

# Machine Learning and Enhanced Sampling Simulations for Computing the Potential of Mean Force and Standard Binding Free Energy

Martina Bertazzo,<sup>1</sup> Dorothea Gobbo,<sup>1</sup> Sergio Decherchi,\* and Andrea Cavalli\*

Cite This: *J. Chem. Theory Comput.* 2021, 17, 5287–5300

Read Online

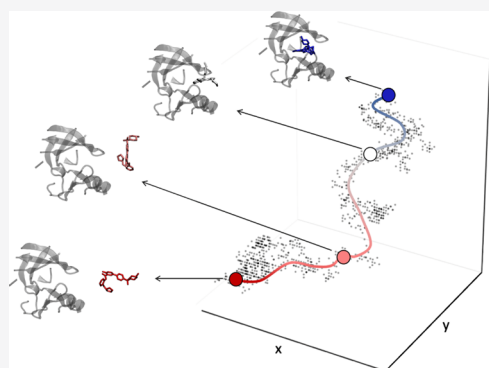
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** Computational capabilities are rapidly increasing, primarily because of the availability of GPU-based architectures. This creates unprecedented simulative possibilities for the systematic and robust computation of thermodynamic observables, including the free energy of a drug binding to a target. In contrast to calculations of relative binding free energy, which are nowadays widely exploited for drug discovery, we here push the boundary of computing the binding free energy and the potential of mean force. We introduce a novel protocol that leverages enhanced sampling, machine learning, and ad hoc algorithms to limit human intervention, computing time, and free parameters in free energy calculations. We first validate the method on a host–guest system, and then we apply the protocol to glycogen synthase kinase 3 beta, a protein kinase of pharmacological interest. Overall, we obtain a good correlation with experimental values in relative and absolute terms. While we focus on protein–ligand binding, the strategy is of broad applicability to any complex event that can be described with a path collective variable. We systematically discuss key details that influence the final result. The parameters and simulation settings are available at PLUMED-NEST to allow full reproducibility.



## 1. INTRODUCTION

In (bio)chemistry, free energy is still the most relevant and challenging physiochemical parameter to predict computationally. When studying the formation of biomolecular complexes under equilibrium conditions, the binding free energy is directly related to the affinity of the interacting partners. In drug discovery, accurate binding free energy estimations (within 1 kcal/mol) are crucial to identifying novel drug candidates.<sup>1,2</sup> As such, a significant portion of the computer-aided drug discovery community is working to improve the accuracy, precision, and robustness of binding free energy predictions by refining the force field parameters<sup>3–9</sup> and enhancing the sampling of slow degrees of freedom.<sup>10,11</sup> In this context, enhanced sampling algorithms<sup>12,13</sup> are increasingly combined with machine learning for more accurate free energy predictions.<sup>14–17</sup>

The advances in free energy perturbation (FEP) have enabled the frequent application of FEP in drug discovery to estimate the relative binding free energy (RBFEE).<sup>18–20</sup> When FEP simulations are applied to predict RBFEEs, the ligand is alchemically transformed into another one through intermediate steps. Because free energy is a state function, the choice of the intermediate states is arbitrary, making the approach very flexible.<sup>21,22</sup> Recent progress in computer hardware and software has made it feasible to apply FEP (or other alchemical) methodologies to absolute binding free

energy (ABFE) predictions,<sup>1,23–26</sup> creating the possibility of directly comparing the binding affinities across chemically different molecules that bind the same target or targets of the same family.<sup>27,28</sup> Although attractive, the routine application of FEP approaches to ABFE calculations is still limited because they do not fully consider how key phenomena (e.g., induced fit and desolvation) contribute to the binding affinity.<sup>29,30</sup> Their broad application to drug discovery is also limited by the higher computational cost of ABFE relative to RBFEE studies. Additionally, FEP provides minimal details about binding intermediates, transient pockets, and molecular mechanisms because these calculations rely on unphysical paths (i.e., the alchemical transformations).

A comprehensive representation of protein–ligand binding events can be provided by free energy methods based on physical paths, including steered molecular dynamics (MD) (via Jarzynski’s equation<sup>31</sup>), umbrella sampling,<sup>32</sup> metadynamics,<sup>33,34</sup> and so on. With these methods, one can simulate the complete association/dissociation process of a drug

Received: February 18, 2021

Published: July 14, 2021



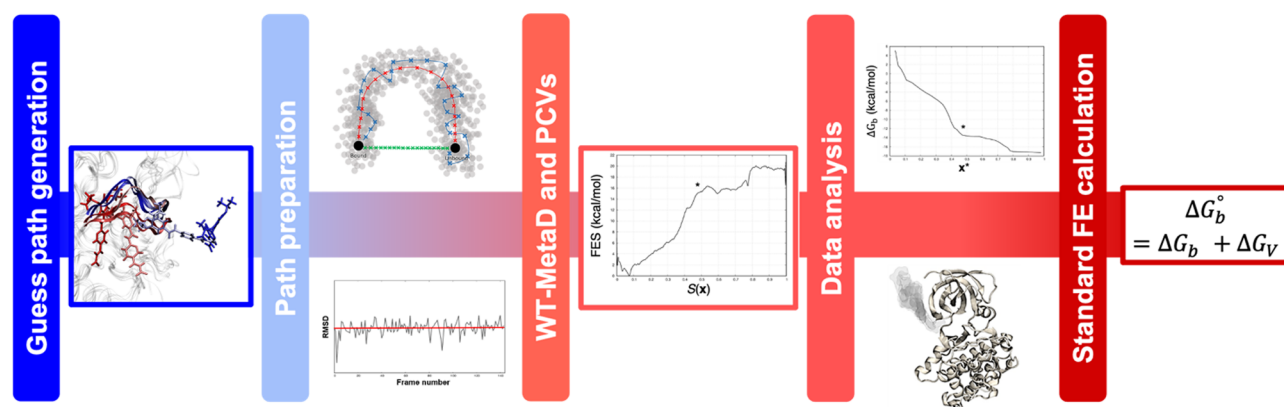


Figure 1. Graphical summary of the operational workflow.

binding to a target in explicit water, taking the structural flexibility of the receptor into account. The free energy is then calculated from the potential of mean force (PMF), leading to good thermodynamic and kinetic estimations.<sup>35</sup> Although attractive, there are at least two key challenges to be addressed to make these approaches widely applicable to drug discovery: (i) the choice of the “optimal” collective variables (CVs) to recapitulate the physical path and capture the slow degrees of freedom of the system and (ii) the definition of the operational workflow to set up, run, and analyze the simulations and eventually provide thermodynamics and kinetics data. For point (i), we here exploit the path CVs (PCVs),<sup>36</sup> which have been extensively used to study protein–ligand binding. For point (ii), a simple, complete, and semiautomatic approach to path-based applications has not yet been reported, although attempts in this direction date back to 2010.<sup>37</sup> We here devise an operational workflow that encompasses: (i) an enhanced sampling method to generate an initial guess path (we use adiabatic bias molecular dynamics unbinding simulations (ABMD)<sup>38</sup> or steered MD); (ii) a machine-learning method to extract an approximate minimum free energy path from the initial guess path; (iii) a steered MD-based ad hoc method here introduced to make uniform the root mean square displacement (RMSD) between consecutive frames to eventually define the PCVs; (iv) well-tempered metadynamics<sup>39</sup> (WT-MetaD) using the PCVs to obtain the PMF; (v) a technique based on the solvent excluded surface to compute the standard volume correction via NanoShaper;<sup>40</sup> and (vi) the calculation of the standard binding free energy via the ratio of partition functions. We discuss this last aspect as reported by Doudou et al.,<sup>41</sup> because it requires identifying the frame discriminating between the bound and unbound states, and this choice influences the outcome. Here too, we suggest a system-independent and semiautomatic procedure to identify the most reliable separating frame by analyzing the binding free energy profile. The pipeline depicted in Figure 1 is validated by computing the binding affinities of two diverse series of compounds, targeting two well-known benchmark systems, namely, a host–guest (HST–GST) complex<sup>43</sup> and the glycogen synthase kinase 3 beta (GSK-3β),<sup>44</sup> a system of pharmacological interest. The computational results and experiments correlated well. In addition to evaluating the RBFE correlations, we also discuss the accuracy of our estimates in absolute terms. This semiautomatic method is of wide applicability for path-based free energy methods, limiting the number of free parameters and human intervention.

## 2. METHODS

Here, we introduce a computational strategy to compute the PMF along the PCVs.<sup>36</sup> While the proposed protocol is of general applicability, we focus on the protein–ligand binding free energy. The procedure is detailed in the following paragraphs, and it can be summarized in the main steps listed below:

- i) Generation of a MD trajectory describing the rare event under investigation, for example, the association/dissociation of protein–ligand complexes;
- ii) Identification of a preliminary minimum free energy path by a machine-learning path-finding algorithm<sup>42</sup> and optimization of the distance between consecutive frames (i.e., RMSD) by the equidistant waypoints algorithm (reported here for the first time);
- iii) Reconstruction of the PMF by WT-MetaD on PCVs;
- iv) Estimation of the standard binding free energy by processing the PMF plus the standard volume correction via a NanoShaper-based<sup>40</sup> technique, purposely developed for the present study.

### 2.1. Path Generation via Enhanced MD Simulations.

The characterization of the free energy profiles underlying binary complexes’ dissociation processes was considered a testbed for our computational strategy. In particular, we considered two sets that have been well characterized by both experiments and computations: the cucurbit[8]uril (CB8) HST–GST system proposed in the SAMPL6 challenge and a congeneric series of ATP-competitive inhibitors against the GSK-3β.<sup>43,44</sup> In Section 2.5, we report the standard protocol used to set up both systems. To generate the putative dissociation paths connecting the bound and unbound states of the ligand, steered MD<sup>45,46</sup> and ABMD<sup>38</sup> simulations were performed on the CB8 and GSK-3β systems, respectively. In the steered MD simulations of the HST–GST complexes, the center of mass (COM) of the guest (GST) was pulled out from the CB8 cavity by applying a spring constant of 2000 kJ mol<sup>-1</sup> nm<sup>-2</sup> and a pull rate of 0.0001 nm ps<sup>-1</sup>, except for G2 and G3, whose larger molecular structures required a slightly higher pull rate. When asymmetric GST molecules are involved, there are slight differences between the two PMF profiles projected over both exit directions from the CB<sub>n</sub> cavity, as reported in previous studies on CB<sub>n</sub> complexes.<sup>47,48</sup> Hence, we performed two independent steered MD simulations along the two exit directions from CB8 for the asymmetric GST molecules included in our selection (G2 and G3). After 10 ns of steered MD simulations, a final COM

distance between CB8 and the GST molecules of approximately 10 Å was achieved for all complexes. GROMACS 2016.<sup>49</sup> was used to perform the pulling simulations in the NVT ensemble. The dissociation processes for the congeneric series of pyrazine-derivative inhibitors of GSK-3 $\beta$  were previously studied by ABMD coupled with an electrostatics-driven CV (i.e., eLABMD).<sup>50</sup> eLABMD is an enhanced sampling simulation technique that smoothly drives the system toward the desired end state while minimally perturbing its natural evolution, determined by thermal fluctuations because of finite temperature.<sup>38</sup> As such, once the system-dependent force constant affecting the magnitude of the backward fluctuations of the reaction coordinate is tuned correctly, a rare event can be accurately sampled, including the metastable states. Here, we selected one of the 20 unbinding trajectories reported in ref 50, by considering: (i) the computational unbinding time near the average one, as defined in ref 49.; (ii) the achievement of the complete ligand solvation (assessed in this study by looking at the protein–ligand contacts within 6 Å); and (iii) the physical soundness of the dissociation pathway. Here, we evaluated the exit direction of the ligand and the time spent sampling the unbound, prebound, and bound states, whose relevance to this study is discussed in Section 4.

**2.2. Approximate Minimum Free Energy Path Finding and Optimization of the Interframe Distance with the Equidistant Waypoints Algorithm.** At this stage, enhanced sampling (or plain MD) trajectories are already available. To accelerate the path-building phase, we do not refine the path by running further simulations. Instead, we clean up the path using the available samples; that is, we find an approximate minimum free energy path. This strategy is very flexible because it does not require running further simulations (e.g., the string method), and it can deal with presampled trajectories from plain MD or enhanced sampling (in the latter case, more care is needed). The execution time of this step is negligible compared to a MD run. To find an approximate minimum free energy path, we use the principal path algorithm, as previously formulated.<sup>42</sup> Given a points cloud, this machine-learning method connects two points defined a priori in data space and tries to pass through the local support of the data distribution, capturing the most “abstract” morphing path between these points. The method searches for a smooth out-of-sample geodesic ruled by the data sample density. It was inspired by the string method,<sup>51</sup> but with several differences:

- The string method is an online method similar to a stochastic gradient descent which is run simultaneously to a MD simulation. The principal path method is a batch method and is applied a posteriori irrespective of the MD sampling technique.
- The principal path can be applied to any kind of data, provided a points cloud or distance matrix is available, making it a machine-learning method, particularly a kernel method.
- The string method has no variational formulation, whereas a functional form is explicit in the principal path method. Indeed, we have shown that the string method iterations minimize the principal path functional in an approximate way.<sup>42</sup>

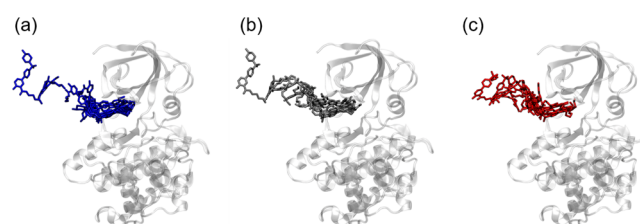
Finally, the method might also be seen as a plain elastic band,<sup>52</sup> where the potential function is substituted with the k-means cost function (discussed below).

The principal path algorithm is formally a regularized version of the k-means clustering algorithm. However, its purpose is significantly different as the principal path searches for a smooth one-dimensional manifold discretized by the waypoints. In particular, if we consider a set of points  $X = \{x_j \in R^d\}$ ,  $j = 1, \dots, N$ , and two points  $w_0, w_{N_c+1} \in R^d$ , the path connecting these two points is defined as an ordered set  $W$  of  $N_c$  waypoints  $w \in R^d$ . The principal path is found by minimizing the standard k-means cost function with the addition of a quadratic regularization term, which restrains the distance between consecutive waypoints and controls the level of smoothness of the path:

$$\min_W \sum_{i=1}^N \sum_{j=1}^{N_c} \|x_i - w_j\|^2 \delta(u_i, j) + s \sum_{i=0}^{N_c} \|w_{i+1} - w_i\|^2 \quad (1)$$

where  $\delta(u_i, j)$  is the Kronecker delta, which gives the membership of the  $i$ th sample to  $j$ th cluster/waypoint and  $u_i$  is a membership function that gives the cluster index; hence  $\delta(u_i, j)$  is different from 0 only if the  $i$ th sample belongs to the  $j$ th cluster. The first term coincides with the standard  $k$ -means cost function, and the second term introduces a set of harmonic restraints applied to consecutive waypoints. The hyper-parameter  $s$  regulates the trade-off between the data fitting and the smoothness of the path, as shown in Figure S1 in the Supporting Information.

By applying the path algorithm to real conformations sampled by MD, the closest physical frames of the original simulation to the ones calculated by the principal path are identified, thus defining a complete sequence of consecutive conformations capturing the event sampled by MD. At this stage, a clean path is available, but because of the peculiarities of molecular simulations, the distance in terms of RMSD between the neighboring snapshots identified is far from being uniform (Figure 2b). This aspect appears to be particularly



**Figure 2.** Graphical representation of the path preparation for one of the GSK-3 $\beta$  complexes studied in this work. (a) Guess unbinding path generated via eLABMD; (b) approximate minimum free energy path defined via the path-finding algorithm; (c) optimized unbinding path in terms of spacing between pairs of successive frames. A reduced number of frames are reported to simplify the representation of the system in the three stages.

relevant when the PCVs,<sup>36</sup>  $S(\mathbf{x})$  and  $Z(\mathbf{x})$ , are chosen to trace the principal path. As reported in the original paper introducing the PCVs,<sup>36</sup> consecutive frames must be as equidistant as possible to ensure the smooth progression of  $S(\mathbf{x})$  along the path and, more importantly, the proper mapping between the formal variable  $S(\mathbf{x})$  and the underlying metric space. Thus, we devised an algorithm based on a series of 20 ps-long steered MD simulations to make uniform the spacing between pairs of successive frames by placing additional and equidistant configurations as needed (Figure 2c). This automated procedure is close in spirit, even if devised

independently, to a procedure reported by Bernetti et al.<sup>53</sup> The principal path algorithm was developed in MATLAB, and the algorithm to make uniform the distance between consecutive frames was developed in Python 3 (see Algorithm 1 pseudocode in the Supporting Information). The code is available upon request. The steered MD simulations were run in GROMACS 2016.5<sup>49</sup> patched with PLUMED 2.5.<sup>54</sup>

The interframe distance was computed as the RMSD of the heavy atoms of the GST molecule for the CB8 complexes, while the RMSD alignment was performed on a selection of 16 C atoms defining the ring core of the CB8 molecule. For the GSK-3 $\beta$  system, the heavy atoms of the ligand and the protein residues located within 4 Å of the ligand in the bound state were considered for RMSD computation. For alignment, we used 25 C $_{\alpha}$  atoms belonging to residues that were uniformly distributed on the protein structure showing a small coordinate displacement during a plain MD simulation. The GSK-3 $\beta$  residues considered in this study for the RMSD alignment were Met162, Tyr163, Gln164, Leu165, Phe166, Arg167, Ser168, Leu169, Ala170, Tyr171, Ile172, Ser237, Ile238, Asp239, Val240, Trp241, Ser242, Ala243, Gly244, Cys245, Leu320, Leu329, Pro331, Leu332, and Ala334 according to PDB code 4ACC. The target RMSD threshold between consecutive frames along the path was set equal to 1 Å for all systems.

**2.3. WT-MetaD and PCVs.** The free energy surfaces underlying the unbinding processes under investigation were reconstructed by WT-MetaD<sup>39</sup> along the PCVs,  $S(\mathbf{x})$  and  $Z(\mathbf{x})$ :

$$S(\mathbf{x}) = \frac{\sum_{i=1}^P i e^{-\lambda \|\mathbf{x} - \mathbf{x}_i\|^2}}{\sum_{i=1}^P e^{-\lambda \|\mathbf{x} - \mathbf{x}_i\|^2}} \quad (2)$$

and

$$Z(\mathbf{x}) = -\lambda^{-1} \ln \left( \sum_{i=1}^P e^{-\lambda \|\mathbf{x} - \mathbf{x}_i\|^2} \right) \quad (3)$$

In eqs 2 and 3,  $\mathbf{x}$  represents the current system configuration,  $\|\mathbf{x} - \mathbf{x}_i\|^2$  is the distance between the current configuration and the  $i^{\text{th}}$  frame of the path,  $P$  is the number of frames included in the original path, and  $\lambda$  is a parameter that modulates the smoothness of the path representation. Here,  $\lambda$  was set equal to 230.0 nm<sup>-2</sup> according to the following heuristic equation that only depends on quantities available before running WT-MetaD:

$$\lambda = \frac{2.3P}{\sum_{i=1}^P \|\mathbf{x} - \mathbf{x}_i\|^2} \quad (4)$$

WT-MetaD was initialized from the bound state and then ran to a maximum of 1  $\mu$ s. By fixing the maximum computing time to be invested in sampling the potential energy surfaces via WT-MetaD, we accumulated a total simulation time of 8  $\mu$ s and  $\sim$ 8.5  $\mu$ s for the CB8 and GSK-3 $\beta$  systems, respectively. Gaussians with a nominal height of 0.2 kcal/mol were used together with a bias factor of 15. The width of the Gaussians was set to 0.2 and 0.01 nm<sup>2</sup> along  $S(\mathbf{x})$  and  $Z(\mathbf{x})$ , respectively, whereas the available space along the  $Z(\mathbf{x})$  dimension was limited by placing a wall at  $Z(\mathbf{x})$  equal to 0.05 nm<sup>2</sup>. The Gaussians deposition time was set to 500 MD steps. All WT-MetaD simulations were performed using GROMACS 2016.5<sup>49</sup> patched with PLUMED 2.5.<sup>54</sup> WT-MetaD simu-

lations run on one GPU node (2 CPU Intel Xeon E5-2650 v4 @ 2.20GHz 12 Cores each, 2 NVIDIA Tesla P100-PCIE-12GB), performing 100 and 30 ns/day for the CB8 and GSK-3 $\beta$  systems, respectively. Simulation of one ligand of GSK-3 $\beta$  for 1  $\mu$ s of WT-MetaD costs approximately 30 days of one node computing time.

**2.4. Standard Binding Free Energy Computation.** The standard binding free energy,  $\Delta G_b^{\circ}$ , was calculated as reported by Doudou et al.<sup>41</sup>

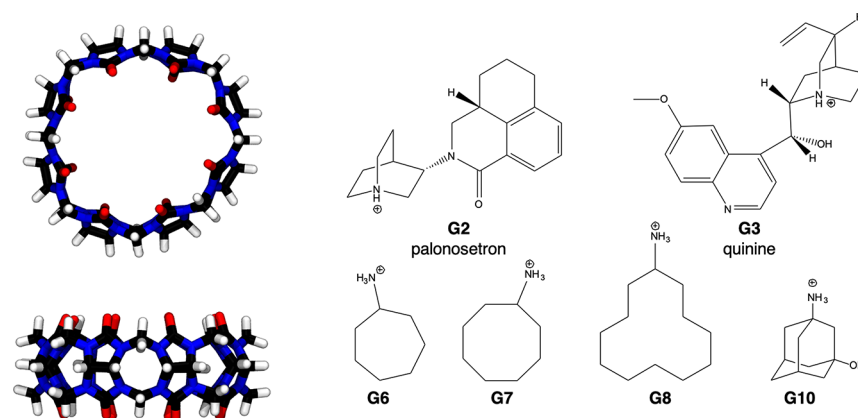
$$\Delta G_b^{\circ} = \Delta G_b + \Delta G_V = -RT \ln \left( \frac{Q_{\text{site}}}{Q_{\text{bulk}}} \right) - RT \ln \left( \frac{V_{\text{bulk}}}{V^{\circ}} \right) \quad (5)$$

The first term of eq 5 represents the probability ratio between the bound and unbound states of the ligand. The second term is the standard volume correction.  $Q_{\text{site}}$  and  $Q_{\text{bulk}}$  denote the partition functions for the bound and unbound regions, respectively, whose ratio is computed by integrating the FES as reconstructed by WT-MetaD,  $F(S, Z)$ , and defined to be zero at its lowest point (i.e., the ligand bound state) (eq 6). The frame separating the bound and unbound states was identified as the first inflection point obtained by plotting the binding free energy,  $\Delta G_b$ , as a function of the bound/unbound frame. In Sections 3.2.1 and 4, the procedure is further explained and discussed together with a representative example.

In detail, the argument of the logarithm in the first term is:

$$\frac{Q_{\text{site}}}{Q_{\text{bulk}}} = \frac{\int_{\text{site}} \exp \left( -\frac{F(S, Z)}{RT} \right) dS dZ}{\int_{\text{bulk}} \exp \left( -\frac{F(S, Z)}{RT} \right) dS dZ} \quad (6)$$

The second term of eq 5 quantifies the free energy for changing from the standard-state volume  $V^{\circ}$  equal to 1661 Å<sup>3</sup>, corresponding to a concentration of 1 M, to the sampled unbound volume,  $V_{\text{bulk}}$ . This correction term is needed to include in the free energy estimate the effect of the limited conformational space available to the ligand when in the unbound state. In this study, the unbound volume,  $V_{\text{bulk}}$ , was quantified using NanoShaper<sup>40</sup> (freely available at <https://gitlab.iit.it/SDecherchi/nanoshaper> and also available in the BiKi Life Sciences software package<sup>55</sup>), by considering the  $S(\mathbf{x})$  frames of the principal path describing the dissociated state of the binary complex. In detail, we collected all the frames belonging to the unbound state in a single pdb file, and then we computed the solvent excluded surface on this union of the ligand atoms. This union surface gives an accurate approximation of the volume traced by the ligand in the unbound state (Supporting Information). In the Results section, the standard binding free energy,  $\Delta G_b^{\circ}$ , for each complex refers to the time average of the last portion of each WT-MetaD simulation, whose length was determined by two conditions:<sup>56</sup> (i) in the considered window, the system is fully diffusive along  $S(\mathbf{x})$  and (ii) the residual height of the hills must be less than 10% of the initial height (i.e., 0.2 kcal/mol in this study). An estimate of the sampling error is computed as the standard error of the time fluctuation of  $\Delta G_b^{\circ}$  over the converged portion of each WT-MetaD simulation. All the data and PLUMED input files to reproduce the results are available on PLUMED-NEST ([www.plumed-nest.org](http://www.plumed-nest.org)), the public repository of the PLUMED consortium,<sup>57</sup> as plumID:21.004.



**Figure 3.** Top and side perspective views of the 3D structure of the CB8 host. Carbon atoms are represented in black, hydrogens in white, nitrogens in blue, and oxygens in red. GST molecules are shown as 2D chemical structures with an explicit protonation state.

**Table 1. Prioritization of the GST Molecules on Standard Binding Free Energies Obtained by WT-MetaD and ITC Experiments<sup>a</sup>**

GST ID	$\Delta G_b$	$\Delta G_v$	$\Delta G_b^\circ$	Rank <sub>comp</sub>	$\Delta G_{b,exp}^\circ$	Rank <sub>exp</sub>
G2	-7.7	0.0	-7.7 ± 0.1	6	-7.66 ± 0.05	5
G3	-9.6	-0.1	-9.7 ± 0.1	3	-6.45 ± 0.06	6
G6	-9.3	0.4	-8.9 ± 0.1	5	-8.34 ± 0.05	3
G7	-10.8	0.3	-10.6 ± 0.1	2	-10.0 ± 0.1	2
G8	-12.9	0.4	-12.5 ± 0.1	1	-13.50 ± 0.04	1
G10	-9.4	0.2	-9.1 ± 0.1	4	-8.22 ± 0.07	4

<sup>a</sup>Free energy terms (i.e.,  $\Delta G_b$ ,  $\Delta G_v$ ,  $\Delta G_b^\circ$ , and  $\Delta G_{b,exp}^\circ$ ) are reported in kcal/mol. Pearson correlation coefficient: 0.84. Bootstrap Pearson correlation coefficient: 0.72 ± 4e-5 (bootstrap standard error and 10,000 samples). Spearman coefficient: 0.6. RMSE: 1.5 kcal/mol. ME: 0.7 kcal/mol. The experimental data refer to Rizzi et al.<sup>43</sup>

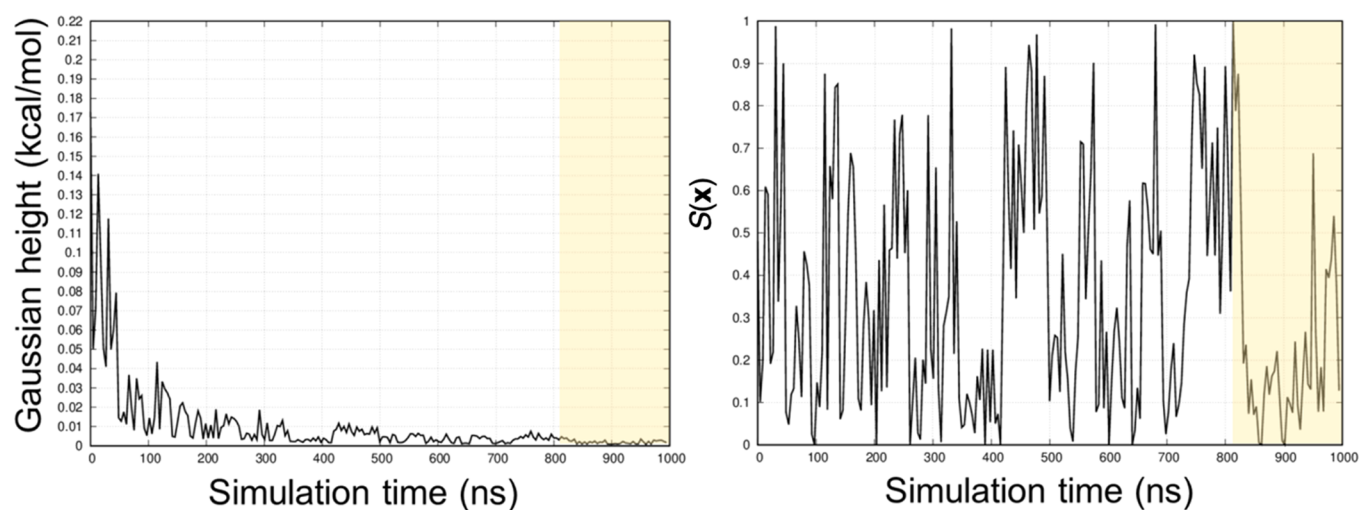
**2.5. System Setup.** All the CB8 complex structures were obtained from the SAMPL6 repository.<sup>43</sup> The host and GST molecules were modeled according to the general Amber force field (GAFF)<sup>58</sup> version 1.8. The AM1-BCC<sup>59,60</sup> point charges were used as supplied by the SAMPL6 organizers. Water is described by the TIP4P-Ew model.<sup>61</sup> Sodium and chloride ions are added to neutralize the system and to maintain the corresponding experimental ionic strength at which the experimental binding affinity was measured, that is, 25 mM Na<sub>3</sub>PO<sub>4</sub> buffer at pH 7.4.<sup>43</sup> All complexes are minimized by 5000 steepest descent steps and then equilibrated. The equilibration protocol requires the thermalization of the system at 300 K in three steps using the Bussi–Parrinello thermostat<sup>62</sup> for a total of 0.3 ns of dynamics. Subsequently, 1 ns of MD in the NPT ensemble is performed until the average pressure of the system is equilibrated to 1 atm according to the Parrinello–Rahman barostat.<sup>63</sup> All MD simulations are performed using GROMACS 2016.5<sup>49</sup> patched with PLUMED 2.5.<sup>54</sup> Production runs are performed in the NVT ensemble, setting 2 fs as the time step. Velocities are randomly assigned before each production run. A cutoff of 12 Å was used for nonbonded interactions, while long-range electrostatic interactions were treated with the particle mesh Ewald<sup>64</sup> scheme, using a grid spacing of 1.6 Å. A temperature of 300 K was controlled using the V-rescale thermostat,<sup>62</sup> while bond lengths for chemical bonds involving hydrogens were restrained to their equilibrium values with the LINCS algorithm.<sup>65</sup> For the detailed protocol used to set up the GSK-3 $\beta$  systems, we refer the reader to ref 50.

### 3. RESULTS

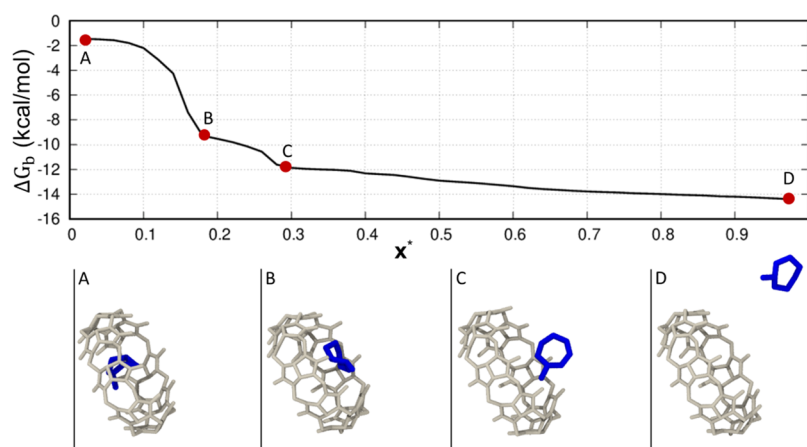
First, we describe the computational strategy applied to the HST–GST system (the CB8–G6 complex). We detail the procedure implemented to identify the  $S(\mathbf{x})$  frame to compute the rate of partition functions between the bound and unbound regions (eq 6). In the second part of this section, we outline the results for the GSK-3 $\beta$  complex system.

**3.1. HST–GST Benchmark System.** As a testbed for our computational strategy, we chose the HST–GST system from the cucurbit[*n*]uril (CB*n*) family, proposed in the SAMPL6 binding challenge. We identified six positively charged GST molecules (Figure 3) displaying a wide range of binding affinities for the host, from -13.5 to -6.45 kcal/mol (Table 1), including a few cases whose experimental binding free energies differed by less than 1 kcal/mol. The highly symmetric CB8 comprises eight identical glycouril monomers linked by pairs of methylene bridges, resulting in its characteristic ring shape (Figure 3). Because of their top-bottom symmetry, asymmetric GSTs have at least two symmetry-equivalent binding modes. For this reason, G2 and G3 were included in our selection. All HST–GST complexes included in the data set display 1:1 experimental stoichiometry.

**3.2. HST–GST Binding Free Energy.** As previously mentioned, steered MD was chosen as an enhanced simulation technique to generate preliminary paths of the HST–GST systems. The path algorithm<sup>42</sup> was subsequently applied to identify an approximate minimum free energy path, describing the dissociation process for every complex in the benchmark data set, thus detecting the milestone frames. Once defined, the principal path was then subjected to the steered MD-based procedure to optimize the definition of the PCVs. Each



**Figure 4.** Gaussian hills height (left) and  $S(x)$  progression (right) as a function of the simulation time for the CB8-G6 complex. For consistency, the reaction coordinate was rescaled, setting  $S(x) = 0$  (bound state) and  $S(x) = 1$  (unbound state) for every complex reported in this study. The shaded region refers to the portion of the WT-MetaD simulation considered in the computation of the binding free energy.



**Figure 5.** Sensitivity of the binding free energy,  $\Delta G_b$ , with varying  $x^*$  frame for the CB8-G6 system.

optimized principal path was sampled for 1  $\mu$ s with WT-MetaD combined with PCVs. Although the small size of the CB8 complexes could have required shorter simulations, we aimed to thoroughly inspect the statistics thanks to the relatively limited computational cost.

The WT-MetaD convergence was primarily assessed by looking at the downward trend of the Gaussian hills height as a function of the simulation time and the diffusing behavior of the  $S(x)$  variable over the simulation time. For the former, a residual height of the hills of approximately 10% of the initial height was considered an acceptable threshold to assess the WT-MetaD convergence.<sup>56</sup> Figure 4 reports the Gaussian height and the progression of the  $S(x)$  variable as a function of the simulation time for the CB8-G6 complex. For this system, we established the convergence after 800 ns of simulation, when both conditions were fulfilled. In the Supporting Information, we also report the evolution of the  $\Delta G_b$  along the simulation time, given its relevance when assessing the convergence of metadynamics-based simulations.

**3.2.1. Identification of the Bound/Unbound  $x^*$  Frame.** As outlined in eqs 5 and 6, the computation of the binding free energy requires the evaluation of the probability ratio between the bound and unbound states of the ligand, which in turn

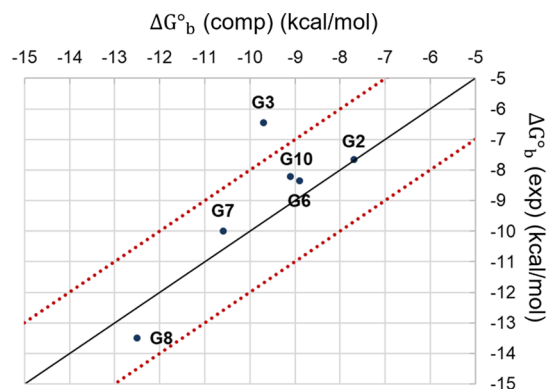
demands the identification of the  $S(x)$  frame discriminating the two states.

To establish a transferrable strategy for semiautomatically identifying the bound/unbound  $x$  frame (hereafter referred to as  $x^*$ ) for every CB8 complex in the data set, we monitored the behavior of the binding free energy,  $\Delta G_b$ , computed following eqs 5 and 6, changing the  $x^*$  frame.

Figure 5 reports the behavior of  $\Delta G_b$  as a function of  $x^*$  for the representative case of the CB8-G6 system, where the rescaled  $x^*$  equal to 0 corresponds to the docked state of the GST molecule (point A). Proceeding to the solvated state, the  $\Delta G_b$  smoothly decreases, displaying two main intermediate inflection points corresponding to two intermediate states, that is, one of the partially docked conformations (point B) and the partially solvated states (point C) of the G6 molecule, respectively. As expected, an almost constant  $\Delta G_b$  value is observed when the GST molecule is fully solvated, because of the energetically equivalent conformations adopted by the solvated GST molecule (point D). To identify the  $x^*$  discriminating the bound and unbound states of the system, we visually inspected the plot showing the  $\Delta G_b$  changing the  $x^*$  frame, and we picked the frame corresponding to the first inflection point from the bound state showing the ligand partially undocked from the binding site (point B in Figure 5).

This criterion is similar in spirit to the “elbow criterion” used in machine learning for selecting the “right” number of clusters in k-means. The physical meaningfulness of  $\alpha^*$  is also evaluated. Moreover, the relative position of  $\alpha^*$  is further cross-checked versus the free energy profile. Indeed, the picked  $\alpha^*$  has also to be compatible with a transition-state-like nature on the PMF (see Figure 8a in the Discussion). If these checks fail, in the general case, we search for the following inflection point until these criteria are satisfied. At this stage, this procedure is still manually curated, and further work is needed to render it completely programmatic. Once  $\alpha^*$  is identified and checked against the qualitative and quantitative criteria, the probability ratio between the docked and solvated states of the ligand and the binding free energy,  $\Delta G_b$ , can be computed (eqs 5 and 6). In the Discussion section, we compare the result obtained for CB8-G6 with another CB8 complex showing a different behavior of  $\Delta G_b$  in terms of  $\alpha^*$ , thus further discussing the strategy used to identify the  $\alpha^*$  frame, because this aspect is critical for the final result.

3.2.2. *Prioritization of the HST–GST Data Set on the Standard Binding Free Energy.* Table 1 and Figure 6 report



**Figure 6.** Scatter plot showing the experimental measurements for the HST–GST data set against the affinity predictions. The two red lines delimit the area within 2 kcal/mol from the diagonal (black line).

the prioritization of the GST molecules on the standard binding free energy resulting from the WT-MetaD simulations and the isothermal titration calorimetry (ITC) measurements released for the SAMPL6 challenge.<sup>43</sup> Each free energy term is labeled according to eq 5. The predicted standard binding free energy,  $\Delta G_b^{\circ}$ , is reported together with the corresponding standard error, namely, the rate between the standard deviation and the square root of the sample size. With respect to the two asymmetric GST molecules (i.e., G2 and G3), the  $\Delta G_b^{\circ}$  values are reported as the average of the binding free energy values of both the exit directions, with the details for both exit directions indicated in the Supporting Information. We evaluated the Pearson correlation coefficient as a measure of the statistical relationship between experimental and computational estimates. From the data set in Table 1, it was equal to 0.84. By running a bootstrap resampling procedure, we assessed the statistical robustness of the correlation. Considering 10,000 bootstrap samples, we obtained an average Pearson correlation coefficient equal to  $0.72 \pm 4e-5$  (i.e., the bootstrap standard error), confirming a good correlation between experimental and computational data. The Spearman coefficient was computed to assess the consistency of the prioritization of the binding affinities from computations and experiments

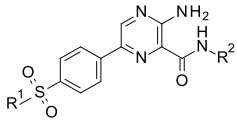
(Spearman coefficient: 0.6), because absolute values are important, but ranking is probably even more cogent in drug discovery campaigns. The data were analyzed according to ref 66. The root mean square error (RMSE) of the predicted binding free energy values with respect to the experimental results is 1.5 kcal/mol. It is worth mentioning that we selected a limited number of GST molecules from the original data set presented in the SAMPL6 challenge. Thus, it is difficult to provide a comprehensive comparison of the performance of our method relative to those of the SAMPL6 challenge reported in ref 43.

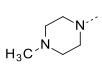
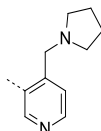
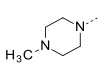

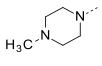
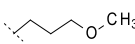
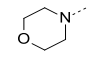

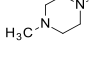
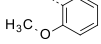
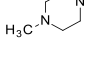
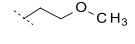
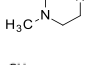
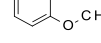
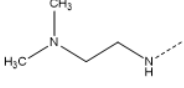

The computational ranking of the CB8 series reported in Table 1 is in fairly good agreement with the experimental one, although with some relevant deviations (see G3, G6, and G10). However, G6 and G10 differ in experimental binding free energy by less than 0.2 kcal/mol, a quantity challenging to predict by computational methods and often within the experimental error. As reported in Table 1, the computational data tend to overestimate the binding free energy (mean error, ME: 0.7 kcal/mol), except for G8. This observation is not surprising, according to the SAMPL6 results obtained by applying methods relying on empirical force fields (i.e., GAFF) to predict the binding free energies of the HST–GST complexes.<sup>43</sup> The accuracy between the computational and experimental datasets reported in Table 1 is around 1 kcal/mol for all the GST molecules except G3 for which the predicted and experimental  $\Delta G_b^{\circ}$  values differ by more than 3 kcal/mol. This deviation might be due to several aspects, such as G3's large chemical structure making convergence more difficult.<sup>67</sup> G3 might also have access to a second probable protonation state in water at the experimental pH,<sup>43</sup> affecting the CB8-G3 binding affinity. Nevertheless, our binding free energy estimate for CB8-G3 is in line with previous computational results relying on other sampling methods, for example, double decoupling method<sup>70</sup> and umbrella sampling<sup>47,70</sup> that systematically overestimated the CB8-G3 binding affinity. Our approach reproduced the result for the CB8-G3 complex (Supporting Information) previously reported by Sun et al.<sup>67</sup> Here, the authors identified the presence of multiple free energy minima corresponding to stable bound states of the G3 molecule in complex with CB8. This result was obtained using WT-MetaD to sample the spherical coordinates,  $\rho$ ,  $\theta$ , and  $\varphi$ . By validating this challenging result against the application of different computational protocols to explore diverse CVs, we increased our confidence in the predictions obtained for the benchmark data set. In addition, we ensured their independence from the starting configuration of the system and the peculiar dissociation path generated by an arbitrary enhanced sampling technique. In the Discussion section, we report additional considerations for the test case of CB8 in complex with the asymmetric GST molecules (G2 and G3). In the Supporting Information, we report all the PMFs, the  $\Delta G_b^{\circ}$  as a function of the  $\alpha^*$ , and the significant plots assessing the convergence of the WT-MetaD simulations for all the CB8 systems.

**3.3. GSK-3 $\beta$  Kinase System.** We then applied the same protocol to a real case study of pharmaceutical interest, the protein kinase GSK-3 $\beta$ . The unbinding kinetics of a strictly congeneric chemical series of pyrazine derivatives was previously characterized by both experiments and computations (i.e., eLABMD). Here, we complete the characterization of those dissociation paths determined via eLABMD by computing the underlying free energy profiles, referring to

the thermodynamic experimental data reported by Berg et al.<sup>44</sup> The chemical structures of the selected ATP-competitive inhibitors of GSK-3 $\beta$  are reported in Table 2. Compounds 1

**Table 2. Chemical Structures of the Selected GSK-3 $\beta$  Inhibitors (1–8)**



Cpd	R <sup>1</sup>	R <sup>2</sup>
1		
2		
3		
4		
5		
6		
7		
8		

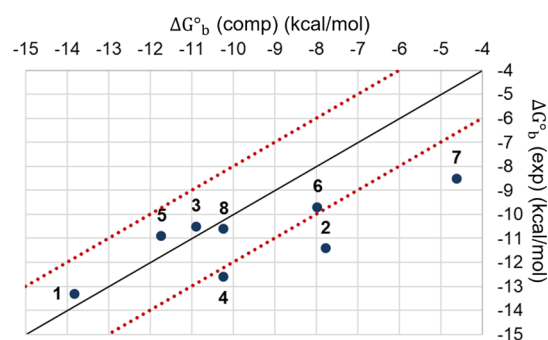
and 4 display two positive and zero charges, respectively. For the remaining inhibitors in the series, one positive charge was assigned to the nitrogen of the 4-methylpiperazine group (R<sup>1</sup>).

**3.4. Prioritization of the GSK-3 $\beta$  Data set on the Standard Binding Free Energy.** The dissociation paths for the GSK-3 $\beta$  complexes were generated via adiabatic bias MD coupled with an electrostatics-based CV, which provides a realistic description of the unbinding processes while sampling the metastable states of the system. This is particularly relevant in this kind of application because we observed that the proper sampling of the intermediate configurations facilitates the identification of the  $\alpha^*$  frame, thus leading to more reliable binding free energy estimates (see Section 4). Once the principal path was identified and optimized in terms of spacing between consecutive frames, we collected 1  $\mu$ s of WT-MetaD along the PCVs for each GSK-3 $\beta$  complex and assessed the convergence of the WT-MetaD simulations as reported for the CB8 systems (Supporting Information). Table 3 and Figure 7 show the prioritization of the GSK-3 $\beta$  chemical series on the standard binding free energies. The Pearson correlation

**Table 3. Prioritization of the GSK-3 $\beta$  Inhibitors on Standard Binding Free Energies Obtained by WT-MetaD and Experiments<sup>a</sup>**

CPD ID	$\Delta G_b$	$\Delta G_V$	$\Delta G_b^o$	Rank <sub>comp</sub>	$\Delta G_b^o$ ,exp	Rank <sub>exp</sub>
1	-13.6	-0.2	-13.8 $\pm$ 0.1	1	-13.3	1
2	-7.1	-0.7	-7.8 $\pm$ 0.3	7	-11.4	3
3	-10.4	-0.5	-10.9 $\pm$ 0.0	3	-10.5	6
4	-10.1	-0.1	-10.2 $\pm$ 0.2	4	-12.6	2
5	-11.5	-0.2	-11.7 $\pm$ 0.1	2	-10.9	4
6	-7.6	-0.3	-8.0 $\pm$ 0.1	6	-9.7	7
7	-4.2	-0.4	-4.6 $\pm$ 0.8	8	-8.5	8
8	-9.9	-0.3	-10.2 $\pm$ 0.9	5	-10.6	5

<sup>a</sup>The free energy terms (i.e.,  $\Delta G_b$ ,  $\Delta G_V$ ,  $\Delta G_b^o$ , and  $\Delta G_b^o$ ,exp) are reported in kcal/mol. Pearson correlation coefficient: 0.78. Bootstrap Pearson correlation coefficient: 0.70  $\pm$  3e-5 (bootstrap standard error, 10,000 samples). Spearman coefficient: 0.6. RMSE: 2.2 kcal/mol. ME: -1.3 kcal/mol. The experimental data refer to Berg et al.<sup>44</sup>



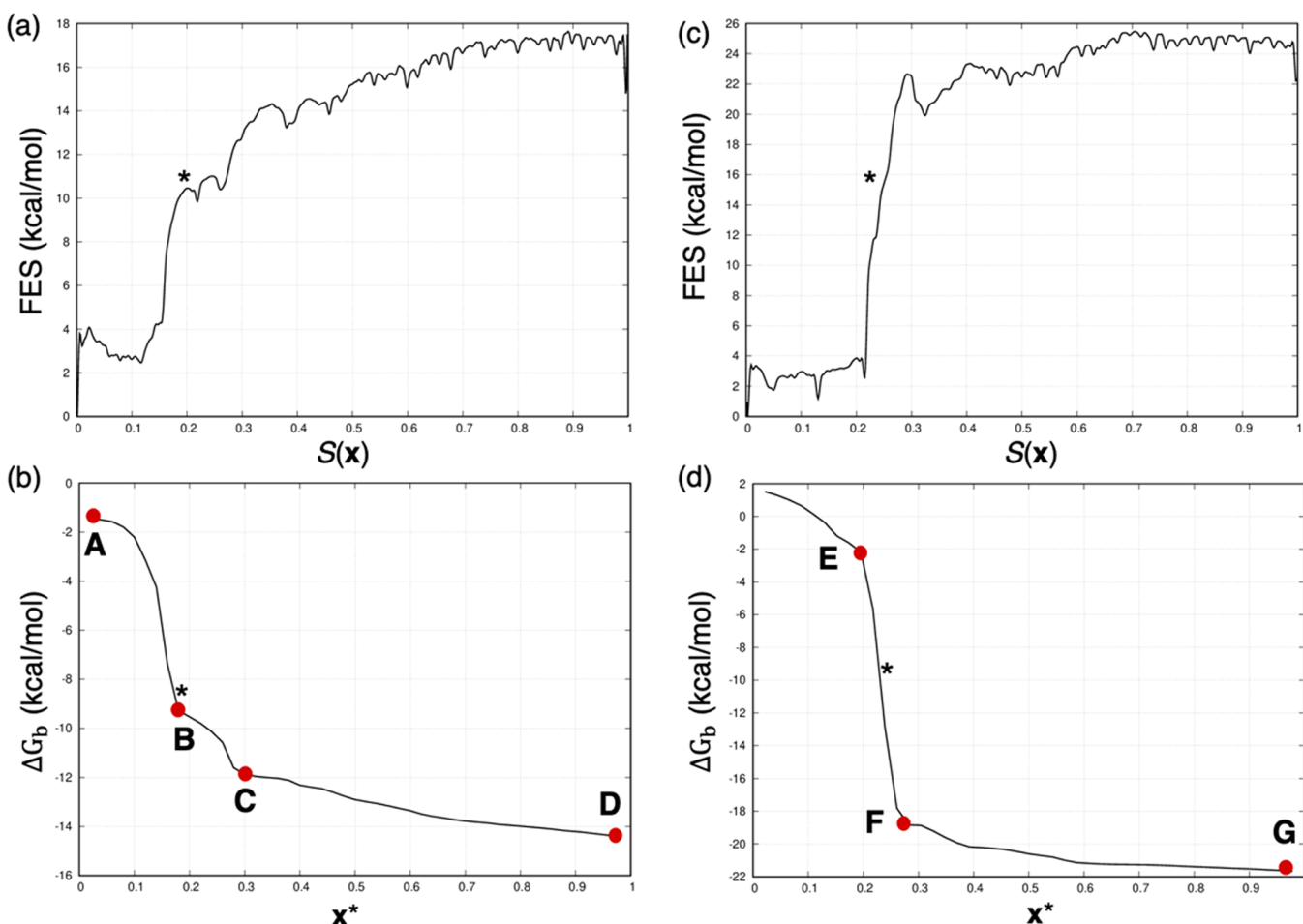
**Figure 7.** Scatter plot showing the experimental measurements for the GSK-3 $\beta$  series against the affinity predictions. The two red lines delimit the area within 2 kcal/mol from the diagonal (black line).

coefficient between computational and experimental values for the GSK-3 $\beta$  data set was 0.78. The bootstrapped estimate of the correlation coefficient was 0.70  $\pm$  3e-5 (bootstrap standard error, 10,000 samples). In absolute terms, the most critical cases in the data set are 2 and 7 followed by 4 and 6, further analyzed in the Discussion section. Consequently, the ranking of the GSK-3 $\beta$  series is not excellent, even though we generally observed a good agreement between computations and experiments (Spearman coefficient: 0.6). The RMSE for the predicted binding free energies relative to the experimental values is 2.2 kcal/mol; the ME results are equal to -1.3 kcal/mol suggesting the general trend toward underestimating the experimental binding free energies.

#### 4. DISCUSSION AND CONCLUSIONS

This study devises a semiautomated approach, of broad applicability, used here to compute the PMF and the standard binding free energy for protein–ligand complexes. The efficient and accurate estimation of the binding free energy remains one of the major open issues in computational drug discovery. In our protocol, a critical step was identifying the  $S(x)$  frame that separated the bound from the unbound states (here referred to as  $\alpha^*$ ) to provide a realistic partition function and thus an accurate standard binding free energy estimation. We did not fully automatize the  $\alpha^*$  identification procedure (this is currently in progress), and we used a collection of cross-checked heuristics for analyzing the plot of the  $\Delta G_b$  versus  $\alpha^*$ , the corresponding free energy profile, and the



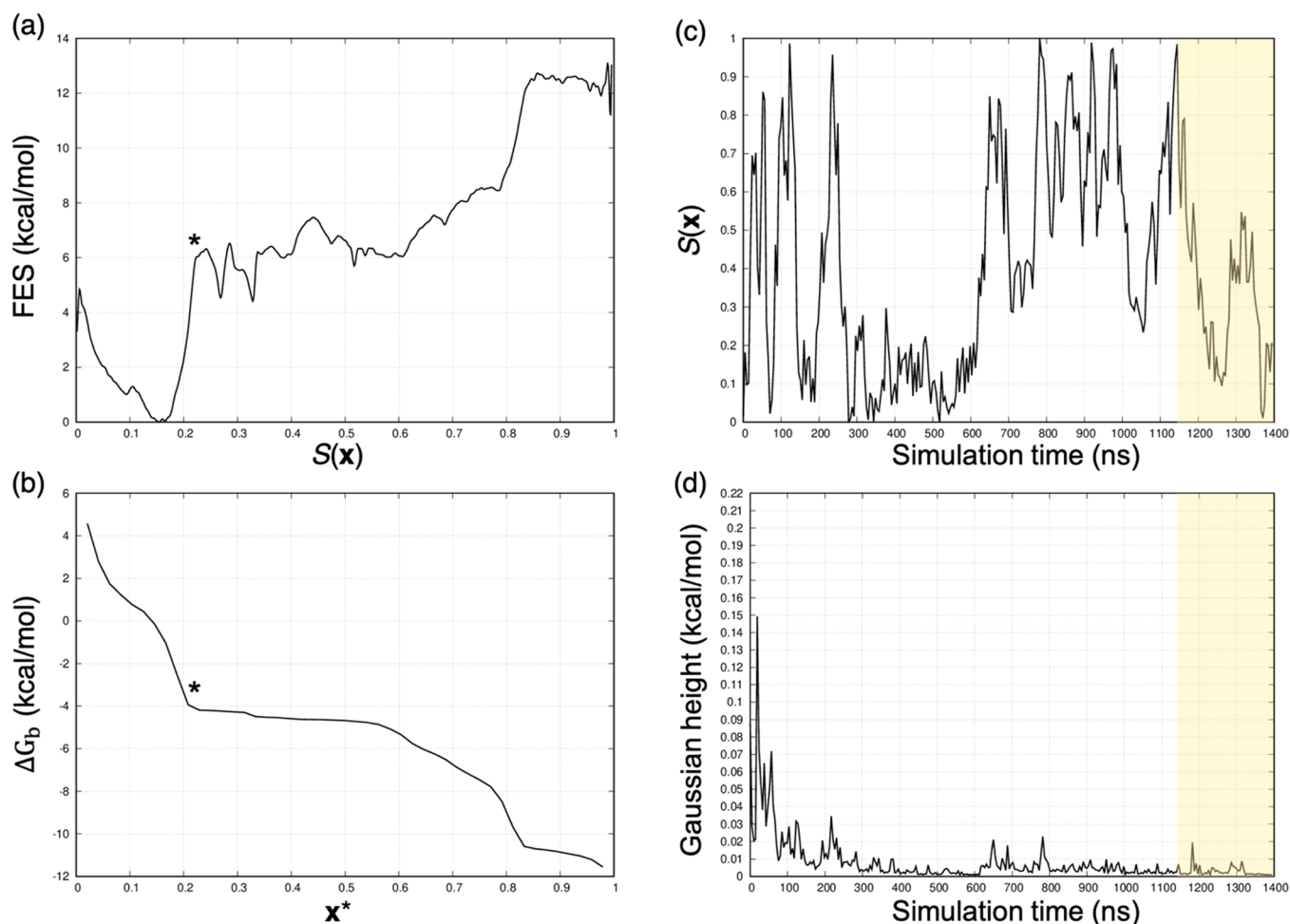


**Figure 8.** Free energy profiles and evolution of the binding free energy,  $\Delta G_b$ , changing the  $x^*$  frame for the CB8-G6 (a and b) and the CB8-G8 (c and d) systems. For CB8-G6, the  $x^*$  frame corresponds to the first inflection point encountered moving from the bound state (b, point B). In the PMF (a), it identifies the first energy barrier. For CB8-G8, because only one inflection point (d, point F) is observed between the bound (d, point E) and unbound (d, point G) states of the system, the  $x^*$  frame is fixed between the two well-defined inflection points (d, points E and F). In this case, in the PMF (c), the  $x^*$  frame corresponds to a not-sampled intermediate state between the lowest energy minimum (bound state) and the plateau corresponding to the solvated state of the system.

physical path. For the CB8-G6 case,  $x^*$  corresponds to a configuration of the system showing the breaking of the key HST–GST contacts together with the GST molecule partially undocked from the binding site (Figure 5 and Figure 8b). The  $x^*$  frame should, in principle, correspond to the transition state of the dissociation mechanism (Figure 8a), making our choice of  $x^*$  similar to the one proposed in previous studies.<sup>56,68</sup> It is worth mentioning that, even though our criterion to identify the  $x^*$  is valid in most cases, there are systems in which some additional considerations may be required. For example, if the first inflection point from the bound state on the  $\Delta G_b$  changing  $x^*$  plot was not compatible with a reasonable free energy change in the free energy profile, we moved to the following one until a transition-state-like free energy barrier was found. This was the case of 4 reported in Section 4 in the Supporting Information. Here, the first inflection point at  $S(x) = 0.2$  did not correspond to a significant free energy change, whereas this criterion was fulfilled considering the inflection point at  $S(x) = 0.4$ , which was picked as  $x^*$  for this system. Another critical case is CB8-G8 investigated here. In Figure 8d, we report the  $\Delta G_b$  for the CB8-G8 case, which shows a single fast intermediate inflection point (F) separating the docked (E) and solvated (G) states of the GST molecule. In these cases,

we heuristically observed that the average point between two well-defined inflection points corresponding to the bound and unbound states (E and F in Figure 8d) is an acceptable approximation of the  $x^*$  frame. We further highlight that identifying the  $x^*$  frame is far from trivial, and it is difficult to detect it by looking at the free energy barriers along the PMF. Thus, human intervention is needed to validate the choice of the  $x^*$  frame by visually inspecting the free energy path. Here, we observed that the steered MD protocol used to generate preliminary dissociation paths for all the HST–GST systems might not have properly sampled the intermediate metastable states for the CB8-G8 system. By failing to sample the transition state region between bound and unbound states, the principal path reconstructed for the CB8-G8 complex does not include any significant intermediate configurations that WT-MetaD can eventually sample. As such, the corresponding PMF is very steep, making it challenging to define a proper  $x^*$  frame (Figure 8c). We emphasize that the  $x^*$  frame identification was far simpler with GSK-3 $\beta$ , where we used adiabatic bias MD to generate preliminary dissociation paths.

To increase the GST molecules' structural variability in the data set and to challenge our approach with noncongeneric compounds, two asymmetric GST molecules (G2 and G3)

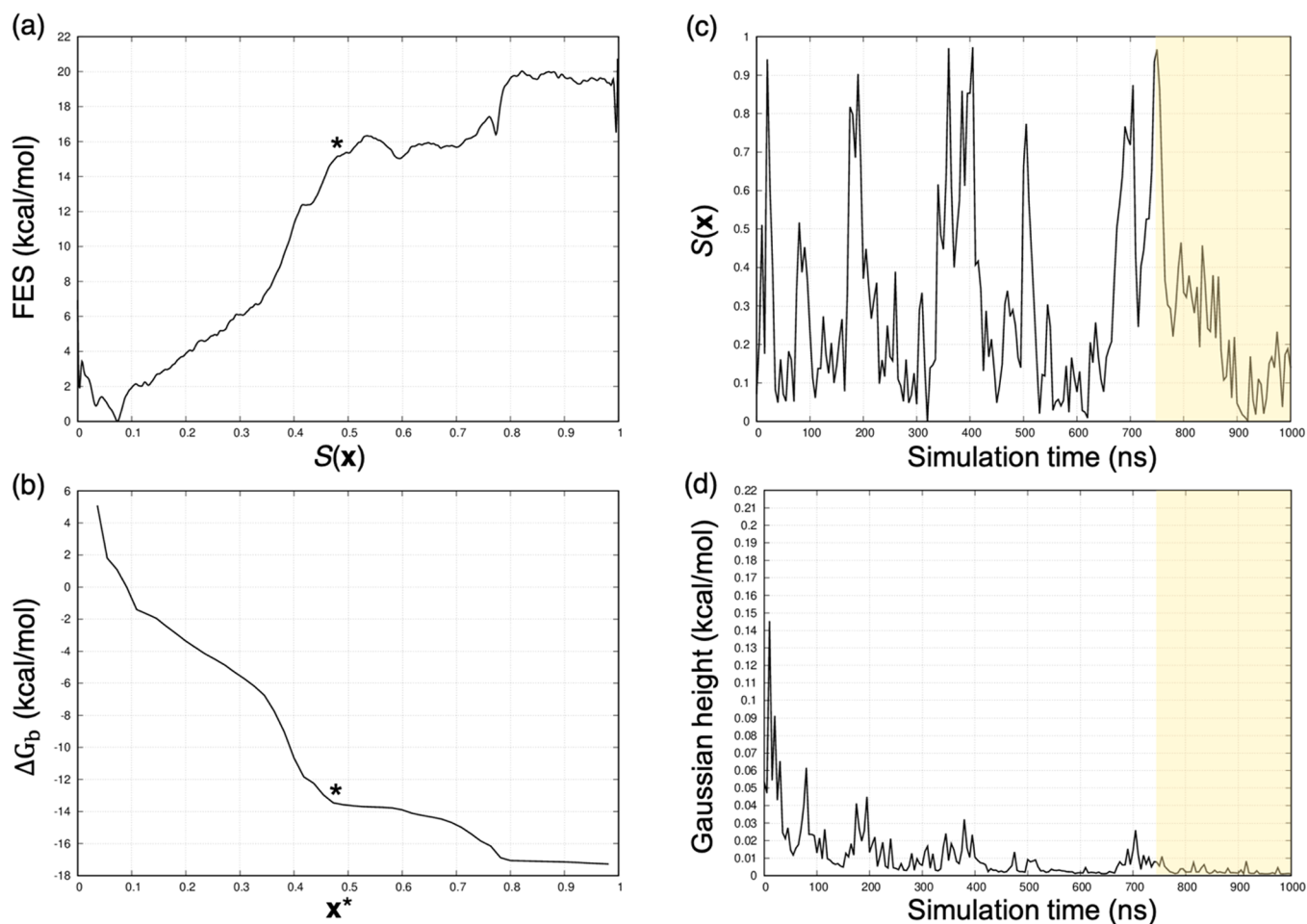


**Figure 9.** GSK-3 $\beta$  in complex with 7. (a) Free energy profile and (b) identification of the  $x^*$  frame. The behavior of (c)  $S(x)$  variable and (d) Gaussian hills against the simulation time. Looking at (c), we observe that the bound state ( $S = 0$ ) for 7 is not properly sampled after 1  $\mu$ s of WT-MetaD simulation. The shaded regions highlight the converged portion of the WT-MetaD simulation considered in the computation of the binding free energy reported in Table 3.

were also investigated. G2 and G3 also display the lowest binding affinities against the cucurbit[ $n$ ]uril host molecule CB8, which determines the partial dissociation of the G2 and G3 GST molecules during NPT equilibration, before the steered MD simulation. As anticipated in the Section 2, steered MD was repeated twice for the CB8 in complex with the asymmetric G2 and G3 to generate dissociation paths involving both the exit directions. Concerning this point, we observed that both G2 and G3 display different PMF profiles, when the GST molecule unbinds in either exit direction, suggesting possible kinetic barriers driving the system toward the most energetically favorable dissociation path. As expected, the different free energy profiles for the asymmetric G2 and G3 lead to binding affinity predictions that differ by less than 1 kcal/mol, thus increasing our confidence in the computational estimates (see Table S1 in the Supporting Information).

For the GSK-3 $\beta$  system, we evaluated the accuracy of the binding free energy estimations in absolute terms by comparing the computational estimate with the experimental reference for each complex. The most critical cases were 2 and 7 for which the free difference between the calculated and experimental values was more than 3.5 kcal/mol. By looking at the behavior of the  $S(x)$  variable, we observed that, after 1  $\mu$ s of WT-MetaD simulation, the bound state of 7 was not

correctly sampled (Figure 9c), possibly because of the suboptimal definition of the prior dissociation path. We would argue that the ABMD parameters that we selected for generating the dissociation paths with all the GSK-3 $\beta$  ligands were not appropriate for this compound. Indeed, we observed unphysical (high energy) ligand conformations during the unbinding event, probably because of the bias strength. We further highlight that the accuracy of the binding free energy estimates critically depends on how extensively the bound and prebound (intermediate) states have been sampled by WT-MetaD, as previously observed in several studies discussing binding free energy estimations for real systems. In light of this, we extended the WT-MetaD for 7 to 1.4  $\mu$ s until a complete transition along the  $S(x)$  path was detected, and the height of the Gaussian hills was low to the point of not allowing further exploration of the  $S(x)$  path (Figure 9c,d). In Table 3, the computational free energy estimate for 7 refers to the extended simulation. In Figure 9a,b, the free energy profile for the dissociation of GSK-3 $\beta$  in complex with 7 is reported together with the plot representing the behavior of the  $\Delta G_b$  against the  $x^*$  frame. Notably, the validation of the computational result against the experimental data might be affected by the experimental low solubility of 7 observed in references 44, 50, thus questioning the accuracy of the experimental



**Figure 10.** GSK-3 $\beta$  in complex with 1. (a) Free energy profile and (b) identification of the  $x^*$  frame. Behavior of (c)  $S(x)$  variable and (d) Gaussian hills against the simulation time. Panel (c) shows the diffusive behavior of the  $S(x)$  variable after 1  $\mu$ s of WT-MetaD simulation. The shaded regions highlight the converged portion of the WT-MetaD simulation considered in the computation of the binding free energy reported in Table 3.

thermodynamic data reported in ref 44. For comparison, Figure 10 reports the GSK-3 $\beta$ /1, where standard binding free energy was estimated with high accuracy.

By evaluating the exploration of the  $S(x)$  path for 2, we again observed that 1  $\mu$ s of WT-MetaD simulation was not enough to let the system extensively explore the complex bound state. However, in contrast to 7, the low height of the Gaussian hills for 2 meant we did not expect that proceeding further in the statistics would significantly change the exploration of the  $S(x)$  path (Supporting Information). As such, we did not extend the statistics for 2. Moreover, the gap of approximately 2 kcal/mol observed for 4 and 6 was evaluated by looking at the simulation convergence. Similarly to 2, we considered both simulations to have reasonably converged.

In conclusion, we introduced a semiautomated pipeline, which combines enhanced sampling simulations with a machine-learning method to predict standard binding free energies. As discussed by Mobley et al.,<sup>69</sup> validated benchmark systems are crucial to understanding how different computational methods perform when attempting to compute the same thermodynamic properties. As such, we tested our workflow on one of the HST–GST systems suggested in the SAMPL6 challenge. Then, the method was applied to a system of pharmaceutical relevance, namely, GSK-3 $\beta$ /ligand complexes.

In both cases, we obtained good binding affinity predictions in both relative and absolute terms. According to the present results, we were able to define the strengths and aspects that need to be improved to make this approach widely and routinely applicable to real drug discovery case studies. In particular, we would suggest using eLABMD as an enhanced sampling technique to define the guess paths. This is because eLABMD can provide an accurate description of the metastable conformations of the system when proper force constants are applied. A straightforward definition of the principal path, requiring minimum human intervention and negligible computational time, can then be obtained in combination with the equidistant waypoint algorithm, which prepares the path for WT-MetaD coupled with PCVs. Moreover, the system-independent procedure implemented to identify the  $x^*$  frame allows one to obtain robust and accurate binding free energy estimates (through the partition function) provided that WT-MetaD simulations are converged. We are working on the automated identification of  $x^*$  based on the finite difference approximation of the derivative of the PMF and a threshold free energy value. Finally, NanoShaper is a valuable tool for computing the sampling volume of the unbound state, thus allowing the accurate estimation of the binding free energy correction. The only step of the workflow that needs to be further investigated is the care in creating an “optimal”

physical path that mimics a minimum free energy path as much as possible. Steered MD is inadequate, and eLABMD proves much better. However, not all eLABMD trajectories are “optimal”, and it is always desirable to increase the “gentleness” of ligand release. We mention here that predicting absolute values via path-based free energy methods is far more computationally expensive and challenging relative to alchemical approaches widely applied in pharmaceutical settings. As such, physical path-based methods may not necessarily be more effective from the drug discovery viewpoint if only a number, namely, the free energy difference, is desired. The amount of information one extracts from the full PMF is not comparable with the output of alchemical methods. In the next future, we aim to apply the computational workflow depicted in Figure 1 to characterize rare events involving other chemical processes somewhat more complex than ligand unbinding.

## ■ ASSOCIATED CONTENT

### SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jctc.1c00177>.

(1) A description of the input matrix to the path-finding algorithm; (2) more details regarding the NanoShaper-based strategy employed to quantify the unbound volume; (3) additional computational data regarding the asymmetric GST molecules, G2 and G3, with relative plots; (4) for all the compounds not included in the main text, the plots representing: the free energy profiles, the  $\Delta G_b$  against the  $x^*$  frame, variation of  $S(x)$ , and the WT-MetaD Gaussian height along the simulation time; (5) plots representing the variation of  $\Delta G_b$  along the simulation time for all the investigated compounds; and (6) the pseudocode of the equidistant waypoints algorithm (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

**Sergio Decherchi** – Computational & Chemical Biology, Fondazione Istituto Italiano di Tecnologia, 16163 Genoa, Italy; BiKi Technologies s.r.l., 16121 Genoa, Italy; [orcid.org/0000-0001-8371-2270](https://orcid.org/0000-0001-8371-2270); Email: [sergio.decherchi@iit.it](mailto:sergio.decherchi@iit.it)

**Andrea Cavalli** – Computational & Chemical Biology, Fondazione Istituto Italiano di Tecnologia, 16163 Genoa, Italy; Department of Pharmacy and Biotechnology (FaBiT), Alma Mater Studiorum – University of Bologna, 40126 Bologna, Italy; [orcid.org/0000-0002-6370-1176](https://orcid.org/0000-0002-6370-1176); Email: [andrea.cavalli@iit.it](mailto:andrea.cavalli@iit.it)

### Authors

**Martina Bertazzo** – Computational & Chemical Biology, Fondazione Istituto Italiano di Tecnologia, 16163 Genoa, Italy; Department of Pharmacy and Biotechnology (FaBiT), Alma Mater Studiorum – University of Bologna, 40126 Bologna, Italy; Present Address: Global Research Informatics/Computational Chemistry, Evotec (France) SAS, 31100 Toulouse, France(M.B.)

**Dorothea Gobbo** – Computational & Chemical Biology, Fondazione Istituto Italiano di Tecnologia, 16163 Genoa, Italy; [orcid.org/0000-0001-9102-710X](https://orcid.org/0000-0001-9102-710X)

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jctc.1c00177>

### Author Contributions

<sup>1</sup>M.B. and D.G. contributed equally to the work.

### Author Contributions

M.B. ran the simulations, analyzed the data, contributed to the methods, and wrote the paper, D.G. ran the simulations, analyzed the data, contributed to the methods, and wrote the paper, S.D. designed the project and the methods, contributed to the analysis of the data, and wrote the paper, and A.C. designed the project and wrote the paper.

### Notes

The authors declare the following competing financial interest(s): Sergio Decherchi and Andrea Cavalli are co-founders and partners of BiKi Technologies, a startup company that develops methods based on MD and related approaches for investigating protein–ligand (un)binding.

## ■ REFERENCES

- (1) Decherchi, S.; Cavalli, A. Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation. *Chem. Rev.* **2020**, *120*, 12788–12833.
- (2) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59*, 4035–4061.
- (3) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, T. D.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 5595–5623.
- (4) Piana, S.; Robustelli, P.; Tan, D.; Chen, S.; Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein–Protein Complexes. *J. Chem. Theory Comput.* **2020**, *16*, 2494–2507.
- (5) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Miguels, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16*, 528–552.
- (6) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1863–1874.
- (7) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E4758–E4766.
- (8) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73.
- (9) Gkeka, P.; Stoltz, G.; Barati Farimani, A.; Belkacemi, Z.; Ceriotti, M.; Chodera, J. D.; Dinner, A. R.; Ferguson, A. L.; Maillet, J.-B.; Minoux, H.; Peter, C.; Pietrucci, F.; Silveira, A.; Tkatchenko, A.; Trstanova, Z.; Wiewiora, R.; Lelièvre, T. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *J. Chem. Theory Comput.* **2020**, *16*, 4757–4775.
- (10) Kuhn, M.; Firth-Clark, S.; Tosco, P.; Mey, A. S. J. S.; Mackey, M.; Michel, J. Assessment of Binding Affinity via Alchemical Free-Energy Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 3120–3130.
- (11) Loeffler, H. H.; Michel, J.; Woods, C. FESetup: Automating Setup for Alchemical Free Energy Simulations. *J. Chem. Inf. Model.* **2015**, *55*, 2485–2490.
- (12) Abrams, C.; Bussi, G. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy* **2014**, *16*, 163–199.

- (13) Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78* (14), 2690–2693.
- (14) Bernetti, M.; Bertazzo, M.; Masetti, M. Data-Driven Molecular Dynamics: A Multifaceted Challenge. *Pharmaceuticals* **2020**, *13*, 253.
- (15) Evans, R.; Hovan, L.; Tribello, G. A.; Cossins, B. P.; Estarellas, C.; Gervasio, F. L. Combining Machine Learning and Enhanced Sampling Techniques for Efficient and Accurate Calculation of Absolute Binding Free Energies. *J. Chem. Theory Comput.* **2020**, *16*, 4641–4654.
- (16) Rufa, D. A.; Bruce Macdonald, H. E.; Fass, J.; Wieder, M.; Grinaway, P. B.; Roitberg, A. E.; Isayev, O.; Chodera, J. D. Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning / molecular mechanics potentials. *bioRxiv* **2020**, No. 227959.
- (17) Scheen, J.; Wu, W.; Mey, A. S. J. S.; Tosco, P.; Mackey, M.; Michel, J. Hybrid Alchemical Free Energy/Machine-Learning Methodology for the Computation of Hydration Free Energies. *J. Chem. Inf. Model.* **2020**, *60*, 5331–5339.
- (18) Chen, H.; Maia, J. D. C.; Radak, B. K.; Hardy, D. J.; Cai, W.; Chipot, C.; Tajkhorshid, E. Boosting Free-Energy Perturbation Calculations with GPU-Accelerated NAMD. *J. Chem. Inf. Model.* **2020**, *60*, 5301–5307.
- (19) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.
- (20) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.
- (21) Shirts, M. R.; Mobley, D. L.; Chodera, J. D., Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? In *Annual Reports in Computational Chemistry*, Spellmeyer, D. C.; Wheeler, R., Eds. Elsevier: 2007; Vol. 3, pp. 41–59.
- (22) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150–160.
- (23) Qian, Y.; Cabeza de Vaca, I.; Vilseck, J. Z.; Cole, D. J.; Tirado-Rives, J.; Jorgensen, W. L. Absolute Free Energy of Binding Calculations for Macrophage Migration Inhibitory Factor in Complex with a Druglike Inhibitor. *J. Phys. Chem. B* **2019**, *123*, 8675–8685.
- (24) Jiang, W.; Roux, B. Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2010**, *6*, 2559–2565.
- (25) Cournia, Z.; Allen, B. K.; Beuming, T.; Pearlman, D. A.; Radak, B. K.; Sherman, W. Rigorous Free Energy Simulations in Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4153–4169.
- (26) Gobbo, D.; Ballone, P.; Decherchi, S.; Cavalli, A. Solubility Advantage of Amorphous Ketoprofen. Thermodynamic and Kinetic Aspects by Molecular Dynamics and Free Energy Approaches. *J. Chem. Theory Comput.* **2020**, *16*, 4126–4140.
- (27) Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. *J. Am. Chem. Soc.* **2017**, *139*, 946–957.
- (28) Singh, N.; Li, W. Absolute Binding Free Energy Calculations for Highly Flexible Protein MDM2 and Its Inhibitors. *Int. J. Mol. Sci.* **2020**, *21*, 4765.
- (29) Lapelosa, M.; Gallicchio, E.; Levy, R. M. Conformational Transitions and Convergence of Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 47–60.
- (30) Michel, J.; Essex, J. W. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 639–658.
- (31) Kerisit, S.; Parker, S. C. Free energy of adsorption of water and calcium on the [10 1 4] calcite surface. *Chem. Commun.* **2004**, 52–53.
- (32) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (33) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.
- (34) Cavalli, A.; Spitaleri, A.; Saladino, G.; Gervasio, F. L. Investigating Drug–Target Association and Dissociation Mechanisms Using Metadynamics-Based Algorithms. *Acc. Chem. Res.* **2015**, *48*, 277–285.
- (35) Pramanik, D.; Smith, Z.; Kells, A.; Tiwary, P. Can One Trust Kinetic and Thermodynamic Observables from Biased Metadynamics Simulations?: Detailed Quantitative Benchmarks on Millimolar Drug Fragment Dissociation. *J. Phys. Chem. B* **2019**, *123*, 3672–3678.
- (36) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, No. 054103.
- (37) Fidelak, J.; Juraszek, J.; Branduardi, D.; Bianciotto, M.; Gervasio, F. L. Free-energy-based methods for binding profile determination in a congeneric series of CDK2 inhibitors. *J. Phys. Chem. B* **2010**, *114*, 9516–9524.
- (38) Marchi, M.; Ballone, P. Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems. *J. Chem. Phys.* **1999**, *110*, 3697–3702.
- (39) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, No. 020603.
- (40) Decherchi, S.; Rocchia, W. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PLoS One* **2013**, *8*, e59744–e59744.
- (41) Doudou, S.; Burton, N. A.; Henchman, R. H. Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force. *J. Chem. Theory Comput.* **2009**, *5*, 909–918.
- (42) Ferrarotti, M. J.; Rocchia, W.; Decherchi, S. Finding Principal Paths in Data Space. *IEEE T. Neur. Net. Lear.* **2019**, *30*, 2449–2462.
- (43) Rizzi, A.; Murkli, S.; McNeill, J. N.; Yao, W.; Sullivan, M.; Gilson, M. K.; Chiu, M. W.; Isaacs, L.; Gibb, B. C.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 host-guest binding affinity prediction challenge. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 937–963.
- (44) Berg, S.; Bergh, M.; Hellberg, S.; Hogdin, K.; Lo-Alfredsson, Y.; Soderman, P.; von Berg, S.; Weigelt, T.; Ormo, M.; Xue, Y.; Tucker, J.; Neelissen, J.; Jerning, E.; Nilsson, Y.; Bhat, R. Discovery of novel potent and highly selective glycogen synthase kinase-3beta (GSK3beta) inhibitors for Alzheimer's disease: design, synthesis, and characterization of pyrazines. *J. Med. Chem.* **2012**, *55*, 9107–9119.
- (45) Izrailev, S.; Stepaniants, S.; Israilewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K., Steered molecular dynamics. In *Computational molecular dynamics: challenges, methods, ideas*; Springer: 1999; 39–65.
- (46) Grubmüller, H.; Heymann, B.; Tavan, P. Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **1996**, *271*, 997–999.
- (47) Song, L. F.; Bansal, N.; Zheng, Z.; Merz, K. M., Jr. Detailed potential of mean force studies on host-guest systems from the SAMPL6 challenge. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1013–1026.
- (48) Hsiao, Y. W.; Soderhjelm, P. Prediction of SAMPL4 host-guest binding affinities using funnel metadynamics. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 443–454.
- (49) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.
- (50) Gobbo, D.; Piretti, V.; Di Martino, R. M. C.; Tripathi, S. K.; Giabbai, B.; Storici, P.; Demitri, N.; Girotto, S.; Decherchi, S.; Cavalli, A. Investigating Drug-Target Residence Time in Kinases through Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **2019**, *15*, 4646–4659.
- (51) E, W.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B* **2002**, *66*, No. 052301.

(52) Elber, R.; Karplus, M. A method for determining reaction paths in large molecules: Application to myoglobin. *Chem. Phys. Lett.* **1987**, *139*, 375–380.

(53) Bernetti, M.; Masetti, M.; Recanatini, M.; Amaro, R. E.; Cavalli, A. An Integrated Markov State Model and Path Metadynamics Approach To Characterize Drug Binding Processes. *J. Chem. Theory Comput.* **2019**, *15*, 5689–5702.

(54) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comp. Phys. Comm.* **2014**, *185*, 604–613.

(55) Decherchi, S.; Bottegoni, G.; Spitaleri, A.; Rocchia, W.; Cavalli, A. BiKi Life Sciences: A New Suite for Molecular Dynamics and Related Methods in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58*, 219–224.

(56) Saladino, G.; Gauthier, L.; Bianciotto, M.; Gervasio, F. L. Assessing the Performance of Metadynamics and Path Variables in Predicting the Binding Free Energies of p38 Inhibitors. *J. Chem. Theory Comput.* **2012**, *8*, 1165–1170.

(57) The PLUMED Consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16*, 670–673.

(58) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(59) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(60) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

(61) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678.

(62) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, No. 014101.

(63) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(64) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(65) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(66) Mey, A. S. J. S.; Allen, B. K.; Macdonald, H. E. B.; Chodera, J. D.; Hahn, D. F.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best practices for alchemical free energy calculations [Article v.1.0]. *Living J. Comp. Mol. Sci.* **2020**, *2*, 18378.

(67) Sun, Z.; He, Q.; Li, X.; Zhu, Z. SAMPL6 host-guest binding affinities and binding poses from spherical-coordinates-biased simulations. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 589–600.

(68) Masetti, M.; Cavalli, A.; Recanatini, M.; Gervasio, F. L. Exploring complex protein-ligand recognition mechanisms with coarse metadynamics. *J. Phys. Chem. B* **2009**, *113*, 4807–4816.

(69) Mobley, D. L.; Gilson, M. K. Predicting binding free energies: Frontiers and benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531–558.

(70) Han, K.; Hudson, P. S.; Jones, M. R.; Nishikawa, N.; Tofoleanu, F.; Brooks, B. R. Prediction of CB[8] host-guest binding free energies in SAMPL6 using the double-decoupling method. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1059–1073.