Principled Approach for Computing Free Energy on Perturbation Graphs with Cycles

Xinqiang Ding* and John Drohan

Department of Chemistry, Tufts University, 62 Talbot Avenue, Medford, MA 02144

E-mail: Xinqiang.Ding@tufts.edu

Abstract

A common approach for computing free energy differences among multiple states is to build a perturbation graph connecting the states and compute free energy differences on all edges of the graph. Such perturbation graphs are often designed to have cycles. Because free energy is a function of states, the free energy around any cycle is zero, which we refer to as the cycle consistency condition. Since the cycle consistency condition relates free energy differences on edges of a cycle, it could be used to improve the accuracy of free energy estimates. Here we propose a Bayesian method called coupled Bayesian multistate Bennett acceptance ratio (CBayesMBAR) that can properly couple the calculations of free energy differences on edges of cycles in a principled way. We apply CBayesMBAR to compute free energy differences among harmonic oscillators and relative protein-ligand binding free energies. In both cases, CBayesMBAR provides more accurate results compared to methods that do not consider the cycle consistency condition. Additionally, it outperforms the cycle closure correction method that also uses cycle consistency conditions.

1 Introduction

It is often needed in computational chemistry to compute free energy differences among multiple thermodynamic states.[1] A common approach for this task is to build a perturbation graph connecting the states, and compute free energy differences on all edges of the graph.[2–6] The free energy difference between any two states can then be obtained by integrating along a path connecting the two states. When the perturbation graph has cycles, there will be multiple paths connecting two states. Because free energy is a function of states, the free energy difference between any two states should be independent of the path chosen. Equivalently, the free energy around any cycle is zero, which we refer to as the cycle consistency condition. The cycle consistency condition, on the one hand, could help diagnose potential systematic problems in the calculation.[7] For example, if the free energy around a cycle is significantly different from zero in practical calculations, it indicates that there might be systematic issues in sampling, such as the system being trapped in a meta-stable state.[8,9] On the other hand, if there are no systematic sampling issues, the cycle consistency condition can be used to improve the accuracy of free energy estimates.

Several methods have been proposed to use the cycle consistency condition to improve the accuracy of free energy estimates. They could be classified into two types. The first type computes the free energies on edges of a cycle independently and then adjusts them to satisfy the cycle consistency condition. Such methods include the cycle closure correction (CCC) method[7] and its weighted variant.[10] The second type couples the calculations of free energies on edges of a cycle and estimate them simultaneously. An example method of this type is the MBARnet[11] method, which computes free energies on edges of a cycle by optimizing an objective function subject to the cycle consistency condition. Although there is little evidence suggesting which type of methods is more accurate, the second type can potentially make better use of the cycle consistency condition because it uses the condition throughout the calculation.

The key to the second type of method lies in effectively coupling the calculations of free

energies along the edges of a cycle using the cycle consistency condition. Intuitively, this coupling should not significantly alter the free energy estimates on edges where the estimates are already highly precise. Instead, it should leverage these high-precision edges and the cycle consistency condition to refine the estimates on edges with lower precision. Here, we propose a probabilistic method that accomplishes this coupling, which we call the coupled Bayesian multistate Bennett acceptance ratio (CBayesMBAR) method. We applied CBayesMBAR to compute free energy differences among four harmonic oscillators and relative protein-ligand binding free energies. In both examples, CBayesMBAR provides more accurate results than those obtained without considering the cycle consistency condition. Furthermore, CBayesMBAR outperformed the widely used cycle closure correction method that also uses the cycle consistency condition. We have implemented CBayesMBAR as part of the BayesMBAR package,[12] which is freely available at `https://github.com/DingGroup/BayesMBAR`

2 Methods

CBayesMBAR is built upon the Bayesian multistate Bennett acceptance ratio (BayesMBAR) method,[12] a Bayesian generalization of the multistate Bennett acceptance ratio (MBAR) method.[13] In BayesMBAR, we formulate free energy estimation as a Bayesian inference problem[14] and derive a posterior distribution of free energy differences given sampled configurations. This posterior distribution is then used to estimate free energy differences and their uncertainties. The Bayesian framework of BayesMBAR naturally facilitates the coupling of multiple BayesMBAR calculations on a perturbation graph with cycles. Here, we first briefly review the BayesMBAR method and then explain how to couple multiple BayesMBAR calculations in CBayesMBAR using the cycle consistency condition.

2.1 BayesMBAR

Let $u_i(x)$, with i ranging from 1 to m, be the reduced potential energy functions[13] of m states. We aim to compute the free energy differences among these states by sampling configurations from their Boltzmann distributions. For the i-th state, we use $\{x_{ik}, k = 1, \ldots, n_i\}$ to represent the n_i configurations sampled from its Boltzmann distribution and we assume that these configurations are uncorrelated. We use $F = (F_1, \ldots, F_m)$ to denote the free energies of the m states. In BayesMBAR, we introduce y_{ik} as the index of the state from which configuration x_{ik} is sampled. Therefore, y_{ik} is equal to i. Although the indices of states for sampled configurations are determined during sampling, they are treated as random variables in BayesMBAR. Specifically, y_{ik} is viewed as a sample from a categorical distribution with parameters $\pi = (n_1/n, \ldots, n_m/n)$, meaning that the probability of sampling a configuration from the i-th state is $p(y = i) = \pi_i = n_i/n$, where $n = \sum_{i=1}^{m} n_i$ is the total number of configurations. The concatenation of state indices and configurations, denoted as (y, x), is viewed as samples from the conditional distribution $p(y, x|F)$, defined as

$$\begin{aligned} p(y = i, x|F) &= p(y = i) \cdot p(x|y = i; F) \\ &= e^{-[u_i(x) - F_i - \log \pi_i]}, \end{aligned} \tag{1}$$

for $i \in \{1, \ldots, m\}$. The free energy F, which was traditionally treated as a parameter in MBAR, is treated as a random variable in BayesMBAR and we assign a prior distribution $p(F)$ to it. Given the prior distribution $p(F)$ and the conditional distribution $p(y, x|F)$ described above, the joint distribution of (F, Y, X) is

$$\begin{aligned} p(F, Y, X) &= p(F) \cdot p(Y, X|F) \\ &= p(F) \cdot \prod_{i=1}^{m} \prod_{k=1}^{n_i} p(y_{ik} = i, x_{ik}|F), \end{aligned} \tag{2}$$

where $Y = \{y_{ik} : i = 1, ..., m; k = 1, ..., n_i\}$ and $X = \{x_{ik} : i = 1, ..., m; k = 1, ..., n_i\}$. The posterior distribution of F given sampled configurations and state indices is therefore

$$\begin{aligned} p(F|Y, X) &= \frac{p(Y|F, X)p(F)}{\int p(Y|F, X)p(F)dF} \\ &\propto p(F) \prod_{i=1}^{m} \prod_{k=1}^{n_i} p(y_{ik} = i|x_{ik}; F), \end{aligned} \tag{3}$$

where

$$p(y = i|x, F) = \frac{p(y = i, x|F)}{\sum_{j=1}^{m} p(y = j, x|F)} = \frac{e^{-[u_i(x) - F_i - \log \pi_i]}}{\sum_{j=1}^{m} e^{-[u_j(x) - F_j - \log \pi_j]}}. \tag{4}$$

BayesMBAR uses the posterior distribution (Eq. 3) to estimate the free energies and their uncertainties. Specifically, it estimates the free energy with either the mode or the mean of the posterior distribution, and computes the uncertainty of the estimate using the standard deviation of the posterior distribution.[15,16] When the prior distribution $p(F)$ is chosen to be a uniform distribution, the mode of the posterior distribution is identical to the MBAR estimate.[13] Our previous work has shown that BayesMBAR provides more accurate uncertainty estimates than the asymptotic analysis used in MBAR, especially when the number of configurations is small.[12] Futhermore, as a Bayesian method, BayesMBAR provides a principled way to incorporate prior information to improve the accuracy of free energy estimates.[12]

2.2 Coupled BayesMBAR

The Bayesian framework used in BayesMBAR enables the coupling of multiple BayesMBAR calculations on a perturbation graph with cycles. For example, consider three thermodynamic states, A, B, and C. To compute their free energy differences, we construct a perturbation graph connecting these states (Fig. 1). In this graph, red, orange, and blue circles represent the three end states, and arrows indicate the paths connecting them. Small black circles along the paths represent intermediate states between the end states. Configurations

are sampled from both the end states and the intermediate states along each path. We use X^h to denote the configurations sampled from states on the h-th path, where h ranges from 1 to 3, and Y^h to denote the state indices of X^h. In other words, X^h and Y^h correspond to X and Y in Eq. 2 and 3. We use F^h to denote the free energy of states on the h-th path, expressed as $F^h = (F^h_1, ..., F^h_{-1})$, where F^h_1 and F^h_{-1} are free energies of the two end states.

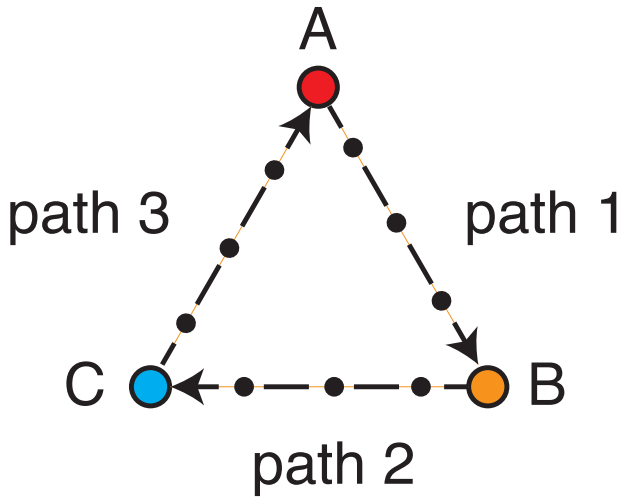

Figure 1: Perturbation graph connecting three thermodynamic states A, B, and C. The small black circles on each path represent intermediate states between the two end states.

In CBayesMBAR, we estimate all free energies of $F = (F^1, F^2, F^3)$ simultaneously using a Bayesian probabilistic framework. Specifically, we assign a prior distribution $p(F)$ to F and view all configurations and state indices $(\{Y^h, X^h\}_{h=1}^3)$ as samples from the conditional distribution $p(\{Y^h, X^h\}_{h=1}^3|F) = \prod_{h=1}^3 p(Y^h, X^h|F)$, where $p(Y^h, X^h|F)$ is defined similarly as in Eq. 2. Then the joint distribution of $(F, \{Y^h, X^h\}_{h=1}^3)$ is

$$p(F, \{Y^h, X^h\}_{h=1}^3) = p(F) \cdot \prod_{h=1}^{3} p(Y^h, X^h|F^h). \tag{5}$$

Based on the joint distribution, we derive the posterior distribution of F given sampled

configurations and state indices as

$$p(F|\{Y^h, X^h\}_{h=1}^3) \propto p(F) \prod_{h=1}^{3} p(Y^h, X^h|F^h). \tag{6}$$

CBayesMBAR uses the posterior distribution in Eq. 6 to estimate the free energies and their uncertainties.

Because the three paths form a cycle, the free energy $F = (F^1, F^2, F^3)$ must satisfy the cycle consistency condition: $F^1_{-1} - F^1_1 + F^2_{-1} - F^2_1 + F^3_{-1} - F^3_1 = 0$. In CBayesMBAR, we encode this condition as prior information in the prior distribution $p(F)$. Specifically, we set $p(F) = 0$ when the cycle consistency condition is violated. When the cycle consistency condition is satisfied, we assume no other prior information about F and set $p(F) \propto 1$, making $p(F)$ a uniform distribution over the space of F that satisfies the cycle consistency condition. Similar to BayesMBAR, CBayesMBAR estimates the free energy using either the mode or the mean of the posterior distribution and calculates the uncertainty using the standard deviation of the posterior distribution. The mode of the posterior distribution is computed by maximizing the posterior density, while the mean and standard deviation are estimated by sampling from the posterior distribution using Hamiltonian Monte Carlo methods.[15–17] Although we used a simple perturbation graph with one cycle in the above example for illustration, the same framework applies to perturbation graphs with multiple cycles, where multiple equality constraints are encoded in the prior distribution $p(F)$. Computational details on calculating the posterior mode, sampling from the posterior distribution, and handling perturbation graphs with multiple cycles are provided in the Supporting Information.

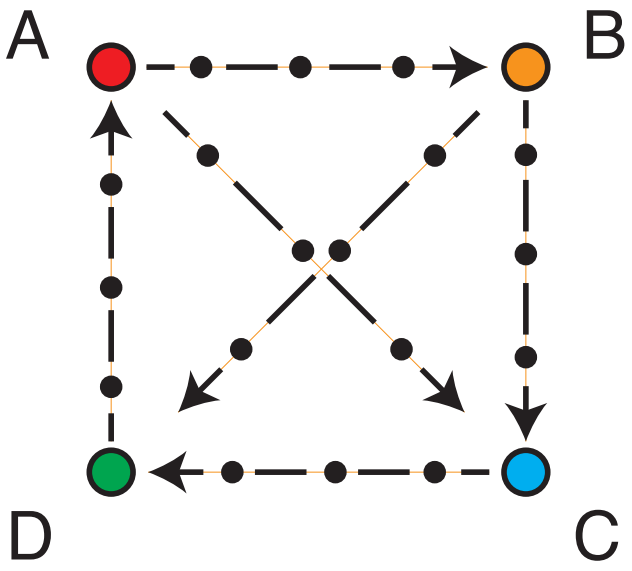

Figure 2: Perturbation graph connecting four harmonic oscillators: A, B, C and D. The small black circles on each path represent intermediate states between the two end states.

3 Results

3.1 Harmonic Oscillators

We begin by applying CBayesMBAR to calculate free energy differences among four 2-D harmonic oscillators, labeled A, B, C, and D (Fig. 1). Each oscillator has a reduced potential energy function given by $u_i(x) = \frac{1}{2}k_i\|x - \mu_i\|^2$, where k_i represents the force constant and μ_i denotes the equilibrium position for $i \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$. The assigned values are: $k_A = 9$, $k_B = 16$, $k_C = 25$, $k_D = 36$, and $\mu_A = (-1, 1)$, $\mu_B = (1, 1)$, $\mu_C = (1, -1)$, $\mu_D = (-1, -1)$. Our perturbation graph (Fig. 2) has paths connecting all state pairs, with three intermediate states introduced along each path. These intermediate states are also modeled as 2-D harmonic oscillators, with force constants and equilibrium positions linearly interpolated between the end states. We sample n configurations from each state on every path, varying n from 10 to 5000. Using these configurations, we estimate free energy differences for all paths using three methods: BayesMBAR, cycle closure correction, and CBayesMBAR. For BayesMBAR, we independently compute free energy differences for each path, using

the posterior mode as the free energy estimate and the standard deviation of the posterior distribution (estimated from 1000 samples) as the uncertainty. For CCC, we adjust the free energy differences obtained with BayesMBAR to satisfy the cycle consistency condition (details are in the Supporting Information). For CBayesMBAR, we simultaneously estimate free energy differences across all paths, coupling the calculations using the cycle consistency condition. As with BayesMBAR, we use the posterior mode for the free energy estimate and the standard deviation of the posterior distribution (estimated from 1000 samples) for uncertainty quantification.

To evaluate the three methods, we compare their free energy estimates against analytically computed exact results. Our first evaluation metric is the root mean square error (RMSE) of the free energy differences across all paths, defined as:

$$\sqrt{\frac{1}{6}\sum_{h=1}^{6}\left(F_{-1}^{h}-F_{1}^{h}-\Delta F_{\text{exact}}^{h}\right)^{2}}.$$

Here, $\Delta F_{\text{exact}}^{h}$ represents the exact free energy difference between the end states of the h-th path, and $F_{-1}^{h}-F_{1}^{h}$ is the estimated value. To ensure statistical robustness, we conduct 100 repetitions of the calculations, compute the mean RMSE over these repetitions, and perform paired t-tests to compare the three methods using the RMSEs obtained (Table 1 and Figures S1-S2). Across all sample sizes n, CBayesMBAR has significantly smaller RMSEs compared to BayesMBAR and CCC, with p-values less than 10^{-4} for all cases. Next, we assess the performance of the three methods in computing the free energy difference between the end states of individual paths using the mean absolute error (MAE) over the 100 repetitions as the metric (Table 2 and Table S1). In all cases, CBayesMBAR has the smallest MAEs among the three methods.

Both BayesMBAR and CBayesMBAR quantify uncertainties in free energy estimates using the standard deviation of the posterior distribution. Table S2 presents the mean estimated uncertainties for free energy differences between the end states of individual paths.

Table 1: The mean RMSE of free energy differences between end states of all paths for harmonic oscillators. All values are in units of k_BT.

n	BayesMBAR	CCC	CBayesMBAR
10	1.66	1.21	1.11
13	1.52	1.09	0.94
18	1.29	0.90	0.79
28	1.01	0.74	0.65
48	0.75	0.54	0.49
99	0.52	0.36	0.33
304	0.31	0.22	0.20
5000	0.08	0.06	0.05

Table 2: The mean absolute error of free energy differences between end states on individual paths for harmonic oscillators. All values are in units of k_BT.

path	n = 10			n = 28		
	BayesMBAR	CCC	CBayesMBAR	BayesMBAR	CCC	CBayesMBAR
A→B	0.85	0.90	0.77	0.49	0.51	0.43
B→C	1.10	0.95	0.88	0.68	0.68	0.61
C→D	1.39	1.11	1.09	0.79	0.66	0.62
D→A	1.14	1.01	0.95	0.62	0.58	0.50
A→C	1.59	1.08	1.01	0.87	0.61	0.57
B→D	1.99	1.22	1.04	1.38	0.81	0.64
path	n = 99			n = 5000		
	BayesMBAR	CCC	CBayesMBAR	BayesMBAR	CCC	CBayesMBAR
A→B	0.28	0.27	0.24	0.04	0.04	0.03
B→C	0.34	0.30	0.28	0.05	0.04	0.04
C→D	0.45	0.33	0.31	0.06	0.05	0.05
D→A	0.38	0.29	0.27	0.05	0.04	0.04
A→C	0.50	0.33	0.31	0.07	0.04	0.04
B→D	0.61	0.35	0.29	0.10	0.06	0.05

Consistently, CBayesMBAR yields smaller uncertainties compared to BayesMBAR, highlighting how coupling free energy estimates via cycle consistency conditions enhances estimation precision. The performance improvement of CBayesMBAR over BayesMBAR in estimating free energy differences along individual paths (Table 2) correlates closely with the uncertainties associated with these estimates (Table S2). For instance, paths with smaller uncertainties (e.g., A $\rightarrow$ B) demonstrate modest improvements in MAE, whereas paths with larger uncertainties (e.g., B $\rightarrow$ D) have more substantial improvements. This observation aligns with the theoretical expectation that cycle-based coupling primarily benefits estimates with lower precision by leveraging high-precision paths and the cycle consistency condition.

3.2 Relative Protein-Ligand Binding Free Energies

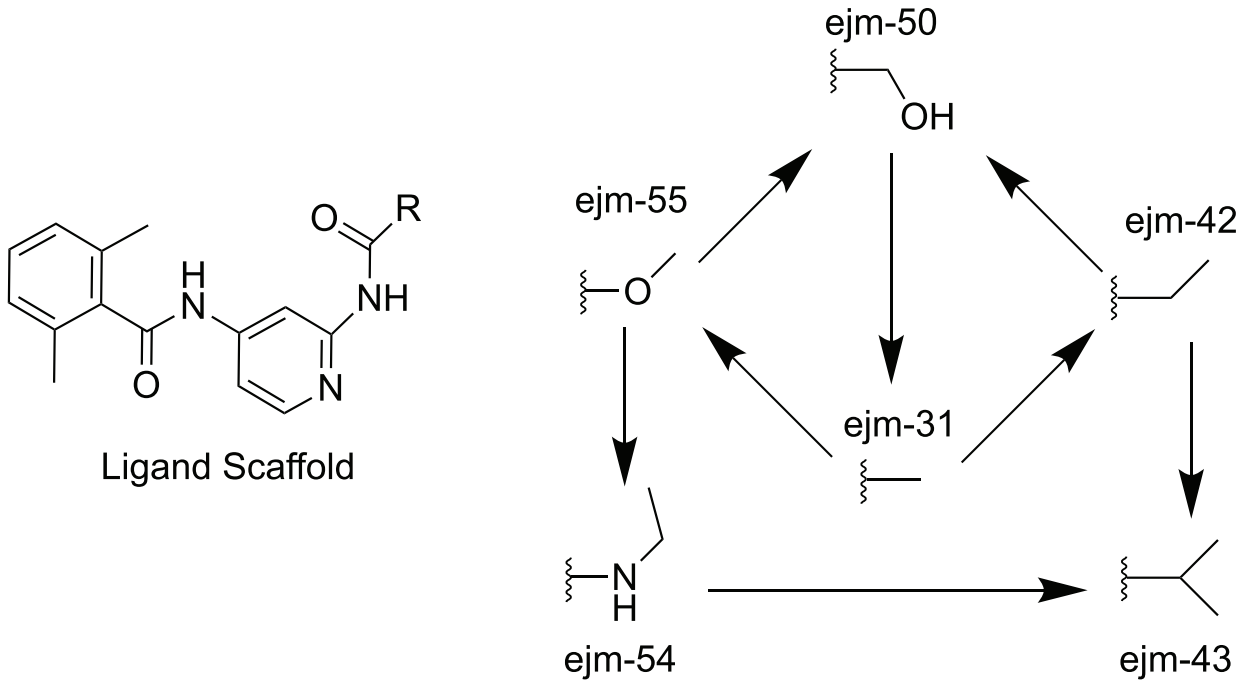

Figure 3: (Left) The common scaffold structure shared by the 6 ligands with the R group representing the chemical group that is different among ligands. (Right) The R group of each ligand and the alchemical perturbation graph connecting the ligands.

Next, we apply CBayesMBAR to compute the relative binding free energies of multiple ligands to a given protein. This type of calculation typically employs an alchemical per-

turbation graph connecting the ligands. Free energy differences along the graph's edges are computed with alchemical methods.[5,18] In our study, we apply CBayesMBAR to calculate the relative binding free energies of 6 ligands to the tyrosine kinase 2 (Tyk2) protein (Protein Data Bank ID: 4GIH[19]), a system widely used as a benchmark in the field.[3] The 6 ligands share a common scaffold structure, with the R group representing the chemical group that differs among them (Fig. 3). We construct an alchemical perturbation graph with multiple cycles to connect the ligands (Fig. 3).

We use the dual topology approach[5] to alchemically change one ligand into the other ligand on each edge of the perturbation graph. To do that, we use the following soft-core potential[20] to smoothly turn on and off the Leonard-Jones potential between a ligand's R group and the system:

$$E_{\mathrm{vdw}}(r) = \lambda_{\mathrm{vdw}} \cdot 4\epsilon \left[\frac{1}{(\alpha \cdot (1-\lambda_{\mathrm{vdw}}) + (r/\sigma)^6)^2} - \frac{1}{\alpha \cdot (1-\lambda_{\mathrm{vdw}}) + (r/\sigma)^6} \right], \tag{7}$$

where $\alpha = 0.5$, r is the distance between two particles, and σ and ϵ are the Leonard-Jones parameters of the two particles. The alchemical variable λ_{vdw} controls the soft-core potential: when $\lambda_{\mathrm{vdw}} = 0$, the Leonard-Jones potential is fully off and when $\lambda_{\mathrm{vdw}} = 1$, it is fully on. Electrostatic interactions between a ligand's R group and the system vary linearly with the alchemical variable λ_{elec}:

$$E_{\mathrm{elec}}(r) = \lambda_{\mathrm{elec}} \frac{q_i q_j}{r}, \tag{8}$$

where q_i and q_j are the charges of the two particles. For an edge from ligand A to ligand B, we employ 13 intermediate alchemical states. These states progressively deactivate the nonbonded interactions involving ligand A's R group with the system, while simultaneously activating those involving ligand B's R group with the system. To prevent singularities, the electrostatic interactions between a ligand's R group and the system are only present when the corresponding Leonard-Jones interactions are fully turned on. The specific values of the

alchemical variables for states along the path from ligand A to ligand B are detailed in Table 3, with states 1 and 15 corresponding to ligands A and B, respectively, and states 2 to 14 representing intermediates.

Table 3: Values of alchemical values for states on the path from ligand A to ligand B.

state index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
λ^A_{elec}	1	0.66	0.33	0	0	0	0	0	0	0	0	0	0	0	0
λ^A_{vdw}	1	1	1	1	0.95	0.9	0.7	0.5	0.3	0.1	0.05	0	0	0	0
λ^B_{elec}	0	0	0	0	0	0	0	0	0	0	0	0	0.33	0.66	1
λ^B_{vdw}	0	0	0	0	0.05	0.1	0.3	0.5	0.7	0.9	0.95	1	1	1	1

For our molecular dynamics simulations, we use the Amber ff14SB force field[21] for the protein, the general Amber force field[22] for ligands, and the TIP3P model[23] for water. Both water and protein phase simulations are conducted in a periodic water box. We calculate electrostatic interactions using the particle mesh Ewald method,[24] while Leonard-Jones interactions are smoothly truncated at 10 Å using a switching function that starts at 9 Å. We sample configurations from each state by running molecular dynamics simulations with OpenMM.[25] These simulations are performed at 298.15 K using the Langevin integrator[26] with a 2 fs time step and a 1 ps^{-1} friction coefficient. We maintain a pressure of 1 atm using a Monte Carlo barostat with Monte Carlo moves attempted every 25 steps. Each state is simulated for 5000 ps, with configurations saved every 2 ps.

We compute the alchemical free energy differences among the ligands using BayesMBAR, CCC, and CBayesMBAR with configurations sampled over simulation times ranging from 20 ps to 5000 ps. To ensure statistically meaningful comparisons, we repeat the calculations, including the molecular dynamics simulations, 10 times. The performance of the three methods were evaluated using the same metrics as in the harmonic oscillator example. Since the exact alchemical free energy differences are unknown, the reference free energy differences are calculated using all configurations sampled from all 10 repeats of the molecular dynamics simulations. The mean RMSEs of the free energy differences between end states of all paths are shown in Table 4. For both the water phase and the protein phase, CBayesMBAR has

significantly smaller RMSEs than BayesMBAR, with p-values less than 0.05 for all cases (Fig. S3 and S4). Compared to CCC, the RMSE of CBayesMBAR is either significantly (p-value ≤ 0.05) smaller or statistically indistinguishable from that of CCC (Fig. S5 and S6).

Table 4: The mean RMSE of free energy differences between end states of all paths in the alchemical perturbation graph. All values are in units of kcal/mol.

simulation time (ps)	water phase			protein phase		
	BayesMBAR	CCC	CBayesMBAR	BayesMBAR	CCC	CBayesMBAR
20	0.51	0.38	0.34	0.76	0.59	0.56
26	0.43	0.32	0.30	0.69	0.55	0.51
36	0.36	0.25	0.24	0.62	0.50	0.46
54	0.31	0.22	0.20	0.55	0.45	0.42
92	0.26	0.21	0.20	0.43	0.37	0.35
182	0.17	0.14	0.14	0.32	0.26	0.26
514	0.12	0.09	0.08	0.25	0.21	0.19
5000	0.03	0.03	0.02	0.14	0.13	0.12

We also compare the three methods using the MAEs of their free energy estimates on individual paths. The results are shown in Table 5 and S3 for the water phase and Table 6 and S4 for the protein phase. Although CBayesMBAR and CCC have reduced RMSEs aggregated over all paths compared to BayesMBAR (Table 4), their improvement over BayesMBAR on individual paths is not as consistent as in the harmonic oscillator example. CBayesMBAR and CCC have smaller MAEs than BayesMBAR for some paths, while for other paths, they have larger MAEs. This indicates that incorporating the cycle consistency condition does not uniformly enhance the accuracy of free energy estimates on all paths, although it does improve the overall accuracy.

Mean estimated uncertainties of free energy estimates on individual paths are presented in Table S5 and S6 for the water and protein phases, respectively. Similar to the harmonic oscillator example, CBayesMBAR demonstrates smaller uncertainties than BayesMBAR, underscoring improved precision in free energy estimates by leveraging the cycle consistency condition. Comparing the uncertainties and the MAEs of the free energy estimates on individual paths, we observe that whether the cycle consistency condition improves the accuracy

of free energy estimates on a path depends on the initial precision of the estimate on the path relative to that of other paths in the cycle. For paths with low precision, CBayesMBAR and CCC significantly improve the accuracy of the estimates, while for paths with high precision, CBayesMBAR and CCC could perform worse than BayesMBAR. This phenomenon occurs because coupling allows noise from low-precision paths to pass into higher precision paths, thereby slightly reducing the accuracy of the estimates on the latter. Notably, for paths where CBayesMBAR and CCC perform worse than BayesMBAR, CBayesMBAR tends to have smaller MAEs than CCC. This suggests that the cycle consistency condition is more effectively utilized in CBayesMBAR than in CCC to improve the accuracy of free energy estimates on individual paths.

Table 5: The mean absolute error of free energy differences between end states of individual paths in the water phase. All values are in units of kcal/mol.

path	20 ps			54 ps		
	BayesMBAR	CCC	CBayesMBAR	BayesMBAR	CCC	CBayesMBAR
ejm_31→ejm_42	0.16	0.33	0.21	0.11	0.16	0.09
ejm_42→ejm_50	0.50	0.35	0.33	0.34	0.17	0.15
ejm_50→ejm_31	0.64	0.29	0.30	0.31	0.18	0.19
ejm_31→ejm_55	0.24	0.39	0.27	0.15	0.19	0.15
ejm_55→ejm_50	0.50	0.33	0.34	0.40	0.26	0.23
ejm_55→ejm_54	0.45	0.34	0.36	0.28	0.21	0.22
ejm_54→ejm_43	0.37	0.28	0.23	0.20	0.16	0.17
ejm_43→ejm_42	0.28	0.25	0.24	0.09	0.17	0.14
path	514 ps			5000 ps		
	BayesMBAR	CCC	CBayesMBAR	BayesMBAR	CCC	CBayesMBAR
ejm_31→ejm_42	0.04	0.04	0.04	0.01	0.02	0.01
ejm_42→ejm_50	0.10	0.07	0.06	0.03	0.02	0.03
ejm_50→ejm_31	0.12	0.05	0.05	0.04	0.02	0.02
ejm_31→ejm_55	0.05	0.09	0.06	0.03	0.03	0.02
ejm_55→ejm_50	0.18	0.10	0.08	0.04	0.03	0.03
ejm_55→ejm_54	0.08	0.08	0.08	0.02	0.02	0.02
ejm_54→ejm_43	0.12	0.09	0.09	0.02	0.02	0.02
ejm_43→ejm_42	0.07	0.08	0.08	0.01	0.01	0.01

Table 6: The mean absolute error of free energy differences between end states of individual paths in the protein phase. All values are in units of kcal/mol.

path	20 ps			54 ps		
	BayesMBAR	CCC	CBayesMBAR	BayesMBAR	CCC	CBayesMBAR
ejm_31→ejm_42	0.22	0.43	0.31	0.21	0.33	0.25
ejm_42→ejm_50	0.64	0.44	0.43	0.42	0.41	0.39
ejm_50→ejm_31	0.82	0.39	0.41	0.73	0.42	0.37
ejm_31→ejm_55	0.35	0.40	0.26	0.20	0.30	0.16
ejm_55→ejm_50	0.96	0.55	0.50	0.57	0.39	0.40
ejm_55→ejm_54	0.41	0.49	0.47	0.33	0.41	0.39
ejm_54→ejm_43	0.68	0.59	0.55	0.42	0.43	0.43
ejm_43→ejm_42	0.96	0.78	0.81	0.54	0.43	0.45
path	514 ps			5000 ps		
	BayesMBAR	CCC	CBayesMBAR	BayesMBAR	CCC	CBayesMBAR
ejm_31→ejm_42	0.10	0.14	0.11	0.06	0.10	0.07
ejm_42→ejm_50	0.12	0.10	0.13	0.09	0.07	0.08
ejm_50→ejm_31	0.27	0.15	0.14	0.11	0.07	0.07
ejm_31→ejm_55	0.16	0.19	0.17	0.14	0.15	0.14
ejm_55→ejm_50	0.26	0.17	0.15	0.17	0.11	0.09
ejm_55→ejm_54	0.20	0.23	0.22	0.14	0.14	0.14
ejm_54→ejm_43	0.31	0.25	0.21	0.15	0.12	0.12
ejm_43→ejm_42	0.21	0.13	0.14	0.13	0.11	0.11

4 Conclusion & Discussion

In this work, we introduce CBayesMBAR, a new method for computing free energy differences on perturbation graphs with cycles. As a Bayesian approach, CBayesMBAR integrates the cycle consistency condition to couple free energy estimates across all edges of a perturbation graph in a principled manner. It incorporates the cycle consistency condition into the prior distribution of free energies and combines this with sampled configurations to form the posterior distribution, which is then used to estimate the free energies and their uncertainties. This ensures that free energy estimates directly satisfy the cycle consistency condition. Uncertainty estimates, derived from the standard deviation of the posterior distribution, reflect both the sampled configurations and the cycle consistency condition. Through two example applications, we demonstrate that CBayesMBAR significantly improves the accuracy of free energy estimates on perturbation graphs with cycles, outperforming both BayesMBAR and CCC. The computational cost of CBayesMBAR is comparable to that of multiple independent BayesMBAR calculations and is negligible compared to the cost of molecular dynamics simulations. For example, when computing relative protein-ligand binding free energies, CBayesMBAR takes about one minute to compute free energy differences and their uncertainties for all edges in the alchemical perturbation graph, using 2500 configurations from each state and running on a single RTX A5000 graphic processing unit.

Acknowledgement

The author thanks the Tufts University High Performance Compute Cluster that was utilized for the research reported in this paper.

Supporting Information Available

Linear constraints imposed by the cycle consistency condition in a perturbation graph with multiple cycles, computational details of the CCC method, computational details of CBayesMBAR for computing posterior mode and sampling from the posterior distribution, RMSEs of the free energy differences between end states of all paths for harmonic oscillators, mean absolute error of free energy differences between end states on individual paths for harmonic oscillators, mean estimated uncertainty of free energy differences between end states of individual paths calculated using BayesMBAR and CBayesMBAR for harmonic oscillators, RMSEs of the alchemical free energy differences between end states of all paths in the water and protein phase, mean absolute error of free energy differences in the water and protein phase between end states of individual paths and mean estimated uncertainty of free energy differences in the water and protein phase between end states of individual paths calculated using BayesMBAR and CBayesMBAR.

References

(1) Chipot, C.; Pohorille, A. *Free Energy Calculations*; Springer, 2007; Vol. 86.

(2) Liu, S.; Wu, Y.; Lin, T.; Abel, R.; Redmann, J. P.; Summa, C. M.; Jaber, V. R.; Lim, N. M.; Mobley, D. L. Lead Optimization Mapper: Automating Free Energy Calculations for Lead Optimization. *J. Comput. Aided Mol. Des.* **2013**, *27*, 755–770.

(3) 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 Lig-

and 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.

(4) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, T. D. J.; 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.

(5) Mey, A. S. J. S.; Allen, B. K.; Bruce McDonald, H. E.; 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 v1.0]. *Living J. Comp. Mol. Sci.* **2020**, *2*, 18378.

(6) Robo, M. T.; Hayes, R. L.; Ding, X.; Pulawski, B.; Vilseck, J. Z. Fast Free Energy Estimates from λ-Dynamics with Bias-Updated Gibbs Sampling. *Nature Communications* **2023**, *14*, 8515.

(7) Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *J. Chem. Theory Comput.* **2013**, *9*, 1282–1293.

(8) Mobley, D. L.; Chodera, J. D.; Dill, K. A. Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.

(9) Ding, X.; Vilseck, J. Z.; Hayes, R. L.; Brooks, C. L. I. Gibbs Sampler-Based λ-Dynamics and Rao–Blackwell Estimator for Alchemical Free Energy Calculation. *J. Chem. Theory Comput.* **2017**, *13*, 2501–2510.

(10) Li, Y.; Liu, R.; Liu, J.; Luo, H.; Wu, C.; Li, Z. An Open Source Graph-Based Weighted

Cycle Closure Method for Relative Binding Free Energy Calculations. *J. Chem. Inf. Model.* **2023**, *63*, 561–570.

(11) Giese, T. J.; York, D. M. Variational Method for Networkwide Analysis of Relative Ligand Binding Free Energies with Loop Closure and Experimental Constraints. *J. Chem. Theory Comput.* **2021**, *17*, 1326–1336.

(12) Ding, X. Bayesian Multistate Bennett Acceptance Ratio Methods. *J. Chem. Theory Comput.* **2024**, *20*, 1878–1888.

(13) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

(14) Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*; Springer Science & Business Media, 2013.

(15) Neal, R. M. MCMC Using Hamiltonian Dynamics. *Handb. Markov Chain Monte Carlo* **2011**, *2*, 2.

(16) Hoffman, M. D.; Gelman, A.; others The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.

(17) Lao, J.; Louf, R. Blackjax: A Sampling Library for JAX. 2020.

(18) Essex, J. W.; Severance, D. L.; Tirado-Rives, J.; Jorgensen, W. L. Monte Carlo Simulations for Proteins: Binding Affinities for Trypsin-Benzamidine Complexes via Free-Energy Perturbations. *J. Phys. Chem. B* **1997**, *101*, 9663–9669.

(19) Liang, J.; Tsui, V.; Van Abbema, A.; Bao, L.; Barrett, K.; Beresini, M.; Berezhkovskiy, L.; Blair, W. S.; Chang, C.; Driscoll, J.; Eigenbrot, C.; Ghilardi, N.; Gibbons, P.; Halladay, J.; Johnson, A.; Kohli, P. B.; Lai, Y.; Liimatta, M.; Mantik, P.; Menghrajani, K.; Murray, J.; Sambrone, A.; Xiao, Y.; Shia, S.; Shin, Y.; Smith, J.;

Sohn, S.; Stanley, M.; Ultsch, M.; Zhang, B.; Wu, L. C.; Magnuson, S. Lead Identification of Novel and Selective TYK2 Inhibitors. *Eur. J. Med. Chem.* **2013**, *67*, 175–187.

(20) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(21) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(22) 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.

(23) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(24) 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.

(25) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

(26) Zhang, Z.; Liu, X.; Yan, K.; Tuckerman, M. E.; Liu, J. Unified Efficient Thermostat Scheme for the Canonical Ensemble with Holonomic or Isokinetic Constraints via Molecular Dynamics. *J. Phys. Chem. A* **2019**, *123*, 6056–6079.