JCTC Journal of Chemical Theory and Computation

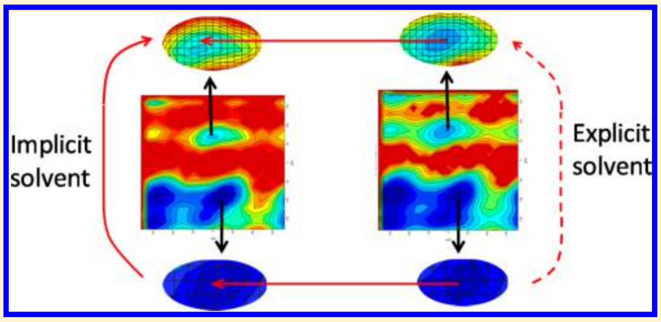
Article

pubs.acs.org/JCTC

Connecting Free Energy Surfaces in Implicit and Explicit Solvent: An Efficient Method To Compute Conformational and Solvation Free Energies

Nanjie Deng,*[†,‡] Bin W. Zhang,[†,‡] and Ronald M. Levy*[†,‡]

[†]Center for Biophysics & Computational Biology and Institute for Computational Molecular Sciences and [‡]Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States



ABSTRACT: The ability to accurately model solvent effects on free energy surfaces is important for understanding many biophysical processes including protein folding and misfolding, allosteric transitions, and protein−ligand binding. Although all-atom simulations in explicit solvent can provide an accurate model for biomolecules in solution, explicit solvent simulations are hampered by the slow equilibration on rugged landscapes containing multiple basins separated by barriers. In many cases, implicit solvent models can be used to significantly speed up the conformational sampling; however, implicit solvent simulations do not fully capture the effects of a molecular solvent, and this can lead to loss of accuracy in the estimated free energies. Here we introduce a new approach to compute free energy changes in which the molecular details of explicit solvent simulations are retained while also taking advantage of the speed of the implicit solvent simulations. In this approach, the slow equilibration in explicit solvent, due to the long waiting times before barrier crossing, is avoided by using a thermodynamic cycle which connects the free energy basins in implicit solvent and explicit solvent using a localized decoupling scheme. We test this method by computing conformational free energy differences and solvation free energies of the model system alanine dipeptide in water. The free energy changes between basins in explicit solvent calculated using fully explicit solvent paths agree with the corresponding free energy differences obtained using the implicit/explicit thermodynamic cycle to within 0.3 kcal/mol out of ~3 kcal/mol at only ~8% of the computational cost. We note that WHAM methods can be used to further improve the efficiency and accuracy of the implicit/explicit thermodynamic cycle.

■ INTRODUCTION

Free energy differences between conformational basins provide the thermodynamic driving force for many biophysical processes including protein folding/misfolding, allosteric transitions, and protein−ligand binding. The ability to accurately compute free energy differences is therefore of both fundamental and practical importance to biophysics.[1−3] Molecular dynamics simulations in explicit solvent provide the most detailed information about solvation effects on biomolecules and are widely used to estimate conformational free energies. However, accurate free energy calculations require extensive sampling of conformational space, which is challenging because of the slow equilibration in an explicit solvent and the complexity of the energy landscape containing multiple basins separated by barriers. While Replica Exchange Molecular Dynamics (REMD)[4,5] and other advanced sampling methods (e.g., metadynamics,[6] Accelerated MD,[7] adaptive umbrella sampling,[8] transition path sampling,[9] Milestoning,[10] and Markov State Model[11−14]) have been developed to enhance the sampling of the conformational space, using these powerful techniques in explicit solvent can still be computationally demanding: for example in a temperature REMD simulation of a solvated system, the number of temperature replicas required to achieve an adequate acceptance ratio is typically quite large, and scales as $\propto \sqrt{N}$, where N is the number of degrees of freedom. To increase the efficiency of sampling in REMD simulations in explicit solvent, specialized techniques like Replica Exchange with Solute Tempering have been developed and applied to protein folding and ligand binding studies.[15,16]

During the past decade implicit solvent models have increasingly been used in free energy calculations to circumvent some of the problems associated with explicit solvent simulations.[17−22] When performing molecular dynamics simulations with implicit solvent models, not only is the computation of each step faster because the number of degrees of freedom is much smaller than when solvent is included in the model explicitly, but perhaps more importantly from the perspective of computational efficiency, the solvent contribution to the solute potential of mean force is calculated analytically as a function of the solute coordinates so that the solvent fluctuations are already averaged. The absence of water friction in implicit solvent is also potentially helpful in sampling

Received: March 20, 2015
Published: April 27, 2015

DOI: 10.1021/acs.jctc.5b00264
J. Chem. Theory Comput. 2015, 11, 2868−2878

the solute conformational space, but for some problems the water may actually act as a lubricant. Lastly, because implicitly solvated systems contain fewer degrees of freedom, they are better suited for REMD simulations. However, because the effects of a molecular solvent are modeled in an averaged, mean field fashion, implicit solvent simulations can be less accurate than their explicit solvent counterpart, for instance in systems where a few specific waters play important roles in the solute energetics and dynamics.[23−26]

Here we present an approach to connect free energy surfaces in explicit and implicit solvents for the purpose of constructing a thermodynamic cycle that combines desirable features of explicit solvent models (increased accuracy) with those of implicit solvent models (speed). In a MD calculation of the conformational free energy difference between two or more basins separated by barriers, the computationally most expensive step comes from the need to sample the reversible crossing of the barrier for a sufficient number of times to achieve equilibration; the sampling within individual free energy basins is often fast even in explicit solvent simulations. On the other hand, the sampling of the barrier crossing can be more readily achieved using computationally less expensive implicit solvent simulations. The idea here is to use the fast implicit solvent simulation to generate an initial estimate of the full free energy surface and then compute the effects of explicit solvent as a "correction" to the implicit solvent results by using a thermodynamic cycle that connects the free energy surfaces of the individual conformational basins obtained from the implicit and explicit solvent models. Here the connection between the two free energy surfaces is realized using localized decoupling simulations; it can also be done using various end-point methods. The key advantage of this approach is that the sampling of the full free energy surface in explicit solvent is replaced by a combination of implicit solvent simulations of the barrier crossing, implicit and explicit solvent simulations within each basin, and a small number of localized decoupling simulations which link the free energy surfaces and are computationally much less expensive than the fully explicit solvent simulations of the free energy changes.

We test this approach using solvated alanine dipeptide as an example. The method yields conformational free energy differences between pairs of basins that are within ∼0.2 kcal/mol of those obtained from exhaustive explicit solvent simulations (where the total free energy changes are ∼3 kcal/mol) at just ∼8% of the computational cost of the direct MD sampling in explicit solvent. In addition, we show that our method of connecting free energy surfaces can also be used to obtain accurate solvation (transfer) free energies of solute molecules in water with complex free energy landscapes.

■ METHODS

We consider a solute molecule in solution containing two free energy basins A and B separated by a barrier: see Figure 1a. System 0 stands for the solute in implicit solvent and system 1 for the solute in explicit solvent. The two free energy basins are divided into N_A and N_B cells, which are based on a set of suitably chosen order parameters (see Figure 2, for the alanine dipeptide example). Using the thermodynamic cycle depicted in Figure 1b, the free energy difference between basin A and basin B in explicit solvent, $\Delta G_{1,A\to B}$, can be obtained without sampling reversible transitions between the two basins in the explicit solvent system 1. We start from the expression

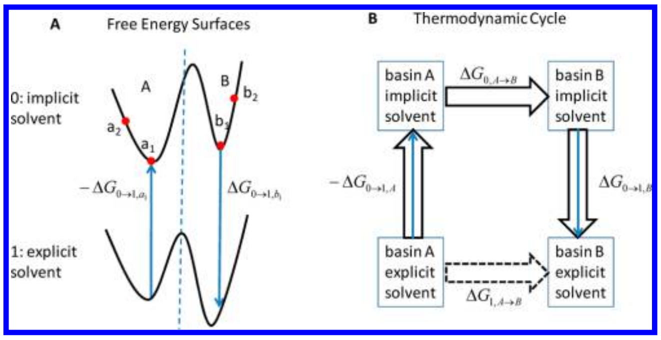


Figure 1. (a) A schematic diagram illustrating the calculation of conformational free energy differences by connecting free energy surfaces in explicit and implicit solvents via localized decoupling. (b) Thermodynamic cycle. Note that $\Delta G_{0\to 1,a_1}$ is the transfer free energy for a cell a_1 in basin A from implicit solvent to explicit solvent, while $\Delta G_{0\to 1,A}$ is the transfer free energy for the whole basin A from implicit to explicit solvent which depends on both the localized transfer free energy $\Delta G_{0\to 1,a_1}$ and a curvature term, see text.

$$\begin{aligned}\Delta G_{1,A\to B} &= -kT\ \ln\frac{Z_{1,B}}{Z_{1,A}} = -kT\ \ln\left(\frac{Z_{0,A}}{Z_{1,A}}\frac{Z_{0,B}}{Z_{0,A}}\frac{Z_{1,B}}{Z_{0,B}}\right)\\ &= -\Delta G_{0\to 1,A} + \Delta G_{0,A\to B} + \Delta G_{0\to 1,B}\end{aligned} \tag{1}$$

Here $Z_{n,X} = \int_{r\in X} e^{-H_n(r)/kT}dr$ is the configuration integral for basin X in system n. $\Delta G_{0,A\to B}$ is the free energy difference between basin A and basin B in the implicit solvent system 0, which is readily obtained from direct simulation in the implicit solvent. $\Delta G_{0\to 1,A}$ is the free energy for transferring the free energy surface of basin A in implicit solvent to explicit solvent; $\Delta G_{0\to 1,B}$ is the corresponding transfer free energy for basin B

$$\Delta G_{0\to 1,A} = -kT\ \ln\frac{Z_{1,A}}{Z_{0,A}},\qquad \Delta G_{0\to 1,B} = -kT\ \ln\frac{Z_{1,B}}{Z_{0,B}} \tag{2}$$

By dividing each of the basins in the two free energy surfaces into multiple cells and performing decoupling simulations focusing on one of the cells in each basin, a_1 and b_1, the transfer free energies can be written

$$\begin{aligned}\Delta G_{0\to 1,A} &= -kT\ \ln\frac{Z_{1,A}}{Z_{0,A}} = -kT\ \ln\left(P_{0,a_1}^{A}e^{-\Delta G_{0\to 1,a_1}/kT}\frac{1}{P_{1,a_1}^{A}}\right)\\ \Delta G_{0\to 1,B} &= -kT\ \ln\frac{Z_{1,B}}{Z_{0,B}} = -kT\ \ln\left(P_{0,b_1}^{B}e^{-\Delta G_{0\to 1,b_1}/kT}\frac{1}{P_{1,b_1}^{B}}\right)\end{aligned} \tag{3}$$

Equation 3 expresses the transfer free energy of a basin from implicit to explicit solvent in terms of the transfer free energy of a cell within the basin and the relative probabilities of occupying that cell within the basin in the two solvent environments.

In this paper we use capital letters A and B for basins and lower case letters a_1 and b_1 for cells in the corresponding basins. For example P_{0,a_1}^{A} and P_{0,b_1}^{B} are the population fractions of the cells a_1 and b_1 normalized with respect to the populations of basin A and basin B, respectively, on surface "0" (the implicit solvent surface). The superscripts A and B in P_{0,a_1}^{A} and P_{0,b_1}^{B} indicate that the population fractions are normalized relative to the total population of the corresponding basins: for example,

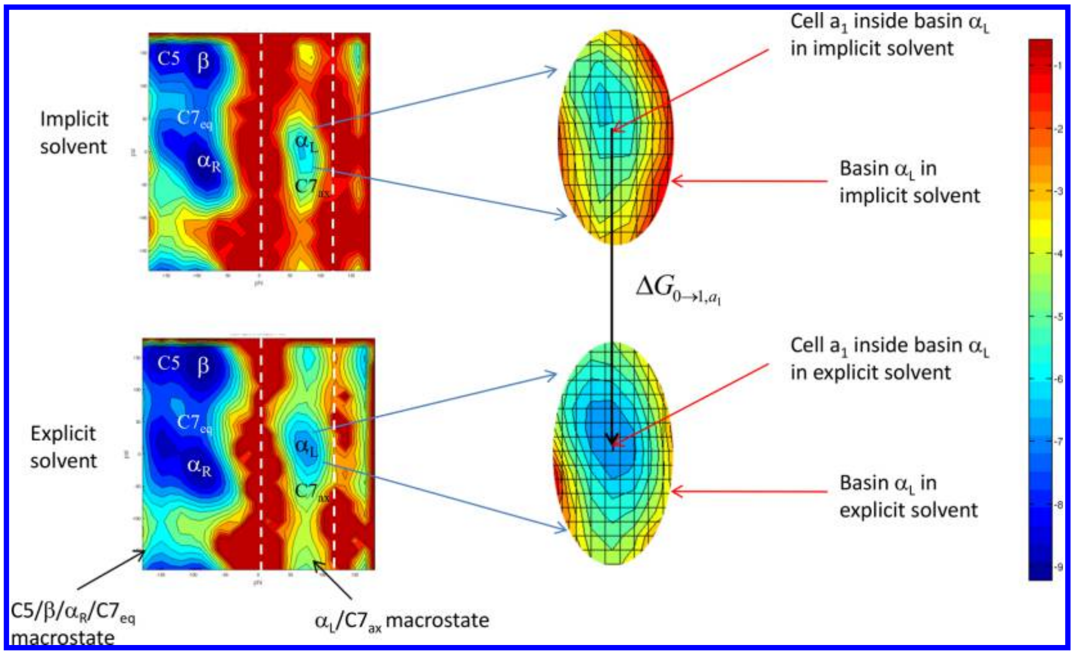


Figure 2. Dividing the free energy surface of a basin (e.g., α_L) into multiple cells and calculating the free energy of transferring a cell (a_1) of the α_L basin in implicit solvent to explicit solvent. The examples shown here are the free energy surfaces of alanine dipeptide in an implicit solvent (AGBNP2) and explicit solvent (TIP3P), projected onto the plane of $\phi-\psi$ dihedral angles. Also shown are the C5/β/α_R/C7$_{eq}$ macrostate ($-180° < \phi < 0°$) and α_L/C7$_{ax}$ macrostate ($0° < \phi < 120°$), each containing basins that interconvert rapidly.

P^B_{0,b_1} is the population of cell b_1 relative to the population of basin B and *not* relative to the total population of the complete free energy surface for alanine dipeptide which is normalized to 1.

To verify eq 3, first suppose that there are only two cells in basin A, a_1 and a_2. Then

$$\begin{aligned}\Delta G_{0\to 1,A} &= -kT\ln\frac{Z_{1,A}}{Z_{0,A}} = -kT\ln\left(\frac{Z_{1,a_1}}{Z_{0,A}}+\frac{Z_{1,a_2}}{Z_{0,A}}\right)\\ &= -kT\ln\left(\frac{Z_{1,a_1}}{Z_{0,a_1}}\frac{Z_{0,a_1}}{Z_{0,A}}+\frac{Z_{1,a_2}}{Z_{0,a_2}}\frac{Z_{0,a_2}}{Z_{0,A}}\right)\\ &= -kT\ln(e^{-\Delta G_{0\to 1,a_1}/kT}P^A_{0,a_1}+e^{-\Delta G_{0\to 1,a_2}/kT}P^A_{0,a_2})\end{aligned} \tag{4}$$

We also have

$$\begin{aligned}\Delta G_{0\to 1,a_2} &= \Delta G_{0,a_2\to a_1} + \Delta G_{0\to 1,a_1} + \Delta G_{1,a_1\to a_2}\\ &= -kT\ln\frac{P_{0,a_1}}{P_{0,a_2}} + \Delta G_{0\to 1,a_1} - kT\ln\frac{P_{1,a_2}}{P_{1,a_1}}\end{aligned} \tag{5}$$

Here $P_{n,x}$ is the population of cell x in the system n. For example P_{0,a_1} represents the population of cell a_1 of basin A in implicit solvent relative to the population of the entire free energy surface (i.e., the total population of all the basins in both macrostates which is normalized to 1). Note the difference between P_{0,a_1} and P^A_{0,a_1}; the latter represents the population fraction of the same cell in implicit solvent relative to the population of basin A.

Substituting eq 5 into eq 4 and simplifying

$$\begin{aligned}\Delta G_{0\to 1,A} &= -kT\ln\left(e^{-\Delta G_{0\to 1,a_1}/kT}P^A_{0,a_1} + e^{-\Delta G_{0\to 1,a_1}/kT}\frac{P_{1,a_2}}{P_{1,a_1}}P^A_{0,a_1}\right)\\ &= -kT\ln\left[P^A_{0,a_1}e^{-\Delta G_{0\to 1,a_1}/kT}\left(1+\frac{P_{1,a_2}}{P_{1,a_1}}\right)\right]\end{aligned} \tag{6}$$

Extending eq 6 to the situation in which basin A has N_A cells, we have

$$\begin{aligned}\Delta G_{0\to 1,A} &= -kT\ln\left[P^A_{0,a_1}e^{-\Delta G_{0\to 1,a_1}/kT}\left(1+\frac{P_{1,a_2}}{P_{1,a_1}}+\cdots+\frac{P_{1,a_{N_A}}}{P_{1,a_1}}\right)\right]\\ &= -\,kT\ln\left(P^A_{0,a_1}e^{-\Delta G_{0\to 1,a_1}/kT}\frac{P_{1,A}}{P_{1,a_1}}\right)\\ &= -kT\ln\left(P^A_{0,a_1}e^{-\Delta G_{0\to 1,a_1}/kT}\frac{1}{P^A_{1,a_1}}\right)\end{aligned} \tag{7}$$

which is eq 3.

Substituting eq 3 into eq 1 yields the formula for the conformational free energy difference between basin A and basin B in explicit solvent, system 1:

$$\begin{aligned}\Delta G_{1,A\to B} &= -\Delta G_{0\to 1,A} + \Delta G_{0,A\to B} + \Delta G_{0\to 1,B}\\ &= kT\ln\frac{P^A_{0,a_1}}{P^A_{1,a_1}} - \Delta G_{0\to 1,a_1} + \Delta G_{0,A\to B}\\ &\quad - kT\ln\frac{P^B_{0,b_1}}{P^B_{1,b_1}} + \Delta G_{0\to 1,b_1}\end{aligned} \tag{8}$$

All the quantities on the right-hand side of eq 8 are readily obtainable without the expensive explicit solvent simulation of the barrier crossing from A to B. The population fractions P^A_{1,a_1} and P^B_{1,b_1} are cell populations in the explicit solvent. Since they are normalized respectively to the populations of basin A and B, P^A_{1,a_1} and P^B_{1,b_1} can be obtained by local sampling within each basin in the explicit solvent system 1, without crossing the barrier. P^A_{0,a_1} and P^B_{0,b_1} are the corresponding cell populations in the implicit solvent system 0, which are readily obtainable. $\Delta G_{0\to 1,a_1}$ and $\Delta G_{0\to 1,b_1}$ are transfer free energies from the implicitly solvated system 0 to the explicitly solvated system 1 for cells a_1 and b_1, respectively, which can be calculated using two restrained decoupling simulations (vertical lines in Figure 1 and Figure 2). Because they are restricted to single cells (or to a small number of cells more generally) on the free energy surfaces, the two decoupling simulations are computationally fast.

On the right-hand side of eq 8, the three terms $\Delta G_{0\to 1,a_1}$, $\Delta G_{0\to 1,b_1}$, and $\Delta G_{0,A\to B}$ have straightforward meanings. The other two terms $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$ and $-kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$ contain information about the differences in the curvatures of the two free energy surfaces within the two corresponding basins. The more different the curvatures of the free energy surfaces within the basins between the implicit solvent and explicit solvent surfaces, the larger the magnitudes of these two terms. For example, suppose that cell a_1 is located at a relatively deep minimum of basin A in implicit solvent 0 but is not at the minimum of basin A in explicit solvent system 1. In this case the quantity $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$ will be large, making a substantial contribution to the conformational free energy $\Delta G_{1,A\to B}$ in eq 8. On the other hand, if the two free energy surfaces have similar curvatures within basin A, then the quantity $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$ will be small.

In a similar fashion to eq 8, for a solute molecule containing two macrostates A and B (e.g., alanine dipeptide), the total transfer free energy from vacuum or implicit solvent (system 0) to explicit solvent (system 1) can be written in terms of local transfer free energies such as $\Delta G_{0\to 1,a_1}$ and $\Delta G_{0\to 1,b_1}$ together with the local cell populations normalized to individual basins in explicit solvent (system 1):

$$\begin{aligned}\Delta G_{0\to 1} &= -kT\ \ln\left(\frac{P^A_{0,a_1}}{P^A_{1,a_1}}P_{0,A}e^{-\Delta G_{0\to 1,a_1}/kT} + \frac{P^B_{0,b_1}}{P^B_{1,b_1}}P_{0,B}e^{-\Delta G_{0\to 1,b_1}/kT}\right)\\ &= -kT\ \ln\left(\frac{P_{0,a_1}}{P^A_{1,a_1}}e^{-\Delta G_{0\to 1,a_1}/kT} + \frac{P_{0,b_1}}{P^B_{1,b_1}}e^{-\Delta G_{0\to 1,b_1}/kT}\right)\end{aligned} \tag{9}$$

Here P_{0,a_1} and P_{0,b_1} are the populations of the two cells a_1 and b_1 in macrostate A and macrostate B in system 0 (vacuum or implicit solvent) normalized to the *total population* (i.e., normalized to 1), while P^A_{1,a_1} and P^B_{1,b_1} are the populations of the two cells in system 1 (explicit solvent) normalized within the *individual macrostates*. Equation 9 expresses the total solvation (transfer) free energy as the logarithm of the sum of Boltzmann factors of cell transfer free energies weighed by prefactors that depend both on the relative populations of the macrostates A and B and on the changes in the curvature of the corresponding free energy surfaces with solvent environment. It can be shown that, in one limiting case when the two free energy surfaces in system 0 and system 1 have the identical shape, the localized solvation free energies (e.g., $\Delta G_{0\to 1,a_1}$ and $\Delta G_{0\to 1,b_1}$) for every cell are the same, and the sum of their prefactors i.e. $(P_{0,a_1}/P^A_{1,a_1}) + (P_{0,b_1}/P^B_{1,b_1})$ is equal to one.

Equation 9 can be used to efficiently compute the solvation free energy of a solute molecule with multiple basins (here system 0 corresponds to vacuum). In the standard method of computing solvation free energies, the solute in water is gradually decoupled from the solvent using a number (e.g., ≥ 10) of alchemical λ windows; in each of the λ windows the simulation needs to extensively sample all the cells on the entire free energy surface to yield converged solvation free energies. Using eq 9, the localized decoupling simulations used to obtain quantities $\Delta G_{0\to 1,a_1}$ and $\Delta G_{0\to 1,b_1}$ only need to sample conformations at intermediate alchemical λ windows within the individual cells, which is very fast to converge. While the cell populations P^A_{1,a_1} and P^B_{1,b_1} do require extensive sampling within each of the individual macrostates, these simulations are performed only in the fully coupled states rather than doing so in every alchemical λ window as is done in the standard method. Equation 9 can be easily extended to systems with multiple free energy macrostates.

To verify eq 9, suppose that a solute molecule is divided into N conformational cells. The total transfer free energy can be written as

$$\begin{aligned}\Delta G_{0\to 1} &= -kT\ \ln\frac{Z_1}{Z_0} = -kT\ \ln\frac{\sum_{i=1}^{N}Z_{1,i}}{Z_0}\\ &= -kT\ \ln\frac{\sum_{i=1}^{N}\frac{Z_{0,i}}{Z_{0,i}}Z_{1,i}}{Z_0} = -kT\ \ln\sum_{i=1}^{N}\frac{Z_{0,i}}{Z_0}\frac{Z_{1,i}}{Z_{0,i}}\\ &= -kT\ \ln\sum_{i=1}^{N}P_{0,i}e^{-\Delta G_{0\to 1,i}/kT}\end{aligned} \tag{10}$$

where $Z_{n,i}$ stands for the configurational integral of cell i in system n, $P_{n,i}$ is the population of cell i in system n, $\Delta G_{0\to 1,i}$ is the solvation free energy for cell i. For a molecule with two macrostates A and B, each macrostate consisting of one or more basins (e.g., for alanine dipeptide the $C5/\beta/\alpha_R/C7_{eq}$ macrostate contains four basins and the $\alpha_L/C7_{ax}$ contains two), we have

$$\Delta G_{0\to 1} = -kT\ \ln\left(\sum_{i\in A}^{N_A}P_{0,i}e^{-\Delta G_{0\to 1,i}/kT} + \sum_{j\in B}^{N_B}P_{0,j}e^{-\Delta G_{0\to 1,j}/kT}\right) \tag{11}$$

Now using the results of eqs 4 and 5, we obtain

$$\begin{aligned}\sum_{i\in A}^{N_A}P_{0,i}e^{-\Delta G_{0\to 1,i}/kT} &= P_{0,a_1}e^{-\Delta G_{0\to 1,a_1}/kT}\left(1 + \frac{P_{1,a_2}}{P_{1,a_1}} + \cdots + \frac{P_{1,a_{N_A}}}{P_{1,a_1}}\right)\\ &= P_{0,a_1}e^{-\Delta G_{0\to 1,a_1}/kT}\frac{P_{1,A}}{P_{1,a_1}} = P_{0,a_1}e^{-\Delta G_{0\to 1,a_1}/kT}\frac{1}{P^A_{1,a_1}}\end{aligned} \tag{12}$$

$$\begin{aligned}\sum_{i\in B}^{N_B}P_{0,j}e^{-\Delta G_{0\to 1,j}/kT} &= P_{0,b_1}e^{-\Delta G_{0\to 1,b_1}/kT}\left(1 + \frac{P_{1,b_2}}{P_{1,b_1}} + \cdots + \frac{P_{1,b_{N_B}}}{P_{1,b_1}}\right)\\ &= P_{0,b_1}e^{-\Delta G_{0\to 1,b_1}/kT}\frac{P_{1,B}}{P_{1,b_1}} = P_{0,b_1}e^{-\Delta G_{0\to 1,b_1}/kT}\frac{1}{P^B_{1,b_1}}\end{aligned}$$

Substitute eq 12 into eq 11, we obtain eq 9.

The total solvation (transfer) free energy in eq 9 can also be written as contributions from macrostates (or basins) to yield more insights into solvation. The two prefactors in the right-hand side of eq 9 can be written as $(P_{0,a_1}/P^A_{1,a_1}) = (P^A_{0,a_1}/P^A_{1,a_1})$

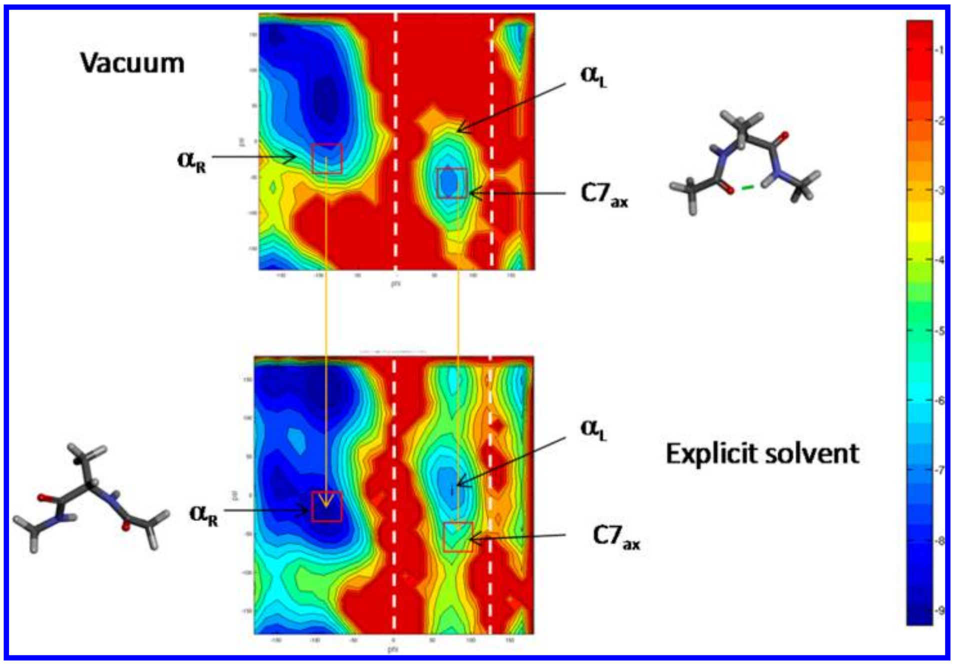


Figure 3. Calculating the free energy difference between α_R and C7$_{ax}$ basins by connecting the free energy surfaces of alanine dipeptide in vacuum and TIP3P water. The intramolecular hydrogen bond in the C7$_{ax}$ conformer is shown in dashed line. Note that the α_L conformer has a very small population in vacuum. The C5/β/α_R/C7$_{eq}$ and α_L/C7$_{ax}$ macrostates are also indicated.

$P_{0,A}$, and $(P_{0,b_1}/P^B_{1,b_1}) = (P^B_{0,b_1}/P^B_{1,b_1})P_{0,B}$. Substituting these into eq 9 yields

$$\begin{aligned}\Delta G_{0\to1} &= -kT\ln\left(\frac{P_{0,a_1}}{P^A_{1,a_1}}e^{-\Delta G_{0\to1,a_1}/kT} + \frac{P_{0,b_1}}{P^B_{1,b_1}}e^{-\Delta G_{0\to1,b_1}/kT}\right)\\ &= -kT\ln\left(\frac{P^A_{0,a_1}}{P^A_{1,a_1}}e^{-\Delta G_{0\to1,a_1}/kT}P_{0,A} + \frac{P^B_{0,b_1}}{P^B_{1,b_1}}e^{-\Delta G_{0\to1,b_1}/kT}P_{0,B}\right)\end{aligned} \quad (13)$$

From, eq 3, we have

$$\begin{aligned}&\frac{P^A_{0,a_1}}{P^A_{1,a_1}}e^{-\Delta G_{0\to1,a_1}/kT} = e^{-\Delta G_{0\to1,A}/kT}\\ &\frac{P^B_{0,b_1}}{P^B_{1,b_1}}e^{-\Delta G_{0\to1,b_1}/kT} = e^{-\Delta G_{0\to1,B}/kT}\end{aligned} \quad (14)$$

Substitute eq 14 into eq 13 yields

$$\Delta G_{0\to1} = -kT\ln(e^{-\Delta G_{0\to1,A}/kT}P_{0,A} + e^{-\Delta G_{0\to1,B}/kT}P_{0,B}) \quad (15)$$

The total solvation (transfer) free energy is therefore written as the logarithm of the sum of population weighted Boltzmann factors of the transfer free energy for each macrostate from system 0 to system 1. For a macrostate on the free energy surface to make a substantial contribution to the total solvation (transfer) free energy, it needs to have a non-negligible population in system 0 (vacuum or implicit solvent) and a favorable macrostate transfer free energy (e.g., $\Delta G_{0\to1,A}$) relative to other macrostates.

In computing the local transfer free energy $\Delta G_{0\to1,a_1}$ using (localized) decoupling simulation methods, the explicitly solvated solute needs to be gradually decoupled from the solvent and simultaneously coupled to the implicit solvent environment. Alternatively, such transformations from explicit to implicit solvent can be done using the solute in vacuum as an intermediate state, i.e.

$$\Delta G_{0\to1,a_1} = \Delta G_{0\to\mathrm{vac},a_1} + \Delta G_{\mathrm{vac}\to1,a_1} \quad (16)$$

Here $\Delta G_{0\to\mathrm{vac},a_1}$ is the local transfer free energy of cell a$_1$ from the implicit solvent to vacuum, which can be done using analytical formulas for the solute structures within cell a$_1$. The second term $\Delta G_{\mathrm{vac}\to1,a_1}$ is the local transfer free energy of cell a$_1$ from vacuum to the explicit solvent, which can be done in a number of ways: (1) local decoupling simulation by restricting the solute conformation to cell a$_1$ or (2) end-point methods such as 3D-RISM,[27,28] again by only considering the solute conformations in cell a$_1$.

We chose alanine dipeptide in water as a model system[29−32] to illustrate the thermodynamic cycle that couples implicit with explicit free energy surfaces. Local decoupling calculations in explicit solvent (TIP3P water model[33]) were performed for the model system at 300 K to estimate transfer free energies for local cells on the free energy surface. The solute is modeled by the OPLS-AA force field.[34,35] In the decoupling calculation a restrained, solvated solute is gradually decoupled from the aqueous solution by turning off the Coulomb interaction first using 11 lambda windows, λ = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0; the Lennard-Jones interactions are then turned off in 17 lambda windows, λ = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.94, 0.985, 1.0. To ensure that the decoupling simulations are localized to the chosen cell on the free energy surface, harmonic dihedral angle restraints are applied to the ϕ and ψ angles. The solvation free energy is determined using thermodynamic integration (TI). The MD sampling at each λ was performed using the GROMACS[36,37] version 4.6.4 for 1 ns. For alanine dipeptide in water, we divide the free energy surface on the ϕ−ψ plane into 120 (along ϕ axis) × 120 (along ψ axis) = 14,400 cells; each of the cells has an area of 3° × 3° (Figure 2). The dimension of the cells is

DOI: 10.1021/acs.jctc.5b00264
J. Chem. Theory Comput. 2015, 11, 2868−2878

Table 1. Computing the Free Energy Differences between Different Basins or Macrostates in Explicit Solvent Using Eq 8, by Connecting the Free Energy Surfaces in Vacuum and in Explicit Solvent[a]

transition A → B	$\Delta G_{1,A\to B}$ exhaustive simulation	$\Delta G_{1,A\to B}$ eq 8	$kT\ln(P^{A}_{0,a_1}/P^{A}_{1,a_1})$	$-\Delta G_{0\to 1,a_1}$	$\Delta G_{0,A\to B}$	$-kT\ln(P^{B}_{0,b_1}/P^{B}_{1,b_1})$	$\Delta G_{0\to 1,b_1}$
$\alpha_R \to C7_{ax}$	4.04 ± 0.17	4.06 ± 0.16	−0.56 ± 0.1	14.4 ± 0.01	−0.34 ± 0.1	−0.04 ± 0.05	−9.4 ± 0.02
$\beta \to C7_{ax}$	4.8 ± 0.17	4.83 ± 0.17	−0.16 ± 0.1	12.23 ± 0.26	2.2 ± 0.01	−0.04 ± 0.05	−9.4 ± 0.02
$C5 \to C7_{ax}$	3.95 ± 0.16	3.91 ± 0.17	0.01 ± 0.01	10.90 ± 0.27	2.44 ± 0.02	−0.04 ± 0.05	−9.4 ± 0.02
$C7_{eq} \to C7_{ax}$	3.15 ± 0.16	3.19 ± 0.17	0.4 ± 0.05	9.23 ± 0.26	3.0 ± 0.01	−0.04 ± 0.05	−9.4 ± 0.02
$C5/\beta/\alpha_R/C7_{eq} \to \alpha_L/C7_{ax}$ a_1 in basin α_R b_1 in basin $C7_{ax}$	2.76 ± 0.23	2.66 ± 0.18	−2.9 ± 0.25	14.4 ± 0.01	2.99 ± 0.1	−2.38 ± 0.07	−9.4 ± 0.02
$C5/\beta/\alpha_R/C7_{eq} \to \alpha_L/C7_{ax}$ a_1 in basin C5 b_1 in basin $C7_{ax}$	2.76 ± 0.23	2.65 ± 0.16	0.54 ± 0.25	10.9 ± 0.02	2.99 ± 0.1	−2.38 ± 0.07	−9.4 ± 0.02

[a] In calculating free energy differences between basins, cells a_1 and b_1 are located in the centers of basins A and B, respectively. For free energy differences between macrostates, the locations of cells a_1 and b_1 are specified in the first column of the corresponding row. Unit: kcal/mol.

chosen to match the range of $\phi-\psi$ angles sampled in the localized decoupling simulations performed under the harmonic dihedral angle restraints. We have experimented with different cell sizes and examined their impact on the accuracy and efficiency of the calculated free energies. Using smaller cells, which is enabled by using stronger $\phi-\psi$ angle restraints, makes the localized decoupling simulations converge even faster, but it also makes the sampling error in the estimated cell populations (e.g., P^{A}_{1,a_1} and P^{B}_{1,b_1}) larger. Conversely, using a larger cell reduces sampling errors in cell populations, at the expense of slowing down the convergence in the decoupling simulations. We find that in the case of the alanine dipeptide example, the chosen cell size of ~3° × 3° yields a satisfactory balance in terms of efficiency.

Simulations of alanine dipeptide in implicit solvent and in vacuum were performed using the IMPACT[38] program with the OPLS-AA force field.[34] Ten and 25 μs MD simulations were run in AGBNP2[39−41] implicit solvent and in vacuum, respectively, to obtain converged free energy estimates in implicit solvent and vacuum environments.

RESULTS

We test our approach described above using alanine dipeptide in water as an example. Figure 2 and Figure 3 show its free energy surfaces and the prominent basins in three environments: TIP3P explicit solvent, AGBNP2 implicit solvent, and vacuum. It can be seen that the differences in the PMFs between the explicit solvent and the implicit solvent are relatively small, while the PMF in the vacuum is qualitatively different from those of the other two, due to the solvation effects on the conformational equilibrium. For example, in vacuum, the $C7_{ax}$ conformer centered at ($\phi = 75°$, $\psi = -58°$) is a free energy minimum due to the N−H···O intramolecular hydrogen bond. In the explicit solvent, however, the $C7_{ax}$ conformer is not located at a minimum (Figure 3). This is because in explicit solvent both the intramolecular hydrogen bonds and the competing solute-water intermolecular hydrogen bonds contribute to the total conformational free energy. While the $C7_{ax}$ conformer features a favorable intramolecular hydrogen bond, it is relative solvation free energy in explicit solvent of −9.4 kcal/mol is less favorable compared with those of other conformers, such as the α_R conformer, which has a relative solvation free energy = −14.4 kcal/mol (see Table 1). Conversely, while the α_R conformer is a minimum in the explicit solvent, it is not a minimum in vacuum (Figure 3). A common feature shared by the three free energy surfaces is that the $C5/\beta/\alpha_R/C7_{eq}$ macrostate ($-180° < \phi < 0°$) is separated from the $\alpha_L/C7_{ax}$ macrostate ($0° < \phi < 120°$) by a significant barrier (The macrostates contain basins that interconvert readily, see Figure 2.). The barrier causes slow equilibration between the two macrostates: the mean first passage time (MFPT) from the $C5/\beta/\alpha_R/C7_{eq}$ macrostate to the $\alpha_L/C7_{ax}$ macrostate is ≈100 ns in explicit solvent, ≈ 178 ns in implicit solvent, and ≈667 ns in vacuum. Although the MFPT is longer in the implicit solvent and in vacuum, the sampling of barrier crossing in these two media is computationally much less costly compared to that in the explicit solvent, because the time per step to do the MD integration in implicit solvent or vacuum is approximately 100 times faster than in explicit solvent (see the Discussion for a comparative analysis of computational times in implicit and explicit solvent)

To efficiently obtain accurate free energy differences in explicit solvent either between basins, or macrostates which are separated by large barriers, we apply the method of connecting implicit/explicit solvent surfaces using eq 8. As shown in Table 1, the free energy differences calculated by connecting the free energy surfaces in vacuum and in explicit solvent agree with those obtained from an exhaustive 4 μs explicit solvent simulation to within 0.2 kcal/mol (out of a free energy difference of ~3−4 kcal/mol). The computational cost of using eq 8 in terms of CPU hours is only 8.3% of that of the exhaustive explicit solvent simulation. Table 1 also lists the values of the different terms of the right-hand side of eq 8. P^{A}_{1,a_1} and P^{B}_{1,b_1} are estimated by running two relatively short explicit solvent simulations (100 ns each) starting from within basin A and B, respectively. P^{A}_{0,a_1}, P^{B}_{1,b_1}, and $\Delta G_{0,A\to B}$ are obtained from direct simulations of alanine dipeptide in vacuum. The two local solvation free energy differences $\Delta G_{0\to 1,a_1}$ and $\Delta G_{0\to 1,b_1}$ are obtained from two restrained decoupling simulations localized to two cells a_1 and b_1 in basins A and B, respectively. Cells a_1 and b_1 can be the centers of the basins, as was done in the top four rows of Table 1. However, they can be any cells within the corresponding basins: in the bottom two rows of Table 1, the free energy difference between the $C5/\beta/\alpha_R/C7_{eq}$ and $\alpha_L/C7_{ax}$ macrostates (Figure 3) are computed using different choices of cell a_1 in the $C5/\beta/\alpha_R/C7_{eq}$ macrostate and almost identical values of $\Delta G_{1,A\to B}$ were obtained.

Table 2. Computing the Free Energy Differences between Different Basins or Macrostates in Explicit Solvent Using Eq 8, by Connecting the Free Energy Surfaces in Implicit and Explicit Solvent[a]

transition A → B	$\Delta G_{1,A\rightarrow B}$ exhaustive simulation	$\Delta G_{1,A\rightarrow B}$ eq 8	$kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$	$-\Delta G_{0\rightarrow 1,a_1}$	$\Delta G_{0,A\rightarrow B}$	$-kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$	$\Delta G_{0\rightarrow 1,b_1}$
$\alpha_R \rightarrow C7_{ax}$	4.04 ± 0.17	4.05 ± 0.06	−0.04 ± 0.04	2.25 ± 0.02	3.33 ± 0.01	0.24 ± 0.09	−1.73 ± 0.09
$\alpha_R \rightarrow \alpha_L$	1.75 ± 0.09	1.70 ± 0.08	−0.04 ± 0.04	2.25 ± 0.02	1.87 ± 0.01	0.1 ± 0.03	−2.48 ± 0.1
$\beta \rightarrow \alpha_L$	2.51 ± 0.09	2.47 ± 0.08	−0.02 ± 0.01	2.65 ± 0.03	2.20 ± 0.02	0.1 ± 0.03	−2.48 ± 0.1
$\beta \rightarrow C7_{ax}$	4.80 ± 0.17	4.78 ± 0.05	−0.02 ± 0.01	2.65 ± 0.03	3.64 ± 0.02	0.24 ± 0.09	−1.73 ± 0.09
$C5 \rightarrow \alpha_L$	1.64 ± 0.10	1.77 ± 0.12	0.01 ± 0.004	2.39 ± 0.11	1.77 ± 0.02	0.1 ± 0.03	−2.48 ± 0.1
$C5/\beta/\alpha_R/C7_{eq} \rightarrow \alpha_L/C7_{ax}$ a_1 in basin α_R b_1 in basin $C7_{ax}$	2.76 ± 0.23	2.71 ± 0.06	0.1 ± 0.03	2.25 ± 0.02	2.8 ± 0.02	−0.71 ± 0.09	−1.73

[a] In calculating free energy differences between basins, cells a_1 and b_1 are located in the centers of basins A and B, respectively. For free energy differences between macrostates, the locations of cells a_1 and b_1 are specified in the first column of the corresponding row. Unit: kcal/mol.

Table 3a. Transfer Free Energy of Alanine Dipeptide from Vacuum to Explicit Solvent Using Eq 9[a]

$P_{0,a_1}/P^A_{1,a_1}$	$\Delta G_{0\rightarrow 1,a_1}$	$P_{0,b_1}/P^B_{1,b_1}$	$\Delta G_{0\rightarrow 1,b_1}$	$\Delta G_{0\rightarrow 1}$ eq 9	$\Delta G_{0\rightarrow 1}$ exhaustive simulation
0.0067 ± 0.0005	−14.4 ± 0.01	0.36 ± 0.02	−9.4 ± 0.02	−11.42 ± 0.05	−11.68 ± 0.12

[a] A is the $C5/\beta/\alpha_R/C7_{eq}$ macrostate, and B is the $\alpha_L/C7_{ax}$ macrostate; a_1 is a cell in basin α_R, and b_1 is a cell in $C7_{ax}$. Unit: kcal/mol.

In several cases in Table 1, the net transfer (solvation) free energy difference $-\Delta G_{0\rightarrow 1,a_1} + \Delta G_{0\rightarrow 1,b_1}$ between the two cells, and the free energy difference between the basins or macrostates in vacuum, $\Delta G_{0,A\rightarrow B}$, make the dominant contributions to the free energy difference between the basins or macrostates $\Delta G_{1,A\rightarrow B}$ in explicit solvent, while the surface curvature related terms $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$ and/or $-kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$ are small. One such example is the free energy difference between the C5 and $C7_{ax}$ basins $C5 \rightarrow C7_{ax}$; here the center of C5 used as cell a_1 is at the minimum of the basin in both vacuum and explicit solvent; therefore, the $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$ term is very small (Table 1). On the other hand, when the two free energy basins have different curvatures, the curvature terms can make a large contribution to $\Delta G_{1,A\rightarrow B}$. This occurs for example when calculating the free energy difference between the $C5/\beta/\alpha_R/C7_{eq}$ and $\alpha_L/C7_{ax}$ macrostates using the cell at the center of α_R basin as a_1: here, α_R is at a minimum in explicit solvent but not at a minimum in vacuum, hence the corresponding curvature term $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1}) = -2.9$ kcal/mol is large.

It should be noted that the magnitudes of the surface curvature terms $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$ and/or $-kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$ depend on the curvature differences in the entire basin or macrostate, not just near the cells a_1 and b_1. For example, in the top four rows of Table 1, $-kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$ is small (= 0.04 kcal/mol); but in the bottom two rows $-kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$ is large (= −2.38 kcal/mol), even though the same cell b_1, i.e. the center of basin $C7_{ax}$, is used in all the rows in Table 1. The reason is the following: in the top four rows of Table 1, region B corresponds to the $C7_{ax}$ basin, which occupies a small area within the $\alpha_L/C7_{ax}$ macrostate on the free energy surface. Within the $C7_{ax}$ basin the free energy surfaces are quite flat in both vacuum and explicit solvent, giving rise to a small curvature term $-kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$. On the other hand, in the bottom two rows, region B involves the entire $\alpha_L/C7_{ax}$ macrostate, which covers a much larger area on the free energy surface, where the differences in the curvatures between the vacuum and explicit solvent surfaces are substantial (Figure 3). Therefore, the surface curvature term $-kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$ is large.

Next we use implicit solvent as the reference system 0 in eq 8 and connect its free energy surface to that of the explicit solvent system 1 to compute the free energy differences in the explicit solvent. Since the implicit solvent surface is more similar to that of the explicit solvent than the vacuum surface, it facilitates the estimates of free energy differences involving more conformers in the $\alpha_L/C7_{ax}$ macrostate, e.g. the α_L basin which has very small populations in vacuum. Table 2 shows the results for the free energy change between different pairs of basins. The free energies computed using eq 8 are within 0.2 kcal/mol from the corresponding values obtained using exhaustive explicit solvent simulations. As can be seen from Table 2, although the implicit solvent estimated free energy differences $\Delta G_{0,A\rightarrow B}$ are already comparable to those from the exhaustive explicit solvent simulations, using the thermodynamic cycle of eq 8 leads to improvement over the estimates based only on the implicit solvent free energy surface. While the improvement mostly comes from the net transfer free energy change $(-\Delta G_{0\rightarrow 1,a_1} + \Delta G_{0\rightarrow 1,b_1})$, for the free energy difference between the two macrostates $C5/\beta/\alpha_R/C7_{eq} \rightarrow \alpha_L/C7_{ax}$, the surface curvature term $kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$ also contributes −0.71 kcal/mol (Table 2). It is noted that the surface curvature related terms $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$ in Table 2 are all very small. This is because in Table 2 the different basins A are located in the $C5/\beta/\alpha_R/C7_{eq}$ macrostate of the free energy surfaces, where the curvatures in both explicit solvent and implicit solvent are quite similar: see Figure 2. The other surface curvature terms $kT\ \ln(P^B_{0,b_1}/P^B_{1,b_1})$, which are associated with individual cells in basin B, while small are still an order of magnitude larger than the corresponding term for the $C5/\beta/\alpha_R/C7_{eq}$ macrostate. This is consistent with the larger differences in the curvatures of the $\alpha_L/C7_{ax}$ macrostates in the two solvents (implicit and explicit). Note that when a surface curvature term $kT\ \ln(P^A_{0,a_1}/P^A_{1,a_1})$ (associated with basin or macrostate A) is small, it also means that the transfer free energy is only weakly dependent on conformation for structures within that basin or macrostate.

DOI: 10.1021/acs.jctc.5b00264
J. Chem. Theory Comput. 2015, 11, 2868−2878

Table 3b. Transfer Free Energy of Alanine Dipeptide from AGBNP2 Implicit Solvent to Explicit Solvent Using Eq 9[a]

$P_{0,a_1}/P^A_{1,a_1}$	$\Delta G_{0\to1,a_1}$	$P_{0,b_1}/P^B_{1,b_1}$	$\Delta G_{0\to1,b_1}$	$\Delta G_{0\to1}$ eq 9	$\Delta G_{0\to1}$ exhaustive simulation
1.39 ± 0.12	−2.39 ± 0.11	0.023 ± 0.004	−2.48 ± 0.1	−2.60 ± 0.1	−2.75 ± 0.12

[a] A is the C5/β/α_R/C7$_{eq}$ macrostate, and B is the α_L/C7$_{ax}$ macrostate; a$_1$ is a cell in basin C5, and b$_1$ is a cell in α_L. Unit: kcal/mol.

Lastly, we test the accuracy and efficiency of using eq 9 to compute the solvation free energy of alanine dipeptide in explicit solvent (Table 3a) and the transfer free energy from implicit solvent to explicit solvent (Table 3b) by connecting free energy surfaces. Here the benchmark for the solvation free energy estimates is obtained using standard TI methods in which 200 ns unrestrained decoupling simulations are performed for each of the 28 alchemical λ windows; the total simulation time exceeds 5 μs, which requires ~9,000 CPU hours. In contrast, using the approach of eq 9, it takes ~7% of the computational cost (in CPU hours) to obtain the solvation free energy $\Delta G_{0\to1}$ = −11.42 kcal/mol, compared with the benchmark $\Delta G_{0\to1}$(exhaustive) = −11.68 kcal/mol. Here the two localized decoupling simulations associated with each of the two free energy macrostates (C5/β/α_R/C7$_{eq}$ and α_L/C7$_{ax}$) converge very rapidly, e.g. within <1 ns per alchemical λ window. Table 3b shows the transfer free energy of alanine dipeptide from AGBNP2 implicit solvent to explicit solvent obtained using the approach of eq 9. This quantity, which equals the difference in the solvation free energies obtained using implicit solvent and explicit solvent (i.e., here $\Delta G_{0\to1}$ = $\Delta G_{\text{vac}\to\text{explicit}} - \Delta G_{\text{vac}\to\text{implicit}}$) thus provides a measure of the quality of the implicit solvent model; for a perfect implicit solvent model, the transfer free energy $\Delta G_{0\to1}$ from implicit to explicit solvent should be zero. The result obtained using eq 9 is in close agreement with the benchmark value obtained from direct exhaustive simulations; the value of $\Delta G_{0\to1}$ suggests that the AGBNP2 implicit solvent model underestimates the overall solvation free energy by ≈−2.6 kcal/mol (out of the total solvation free energy of −11.68 kcal/mol. See Tables 3a and 3b.).

We also rewrite the total solvation (transfer) free energies using eq 15 to estimate the contributions from each of the macrostates of alanine dipeptide to the total solvation free energy (data not shown). While the free energies of transferring each of the two macrostates ($\Delta G_{0\to1,A}$ and $\Delta G_{0\to1,B}$) have the same order of magnitudes, macrostate C5/β/α_R/C7$_{eq}$ has a much larger population in system 0 than macrostate α_L/C7$_{ax}$. Therefore, for alanine dipeptide the total solvation (transfer) free energies are dominated by the contribution from the C5/β/α_R/C7$_{eq}$ macrostate.

■ DISCUSSION

Despite advances in computer hardware, calculating accurate conformational free energy differences in explicit solvent using direct MD simulations is computationally costly, especially on rugged free energy surfaces with large barriers between populated minima. For alanine dipeptide in water, an accurate calculation of the free energy difference between any two basins, one in each of the two macrostates, using direct simulation requires the simulation time to be much longer than the slowest relaxation time of the system, such that a sufficient number of reversible transitions between the basins are sampled in the simulation. We estimated that roughly ≈40 reversible transitions are needed to achieve an error of ≤0.3 kcal/mol in the calculated free energy differences between the basins in the C5/β/α_R/C7$_{eq}$ macrostate and those in the α_L/C7$_{ax}$ macrostate. This translates into ~4 μs of simulation time, which takes about 6400 CPU hours in explicit solvent (the solvated alanine dipeptide contains ~1,700 atoms; the MD throughput on a single CPU is ~0.62 ns/h). In contrast, achieving the same amount of reversible transitions across the basins (which corresponds to a simulation time of 40 × 178 ns = 7.12 μs; the mean first passage time in implicit solvent is 178 ns) requires only 107 CPU hours in implicit solvent (the implicitly solvated alanine dipeptide contains 22 atoms, and the MD throughput on a single CPU is 66.7 ns/h). Therefore, using the new approach, calculating the barrier crossing in implicit solvent, and then connecting the free energy surfaces in implicit solvent and explicit solvent, the same accuracy in the free energy differences can be achieved using ~8% of the CPU hours used in the direct MD simulation in explicit solvent only. As mentioned earlier, although the mean first passage time (MFPTs) to transit between macrostates in the implicit solvent and vacuum are longer (178 and 667 ns, respectively) than that in the explicit solvent (100 ns) due to the differences in the barrier heights in these media and friction effects, the sampling of the full free energy surfaces is still computationally much less costly in implicit solvent or vacuum because of the much faster computation of each MD step in these two media.

We also compared the computational efficiency of the new approach of connecting the free energy surfaces using eq 8 with one of the most widely used enhanced sampling methods, umbrella sampling. The full free energy surface in explicit solvent can be computed using 24 × 24 = 576 umbrella windows on the ϕ−ψ surface, i.e. each window is a 15° × 15° cell. For each cell, 2 ns MD simulation is performed to collect biased distribution data, resulting in a total simulation time of ~1 μs in explicit solvent. The total computing time is ~2000 CPU hours. This is about 3.5 times the CPU time used by our approach of connecting the free energy surfaces of implicit and explicit solvent.

The approach expressed in eq 8 contains three elements: (1) The computationally costly estimates of free energy difference between basins or macrostates in explicit solvent is replaced by simulations in an implicit solvent. (2) The correction to the implicit solvent result is computed using two local decoupling simulations focusing on two cells, one located in each of the basins or macrostates of interest. (3) Two local explicit solvent simulations are needed to estimate the populations of the cells within each of the basins or macrostates in explicit solvent, to evaluate the differences in the surface curvature terms in eq 8. All three parts are fast: the simulation in implicit solvent is much faster than its explicit solvent counterpart both in terms of CPU hours and wall clock time; the two local decoupling simulations are also fast since extensive solute conformational sampling is not needed because the solute is restrained to the cells on the free energy surface during the alchemical transformations. Part (3) above takes more computing time than parts (1) and (2) but is still only a fraction of that in the direct simulation.

While the calculated free energy differences using eq 8 are in principle invariant to the locations of the cells a$_1$ and b$_1$ within basins/macrostates, their choice can affect the curvature terms and local transfer free energies and affect the convergence of calculations of these terms (see for example Table 1, bottom two rows). In general, the cells should be chosen to correspond to the centers of the free energy basin: their larger populations will make the estimation of quantities like P^A_{1,a_1} and P^B_{1,b_1} converge faster. Note that when the free energy surfaces of the implicit solvent and the explicit solvent differ significantly, a cell which is at the local minimum in one surface can be far from local minima in the other surface. In this case, for the surface where the cell is far from the local minimum, the estimation of the cell population will converge slowly. Since the sampling in explicit solvent is generally much slower than that in implicit solvent, the cells should be chosen to be near the expected local free energy minima in the explicit solvent surface. Note that in practice, however, the precise location of the free energy minima in explicit solvent may not be known a priori.

In our approach of connecting the free energy surfaces, the quality of the implicit solvent model will affect the efficiency of the calculation. The closer the implicit and explicit solvent surfaces are to each other, the easier it is to obtain the converged estimates of the localized transfer free energy estimates of the cells between implicit and explicit solvent. Conversely, a poor implicit solvent model will make it more costly to converge the localized transfer free energies. There is also the related issue of what cells to choose to do the restrained decoupling calculations, as discussed above. Taken together, using advanced implicit solvent model may lead to more efficient calculation of the free energy differences using our method, but the computational complexity of the implicit solvent model needs to be taken into account.[42,43]

Our method of connecting free energy surfaces also yields rapid, accurate estimates of solvation free energies. As shown by Mobley et al.[44] for flexible solutes sufficient sampling of the conformational degrees of freedom of the solute is crucial for calculating solvation free energies (transfer free energies from vacuum to solvent), especially in the case when the PMF in the gas phase and in the solvent differ significantly. In the standard TI calculation of solvation free energy, the entire free energy surface needs to be sampled throughout the decoupling process which converges very slowly. In contrast, in our approach eq 9, the decoupling simulations are confined to a few localized cells, for which the decoupling simulations converge rapidly. As we have shown in the model system alanine dipeptide, accurate solvation free energies can be obtained using ∼7% of the CPU time required for the standard TI calculation performed entirely in explicit solvent.

We note that in the current implementation of the thermodynamic cycle approach eq 8, only two cell transfer free energies are computed in order to reconstruct a free energy surface in explicit solvent containing two macrostates (each of which contains several basins). When additional cell transfer free energies are calculated, the estimated transfer free energy values are coupled through the curvature of the free energy surfaces. This "redundancy" opens the possibility of using WHAM (Weighted Histogram Analysis Method) methods to improve the accuracy and efficiency of the method by enforcing self-consistency. We will report on this in a future communication.

We have presented a method for computing conformational free energy differences and solvation (transfer) free energies in explicit solvent using a thermodynamic cycle that connects the implicit/explicit free energy surfaces. In this approach the computational efficiency of the implicit solvent model and the accuracy of the explicit solvent simulation are combined to yield accurate estimates of free energy differences in explicit solvent without direct sampling of the full free energy surfaces in explicit solvent. Using alanine peptide in water as a model system, we have shown that accurate conformational free energy differences and solvation free energies are obtained to within 0.3 kcal/mol from the corresponding results from exhaustive explicit solvent simulations using ∼8% of the computational cost of the direct simulations in explicit solvent. This method has the potential to be applied to determine free energy changes associated with biomolecular binding and conformational transitions in complex systems.

It is recognized that certain challenges need to be resolved in applying our approach of connecting free energy surfaces for larger, more complex solutes. First, the appropriate reaction coordinates associated with the slowest relaxation processes of the system need to be identified in order to characterize the free energy surfaces. The definition of basins/macrostates should in principle correspond to the collection of metastable states which reflect the biological function relevant to the system, i.e. the functionally important conformations.[45] Second, for large solutes, the local cell decoupling simulations used in eq 8 become more difficult. At some point, other end-point methods for estimating solvation free energies may need to be used in place of the cell decoupling simulations. Among these methods are 3D-RISM,[46,47]inhomogeneous fluid solvation theories,[48] GIST,[25] and solution theory in the energy representation.[49] As already mentioned, the possibility of using WHAM methods to solve for the surface curvatures and coupling free energies self-consistently can also be employed when the coupling free energies are estimated by these and other end-point methods.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: nanjie.deng@gmail.com (N.D.).
*E-mail: ronlevy@temple.edu (R.M.L.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work has been supported by the R01 GM30580 to R.M.L. The free energy calculations were performed using NSF XSEDE computing resources.

■ REFERENCES

(1) Simonson, T. Free Energy Calculations. In *computational biochemistry and biophysics*; Marcel Dekker: New York, 2000.

(2) *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Study ed.; Chipot, C., Pohorille, A., Eds.; Springer series in chemical physics; Springer: New York, 2007.

(3) Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10* (7), 2632−2647.

(4) Hansmann, U. H. E. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.* **1997**, *281* (1−3), 140−150.

(5) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141−151.

(6) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562−12566.

(7) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120* (24), 11919.

(8) Bartels, C.; Karplus, M. Multidimensional Adaptive Umbrella Sampling: Applications to Main Chain and Side Chain Peptide Conformations. *J. Comput. Chem.* **1997**, *18* (12), 1450−1462.

(9) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition Path Sampling and the Calculation of Rate Constants. *J. Chem. Phys.* **1998**, *108* (5), 1964.

(10) Faradjian, A. K.; Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *J. Chem. Phys.* **2004**, *120* (23), 10880.

(11) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131* (12), 124101.

(12) Andrec, M.; Andrec, M.; Gallicchio, E.; Levy, R. M. Protein Folding Pathways from Replica Exchange Simulations and a Kinetic Network Model. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (19), 6801−6806.

(13) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways from Short off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (45), 19011−19016.

(14) Deng, N.; Dai, W.; Levy, R. M. How Kinetics within the Unfolded State Affects Protein Folding: An Analysis Based on Markov State Models and an Ultra-Long MD Trajectory. *J. Phys. Chem. B* **2013**, *117* (42), 12787−12799.

(15) Wang, L.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* **2011**, *115* (30), 9431−9438.

(16) Wang, L.; Berne, B. J.; Friesner, R. A. On Achieving High Accuracy and Reliability in the Calculation of Relative Protein-Ligand Binding Affinities. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (6), 1937−1942.

(17) Roux, B.; Simonson, T. Implicit Solvent Models. *Biophys. Chem.* **1999**, *78* (1−2), 1−20.

(18) Feig, M.; Brooks, C. L. Recent Advances in the Development and Application of Implicit Solvent Models in Biomolecule Simulations. *Curr. Opin. Struct. Biol.* **2004**, *14* (2), 217−224.

(19) Tan, C.; Yang, L.; Luo, R. How Well Does Poisson−Boltzmann Implicit Solvent Agree with Explicit Solvent? A Quantitative Analysis. *J. Phys. Chem. B* **2006**, *110* (37), 18680−18687.

(20) Chen, J.; Brooks, C. L.; Khandogin, J. Recent Advances in Implicit Solvent-Based Methods for Biomolecular Simulations. *Curr. Opin. Struct. Biol.* **2008**, *18* (2), 140−148.

(21) Aguilar, B.; Shadrach, R.; Onufriev, A. V. Reducing the Secondary Structure Bias in the Generalized Born Model via R6 Effective Radii. *J. Chem. Theory Comput.* **2010**, *6* (12), 3613−3630.

(22) Gallicchio, E.; Lapelosa, M.; Levy, R. M. Binding Energy Distribution Analysis Method (BEDAM) for Estimation of Protein−Ligand Binding Affinities. *J. Chem. Theory Comput.* **2010**, *6* (9), 2961−2977.

(23) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein-Ligand Binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (3), 808−813.

(24) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, *130* (9), 2817−2831.

(25) Nguyen, C. N.; Kurtzman Young, T.; Gilson, M. K. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor cucurbit[7]uril. *J. Chem. Phys.* **2012**, *137* (4), 044101.

(26) Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T. Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. *J. Chem. Theory Comput.* **2014**, *10* (7), 2769−2780.

(27) Hirata, F.; Redfern, P.; Levy, R. M. Viewing the Born Model for Ion Hydration through a Microscope. *Int. J. Quantum Chem.* **1988**, *34* (S15), 179−190.

(28) *Molecular Theory of Solvation*; Hirata, F., Ed.; Understanding chemical reactivity; Kluwer Academic Publishers: Dordrecht; Boston, 2003.

(29) Tobias, D. J.; Brooks, C. L. Conformational Equilibrium in the Alanine Dipeptide in the Gas Phase and Aqueous Solution: A Comparison of Theoretical Results. *J. Phys. Chem.* **1992**, *96* (9), 3864−3870.

(30) Vargas, R.; Garza, J.; Hay, B. P.; Dixon, D. A. Conformational Study of the Alanine Dipeptide at the MP2 and DFT Levels. *J. Phys. Chem. A* **2002**, *106* (13), 3213−3218.

(31) Chekmarev, D. S.; Ishida, T.; Levy, R. M. Long-Time Conformational Transitions of Alanine Dipeptide in Aqueous Solution: Continuous and Discrete-State Kinetic Models. *J. Phys. Chem. B* **2004**, *108* (50), 19487−19495.

(32) Miao, Y.; Sinko, W.; Pierce, L.; Bucher, D.; Walker, R. C.; McCammon, J. A. Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *J. Chem. Theory Comput.* **2014**, *10* (7), 2677−2689.

(33) 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* (2), 926.

(34) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225−11236.

(35) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474−6487.

(36) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435−447.

(37) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29* (7), 845−854.

(38) Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **2005**, *26* (16), 1752−1780.

(39) Gallicchio, E.; Paris, K.; Levy, R. M. The AGBNP2 Implicit Solvation Model. *J. Chem. Theory Comput.* **2009**, *5* (9), 2544−2564.

(40) Gallicchio, E.; Levy, R. M. AGBNP: An Analytic Implicit Solvent Model Suitable for Molecular Dynamics Simulations and High-Resolution Modeling. *J. Comput. Chem.* **2004**, *25* (4), 479−499.

(41) Felts, A. K.; Gallicchio, E.; Chekmarev, D.; Paris, K. A.; Friesner, R. A.; Levy, R. M. Prediction of Protein Loop Conformations Using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. *J. Chem. Theory Comput.* **2008**, *4* (5), 855−868.

(42) Dzubiella, J.; Swanson, J. M. J.; McCammon, J. A. Coupling Hydrophobicity, Dispersion, and Electrostatics in Continuum Solvent Models. *Phys. Rev. Lett.* **2006**, *96* (8).

(43) Zhou, S.; Cheng, L.-T.; Dzubiella, J.; Li, B.; McCammon, J. A. Variational Implicit Solvation with Poisson−Boltzmann Theory. *J. Chem. Theory Comput.* **2014**, *10* (4), 1454−1467.

(44) Mobley, D. L.; Dill, K. A.; Chodera, J. D. Treating Entropy and Conformational Changes in Implicit Solvent Simulations of Small Molecules. *J. Phys. Chem. B* **2008**, *112* (3), 938−946.

(45) Deng, N.; Zheng, W.; Gallicchio, E.; Levy, R. M. Insights into the Dynamics of HIV-1 Protease: A Kinetic Network Model Constructed from Atomistic Simulations. *J. Am. Chem. Soc.* **2011**, *133* (24), 9387−9394.

(46) Kinoshita, M.; Okamoto, Y.; Hirata, F. Calculation of Solvation Free Energy Using RISM Theory for Peptide in Salt Solution. *J. Comput. Chem.* **1998**, *19* (15), 1724−1735.

(47) Truchon, J.-F.; Pettitt, B. M.; Labute, P. A Cavity Corrected 3D-RISM Functional for Accurate Solvation Free Energies. *J. Chem. Theory Comput.* **2014**, *10* (3), 934−941.

(48) Huggins, D. J.; Payne, M. C. Assessing the Accuracy of Inhomogeneous Fluid Solvation Theory in Predicting Hydration Free Energies of Simple Solutes. *J. Phys. Chem. B* **2013**, *117* (27), 8232−8244.

(49) Takemura, K.; Guo, H.; Sakuraba, S.; Matubayasi, N.; Kitao, A. Evaluation of Protein-Protein Docking Model Structures Using All-Atom Molecular Dynamics Simulations Combined with the Solution Theory in the Energy Representation. *J. Chem. Phys.* **2012**, *137* (21), 215105.