HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2018 February 27.

Published in final edited form as:

J Chem Theory Comput. 2017 June 13; 13(6): 2501–2510. doi:10.1021/acs.jctc.7b00204.

Gibbs Sampler Based λ-Dynamics and Rao-Blackwell Estimator for Alchemical Free Energy Calculation

Xinqiang Ding[†], **Jonah Z. Vilseck**[‡], **Ryan L. Hayes**[‡], and **Charles L. Brooks III**[†,‡,¶]

[†]Department of Computational Medicine & Bioinformatics, University of Michigan

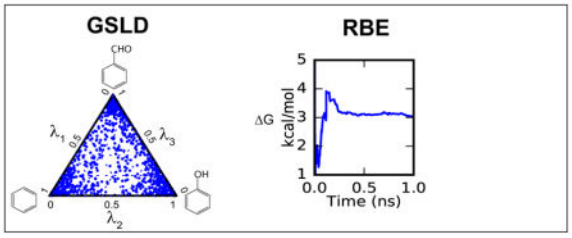
[‡]Department of Chemistry, University of Michigan

[¶]Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109, United States

Abstract

λ-dynamics is a generalized ensemble method for alchemical free energy calculations. In traditional λ-dynamics, the alchemical switch variable λ is treated as a continuous variable ranging from 0 to 1 and an empirical estimator is utilized to approximate the free energy. In the present paper, we describe an alternative formulation of λ-dynamics that utilizes the Gibbs sampler framework which we call Gibbs Sampler λ-dynamics (GSLD). GSLD, like traditional λ-dynamics, can be readily extended to calculate free energy differences between multiple ligands in one simulation. We also introduce a new free energy estimator, the Rao-Blackwell estimator (RBE) for use in conjunction with GSLD. Compared with the current empirical estimator, the advantage of RBE is that RBE is an unbiased estimator and its variance is usually smaller than the current empirical estimator. We also show that the multistate Bennett acceptance ratio (MBAR) equation or the unbinned weighted histogram analysis method (UWHAM) equation can be derived using the RBE. We illustrate the use and performance of this new free energy computational framework by application to a simple harmonic system as well as relevant calculations of small molecule relative free energies of solvation and binding to a protein receptor. Our findings demonstrate consistent and improved performance compared with conventional alchemical free energy methods.

Graphical Abstract



Correspondence to: Charles L. Brooks, III.

Supporting Information Available

The supporting figures and tables are included in the the following file: **SI.pdf**: supporting figures and tables mentioned but not shown in the article. This material is available free of charge via the Internet at http://pubs.acs.org/.

1 Introduction

Free energy calculation is fundamental for understanding many important biophysical processes, such as protein conformational changes, protein-protein interactions, and protein-ligand binding processes.[1,2] Calculating protein-ligand binding free energy has important applications in drug discovery, especially in the lead compound generation and optimization stages.[3–5] These stages only require calculating protein-ligand relative binding free energy, which has been shown to be easier than calculating protein-ligand absolute binding free energy.[3,4]

One widely used methodology for calculating protein-ligand relative binding free energy is the alchemical free energy approach.[3–5] This approach utilizes the thermodynamic cycle shown in Figure 1.[2] This thermodynamic cycle specifies that $\Delta\Delta G^{\text{binding}}_{L_0\to L_1}=\Delta G^{\text{binding}}_{L_1}-\Delta G^{\text{binding}}_{L_0}=\Delta G^{\text{bound}}_{L_0\to L_1}-\Delta G^{\text{unbound}}_{L_0\to L_1}$. In order to calculate the relative binding free energy between ligand L_0 and L_1 with receptor R, i.e, $\Delta\Delta G^{\text{binding}}_{L_0\to L_1}$, the alchemical free energy method calculates $\Delta G^{\text{unbound}}_{L_0\to L_1}$ and $\Delta G^{\text{bound}}_{L_0\to L_1}$ by employing alchemical transformations morphing ligand L_0 into ligand L_1 in both unbound and bound environments, respectively.

Several alchemical free energy calculation methods have been developed over the last several decades, such as free energy perturbation,[6,7] thermodynamic integration,[2,8] enveloping distribution sampling[9,10] and λ-dynamics.[11–17] λ-dynamics is a generalized ensemble method in which the alchemical transformation variable λ is a continuous variable ranging from 0 to 1, with λ = 0, 0 < λ < 1, and λ = 1 corresponding to the ligand being in L_0 state, intermediate hybrid states, and L_1 state, respectively. The potential energy corresponding to λ is

$$V(\lambda,\{x_i\}_{i=0}^{1},X)=(1-\lambda)V_0(x_0,X)+\lambda(V_1(x_1,X)+G_1^b)+V_{\text{env}}(X), \quad (1)$$

where X, x_0 and x_1 are atomic coordinates associated with the environment, the ligand L_0 and the ligand L_1, respectively. $V_i(x_i, X)$ is the potential energy between ligand L_i and the environment and $V_{\text{env}}(X)$ is the potential energy of the environment. G_1^b is a biasing potential to ensure that the two physical states, corresponding to λ = 0 and λ = 1, are both sampled in the simulation. The biasing potential G_1^b is determined iteratively by running multiple short simulations.[14,15,18,19] The dynamics of the system (λ, $\{x_i\}_{i=0}^{1}$, X) is generated from the extended Hamiltonian:

$$H(\lambda,\{x_i\}_{i=0}^{1},X)=T_{x,X}+T_\lambda+V(\lambda,\{x_i\}_{i=0}^{1},X) \quad (2)$$

where $T_{x,X}$ and T_λ are the kinetic energy associated with coordinates ($\{x\}_{i=0}^{1}$, X) and λ, respectively. The free energy difference between ligand L_0 and L_1, with the biasing potential G_1^b, is

$$\Delta G = -\beta^{-1} \ln \frac{P(\lambda=1)}{P(\lambda=0)}, \quad (3)$$

where β is the inverse temperature; $P(\lambda = 0)$ and $P(\lambda = 1)$ are probability denisties of λ at points 0 and 1, respectively. In practice, this free energy difference ΔG is estimated using the following empirical estimator based on the trajectory of λ:

$$\Delta \hat{G} = -\beta^{-1} \ln \frac{P(\lambda > \lambda_{\text{cutoff}})}{P(\lambda < 1 - \lambda_{\text{cutoff}})}, \quad (4)$$

where λ_{cutoff} ($0 < \lambda_{\text{cutoff}} < 1$) is a cutoff value which is chosen to be close to 1.[14]

Although the empirical estimator is straightforward to evaluate based on the λ trajectory, it is not necessarily optimal. One issue is that the empirical estimator is systematically biased as it uses $P(\lambda < 1 - \lambda_{\text{cutoff}})$ and $P(\lambda > \lambda_{\text{cutoff}})$ to approximate $P(\lambda = 0)$ and $P(\lambda = 1)$, respectively. Additionally, the bias depends on the cutoff value λ_{cutoff}, which is chosen empirically and is difficult to quantify as it may vary among different systems.

In the current work, we present a novel form of λ-dynamics called the Gibbs sampler based λ-dynamics (GSLD) with the Rao-Blackwell estimator (RBE). The Gibbs sampler framework for calculating free energy differences between two ligands was first suggested by Chodera and Shirts.[20] In their work, λ was treated as a discrete variable and MBAR[21] was used to estimate the free energy change. In this study, we show that GSLD and RBE can treat λ as either a discrete variable or a continuous variable when calculating free energy differences between two ligands. When λ is treated as a continuous variable, GSLD and RBE can be generalized to simultaneously calculate free energies of multiple ligands in one simulation, as in the generalization of λ-dynamics.[14] We explore these new methods through applications to three model systems in this paper. This paper is organized as follows. In section 2, we describe GSLD and its generalization to multiple ligands. Then we introduce the RBE and show that the MBAR/UWHAM equations[21–23] can be derived from the RBE. In section 3, we give detailed setup information for the setup and simulation of the three systems with which we tested the methods. Our results for these three systems are presented in section 4. We conclude with a discussion of how the GSLD and RBE can be used for other applications.

2 Methods

2.1 Gibbs Sampler Based λ-Dynamics

As a generalized ensemble method, GSLD samples from the joint distribution of λ and the atomic coordinates of the system using the Gibbs sampler. In this section, we first briefly introduce the Gibbs sampler. We then use the Gibbs sampler to formulate pairwise GSLD. We conclude by showing how the GSLD can be generalized to work for multiple ligands.

2.1.1 The Gibbs Sampler—The Gibbs sampler, which is widely used in both statistics and machine learning, is a Markov Chain Monte Carlo (MCMC) method for sampling from multivariate distributions.[24,25] To sample (X, Y) from the joint distribution: $(X, Y) \sim P(X, Y)$, the Gibbs sampler generates a Markov chain of states $\{(X_t, Y_t), t = 0, 1, 2, \ldots, N\}$ using the following procedure:

- **Step 0:** initialize the starting state (X_0, Y_0).
- **Step t:** sample from the conditional distribution
  - **Updating X**: given the state (X_{t-1}, Y_{t-1}) from step $t - 1$, sample X_t from the conditional distribution of $X_t \sim P(X_t|Y_{t-1})$.
  - **Updating Y**: given X_t from the above update step, sample Y_t from the conditional distribution of $Y_t \sim P(Y_t|X_t)$. The resulting sample (X_t, Y_t) is the state for step t.

Because the above procedure satisfies the detailed balance condition with respect to the joint distribution: $(X, Y) \sim P(X, Y)$, the sampled states $\{(X_t, Y_t), t = 0, 1, 2, \ldots, N\}$ converge to the joint distribution.[24,25] The update steps require sampling from both conditional distributions: $X_t \sim P(X_t|Y_{t-1})$ and $Y_t \sim P(Y_t|X_t)$. If direct sampling from the conditional distribution is possible, independent samples can be directly drawn using numerical pseudorandom number generators. Otherwise, samples can be drawn using other Monte Carlo methods or Hamiltonian dynamics, as long as the method satisfies the detailed balance condition with respect to the corresponding conditional distribution.[25,26] This property of the Gibbs sampler makes it quite flexible on choosing appropriate sampling methods based on the conditional distributions.

2.1.2 Pairwise GSLD—Pairwise GSLD calculates the free energy difference between two ligands: ligand L_0 and ligand L_1. In pairwise GSLD, λ can be treated as either a continuous variable or a discrete variable.

Continuous λ: When λ is treated as a continuous variable, pairwise GSLD samples from the joint distribution of (λ, $\{x_i\}_{i=0}^{1}$, X):

$$P(\lambda, x_0, x_1, X) = \frac{\exp(-\beta\left[(1-\lambda)V_0(x_0, X) + \lambda(V_1(x_1, X) + G_1^b) + V_{\text{env}}(X)\right])}{Z}, \quad (5)$$

where Z is the partition function of the generalized ensemble and G_1^b is a biasing potential. G_1^b is determined automatically in the current simulations using a Wang-Landau like algorithm[27] which is described in Appendix A. The Gibbs sampler for sampling from the above joint distribution is as follows:

- **Step 0:** initialize the starting state (λ^0, $\{x_i^0\}_{i=0}^1$, X^0).

  Step t: sample from the conditional distributions:

  - **Updating** ($\{x_i\}_{i=0}^1$, X): given the state (λ^{t-1}, $\{x_i^{t-1}\}_{i=0}^1$, X^{t-1}) from step $t-1$, sample ($\{x_i^t\}_{i=0}^1$, X^t) from the conditional distribution: $P(\{x_i^t\}_{i=0}^1, X^t|\lambda^{t-1}) \propto \exp(-\beta\left[(1-\lambda^{t-1})V_0(x_0^t, X^t)+\lambda^{t-1}\left(V_1(x_1^t, X^t)+G_1^b\right)+V_{\text{env}}(X^t)\right])$, which is the canonical ensemble distribution at the inverse temperature β. A sample can be drawn from this distribution using molecular dynamics simulation.

  - **Updating** λ: given the atomic coordinates ($\{x_i^t\}_{i=0}^1$, X^t) sampled from the above update step, sample λ^t directly from the conditional distribution $P(\lambda^t|\{x_i^t\}_{i=0}^1, X^t)$ using numerical pseudo-random number generator. The conditional distribution $P(\lambda^t|\{x_i^t\}_{i=0}^1, X^t)$ is:

$$P(\lambda^t|\{x_i^t\}_{i=0}^1, X^t)=\frac{\exp(-\beta\left[(1-\lambda^t)V_0(x_0^t, X^t)+\lambda^t(V_1(x_1^t, X^t)+G_1^b)+V_{\text{env}}(X^t)\right])}{\int_0^1\exp(-\beta\left[(1-\lambda^t)V_0(x_0^t, X^t)+\lambda^t(V_1(x_1^t, X^t)+G_1^b)+V_{\text{env}}(X^t)\right])d\lambda}$$
$$=\frac{\beta\cdot(\Delta V_{0\to1}^t+G_1^b)\exp(-\lambda^t\cdot\beta\cdot[\Delta V_{0\to1}^t+G_1^b])}{1-\exp(-\beta\cdot[\Delta V_{0\to1}^t+G_1^b])} \quad (0\le\lambda^t\le1), \tag{6}$$

    where $\Delta V_{0\to1}^t=V_1(x_1^t, X^t)-V_0(x_0^t, X^t)$. This is an exponential distribution of λ^t restricted on the interval of [0, 1]. Therefore, sampling λ^t directly from this distribution can be done using the inverse transformation method:

$$\lambda^t=-\frac{1}{\beta\cdot(\Delta V_{0\to1}^t+G_1^b)}\ln\left[1-\left[1-e^{\beta\cdot(\Delta V_{0\to1}^t+G_1^b)}\right]\cdot u\right] \tag{7}$$

    where u is a random sample from the uniform distribution on [0, 1]. The resulting sample (λ^t, $\{x_i^t\}_{i=0}^1$, X^t) is the state for step t.

Discrete λ: When λ is a discrete variable specified by the set {l_1, l_2, … l_M}, GSLD samples from the joint distribution

$$P(\lambda=l_j, x_0, x_1, X) \propto \exp(-\beta \left[V_0(x_0, X, 1-l_j)+V_1(x_1, X, l_j)+G_j^b+V_{\text{env}}(X) \right]), \quad (8)$$

where G_j^b is the biasing potential added to the state corresponding to $\lambda = l_j$. Sampling from this distribution is done in the same way as the case where λ is continuous except that the conditional distribution $P(\lambda^t|\{x_i^t\}_{i=0}^1, X^t)$ becomes a multinomial distribution:

$$P(\lambda^t=l_j|\{x_i^t\}_{i=0}^1, X^t)=\frac{\exp(-\beta \left[V_0(x_0^t, X^t, 1-l_j)+V_1(x_1^t, X^t, l_j)+G_j^b \right])}{\sum_{k=1}^M \exp(-\beta \left[V_0(x_0^t, X^t, 1-l_k)+V_1(x_1^t, X^t, l_k)+G_k^b \right])} \quad (9)$$

from which samples can also be drawn directly using numerical methods. The biasing potentials G_j^b are determined similarly as the case when λ is continuous. We note that equation 9 is similar to the distribution calculated using the infinite swap limit in replica exchange methods.[28–31]

The advantage of using λ as a discrete variable is that the pairwise GSLD still works when the potential energy $V_i(x_i, X, \lambda)$ is λ dependent, such as when a soft-core Lennard-Jones potential[32] is employed to facilitate sampling. When λ is continuous, using λ dependent $V_i(x_i, X, \lambda)$ will make the normalization constant of the conditional distribution $P(\lambda|\{x_i\}_{i=0}^1, X)$ not analytically integrable and prevent direct sampling from the conditional distribution $P(\lambda|\{x_i\}_{i=0}^1, X)$. However, as shown below, the advantage of using λ as a continuous variable is that the GSLD can be generalized for multiple ligands.

2.1.3 Generalizing GSLD for Multiple Ligands—Like λ-dynamics, GSLD can be generalized to calculate the free energies for multiple ligands in one simulation. Assuming there are n ligands, the fraction of the ith ligand in the hybrid state is represented by λ_i, for $i = 1, 2, \ldots, n$. The hybrid state is specified by the value of $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ which satisfies the conditions $\sum_{i=1}^n \lambda_i=1$ and $0 \le \lambda_i \le 1$, $i = 1, 2, \ldots, n$. The hybrid state's potential energy is defined as: $V(\{\lambda_i\}_{i=1}^n, \{x_i\}_{i=1}^n, X)=\sum_{i=1}^n \lambda_i(V_i(x_i, X)+G_i^b)+V_{\text{env}}(X)$, where x_i and X are atomic coordinates associated with the ith ligand and environment, respectively; G_i^b is the biasing potential added for the ith ligand and can be determined similarly as in the pairwise GSLD. Sampling from the generalized ensemble distribution: $P(\{\lambda_i\}_{i=1}^n, \{x_i\}_{i=1}^n, X) \propto \exp(-\beta \cdot V(\{\lambda_i\}_{i=1}^n, \{x_i\}_{i=1}^n, X))$ can be done using the following Gibbs sampler procedure:

- **Step 0**: initialize the starting state ($\{\lambda_i^0\}_{i=1}^n, \{x_i^0\}_{i=1}^n, X^0$).
- **Step t:** sample from the conditional distributions.

- **Updating** ($\{x_i\}_{i=1}^n$, X): given the state ($\{\lambda_i^{t-1}\}_{i=1}^n, \{x_i^{t-1}\}_{i=1}^n$, X^{t-1}) from step $t-1$, sample ($\{x_i^t\}_{i=1}^n$, X^t) from the conditional distribution $P(\{x_i^t\}_{i=1}^n, X^t|\{\lambda_i^{t-1}\}_{i=1}^n)$ using molecular dynamics simulation.
- **Updating** $\{\lambda_i\}_{i=1}^n$: given the sample ($\{x_i^t\}_{i=1}^n$, X^t) from the above update step, the conditional distribution of $\{\lambda_i^t\}_{i=1}^n$ in the set $S=\{(\lambda_1,\ldots,\lambda_n)|\sum_{i=1}^n \lambda_i=1 \text{ and } \lambda_i \geq 0, i=1,\ldots n\}$ is given by

$$P(\{\lambda_i^t\}_{i=1}^n|\{x_i^t\}_{i=1}^n, X^t)=\frac{\exp(-\beta\left[\sum_{i=1}^n \lambda_i^t\left[V_i(x_i^t,X^t)+G_1^b\right]+V_{\text{env}}(X^t)\right])}{Z}. \quad (10)$$

where

$$Z=\int_S \exp(-\beta\left[\sum_{i=1}^n \lambda_i^t\left[V_i(x_i^t,X^t)+G_i^b\right]+V_{\text{env}}(X^{(t)})\right])dm_S(\lambda)$$
$$=e^{-\beta V_{\text{env}}(X^t)}\sum_{i=1}^n \frac{e^{-\beta\left[V_i(x_i^t,X^t)+G_i^b\right]}}{\beta^{n-1}\prod_{j\neq i}(\left[V_j(x_j^t,X^t)+G_j^b\right]-\left[V_i(x_i^t,X^t)+G_i^b\right])}, \quad (11)$$

and $dm_S(\lambda)$ is the infinitesimal volume element of the simplex S. Because $\sum_{i=1}^n \lambda_i^t=1$, the conditional distribution $P(\{\lambda_i^t\}_{i=1}^n|\{x_i^t\}_{i=1}^n, X^t)$ has only $n-1$ degrees of freedom. Sampling from this conditional distribution is equivalent to sampling from the $n-1$ dimensional distribution:

$$P(\{\lambda_i^t\}_{i=1}^{n-1}|\{x_i^t\}_{i=1}^n, X^t) \propto \exp(-\beta\left[\sum_{i=1}^{n-1}\lambda_i\left[V_i(x_i^t,X^t)+G_i^b-V_n(x_n^t,X^t)-G_n^b\right]\right]), \quad (12)$$

where $0 \leq \sum_{i=1}^{n-1}\lambda_i^t \leq 1$, and $\lambda_i^t \geq 0$. The environment atom energy term, $V_{\text{env}}(X^t)$, does not appear in equation (12) because it is part of both the numerator and denominator of equation (10) and can be canceled out as a constant when ($\{x_i^t\}_{i=1}^n$, X^t) is fixed. Sampling from this $n-1$ dimensional distribution $P(\{\lambda_i^t\}_{i=1}^{n-1}|\{x_i^t\}_{i=1}^n, X^t)$ is done

using the rejection method. In the rejection method, each $\{\lambda_i^t\}_{i=1}^{n-1}$ is sampled independently from the distribution: $P(\lambda_i^t) \propto \exp(-\beta\lambda_i \left[V_i(x_i^t, X^t)+G_i^b-V_n(x_n^t, X^t)-G_n^b\right])$, where $0 \leq \lambda_i^t \leq 1$. If the sample $\{\lambda_i^t\}_{i=1}^{n-1}$ satisfies the condition $0 \leq \sum_{i=1}^{n-1}\lambda_i^t \leq 1$, it is accepted, otherwise the sample $\{\lambda_i^t\}_{i=1}^{n-1}$ is rejected. This procedure is repeated until a sample $\{\lambda_i^t\}_{i=1}^{n-1}$ is accepted. Set $\lambda_n^t=1-\sum_{j=1}^{n-1}\lambda_j^t$ and the resulting sample ($\{\lambda_i^t\}_{i=1}^{n}$, $\{x_i^t\}_{i=1}^{n}$, X^t) is the state for step *t*.

2.2 Rao-Blackwell Estimator (RBE)

Although the empirical estimator used in λ-dynamics can also be utilized in GSLD to estimate the free energy, it is not an optimal estimator and may contain a system dependent bias. RBE is introduced here to eliminate these potential issues. RBE is the estimator derived by applying the Rao-Blackwellization transformation to the empirical estimator. Rao-Blackwellization is a statistical method, inspired by the Rao-Blackwell theorem,[33,34] to transform a crude estimator into a better estimator that has smaller mean squared error for estimating the quantity of interest.[35]

For pairwise GSLD with continuous λ, the quantity of interest is the free energy $\Delta G = -\beta^{-1}$ ln $[P(\lambda = 1)/P(\lambda = 0)]$. To estimate ΔG, the empirical estimator approximates $P(\lambda = 1)$ and $P(\lambda = 0)$ directly by calculating the fraction of λs which are close to 1 and 0, respectively, based on the λ trajectory. In contrast, the RBE ignores the λ trajectory and only uses the atomic coordinate trajectory. It is based on the fact that $P(\lambda = 1)$ and $P(\lambda = 0)$ are equal to the expectation of the conditional probability of λ with respect to the atomic coordinates, i.e, $P(\lambda=1)=\mathbb{E}_{\{\{x_i\}_{i=0}^1,X\}}\left[P(\lambda=1)|\{x_i\}_{i=0}^1,X\right]$ and $P(\lambda=0)=\mathbb{E}_{\{\{x_i\}_{i=0}^1,X\}}[P(\lambda=0)|\{x_i\}_{i=0}^1,X]$. Therefore, RBE uses the following formula to estimate the free energy ΔG:

$$\begin{aligned}\Delta G_{\mathrm{RBE}} &= -\beta^{-1}\ln\frac{P(\lambda=1)}{P(\lambda=0)} \\ &= -\beta^{-1}\ln\frac{\mathbb{E}_{\{\{x_i\}_{i=0}^1,X\}}[P(\lambda=1|\{x_i\}_{i=0}^1,X)]}{\mathbb{E}_{\{\{x_i\}_{i=0}^1,X\}}[P(\lambda=0|\{x_i\}_{i=0}^1,X)]} \\ &= -\beta^{-1}\ln\frac{1/N\cdot\sum_{t=0}^{N}P(\lambda=1|\{x_i^t\}_{i=0}^1,X^t)}{1/N\cdot\sum_{t=0}^{N}P(\lambda=0|\{x_i^t\}_{i=0}^1,X^t)}\end{aligned} \quad (13)$$

where

$$P(\lambda=1|\{x_i^t\}_{i=0}^1, X^t)=\frac{\beta\cdot(\Delta V_{0\to1}^t+G_1^b)\cdot\exp(-\beta\cdot[\Delta V_{0\to1}^t+G_1^b])}{1-\exp(-\beta\cdot[\Delta V_{0\to1}^t+G_1^b])}$$
$$P(\lambda=0|\{x_i^t\}_{i=0}^1, X^t)=\frac{\beta\cdot(\Delta V_{0\to1}^t+G_1^b)}{1-\exp(-\beta\cdot[\Delta V_{0\to1}^t+G_1^b])}, \quad (14)$$

and N is the number of samples.

For the generalized GSLD with multiple ligands, the RBE can be derived similarly. To estimate the free energy of the ith ligand given by $G(\lambda_i = 1, \lambda_{j\neq i} = 0) = -\beta^{-1} \ln P(\lambda_i = 1, \lambda_{j\neq i} = 0)$, the RBE uses the following formula:

$$\begin{aligned}
&G_{\mathrm{RBE}}(\lambda_i=1,\lambda_{j\neq i}=0)=-\beta^{-1}\ln P(\lambda_i=1,\lambda_{j\neq i}=0)\\
&=-\beta^{-1}\ln\mathbb{E}_{\{\{x_k\}_{k=1}^n,X\}}\left[P(\lambda_i=1,\lambda_{j\neq i}=0|\{\{x_k\}_{k=1}^n,X\})\right]\\
&=-\beta^{-1}\ln\left[1/N\cdot\sum_{t=0}^{N}P(\lambda_i=1,\lambda_{j\neq i}=0|\{\{x_k\}_{k=1}^n,X\})\right]\\
&=-\beta^{-1}\ln\left[1/N\cdot\sum_{t=0}^{N}\frac{\exp(-\beta[V_i(x_i^t,X^t)+G_i^b])}{Z}\right],
\end{aligned} \quad (15)$$

where Z is given in equation 11 in **section 2.1.3**.

As shown in the above formulas, the RBE estimator ΔG_{RBE} does not depend on the empirical cutoff value of $\lambda_{\mathrm{cutoff}}$. Based on the Rao-Blackwell theorem, ΔG_{RBE} is an unbiased estimator. In addition, if the samples from GSLD are independent, the mean squared error of RBE is guaranteed to be smaller than or equal to that of the empirical estimator. Although the samples from GSLD are usually not truly independent, the advantage of RBE can often be justified empirically.[36]

2.3 Derivation of the MBAR/UWHAM equations using RBE

Although RBE is originally introduced to estimate free energies based on sampling from GSLD, RBE can also be used when multiple equilibrium states are sampled independently. When RBE is applied to this case, it generates the MBAR/UWHAM equations,[21–23] which are widely used in current alchemical free energy methods.

Let us assume there are M equilibrium states with potential energy function of V_i, $i = 1, 2, \ldots, M$. Each equilibrium state is sampled independently. The conformations sampled from state i are represented as x_i^k, $k = 1, 2, \ldots, n_i$, where n_i is the number of conformations from state i. The total number of conformations is $N=\sum_{j=1}^{M} n_j$. The free energy of state i is represented as G_i^*. We use $\lambda \in \{1, 2, \ldots, M\}$ as an index variable to represent the M equilibrium states, with $\lambda = i$ corresponding to state i. To calculate the free energies for all the equilibrium states, all the conformations $\{x_i^k, i = 1, 2, \ldots, M, k = 1, 2, \ldots, n_i\}$ are pooled

together and viewed as samples from the generalized ensemble $P(\lambda=i,x) \propto e^{-\beta[V_i(x)+G_i^b]}$, where G_i^b is the biasing energy added to state *i* to adjust the relative weight of state *i* to be proportional to n_i, i.e, G_i^b needs to satisfy the condition:

$$G_i=G_i^*+G_i^b=-\beta^{-1}\ln\frac{n_i}{N}, \quad (16)$$

where G_i is the free energy of state *i* with the biasing potential of G_i^b and G_i^* is the unbiased free energy of state *i*. We note that the biasing potentials G_i^b in equation 16 are unknown variables. They are introduced to make the equation 16 valid, which is the requirement for applying the RBE. These unknown biasing potentials G_i^b can be calculated after the values of G_i^* are solved. The RBE for this generalized ensemble is:

$$\begin{aligned} G_i &= -\beta^{-1}\ln P(\lambda=i) \\ &= -\beta^{-1}\ln\frac{1}{N}\sum_{j=1}^{M}\sum_{k=1}^{n_i}P(\lambda=i|x_j^k) \\ &= -\beta^{-1}\ln\frac{1}{N}\sum_{j=1}^{M}\sum_{k=1}^{n_j}\frac{e^{-\beta\left[V_i(x_j^k)+G_i^b\right]}}{\sum_{l=1}^{M}e^{-\beta\left[V_l(x_j^k)+G_l^b\right]}} \end{aligned} \quad (17)$$

Combining equation 16 with equation 17, we have:

$$G_i^*=-\beta^{-1}\ln\sum_{j=1}^{M}\sum_{k=1}^{n_j}\frac{e^{-\beta\left[V_i(x_j^k)\right]}}{\sum_{l=1}^{M}n_l\cdot e^{-\beta\left[V_l(x_j^k)-G_l^*\right]}} \quad (18)$$

which is the same as the MBAR/UWHAM equations.[21–23] Previously, the MBAR/UWHAM equations were derived as either a result of the maximum likelihood principle or an unbinned extension of the weighted histogram analysis method (WHAM).[21–23] Here we have shown that the MBAR/UWHAM equations can also be derived using RBE.

3 Model Systems and Computational Details

To illustrate how GSLD works and the advantage of RBE over the empirical estimator typically used in λ-dynamics, we applied GSLD and RBE to three test cases: **(a)** calculation of the free energy difference between two states of a harmonic oscillator system, **(b)** calculation of the relative hydration free energies of three benzene derivatives, and **(c)** calculation of the binding free energy difference between benzene and p-xylene bound to the L99A mutant of the protein T4 lysozyme.[37,38] The simulations in these calculations were run using CHARMM[39] compiled with OpenMM.[40] Each calculation was repeated 10 times. Error bars were calculated as the standard variation of the results from these 10 independent repeats.

3.1 Harmonic System

The harmonic system consists of a one dimensional particle that switches between two states: state 0 and state 1. Each state has a harmonic potential energy. The purpose is to calculate the free energy difference of the particle when it changes from state 0 to state 1, i.e, $\Delta G = G_1 - G_0$. Specifically, state 0 has a potential energy given by $\frac{1}{2}k_0(x-x_0^e)^2$, and state 1 has a potential energy given by $\frac{1}{2}k_1(x-x_1^e)^2$. In order to prevent the particle from moving too far from the equilibrium position, a restraining potential is added for each state. This restraining potential is not scaled by λ. The resulting hybrid potential energy is:

$$\begin{aligned} V(\lambda, x_0, x_1) = &(1-\lambda)\cdot\frac{1}{2}k_0(x_0-x_0^e)^2 \\ &+\lambda\cdot\frac{1}{2}k_1(x_1-x_1^e)^2 \\ &+\frac{1}{2}k_{\text{env}}(|x_0|-x_{\text{env}}^e)^2\{|x_0| \geq x_{\text{env}}^e\}+\frac{1}{2}k_{\text{env}}(|x_1|-x_{\text{env}}^e)^2\{|x_1| \geq x_{\text{env}}^e\}, \end{aligned}$$

where 𝟙;{condition} is equal to 1 if the condition is true, otherwise it is equal to 0. GSLD is used to sample from the joint distribution of $(\lambda, \{x_i\}_{i=0}^1)$:$P(\lambda, \{x_i\}_{i=0}^1) \propto \exp(-\beta \cdot V(\lambda, \{x_i\}_{i=0}^1))$. Given the value of λ, sampling the coordinates ($\{x_i\}_{i=0}^1$) is accomplished by running Langevin dynamics for 1 ps with a step size of 1 fs, temperature of 300 K, and friction coefficient of 10 ps^{-1}. The total simulation time is 10 ns. The parameters used for $x_0^e, x_1^e, x_{\text{env}}^e$and k_{env} are −2.0 Å, 2.0 Å, 4.0 Å, and 2.5 kcal/mol · $Å^{-2}$, respectively. Two variations of the model system that correspond to setting different values for k_0 and k_1 are used: a symmetrical system with $k_0 = k_1 = 0.75$ kcal/mol·$Å^{-2}$, and an asymmetrical system with $k_0 = 0.75$ kcal/mol · $Å^{-2}$ and $k_1 = 0.075$ kcal/mol · $Å^{-2}$.

3.2 Relative Hydration Free Energies for Three Benzene Derivatives

Relative hydration free energies for three benzene derivatives: benzene, phenol, and benzaldehyde were calculated from the difference between alchemical free energy changes computed in vacuum and in water. The topology and parameter files for the hybrid ligand were generated using MATCH[41] and in-house developed scripts based on the CHARMM General Force Field (CGenFF).[42] The simulation in water was done in a water box consisting of 800 TIP3P[43] water molecules with cubic periodic boundary conditions. The water box had a size of 30.0 Å × 30.0 Å × 30.0 Å. A nonbonded cutoff of 14 Å was used, and the van der Waals switching function and electrostatic force switching function[44] were used between 12 Å and 14 Å. Sampling from the conditional distribution $P(x, X|\lambda)$ was accomplished by running Langevin dynamics at 298.15 K for 0.2 ps. The time step size was 2 fs and the friction coefficient was 10 ps^{-1}. The length of all bonds involving hydrogen atoms was fixed during the simulation using the SHAKE algorithm.[45] The three relative hydration free energies were first calculated by three independent pairwise GSLDs. Then they were calculated simultaneously using the generalized GSLD for multiple ligands. For comparison, the three relative hydration free energies were also calculated using the FEP/

MBAR method, in which 11 states corresponding to λ = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1 were used.

3.3 Relative Binding Free Energy between Benzene and p-Xylene with T4 Lysozyme

The L99A mutant of T4 lysozyme has been a model protein system for testing free energy calculation methods.[46–48] It has experimental binding free energy data for a series of benzene derivatives including benzene and p-xylene.[37,38] The relative binding free energy between benzene and p-xylene was calculated using the difference between the alchemical free energy change in water and in the protein environment. The alchemical free energy change in water was calculated using pairwise GSLD with continuous λ. Calculating the alchemical free energy in the protein environment is challenging, even though the binding site of T4 lysozyme is a relatively simple non-polar pocket and the alchemical change from benzene to p-xylene is small. This challenge arises from the fact that T4 lysozyme has a conformational change for the side-chain dihedral angle χ (N-CA-CB-CG1) of residue Val111, which accompanies the alchemical transformation from benzene to p-xylene.[46] When T4 lysozyme binds with benzene (PDB ID: 181L), the dihedral angle stays in the *trans* conformation ($\chi \approx -180°$). When it binds with p-xylene (PDB ID: 187L), the dihedral angle changes into the *gauche* conformation ($\chi \approx -60°$). Failing to sample these two relevant conformations in a free energy calculation would cause a quasi-nonergodicity problem, i.e, the calculated free energy will depend on which conformation is used as the starting conformation.[46,48] To address the problem, several methods have been developed. These methods include enhanced sampling methods such as the 2-dimensional replica exchange method (REM)[47] and the free energy perturbation/replica exchange with solute tempering (FEP/REST) method,[48] and the potential of mean force (PMF) method, which was first introduced by Tobias and Brooks for addressing a similar problem in 1989[49] and rediscovered as the “confine-and-release” method by Mobley et al. in 2007.[46] Here we combined the PMF method with GSLD to calculate the alchemical free energy changes between benzene and p-xylene in the protein environment.

To make our computational protocol clear, we reformulated the PMF method[46,49] using conditional probability as shown in Appendix B. The free energy change $\Delta G(\chi^*)$ was calculated using pairwise GSLD with a harmonic restraint potential on χ to keep it near χ^* during the pairwise GSLD simulation. The force constant of the harmonic restraint potential was 1195.3 kcal/mol · radius^{-2}. In our calculations, we chose χ^* to be −180° and −60°, although the final calculated result ΔG did not depend on the choice of χ^*. In the pairwise GSLD, λ was chosen to be a discrete variable specified by the set $\{l_1, l_2, \ldots, l_{16}\}$. $\lambda = l_1$ corresponds to the physical state that the ligand is benzene and $\lambda = l_{16}$ corresponds to the physical state that the ligand is p-xylene. When λ was changed from l_1 to l_{16}, the ligand was alchemically transformed from benzene into p-xylene. During the alchemical transformation, the partial charges on benzene atoms were turned off first. Then the benzene atoms were transformed into p-xylene atoms before the partial charges on p-xylene atoms were turned on. A soft-core Lennard-Jones potential was used during the transformation.[50] The formula used for both electrostatic potential and the soft-core Lennard-Jones potential is shown in Table S1. The potential energy scaling factors used for each state $\lambda = l_i$ are also shown in Table S1. The free energy $-\beta^{-1}\ln P(\chi^*|\lambda = l_1)$ and the free energy $-\beta^{-1}\ln P(\chi^*|\lambda =$

l_{16}) were computed by calculating the potential of mean force (PMF) with respect to χ when T4 lysozyme binds with benzene ($\lambda = l_1$) and with p-xylene ($\lambda = l_{16}$), respectively. The simulations was run inside a TIP3P water box with a size of 79.0Å × 56.4Å × 55.4Å and rectangular periodic boundary conditions were used. The water box had 7112 water molecules in total. The CHARMM36 force field[51] was used for T4 lysozyme and the CHARMM General Force Field (CGenFF)[42] was used for the ligands. The nonbonded interaction options were the same as that used in the relative hydration free energy calculations.

4 Results

4.1 The Harmonic System

As shown in Figure 2(A), GSLD is able to sample the continuous λ well for both symmetrical and asymmetrical systems. Figure 2(B) shows the estimated free energy changes ΔG using the Rao-Blackwell estimator and two empirical estimators with cutoff values of 0.9 and 0.99. For the symmetrical system, the true value for the free energy changes is equal to 0 kcal/mol because of the symmetry. The RBE and the empirical estimator with cutoff of 0.9 converge to 0 kcal/mol within 2 ns, whereas the empirical estimator with cutoff of 0.99 needs 10 ns of simulation to converge to 0 kcal/mol. Moreover the RBE has the smallest variance among the three estimators. For the asymmetrical system, the empirical estimator with a cutoff of 0.9 converges to −0.41 ± 0.03 kcal/mol and the empirical estimator with a cutoff of 0.99 converges to −0.50 ± 0.06 kcal/mol, whereas the result from numerical integration is −0.56 kcal/mol. This shows that the results of empirical estimators can be biased and the bias depends on the value of the cutoff. Increasing the cutoff value decreases the estimation bias, but it increases the estimation variance because a higher cutoff decreases the number of valid samples used by the empirical estimator. In contrast, the result of RBE converges to −0.56 ± 0.02 kcal/mol, which is closest to the true value and also has the smallest variance. The detailed numerical results can be found in the Table S2. Overall, the results suggest that, for this harmonic system, the GSLD is able to extensively sample the alchemical states and the RBE is better than the empirical estimator in terms of both bias and variance.

4.2 Relative Hydration Free Energies for Three Benzene Derivatives

Results of pairwise GSLD simulations in vacuum and in water are shown in Figure S1 and Figure 3, respectively. The pairwise GSLD is able to sample the alchemical states very well for both the simulations in vacuum and the simulations in water. For the simulation in vacuum, the RBE outperforms empirical estimators in terms of both bias and variance, as in the harmonic system. For the simulation in water, the RBE has a similar variance to that of the empirical estimators, because samples from the simulation in water are more correlated than those from the simulations in vacuum. Nevertheless, the RBE is still better than the empirical estimators in terms of the bias. As shown in Figure 3(B), the empirical estimator depends on the cutoff. As the cutoff increases from 0.9 to 0.99, the empirical estimator results move towards to the RBE results. As an example, for the alchemical change from benzene to benzaldehyde, when the cutoff increases from 0.9 to 0.99, the empirical estimator result changes from 2.20 ± 0.08 kcal/mol to 2.60 ± .08 kcal/mol. The RBE result is 3.04

± 0.09 kcal/mol, which is indistinguishable from the FEP/MBAR result 3.01 ± 0.02 kcal/mol. The detailed numerical values from pairwise GSLD and FEP/MBAR can be found in the Table S3 and S4.

The simulation results in vacuum and in water from generalized GSLD for multiple ligands are shown in Figure S2 and Figure 4, respectively. The ternary plots[52] of (λ_1, λ_2, λ_3) trajectories show that the generalized GSLD is able to explore the hybrid ligand configuration space of (λ_1, λ_2, λ_3): the unit simplex $\{(\lambda_1, \lambda_2, \lambda_3) | \sum_{i=1}^{3} \lambda_i = 1, 0 \le \lambda_i \le 1$ for $i = 1, 2, 3\}$, in both vacuum and water. In vacuum, the configuration space (λ_1, λ_2, λ_3) is sampled rather uniformly, while in water, the configuration space is sampled mostly close to the physical states, i.e. the corners of the ternary plot in Figure 4. This difference is because the biasing potential energy used in this study is a linear biasing potential $\lambda_i G_i^b$. With the linear biasing potential, the biased free energy landscape over the configuration space (λ_1, λ_2, λ_3) in vacuum is almost flat. In water, the corresponding biased free energy landscape is not flat due to the polarization energy of the solvent interacting with reactant and product states, and the biased free energies of the physical states is lower than the intermediate non-physical states, which explains why the sampled (λ_1, λ_2, λ_3) are mostly around the physical states. Based on the trajectory from the generalized GSLD simulation, the calculated free energy using RBE and empirical estimators are shown in Figure S2 (B) and Figure 4(B). These results suggests again that, compared with the empirical estimators, the RBE is a better estimator as it has no bias and a smaller variance. The detailed numerical results from the generalized GSLD for multiple ligands is shown in the Table S5.

The calculated relative hydration free energies for the three benzene derivatives using pairwise GSLD, generalized GSLD for multiple ligands and FEP/MBAR methods are combined in Table 1. The results from all three methods agree well with each other. The total simulation time in water for calculating all three relative hydration free energies is 9 ns for pairwise GSLD, 3 ns for generalized GSLD for multiple ligands and 33 ns for FEP/MBAR methods, which suggests the efficacy of the generalized GSLD for multiple ligands.

4.3 Relative Binding Free Energy of Benzene and p-Xylene with T4 Lysozyme

The λ trajectories from the simulation with T4 lysozyme using pairwise GSLD and the free energy estimations using RBE are shown in Figure 5. For both the case where χ is restricted to the *trans* conformation ($\chi^* = -180°$) and the case where χ is restricted to the *gauche* ($\chi^* = -60°$) conformation, the pairwise GSLD is able to sample the alchemical switching variable λ well and the RBE estimations converge in 10 ns of simulation. When χ is restricted to the *trans* conformation, the estimated free energy converges to −8.40 ± 0.46 kcal/mol. When χ is restricted in the *gauche* conformation, the estimated free energy converges to −10.60 ± 0.36 kcal/mol. These two free energy estimations are different by 2.20 kcal/mol because the dihedral angle χ is restricted to different conformations. Based on the PMF method, in order to get the free energy corresponding to the case where χ is not restricted, the restricting free energies ($-\beta^{-1} \ln P(\chi^* | \lambda = l_1)$ and $-\beta^{-1} \ln P(\chi^* | \lambda = l_{16})$) need to be considered and used to correct the free energy $\Delta G(\chi^*)$ using equation 21 in Appendix B. These corrections are shown in Table 2. After the corrections, the estimated free energy ΔG is −9.27 ± 0.50 kcal/mol when $\chi^* = -180°$ and −9.01 ± 0.40 kcal/mol when $\chi^* = -60°$.

Therefore, after the corrections, the estimated free energy differences (ΔG) agree very well within statistical uncertainty. Based on the these corrected values, the relative binding free energies ($\Delta\Delta G$) are 0.27 ± 0.56 kcal/mol and 0.43 ± 0.46 kcal/mol when $\chi^* = -180°$ and $\chi^* = -60°$, respectively. These results are close to the relative binding free energy from experiment, which is 0.52 ± 0.22 kcal/mol.[37,38]

5 Discussion and Conclusion

Although the GSLD and RBE are applied only for calculating relative hydration free energy and relative binding free energy in this study, they could also be used for other purposes. One of the applications would be for calculating the pK_a value of protein amino acids by combining with the constant pH molecular dynamics methods (CPHMD),[18,53,54] as several CPHMD methods are based on λ-dynamics. Furthermore, the GSLD framework presented here is not limited to alchemical free energy calculations. The λ variable could be replaced by the pH values, which would correspond to pH generalized ensemble simulations. In these cases, we can also derive the corresponding RBE similarly.

In this study, we have presented the formalism for the Gibbs sampler based λ-dynamics (GSLD) and the Rao-Blackwell estimator (RBE) for alchemical free energy calculations. These methods were successfully demonstrated for three test cases of increasing complexity. The GSLD, a generalized ensemble sampling method, works for the case where λ is a discrete variable and for the case where λ is considered to be continuous. When λ is continuous, the GSLD can be generalized to calculate free energies for multiple ligands simultaneously in one simulation. The RBE not only eliminates the bias problem of the empirical estimator used in the original λ-dynamics, but also has smaller estimation variance than the empirical estimator. Moreover, we have also shown that the RBE can be used to derive the MBAR/UWHAM equations, which provides new understanding for the MBAR/UWHAM method.[21–23]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by grants from NIH (GM037554 and GM107233)

References

1. Jorgensen WL. The Many Roles of Computation in Drug Discovery. Science. 2004; 303:1813–1818. [PubMed: 15031495]
2. Bash PA, Singh UC, Langridge R, Kollman PA. Free Energy Calculations by Computer Simulation. Science. 1987; 236:564–8. [PubMed: 3576184]
3. Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS. Alchemical free energy methods for drug discovery: progress and challenges. Curr Opin Struct Biol. 2011; 21:150–160. [PubMed: 21349700]
4. Shirts, MR., Mobley, DL., Chodera, JD. Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time?. Elsevier; 2007. p. 41-59.

5. Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, Murcko M, Frye L, Farid R, Lin T, Mobley DL, Jorgensen WL, Berne BJ, Friesner RA, 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 Phys Chem B. 2015; 137:2695–2703.
6. Zwanzig RW. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. J Chem Phys. 1954; 22:1420–1426.
7. Zwanzig RW. High-Temperature Equation of State by a Perturbation Method. II. Polar Gases. J Chem Phys. 1955; 23:1915–1922.
8. Straatsma TP, Berendsen HJC. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. J Chem Phys. 1998; 89:5876–5886.
9. Christ CD, van Gunsteren WF. Enveloping distribution sampling: A method to calculate free energy differences from a single simulation. J Chem Phys. 2007; 126:184110. [PubMed: 17508795]
10. Riniker S, Christ CD, Hansen N, Mark AE, Nair PC, van Gunsteren WF. Comparison of enveloping distribution sampling and thermodynamic integration to calculate binding free energies of phenylethanolamine N-methyltransferase inhibitors. J Chem Phys. 2011; 135:024105. [PubMed: 21766923]
11. Kong X, Brooks CL III. λ-dynamics: A new approach to free energy calculations. J Chem Phys. 1998; 105:2414–2423.
12. Guo Z, Brooks CL III, Kong X. Efficient and flexible algorithm for free energy calculations using the λ-dynamics approach. J Phys Chem B. 1998; 102:2032–2036.
13. Guo Z, Brooks CL III. Rapid screening of binding affinities: application of the λ-dynamics method to a trypsin-inhibitor system. J Phys Chem B. 1998; 120:1920–1921.
14. Knight JL, Brooks CL III. λ-Dynamics free energy simulation methods. J Comput Chem. 2009; 30:1692–1700. [PubMed: 19421993]
15. Knight JL, Brooks CL III. Multisite λ Dynamics for Simulated Structure–Activity Relationship Studies. J Chem Theory Comput. 2011; 7:2728–2739. [PubMed: 22125476]
16. Armacost KA, Goh GB, Brooks CL III. Biasing Potential Replica Exchange Multisite λ-Dynamics for Efficient Free Energy Calculations. J Chem Theory Comput. 2015; 11:1267–1277. [PubMed: 26579773]
17. Zheng L, Chen M, Yang W. Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. Proc Natl Acad Sci U S A. 2008; 105:20227–20232. [PubMed: 19075242]
18. Goh GB, Hulbert BS, Zhou H, Brooks CL III. Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism. Proteins: Struct Funct Bioinf. 2014; 82:1319–1331.
19. Hayes RL, Armacost KA, Vilseck JZ, Brooks CL. Adaptive Landscape Flattening Accelerates Sampling of Alchemical Space in Multisite Îż Dynamics. J Phys Chem B. 2017; doi: 10.1021/acs.jpcb.6b09656
20. Chodera JD, Shirts MR. Replica exchange and expanded ensemble simulations as gibbs sampling: Simple improvements for enhanced mixing. J Chem Phys. 2011; 135:194110. [PubMed: 22112069]
21. Shirts MR, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys. 2008; 129:124105. [PubMed: 19045004]
22. Tan Z, Gallicchio E, Lapelosa M, Levy RM. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. J Chem Phys. 2012; 136:144102. [PubMed: 22502496]
23. Zhang BW, Xia J, Tan Z, Levy RM. A stochastic solution to the unbinned WHAM equations. J Phys Chem Lett. 2015; 6:3834–3840. [PubMed: 26722879]
24. Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans Pattern Anal Mach Intell. 1984; PAMI-6:721–741.
25. Smith AF, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J R Stat Soc B. 1993:3–23.

26. Ritter C, Tanner MA. Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. J Am Stat Assoc. 1992; 87:861–868.
27. Wang F, Landau DP. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. Phys Rev Lett. 2001; 86:2050–2053. [PubMed: 11289852]
28. Zhang BW, Dai W, Gallicchio E, He P, Xia J, Tan Z, Levy RM. Simulating replica exchange: Markov state models, proposal schemes, and the infinite swapping limit. J Phys Chem B. 2016; 120:8289–8301. [PubMed: 27079355]
29. Plattner N, Doll J, Dupuis P, Wang H, Liu Y, Gubernatis J. An infinite swapping approach to the rare-event sampling problem. J Chem Phys. 2011; 135:134111. [PubMed: 21992286]
30. Dupuis P, Liu Y, Plattner N, Doll JD. On the infinite swapping limit for parallel tempering. Multiscale Model Simul. 2012; 10:986–1022.
31. Plattner N, Doll J, Meuwly M. Overcoming the rare event sampling problem in biological systems with infinite swapping. J Chem Theory Comput. 2013; 9:4215–4224. [PubMed: 26592410]
32. Beutler TC, Mark AE, van Schaik RC, Gerber PR, van Gunsteren WF. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. Chem Phys Lett. 1994; 222:529–539.
33. Rao CR. Information and accuracy attainable in the estimation of statistical parameters. Bull Calcutta Math Soc. 1945; 37:81–91.
34. Blackwell D. Conditional expectation and unbiased sequential estimation. Ann Math Stat. 1947:105–110.
35. Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc. 1990; 85:398–409.
36. Pearl J. Evidential reasoning using stochastic simulation of causal models. Artificial Intelligence. 1987; 32:245–257.
37. Morton A, Matthews BW. Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: linkage of dynamics and structural plasticity. Biochemistry. 1995; 34:8576–8588. [PubMed: 7612599]
38. Morton A, Baase WA, Matthews BW. Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 lysozyme. Biochemistry. 1995; 34:8564–8575. [PubMed: 7612598]
39. Brooks BR, Brooks CL III, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The biomolecular simulation program. J Comput Chem. 2009; 30:1545–1614. [PubMed: 19444816]
40. Eastman P, Friedrichs MS, Chodera JD, Radmer RJ, Bruns CM, Ku JP, Beauchamp KA, Lane TJ, Wang LP, Shukla D, Tye T, Houston M, Stich T, Klein C, Shirts MR, Pande VS. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. J Chem Theory Comput. 2012; 9:461–469. [PubMed: 23316124]
41. Yesselman JD, Price DJ, Knight JL, Brooks CL III. MATCH: An atom-typing toolset for molecular mechanics force fields. J Comput Chem. 2012; 33:189–202. [PubMed: 22042689]
42. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. 2010; 31:671–690. [PubMed: 19575467]
43. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983; 79:926–935.
44. Steinbach PJ, Brooks BR. New spherical-cutoff methods for long-range forces in macromolecular simulation. J Comput Chem. 1994; 15:667–683.
45. Van Gunsteren WF, Berendsen HJC. Algorithms for macromolecular dynamics and constraint dynamics. Mol Phys. 2006; 34:1311–1327.
46. Mobley DL, Chodera JD, Dill KA. The Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. J Chem Theory Comput. 2007; 3:1231–1235. [PubMed: 18843379]

47. 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. [PubMed: 21857813]
48. Wang L, Berne BJ, Friesner RA. On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. Proc Natl Acad Sci U S A. 2012; 109:1937–1942. [PubMed: 22308365]
49. Tobias DJ, Brooks CL III, Fleischman SH. Conformational flexibility in free energy simulations. Chem Phys Lett. 1989; 156:256–260.
50. Shirts MR, Pande VS. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. J Chem Phys. 2005; 122:134508. [PubMed: 15847482]
51. Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, MacKerell AD Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ 1and χ 2Dihedral Angles. J Chem Theory Comput. 2012; 8:3257–3273. [PubMed: 23341755]
52. Harper, M., Weinstein, B., Simon, C., chebee7i, Swanson-Hysell, N., The Gitter, B., Maximiliano, G., Guido, Z. python-ternary: Ternary Plots in Python. 2015. https://zenodo.org/record/34938#.WJeZDxIrKGg,https://github.com/marcharper/python-ternary
53. Goh GB, Knight JL, Brooks CL III. Constant pH Molecular Dynamics Simulations of Nucleic Acids in Explicit Solvent. J Chem Theory Comput. 2011; 8:36–46.
54. Wallace JA, Shen JK. Continuous Constant pH Molecular Dynamics in Explicit Solvent with pH-Based Replica Exchange. J Chem Theory Comput. 2011; 7:2617–2629. [PubMed: 26606635]

Appendix

A. A Wang-Landau like algorithm to automatically determine the biasing potential G1b used in pairwise GSLD when λ is continuous

The purpose of the biasing potential G_1^b used in the pairwise GSLD when λ is continuous is to make the biased free energy landscape over the λ space flat, i.e. to make the simulation spend about equal time at all λ values between 0 and 1. In current study, a linear biasing potential λG_1^b is utilized, because with the linear biasing potential the biased free energy landscape over λ space is quite flat, i.e. the energy barrier between the two physical states λ = 0 and λ = 1 is small enough that the λ is well sampled across the interval [0, 1]. If the linear biasing potential energy cannot make the biased free energy landscape over the λ space flat enough, a quadratic form of biasing potential can be utilized as in Hayes et al.'s flattening method.[19] The biasing potential G_1^b is determined automatically using the following Wang-Landau like algorithm:

- Set the initial biasing potential $G_1^b{=}0\,\mathrm{kcal/mol}$, the decay parameter a such that $0 < a < 1$ (a = 0.998 in this study), the biasing potential increment Δ in each step (Δ = 2.0 kcal/mol in this study) and the number of steps R (R = 3000 in this study). Initialize the starting state (λ^0, $\{x_i^0\}_{i=0}^1$, X^0).
- For t = 1 to R:

  Sample ($\{x_i^t\}_{i=0}^1$, X^t) from the conditional distribution: $P(\{x_i^t\}_{i=0}^1, X^t|\lambda^{t-1})$ by running molecular dynamics simulations and then sample λ^t from the

conditional distribution $P(\lambda^t|\{x_i^t\}_{i=0}^1, X^t)$. Set $G_1^b(t)=G_1^b(t-1)+(\lambda^t-0.5)*\Delta(t)$ and $\Delta(t) = \alpha * \Delta(t-1)$.

- The final value of G_1^b from the above step is fixed and used as the biasing potential in following simulations.

B. Reformulation of the PMF method using conditional probability

The PMF method requires prior knowledge of which slow degree of freedom is affecting the free energy calculation. In the context of T4 lysozyme, the slow degree of freedom is the side-chain dihedral angle N-CA-CB-CG1 (χ) of residue Val111. The joint distribution of (χ, λ) : $P(\chi, \lambda)$ is of most interest, as it encapsulates all the relevant information required to calculate the free energy $\Delta G = -\beta^{-1}\ln(P(\lambda = l_{16})/P(\lambda = l_1))$. Based on the chain rule of conditional probability, we have the following equations:

$$\begin{aligned} P(\chi=\chi^*,\lambda=l_{16})&=P(\chi=\chi^*|\lambda=l_{16})P(\lambda=l_{16})=P(\lambda=l_{16}|\chi=\chi^*)P(\chi=\chi^*)\\ P(\chi=\chi^*,\lambda=l_1)&=P(\chi=\chi^*|\lambda=l_1)P(\lambda=l_1)=P(\lambda=l_1|\chi=\chi^*)P(\chi=\chi^*) \end{aligned} \quad (19)$$

Combining the above two equation gives us:

$$\frac{P(\lambda=l_{16})}{P(\lambda=l_1)}=\frac{P(\lambda=l_{16}|\chi=\chi^*)}{P(\lambda=l_1|\chi=\chi^*)}\cdot\frac{P(\chi=\chi^*|\lambda=l_1)}{P(\chi=\chi^*|\lambda=l_{16})}. \quad (20)$$

Therefore, we can calculate the free energy ΔG as

$$\begin{aligned} \Delta G&=-\beta^{-1}\ln\frac{P(\lambda=l_{16})}{P(\lambda=l_1)}\\ &=-\beta^{-1}\ln\frac{P(\lambda=l_{16}|\chi=\chi^*)}{P(\lambda=l_1|\chi=\chi^*)}-\beta^{-1}\ln\frac{P(\chi=\chi^*|\lambda=l_1)}{P(\chi=\chi^*|\lambda=l_{16})}\\ &=\Delta G(\chi=\chi^*)+[-\beta^{-1}\ln P(\chi=\chi^*|\lambda=l_1)]-[-\beta^{-1}\ln P(\chi=\chi^*|\lambda=l_{16})], \end{aligned} \quad (21)$$

where $\Delta G(\chi = \chi^*)$ is alchemical free energy change when χ is fixed at the value χ^*; $-\beta^{-1}\ln P(\chi = \chi^*|\lambda = l_1)$ is the free energy required to restrict the dihedral angle χ at the value χ^* when T4 lysozyme binds with benzene, i.e, $\lambda = l_1$; $-\beta^{-1}\ln P(\chi = \chi^*|\lambda = l_{16})$ is the corresponding free energy required when T4 lysozyme binds with p-xylene, i.e, $\lambda = l_{16}$. The above equation holds regardless of the value of χ^*.

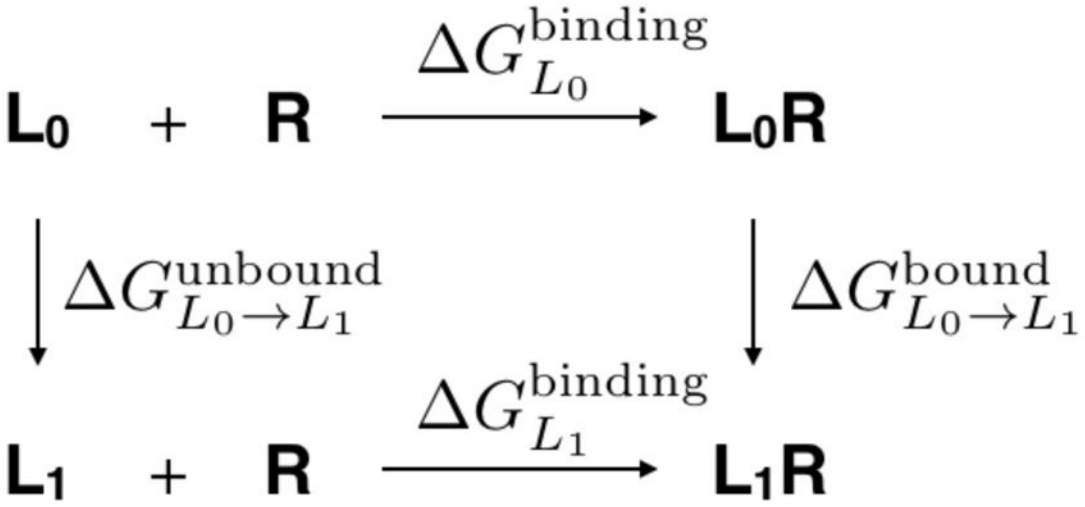


Figure 1.

The thermodynamic cycle used for calculating a relative binding free energy between ligand L_0 and L_1 with a receptor R.

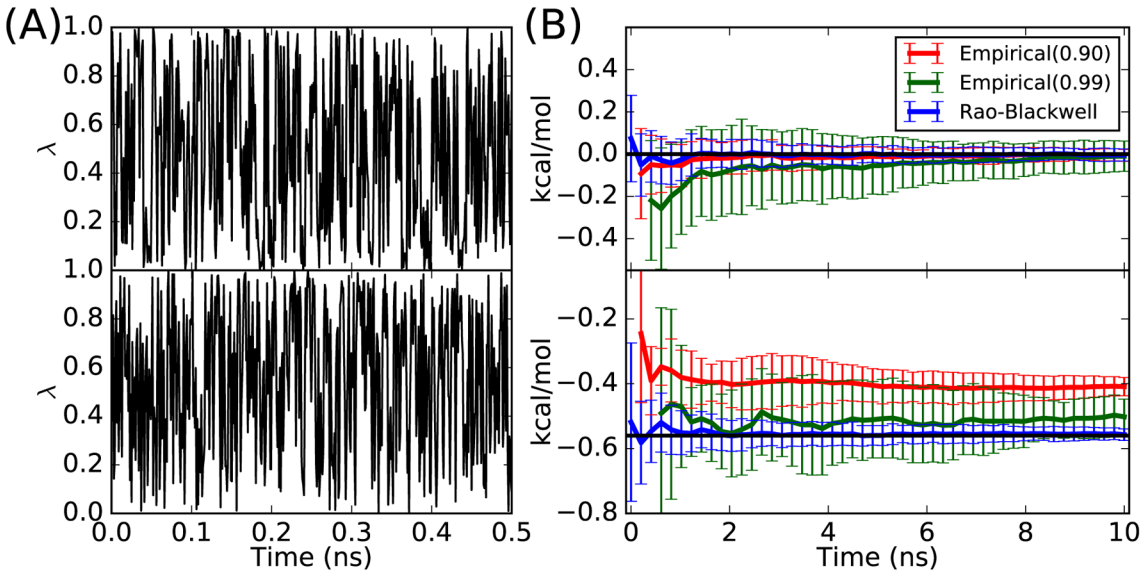


Figure 2.

(A) λ trajectories for the symmetrical harmonic system (top) and the asymmetrical harmonic system (bottom); (B) Free energy estimations for the symmetrical system (top) and the asymmetrical system (bottom) using the empirical estimators with a cutoff of 0.9 and 0.99 and the Rao-Blackwell estimator. The horizontal black line is the calculated free energy change using numerical integration.

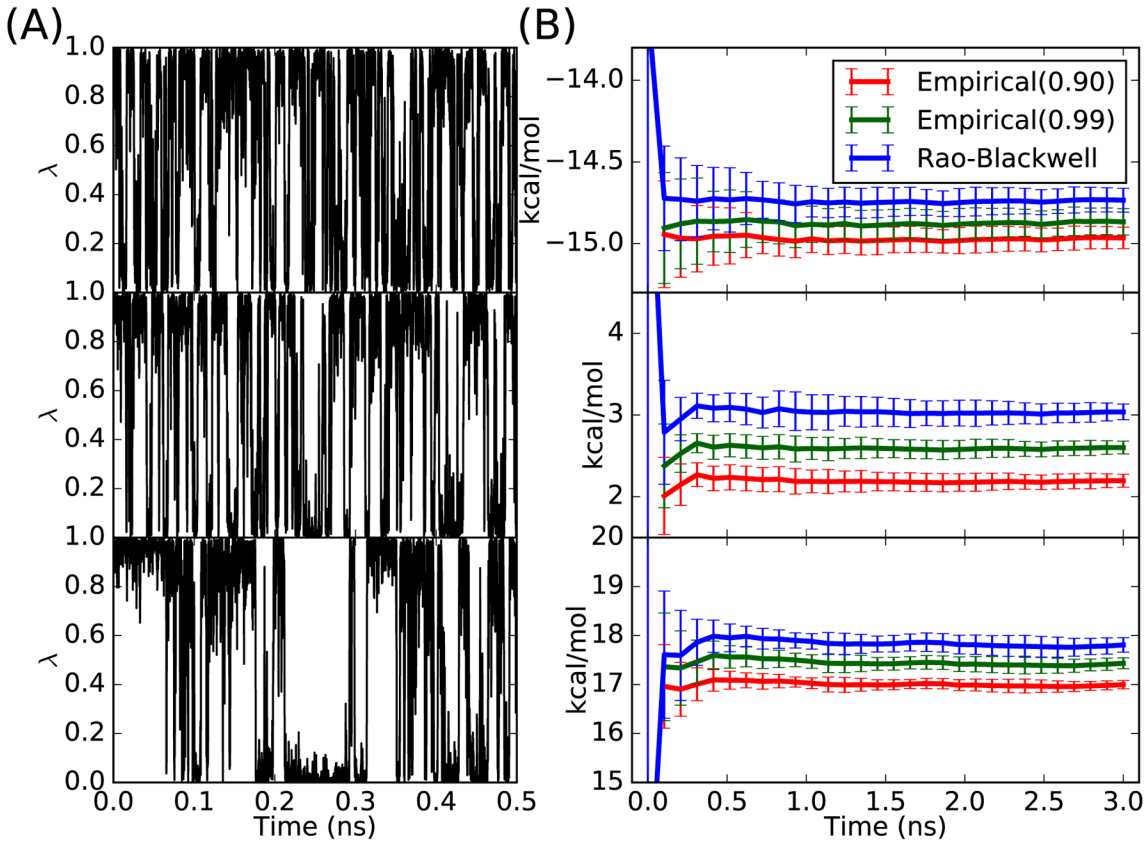


Figure 3.

(A) λ trajectories from simulations in water using GSLD for alchemical changes benzene to phenol (top), benzene to benzaldehyde (middle), and phenol to benzaldehyde (bottom). **(B)** Estimated alchemical free energy changes in water using empirical estimators with different cutoff values and the Rao-Blackwell estimator for alchemical changes benzene to phenol (top), benzene to benzaldehyde (middle), and phenol to benzaldehyde (bottom).

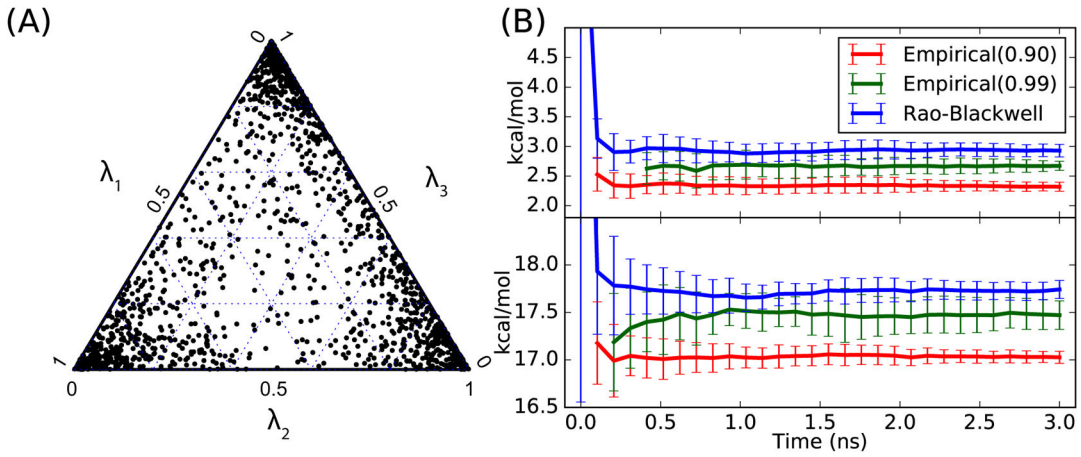


Figure 4.

(A) Ternary plot of (λ_1, λ_2, λ_3) sampled using GSLD for multiple ligands in water. **(B)** Estimated free energy changes in water for alchemical changes: benzene to benzaldehyde (top) and phenol to benzaldehyde (bottom) using empirical estimator with different cutoff values and RBE.

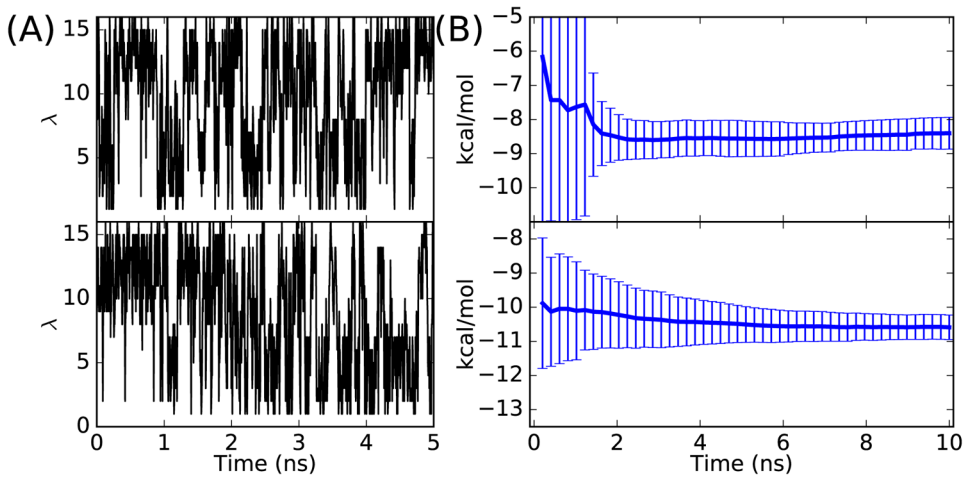


Figure 5.

(A) λ trajectories for simulations with T4 lysozyme using pairwise GSLD for the χ^* = −180° (top) and χ^* = −60° (bottom); (B) Free energy estimation using RBE for χ^* = −180° (top) and χ^* = −60° (bottom).

Table 1

Comparison of Relative Hydration Free Energies ($\Delta\Delta G$ in kcal/mol) for The Three Benzene Derivatives. The total simulation time in water for each method is shown in parenthesis.

substituents change	$\Delta\Delta G_{exp}$	Pairwise GSLD $\Delta\Delta G$(9ns)	GSLD for Multiple Ligands $\Delta\Delta G$(3ns)	FEP/MBAR $\Delta\Delta G$(33ns)
Benzene → Phenol	−5.77	−4.46 ± 0.08	−4.53 ± 0.15	−4.46 ± 0.03
Benzene → Benzaldehyde	−3.18	−3.11 ± 0.11	−3.22 ± 0.11	−3.13 ± 0.03
Phenol → Benzaldehyde	2.59	1.39 ± 0.17	1.31 ± 0.10	1.34 ± 0.14

Table 2

Alchemical Free Energy Changes (kcal/mol) Between Benzene and p-Xylene Binding with T4 Lysozyme Calculated Using Pairwise GSLD with Corrections from PMFs.

χ^*	$\Delta G(\chi^*)$	$-\beta^{-1}\ln P(\chi^*\|\lambda = 0)$	$-\beta^{-1}\ln P(\chi^*\|\lambda = 1)$	ΔG	$\Delta\Delta G$
trans ($\chi^* = -180°$)	−8.40 ± 0.46	−0.47 ± 0.01	0.4 ± 0.03	−9.27 ± 0.50	0.27 ± 0.56
gauche ($\chi^* = -60°$)	−10.60 ± 0.36	1.14 ± 0.03	−0.45 ± 0.01	−9.01 ± 0.40	0.43 ± 0.46

The alchemical free energy change ΔG in water is −9.44 ± 0.06 kcal/mol and the experimental relative binding free energy $\Delta\Delta G$ is 0.52 ± 0.22 kcal/mol.