volume, over all intermediate and final states.

Binding free energies over all sites: We introduce a final measure, the all-site binding free energy, which is the binding free energy over all the bound clusters considered together. Here, we are interested in the binding affinity over the volume defined by all known specific binding clusters previous identified. The partition function is given by:

$$Q_{all\ sites} = \int\limits_{V_{all\ sites}} e^{-U(\vec{\mathbf{x}})/k_B T} d\vec{\mathbf{x}} \tag{8}$$

where $V_{all\ sites}$ represents the volume of all individual binding sites combined. This should be nearly equal to the binding affinity over the entire protein ($\Delta G_{overall}$), and thus may be

more comparable for many experimental definitions of binding affinity such as by isothermal calorimetry (ITC) or surface plasmon resonance (SPR) than the overall binding affinity using MBAR. This definition does exclude probability density outside of a localized binding site but still in contact with the protein, but these interactions should be negligible because of the low density. Because of the granularity of the boxes, this definition may also exclude some probability density at the edge of clusters that spills into neighboring boxes without reaching the density cutoff, an approximation that we analyze later. MBAR is applied to the samples that occur over the joint volume of all binding sites, over all intermediate and final states. Because the partition function in MBAR is a weighted sum over all samples, each sample can be assigned to a binding cluster, and we strictly satisfy:

$$Q_{all\ sites} = \sum_{i=1}^{N_{clusters}} Q_{site,i} \tag{9}$$

or alternatively:

$$\Delta G_{all\ sites} = -k_B T \ln \left(\sum_{i=1}^{N_{clusters}} e^{-\Delta G_{site,i}/k_B T} \right) \tag{10}$$

In this study, there are a few cases where more than one cluster has samples in the same grid volume, which means that relationship in Eqs. 9 and 10 is only approximately correct because of double counting. However, for this grid size, the differences are less than 0.1 kcal/mol, so we do not attempt to define binding site volumes using a finer grid spacing or split the boxes between clusters.

Results and discussion

Binding sites are consistently identified in repeated trials

To test the statistical robustness of our methodology, we performed ten independent simulation runs of the 1-methylpyrrole/T4 lysozyme L99A system. We analyzed the configuration distribution from all fully coupled states for each independent run individually and compared them.

Between six and twelve clusters were identified for each of the ten simulations, with a total of seventeen independent clusters observed among all simulations. For statistical consistency, we are interested mainly in the most common clusters. After we discarded the six singletons which occurred in only one simulation, eleven sites were left that appeared in multiple simulations. The occupancy O_i of a specific site i, the probability of observing a ligand in this binding site, over the $N_{trials} = 10$ trials is defined as:

$$O_i = \frac{1}{N_{trials}} \sum_{j=1}^{N_{trials}} \frac{N_{i,j}}{N_{total,j}} \tag{11}$$

$N_{i,j}$ is the number of samples observed in site i in trial j, and is set to zero if no cluster is found at that site during that trial. $N_{total,j}$ is the total of number of samples in the observed clusters over all trials. This is a slight approximation, as if a cluster is not observed, the volume still has nonzero density. However, since the cutoff for a cluster is $<1\ \%$, the approximation does not appreciably change the results.

Table 1 Computed site occupancies and free energies

Site	Frequency	From direct observation		From free energy calculation	
		ΔG_{site}	Occupancy	ΔG_{site}	Occupancy
1	10	-3.239 ± 0.292	0.467 ± 0.046	-3.482 ± 0.261	0.364 ± 0.101
2	10	-2.784 ± 0.213	0.211 ± 0.024	-3.043 ± 0.182	0.173 ± 0.044
3	10	-2.142 ± 0.176	0.075 ± 0.010	-2.612 ± 0.206	0.084 ± 0.027
4	8	-2.103 ± 0.154	0.060 ± 0.008	-2.587 ± 0.152	0.080 ± 0.019
5	8	-1.889 ± 0.149	0.048 ± 0.008	-2.566 ± 0.131	0.077 ± 0.016
6	6	-1.804 ± 0.104	0.042 ± 0.005	-2.538 ± 0.119	0.074 ± 0.014
7	5	-1.708 ± 0.109	0.035 ± 0.005	-1.893 ± 0.123	0.025 ± 0.005
8	7	-1.596 ± 0.138	0.029 ± 0.008	-2.599 ± 0.103	0.082 ± 0.013
9	5	-1.263 ± 0.114	0.016 ± 0.005	-1.820 ± 0.091	0.022 ± 0.003
10	4	-1.347 ± 0.098	0.010 ± 0.003	-1.613 ± 0.118	0.016 ± 0.003
11	3	-0.765 ± 0.001	0.007 ± 0.000	-0.672 ± 0.019	0.003 ± 0.000

Quantitative analysis of the eleven putative binding sites identified from ten simulation runs. Frequency is the number of trial runs (out of ten) in which this site was visited. Occupancies from direct observation of the fully interacting states are calculated via Eq. 11, while free energies are estimated from these occupancies by Eq. 13. Free energies are computed at each binding site using MBAR and samples collected from all intermediates confined to the binding sites (Eq. 7), with occupancy estimated from the calculated free energies via Eq. 12. Of eleven putative binding sites discovered in total, three are observed in all ten runs. Site 1, the most populated site in all runs, is located in the binding pocket, indicating that we can identify the binding sites consistently. All free energies in kcal/mol. Error bars are standard deviations over the ten runs
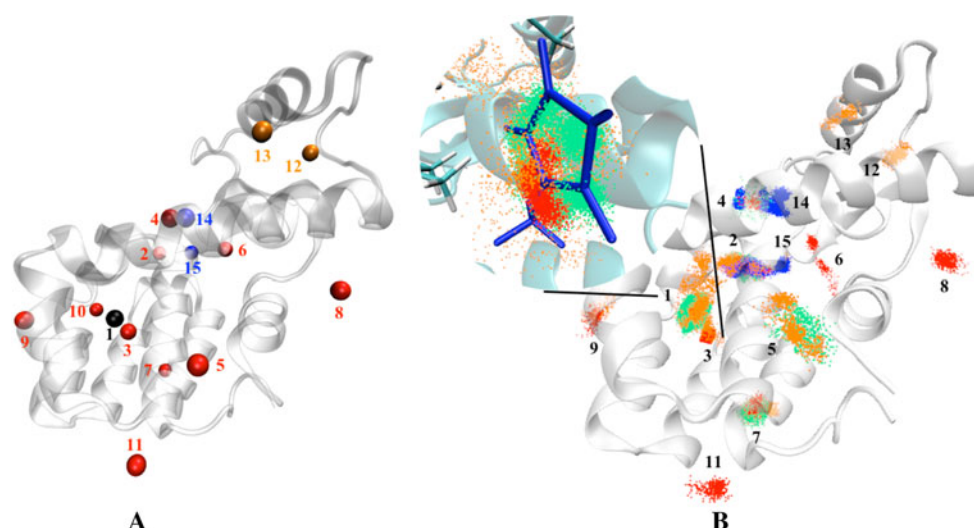
Fig. 3 Fifteen binding sites identified from all simulation runs. **A** The centroid of each site is represented by a sphere, with diameter of 2 Å (the grid resolution). *Black* indicates the crystallographic binding site. *Black* and *red* sites together are the eleven sites for 1-methylpyrrole, with benzene sites as a subset of these. *pink* and *blue* represent additional sites exclusively for *p*-xylene and phenol, respectively.

B The binding site predictions for one run of 1-methylpyrrole (*red*), benzene (*green*), *p*-xylene (*orange*) and phenol (*blue*). Each point represents the center of geometry of ligands at the fully coupled states after grid-based density filtering and clustering. In the inset of the nonpolar binding pocket, all the protein residues within 6 Å of the ligand are shown

Table 1 shows the analysis of the eleven sites identified from our ten runs, with their physical locations in the protein shown by the first eleven positions in Fig. 3a. In Fig. 3a, the volume describing the binding site is represented by a sphere with diameter of 2 Å (the grid resolution). Black indicates the experimental binding site. The eleven sites are numbered based on the frequency of each cluster appearing in the ten trials, and by occupancy if frequency is the same. Of the eleven sites, three are observed in all ten runs, two of which had fractional occupancies greater than 0.2 in all ten runs.

Importantly, site 1 is the most populated in all ten independent runs and is located at the crystallographic ligand binding site, indicating that we can identify this experimentally observed binding site consistently. Site 2 is also observed in all runs and has an average occupancy of more than 0.2. Though not as populated, site 3 is also observed in all runs. However, as indicated from Fig. 3a, site 3 is very close to site 1 and could be interpreted as "spillover" from site 1. All the other sites occur with much lower probability and can be best interpreted as weaker nonspecific binding sites. The clusters in Fig. 3b show the binding site predictions (with the same numbering system) for all four molecules after conducting the grid-based density analysis, each point representing a conformation at the fully coupled states, with only one of the ten runs shown (in red) for 1-methylpyrrole. As shown in Fig. 3b, the volume of site 1 for 1-methylpyrrole is relatively small despite having almost half of the total samples, indicating that density at the binding site is highly localized.

Free energy differences are simply $k_B T$ times the natural logarithm ratios of the relative probabilities of the two states. We should therefore be able to directly compare the ranking of the sites by occupancy (measured by probabilities of being found in each location in the fully coupled states) to the free energies calculated for each site estimated by MBAR. Free energies of binding to each site are computed as described Section II E using Eq. 7, and are shown in Table 1, where they can be compared directly to the occupancies. The ranking of the free energies of the sites agrees with that of the occupancies in almost all cases, though there are some differences somewhat outside of statistical error. The free energy difference between the two most populated binding sites is only 0.44 kcal/mol, suggesting that there may exist at least one potential binding site other than the experimental binding site. The fact that low-frequency clusters are not consistently observed in all simulations indicate that the simulations are not entirely converged. This may explain the difference in binding affinity between rarer clusters, although the convergence of the dominant binding sites does appear adequate based on agreement between the two ways of calculating relative affinity between clusters.

To better understand the consistency between free energies and occupancies, we can estimate an occupancy for each site based on its free energy. We estimated the occupancies O_i from the free energies ΔG_i as:

$$O_i = \frac{e^{-\Delta G_i/k_B T}}{\sum_{j=1}^{N_{sites}} e^{-\Delta G_j/k_B T}} \tag{12}$$

where G_i is the ΔG_{site} for binding site i. Uncertainties for each site free energy are the standard deviation of the free energy over the ten independent runs, and are the uncertainty in a single calculation, not in the mean.

We can also estimate each cluster's free energy based on the directly observed occupancy of the cluster in the fully interacting states. Each cluster's free energy relative to the unbound state is equal to:

$$\Delta G_i = -k_B T \ln \frac{O_i}{O_{far}} \qquad (13)$$

where ΔG_i and O_i are ΔG_{site} and occupancy for site i. O_{far} is the occupancy of the "cluster" of samples far away from the protein as to be effectively noninteracting. This cluster serves as a reference, because the transfer of the ligand from solvent to this volume should be $\Delta G_{site} = 0$. We define this cluster as those samples found between $r = r_{cutoff}$ and $r_{cutoff} - 5$ Å in the fully coupled state.

As shown in Table 1, the occupancies calculated both ways as well as the free energies calculated both ways are in relatively good agreement within statistical error, indicating that our definition of the occupancy and the free energy calculation methodology are consistent. The free energy calculations in principle contain more information, since they incorporate the potential energies as well as the location information the occupancies contain and also include samples from multiple intermediate states. Interestingly, however, the uncertainties in occupancies and free energies calculated starting from *either* directly observed occupancies or using MBAR are similar.

The dominant binding site can be identified accurately across multiple molecules

To test the accuracy of our methodology in identifying binding sites across a range of ligands, we examined the predicted sites of four ligands binding to the same protein, one of which (phenol) is known not to bind experimentally. The same simulation parameters were used, except only one simulation was run for each of these additional ligands.

Figure 4 shows the site occupancies for four molecules. For 1-methylpyrrole, the statistical error in a single run (not in the mean of all runs) was calculated over the 10 runs, while values for only one run were used for the other three ligands. Since many of the same binding sites were observed in simulations of the different molecules, we used the same numbering systems described in the previous section for the 1-methylpyrrole runs, adding newly identified sites to the initial eleven sites.

As shown in Fig. 4, since the three binders share similar binding patterns, the total number of potential binding sites identified on the protein only increases by four when
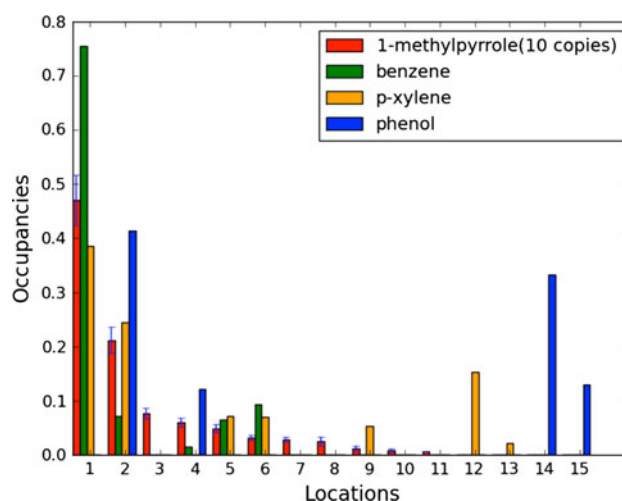

Fig. 4 Binding site fractional occupancies. The three binders share similar binding patterns, and are labeled by extending the numbering scheme from the 1-methylpyrrole simulations. Site 1, located at the experimental binding site, is the most populated site for all three binders. However, no samples above background are observed in the binding site for the nonbinder, phenol. *Error bars* in 1-methylpyrrole are standard deviations over the ten runs

additional ligands are analyzed, with two of the sites from the non-binder, phenol. These four additional sites are the last four numbered sites in Fig. 3a. Orange and blue represent additional sites observed for *p*-xylene and phenol, respectively. The green, orange and blue clusters in Fig. 3b are the binding site predictions for benzene, *p*-xylene and phenol. Each point represents a conformation at the fully coupled state, with the low density sites filtered out. The binding site at the crystallographically observed binding cavity (site 1) is identified as the most populated site for all three binders. Additionally, no binding cluster of any density above background is identified at this location for simulations of the non-binder. This suggests that, at least for this model system and small set of ligands, we can identify the experimental binding site accurately and consistently and differentiate the binders from non-binders.

Binding poses can also be identified

Pose prediction at site 1 for 1-methylpyrrole

After the binding site (site 1) was successfully identified, we further examined the poses found at that site. From the 10 runs of the 1-methylpyrrole/T4 lysozyme L99A system, we took the set of ligands in the most populated cluster, which is also the experimental binding cluster, and examined the poses of the ligand configurations in this site.

We initially attempted to analyze the poses based on the hydrophobic interaction contacts made between the ligand and the protein predicted by the LIGPLOT program.

Table 2 Average ligand RMSD (in Å) from crystal structures of AutoDock and the methodology presented in this paper

Molecules	Rigid AutoDock	Flexible AutoDock	Our methodology
1-Methylpyrrole	1.84	1.87	1.93 ± 0.09
Benzene	1.62	2.30	2.32
p-Xylene	2.32	3.76	3.14
Phenol[a]	11.21	12.87	N/A

For AutoDock, the average RMSD was calculated over 50 top poses, while for our methodology, this RMSD was calculated over all poses in the binding site cluster, with the standard deviation over 10 repetitions for 1-methylpyrrole. For the nonbinder phenol, since there is no crystal structure available, we use the co-crystal ligand benzene with phenol in order to identify whether docking incorrectly places the ligands in the binding site. The percentage of ligands in the binding site volume may be higher than that within 2 Å RMSD because of local protein rearrangement during the simulation. All RMSDs are symmetry corrected

[a] As compared to the binding cavity in benzene co-crystal structure

Table 3 Percentages (%) of poses with RMSD from crystal structure less than 2 Å for AutoDock and the methodology presented in this paper

Molecules	Rigid AutoDock	Flexible AutoDock	Our methodology
1-Methylpyrrole	46.0	50.0	43.3 ± 2.8
Benzene	52.0	30.0	33.4
p-Xylene	36.0	20.0	19.1
Phenol[a]	2.0	4.0	0.0

The standard error for 1-methylpyrrole was calculated over the ten runs. For the nonbinder phenol, since there is no crystal structure available, we replaced the benzene co-crystal ligand with phenol and computed RMSD to the resulting structure. All RMSDs are symmetry corrected

[a] As compared to the binding cavity in benzene co-crystal structure

Although there were a number of hydrophobic interactions correlated with low RMSD configurations, there was no single hydrophobic interaction pattern that could be conclusively identified with low RMSD (< 1 Å) binding, suggesting that it is not possible to identify the most representative pose by hydrophobic interaction patterns alone for this system. This was determined by using one run of 1-methylpyrrole system as a training set to determine patterns of contacts associated with low RMSD and then testing these patterns on a second run to see if low RMSD structures were identified.

We hypothesize that it is difficult to determine binding patterns from contacts is this case because it is an engineered ligand binding system with a large hydrophobic binding surface (up to 20 contacts, depending on the definition of contact), with similar contributions to binding energy. Such a consensus pose procedure based on observed contacts is more likely to work for systems with important hydrogen-bonding patterns systems and more complex ligands, a hypothesis that we plan to test in future studies.

We therefore focused on identifying poses based on RMSD from crystal structure. We calculated the RMSD for all four molecules with respect to the co-crystal poses (Tables 2, 3). All RMSD values are symmetry corrected. Although we ran all docking and simulations with the benzene co-crystal structure, we calculated RMSDs from the experimental crystal structures of 1-methylpyrrole and p-Xylene (PDB accession codes 2OU0 and 3GUM) after aligning the alpha carbons to incorporate the conformational differences between the complexes.

Figure 5 shows 100 typical poses of each binder at the binding site are shown. 1-Methylpyrrole is primarily

oriented the same way in all configurations, as can be seen by the essentially stationary single nitrogen. Benzene has somewhat more conformational heterogeneity, as can be expected from a highly symmetrical ligand, but still has a relatively localized binding density. However, p-xylene has significant conformational heterogeneity in the binding site, which we discuss in the next section.

The role of Val111 in binding

One of the challenges involved in simulations of ligand binding is capturing correlated motions involving both ligand and protein. T4 lysozyme L99A is a good model system to test the power of this methodology to overcome this sampling problem. Previous simulations have shown that p-xylene cannot bind to the same configuration of the binding cavity as smaller ligands; instead, a rotamer change of Val111 is first required. In simulations with p-xylene placed in the binding cavity, the occluded nature of the pocket makes this rotamer motion is extremely slow, often occurring on time scales beyond that of typical simulations [15]. In this study, we monitored movement of Val111 during the HREMD simulations of p-xylene and benzene. Figure 6 shows the RMSD of the two ligands from their crystal structure with respect to the RMSD of Val111 from the crystal structure for (a) p-xylene and (b) benzene as well as the ligand RMSD of the ligands versus against the Val111 χ dihedral angle ($C - C_\alpha - C_\beta - C_\gamma$) in (c) and (d). Each dot is a conformation at each iteration. Because we are comparing the ligand pose to the crystal structure pose, low ligand RMSD corresponds to the ligand being in the crystallographic binding site.

As shown in Fig. 6a, the ligand binding and the conformational change of Val111 for p-xylene are highly correlated. When p-xylene enters the binding site, Val111 is necessarily displaced; if it is not, no binding occurs. For
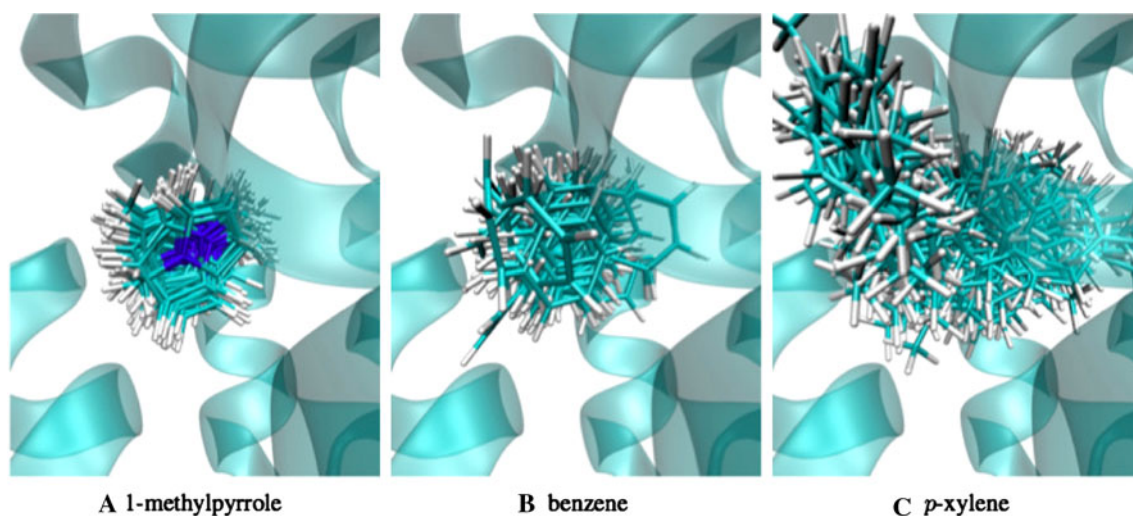
Fig. 5 Superimposed poses (100 each) at the experimental binding site for all three binders (1-methylpyrrole, benzene and *p*-xylene). For 1-methylpyrrole and benzene, configurational noise is limited, while *p*-xylene transitions between two different clusters during the simulation
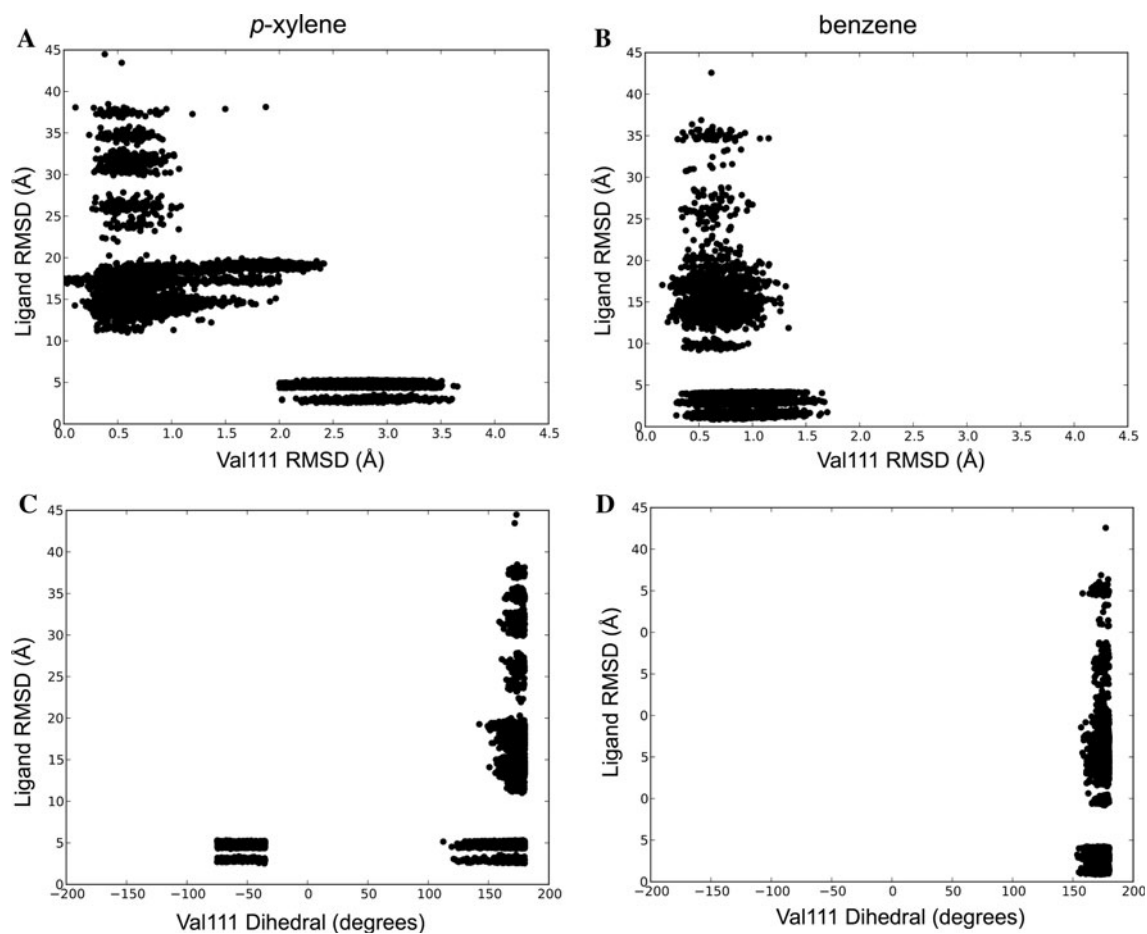

Fig. 6 Correlation between ligand binding site occupation and Val111 displacement for p-xylene and benzene. RMSD of the ligand from the crystal structure with respect to the RMSD of Val111 from the crystal structure (*upper graphs*) and the Val111 χ dihedral angle $(C - C_\alpha - C_\beta - C_\gamma)$ (*lower graphs*) for *p*-xylene (*left side*, **a** and **c**) and benzene (*right side* **b** and **d**). All calculations are of fully interacting ligands. Val111 must move for *p*-xylene binding to occur, either by a torsional angle rotation or by backbone motion, but benzene only binds to the unbound crystallographic configuration of Val111
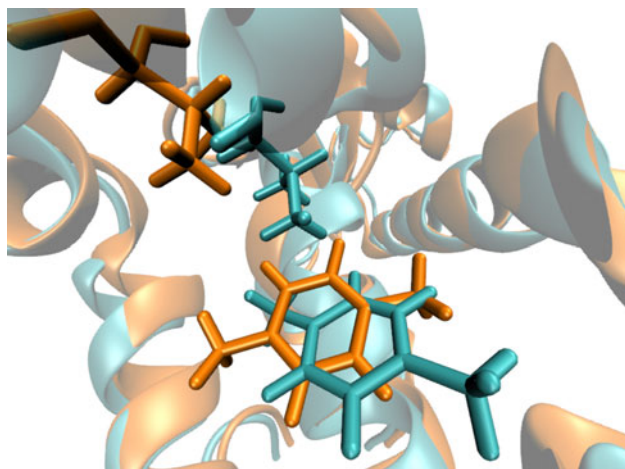
Fig. 7 Two representative structures observed in the simulation of *p*-xylene. *Cyan* and *orange* are crystal-like (RMSD = 0.3 Å from crystal) and alternative (RMSD = 2.87 Å from crystal) structures, respectively. In the crystal-like structure, Val111 dihedral changes from the configuration found in the apo or small binder crystals. In the alternative structure, Val111 shifts away via backbone motion

benzene binding (Fig. 6b), Val111 stays in the initial location regardless of whether the ligand is bound or not. This demonstrates that our HREMD decoupling strategy can significantly accelerate such coupled configurational changes on binding that would normally require long simulations of at least several nanoseconds in standard MD simulations [15]. HREMD does this by removing the ligand from the pocket so that the dihedral transition can occur.

If we look directly at the Val111 χ dihedral angle (C–C_α–C_β–C_γ), the correlation between binding of ligand and the conformational change of Val111 is not complete. There are in fact configurations that have low *p*-xylene RMSD, but where the dihedral corresponds to the bound benzene structure (181L), unlike the *p*-xylene structure (3GUM). This is possible because the protein backbone shifts out, allowing Val111 to move, a binding mode not observed in previous free energy calculations [15]. Figure 7 shows two low RMSD structures from each of the two clusters. Cyan and orange are used for the dihedral shift (RMSD = 0.34 Å) and alternative backbone shift (RMSD = 2.87 Å) structures, respectively. It is not clear if this observed difference in binding modes from previous simulations is due to force field errors, implicit solvent deficiencies, lack of protein relaxation, or some other unknown reason.

To quantify the relative frequency of the two binding modes, we clustered all the conformations in the binding site of *p*-xylene. Only two clusters with more than 10 % of all the conformations are present, with respective occupancies of 0.53 and 0.32. By comparing to the *p*-xylene crystal structure, we found that cluster one has a 0.56 Å

average RMSD with respect to the crystal structure while cluster two has a 3.03 Å average RMSD. There are thus two primary binding modes in this location-defined cluster that can be distinguished by their orientation.

One unrelated but important observation from Fig. 6 is that there are no ligand observations in the range of 5 and 10 Å for either benzene or *p*-xylene in the interacting state, indicating that there is no observed physical entry route for the ligand in the simulation. Instead, it hops back and forth between bulk and the binding site via the unphysical decoupling pathway.

Comparison of docking and our modified HREMD methodology

It is instructive to compare the performance of docking methods to our methodology. The T4 lysozyme L99A system has proven a challenging case for UCSF's DOCK program as well as other docking programs [43–46]. Therefore, as an additional check we attempted molecular docking to identify binding sites and poses, in our case using AutoDock. We first compared the average ligand RMSD from the crystal structures for all binders in both cases. For AutoDock, the average RMSD was calculated over 50 top poses, while for our modified HREMD, the average RMSD was calculated over all poses in the highest probability binding site. We also compared the percentages of poses with RMSD (from the experimental co-crystal structure for each ligand after alpha carbon alignment) with values less than 2 Å. Since there is no crystal structure for the nonbinder phenol, we used the benzene co-crystal and replaced the benzene with phenol and used RMSDs to that modeled crystal structure to see if either approach incorrectly placed phenol into the binding site. Results are shown in Tables 2 and 3.

We note that the percentage of ligands in the binding site volume (as seen in Fig. 4) may be higher than the percentage within 2 Å RMSD of the crystal structure because of local protein rearrangement during the simulation. For example, ∼40 % of *p*-xylene configurations were in the binding site volume, and the average RMSD of the alternate configurations was 3.03 Å compared to the average RMSD of 0.56 Å without protein rearrangement. If such protein rearrangements observed in simulation are accepted as potentially physical, then these alternate configurations should also be considered part of the binding ensemble at this site.

Surprisingly, AutoDock and the more sophisticated methodology presented here produced comparable results for the binding site locations. Fraction within a given RMSD does not mean exactly the same thing when comparing the two methods. In the docking runs, only 50 poses were generated out of hundreds of thousands of attempts

Table 4 Comparisons between calculated and experimental binding free energies of four different molecules in kcal/mol

Molecules	ΔG_{site}	$\Delta G_{all\ sites}$	$\Delta G_{overall}$	$\Delta G_{explicit}$	$\Delta G_{experimental}$
1-Methylpyrrole	-3.48 ± 0.26	-4.15 ± 0.25	-5.05 ± 0.21	-4.32 ± 0.08	-4.44
Benzene	-4.26 ± 0.71	-5.15 ± 0.80	-6.01 ± 0.81	-4.56 ± 0.20	-5.19
p-Xylene	-4.01 ± 0.89	-4.94 ± 0.85	-5.72 ± 0.95	-3.54 ± 0.17	-4.67
Phenol	-1.03 ± 0.32	-1.78 ± 0.47	-2.32 ± 0.58	-1.26 ± 0.09	> -2.74

ΔG_{site} is the binding free energy to the most populated cluster, which except for phenol is the binding cavity. The binding energy of phenol to the binding cavity is -0.16 ± 0.53 kcal/mol. $\Delta G_{all\ sites}$ is the binding energy over all specifically-bound clusters, while $\Delta G_{overall}$ is over the entire protein. $\Delta G_{explicit}$ are explicit solvent simulations from Ref. [15]

while in our simulations, all poses in the binding configuration are counted. Instead, it should be considered only an indication of whether the crystallographic binding site can be identified. Rigid docking outperforms flexible docking substantially for two binders, which is especially interesting in the case of p-xylene. Since we know that Val111 must readjust from the small-binder crystal structure in both experiment and simulations for binding to actually occur, the better performance of rigid docking indicates that the good performance may be a statistical fluke, and that it is only recognizing a hollow hydrophobic site. Tests on wider sets of ligands as we are currently carrying out will be required to further compare the methods.

Binding free energies can be accurately calculated

Though the initial goal of this study was not to calculate binding free energies, the fact that our methodology was modified from a free energy calculation tool made it straightforward. We calculated the free energies of ligand binding to different sites, as shown in Table 1. The ordering of the sites using free energies matches the ordering using occupancies well, though not perfectly. The free energy of ligand binding to the most populated binding site is substantially more favorable than those of other sites, confirming that a single site is dominant, though not overwhelmingly so, at only 2–3 times the occupancy of the next most frequently occupied site.

Additionally, we were able to calculate the overall free energies of different ligands associated with the protein, over the entire simulation volume, as shown in Table 4. The overall free energies generally match the experimental values to within statistical noise. In Table 4, we also compare all-site binding free energies and binding free to the dominant binding site to the the overall free energies. For the non-binder phenol ΔG_{site} is close to zero since the experimental site was not observed as the one of the predicted potential clusters. The errors for the 10 replica set of 1-methylpyrrole simulations are calculated using the

standard deviation in the free energy over the ten simulations, while the errors for the rest are calculated using the statistical uncertainty estimate for MBAR.

As a comparison, we also include in Table 4 the explicit solvent calculations of the same ligands (with the same forcefield except for the use of explicit, rather than implicit, solvent) from Mobley et al. [15], which were calculated assuming binding to only a single site. We observe that these binding calculations are relatively consistent with our results. They are in particularly close agreement with the free energy of binding to the highest occupancy site, though the statistical noise is somewhat too high to reach any strong conclusions. Gallicchio et al. [41], using a different choice of force field and implicit solvation model, but also assuming a single binding site, calculated a binding free energy of -4.01 ± 0.04 kcal/mol for benzene and -1.40 ± 0.03 for phenol. This agrees with our single site calculation for benzene, but is more favorable for binding for phenol. The number for phenol in Table 4 is for the most favorable binding site for phenol, not the hydrophobic pocket, which has a binding affinity -0.16 kcal/mol. The binding free energies of other molecules examined by Gallicchio et al. were also underestimated, similar to the explicit solvent calculations of Mobley et al. This underestimation may be due to experimental contribution of alternate sites to the free energy of binding not examined in these simulations, but may also be explained by a host of other force field issues.

In the limit of tight binding and a sufficiently small simulation volume, the overall free energy should be slightly more favorable than the all-sites free energy, because the overall free energy also includes the completely nonspecific binding to the protein and the low concentration in the simulated volume near the protein. However, in this study this discrepancy approaches 1 kcal/mol. This difference appears to in part be because of the granularity of the clustering algorithm, which omits density outside the cluster if it falls below 8 times the average density background. We performed an alternate binding calculation for the 1-methylpyrrole case in which we set the energies of all samples not in the set of grid

cubes assigned to binding site clusters equal to energies drawn from the samples away from the protein. In this case, the overall binding affinity changed from -5.05 ± 0.21 to -4.19 ± 0.19 kcal/mol, indicating that the difference between the all-site free energy and overall free energy was due to samples associated with the protein, not samples at other locations in the box. However, it is still unclear how much of the weight is due to samples from the binding sites that were not included in the clustering because of the grid granularity and how much is due to samples weakly associated to the protein but not part of any binding cluster. With these missing densities, all-site binding affinities would be shifted somewhat towards the overall binding affinity, and the individual site binding affinities would also become slightly more favorable.

One of the difficulties in GPU-accelerated MD simulations is parallelization of a single simulation across multiple GPUs. The highly parallelized replica structure of HREMD made it suitable to run on multiple GPUs, since we can parallelize up to one GPU per replica. As a result, we were able to generate 15-ns simulations for all 24 alchemical states in about 6.3 days of wall time, using 6 GPUs at 4 replicas per GPU, running at approximately 10 ns/day/GPU in GPU time per single replica. This time scale makes such calculations already potentially useful for drug discovery. Optimized OpenMM GPU code without the alchemical state code achieved 40 ns/day on the same same machine and on the same systems. This indicates that with properly optimized code and given the rapid development of GPU processor technology, the wall-clock time for studies such as this will decrease significantly in the very near future.

Some parameters involved in our simulations, such as the number of fully coupled states, the number of fully uncoupled states and the Monte Carlo displacement, could potentially be further optimized, as our initial optimization tests of these parameters was done with a sparse grid of parameter choices. The results (in Supporting Information) suggest that in most cases, the sampling is not particularly sensitive to these parameters, though a full optimization is beyond the scope of the current study. A rigorous exploration of these parameters over longer time scale may reveal additional ways to further improve the efficiency of the methods presented in this study. There are a large number of other potential ways to improve the efficiency of these simulations. For example, choosing $c = 6$ or $c = 12$ instead of $c = 1$ is likely to be somewhat more efficient [63], requiring fewer intermediates for rapid mixing between states. Other possibilities include optimization of the OpenMM CUDA implementation and adding Monte Carlo moves of ligand and protein torsional angles. Such improvements could further bring the convergence time

down from days to hours, making such simulations a more useful tool in drug design pipelines.

We have found that optimized HREMD simulations in implicit solvent can identify binding sites and binding modes in a model system without prior knowledge of the binding site, even in a highly buried binding pocket. Since we start the simulations from random starting configurations, no binding site information is needed. As a result, our methodology can potentially be used to conduct low-throughput virtual screening, even when no binding site information is available. In low-throughput virtual screening, especially in the lead optimization stage, the accuracy presented here may be sufficient, and the relatively moderate computational cost will either now or soon be accessible.

However, it is important to recognize that this is a test of only four molecules and a single, relatively small protein. The demonstrated ability of modified HREMD methods presented here to sample multiple binding sites will be independent of the system. However, the success in finding the binding site and the agreement of binding affinities may not be nearly as transferable. This study is meant as an exploration of the utility of modified HREMD to sample between binding sites, and is only a proof-of-principle.

Despite the general success of this methodology, there are a few flaws in the clustering approach presented here. One problem is that more than one cluster can contain samples in the same grid volume, leading to the inability to uniquely decompose a binding site into separate clusters. However, this leads to a relatively low amount of error, less than 0.1 kcal/mol in this study. Another problem is that there are some samples belonging to the binding cluster that are omitted because they partially fall into another box that falls below the overall density cutoff. Overcoming these problems would require either additional data in order to use a smaller grid, or a more robust density-based clustering algorithm, technical problems that can presumably be overcome with sufficient work, but which are not required for the level of precision presented in this study.

We find that for at least the moderate affinity ligands in this study, the free energy of binding sites other than the most likely binding site contributes nonnegligibly to the total free energy, with these alternate binding sites contributing between 0.7 and 0.9 kcal/mol to the overall binding free energy. Although this contribution is likely to be less in tight binding molecules that have a very high affinity binding mode, this observation does mean that the exact binding affinity can depend significantly on the way the binding site is defined and the method used to calculate it. This contribution from alternate sites may possibly be a reason that binding affinities computed in the studies of Mobley et al. [15] and Gallicchio et al. [41], in which only the crystallographic buried binding cavity of was

considered, were consistently less favorable than experiment by about this amount. However, there are certainly no lack of other possible explanations for this discrepancy. The existence of a distribution of binding sites, if is a feature of experimental systems, and not merely an artifact of the simulation, may also be important for fragment-based drug design studies, as there may be multiple binding sites that are worth targeting in a single protein.

We also compared the alternative binding sites observed directly with the experimental electron densities deposited in the Protein Data Bank to see if unassigned densities could be correlated with these putative binding sites. We examined all binding sites with occupancy of above 0.1 in the simulations, as density lower than this is unlikely to be observed above noise. For benzene, no alternative sites have occupancies larger than 0.1, so no search is necessary. For p-xylene, we did not observe any apparent electron densities in the volumes of the two putative sites with occupancies larger than this threshold of 0.1 occupancy. For 1-methylpyrrole, two ligands were proposed in the crystal structure, one of which is an alternative site with a lower density than the binding site. However, this alternative site was not predicted by our methodology. For the single computationally predicted alternative site with 1-methylpyrrole with an occupancy higher than 0.1, we observed some unassigned electron density in the crystal structure at that location, but it was not distinguishable from water. Interestingly, the electron density of Met106 in this alternative binding site was ambiguous in the crystal structure, with two different conformations of Met106 proposed. However, this may be a coincidence and may not be related to potential experimental partial occupancy of the ligand at this site. Our simulations do appear to be fairly well converged, at least with respect to the two most populated binding sites, which suggests that either the force field and/or implicit solvent model is creating spurious density, or there is some physical reason for this binding site not being present in experimental crystal structures.

Conclusions

In this study, we used a modified version of Hamiltonian replica exchange molecular dynamics among alchemical intermediates combined with Monte Carlo ligand translation/rotation moves to identify putative binding sites and poses in the T4 lysozyme L99A model system starting from random initial ligand positions. Our results suggest that this methodology can identify the binding sites consistently and accurately. Moreover, we can identify the correct binding orientations within these binding sites relatively accurately. Last but not least, we can not only calculate the overall free energies of binding using MBAR, but can also decompose the contributions to the overall binding free energy both in terms of individual binding sites and all binding sites combined, demonstrating the extent to which the ensemble of weak binders may contribute nonnegligibly to the overall free energy. With the wider availability of GPU simulation resources, this methodology may be a stepping-off point for further improved drug discovery methods when no co-crystal structures of the target of interest are available.

References

1. Schneider G (2010) Virtual screening: an endless staircase? Nat Rev Drug Discov 9(4):273–276
2. B-Rao C, Subramanian J, Sharma SD (2009) Managing protein flexibility in docking and its applications. Drug Discov Today 14(7–8):394–400
3. Lie MA, Thomsen R, Pedersen CNS, Schiøtt B, Christensen MH (2011) Molecular docking with ligand attached water molecules. J Chem Inf Model 51(4):909–917
4. Thompson DC, Humblet C, Joseph-McCarthy D (2008) Investigation of MM-PBSA rescoring of docking poses. J Chem Inf Model 48(5):1081–1091
5. Graves AP, Shivakumar DM, Boyce SE, Jacobson MP, Case DA, Shoichet BK (2008) Rescoring docking hit lists for model cavity sites: predictions and experimental testing. J Mol Biol 377(3):914–934
6. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 57(2):225–242
7. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49(20):5912–5931
8. Deng W, Verlinde CLMJ (2008) Evaluation of different virtual screening programs for docking in a charged binding pocket. J Chem Inf Model 48(10):2010–2020
9. Levitt DG, Banaszak LJ (1992) POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10(4):229–234
10. Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15(6):359–363
11. Patrick Brady Jr G, Stouten PFW (2000) Fast prediction and visualization of protein binding pockets with PASS. J Comput Aid Mol Des 14(4):383–401
12. Halgren TA (2009) Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model 49(2):377–389
13. Verdonk ML, Cole JC, Watson P, Gillet V, Willett P (2001) SuperStar: improved knowledge-based interaction fields for protein binding sites. J Mol Biol 307(3):841–859

14. Bliznyuk AA, Gready JE (1999) Simple method for locating possible ligand binding sites on protein surfaces. J Comput Chem 20(9):983–988
15. Mobley DL, Graves AP, Chodera JD, McReynolds AC, Shoichet BK, Dill KA (2007) Predicting absolute ligand binding free energies to a simple model site. J Mol Biol 371(4):1118–1134
16. Jiang W, Roux B (2010) Free energy perturbation Hamiltonian replica-exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. J Chem Theory Comput 6(9):2559–2565
17. Deng Y, Roux B (2009) Computations of standard binding free energies with molecular dynamics simulations. J Phys Chem B 113(8):2234–2246
18. Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS (2011) Alchemical free energy methods for drug discovery: progress and challenges. Curr Opin Struc Biol 21(2):150–160
19. Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL, Ensign DL, Bruns CM, Pande VS (2009) Accelerating molecular dynamic simulation on graphics processing units. J Comput Chem 30(6):864–872
20. Eastman P, Pande V (2010) OpenMM: a hardware-independent framework for molecular simulations. Comput Sci Eng 12(4):34–39
21. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4(2):187–217
22. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Commun 91(1–3):43–56
23. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, DeBolt S, Ferguson D, Seibel G, Kollman P (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput Phys Commun 91(1–3):1–41
24. Clark SW, Tempczyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 112(16):6127–6129
25. Onufriev A, Bashford D, Case DA (2000) Modification of the Generalized Born Model Suitable for Macromolecules. J Phys Chem B 104(15):3712–3720
26. Michel J, Verdonk ML, Essex JW (2006) Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization? J Med Chem 49(25):7427–7439
27. Shaw DE, Chao JC, Eastwood MP, Joseph G, Grossman JP, Richard HC, Lerardi DJ, István K, Klepeis JL, Layman T, McLeavey C, Deneroff MM, Moraes MA, Mueller R, Priest EC, Shan Y, Spengler J, Theobald M, Towles B, Wang SC, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ (2008) Anton, a special-purpose machine for molecular dynamics simulation. Commun ACM 51(7):91
28. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330(6002):341–346
29. Mobley DL (2012) Let's get honest about sampling. J Comput Aid Mol Des 26(1):93–95
30. Purisima EO, Hogues H (2012) Protein-ligand binding free energies from exhaustive docking. J Phys Chem B 116(23):6872–6879
31. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314(1–2):141–151
32. Fukunishi H, Watanabe O, Takada S (2002) On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. J Chem Phys 116(20):9058
33. Hamelberg D, Mongan J, Andrew MJ (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120(24):11919–11929
34. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J Comput Phys 23(2):187–199
35. Barducci A, Bonomi M, Parrinello M (2011) Metadynamics. WIREs Comput Mol Sci 1(5):826–843
36. Deighan M, Bonomi M, Pfaendtner J (2012) Efficient Simulation of Explicitly solvated proteins in the well-tempered ensemble. J Chem Theory Comput 8(7):2189–2192
37. Kokubo H, Tanaka T, Okamoto Y (September 2013) Two-dimensional replica-exchange method for predicting protein-ligand binding structures. J Comput Chem 34(30):2601–2614
38. Chodera JD, Shirts MR (2011) Replica exchange and expanded ensemble simulations as Gibbs sampling: simple improvements for enhanced mixing. J Chem Phys 135(19):194110
39. Peter KE, Friedrichs MS, Chodera JD, Radmer RJ, Bruns CM, Ku JP, Beauchamp KA, Lane TJ, Wang L-P, Shukla D, Tye T, Houston M, Stich T, Klein C, Shirts MR, Pande VS (2013) OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. J Chem Theory Comput 9(1):461–469
40. Shirts MR, Chodera JD (2008) Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys 129(12):124105
41. Gallicchio E, Lapelosa M, Levy RM (2010) Binding energy distribution analysis method (bedam) for estimation of protein ligand binding affinities. J Chem Theory Comput 6(9):2961–2977
42. Boyce SE, Mobley DL, Rocklin GJ, Graves AP, Dill KA, Shoichet BK (2009) Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. J Mol Biol 394(4):747–763
43. Wei BQ, Baase WA, Weaver LH, Matthews BW, Shoichet BK (2002) A model binding site for testing scoring functions in molecular docking. J Mol Biol 322(2):339–355
44. Wei BQ, Weaver LH, Ferrari AM, Matthews BW, Shoichet BK (2004) Testing a flexible-receptor docking algorithm in a model binding site. J Mol Biol 337(5):1161–1182
45. Ferrari AM, Wei BQ, Costantino L, Shoichet BK (2004) Soft docking and multiple receptor conformations in virtual screening. J Med Chem 47(21):5076–5084
46. Graves AP, Brenk R, Shoichet BK (2005) Decoys for docking. J Med Chem 48(11):3714–3728
47. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. J Comput Chem 26(16):1668–1688
48. Mobley DL, Dumont E, Chodera JD, Dill KA (2007) Comparison of charge models for fixed-charge force fields: small-molecule hydration free energies in explicit solvent. J Phys Chem B 111(9):2242–2254
49. Jakalian A, Bush BL, Jack DB, Bayly CI (2000) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. J Comput Chem 21(2):132–146
50. Jakalian A, Jack DB, Bayly CI (2002) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem 23(16):1623–1641
51. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25(9):1157–1174

52. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model 25(2):247–260

53. Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. Proteins 8(3):195–202

54. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30(16):2785–2791

55. Eastman P, Pande VS (2010) CCMA: a robust, parallelizable constraint method for molecular simulations. J Chem Theory Comput 6(2):434–437

56. Mobley DL, Chodera JD, Dill KA (2006) On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. J Chem Phys 125(8):084902

57. Boresch S, Tettinger F, Leitgeb M, Karplus M (2003) Absolute binding free energies: a quantitative approach for their calculation. J Phys Chem A 107(35)

58. Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE (2011) How does a drug molecule find its target binding site? J Am Chem Soc 133(24):9181–9183

59. Harvey MJ, Giupponi G, De Fabritiis G (2009) ACEMD: accelerating biomolecular dynamics in the microsecond time scale. J Chem Theory Comput 5(6):1632–1639

60. Zacharias M, Straatsma TP, McCammon JA (1994) Separation-shifted scaling, a new scaling method for Lennard–Jones interactions in thermodynamic integration. J Chem Phys 100(12):9025

61. Beutler TC, Mark AE, van Schaik RC, Gerber PR, van Gunsteren WF (1994) Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. Chem Phys Lett 222(6):529–539

62. Shirts MR, Pande VS (2005) Solvation free energies of amino acid side chains for common molecular mechanics water models. J Chem Phys 122:134508

63. Pham TT, Shirts MR (2011) Identifying low variance pathways for free energy calculations of molecular transformations in solution phase. J Chem Phys 135(3):034114

64. Sindhikara D, Emerson DJ, Roitberg AE (2010) Exchange often and properly in replica exchange molecular dynamics. J Chem Theory Comput 6:2804–2808

65. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallogr A 32(5):922–923

66. Kabsch W (1978) A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr A 34(5): 827–828

67. Ho BK http://boscoh.com/protein/matchpy.html

68. Sander J, Ester M, Kriegel H-P, Xu X (1998) Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. Data Min Knowl Discov 2(2):169–194

69. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8(2):127–134

70. Shirts MR, Chodera JD pymbar, https://simtk.org/home/pymbar