JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Effects of Biomolecular Flexibility on Alchemical Calculations of Absolute Binding Free Energies

Morgan Lawrenz,[*,†] Riccardo Baron,[*,†] Yi Wang,[†] and J. Andrew McCammon[†,‡]

[†]Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics, and [‡]Department of Pharmacology, Howard Hughes Medical Institute, University of California, San Diego, La Jolla, California 92093, United States

Supporting Information

ABSTRACT: The independent trajectory thermodynamic integration (IT-TI) approach (Lawrenz, M., et al. *J. Chem. Theory. Comput.* **2009**, *5*, 1106−1116) for free energy calculations with distributed computing is employed to study two distinct cases of protein−ligand binding: first, the influenza surface protein N1 neuraminidase bound to the inhibitor oseltamivir, and second, the *Mycobacterium tuberculosis* enzyme RmlC complexed with the molecule CID 77074. For both systems, finite molecular dynamics (MD) sampling and varied molecular flexibility give rise to IT-TI free energy distributions that are remarkably centered on the target experimental values, with a spread directly related to protein, ligand, and solvent dynamics. Using over 2 μs of total MD simulation, alternative protocols for the practical, general implementation of IT-TI are investigated, including the optimal use of distributed computing, the total number of alchemical intermediates, and the procedure to perturb electrostatics and van der Waals interactions. A protocol that maximizes predictive power and computational efficiency is proposed. IT-TI outperforms traditional TI predictions and allows a straightforward evaluation of the reliability of free energy estimates. Our study has broad implications for the use of distributed computing in free energy calculations of macromolecular systems.

1. INTRODUCTION

Alchemical free energy methods often employ molecular dynamics (MD) simulations of unphysical intermediate microstates in order to calculate the free energy difference between two physically relevant canonical ensembles. Examples are the relative binding free energy difference between different ligands to a receptor or the free energy change upon transferral of a ligand and protein from the unbound to the bound state. The latter is often referred to as the *absolute binding free energy* described by the thermodynamic cycle in Scheme 1.[1−7] Although MD-based free energy calculations rely on rigorous statistical mechanics principles,[5,6,8,9] their practical application is still challenging for systems with numerous degrees of freedom. MD sampling may be trapped in confined regions of conformational space due to the frustrated nature of protein and ligand energy landscapes, thus leading to insufficient statistics.

The use of independent MD simulations recently proved to be an appealing strategy to alleviate this issue, particularly with the rapid and steady increase of computational power in the form of multiple CPU and GPU clusters. This approach was applied to a number of systems in different flavors. Fujitani et al. employed multiple free energy perturbation (FEP) calculations to estimate absolute free energies of FKBP ligand binding.[10] Zagrovic and van Gunsteren used multiple one-step perturbation runs to calculate relative free energies of PDE5 ligand binding.[11] In Mobley et al. and Boyce et al., multiple FEP calculations for different docked ligand binding poses were used to predict relative and absolute binding free energies for ligands to engineered binding sites of T4 lysozyme.[12,13] Lawrenz et al. employed independent trajectory thermodynamic integration (IT-TI) to obtain accurate absolute free energies for peramivir binding to N1 neuraminidase, as well as relative binding free energies of alchemically modified compounds.[14] The latter study also emphasized the importance of solvent effects in this context. Accurate free energies are needed for all states of the thermodynamic cycle of interest (see Scheme 1) to achieve high predictive power, as realized since the pioneering applications of alchemical approaches.[2,3,15,16] Here, we use IT-TI to compute absolute binding free energies for two ligands to two protein drug targets with different active site structural and chemical properties.

First, we consider the influenza surface protein N1 neuraminidase binding to oseltamivir.[17] N1 neuraminidase facilitates viral shedding from infected cells and is a key target for treatment of pandemic flu. The N1 active site is composed of flexible loops[14,18] and is highly solvent exposed (see Figure 1a,c). The ligand oseltamivir has zero net charge but contains one ammonium group and one carboxyl group (Figure 1e); the latter forms salt bridges with the arginine triad binding motif (R118, R292, and R371 in Figure 1c).[17] Electrostatic interactions have been characterized as the dominant contribution to ligand binding.[14,19] Oseltamivir is flexible due to 10 nonstericallly hindered rotatable bonds, including a branched aliphatic tail that occupies a hydrophobic subpocket.[14,20]

Second, we study the *Mycobacterium tuberculosis* (*M. tuberculosis*) enzyme dTDP-6-deoxy-*d*-xylo-4-hexopyranosid-4-ulose 3,5-epimerase (RmlC), which is crucial for assembly of the mycobacterial waxy, impermeable cell wall and is a viable drug target.[21] In this case, the bound ligand, compound identifier (CID) 77074, was a top hit from virtual screening, followed by experimental validation.[21] The RmlC binding site is organized into β-sheets and is smaller and

Received: April 4, 2011
Published: June 02, 2011

narrower than that in N1 (compare Figure 1a,c and Figure 1b,d). Aromatic residues Y138, F26, and H119 stack against the ligand aromatic rings (see Figure 1d). The ligand itself contains seven rotatable bonds, with limited flexibility due to the presence of two bulky ring groups (Figure 1f).

We investigate to which extent protein, ligand, and solvent dynamics influence the reliability of absolute binding free energies computed with TI. Using IT-TI, we see that finite sampling and varied molecular flexibility of the two investigated protein−ligand systems give rise to distributions of free energy estimates. This observation is in line with previous suggestions for N1 neuraminidase based on more reduced statistics.[14] We show that the features of these distributions—while remarkably centered around the target experimental values—are linked to protein, ligand, and solvent dynamic sampling. Additionally, we use statistics from over 2 μs of overall IT-TI simulation time to compare different approaches for optimal distributed computing and alternative protocols for the practical application of TI. We suggest a protocol that is optimal for two systems with different dynamic properties. Future work will investigate whether this protocol might be optimal for protein−ligand binding in general.

2. MATERIALS AND METHODS

2.1. Molecular Models and Simulations. Initial coordinates were available for N1 bound to the ligand oseltamivir on the basis of X-ray crystallography experiments (PDB: 2HU0).[17] For RmlC, initial coordinates for its complex with CID 77074, or 1-(3-(5-allyl-5*H*-[1,2,4]triazino[5,6-*b*]indol-3-ylthio)propyl)-1*H*-benzo[*d*]imidazol-2(3*H*)-one, were based on the unbound X-ray structure (PDB: 2IXC) and an experimentally verified computational docking pose.[21] A monomer of the natively tetrameric protein N1 was simulated, as in previous studies,[14] while the RmlC protein was simulated as a dimer, for half the N1 simulation time, because its active site spans the interface between two monomers (see Figure 1). Thus, RmlC analyses were performed by concatenating two monomer trajectories for identical overall sampling times for each system. See Table 1 for a summary of MD sampling periods.

Molecular models were based on the AMBER FF99SB force field[22] and the compatible TIP3P model for water.[23] The cubic simulation box contained N1, 15 126, and RmlC, 24 305 water molecules, added to the system using AmberTools. Both systems were neutralized with (N1, 1, or RmlC. 24) Na^+ counterions with AMBER rescaled parameters.[24] The importance of a protein-bound Ca^{2+} ion in N1 ligand binding was recently highlighted.[25] Ligands were parametrized using the generalized Amber force field (GAFF)[26] parameters for angles, bonds, and torsions, and RESP[27] fitting of Gaussian03[28] calculated electrostatic potentials at the Hartree−Fock/6-31G* level. All simulations were performed using the NAMD software[29] (version 2.7b1). A 2 fs time step was employed, with hydrogen-containing protein bonds constrained using RATTLE[30] and water geometries constrained using SETTLE.[31] The particle mesh Ewald (PME) approximation[32] (1 $Å^{-3}$ grid density) was employed for electrostatics. Short-range nonbonded interactions were evaluated every 2 fs and long-range electrostatics every 4 fs (nonbonded interaction cutoff, 12 Å; switching distance, 10 Å).[29] After incremental heating to 300 K, the system was equilibrated for 2 ns in the *N*, *p*, *T* ensemble with Langevin pressure and temperature controls[33] before each *N*, *V*, *T* independent simulation was initialized with a random velocity.

2.2. Free Energy Calculations. Free energy changes along the thermodynamic cycle in Scheme 1 were evaluated using thermodynamic integration (TI) as[34]

$$\Delta F_{0\to 1} = \int_0^1 d\lambda \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda \qquad (1)$$

where in this study $\Delta F_{0\to 1}$ is either the $\Delta F^\circ_{protein}$ or ΔF°_{water} standard Helmoltz free energy in Scheme 1 and U is the total potential energy of the system. The ligand is decoupled from

Scheme 1. Thermodynamic Cycle Underlying Alchemical Absolute Binding Free Energy Calculations

$$\begin{array}{ccc} PL_{(water)} & \rightarrow & P_{(water)} + L_{(gas)} \quad \Delta F^\circ_{protein} \\ L_{(water)} & \rightarrow & L_{(gas)} \quad \Delta F^\circ_{water} \\ P_{(water)} + L_{(water)} & \rightarrow & PL_{(water)} \quad \Delta F^\circ_{bind} = \Delta F^\circ_{water} - \Delta F^\circ_{protein} \end{array}$$

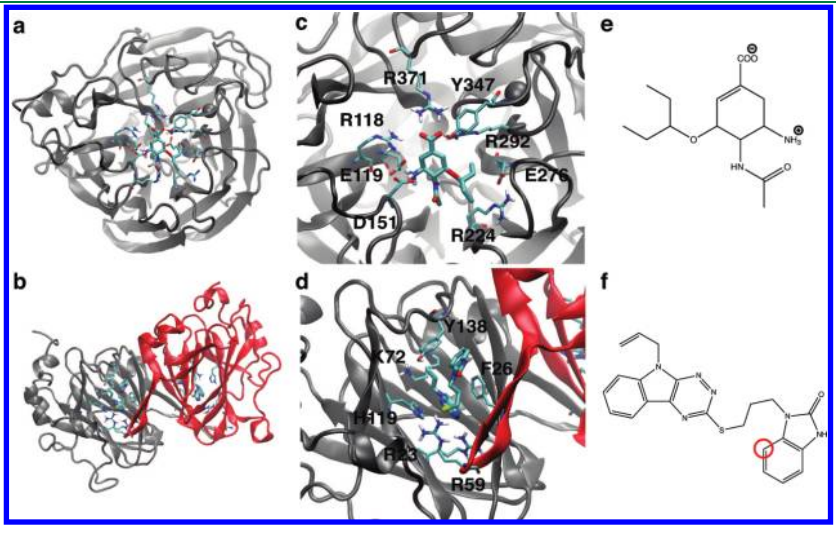

Figure 1. Protein−ligand structures of the investigated systems. Overall views of the N1 monomer (a) and RmlC dimer (b) structures are shown, with the RmlC monomers in b colored to highlight the interface. The active site residues of the two proteins are labeled for N1 in (a) and for RmlC in (d). Ligand chemical structures are depicted for the N1 ligand oseltamivir (e) and the RmlC inhibitor 77074 (f), the latter with the restrained atom (see Materials and Methods) highlighted in red. For oseltamivir, the center of mass was restrained, not a single atom (see Materials and Methods).

Table 1. Protocols for IT-TI Calculations

reference name	elec/vdW no. λ	initialization	nonbonded	decoupling	runs × time/λ (ns)	total time (ns)
medium cont/simul/14λ[a]	9/14	continuous	simultaneous	inter + intra	20 × 1	280
long cont/simul/14λ	9/14	continuous	simultaneous	inter + intra	10 × 2	280
medium cont/simul/inter/14λ	9/14	continuous	simultaneous	inter only	20 × 1	280
medium cont/sep/14λ	9/5	continuous	separate	inter + intra	20 × 1	280
medium cont/sep/19λ[b]	9/10	continuous	separate	inter + intra	20 × 1	80
medium parall/simul/14λ	9/14	parallel[c]	simultaneous	inter + intra	20 × 1	280
medium parall/simul/19λ	9/19	parallel[c]	simultaneous	inter + intra	20 × 1	380
medium parall/simul/inter/14λ	9/14	parallel[c]	simultaneous	inter only	20 × 1	280
medium parall/sep/14λ	9/5	parallel[c]	separate	inter + intra	20 × 1	280
medium parall/sep/19λ	9/10	parallel[c]	separate	inter + intra	20 × 1	380
long parall/sep/19λ	9/10	parallel[c]	separate	inter + intra	10 × 2	380

[a] For N1, 14 λ = [0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0,4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]; for RmlC, [0, 0.1, 0.2, 0.25, 0.3, 0.37, 0.45, 0.5, 0.65, 0.75, 0.8, 0.9, 0.97, 1]. [b] For N1, 19 λ adds [0.55, 0.65, 0.75, 0.85, 0.95]; for RmlC, [0.05, 0.15, 0.6, 0.7, 0.85]. [c] Protocols are well-suited for distributed computing.

the surrounding environment with the coupling parameter λ that changes from 0 to 1 to linearly scale all ligand nonbonded potential energy terms as

$$U(X;\lambda) = U_{\text{unperturbed}}(X) + \lambda U_{\text{decoupled}}(X) + (1-\lambda)U_{\text{coupled}}(X) \quad (2)$$

where X denotes the overall system configurational space assuming equilibrium conditions. In all cases the soft-core potential by Zacharias et al. was employed to enhance sampling and eliminate instabilities (shift parameter $\delta = 5$).[35] The $\partial U/\partial\lambda$ values of eq 1 were printed for each λ every 0.1 ps, and their forward cumulative average was monitored to evaluate convergence (generally reached within equilibration periods of 500 ps). Numerical integration of eq 1 was performed using an interpolated cubic spline.

A harmonic restraining potential $U(r_L) = (1/2)k_h(r_L - r_0)^2$ was applied to restrict ligand sampling r_L to a finite volume V_{pocket} within the active site throughout the TI calculations of $\Delta F^\circ_{\text{protein}}$. Reasonable k_h values were obtained from average fluctuations of the ligand position ($\langle\delta r^2\rangle$) during a free 2 ns N, p, T MD run as $k_h = (3RT)/(\langle\delta r^2\rangle)$,[14,36] with R the molar gas constant and T the absolute temperature of 300 K. A k_h = 2.9 kcal·mol^{-1}·Å^{-2} was used for restraint of the oseltamivir center of mass and k_h = 0.74 kcal·mol^{-1}· Å^{-2} for restraint of a central atom (highlighted in Figure 1f) in the 77074 ligand.

Then, the standard-state free energy was taken into account[8,9,37] for $\Delta F^\circ_{\text{protein}}$ through an analytical correction for transferral of the ligand from the restricted volume V_{pocket} to the bulk V° as

$$\Delta F^\circ_{\text{protein}} = \int_0^1 d\lambda \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda + RT \ln\left(\frac{V_{\text{pocket}}}{V^\circ}\right) \quad (3)$$

To reflect protein−ligand binding at a standard ligand concentration of 1 M, V° = 1661 Å^3, with T = 300 K. V_{pocket} was explicitly determined from multiple MD trajectories using the VMD VolMap plugin.[38] This procedure gave average (RT ln ($V^{\text{pocket}}/V^\circ$)) corrections of −1.25 kcal·mol^{-1} for the N1 system and −1.07 kcal·mol^{-1} for the RmlC system. We note that the magnitude of such corrections is significant (up to 10% of the ΔF_{bind} values for both systems) and should not be neglected.[9,37] For each RmlC calculation, the $\Delta F^\circ_{\text{protein}}$ was halved to obtain an average value for one active site.

One can obtain IT-TI $\Delta F^\circ_{\text{bind}}$ estimates from all combinations of K-independent $\Delta F^\circ_{\text{water}}$ estimates and J-independent $\Delta F^\circ_{\text{protein}}$ estimates as

$$\Delta F^\circ_{\text{bind},(k,j)} = [\Delta F^\circ_{\text{water},k} - \Delta F^\circ_{\text{protein},j}]^{k=1\ldots K}_{j=1\ldots J} \quad (4)$$

Here, a total of $N = KJ$ estimates of $\Delta F^\circ_{\text{bind}}$ are generated and binned in windows of width RT = 0.6 kcal·mol^{-1}. The linear average of the N-independent binding free energy estimates, $\Delta\overline{F}^\circ_{\text{bind}}$, is reported throughout the paper.

2.3. Alternative IT-TI Protocols. Alternative approaches for IT-TI distributed computing were investigated by using more, *medium* independent simulations or fewer, *long* independent simulations. Effects of varied user-defined inputs for TI were also explored, as summarized in Table 1. For independent TI calculations, the λ intermediate simulations were either initialized continuously (*cont* protocols) or in parallel (*parall* protocols). In the first case, simulations at λ = 0 started from the configuration (coordinates and velocities) from a 2 ns N, p, T equilibrated system; at each increasing λ value, the end configuration from the previous λ simulation was used. These IT-TI protocols are less-suited for distributed computing because the MD initialization requires information from sequential runs, but this approach does allow more equilibrated starting structures at successive λ values. Instead, for the parall protocols, all λ simulations were independently initialized from the same N, p, T equilibrated structure with a random velocity. This approach is well-suited for distributed computing, because the MD initialization is independent among each λ simulation. Ligand electrostatics and van der Waals interactions were perturbed, as in eq 2, in three alternative ways (see Table 1). First, electrostatic interactions were decoupled for $0 \le \lambda \le 0.5$ and van der Waals more slowly for $0 \le \lambda \le 1$ (*simul* protocol). Second, the same components were scaled separately, with electrostatic interactions for $0 \le \lambda \le 0.5$ and then van der Waals for $0.5 \le \lambda \le 1$ (*sep* protocol). Third, only the intermolecular terms were decoupled (*inter* protocol).

2.4. Error Analysis of IT-TI Predictions. We evaluated the *accuracy* and *precision* of our IT-TI estimates. Accuracy is described by the match of $\Delta\overline{F}^\circ_{\text{bind}}$ with respect to a reference experimental value, here assumed to be characterized by zero uncertainty. Precision is reflected in the spread of the IT-TI $\Delta F^\circ_{\text{bind}}$ estimates and is described by the standard deviation σ_{bind}. Here σ_{bind} has two components, σ_{water} from the

$\Delta\overline{F}^{\circ}_{water}$ calculations and $\sigma_{protein}$ from the $\Delta F^{\circ}_{protein}$ calculations. Accuracy is limited by systematic errors, which are due to, for example, empirical force field and water models, as well as numerical approximations in the MD algorithms. Both accuracy and precision is affected by random errors from finite sampling. We can capture the statistical uncertainty on the IT-TI $\Delta\overline{F}^{\circ}_{bind}$ due to random errors from N independent calculations with the standard error $\delta = \sigma/\sqrt{N}$, as previously suggested.[14] Note that this metric approaches zero for large N. We computed this uncertainty for the J estimates of $\Delta F^{\circ}_{protein}$ and for K estimates of ΔF°_{water} (eq 4) and propagate for the uncertainty on $\Delta\overline{F}^{\circ}_{bind}$ as

$$\delta_{bind} = \sqrt{\frac{\sigma_{water}^2}{\sqrt{K}} + \frac{\sigma_{protein}^2}{\sqrt{J}}}$$

2.5. Analysis of Conformational Sampling. MD snapshots were saved every 2 ps for analysis, with all protein backbone atoms first aligned to a reference structure. Active site residues for both systems were identified as those within 5 Å of the ligand.

For each system, principal component analysis (PCA)[39,40] of protein fluctuations was performed by calculating the covariance matrix for active site heavy atoms with GROMACS (version 4.0.4 compiled in double precision),[41] using all λ simulations in all J = 10 *long cont/simul/14λ* calculations, for 280 ns of total simulation time (Table 1). Then, projections for independent λ simulations were generated along 20 out of the total 528 principal components (PCs) of this matrix, accounting for 75% of the protein fluctuations. Projections for trajectories using other IT-TI protocols for a given system are along these same PCs for comparison, with projections along the four most dominant PC described in detail. We also project previously performed λ = 0 apo and holo MD simulations onto these PCs for reference, with 400 and 10 ns each for apo and holo simulation with N1 and RmlC, respectively. Details of these N1 simulations have been previously reported.[25] For a fair comparison of the two systems, all projections were reweighted as w^{-1}, w being the number of active site atoms (N1, 176; RmlC, 161). For hydration analysis, water−water hydrogen bonds within 5 Å of the ligand in the active site were monitored. Hydrogen bonds were defined to have a maximum hydrogen−acceptor distance of 3.5 Å and a minimum donor−hydrogen−acceptor angle of 120°. The software VMD,[38] xmgrace, as well as python scripts based on matplotlib and NumPy libraries were used for analysis and graphical representations.

3. RESULTS AND DISCUSSION

3.1. IT-TI Free Energy Distributions and Their Dependence on Biomolecular Flexibility. Because independent TI estimates of ΔF°_{bind} vary with the specific set of simulations performed, IT-TI generates distributions of free energy estimates and provides an average value $\Delta\overline{F}^{\circ}_{bind}$ with a reliable measure of uncertainty (δ_{bind}).[14] We evaluate the accuracy of our predicted $\Delta\overline{F}^{\circ}_{bind}$ values with reference free energies derived from the experimental K_i as $\Delta F_{expt} = RT\ln(K_i)$. For N1-oseltamivir and RmlC-77074 binding, ΔF_{expt} values of −13.7 and −9.9 kcal·mol^{-1} were reported, respectively.[21,42] We compute N IT-TI estimates of ΔF°_{bind} from K independent calculations of ΔF°_{water} and J calculations of $\Delta F^{\circ}_{protein}$ (eq 4). The K = 20 ΔF°_{water} results have a much smaller spread relative to the J = 20 $\Delta F^{\circ}_{protein}$ results, with σ_{water} = 0.4 and 0.2 kcal·mol^{-1} compared to $\sigma_{protein}$ = 4.4 and 3.2 kcal·mol^{-1} for the *medium cont/simul/14λ* N1-oseltamivir and RmlC-77074 calculations, respectively. Thus, the shape of the ΔF°_{bind} distributions is dominated by the variation of the J $\Delta F^{\circ}_{protein}$ results, as expected due to the numerous degrees of freedom and complex energy landscape in this state.

Table 2. Summary of IT-TI Results with Varied Protocols

	$\Delta\overline{F}^{\circ}_{bind} \pm \delta_{bind}$ (kcal·mol^{-1})	
reference name	N1	RmlC
medium cont/simul/14λ	−12.2 ± 1.0	−11.0 ± 0.4
long cont/simul/14λ	−10.4 ± 1.6	−9.4 ± 0.4
medium cont/simul/inter/14λ	−14.9 ± 1.2	
medium cont/sep/14λ	−10.4 ± 1.2	
medium cont/sep/19λ	−13.7 ± 1.1	
medium parall/simul/14λ	−10.6 ± 0.6	
medium parall/simul/19λ	−11.2 ± 0.6	
medium parall/simul/inter/14λ	−12.2 ± 0.7	
medium parall/sep/14λ	−11.1 ± 0.6	
medium parall/sep/19λ	**−14.3 ± 0.5**	**−11.8 ± 0.3**
long parall/sep/19λ	**−12.8 ± 0.6**	**−10.8 ± 0.2**

Both J = 20 medium and J = 10 long independent protein simulations were used to compute ΔF°_{bind}, to test the computational efficiency of more, shorter runs compared with fewer, longer independent runs. See Table 2 for a summary of all IT-TI predictions. Figure 2 shows distributions of ΔF°_{bind} estimates obtained using the cont/simul/14λ protocol in Table 1. The distributions are remarkably different for the two systems investigated. The N1-oseltamivir results in Figure 2a,c have a broad range, from very favorable (−20 kcal·mol^{-1}) to unfavorable (>0 kcal·mol^{-1}). As reported in Table 2, estimates from medium runs give $\Delta\overline{F}^{\circ}_{bind}$ = −12.2 ± 1.0 kcal·mol^{-1}. Corresponding results for the long simulations display a marked shift of $\Delta\overline{F}^{\circ}_{bind}$ to 3.3 kcal·mol^{-1} away from ΔF_{expt} and an increase of δ_{bind}, with $\Delta\overline{F}^{\circ}_{bind}$ = −10.4 ± 1.6 kcal·mol^{-1}. The use of more, independent simulations improved the free energy results in this case. In contrast, the RmlC-77074 distributions are centered near ΔF_{expt} (see Figure 2b,d), and the use of fewer, independent runs with longer sampling times gave the most accurate and precise results. A close match with experiment is found for the long simulation results in Figure 2d, with $\Delta\overline{F}^{\circ}_{bind}$ = −9.4 ± 0.4 kcal·mol^{-1} (Table 2). Overall, the δ_{bind} of the RmlC-77074 results is significantly smaller than the δ_{bind} of the N1-oseltamivir results, due to a much smaller spread σ_{bind}.

To probe underlying causes of the different free energy results for the two systems, we analyzed protein sampling during the independent simulations with PCA of protein fluctuations. Figure 3 shows projections along the two most dominant principal components at 5 λ values, [0, 0.2, 0.5, 0.8, 1], depicting changes in protein sampling along the perturbation in eq 2. For comparison, the same data for longer apo and holo N1 and RmlC simulations with λ = 0 (see Materials and Methods) onto the same PC are reported (Figure 3). For N1, the J simulations slowly equilibrated into varied portions of phase space, resulting in non-overlapping projections at λ = 1. Many of the simulations also exclusively sampled motions that are not visited in the reference holo or apo simulations. Our analysis indicates highly frustrated N1 sampling as the λ simulations are continuously

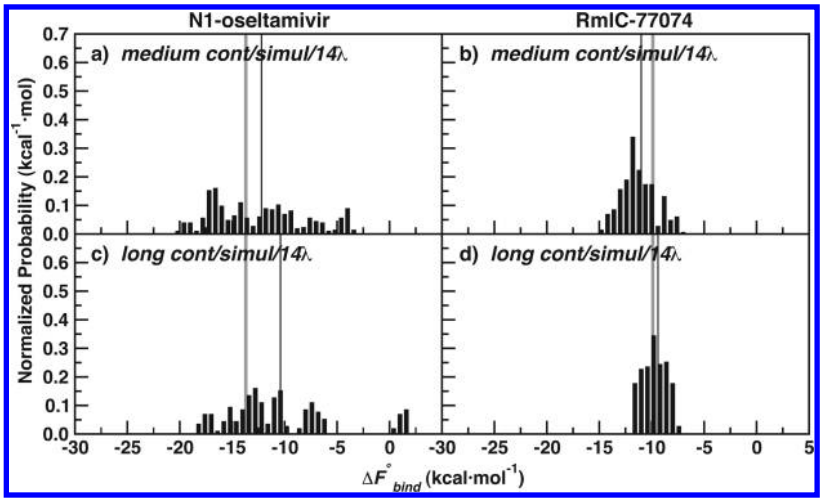

Figure 2. Normalized distributions of N1-oseltamivir (left) and RmlC-77074 (right) IT-TI results for medium and long cont/simul/14λ TI protocols (Table 1). ΔF_{expt} for both systems is also depicted (gray line), along with $\Delta\overline{F}^{\circ}_{bind}$ (thin black line).

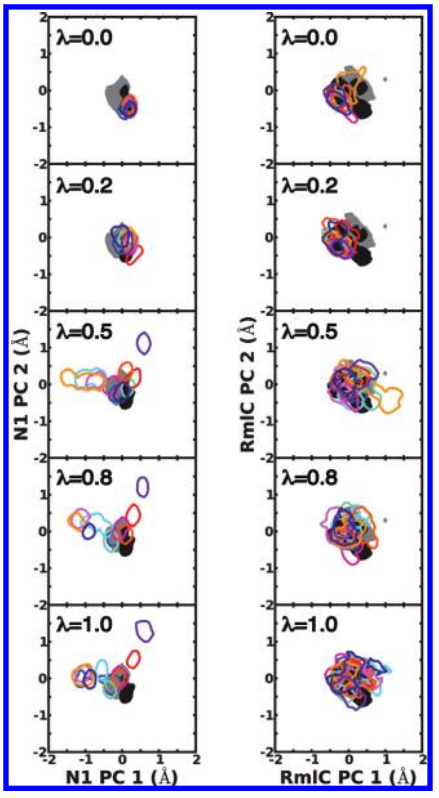

Figure 3. Receptor flexibility for N1 (left) and RmlC (right) as captured by two dominant principal components (PCs) of active site residue fluctuations from long cont/simul14λ simulations. Contours depicting projections for 90% of the apo (filled gray) and holo (filled black) MD simulations, as well as each of $J = 10$ independent trajectories (unfilled color) are shown at λ values [0, 0.2, 0.5, 0.8, 1]. Projections are reweighted to allow direct comparison between the two systems. See Materials and Methods for details.

initialized, contributing to varied free energy estimates and a large $\sigma_{protein}$ component of σ_{bind}. A different picture emerges for RmlC-77074, which had a significantly smaller $\sigma_{protein}$ compared to N1-oseltamivir. In this case, the J independent simulations accessed similar regions of conformational space, as inferred from overlapping projections (see Figure 3). The RmlC projections also significantly overlap with projections from apo and holo reference simulations. These observations hold similarly when analyzing projections along other, less dominant PC from PCA (not shown) and link the varied spreads of IT-TI free energy estimates, and corresponding uncertainties δ_{bind}, to protein conformational sampling.

Differences in N1 and RmlC dynamics are also revealed in the sampling of specific binding site residue torsions. Comparison of torsion sampling at $\lambda = 0$ and at $\lambda = 1$ reveals that 9 out of 15 monitored N1 active site residues, but only 2 out of 11 RmlC residues, increased flexibility and sampled multiple conformations upon ligand decoupling. The torsional angles were also sampled in populations that vary among the $J = 10$ independent runs, particularly for charged N1 residues R224, R371, R118, E277, and E119 (Supporting Information Figure S1). As seen in the PCA, the N1 system is challenged to access its conformational space within a single simulation, while for RmlC, sampling is more complete within and similar among independent IT-TI simulations. We note that, in addition to protein sampling and $\sigma_{protein}$, water sampling and σ_{water} contributes to the varied σ_{bind} for the two systems; the more flexible ligand oseltamivir has more diverse sampling than the sterically hindered 77074, reflected in the larger σ_{water} for this ligand (see above). Altogether, these sampling behaviors yield the different uncertainties δ_{bind} on $\Delta\overline{F}^{\circ}_{bind}$ estimates for the two systems (see Table 2).

3.2. IT-TI Free Energy Distributions and Their Dependence on Solvent Effects. Hydration dynamics and solvent fluctuations also contribute to the spread of the IT-TI free energy distributions, in addition to protein and ligand flexibility described in the previous section. Here we report an example from the N1-oseltamivir IT-TI results in closer detail. In Figure 2c, an outlier, unfavorable $\Delta F^{\circ}_{protein}$ estimate was computed (see histogram bars around $\Delta F^{\circ}_{bind} = 0$). This result can be linked to pronounced solvent fluctuations during the $\Delta F^{\circ}_{protein}$ calculation at λ values 0.2 and 0.25. At these intermediate states, water molecules diffuse into the active site, very close to the partially decoupled ligand carboxyl and ammonium groups, and an increased number of active site water−water hydrogen bonds is observed (Figure 4c). This coincides with a shift in the electrostatics component of $\partial U/\delta\lambda$ (Figure 4b), giving a less

positive integrated $\Delta F^{\circ}_{protein}$ value and unfavorable ΔF°_{bind} estimates (eq 4).

These observations are fully consistent with the dynamic nature of protein hydration and dewetting fluctuations in binding cavities recently reported in the literature[43−45] and their thermodynamic relevance.[46−48] Because time scales of these solvent fluctuations may reach several hundred picoseconds, it is expected that our individual nanosecond TI runs may have diverse solvent behavior among the 10 performed. Here, the advantage of using IT-TI is illustrated, since a single TI calculation could yield a falsely unfavorable ΔF°_{bind} estimate for N1-oseltmivir. Multiple estimates of ΔF°_{bind} allow recovery of the probability distribution from multiple, independent simulations that sample both rare and dominant events. With enough independent estimates, this distribution should reflect that of the true physical system. We also note that the solvent-exposed N1 has a consistent number of water molecules in the active site throughout the TI calculations (Figure 4c), highlighting the importance of water in both the bound and unbound states. Instead, the RmlC binding site has a more abrupt influx of water near $\lambda = 1$ (Supporting Information Figure S2).

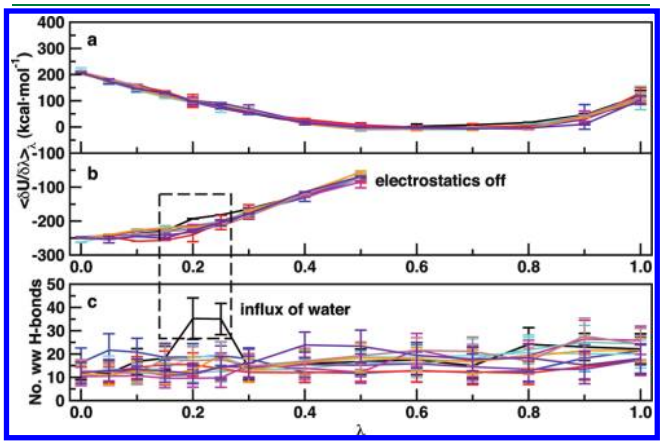

Figure 4. N1 active site hydration behavior and corresponding $\langle \partial U/\delta\lambda \rangle_{\lambda}$ values from the long cont/simul/14λ protocol. $J = 10$ independent estimates of $\langle \partial U/\delta\lambda \rangle_{\lambda}$ are color-coded and interpolated for (a) van der Waals and (b) electrostatics components, with (c) the average and standard deviation of water−water hydrogen bonds within 5 Å of oseltamivir at each λ. The black curve in all panels indicates the TI calculation which gave an unfavorable $\Delta F^{\circ}_{protein}$ result. See Materials and Methods for details.

3.3. N1-oseltamivir Protocol Investigation. In an effort to improve consistency of the N1-oseltamivir $\langle \partial U/\delta\lambda \rangle_{\lambda}$ values in Figure 4 and the free energy estimates in Figure 2, we conducted a series of IT-TI protocol changes for the N1-oseltamivir system. The varied N1-oseltamivir medium protocols implemented for the IT-TI calculations are described in Table 1, with the corresponding $\Delta\overline{F}^{\circ}_{bind} \pm \delta_{bind}$ listed in Table 2. Here, we aim for improved precision and accuracy over the medium cont/simul/14λ results (Figure 2a and Figure 5a). We compare the spread, σ_{bind}, of the IT-TI distributions in Figure 5 to the $\sigma_{bind} = 4.4$ kcal·mol^{-1} of the medium cont/simul/14λ results.

All protocols with continuous initialization of λ intermediates gave free energy distributions with a large spread, with $4.4 \leq \sigma_{bind} \leq 5.4$ kcal·mol^{-1} (see Figure 5a,c,e,g). In these cases, the $\Delta\overline{F}^{\circ}_{bind}$ values were also consistently less favorable than ΔF_{expt}, with the exception of protocol cont/sep/19λ in Figure 5g. In the latter case, estimates are shifted to more favorable values and $\Delta\overline{F}^{\circ}_{bind}$ matches the ΔF_{expt} of −13.7 kcal·mol^{-1} with $\sigma_{bind} = 4.8$ kcal·mol^{-1} (Table 2). PCA of these cont simulations (Supporting Information Figure S3) indicated frustrated sampling, with projections similar to those seen in Figure 3. Decoupling of only ligand intermolecular nonbonded components in both protocols cont/simul/inter/14λ and parall/simul/inter/14λ reduced precision and made little difference in accuracy (Figure 5c,d and Table 2).

Overall, the σ_{bind} of the IT-TI distributions is significantly reduced with parallel initialization of each λ simulation. These estimates, in Figure 5b,d,f,h, had σ_{bind} values ≤ 3.0 kcal·mol^{-1}. However, only the protocol parall/sep/19λ gave an accurate $\Delta\overline{F}^{\circ}_{bind}$, close to ΔF_{expt} at −14.3 kcal·mol^{-1} with $\sigma_{bind} = 2.1$ kcal·mol^{-1} (Table 2). This improvement in accuracy is observed

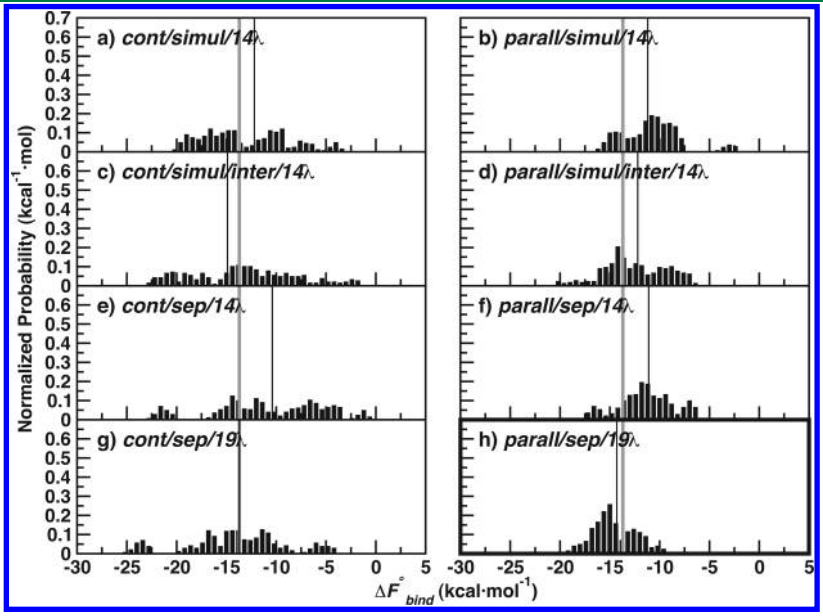

Figure 5. Normalized distributions for N1-oseltamivir IT-TI results with various medium decoupling protocols. Panels are labeled with the procedures from Table 1, and ΔF_{expt} is depicted (gray line), along with $\Delta\overline{F}^{\circ}_{bind}$ (thin black line). $\Delta\overline{F}^{\circ}_{bind} \pm \delta_{bind}$ values are reported in Table 2.

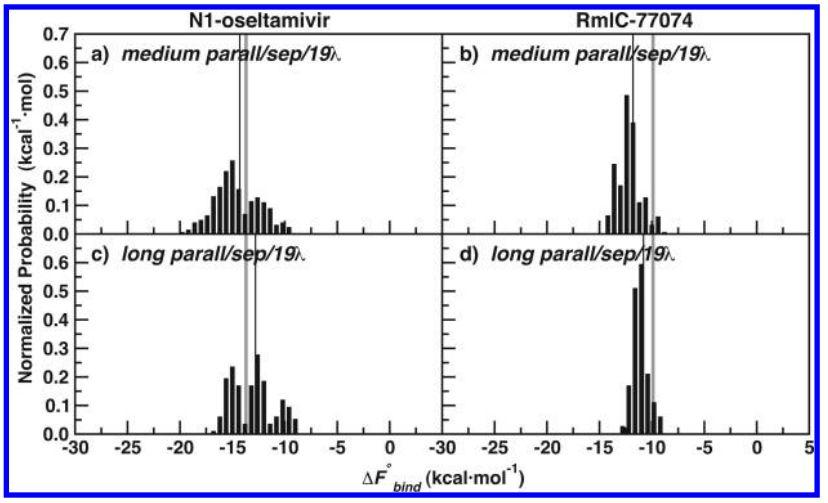

Figure 6. Normalized distributions of N1-oseltamivir (left) and RmlC-77074 (right) IT-TI results for medium and long parall/sep/19λ protocols (Table 2). ΔF_{expt} for both systems is also depicted (gray line), along with $\Delta\overline{F}^{\circ}_{bind}$ (thin black line).

for both the parall and cont protocols with separate decoupling of nonbonded components and 19 λ values (compare Figure 5e,f with Figure 5g,h). For protocols cont/sep/14λ and parall/sep/14λ, the van der Waals interactions are decoupled with only 5 λ intermediates, while the cont/simul/14λ and parall/simul/14λ protocols employed 14 λ (see Table 1). Five additional λ values for the cont/sep and parall/sep protocols, for 10 van der Waals λ intermediates and 19 total λ values, improved interpolation of the van der Waals $\langle \partial U/\delta\lambda \rangle_{\lambda}$ values for more accurate integration. These results are highlighted in bold in Figure 5h and Table 2.

The diverse outcomes obtained using alternative IT-TI protocols show that free energy calculations depend on a broad variety of user-defined choices. An optimal protocol was designed for N1-oseltamivir binding and suggests that (i) TI intermediates can be more conveniently placed at target λ values once a preliminary knowledge of the $\langle \partial U/\delta\lambda \rangle_{\lambda}$ vs λ curve is known; (ii) these λ values may be run in parallel, initialized from a λ = 0 holo configuration—an approach particularly suited for distributed computing; and (iii) separate decoupling of both inter- and intramolecular nonbonded components gives more accurate free energy estimates. We note that the approach described in (ii) may not be optimal for cases when the apo and holo states are separated by large conformational changes.

3.4. Application of Optimized Protocol to Both N1-Oseltamivir and RmlC-77074 Test Systems. We applied our optimal IT-TI protocol for N1-oseltamivir binding to RmlC-77074 and, again, compared estimates from two approaches for distributed computing. For N1-oseltamivir, both approaches gave more accurate IT-TI results with reduced uncertainty compared to Figure 2. In Figure 6a, we see that more, medium simulations gave a more favorable $\Delta\overline{F}^{\circ}_{bind}$ estimate of -14.3 ± 0.5 kcal·mol^{-1}, compared to $\Delta\overline{F}^{\circ}_{bind} = -12.8 \pm 0.6$ kcal·mol^{-1} computed with fewer, long simulations in Figure 6c. The additional simulations enhanced N1 sampling and improved the reliability of the IT-TI estimates (Table 2). For RmlC-77074, the results in Figure 6b,d reflect similar accuracy and precision compared to Figure 2, particularly for the long simulations. Here the $\Delta\overline{F}^{\circ}_{bind} = -10.8 \pm 0.2$ kcal·mol^{-1} (Table 2), with a very small δ_{bind} due to consistent sampling among the independent simulations.

We can directly compare the 2-D projections of the long parall/sep/19λ simulations for both systems in Figure 7 with

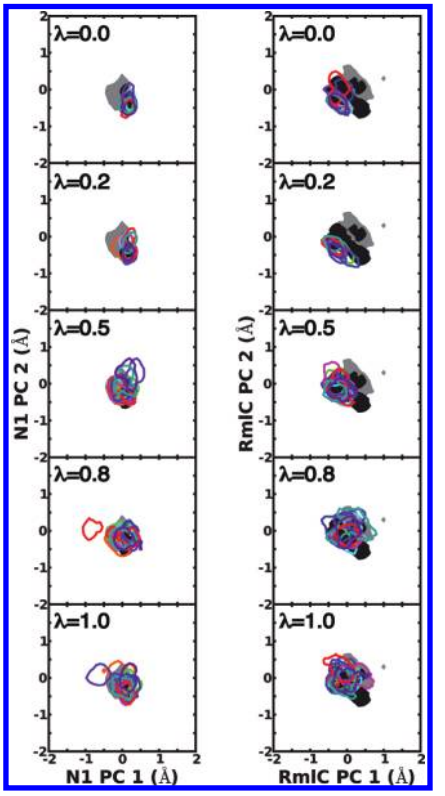

Figure 7. Receptor flexibility for N1 (left) and RmlC (right) as captured by two dominant PCs of active site residue fluctuations from long parall/sep/19λ simulations. See Figure 3 for color coding.

those in Figure 3. With the optimized protocol, the highly frustrated N1 sampling of Figure 3 is largely alleviated; the J = 10 projections overlap significantly with each other as well as with the reference apo and holo projections. One N1 simulation sampled outlier motions, but these exchanged with motions near the holo state during the same simulation (red contours in Figure 7 at λ = 0.8). The RmlC projections are similar to those in Figure 3, with consistent overlap among the independent runs and access of both apo and holo motions. In addition to more consistent protein sampling, both systems have reduced

fluctuations of water−water hydrogen bonds in the active site when using the parall/sep protocol (Supporting Information Figure S2). This is primarily due to parallel initialization, rather than the separate decoupling protocol.

4. CONCLUSION

We investigated the (thermo)dynamics underlying two protein−ligand binding processes that involve very different protein, ligand, and water sampling. The distributions of binding free energy estimates produced from independent-trajectory thermodynamic integration (IT-TI) were remarkably centered on the target experimental values but had very different spreads for the two protein−ligand systems. The solvent exposed active site of N1, with many flexible, charged binding residues, has a larger population of microstates, with nontrivial barriers, that are easily over- or undersampled during a single TI calculation. The frustrated sampling observed for this system resulted in a broad range of free energy estimates from IT-TI, and additional independent runs, rather than longer sampling times, gave more reliable results. The RmlC microstates are more accessible, even within short (ns) simulations. Here, the IT-TI distributions had a smaller spread, and extension of simulation time improved the reliability of the results.

With tests of varied protocols for TI, we find that, for both protein systems, each alchemical intermediate may be run in parallel, each initialized with a $\lambda = 0$ configuration, for more consistent free energy results with maintained accuracy. This parallel approach also allows faster completion of the calculations compared to calculations with continuously initialized intermediates and is ideal for distributed computing. Additionally, separate decoupling of the inter- and intramolecular nonbonded terms gave optimal accuracy and precision overall, but only when employed with adequate intermediates for smooth interpolation of both electrostatics and van der Waals $\langle \partial U/\partial \lambda \rangle_\lambda$ values during integration (eq 1). We suggest an approach for performance of IT-TI calculations that maximizes reliability and computational efficiency with available sampling times.

■ ASSOCIATED CONTENT

Supporting Information. Analysis of torsional angle sampling for N1 and RmlC active site residues (Figure S1), N1 and RmlC active site hydration behavior for varied IT-TI protocols (Figure S2), and PCA of N1 active site residue fluctuations from varied IT-TI protocols (Figure S3). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mlawrenz@ucsd.edu (M.L.), rbaron@mccammon.ucsd.edu (R.B.). Tel.: +1-858-822-2771 (M.L.), +1-858-534-2913 (R.B.).

■ ACKNOWLEDGMENT

We thank the members of the McCammon group for useful discussions. This work was supported, in part, by the National Institutes of Health, the National Science Foundation, the National Biomedical Computational Resource, and the Howard Hughes Medical Institute. We thank the Center for Theoretical Biological Physics (NSF Grant PHY-0822283) and the Texas Advanced Computer Center (Grant TG-MCA93S013) for distributed computing resources. We also thank Dr. Ross C. Walker at the San Diego Supercomputing Center for additional computational resources.

■ REFERENCES

(1) Tembe, B.; McCammon, J. *Comput. Chem.* **1984**, *8*, 281–283.
(2) Beveridge, D.; DiCapua, F. *Annu. Rev. Biophys. Chem.* **1989**, *18*, 431–492.
(3) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395–2417.
(4) Jorgensen, W. *Science* **2004**, *303*, 1813–1818.
(5) Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem.* **2010**, 10235–10253.
(6) Gilson, M. K.; Zhou, H.-X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
(7) van Gunsteren, W. F.; Beutler, T. C.; Fraternali, F.; King, P. M.; Mark, A. E.; Smith, P. E. *Computation of Free Energy in Practice: Choice of Approximations and Accuracy Limiting Factors*; ESCOM Science: Leiden, The Netherlands, 1993; Vol. 2
(8) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.
(9) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535–9551.
(10) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 084108.
(11) Zagrovic, B.; van Gunsteren, W. *J. Chem. Theory Comput.* **2007**, *3*, 301–311.
(12) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118–1134.
(13) Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A.; Shoichet, B. K. *J. Mol. Biol.* **2009**, *394*, 747–763.
(14) Lawrenz, M.; Baron, R.; McCammon, J. *J. Chem. Theory Comput.* **2009**, *5*, 1106–1116.
(15) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050–3054.
(16) Wong, C.; McCammon, J. *J. Am. Chem. Soc.* **1986**, *108*, 3830–3832.
(17) Russell, R. J.; Haire, L. F.; Stevens, D. J.; Collins, P. J.; Lin, Y. P.; Blackburn, G. M.; Hay, A. J.; Gamblin, S. J.; Skehel, J. J. *Nature* **2006**, *443*, 45–49.
(18) Amaro, R. E.; Minh, D. D. L.; Cheng, L. S.; Lindstrom, W. M.; Olson, A. J.; Lin, J.-H.; Li, W. W.; McCammon, J. A. *J. Am. Chem. Soc.* **2007**, *129*, 7764–7765.
(19) Le, L.; Lee, E. H.; Hardy, D. J.; Truong, T. N.; Schulten, K. *PLoS Comput. Biol.* **2010**, *6*, No. e1000939.
(20) Stoll, V.; Kent, S. D.; Maring, C. J.; Muchmore, S.; Giranda, V.; Gu, Y. Y.; Wang, G.; Chen, Y.; Sun, M.; Zhao, C.; Kennedy, A. L.; Madigan, D. L.; Xu, Y.; Saldivar, A.; Kati, W.; Laver, G.; Sowin, T.; Sham, H. L.; Greer, J.; Kempf, D. *Biochemistry* **2003**, *42*, 718–727.
(21) Sivendran, S.; Jones, V.; Sun, D.; Wang, Y.; Grzegorzewicz, A. E.; Scherman, M. S.; Napper, A. D.; McCammon, J. A.; Lee, R. E.; Diamond, S. L.; McNeil, M. *Bioorg. Med. Chem.* **2010**, *18*, 896–908.
(22) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.
(23) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
(24) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.
(25) Lawrenz, M.; Wereszczynski, J.; Amaro, R.; Walker, R.; Roitberg, A.; McCammon, J. A. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2523–2532.
(26) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Case, P. A. K. D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
(27) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
(28) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G; Robb, M.; Cheeseman, J.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K.; Burant, J.; Millam, J.; Iyengar, S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.;

Scalmani, G.; Rega, N.; Petersson, G.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J.; Hratchian, H.; Cross, J.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.; Yazyev, O.; Austin, A.; Cammi, R.; Pomelli, C.; Ochterski, J.; Ayala, P.; Morokuma, K.; Voth, G.; Salvador, P.; Dannenberg, J.; Zakrzewski, V.; Dapprich, S.; Daniels, A.; Strain, M.; Farkas., O.; Malick, D.; Rabuck, A.; Raghavachari, K.; Foresman, J.; Ortiz, J.; Cui, Q.; Baboul, A.; Clifford, S.; Cioslowski, J.; Stefanov, B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R.; Fox, D.; Keith, T.; Laham, A.; Peng, C.; Nanayakkara, A.; Challacombe, M.; Gill, P.; Johnson, B.; Chen, W.; Wong, M.; Gonzalez, C.; Pople, J. *Gaussian 03*, Revision C.02, 2003; Gaussian, Wallingford, CT, 2004.

(29) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(30) Andersen, H. *J. Comput. Phys.* **1983**, *52*, 24–34.

(31) Shuichi, M.; Peter, A. *J. Comput. Chem.* **1992**, *13* (952−962), No. 148324.

(32) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(33) Feller, S.; Zhang, Y.; Pastor, R.; Brooks, B. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(34) Kirkwood, J. *J. Chem. Phys.* **1935**, 300–313.

(35) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025–9031.

(36) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.

(37) General, I. J. *J. Chem. Theory Comput.* **2010**, *6*, 2520–2524.

(38) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14* (33−38), 27–28.

(39) Garca, A. E. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699.

(40) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct., Funct., Bioinf.* **1993**, *17*, 412–425.

(41) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(42) Kati, W. M.; Montgomery, D.; Carrick, R.; Gubareva, L.; Maring, C.; McDaniel, K.; Steffy, K.; Molla, A.; Hayden, F.; Kempf, D.; Kohlbrenner, W. *Antimicrob. Agents Chemother.* **2002**, *46*, 1014–1021.

(43) Setny, P.; Geller, M. *J. Chem. Phys.* **2006**, *125*, 144717.

(44) Baron, R.; McCammon, J. *Biochemistry* **2007**, *46*, 10629–10642.

(45) Young, T.; Hua, L.; Huang, X.; Abel, R.; Friesner, R.; Berne, B. J. *Proteins* **2010**, *78*, 1856–1869.

(46) Baron, R.; Setny, P.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 12091–12097.

(47) Setny, P.; Baron, R.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 2866–2871.

(48) Hummer, G. *Nature Chem.* **2010**, *2*, 906–907.