Gaussian accelerated molecular dynamics in OpenMM

Matthew Copeland[1,+], Hung N. Do[1,+], Lane Votapka[2], Keya Joshi[1], Jinan Wang[1], Rommie E. Amaro[2] and Yinglong Miao[1,*]

[1]Center for Computational Biology and Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66047; [2]Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093

[+] These authors contributed equally to this work
[*] To whom correspondence should be addressed: miao@ku.edu

Virtual Special Issue (VSI) of The Journal of Physical Chemistry B entitled "Biomolecular Electrostatic Phenomena".

Abstract

Gaussian accelerated molecular dynamics (GaMD) is a computational technique that provides both unconstrained enhanced sampling and free energy calculations of biomolecules. Here, we present the implementation of GaMD in the OpenMM simulation package and validate it on model systems of alanine dipeptide and RNA folding. For alanine dipeptide, 30ns GaMD production simulations reproduced free energy profiles of 1000ns conventional molecular dynamics (cMD) simulations. In addition, GaMD simulations captured folding pathways of three hyperstable RNA tetraloops (UUCG, GCAA, and CUUG) and binding of the rbt203 ligand to the HIV-1 Tar RNA, both of which involved critical electrostatic interactions such as hydrogen bonding and base stacking. Together with previous implementations, GaMD in OpenMM will allow for wider applications in simulations of proteins, RNA, and other biomolecules.

Keywords: Gaussian accelerated molecular dynamics (GaMD), OpenMM, enhanced sampling, electrostatics, biomolecules.

Introduction

Molecular dynamics (MD) is a powerful computational technique for simulating biomolecular dynamics at an atomistic level[1]. Due to advancements in computing hardware and software, timescales accessible to MD simulations have increased, while costs have decreased[2-3]. However, conventional MD (cMD), which makes no use of any enhanced sampling schemes, is often limited to tens to hundreds of microseconds[3-10] for simulations of biomolecular systems, and cannot attain the timescales required to observe many biological processes of interest, which typically occur over milliseconds or longer, due to high energy barriers (e.g., 8-12 kcal/mol)[3-10].

Many enhanced sampling techniques have been developed during the last several decades to overcome the challenges mentioned above[11-15]. One class of enhanced sampling techniques use predefined collective variables (CVs) or reaction coordinates (RCs), including umbrella sampling[16-17], metadynamics[18-19], adaptive biasing force[20-21] and steered MD[22]. However, it can be challenging to define proper CVs prior to simulation[3], and predefined CVs might significantly limit the sampling of conformational space during simulations[3]. Another class of enhanced sampling techniques, including replica exchange MD (REMD)[23-24] or parallel tempering[25], self-guided Langevin MD[26-28] and accelerated MD (aMD)[29-30], do not require predefined CVs. The latter class of unconstrained enhanced sampling techniques remain attractive to improve the sampling of biomolecular dynamics and obtain sufficient accuracy in free energy calculations.

Gaussian accelerated molecular dynamics (GaMD) is an unconstrained enhanced sampling technique that works by applying a harmonic boost potential to smooth biomolecular potential energy surface[31]. Since this boost potential usually exhibits a near

Gaussian distribution, cumulant expansion to the second order ("Gaussian approximation") can be applied to achieve proper energy reweighting[32]. GaMD allows for simultaneous unconstrained enhanced sampling and free energy calculations of large biomolecules[31]. GaMD has been successfully demonstrated on enhanced sampling of ligand binding[31, 33-36], protein folding[31, 35], protein conformational changes[34, 37-40], protein-membrane[41], protein-protein[42-44], and protein-nucleic acid[45-46] interactions. Furthermore, GaMD has been combined with REMD[47-48] to further improve conformational sampling and free energy calculations[3]. In addition, "selective GaMD" algorithms, including Ligand GaMD (LiGaMD)[49], Peptide GaMD (Pep-GaMD)[50], and Protein-Protein Interaction-GaMD (PPI-GaMD)[44] have been developed to enable repetitive binding and dissociation of small-molecule ligands, highly flexible peptides, and proteins within microsecond simulations, which allow for highly efficient and accurate calculations of ligand/peptide/protein binding free energy and kinetic rate constants[3]. Recently, GaMD has been combined with Deep Learning and free energy profiling in a workflow (GLOW) to predict molecular determinants and map free energy landscapes of biomolecules[37]. GaMD has been implemented in widely used simulation packages including AMBER[31], NAMD[35], GENESIS[48], and TINKER-HP[51].

In this work, we have implemented GaMD in the OpenMM simulation package[52]. OpenMM is an open-source scientific software package for performing MD simulations on a range of high-performance computing architectures[52]. OpenMM was designed to be simple and easy to use, hardware independent, and extensible so that new hardware architectures can be accommodated and new functionality can be easily added[52]. In fact, accelerated MD (aMD) has been previously implemented in OpenMM[53]. We validated the

implementation on the model simulations of alanine dipeptide, three hyperstable RNA tetraloops of UUCG, GCAA, and CUUG, and rbt203 ligand binding to the HIV-1 Tar RNA.

Methods

Gaussian accelerated molecular dynamics (GaMD)

GaMD works by adding a harmonic boost potential to smooth the potential energy surface when the system potential drops below a reference energy E[31]:

$$\Delta V(r) = \begin{cases} \frac{1}{2}k\left(E - V(\vec{r})\right)^2, & V(\vec{r}) < E \\ 0, & V(\vec{r}) \geq E, \end{cases} \quad (1)$$

where k is the harmonic force constant. The two adjustable parameters E and k can be determined based on three enhanced sampling principles. First, for any two arbitrary potential values $V_1(\vec{r})$ and $V_2(\vec{r})$ found on the original energy surface, if $V_1(\vec{r}) < V_2(\vec{r})$, ΔV should be a monotonic function that does not change the relative order of the biased potential values; i.e., $V_1^*(\vec{r}) < V_2^*(\vec{r})$. Second, if $V_1(\vec{r}) < V_2(\vec{r})$, the potential difference observed on the smoothed energy surface should be smaller than that of the original, i.e., $V_2^*(\vec{r}) - V_1^*(\vec{r}) < V_2(\vec{r}) - V_1(\vec{r})$. The reference energy needs to be set in the following range:

$$V_{max} \leq E \leq V_{min} + \frac{1}{k}, \quad (2)$$

where V_{max} and V_{min} are the system minimum and maximum potential energies. To ensure that **equation (2)** is valid, k must satisfy: $k \leq \frac{1}{V_{max} - V_{min}}$. Let us define $k \equiv k_0 \frac{1}{V_{max} - V_{min}}$, then $0 \leq k_0 \leq 1$. Third, the standard deviation of ΔV needs to be small

enough (i.e., narrow distribution) to ensure proper energetic reweighting[32]: $\sigma_{\Delta V} = k(E - V_{avg})\sigma_V \leq \sigma_0$, where V_{avg} and σ_V are the average and standard deviation of the system potential energies, $\sigma_{\Delta V}$ is the standard deviation of ΔV with σ_0 as a user-specified upper limit (e.g., $10k_BT$) for proper reweighting. When E is set to the lower bound $E = V_{max}$, k_0 can be calculated as:

$$k_0 = \min(1.0,\ k_0') = \min\left(1.0, \frac{\sigma_0}{\sigma_V}\frac{V_{max} - V_{min}}{V_{max} - V_{avg}}\right), \tag{3}$$

Alternatively, when the threshold energy E is set to its upper bound $E \leq V_{min} + \frac{1}{k}$, k_0 is set to:

$$k_0 = k_0'' \equiv \left(1.0 - \frac{\sigma_0}{\sigma_V}\right)\frac{V_{max} - V_{min}}{V_{avg} - V_{min}}, \tag{4}$$

if k_0'' is found to be between 0 and 1. Otherwise, k_0 is calculated using **equation (3)**.

For energetic reweighting of GaMD simulations, the probability distribution along a selected reaction coordinate can be calculated from simulations as $p^*(A)$. Given the boost potential $\Delta V(r)$ of each frame in GaMD simulations, $p^*(A)$ can be reweighted to recover the canonical ensemble distribution, $p(A)$, as:

$$p(A_j) = p^*(A_j)\frac{\langle e^{\beta\Delta V(\vec{r})}\rangle_j}{\sum_{i=1}^{M}\langle p^*(A_i)e^{\beta\Delta V(\vec{r})}\rangle_i}, \qquad j = 1, \dots, M \tag{5}$$

where M is the number of bins, $\beta = k_BT$ and $\langle e^{\beta\Delta V(\vec{r})}\rangle_j$ is the ensemble-averaged Boltzmann factor of $\Delta V(\vec{r})$ for simulation frames found in the j^{th} bin. The ensemble-averaged reweighting factor can be approximated using cumulant expansion[31-32]:

$$\langle e^{\beta\Delta V(\vec{r})}\rangle = exp\left\{\sum_{k=1}^{\infty}\frac{\beta^k}{k!}C_k\right\}, \tag{6}$$

where the first two cumulants are given by:

$$C_1 = \langle \Delta V \rangle, \qquad (7)$$
$$C_2 = \langle \Delta V^2 \rangle - \langle \Delta V \rangle^2 = \sigma_v^2.$$

The boost potential obtained from GaMD simulations usually shows near-Gaussian distribution[54]. Cumulant expansion to the second order thus provides a good approximation for computing the reweighting factor[31-32]. The reweighted free energy $F(A) = -k_B T \ln p(A)$ is calculated as:

$$F(A) = F^*(A) - \sum_{k=1}^{2} \frac{\beta^k}{k!} C_k + F_c, \qquad (8)$$

where $F^*(A) = -k_B T \ln p^*(A)$ is the modified free energy obtained from GaMD simulation and F_c is a constant.

Implementation of GaMD in OpenMM

In recent years, the OpenMM simulation engine[52] has been developed to enable fast and extensible MD simulations. OpenMM features a convenient API layer, which allows users to access OpenMM's functions from external programs, including code written in Python. OpenMM also possesses lower layers to make the most efficient use of CPU and GPU hardware capabilities.

Part of OpenMM's extensibility includes the built-in CustomIntegrator object, which allows developers to design integration algorithms from within the high-level API layer, not requiring them to delve into the complexities of the lower OpenMM code layers. The CustomIntegrator accepts a set of variables and instructions in the form of character strings. OpenMM passes these strings to a just-in-time compiler[55] to be converted to CPU or GPU platform code at runtime - enabling both highly efficient and highly customizable code. We used the CustomIntegrator to implement several variations of the GaMD algorithm within Python.

For GaMD in OpenMM, multiple modes are available for applying boost potential to biomolecules: (1) boosting the dihedral energetic term only, (2) boosting the total potential energy only, (3) boosting the non-bonded terms in the potential energy, and (4) boosting a combination of two of the aforementioned terms, called "dual-boost" (i.e., "total energy – dihedral energy dual boost"). The GaMD boost potential is computed based on statistics of the system as detailed in the previous section. In addition, both the "lower-bound" and "upper-bound" integration schemes are implemented in GaMD OpenMM.

GaMD simulations generally include three stages: (i) short cMD, (ii) GaMD equilibration and (iii) GaMD production. The program first collects potential statistics from a short cMD run. Subsequently, a boost potential is added to the system in the GaMD equilibration stage while updates of the potential statistics continue. After the equilibration stage, the statistics collected is assumed to be sufficient to represent the potential energy landscape of interest. Hence, the reference energy and harmonic force constant are fixed to calculate the boost potential for running the production simulation. Note that in both the cMD and equilibration stages, there are a small number of steps at the beginning of each stage during which we do not collect statistics. These steps, named preparation steps, are performed to allow the system to adapt to the simulation environment. The program starts collecting statistics of the potential energies after the preparation steps.

MD simulations frequently experience interruptions; therefore, it is helpful for simulation utilities to be able to easily restart incomplete simulations. This is accomplished in GaMD OpenMM by leveraging OpenMM's checkpoint utility - the exact state of the simulation, including all variables related to the GaMD portion of the simulation, is saved with a frequent interval in time. Therefore, if an interruption occurs to the simulation, the

GaMD OpenMM program can automatically recover the most recent checkpoint and continue the simulation from where it left off, regardless of which stage of the GaMD process it was in when the interruption occurred.

Our GaMD OpenMM package is open-source and available for download, along with documentation for installation and usage, as well as tutorials, at https://github.com/MiaoLab20/gamd-openmm.git.

Simulation Protocols and Benchmarks

For alanine dipeptide, the AMBER ff99SB force field parameters were used. The LEaP module in the AmberTools package[56-59] was used to build the simulation system for alanine dipeptide. The alanine dipeptide was solvated in a TIP3P[60] water box that extends ~8–10 Å from the solute surface. The final system contained 1912 atoms, with a total of 630 water molecules.

For GaMD simulations of RNA molecules, the AMBER RNA OL3[61] and GAFF2[62] force field parameters were used for the RNA and ligand molecules, respectively. The simulation systems of the UUCG, GCAA, and UUCG tetraloops were prepared starting from the 1F7Y[63], 1ZIH[64], and 1RNG[65] PDB structures, respectively. The PDB structures were solvated in octahedral TIP3P[60] water boxes that extended 12 Å from the RNA surfaces, with approximately 1M KCl added to the solutions by the LEaP module in the AmberTools package[57-59, 66]. The final systems of UUCG, GCAA, and CUUG tetraloops contained 7,805, 6,218, and 7,538 atoms, respectively. Starting from the 1UUD[67] PDB structure, the bound rbt203 ligand was removed from the HIV-1 Tar RNA and placed at a ~15 Å distance away from the RNA surface to prepare the simulation system for ligand binding to the HIV-1 Tar RNA. The RNA-ligand complex was then solvated in a cubical

TIP3P[60] water box that extended 15 Å from the solute surface by the CHARMM-GUI webserver[68-70]. The system charge was neutralized with 0.15 M NaCl and 0.01 M Mg^{2+}, which resulted in a final system size of 40,829 atoms. All RNA systems were simulated at a temperature of 300 K.

Periodic boundary conditions were applied for the simulation systems. Bonds containing hydrogen atoms were restrained with the SHAKE[71] algorithm and a 2*fs* timestep was used. Weak coupling to an external temperature and pressure bath was used to control both temperature and pressure[72]. The electrostatic interactions were calculated using the particle mesh Ewald (PME) summation[73] with a cutoff of 8.0 Å for long-range interactions. After the initial energy minimization and thermalization, dual-boost GaMD was applied to simulate the systems. The system threshold energy for applying the boost potential was set V_{max}. The default parameter values were used for the GaMD simulations except stated otherwise. For alanine dipeptide, three independent simulations were performed with randomized initial atomic velocities, each of which consisted of 2ns short cMD, followed by 4ns GaMD equilibration and then 30 ns GaMD production simulation. After collecting the statistics, the threshold energy E and harmonic force constant k were computed according to **equation (3)**.

For the simulations of RNA tetraloops, three independent dual-boost GaMD simulations were performed for each system, each of which consisted of 2 ns cMD, 8 ns GaMD equilibration after adding the boost potential and 3,000-5,000 ns GaMD production. The σ_{0V} values were lowered to 1.5 kcal/mol from the default 6.0 kcal/mol to observe semi-stable refolding of the RNA tetraloops. For the simulations of the rbt203 ligand binding to the HIV-1 Tar RNA, five independent dual-boost GaMD simulations were performed, each

of which included 1.6 ns cMD, 6.4 ns GaMD equilibration after adding the boost potential and 400-500 ns GaMD production. GaMD simulation frames were saved every 0.1 ps. The VMD[74] and CPPTRAJ[75] tools were used for simulation trajectory analysis. Finally, the PyReweighting toolkit[32] was applied to compute the potential of mean force (PMF) profiles of the backbone dihedrals Φ and Ψ in the alanine dipeptide. The heavy-atom RMSD of RNA tetraloops relative to respective PDB structures (1F7Y for UUCG, 1ZIH for GCAA, and 1RNG for CUUG) and the U3-G6, G3-A6, and C3-G6 center-of-mass (COM) distances were used as RCs to calculate the PMF profiles in the RNA tetraloop simulations. The COM distance between the rbt203 ligand and nucleotide A6 and the COM distance between RNA nucleotides A6 and U7 side chains were selected to compute the PMF profiles in the simulations of ligand binding to the HIV-1 Tar RNA.

Results

Free Energy Profiles of Alanine Dipeptide

For alanine dipeptide, outputs from the dual-boost GaMD simulations were used to compute free energy profiles of the Φ and Ψ dihedrals (**Figure 1A**). The boost potential from three independent 30 ns GaMD production simulations was 9.4 ± 2.7 kcal/mol. The 2nd order cumulant expansion was applied to energetically reweight the GaMD simulations.

The 1D free energy profiles obtained from three 30 ns GaMD simulations agreed quantitatively with the PMF profiles from the 1000 ns cMD simulations (**Figures 1B-E** and **S1**). For Φ, moderate fluctuations were observed near the energy barrier at 0°, and the free energy value increased slightly at ~50° (**Figures 1B** and **S1A-B**). The 2D free energy profiles of backbone dihedrals (Φ, Ψ) in 3 × 30ns GaMD simulations and 3 × 1000ns cMD

simulations are shown in **Figure 1D** and **1E**. Overall, GaMD in OpenMM was able to identify five low-energy conformational states, which centered around (-148°, 0°) and (-69°, -17°) for the right-handed α helix (α_R), (48°, -12°) for the left-handed α helix (α_L), (-150°, 159°) for the β-sheet and (-72°, 162°) for the polyproline II (P_{II}) conformation. The corresponding minimum free energies were approximately 0, 0.74, 3.15, 1.68, and 2.65 kcal/mol. The 2D free energy profile obtained from the GaMD and cMD simulations showed a high degree of similarity (**Figure 1E**).

Folding of the RNA tetraloop structures: UUCG, GCAA, and CUUG

Multiple independent dual-boost GaMD simulations were performed on three RNA tetraloops structures of UUCG (PDB: 1F7Y) [63], GCAA (PDB: 1ZIH) [64], and CUUG (PDB: 1RNG)[65]. Similar averages and standard deviations of the added boost potentials were recorded for the systems, i.e., 9.3 ± 2.5 kcal/mol for UUCG, 10.1 ± 3.1 kcal/mol for GCAA, and 9.0 ± 2.9 kcal/mol for CUUG. Starting from the folded structures, GaMD simulations captured multiple unfolding and semi-stable folding events. A folding event was defined as attaining < 4Å heavy-atom RMSD relative to respective PDB structures[76] of the three RNA tetraloops for more than ~10 ns (**Figures S2-S4**). The 2D PMF free energy profiles were calculated from the respective heavy-atom RMSD to the PDB structures and COM distances between first and last residues of the RNA tetraloops to characterize their folding processes.

The 2D PMF free energy profile of the UUCG folding was calculated from the heavy-atom RMSD of the RNA tetraloop relative to the 1F7Y PDB structure and the distance between nucleotides U3 and G6 (**Figures 2A** and **S2**). Four low-energy

conformational states, including "Folded" (**Figure 2B**), "I1" (**Figure 2C**), "I2" (**Figure 2D**), and "Unfolded" (**Figure 2E**), were uncovered from the GaMD simulations of the UUCG tetraloop. All the low-energy conformational states were compared to the nuclear magnetic resonance (NMR) structure of folded UUCG (PDB: 1F7Y)[63] (**Figure 2**). In the "Folded" low-energy conformation, nucleotides U3 and G6 flipped in and formed hydrogen bonds with one another, and nucleotide U3 base stacked with nucleotide C5. The heavy-atom RMSD of UUCG relative to the 1F7Y PDB was ~1.1 Å, and the COM distance between nucleotides U3 and G6 was ~9.8 Å (**Figure 2B**). The COM distance between nucleotides U3 and G6 increased to ~14.2 Å, and the heavy-atom RMSD increased to ~4.1 Å in the "I1" low-energy conformational state. Nucleotide G6 flipped out, whereas nucleotide C5 flipped in to interact with nucleotide U3 (**Figure 2C**). The heavy-atom RMSD in the "I2" low-energy conformational state was similar to the "I1" low-energy conformational state, although the RNA backbone distorted heavily, which decreased the distance between nucleotides U3 and G6 to ~7.3 Å (**Figure 2D**). In the "Unfolded" low-energy conformational state, the heavy-atom RMSD relative to the 1F7Y PDB structure was ~5.8 Å, and the U3-G6 distance was ~8.0 Å (**Figure 2E**).

For the GCAA RNA tetraloop, four low-energy conformational states were identified from 2D PMF calculated from the GaMD simulations (**Figures 3A** and **S3**), including "Folded" (**Figure 3B**), "I1" (**Figure 3C**), "I2" (**Figure 3D**), and "Unfolded" (**Figure 3E**). They were compared to the MMR structure of folded GCAA (PDB: 1ZIH) [64] in **Figure 3**. In the "Folded" state, the side chains of nucleotides C4-A6 were base stacked and located on the opposite side of nucleotide G3. The heavy-atom RMSD relative to the 1ZIH PDB structure was ~1.0 Å, and the COM distance between nucleotides G3 and A6

was ~9.3 Å (**Figure 3A**). In the "I1" state, the side chain of nucleotide G3 flipped to the same side of nucleotides U4-A6, and the base stacking only existed between nucleotides A5 and A6. The heavy-atom RMSD was ~2.6 Å, and the distance between nucleotides G3 and A6 was ~8.8 Å (**Figure 3B**). In the "I2" state, the base stacking between nucleotides A5 and A6 remained stable. The heavy-atom RMSD relative to the 1ZIH PDB structure was ~3.9 Å, and the G3-A6 distance was ~10.5 Å (**Figure 3C**). In the "Unfolded" state, nucleotide C4 flipped out, while nucleotide G3 formed base stacking with nucleotides A5 and A6. The heavy-atom RMSD was ~4.6 Å, and the G3-A6 distance was ~7.6 Å (**Figure 3D**).

For the CUUG RNA tetraloop, three distinct low-energy conformational states were identified from the 2D PMF (**Figures 4A** and **S4**), namely "Folded" (**Figure 4B**), "I1" (**Figure 4C**), and "Unfolded" (**Figure 4D**). They were also compared to the MMR structure of folded CUUG (PDB: 1RNG) [65] (**Figure 4**). In the "Folded" state, nucleotides C3 and G6 flipped in and formed a hydrogen bond with one another. The heavy-atom RMSD relative to the 1RNG PDB structure was ~1.1 Å, and the COM distance between nucleotides C3 and G6 was ~10.8 Å (**Figure 4B**). The RNA backbone distorted in the "I1" state. The heavy-atom RMSD was ~3.9 Å, and the C3-G6 distance was ~6.9 Å (**Figure 4C**). In the "Unfolded" state, the heavy-atom RMSD relative to the 1RNG PDB structure was ~4.3 Å, and the C3-G6 distance was ~13.2 Å (**Figure 4D**).

Binding of the rbt203 ligand to the HIV-1 Tar RNA

Starting from the 1UUD[67] PDB structure, the bound rbt203 ligand was removed and placed at a ~15 Å distance from the HIV-1 Tar RNA. Five independent 400-500 ns

GaMD simulations captured multiple stable binding events of the rbt203 ligand to the HIV-1 Tar RNA (**Figure S5**). The average added boost potentials were recorded to be 9.1 ± 3.0 kcal/mol. The 2D PMF free energy profiles was calculated from the COM distance between the rbt203 ligand and nucleotide A6 side chain and the COM distance between nucleotides A6 and U7 side chains to characterize ligand binding to the HIV-1 Tar RNA (**Figures 5** and **S6**).

Six low-energy conformational states were uncovered from the GaMD simulations of ligand binding to HIV-1 Tar RNA, including "B1", "B2", "I1", "I2", "I3", and "U". The "B1" and "B2" low-energy conformational states represented the bound conformation of rbt203 in the HIV-1 Tar RNA, while the "U" low-energy conformation state represented the unbound conformation. In the "B1" low-energy conformational state, the distance between rbt203 ligand and nucleotide A6 was ~7.1 Å, and the distance between nucleotides A6 and U7 was ~3.8 Å (**Figure 5A**). Nucleotide U7 flipped in and pointed towards the core of the HIV-1 Tar RNA. The rbt203 ligand interacted with nucleotides A6, U7, C8, U9, G10, C23, and U24 in this low-energy conformational state (**Figure 5A**). In the "B2" low-energy conformational state, the distance between rbt203 ligand and nucleotide A6 was ~9.1 Å, and the distance between nucleotides A6 and U7 was ~8.1 Å (**Figure 5B**). Similar to the "B1" low-energy conformational state, nucleotide U7 also pointed towards the ligand. The rbt203 ligand interacted with nucleotides G5, A6, U7, G10, A11, G12, C13, A19, G20, C21, U22, C23, and U24 (**Figure 5C**). In the "I1" low-energy conformational state, the A6-rbt203 ligand distance was ~20.1 Å, and the A6-U7 distance was ~4.1 Å (**Figure 5D**). The rbt203 ligand was located at the terminal nucleotides of HIV-1 Tar RNA. The interacting nucleotides with rbt203 ligand were G1, G2, C3, C23, U24, G27, C28, and

C29 (**Figure 5D**). In the "I2" low-energy conformational state, the A6-rbt203 ligand distance was ~20.0 Å, and the A6-U7 distance was ~12.4 Å (**Figure 5E**). The rbt203 ligand was at a similar location as in the "I1" low-energy conformational state. The interacting nucleotides in the "I2" low-energy conformational state were G1, G2, C3, A4, U26, G27, C28, and C29 (**Figure 5E**). In the "I3" low-energy conformational state, the A6-rbt203 ligand distance was ~34.1 Å, and the A6-U7 distance was ~12.2 Å (**Figure 5F**). The rbt203 ligand was located at the U15-G18 RNA tetraloop. The rbt203 ligand interacted with nucleotides U15, G16, and G17 of the HIV-1 Tar RNA (**Figure 5F**). In the "U" low-energy conformational state, the rbt203 ligand was found in the bulk solvent, and nucleotide U7 flipped outwards. The distance between nucleotide A6 and rbt203 ligand was ~41.8 Å, and the distance between nucleotides A6 and U7 was ~12.7 Å in this low-energy conformational state (**Figure 5G**).

Discussion

By adding a harmonic boost potential to smoothen the potential energy surface, GaMD provides both unconstrained enhanced sampling and free energy calculation of biomolecules. Important statistical properties of the system potential, such as the average, maximum, minimum and standard deviation values, are used to calculate the simulation acceleration parameters, particularly the threshold energy and force constant for applying the boost potential. In this study, we have implemented GaMD in the OpenMM package. "Selective GaMD" algorithms, including Ligand GaMD, Peptide GaMD and Protein-

Protein Interaction GaMD, have not been implemented in OpenMM, although they are planned to be implemented in the future.

Three independent 30 ns GaMD simulations were able to capture five different low-energy conformational states of the backbone dihedrals (Φ, Ψ) in alanine dipeptide, which were in good agreement with the cMD simulations (**Figure 1D,E**). In addition, the 1D free energy profiles of GaMD and cMD mostly overlapped, except the elevated free energy value at ~50° for Φ and minor fluctuations in the energy barriers at 0° for Φ and −120° for Ψ (**Figures 1** and **S1**). Notably, both the 1D and 2D free energy profiles of GaMD in OpenMM were highly similar to those from previous implementations of GaMD in AMBER[31] and NAMD[35] in terms of the low-energy states and free energy profiles. The alanine dipeptide system provides a sort of benchmark or validation of the correctness of the GaMD approach and any of its implementations. Our present results show that GaMD in OpenMM can reproduce the correct free energy profiles for alanine dipeptide, as we have shown for previous implementations of GaMD[31, 35, 48, 51], providing evidence that we have completed a correct implementation of GaMD in OpenMM, providing users confidence in applying GaMD OpenMM for their own systems of interest.

GaMD in OpenMM successfully captured the unfolding and semi-stable refolding of three hyperstable RNA tetraloops of UUCG, GCAA, and CUUG[76]. The low-energy conformational states obtained illustrated the unfolding pathways of the three hyperstable tetraloops, which were mostly the reverses of the folding pathways uncovered by Chen et. al[76]. For UUCG, starting from the "Folded" low-energy conformational state where the two nucleotides U3 and G6 pointed inwards and interacted with each other (**Figure 2B**), the backbone of the UUCG tetraloop skewed to the right as the nucleotide C5 flipped in,

pushed G6 outwards, and formed interactions with U3 in the "I1" low-energy conformation (**Figure 2C**). As UUCG transited from the "I1" to "I2" conformation, the U3-C5 interaction was broken, and both nucleotides flipped outwards. The RNA core was solely occupied by nucleotide G6, heavily distorting the tetraloop (**Figure 2D**). Finally, the RNA stretched out and became unfolded in the "Unfolded" low-energy conformational state (**Figure 2E**). For GCAA, the unfolding pathway started with the "Folded" low-energy conformation where base stacking was observed between the three nucleotides C4-A6, and only G3 pointed inwards (**Figure 3B**). As GCAA transited from the "Folded" to "I1" conformation, the stacking between nucleotides C4 and A5 was broken, while the tetraloop shrunk in size (**Figure 3C**). The base stacking between nucleotides A5 and A6 remained stable in the "I2" low-energy conformational state, as the RNA began stretching out and nucleotide C4 flipped to the opposite side (**Figure 3D**). Finally, GCAA stretched out and coiled into the "Unfolded" low-energy conformation, where nucleotide G3 flipped outwards and formed base stacking with nucleotide A5, which in turn remained base-stacked with nucleotide A6 (**Figure 3E**). For CUUG, nucleotides C3 and G6 pointed inwards and formed hydrogen bonds with each other in the "Folded" low-energy conformational state (**Figure 4B**). As CUUG transited from the "Folded" to "I1" low-energy conformation, all four nucleotides in the tetraloop pointed outwards as the RNA skewed left and shrunk in its size significantly (**Figure 4C**). To completely unfold, CUUG straightened out its terminal nucleotides and became stretched out in the "Unfolded" low-energy conformation (**Figure 4D**). Overall, the low-energy conformational states and unfolding pathways uncovered from GaMD simulations in OpenMM agreed well with a previous study carried out by Chen et al.[76], particularly the "Folded" conformations, "I1"

of UUCG, “I2” of GCAA, and “I1” of CUUG. Nevertheless, it is also worth noting that GaMD in OpenMM was only able to capture semi-stable refolding events of all three RNA tetraloops, where the heavy-atom RMSD relative to respective PDB structures were in the range of ~1.8-2.5 Å (**Figures S2-S4**). This was primarily because the RNA force field parameters were biased to favor rigid, highly stacked conformations, as described in the previous study[76]. The independent GaMD simulations of each RNA tetraloop have not achieved proper convergences within the 4-5μs simulation time windows as indicated by the different free energy profiles across the simulations of each tetraloop (**Figures S2-S4**). Longer GaMD simulations combined with more accurate RNA force field parameter sets are required to achieve consistent simulations of RNA.

In the 1UUD[67] PDB structure of HIV-1 Tar RNA, the distance between nucleotide A6 and rbt203 ligand is ~8.9 Å, and the distance between nucleotides A6 and U7 is ~7.4 Å. Nucleotide U7 points towards the core of the HIV-1 Tar RNA, and the rbt203 ligand interacts with nucleotides A6, U7, U9, G10, A11, G12, C13, A19, G20, C21, U22, C23, and U24. The distance between nucleotide A6 and rbt203 ligand in the 1UUD PDB structure is comparable to those in the “B1” and “B2” low-energy conformational states, while the distance between nucleotides A6 and U7 is the middle between those in the “B1” and “B2” low-energy conformational states (**Figure 5B,C**). The interacting nucleotides of the rbt203 ligand are highly similar between the GaMD-bound conformations and the 1UUD PDB structure, demonstrating the agreements between GaMD simulation results and experimental data[67].

One recent study by Tang et al.[36] demonstrated that base stacking between ligands and nucleotides is the key interaction that drives ligand binding in single-stranded nucleic

acids. Furthermore, Chen et al.[76] found that preformed G1-A4 and C1-G4 base pairs played a significant role in the accurate folding of the GCAA and CUUG RNA tetraloops.

In addition, we observed that nucleotide U7 flips inwards and points towards the core of the HIV-1 Tar RNA in both the bound "B1" and "B2" low-energy conformations (**Figure 5B,C**) and the 1UUD[67] PDB structure, while flips outwards in the unbound "U" low-energy conformation (**Figure 5G**). The observation of nucleotide U7 "base-flipping"[77] phenomenon during ligand binding illustrated the importance of this nucleotide in the ligand binding to the HIV-1 Tar RNA. Furthermore, two slightly different binding pathways the rbt203 ligand to the HIV-1 Tar RNA could be observed from the free energy profile in **Figure 5**. While both pathways started from the "U", "I3", and "I2" low-energy conformational states, the second pathway arrived abruptly at the bound "B2" low-energy conformational state, whereas the dominant pathway involved a stabilization of the intermediate state as indicated by the transition from "I2" to "I1", before ending at the bound "B1" low-energy conformation (**Figures 5** and **S5**). The dominant pathway is described in detail as follow. Starting from the bulk solvent (**Figure 5G**), the rbt203 ligand approached the HIV-1 Tar RNA first through interactions with the U15-G18 tetraloop (**Figure 5F**). The rbt203 ligand then dissociated back to the bulk solvent and relocated to the terminal nucleotides of the HIV-1 Tar RNA (**Figure 5E**). At this stage, nucleotide U7 flipped inwards and became ready to interact with the rbt203 ligand (**Figure 5D**). Finally, the rbt203 ligand moved from the terminal of HIV-1 Tar RNA to its binding pocket, located at the core nucleotides of the RNA (**Figures 5B** and **5C**). The ligand drew closer to nucleotide A6 in the "B2" low-energy conformational state (**Figure 5C**) and nucleotide U7 in the "B1" low-energy conformational state (**Figure 5B**). Given the fact that the RNA

conformational changes took place after ligand bound, the binding of rbt203 ligand to the HIV-1 Tar RNA is an induced-fit process. On the other hand, it is also worth noting that similar to the RNA folding simulations, GaMD simulations of RNA-ligand binding has not converged within the 500ns simulation time windows as shown by the different free energy profiles calculated from the individual simulations (**Figure S6**). Furthermore, as mentioned above, more accurate RNA force field parameter sets are required to achieve consistent simulations of RNA molecules.

In summary, we have implemented GaMD in OpenMM. It is complementary to previous implementations of GaMD in AMBER[31], NAMD[35], GENESIS[48], and TINKER-HP[51]. As demonstrated on model systems, results of the current work will facilitate the applications of GaMD in enhanced sampling and free energy calculations of a wide range of large biomolecules, especially RNA structures that involve critical electrostatic interactions.

ASSOCIATED CONTENT

Supporting Information

The implementation algorithm and example simulation input file of GaMD in OpenMM and six supporting figures (Figures S1-S7) are available free of charge on the ACS Publications website.

Notes

The authors declare no competing financial interest.

Acknowledgements

We would like to thank Peter Eastman for his help and support with the GaMD implementation of OpenMM. This work used supercomputing resources with allocation award TG-MCB180049 through the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (NSF) grant number ACI-1548562, and project M2874 through the National Energy Research Scientific Computing Center (NERSC), which is a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231, and the Research Computing Cluster at the University of Kansas. Y. Miao acknowledges support by NIH (grant R01GM132572), NSF (grant DBI2121063) and the startup funding in the College of Liberal Arts and Sciences at the University of Kansas.

References

(1) Karplus, M.; McCammon, J. A., Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **2002,** *9* (9), 646-652.
(2) Hollingsworth, S.; Dror, R., Molecular dynamics simulation for all. *Neuron* **2018,** *99*, 1129-43.
(3) Wang, J.; Arantes, P.; Bhattarai, A.; Hsu, R.; Pawnikar, S.; Huang, Y.-m.; Palermo, G.; Miao, Y., Gaussian accelerated molecular dynamics: principles and applications. *WIREs Computational Molecular Science* **2021**, e1521.
(4) Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007,** *450*, 964-72.
(5) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D., ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *Journal of Chemical Theory and Computation* **2009,** *5*, 1632-9.
(6) Johnston, J. M.; Filizola, M., Showcasing modern molecular dynamics simulations of membrane proteins through G protein-coupled receptors. *Current Opinions in Structural Biology* **2011,** *21*, 552-8.
(7) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y., et al., Atomic-level characterization of the structural dynamics of proteins. *Science* **2010,** *330*, 341-6.

(8) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S., To milliseconds and beyond: challenges in the simulation of protein folding. *Current Opinions in Structural Biology* **2013,** *23*, 58-65.
(9) Vilardaga, J.-P.; Bünemann, M.; Krasel, C.; Castro, M.; Lohse, M. J., Measurement of the milisecond activation switch of G protein-coupled receptors in living cells. *Nature Biotechnology* **2008,** *21*, 807-12.
(10) Miao, Y.; Ortoleva, P. J., Viral structural transitions: an all-atom multiscale theory. *Journal of Chemical Physics* **2006,** *125*, 214901.
(11) Spiwok, V.; Sucur, Z.; Hosek, P., Enhanced sampling techniques in biomolecular simulations. *Biotechnology Advances* **2015,** *33*, 1130-40.
(12) Gao, Y. Q.; Yang, L. J.; Fan, Y. B.; Shao, Q., Thermodynamics and kinetics simulations of multi-timescale processes for complex systems. *International Reviews in Physical Chemistry* **2008,** *27*, 201-227.
(13) Liwo, A.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A., Computational techniques for efficient conformational sampling of proteins. *Current Opinion in Structural Biology* **2008,** *18*, 134-139.
(14) Christen, M.; van Gunstere, W., On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: a review. *Journal of Computational Chemistry* **2008,** *29*, 157-66.
(15) Miao, Y.; McCammon, J. A., Unconstrained enhanced sampling for free energy calculations of biomolecules: a review. *Molecular Simulation* **2016,** *42*, 1046-55.
(16) Torrie, G.; Valleau, J., Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *Journal of Computational Physics* **1977,** *23*, 187-199.
(17) Kumar, S.; Rosenberg, J.; Bouzida, D.; Swendsen, R.; Kollman, P., THE weighted histogram analysis method for free-energy calculations on biomolecules. I. THE method. *Journal of Computational Chemistry* **1992,** *13*, 1011-21.
(18) Laio, A.; Gervasio, F., Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **2008,** *71*, 126601.
(19) Besker, N.; Gervasio, F., Using metadynamics and path collective variables to study ligand binding and induced conformational transitions. In *Computational drug discovery and design*, Berlin: Springer: 2012; pp 501-13.
(20) Darve, E.; Rodriguez-Gomez, D.; Pohorille, A., Adaptive biasing force method for scalar and vector free energy calculations. *Journal of Chemical Physics* **2008,** *128*, 144120.
(21) Darve, E.; Wilson, M.; Pohorille, A., Calculating free energies using a scaled-force molecular dynamics algorithm. *Molecular Simulation* **2002,** *28*, 113-44.
(22) Isralewitz, B.; Baudry, J.; Gullingsrud, J.; Kosztin, D.; Schulten, K., Steered molecular dynamics investigations of protein function. *Journal of Molecular Graphics and Modelling* **2001,** *19*, 13-25.
(23) Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999,** *314*, 141-51.
(24) Okamoto, Y., Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *Journal of Molecular Graphics and Modelling* **2004,** *22*, 425-39.

(25) Hansmann, U., Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters* **1997,** *281*, 140-50.
(26) Wu, X.; Brooks, B., Self-guided Langevin dynamics simulation method. *Chemical Physics Letters* **2003,** *381*, 512-8.
(27) Wu, X.; Brooks, B.; Vanden-Eijnden, E., Self-guided Langevin dynamics via generalized Langevin equation. *Journal of Computational Chemistry* **2016,** *37*, 595-601.
(28) Wu, X.; Wang, S., Self-guided molecular dynamics simulation for efficient conformational search. *The Journal of Physical Chemistry B* **1998,** *102*, 7238-50.
(29) Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *Journal of Chemical Physics* **2004,** *120*, 11919-11929.
(30) Voter, A. F., Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events. *Physical Review Letters* **1997,** *78*, 3908.
(31) Miao, Y.; Feher, V. A.; McCammon, J. A., Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. *Journal of Chemical Theory and Computation* **2015,** *11*, 3584-3595.
(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. *Journal of Chemical Theory and Computation* **2014,** *10*, 2677-2689.
(33) Do, H.; Akhter, S.; Miao, Y., Pathways and Mechanism of Caffeine Binding to Human Adenosine A2A Receptor. *Frontiers in Molecular Biosciences* **2021,** *8*, 242.
(34) Bhattarai, A.; Pawnikar, S.; Miao, Y., Mechanism of ligand recognition by human ACE2 receptor. *Journal of Physical Chemistry Letters* **2021,** *12*, 4814-4822.
(35) Pang, Y.; Miao, Y.; McCammon, J. A., Gaussian accelerated molecular dynamics in NAMD. *Journal of Chemical Theory and Computation* **2017,** *13*, 9-19.
(36) Tang, Z.; Akhter, S.; Ramprasad, A.; Wang, X.; Reibarkh, M.; Wang, J.; Aryal, S.; Thota, S.; Zhao, J.; Douglas, J., et al., Recognition of single-stranded nucleic acids by small-molecule splicing modulators. *Nucleic Acids Research* **2021,** *49* (14), 7870-7883.
(37) Do, H.; Wang, J.; Bhattarai, A.; Miao, Y., GLOW: a workflow that integrates Gaussian accelerated molecular dynamics and Deep Learning for free energy profiling. *Journal of Chemical Theory and Computation* **2022,** *18* (3), 1423-1436.
(38) Bhattarai, A.; Devkota, S.; Bhattarai, S.; Wolfe, M. S.; Miao, Y., Mechanisms of gamma-secretase activation and substrate processing. *ACS Central Science* **2020,** *6* (6), 969-983.
(39) Bhattarai, A.; Devkota, S.; Do, H.; Wang, J.; Bhattarai, S.; Wolfe, M.; Miao, Y., Mechanism of Tripeptide Trimming of Amyloid beta-Peptide 49 by gamma-Secretase. *Journal of American Chemical Society* **2022,** *144*, 6215-6226.
(40) Miao, Y.; McCammon, J. A., Graded activation and free energy landscapes of a muscarinic G-protein-coupled receptor. *Proceedings of the National Academy of Sciences of the United States of America* **2016,** *113*, 12162-12167.
(41) Bhattarai, A.; Wang, J.; Miao, Y., G-protein-coupled receptor-membrane interactions depend on the receptor activation state. *Journal of Computational Chemistry* **2020,** *41*, 460-471.
(42) Miao, Y.; McCammon, J. A., Mechanism of the G-protein mimetic nanobody binding to a muscarinic G-protein-coupled receptor. *Proceedings of the National Academy of Sciences of the United States of America* **2018,** *115*, 3036-3041.

(43) Wang, J.; Miao, Y., Mechanistic insights into specific G protein interactions with adenosine receptors. *The Journal of Physical Chemistry B* **2019,** *123*, 6462-6473.
(44) Wang, J.; Miao, Y., Protein-Protein Interaction-Gaussian Accelerated Molecular Dynamics (PPI-GaMD): Characterization of Protein Binding Thermodynamics and Kinetics. *Journal of Chemical Theory and Computation* **2022,** *18* (3), 1275-1285.
(45) East, K. W.; Newton, J. C.; Morzan, U. N.; Narkhede, Y. B.; Acharya, A.; Skeens, E.; Jogl, G.; Batista, V. S.; Palermo, G.; Lisi, G. P., Allosteric Motions of the CRISPR-Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics. *Journal of American Chemical Society* **2020,** *142* (3), 1348-1358.
(46) Ricci, C. G.; Chen, J. S.; Miao, Y.; Jinek, M.; Doudna, J. A.; McCammon, J. A.; Palermo, G., Deciphering Off-Target Effects in CRISPR-Cas9 through Accelerated Molecular Dynamics. *ACS Central Science* **2019,** *5* (4), 651-662.
(47) Huang, Y.-m.; McCammon, J. A.; Miao, Y., Replica exchange Gaussian accelerated molecular dynamics: improved enhanced sampling and free energy calculation. *Journal of Chemical Theory and Computation* **2018,** *14*, 1853-64.
(48) Oshima, H.; Re, S.; Sugita, Y., Replica-exchange umbrella sampling combined with Gaussian accelerated molecular dynamics for free-energy calculation of biomolecues *Journal of Chemical Theory and Computation* **2019,** *15*, 5199-208.
(49) Miao, Y.; Bhattarai, A.; Wang, J., Ligand Gaussian accelerated molecular dynamics (LiGaMD): characterization of ligand binding thermodynamics and kinetics. *Journal of Chemical Theory and Computation* **2020,** *16*, 5526-47.
(50) Wang, J.; Miao, Y., Peptide Gaussian accelerated molecular dynamics (Pep-GaMD): enhanced sampling and free energy and kinetics calculations of peptide binding. *Journal of Chemical Physics* **2020,** *153*, 154109.
(51) Celerse, F.; Inizan, T. J.; Lagardere, L.; Adjoua, O.; Monmarche, P.; Miao, Y.; Derat, E.; Piquemal, J.-P., An Efficient Gaussian-Accelerated Molecular Dynamics (GaMD) Multilevel Enhanced Sampling Strategy: Application to Polarizable Force Fields Simulations of Large Biological Systems. *Journal of Chemical Theory and Computation* **2022,** *18* (2), 968-977.
(52) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D., et al., OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation* **2013,** *9*, 461-469.
(53) Lindert, S.; Bucher, D.; Eastman, P.; Pande, V. S.; McCammon, J. A., Accelerated Molecular Dynamics Simulations with the AMOEBA Polarizable Force Field on Graphics Processing Units. *Journal of Chemical Theory and Computation* **2013,** *9* (11), 4684-4691.
(54) Miao, Y.; McCammon, J. A., Gaussian Accelerated Molecular Dynamics: Theory, Implementation and Applications. *Annu Rep Comp Chem* **2017,** *13*, 231-278.
(55) Kobalicek, P., AsmJit Project: Machine Code Generation for C++. [Online] https://asmjit.com/, 2018.
(56) Case, D. A.; Belfon, K.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E.; Cruzeiro, V. W. D.; Darden, T.; Duke, R. E.; Giambasu, G., AMBER 2020. **2020**.
(57) Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. . *Journal of Chemical Theory and Computation* **2012,** *8* (5), 1542-1555.

(58) Salomon-Ferrer, R.; Gotz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C., Routined microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **2013,** *9*, 3878-3888.
(59) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C., An overview of the Amber biomolecular simulation package. *Wiley Interdiscip Rev Comput Mol Sci* **2013,** *3* (2), 198-210.
(60) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M., Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **1983,** *79*, 926-935.
(61) Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, B.; Cheatham, T. E.; Jurecka, P., Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *Journal of Chemical Theory and Computation* **2011,** *7*, 2866-2902.
(62) He, X.; Man, V. H.; Yang, W.; Lee, T.-S.; Wang, J., A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics* **2020,** *153*, 114502.
(63) Ennifar, E.; Nikulin, A.; Tishchenko, S.; Serganov, A.; Nevskaya, N.; Garber, M.; Ehresmann, B.; Ehresmann, C.; Nikonov, S.; Dumas, P., The crystal structure of UUCG tetraloop. *Journal of Molecular Biology* **2000,** *304* (1), 35-42.
(64) Jucker, F. M.; Heus, H. A.; Yip, P. F.; Moors, E. H.; Pardi, A., A network of heterogeneous hydrogen bonds in GNRA tetraloops. *Journal of Molecular Biology* **1996,** *264* (5), 968-980.
(65) Jucker, F. M.; Pardi, A., Solution Structure of the CUUG Hairpin Loop: A Novel RNA Tetraloop Motif. *Biochemistry* **1995,** *34* (44), 14416-14427.
(66) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; T.E. Cheatham, I.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E., et al. *Amber 2021*, University of California, San Francisco: 2021.
(67) Davis, B.; Afshar, M.; Varani, G.; Murchie, A. I. H.; Karn, J.; Lentzen, G.; Drysdale, M. J.; Bower, J.; Potter, A. J.; Aboul-Ela, F., Rational Design of Inhibitors of HIV-1 Tar RNA Through the Stabilisation of Electrostatic "Hot Spots". *Journal of Molecular Biology* **2004,** *336* (3), 343-356.
(68) Jo, S.; Kim, T.; Iyer, V.; Im, W., CHARMM-GUI: A Web-based Graphical User Interface for CHARMM. *Journal of Computational Chemistry* **2008,** *29*, 1859-1865.
(69) Lee, J.; Cheng, X.; Swails, J.; Yeom, M.; Eastman, P.; Lemkul, J.; Wei, S.; Buckner, J.; Jeong, J.; Qi, Y., et al., CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation* **2016,** *12*, 405-413.
(70) Lee, J.; Hitzenberger, M.; M. Rieger; N.R. Kern; M. Zacharias; Im, W., CHARMM-GUI supports the Amber force fields. *Journal of Chemical Physics* **2020,** *153*, 035103.
(71) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977,** *23* (3), 327-341.

(72) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R., Molecular Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.
(73) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A Smooth Particle Mesh Ewald Method. *Journal of Chemical Physics* **1995,** *103* (19).
(74) Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *Journal of Molecular Graphics and Modelling* **1996,** *14*, 33-38.
(75) Roe, D. R.; Cheatham, I. T. E., PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation* **2013,** *9*, 3084-3095.
(76) Chen, A. A.; Garcia, A. E., High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *PNAS* **2013,** *110* (42), 16820-16825.
(77) Levintov, L.; Paul, S.; Vashisth, H., Reaction Coordinate and Thermodynamics of Base Flipping in RNA. *Journal of Chemical Theory and Computation* **2021,** *17*, 1914-1921.

Figure Captions

Figure 1. (A) Schematic representation of backbone dihedrals Φ and Ψ in alanine dipeptide. **(B-C)** Potential of mean force (PMF) profiles of the (B) Φ and (C) Ψ dihedrals calculated from three 30 ns GaMD simulations combined using cumulant expansion to the 2nd order. **(D)** The 2D PMF profile of backbone dihedrals (Φ, Ψ) from combined three 30ns GaMD simulations trajectories. The low energy wells are labeled corresponding to the right-handed α helix (α_R), left-handed α helix (α_L), β-sheet (β) and polyproline II (P_{II}) conformations. **(E)** The 2D PMF profile of backbone dihedrals (Φ, Ψ) from combined three 1000 ns cMD simulations trajectories. The low energy wells are labeled corresponding to the right-handed α helix (αR), left-handed α helix (αL), β-sheet (β) and polyproline II (PII) conformations.

Figure 2. Folding of the UUCG RNA tetraloop captured by GaMD in OpenMM. (A) 2D free energy profile of the heavy-atom RMSD of UUCG relative to the 1F7Y PDB structure and the center of mass (COM) distance between nucleotides U3 and G6. The low-energy RNA conformational states are labeled "Folded", "I1", "I2", and "Unfolded". **(B)** The "Folded" low-energy conformational state compared to the 1F7Y PDB structure, for which the RMSD is ~1.1 Å and the U3-G6 distance is ~9.8 Å. **(C)** The "I1" low-energy conformational state compared to the 1F7Y PDB structure, for which the RMSD is ~4.1 Å and the U3-G6 distance is ~14.2 Å. **(D)** The "I2" low-energy conformational state compared to the 1F7Y PDB structure, for which the RMSD is ~4.2 Å and the U3-G6 distance is ~7.3 Å. **(E)** The "Unfolded" low-energy conformational state compared to the 1F7Y PDB structure, for which the RMSD is ~5.8 Å and the U3-G6 distance is ~8.0 Å.

The low-energy RNA conformations are colored orange, cyan, magenta, and yellow, and the 1F7Y PDB structure is colored gray.

Figure 3. Folding of the GCAA RNA tetraloop captured by GaMD in OpenMM. (A) 2D free energy profile of the heavy-atom RMSD of GCAA relative to the 1ZIH PDB structure and the COM distance between nucleotides G3 and A6. The low-energy RNA conformational states are labeled "Folded", "I1", "I2", and "Unfolded". **(B)** The "Folded" low-energy conformational state compared to the 1ZIH PDB structure, for which the RMSD is ~1.0 Å and the G3-A6 distance is ~9.0 Å. **(C)** The "I1" low-energy conformational state compared to the 1ZIH PDB structure, for which the RMSD is ~2.6 Å and the G3-A6 distance is ~8.8 Å. **(D)** The "I2" low-energy conformational state compared to the 1ZIH PDB structure, for which the RMSD is ~3.9 Å and the G3-A6 distance is ~11.0 Å. **(E)** The "Unfolded" low-energy conformational state compared to the 1ZIH PDB structure, for which the RMSD is ~4.5 Å and the G3-A6 distance is ~8.0 Å. The low-energy RNA conformations are colored orange, cyan, magenta, and yellow, and the 1ZIH PDB structure is colored gray.

Figure 4. Folding of the CUUG RNA tetraloop captured by GaMD in OpenMM. (A) 2D free energy profile of the heavy-atom RMSD of CUUG relative to the 1RNG PDB structure and the COM distance between nucleotides C3 and G6. The low-energy RNA conformational states are labeled "Folded", "I1", and "Unfolded". **(B)** The "Folded" low-energy conformational state compared to the 1RNG PDB structure, for which the RMSD is ~1.1 Å and the C3-G6 distance is ~10.9 Å. **(C)** The "I1" low-energy conformational state compared to the 1RNG PDB structure, for which the RMSD is ~3.9 Å and the C3-G6 distance is ~6.9 Å. **(D)** The "Unfolded" low-energy conformational state compared to the

1RNG PDB structure, for which the RMSD is ~4.1 Å and the C3-G6 distance is ~13.1 Å. The low-energy RNA conformations are colored orange, cyan, magenta, and yellow, and the 1RNG PDB structure is colored gray.

Figure 5. Binding of the rbt203 ligand to the HIV-1 Tar RNA captured by GaMD in OpenMM. (A) 2D free energy profile of the COM distance between the rbt203 ligand (Lig) and RNA nucleotide A6 and the COM distance between RNA nucleotides A6 and U7 side chains. The low-energy conformational states are labeled "B1", "B2", "I1", "I2", "I3", and "U". **(B)** The "B1" low-energy conformational state, for which the A6-P14 ligand distance is ~8.0 Å and the A6-U7 distance is ~3.5 Å. **(C)** The "B2" low-energy conformational state, for which the A6-P14 ligand distance is ~10.1 Å and the A6-U7 distance is ~10.0 Å. **(D)** The "I1" low-energy conformational state, for which the A6-P14 ligand distance is ~20.1 Å and the A6-U7 distance is ~4.1 Å. **(E)** The "I2" low-energy conformational state, for which the A6-P14 ligand distance is ~20.0 Å and the A6-U7 distance is ~13.5 Å. **(F)** The "I3" low-energy conformational state, for which the A6-P14 ligand distance is ~41.8 Å and the A6-U7 distance is ~12.7 Å. **(G)** The "U" low-energy conformational state, for which the A6-P14 ligand distance is ~34.1 Å and the A6-U7 distance is ~12.2 Å. The low-energy RNA-ligand conformational states are colored orange, green, cyan, magenta, yellow, pink, and marine, and the 1UUD PDB is colored gray.

Figure 1

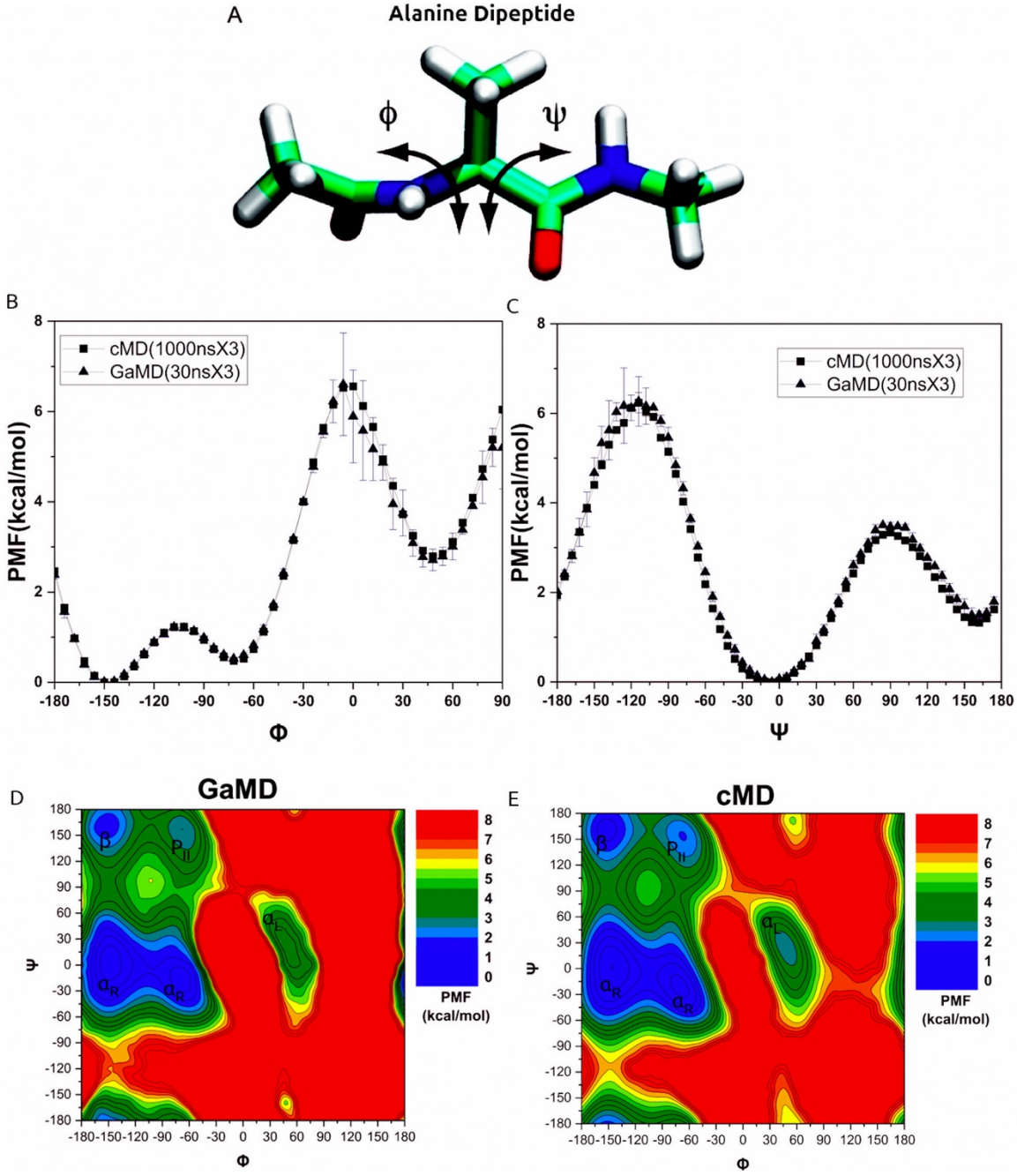

A
Alanine Dipeptide
ϕ
ψ
B
cMD(1000nsX3)
GaMD(30nsX3)
PMF(kcal/mol)
Φ
C
cMD(1000nsX3)
GaMD(30nsX3)
PMF(kcal/mol)
Ψ
D
GaMD
β
PII
αR
αR
αL
Ψ
Φ
PMF
(kcal/mol)
E
cMD
β
PII
αR
αR
αL
Ψ
Φ
PMF
(kcal/mol)

Figure 2

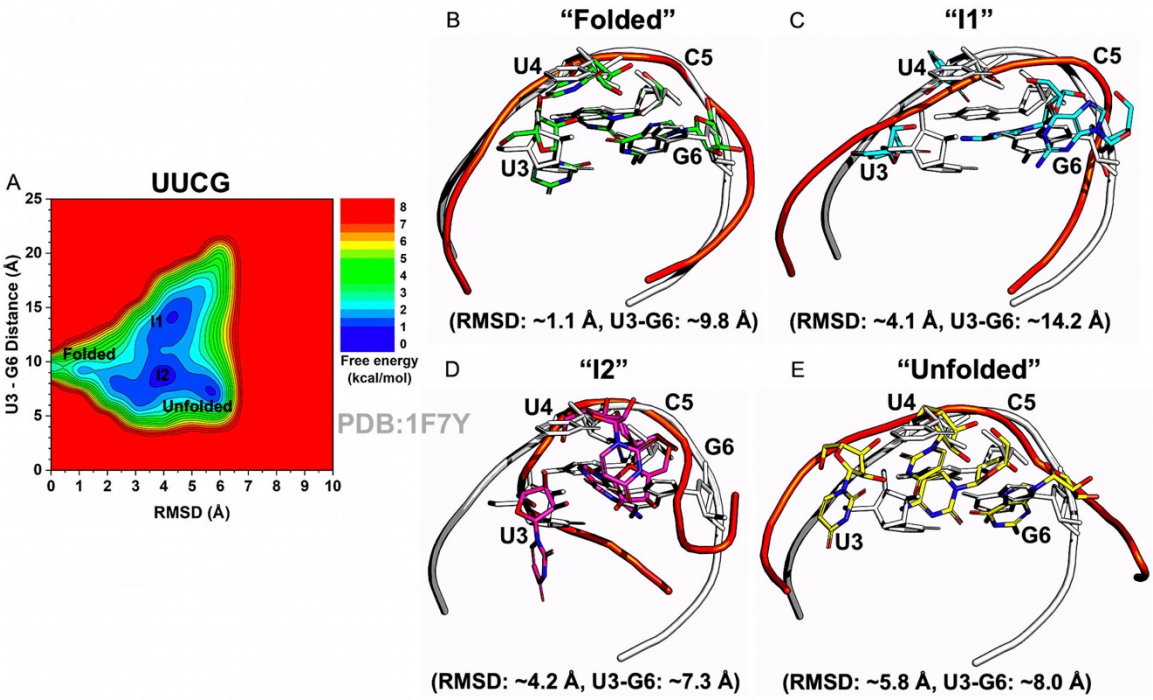

A
UUCG
U3 - G6 Distance (Å)
RMSD (Å)
Folded
I1
I2
Unfolded
Free energy (kcal/mol)
PDB:1F7Y
B
"Folded"
U4
C5
U3
G6
(RMSD: ~1.1 Å, U3-G6: ~9.8 Å)
C
"I1"
U4
C5
U3
G6
(RMSD: ~4.1 Å, U3-G6: ~14.2 Å)
D
"I2"
U4
C5
G6
U3
(RMSD: ~4.2 Å, U3-G6: ~7.3 Å)
E
"Unfolded"
U4
C5
U3
G6
(RMSD: ~5.8 Å, U3-G6: ~8.0 Å)

Figure 3

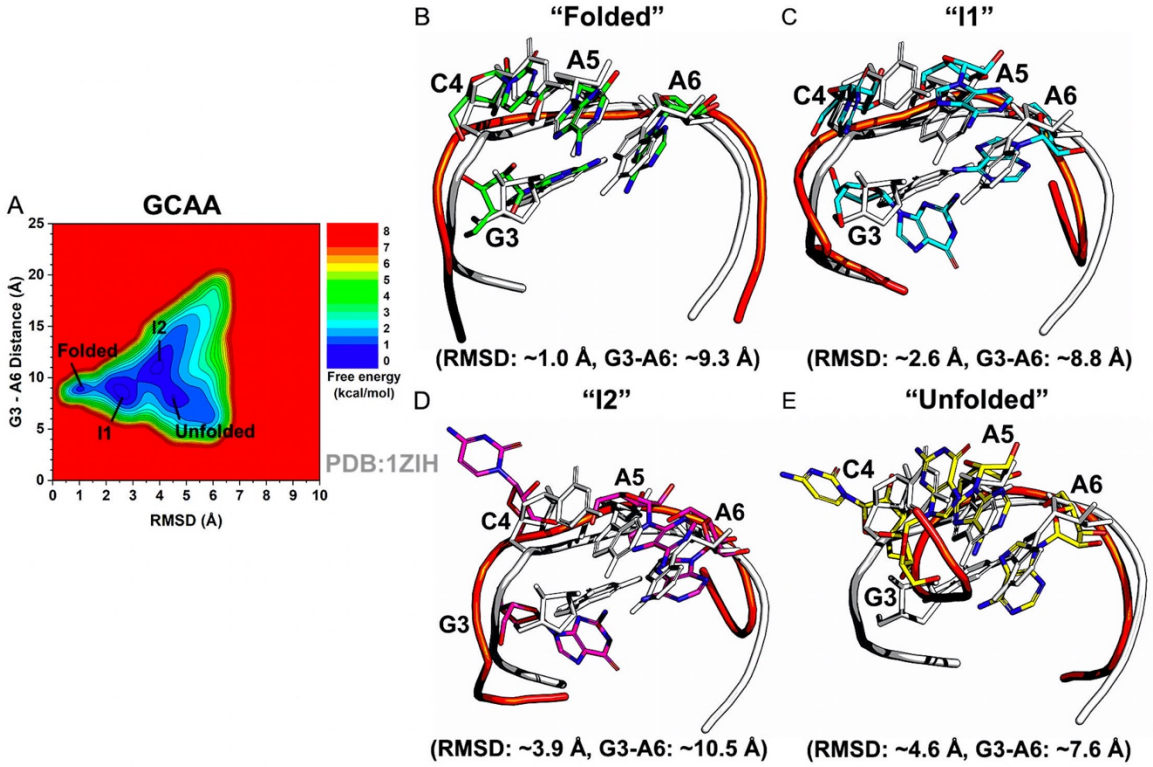

A
GCAA
G3 - A6 Distance (Å)
RMSD (Å)
Folded
I1
I2
Unfolded
Free energy (kcal/mol)
PDB:1ZIH
B
"Folded"
C4
A5
A6
G3
(RMSD: ~1.0 Å, G3-A6: ~9.3 Å)
C
"I1"
(RMSD: ~2.6 Å, G3-A6: ~8.8 Å)
D
"I2"
(RMSD: ~3.9 Å, G3-A6: ~10.5 Å)
E
"Unfolded"
(RMSD: ~4.6 Å, G3-A6: ~7.6 Å)

Figure 4

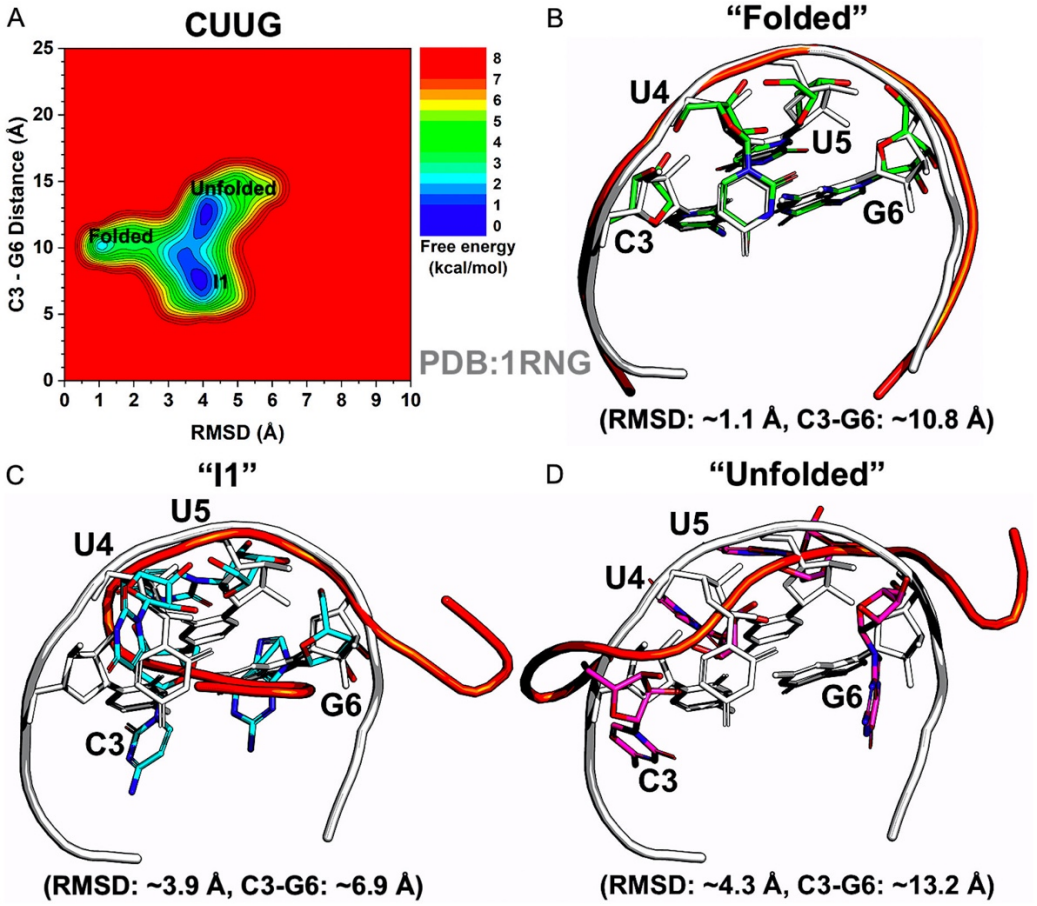
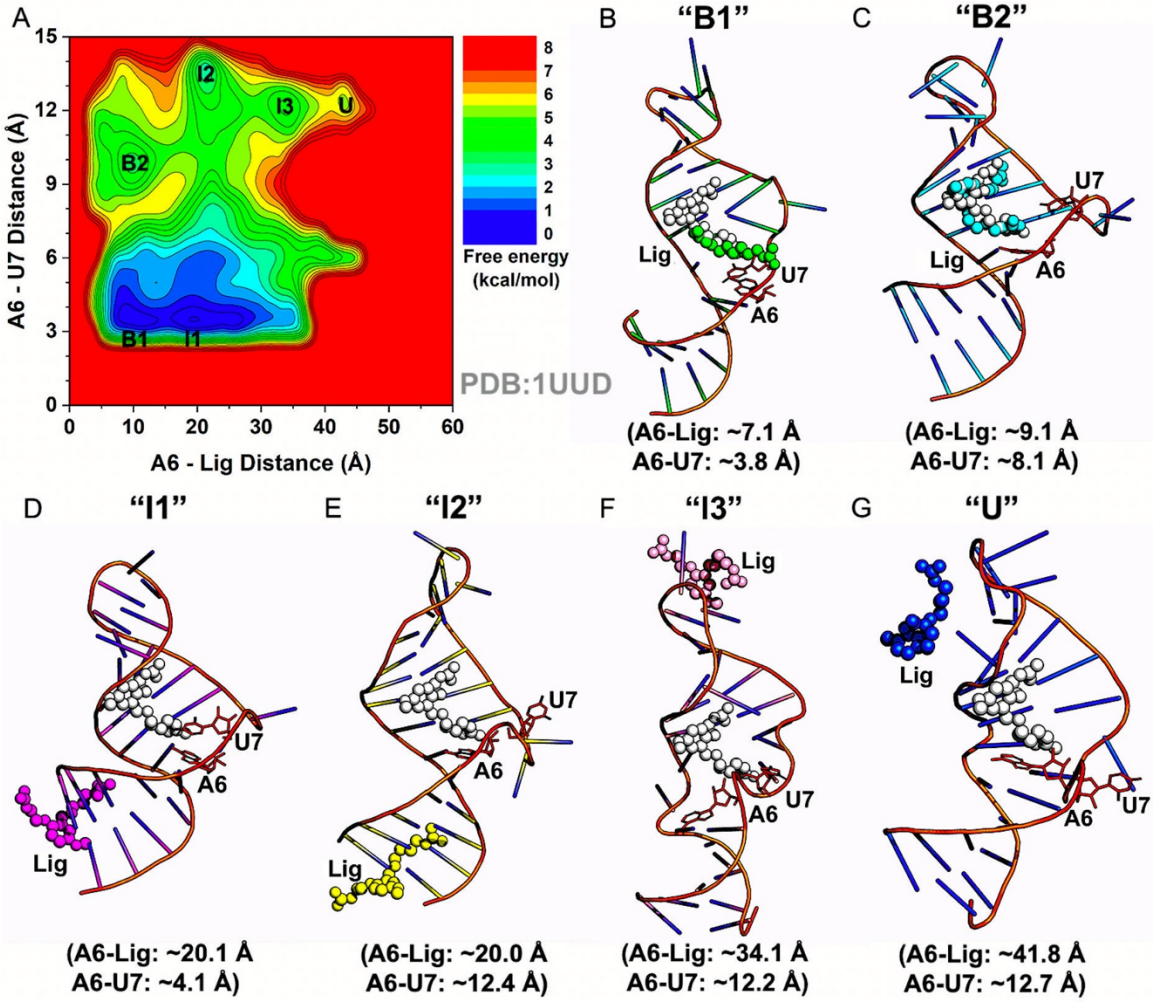

A
CUUG
C3 - G6 Distance (Å)
RMSD (Å)
Unfolded
Folded
I1
Free energy (kcal/mol)
PDB:1RNG
B
"Folded"
U4
U5
G6
C3
(RMSD: ~1.1 Å, C3-G6: ~10.8 Å)
C
"I1"
U5
U4
G6
C3
(RMSD: ~3.9 Å, C3-G6: ~6.9 Å)
D
"Unfolded"
U5
U4
G6
C3
(RMSD: ~4.3 Å, C3-G6: ~13.2 Å)

Figure 5


A
A6 - U7 Distance (Å)
A6 - Lig Distance (Å)
I2
I3
U
B2
B1
I1
Free energy (kcal/mol)
PDB:1UUD
B "B1"
Lig
U7
A6
(A6-Lig: ~7.1 Å
A6-U7: ~3.8 Å)
C "B2"
U7
Lig
A6
(A6-Lig: ~9.1 Å
A6-U7: ~8.1 Å)
D "I1"
U7
A6
Lig
(A6-Lig: ~20.1 Å
A6-U7: ~4.1 Å)
E "I2"
U7
A6
Lig
(A6-Lig: ~20.0 Å
A6-U7: ~12.4 Å)
F "I3"
Lig
A6
U7
(A6-Lig: ~34.1 Å
A6-U7: ~12.2 Å)
G "U"
Lig
A6
U7
(A6-Lig: ~41.8 Å
A6-U7: ~12.7 Å)

Supporting Information

for "**Gaussian Accelerated Molecular Dynamics in OpenMM**"

Implementation algorithm of Gaussian accelerated Molecular Dynamics (GaMD) in OpenMM

```
GaMD {
   For i = 1, …, conventional_md // Stage 1: run short initial conventional molecular dynamics
      if (i >= conventional_md_prep):
         n =  i - conventional_md_prep
         Update(V, Vmax, Vmin)
      if (i >= conventional_md_prep) and (i % averaging_window_interval):
         Update(i, V, Vavg, sigmaV)
   End

   if (i == conventional_md):
      calculate_threshold_energy_with_effective_harmonic_constant(sigma0,  sigmaV,  Vmax,
Vmin, k, E)

   For i = 1, …, gamd_equilibration: // Equilibrate the system after adding boost potential
      If (E > V):
         deltaV = 0.5*k*(E-V)**2
         V = V + deltaV
      EndIf
      Update Vmax, Vmin, Vavg, sigmaV
      If (i >= gamd_equilibration_prep):
         calculate_threshold_energy_with_effective_harmonic_constant(sigma0, sigmaV, Vmax,
Vmin, k, E)
      EndIf
    End

    For i = 1, …, total_simulation_length // run production simulation
         If (E > V):
            deltaV = 0.5*k*(E-V)**2
            V = V + deltaV
         EndIf
   End
}

Subroutine UpdateMaxMin(V,Vmax,Vmin):
   if (V > Vmax) Vmax = V
   if (V < Vmin) Vmin = V
}
Subroutine UpdateAvgSigma(n, V, Vmax, Vmin, Vavg, sigmaV ):
   Vdiff  = V – Vavg
   Vavg = Vavg + Vdiff / n
```

```
    M2 = M2 + Vdiff * (V – Vavg)
    sigmaV = sqrt(M2 / n)
}

// Lower Bound Integrator
Subroutine  calculate_threshold_energy_with_effective_harmonic_constant(sigma0,Vmax,Vmin,
k, E):
    E = Vmax
    k0’ = (sigma0/sigmaV) * (Vmax-Vmin)/(Vmax-Vavg)
    k0 = min(1.0, k0’)
    k = k0/(Vmax-Vmin)
}
// Upper Bound Integrator
Subroutine calculate_threshold_energy_with_effective_harmonic_constant(sigma0,Vmax,Vmin,
k, E):
    k0” = (1-sigma0/sigmaV) * (Vmax-Vmin)/(Vavg-Vmin)
    If 0 < k0” <= 1:
        k0 = k0”
        E = Vmin + (Vmax-Vmin)/k0
    Else:
        E = Vmax
        k0’ = (sigma0/sigmaV) * (Vmax-Vmin)/(Vmax-Vavg)
        k0 = min(1.0, k0’)
    end
    k = k0/(Vmax-Vmin)
}
```

Example Input XML file for GaMD-OpenMM simulation

```
<?xml version="1.0" ?>
<gamd>
    <temperature>298.15</temperature> <!-- unit.kelvin -->

    <system>
        <nonbonded-method>PME</nonbonded-method>
        <nonbonded-cutoff>1.0</nonbonded-cutoff> <!-- unit.nanometers -->
        <constraints>HBonds</constraints>
    </system>

    <barostat>
        <pressure>1.0</pressure> <!-- unit.bar -->
        <frequency>25</frequency>
    </barostat>

    <run-minimization>True</run-minimization>
```

```
  <integrator>
    <algorithm>langevin</algorithm>
    <boost-type>lower-dual</boost-type>
    <sigma0>
      <primary>6.0</primary> <!-- unit.kilocalories_per_mole -->
      <secondary>6.0</secondary> <!-- unit.kilocalories_per_mole -->
    </sigma0>
    <random-seed>0</random-seed>
    <dt>0.002</dt> <!-- unit.picoseconds -->
    <friction-coefficient>1.0</friction-coefficient> <!-- unit.picoseconds**-1 -->
    <number-of-steps>
      <conventional-md-prep>200000</conventional-md-prep>
      <conventional-md>1000000</conventional-md>
      <gamd-equilibration-prep>200000</gamd-equilibration-prep>
      <gamd-equilibration>2000000</gamd-equilibration>
      <gamd-production>3000000</gamd-production>
      <averaging-window-interval>50000</averaging-window-interval>
    </number-of-steps>
  </integrator>

  <input-files>
    <amber>
      <topology>data/dip.top</topology>
      <coordinates type="rst7">data/md-4ns.rst7</coordinates>
    </amber>
  </input-files>

  <outputs>
    <directory>output/</directory>
    <overwrite-output>True</overwrite-output>
    <reporting>
      <energy>
        <interval>500</interval>
      </energy>
      <coordinates>
        <file-type>DCD</file-type>
      </coordinates>
      <statistics>
        <interval>500</interval>
      </statistics>
    </reporting>
  </outputs>
</gamd>
```

Figure S1. Potential of mean force (PMF) profiles of the (**A-B**) Φ **and (C-D)** Ψ dihedrals calculated from three 1000ns cMD simulations (**A,C**) and three 30ns GaMD simulations (**B,D**).

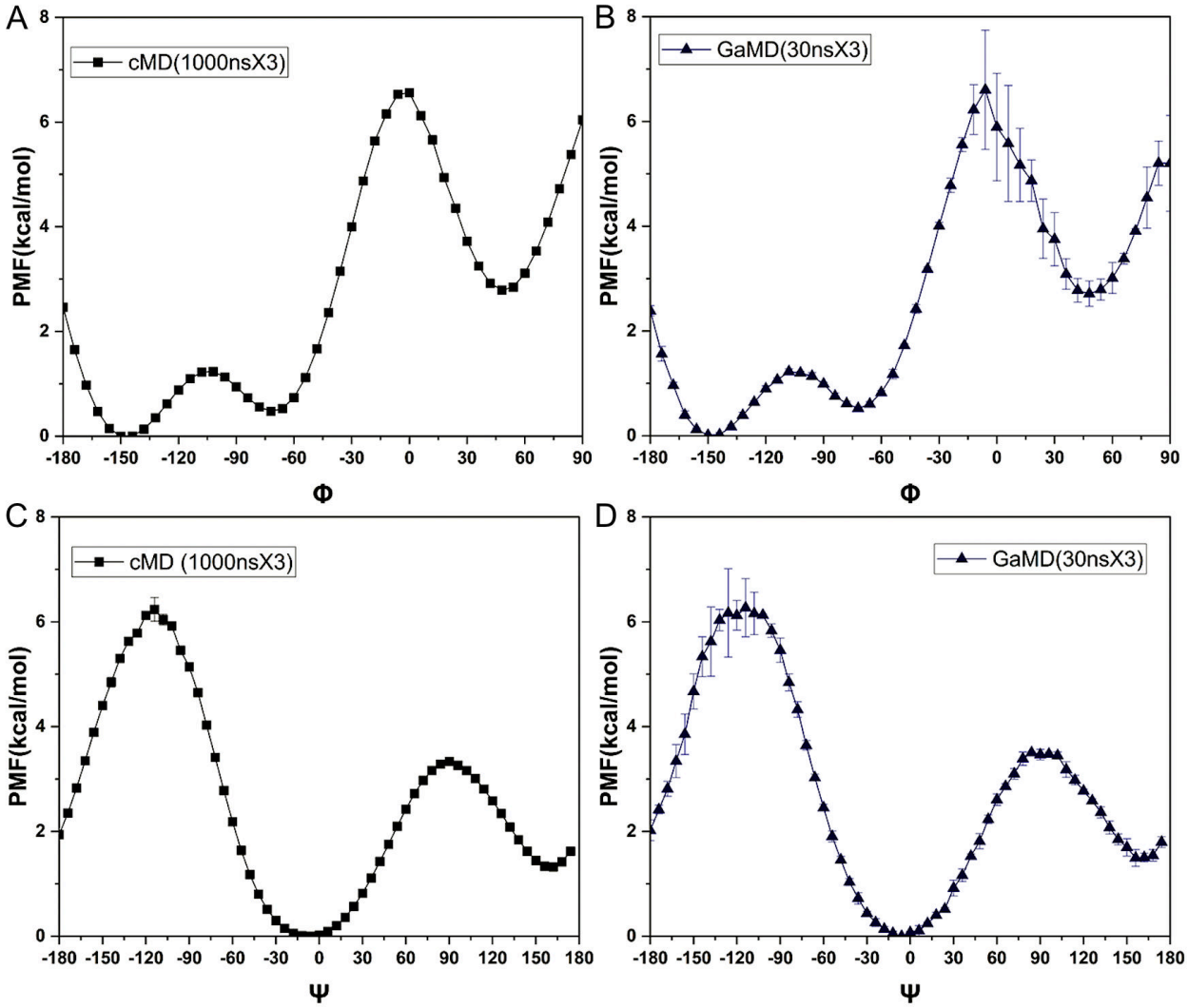


Figure S2. (A-C) 2D free energy profiles of the heavy-atom RMSD of UUCG relative to the 1F7Y PDB structure and the COM distance between nucleotides U3 and G6 calculated from three independent 5,000 ns GaMD simulations of the UUCG RNA tetraloop. The low-energy RNA conformational states are labeled "Folded", "I1", "I2", and "Unfolded". **(D)** Time courses of the heavy-atom RMSD of UUCG relative to the 1F7Y PDB structure calculated from three independent 5,000 ns GaMD simulations of the UUCG RNA tetraloop. **(E)** Time courses of the COM distance between nucleotides U3 and G6 calculated from three independent 5,000 ns GaMD simulations of the UUCG RNA tetraloop.

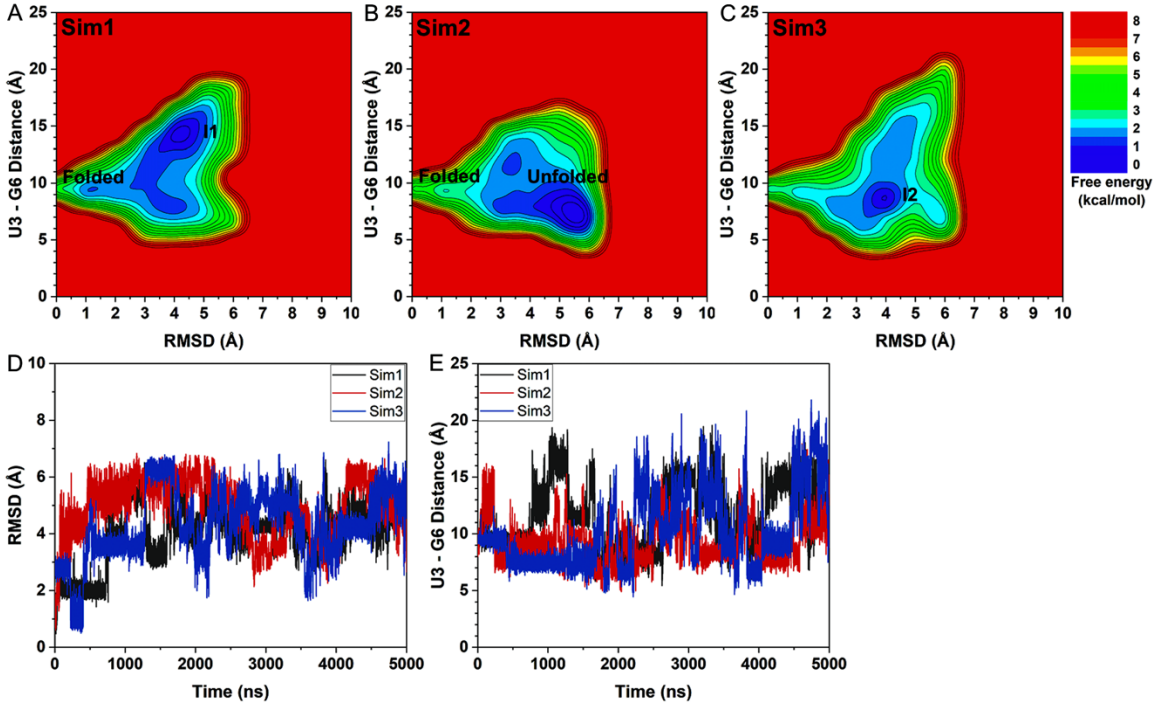


Figure S3. (A-C) 2D free energy profiles of the heavy-atom RMSD of GCAA relative to the 1ZIH PDB structure and the COM distance between nucleotides G3 and A6 calculated from three independent 4,000 ns GaMD simulations of the GCAA RNA tetraloop. The low-energy RNA conformational states are labeled "Folded", "I1", "I2", and "Unfolded". **(D)** Time courses of the heavy-atom RMSD of GCAA relative to the 1ZIH PDB structure calculated from three independent 4,000 ns GaMD simulations of the GCAA RNA tetraloop. **(E)** Time courses of the COM distance between nucleotides G3 and A6 calculated from three independent 4,000 ns GaMD simulations of the GCAA RNA tetraloop.

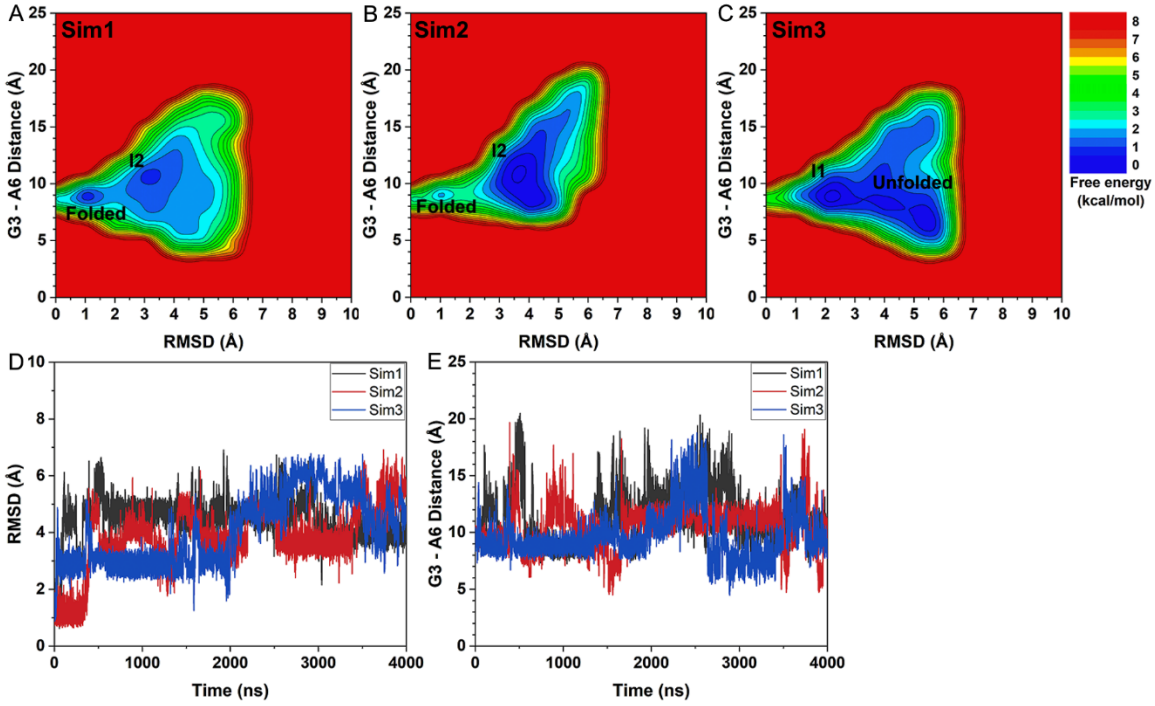


Figure S4. (A-C) 2D free energy profiles of the heavy-atom RMSD of CUUG relative to the 1RNG PDB structure and the COM distance between nucleotides C3 and G6 calculated from three independent 3,000-5,000 ns GaMD simulations of the CUUG RNA tetraloop. The low-energy RNA conformational states are labeled "Folded", "I1", and "Unfolded". **(D)** Time courses of the heavy-atom RMSD of CUUG relative to the 1RNG PDB structure calculated from three independent 3,000-5,000 ns GaMD simulations of the CUUG RNA tetraloop. **(E)** Time courses of the COM distance between nucleotides C3 and G6 calculated from three independent 3,000-5,000 ns GaMD simulations of the CUUG RNA tetraloop.

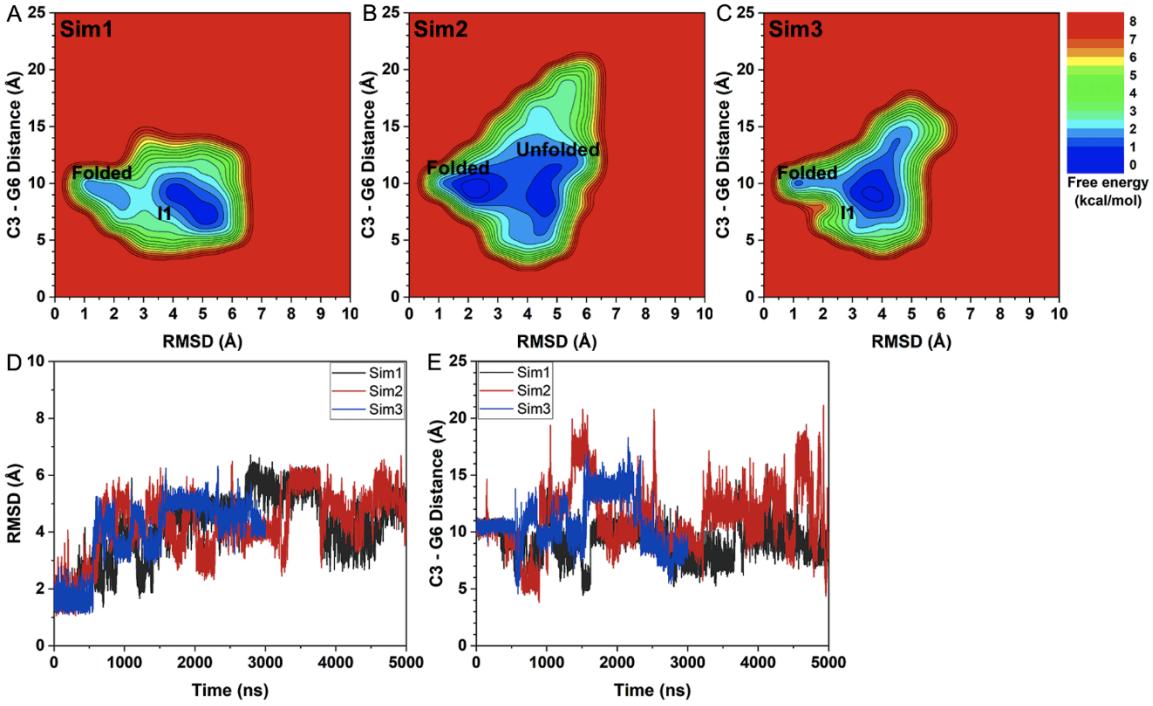


Figure S5. (A) Time courses of the COM distance between the rbt203 ligand and RNA nucleotide A6 calculated from five independent 500 ns GaMD simulations of the rbt203 binding to the HIV-1 Tar RNA. **(B)** Time courses of the COM distance between RNA nucleotides A6 and U7 side chains calculated from five independent 500 ns GaMD simulations of the rbt203 binding to the HIV-1 Tar RNA.

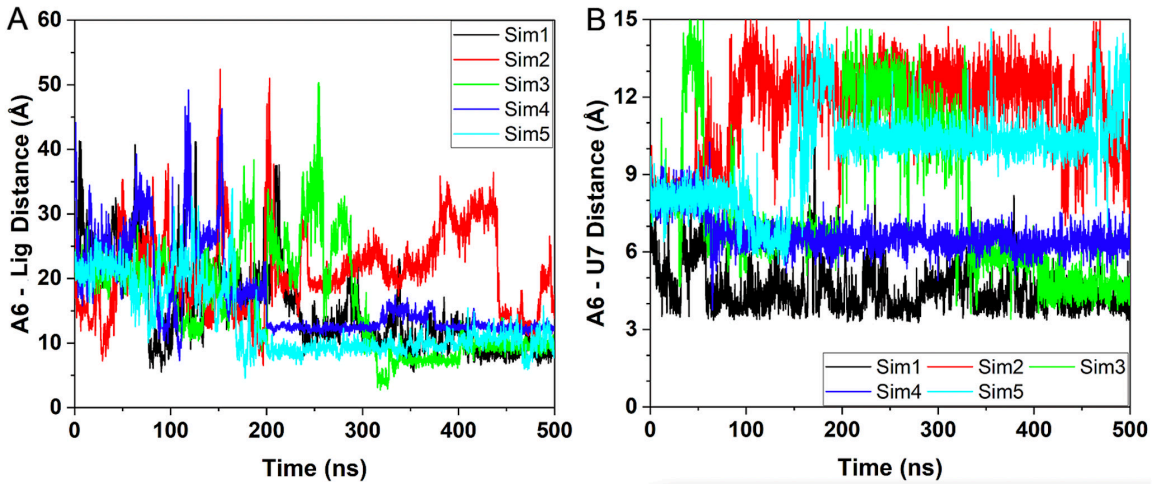


Figure S6. 2D free energy profiles of the COM distance between the rbt203 ligand (Lig) and RNA nucleotide A6 and the COM distance between RNA nucleotides A6 and U7 side chains calculated from five independent 500 ns GaMD simulations of the rbt203 binding to the HIV-1 Tar RNA. The low-energy conformational states are labeled "B1", "B2", "I1", "I2", "I3", and "U".

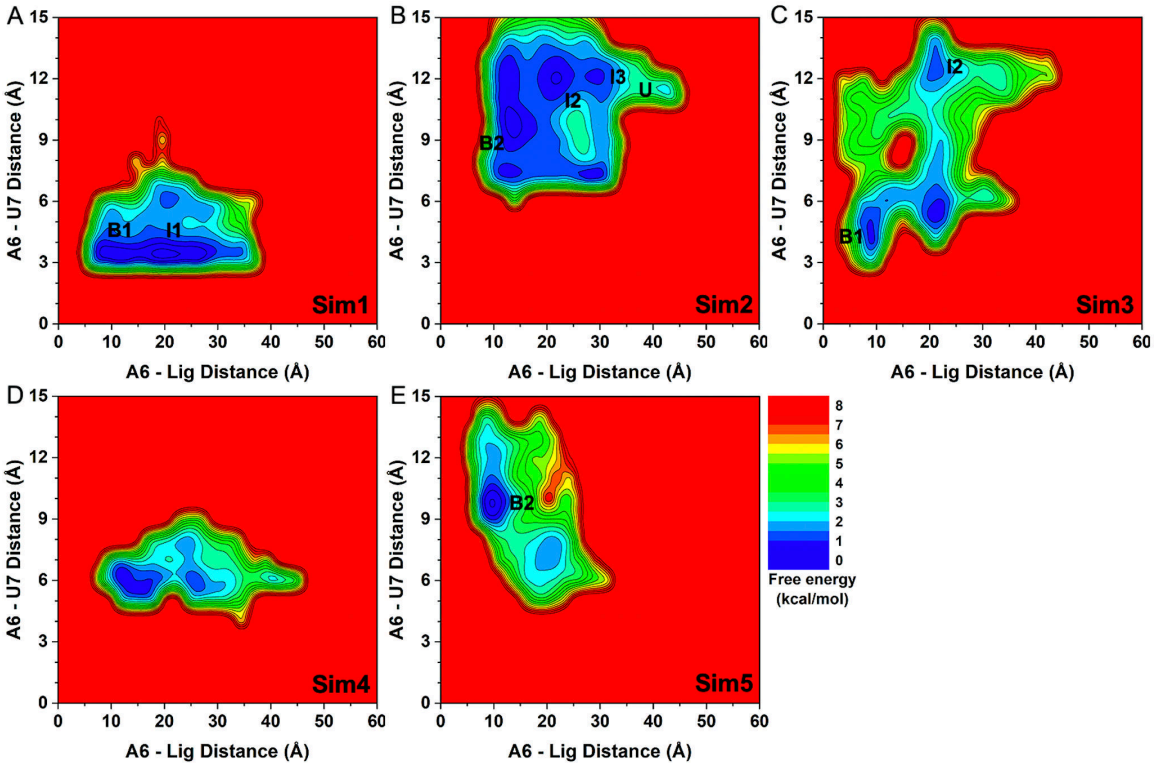


Figure S7. Graphical representation of the GaMD algorithm as implemented in OpenMM. The algorithm is divided into the "conventional MD" and "Gaussian accelerated MD" portions, both of which are further divided into a total of five stages.

Stage 1: Conventional MD preparatory stage: no statistics are collected to allow the system to equilibrate.

Stage 2: Conventional MD stage: boost parameters V_{max}, V_{min}, V_{avg}, and σ_v are collected.

Stage 3: GaMD pre-equilibration stage: boost potential is applied, but boost parameters are not updated.

Stage 4: GaMD equilibration stage: boost potential is applied, and boost parameters are updated.

Stage 5: GaMD production stage: boost potential is applied, boost parameters are held fixed.