Published in final edited form as:
J Phys Chem Lett. 2023 June 01; 14(21): 4970–4982. doi:10.1021/acs.jpclett.3c00926.

Deep Boosted Molecular Dynamics (DBMD): Accelerating Molecular Simulations with Gaussian Boost Potentials Generated Using Probabilistic Bayesian Deep Neural Network

Hung N. Do[1], **Yinglong Miao**[1,*]

[1]Center for Computational Biology and Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66047

Abstract

We have developed a new Deep Boosted Molecular Dynamics (DBMD) method. Probabilistic Bayesian neural network models were implemented to construct boost potentials that exhibit Gaussian distribution with minimized anharmonicity, thereby allowing for accurate energetic reweighting and enhanced sampling of molecular simulations. DBMD was demonstrated on model systems of alanine dipeptide and the fast-folding protein and RNA structures. For alanine dipeptide, 30ns DMBD simulations captured up to 83–125 times more backbone dihedral transitions than 1μs conventional molecular dynamics (cMD) simulations and were able to accurately reproduce the original free energy profiles. Moreover, DBMD sampled multiple folding and unfolding events within 300ns simulations of the chignolin model protein and identified low-energy conformational states comparable to previous simulation findings. Finally, DBMD captured a general folding pathway of three hairpin RNAs with the GCAA, GAAA, and UUCG tetraloops. Based on Deep Learning neural network, DBMD provides a powerful and generally applicable approach to boosting biomolecular simulations. DBMD is available with open source in OpenMM at https://github.com/MiaoLab20/DBMD/.

Graphical Abstract

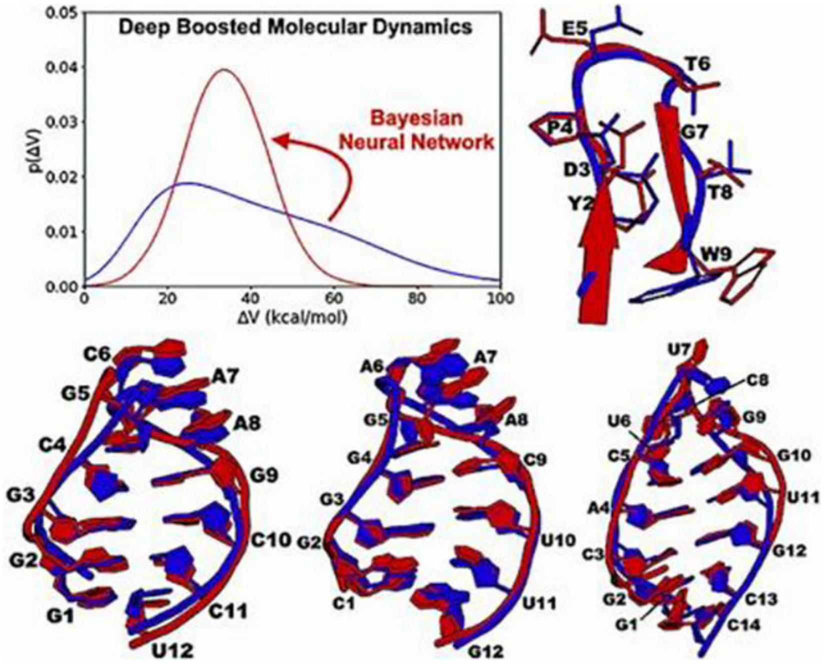
[*]To whom correspondence should be addressed: miao@ku.edu.

Author Contributions

H.N.D. performed research, analyzed data, and wrote the manuscript. Y.M. supervised the project, interpreted data, and wrote the manuscript. All authors contributed towards the final version of the manuscript.

Supporting Information

Algorithm (pseudo-code) and example input file of DBMD in OpenMM, along with description of input parameters. Time courses of the effective harmonic force constants, boost potential parameters, the reaction coordinates selected for the calculations of PMF profiles and distributions of the boost potentials applied in the DBMD simulations of alanine dipeptide, chignolin, and RNA hairpins with GCAA, GAAA, and UUCG tetraloops.



Keywords

Probabilistic neural networks; Molecular dynamics; Protein folding; RNA folding; Free energy profiles

Molecular dynamics (MD) is a powerful computational technique for simulating biomolecular dynamics at an atomistic level[1]. With recent advances in computing hardware and software developments, timescales accessible to MD simulations have significantly increased[2, 3]. However, conventional MD (cMD) is often limited to tens to hundreds of microseconds[4] for simulations of typical biomolecular systems, and cannot attain the timescales required to observe many biological processes of interest, which typically occur over milliseconds or longer with high energy barriers (e.g., 8–12 kcal/mol)[5].

Many enhanced sampling techniques have been developed during the last several decades to overcome the challenges mentioned above[6]. In particular, Gaussian accelerated molecular dynamics (GaMD) is an enhanced sampling that technique works by applying a harmonic boost potential to smooth biomolecular potential energy surface[7]. Since this boost potential exhibits a near Gaussian distribution, cumulant expansion to the second order (“Gaussian approximation”) can be applied to achieve proper energetic reweighting[8]. GaMD allows for simultaneous unconstrained enhanced sampling and free energy calculations of large biomolecules [7]. GaMD has been successfully demonstrated on enhanced sampling of ligand binding, protein folding, protein conformational change, as well as protein-membrane, protein-protein, and protein-nucleic acid interactions[3]. GaMD has been implemented in widely used simulation packages including AMBER[7], NAMD[9], OpenMM[10], GENESIS[11], and TINKER-HP[12].

Recently, Machine Learning/Deep Learning techniques (ML/DL) have been combined with MD methods to enhance the sampling of biomolecular simulations. DeepDriveMD is a DL driven adaptive MD method designed specifically to simulate protein folding[13]. In DeepDriveMD, DL was utilized to reduce the dimensionality of MD simulations to automatically build latent representations that correspond to biophysically relevant collective variables (CVs) and drive MD simulations to automatically sample potentially novel conformational states based on the CVs[13]. DeepDriveMD has been demonstrated to speed up the folding simulations of Fs-peptide and the fast-folding variant of the villin head piece protein by at least 2.3 folds[13]. The State Predictive Information Bottleneck (SPIB) approach was applied as a deep neural network to learn a priori CV for well-tempered metadynamics from undersampled trajectories[14]. The well-tempered metadynamics performed along the biased SPIB-learned CVs were shown to achieve > 40 times acceleration in simulating the left- to right-handed chirality transitions in a synthetic helical peptide and permeation of a small benzoic acid molecule through a synthetic, symmetric phospholipid bilayer[14]. Moreover, denoising diffusion probabilistic models were combined with replica exchange MD to achieve superior sampling of biomolecular energy landscape at temperatures that were not simulated without the assumption of particular slow degrees of freedom[15]. The temperature was treated as a fluctuating random variable and not a control parameter to allow for the direct sampling from the joint probability distribution in configuration and temperature space. The procedure was shown to discover transition and metastable states that were previously unseen at the temperature of interest and bypass the need to perform simulations for a wide range of temperatures[15].

In this work, we have developed a new Deep Boosted Molecular Dynamics (DBMD) method. In DBMD, probabilistic Bayesian neural network models were used to construct boost potentials that exhibit Gaussian distribution with minimized anharmonicity for accurate energetic reweighting and enhanced sampling (Figure 1). DBMD has been demonstrated on model systems of the alanine dipeptide in explicit and implicit solvent, the chignolin fast-folding protein, and three hairpin RNAs with the GCAA, GAAA, and UUCG tetraloops.

DBMD simulations were performed on alanine dipeptide on alanine dipeptide (Figure 2a) in explicit and implicit solvent. Representative distributions of randomly generated boost potentials and the boost potentials generated by DL for alanine dipeptide in explicit and implicit solvent are shown in Figure 2b and 2c, respectively. DL was able to reduce the anharmonicity from 0.153 for the randomly generated boost potentials to 0.019 and 0.006 in two iterations of the explicit-solvent simulation (Figure 2b), and from 0.295 to 0.013 and 0.006 in two iterations of the implicit-solvent simulation (Figure 2c).

The time courses of the effective harmonic force constants (k_{0P} and k_{0D}) as well as the total and dihedral boost potential parameters (V_{min}, V_{max}, and E) during the equilibration of the alanine dipeptide in explicit and implicit solvent are shown in Figure S1. During the one round of 1ns DBMD equilibration in explicit solvent, the total and dihedral effective harmonic force constants k_{0P} and k_{0D} stayed at 0.35 and 1.0, respectively (Figure S1a). The minimum total and dihedral potential energies V_{minP} and V_{minD} also remained constant at −5,966.96 kcal/mol and 5.92 kcal/mol, respectively (Figure S1b–S1c). However, the maximum total and dihedral potential energy V_{maxP} and V_{maxD} increased from −5,742.44 kcal/mol and 25.18 kcal/mol to −5,690.89 kcal/mol and 33.16 kcal/mol, respectively (Figure S1b–S1c). The reference total and dihedral potential energy for applying boosts were the same as the maximum potential energies. The effective harmonic force constants as well as extrema and reference potential energies in the implicit-solvent equilibration followed similar trends as the explicit-solvent simulation (Figure S1e–S1g).

Three independent 30ns DBMD simulations of alanine dipeptide in both explicit and implicit solvent captured more dihedral transitions compared to 1μs cMD simulations (Figure S2). In particular, DBMD sampled ~15, ~14, and ~10 Φ dihedral transitions during the 30ns of Sim1, Sim2, and Sim3, respectively, compared to only ~4 dihedral transitions observed in the 1μs cMD of alanine dipeptide in explicit solvent (Figure S2a–S2d). In the implicit-solvent simulations, Sim1, Sim2, and Sim3 sampled ~17, ~28, and ~28 Φ dihedral transitions during the 30ns simulations, respectively, compared to the ~26 Φ dihedral transitions observed in the 1μs cMD simulation (Figure S2e–S2h). Therefore, DBMD accelerated the explicit-solvent simulations by ~83–125 times and implicit-solvent simulations by ~22–36 times. Furthermore, the boost potentials applied in DBMD simulations of alanine dipeptide followed Gaussian distributions, with low anharmonicity of 6.2×10^{-3} in the explicit-solvent and 1.7×10^{-4} in implicit-solvent simulations (Figure S3a–S3b). The averages and standard deviations of the added boost potentials were recorded to be 11.2 ± 2.8 and 11.3 ± 2.3 kcal/mol in the explicit and implicit solvent simulations, respectively.

The PMF free energy profiles of alanine dipeptide were calculated for the Φ and Ψ dihedral angles. The 1D PMF free energy profiles were in excellent agreement between DBMD and cMD for both Φ and Ψ in explicit and implicit solvent (Figure S3c–S3f). Moreover, the 2D PMF free energy profiles of the (Φ, Ψ) backbone dihedrals showed high degrees of similarity between DBMD and cMD simulations (Figure 2d–2g). In particular, DBMD simulations in explicit solvent sampled five different low-energy conformational states of alanine dipeptide, which centered around (−150°, 159°) in the β-sheet, (−72°, 162°) in the polyproline II (P_{II}), (48°, 18°) in the left-handed α helix (α_L), and (−148°, 0°) and (−69°, −17°) in the right-handed α helix (α_R) conformation (Figure 2d). In implicit solvent, DBMD also identified five low-energy conformational states of alanine dipeptide, including β-sheet centered around (−160°, 150°), P_{II} around (−62°, 140°) and (−90°, 61°), α_L around (56°, 34°), and α_R around (−70°, −27°) (Figure 2e). The 1D and 2D free energy profiles of (Φ, Ψ) calculated from DBMD simulations were in excellent agreements with previous GaMD simulations performed by AMBER[7], NAMD[9], and OpenMM[10]. Therefore, simulations of alanine dipeptide have demonstrated the enhanced sampling capability as well as accuracy of DBMD for both explicit and implicit solvent systems.

Representative distributions of randomly generated dual boost potentials and the boost potentials generated by DL for chignolin folding are shown in Figure 3a. With the use of DL, the anharmonicity reduced from 0.17 for the randomly generated boost potentials to 0.01 and 0.005 in two iterations (Figure 3a).

The time courses of the effective harmonic force constants (k_{0P} and k_{0D}) as well as the total and dihedral boost potential parameters (V_{min}, V_{max}, and E) during the equilibration of the chignolin fast-folding protein in explicit solvent are shown in Figure S4. During the two rounds of 5ns DBMD equilibration, the dihedral effective harmonic force constant k_{0D} remained at 1.0, while the total effective harmonic force constant k_{0P} decreased from 0.94 in round one to 0.89 in round two (Figure S4a). The minimum total potential energy V_{minP} increased from −21,388.36 kcal/mol in round one to −20,761.33 kcal/mol in round two (Figure S4b). The maximum total potential energy V_{maxP} increased from −20,742.03 kcal/mol to −20,234.95 kcal/mol and −19,671.23 kcal/mol at the end of round one and two, respectively (Figure S4b). The minimum dihedral potential energy V_{minD} increased from 87.50 kcal/mol in round one to 94.88 kcal/mol in round two (Figure S4c). The maximum dihedral potential energy V_{maxD} increased from 120.52 kcal/mol to 139.37 kcal/mol at the end of round one and 143.53 kcal/mol at the end of round two (Figure S4c). While the reference dihedral potential energy E_D was identical to the maximum dihedral potential energy V_{maxD}, the reference total potential energy E_P was slightly higher than the maximum total potential energy V_{maxP} (Figure S4b).

Three independent 300ns DBMD simulations of chignolin in explicit solvent starting from its extended conformation were able to capture multiple folding and unfolding events of the protein (Figure S5). In particular, six, seven, and ten different folding-unfolding events were sampled in Sim1, Sim2, and Sim3 of chignolin (Figure S5a). Here, chignolin was considered folded if the C_α-atom RMSD of residues Y2-W9 was ≤ 1.0 Å. Furthermore, the boost potentials applied in DBMD simulations of chignolin followed the Gaussian distribution, with an anharmonicity of 7.1×10^{-3} (Figure S5c) and an average of 23.1 ± 5.1 kcal/mol.

The 2D PMF free energy profile of chignolin folding was calculated using the C_α-atom RMSD relative to the 1UAO[16] PDB structure and Rg of residues Y2-W9 as RCs. Three different low-energy conformational states of chignolin were identified from the free energy profile, namely “Folded”, intermediate “I”, and “Unfolded” (Figure 3b). The “Folded” low-energy conformational state of chignolin centered around 0.4 Å and 4.1 Å of RMSD and Rg, respectively. In this state, terminal residues Y2-D3 formed β-sheets with residues G7-W9 of chignolin, while the loop formed by the backbone atoms of residues P4-T6 closely matched with the 1UAO[16] PDB structure (Figure 3c). In the intermediate “I” low-energy conformational state, the C_α-atom RMSD and Rg were ~4.0 Å and ~5.2 Å. Transitioning from the “Folded” to intermediate “I” state, the β-strands were broken apart due to the opposite movement of residues G1-D3 and T8-G10. However, the core loop of chignolin was somewhat maintained with the hydrophilic side chains of residues E5-T6 exposed to the solvent (Figure 3d). Finally, in the “Unfolded” low-energy conformational state, chignolin was fully extended with all amino acids exposed to the solvent, resulting in a RMSD of ~5.0 Å and Rg of ~6.5 Å (Figure 3e).

Representative distributions of randomly generated dual boost potentials and the boost potentials generated by DL for the hairpin RNAs with the GCAA, GAAA, and UUCG tetraloops are shown in Figures 4a–6a, respectively. With the use of DL, the anharmonicity reduced from 0.135 for the randomly generated boost potentials to 0.016, 0.015, and 0.009

in three iterations of the GCAA RNA system simulation (Figure 4a). For GAAA, DL lowered the anharmonicity from 0.137 for the random boost potentials to 0.012, 0.01, and 0.008 in three iterations (Figure 5a). For UUCG, the anharmonicity reduced from 0.147 to 0.014 to 0.013 and 0.008 (Figure 6a).

The time courses of the effective harmonic force constants (k_{0P} and k_{0D}) as well as the total and dihedral boost potential parameters (V_{min}, V_{max}, and E) during the equilibration of the hairpin RNAs with GCAA, GAAA, and UUCG tetraloop in implicit solvent are shown in Figures S6–S8. During the three rounds of 5ns DBMD equilibration of the GCAA RNA tetraloop system, the total effective harmonic force constant k_{0P} decreased from 0.20 in round one to 0.10 in round two but increased to 0.17 in round three, while the dihedral effective harmonic force constant k_{0D} decreased from 0.84 in round one to 0.56 in round two and 0.51 in round three (Figure S6a). The minimum total potential energy V_{minP} fluctuated from −2867.79 kcal/mol in round one to −2930.46 kcal/mol in round two to −2871.63 kcal/mol in round three (Figure S6b). The maximum total potential energy V_{maxP} also fluctuated between −2509.02 kcal/mol, −2366.52 kcal/mol, and −2458.10 kcal/mol among the three (Figure S6b). The minimum dihedral potential energy V_{minD} fluctuated from 291.14 kcal/mol in round one to 326.13 kcal/mol in round two to 314.70 kcal/mol in round three, whereas the maximum dihedral potential energy V_{maxD} decreased from 400.88 kcal/mol to 391.04 and 368.91 kcal/mol from round one to round three (Figure S6c). The reference total and dihedral potential energies E_P and E_D were mostly identical to the maximum total and dihedral potential energies V_{maxP} and V_{maxD}, except during round one for the E_D (Figure S6c).

For the GAAA RNA tetraloop system, the total and dihedral effective harmonic force constants k_{0P} and k_{0D} decreased from 1.0 and 0.33 in round one to 0.15 and 0.98 in round two to 0.098 and 0.51 in round three (Figure S7a). The minimum total potential energy V_{minP} decreased from −2621.20 kcal/mol in round one to −2746.68 kcal/mol in round two to −2799.21 kcal/mol in round three, whereas the maximum total potential energy V_{maxP} fluctuated between −2380.18 kcal/mol, −2177.94 kcal/mol, and −2265.46 kcal/mol during the three rounds of DBMD equilibration (Figure S7b). The minimum dihedral potential energy V_{minD} increased from 289.55 kcal/mol to 323.12 kcal/mol and 331.20 kcal/mol from round one to round three, while the maximum dihedral potential energy V_{maxD} decreased from 396.04 kcal/mol to 387.99 and 380.68 kcal/mol from round one to three (Figure S7c). The reference total and dihedral potential energies were identical to the maximum total and dihedral energies.

For the UUCG RNA tetraloop system, the total and dihedral effective harmonic force constants k_{0P} and k_{0D} fluctuated between 0.12 and 1.0 in round one to 0.27 and 0.32 in round two to 0.19 and 0.46 in round three (Figure S8a). The minimum total potential energy V_{minP} also fluctuated between −3360.51 kcal/mol, −3189.11 kcal/mol, and −3258.47 kcal/mol from round one to three of DBMD equilibration, whereas the maximum total potential energy V_{maxP} increased from −2900.68 kcal/mol in round one to −2831.27 kcal/mol and −2733.89 kcal/mol in round two and three, respectively (Figure S8b). The minimum dihedral potential energy V_{minD} fluctuated between 341.78 kcal/mol, 337.89 kcal/mol, and 327.41 kcal/mol from round one to three of DBMD equilibration, whereas the maximum dihedral potential energy V_{maxD} decreased from 416.37 kcal/mol to 411.57 kcal/mol to 407.42 kcal/mol from round one to three (Figure S8c). The reference total and dihedral potential energies were the same as the maximum potential energies during the DBMD equilibration of the UUCG RNA tetraloop system.

Multiple independent 2μs DBMD simulations were performed on the hairpin RNAs with GCAA, GAAA, and UUCG tetraloops in implicit solvent, starting from their extended conformations (Figures 4–6). Remarkably, DBMD was able to capture multiple folding and unfolding events for all three hairpin RNAs within 2μs of simulations. In particular, a total of 18, 16, and 11 different stable folding-unfolding events were observed within 2μs DBMD simulations of the RNAs with GCAA, GAAA, and UUCG tetraloops, respectively (Figures S9a–S11a). The DBMD boost potentials exhibited Gaussian distributions, with low anharmonicity of 8.3×10^{-3}, 3.9×10^{-4}, and 2.9×10^{-3} in the GCAA, GAAA, and UUCG RNA tetraloop simulations (Figures S9c, S10c, and S11c). Furthermore, the boost potentials were recorded to be 37.0 ± 4.5 kcal/mol for the GCAA, 32.9 ± 3.1 kcal/mol for GAAA, and 27.6 ± 3.4 for UUCG system, given the different η_P and η_D used for the RNA systems.

The 2D PMF free energy profiles of the hairpin RNAs with tetraloops were calculated using the heavy-atom RMSDs of the whole RNAs relative to respective PDB structures (1ZIH[17] for GCAA, 2ADT[18] for GAAA, and 2KOC[19] for UUCG) and the G1-U12, C1-G12, and G1-C14 center-of-mass (COM) distances as RCs. DBMD sampled three different low-energy conformational states, including "Folded", intermediate "I", and "Unfolded", for the RNA with GCAA tetraloop (Figure 4b), four different low-energy conformational states, namely "Folded", intermediate "I1" and "I2", and "Unfolded", for the GAAA tetraloop (Figure 5b), and three different low-energy conformational states, including "Folded", intermediate "I", and "Unfolded", for the UUCG tetraloop (Figure 6b).

In the "Folded" low-energy conformational state of the 12-mer hairpin RNA with the GCAA tetraloop, the heavy-atom RMSD relative to the 1ZIH[17] PDB structure was ~1.1 Å, and the COM distance between terminal nucleotides G1 and U12 was ~10.7 Å. This "Folded" low-energy conformational state was maintained by the Watson-Crick base pairs between nucleotides G2-C11, G3-C10, and C4-G9 and base stacking between nucleotides C6-A7-A8 of the GCAA tetraloop (Figure 4c). With transition from the "Folded" to the intermediate "I" state, most of the Watson-Crick base pairs distorted, with the side chains of nucleotides G9, C11, and U12 flipping out and exposing themselves to the solvent, while the base stacking between nucleotides C6-A7-A8 of the GCAA tetraloop was intact as observed in a conformation at ~8.3Å heavy-atom RMSD relative to the 1ZIH[17] PDB structure and ~6.5Å G1-U12 COM distance (Figure S12). In the intermediate "I" low-energy conformational state, the RNA began extending, with nucleotides G1-C4 and C10-U12 extending in opposite directions. The base stacking between nucleotides C6-A7-A8 was mostly broken, with nucleotide A8 flipping out to base stack with nucleotide A9. In this state, the heavy-atom RMSD relative to the 1ZIH[17] PDB structure was ~9.2 Å, and the G1-U12 COM distance was ~12.9 Å (Figure 4d). In the "Unfolded" low-energy conformational state, the RNA was completely stretched out, with a heavy-atom RMSD of ~14.5 Å and G1-U12 COM distance of ~48.3 Å (Figure 4e).

In the “Folded” low-energy conformational state of the 12-mer hairpin RNA with the GAAA tetraloop, the heavy-atom RMSD relative to the 2ADT[18] PDB structure was ~1.3 Å, and the COM distance between terminal nucleotides C1 and G12 was ~10.3 Å. Similar to the GCAA system, this “Folded” state of the GAAA system was maintained by the Watson-Crick base pairs between nucleotides C1-G12 and G4-C9 as well as the base stacking between nucleotides A6-A7-A8 of the GAAA tetraloop (Figure 5c). The heavy-atom RMSD increased to ~7.8 Å, whereas the C1-G12 COM distance decreased to ~7.5 Å in the intermediate “I1” low-energy conformation. In this state, both the Watson-Crick base pairs and base stacking in the GAAA tetraloop were broken, with nucleotides G3, G4, A7, A8, C9, U10 flipping out and exposing to the solvent. However, base stacking was observed between nucleotides G5 and A6 of the GAAA tetraloop (Figure 5d). In the “I2” intermediate state, the heavy-atom RMSD the 2ADT[18] PDB structure was ~9.5 Å, and the COM distance between terminal nucleotides C1 and G12 was ~17.9 Å. The 12-mer RNA was mostly distorted, with random base stacking formed between nucleotides G5-G12 and A6-A8. The side chains of the other nucleotides flipped out and exposed to the solvent (Figure 5e). In the “Unfolded” low-energy conformational state, the RNA was completely stretched out, with a heavy-atom RMSD of ~13.5 Å and C1-G12 COM distance of ~45.0 Å (Figure 5f).

The “Folded” low-energy conformational state of the 14-mer hairpin RNA with the UUCG tetraloop has a heavy-atom RMSD relative to the 2KOC[19] PDB structure of ~2.3 Å and COM distance between terminal nucleotides G1 and C14 of ~9.8 Å. In this state, Watson-Crick base pairs were formed between nucleotides G2-C13, C3-G12, A4-U11, and C5-G10. However, unlike the GCAA and GCAA systems, no base stacking was observed between the nucleotides in the UUCG tetraloop (Figure 6c). In the “I” intermediate state, the heavy-atom RMSD increased to ~9.3 Å and the G1-C14 COM distance decreased to ~7.6 Å. The RNA was mostly distorted, with random base stacking formed between nucleotides G2 and G10. Most of the other nucleotides flipped out and exposed to the solvent (Figure 6d). In the “Unfolded” low-energy conformational state, the heavy-atom RMSD relative to the 2KOC[19] PDB structure further increased to ~14.3 Å and the COM distance between nucleotides G1 and C14 increased to ~43.4 Å. The RNA was mostly stretched out (Figure 6e). Therefore, DBMD was able to capture repetitive folding and unfolding of RNA tetraloop structure in 2μs simulations, thereby enabling characterization of the RNA folding free energy landscapes.

In this work, we have developed DBMD, which generates boost potentials with Gaussian distribution using DL to reduce energy barriers and enhanced conformational sampling of biomolecules. Probabilistic Bayesian DL models are trained using potential energies of finished simulation frames to build the boost potentials that exhibit Gaussian distribution with anharmonicity $\gamma < 0.01$. We have demonstrated DBMD on the simulations of alanine dipeptide in explicit and implicit solvent and folding of the chignolin protein and hairpin RNAs with the GCAA, GAAA, and UUCG tetraloops. Overall, DBMD was able to greatly enhance conformational transitions and characterize the protein and RNA folding free energy landscapes.

DBMD captured multiple folding and unfolding events of chignolin within 300 ns of simulations (Figure S5a). Compared to previous aMD[20] simulations of chignolin folding, DBMD sped up the folding-unfolding transition by 1.35 times. Furthermore, DBMD accelerated the folding-unfolding transition by 6 times compared to previous 300ns simulations performed with GaMD in AMBER[7] and NAMD[9]. However, it should be noted that the reference total potential energy E_P was set to $V_{minP} + \frac{V_{maxP} - V_{minP}}{k_{0P}}$ for DBMD simulations of chignolin (Figure S4b). Therefore, the boost potentials in the DBMD simulations of chignolin were recorded to be 23.1 ± 5.1 kcal/mol, which was approximately triple and double the boost potentials applied in the previous all-atom dual-boost lower-bound GaMD simulations of chignolin in AMBER ($\triangle V = 9.54 \pm 2.44\ \mathrm{kcal/mol}$)[7] and NAMD ($\Delta V = 11.2 \pm 2.8\ \mathrm{kcal/mol}$)[9], respectively, where $E_P = V_{maxP}$. Nevertheless, even with the larger average and standard deviation of applied boost potentials, DBMD still managed to obtain relatively a smaller anharmonicity $\left(\gamma = 7.1 \times 10^{-3}\right)$ compared to GaMD in AMBER $\left(\gamma = 9.2 \times 10^{-3}\right)$[7] and NAMD $\left(\gamma = 9.7 \times 10^{-3}\right)$[9], which was one advantage of using DL to generate Gaussian boost potentials. On the other hand, DBMD still provided a 2D free energy profile of the C_α-atom RMSD and Rg of residues Y2-W9 with high degrees of similarity compared to previous GaMD simulations of chignolin performed with AMBER and NAMD[7, 9]. In particular, DBMD sampled all three low-energy conformational states ("Folded", intermediate "I", and "Unfolded") as GaMD in AMBER[7] and the two low-energy conformations ("Folded" and "I") as GaMD in NAMD[9] (Figure 3). Moreover, the folding mechanism uncovered by DBMD was relatively similar to that by GaMD in AMBER[7]. Starting from the extended conformation of the "Unfolded" state (Figure 3e), the terminal residues of chignolin was brought closer due to the interactions between residues P4 and G7 in the intermediate "I" state (Figure 3d). With transition from the intermediate "I" to the "Folded" state, antiparallel β-sheets were formed between residues G1-D3 and G7-G10, with the hydrophilic side chains of residues D3, E5, T6, and T8 exposed to the solvent (Figure 3c).

For the simulations of the hairpin RNAs with GCAA, GAAA, and UUCG tetraloops, the total number of folding and unfolding events captured by AIMBD simulations reduced from the GCAA to GAAA to UUCG simulation system, which was in good agreement with previous studies by Tan et al.[21] and Chen et al.[22]. This also demonstrated the importance of the base stacking within the tetraloop for RNA folding. In particular, while nucleotides C6-A7-A8 of the GCAA tetraloop and A6-A7-A8 of the GAAA tetraloop base-stacked in their respective "Folded" low-energy conformations, no base stacking was observed within the "Folded" hairpin RNA with UUCG tetraloop (Figures 4c–6c). Furthermore, the folding mechanisms uncovered by DBMD were similar among the hairpin RNAs with GCAA, GAAA, and UUCG tetraloop (Figures 4–6). Starting from the extended conformation in the "Unfolded" low-energy conformational states (Figures 4e, 5f, and 6e), Watson-Crick base pairs began to form from terminal nucleotides towards the cores and tetraloops of the RNAs. Finally, base stacking between the nucleotides of the tetraloops were formed to enable the stable folding of the hairpin RNAs (Figures 4c and 5c). This general mechanism of RNA folding showed high degrees of similarity to the previous study by Chen et al.[22], even though they used shorter RNA strands, a different force field parameter set, and a different solvation model.

In conclusion, we have developed DBMD, a DL-based enhanced sampling technique that allows for accurate energetic reweighting and enhanced sampling of biomolecular systems. DBMD is available with open source in OpenMM at https://github.com/MiaoLab20/DBMD/. As demonstrated on the model systems, DBMD captured multiple dihedral transitions of alanine dipeptide as well as folding-unfolding events of the chignolin protein and hairpin RNAs with tetraloops within relatively short simulation lengths. The performances of DBMD in OpenMM on simulations of larger biological systems, including membrane proteins, and with other force fields besides AMBER will be examined in future studies. DBMD is expected to facilitate the simulations and free energy calculations of a wide range of biomolecules.

Methods

Theory of DBMD

In DBMD, boost potentials ΔV are optimized using DL to follow Gaussian distribution with minimized anharmonicity. Considering a system comprised of N atoms with coordinates $r \equiv \{\vec{r}_1, \ldots, \vec{r}_N\}$ and momenta $p \equiv \{\vec{p}_1, \ldots, p_N\}$, the system Hamiltonian can be expressed as:

$$H(r, p) = K(p) + V(r), \tag{1}$$

where $K(p)$ and $V(r)$ are the system kinetic and total potential energies, respectively. To enhance biomolecular conformational sampling, boost potentials can be added to the system potential energies. According to the DBMD algorithm, the boost potential can be calculated as the following[7]:

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

where E is the reference energy for adding boost potential and k is the harmonic force constant. Here, the reference energy can be set in a range: $V_{max} \leq E \leq V_{min} + \frac{1}{k}$. The harmonic force constant is calculated as $k = \frac{k_0}{V_{max} - V_{min}}$, with the effective harmonic force constant $k_0 \in (0,1]$. Accordingly, the reference energy can be expressed as $E = V_{min} + \frac{V_{max} - V_{min}}{k_0}$. Here, $E = V_{max}$ when $k_0 = 1$, and the smaller the k_0 values, the higher the reference energy E. In DBMD, we introduce a parameter called the reference energy factor (η) valued between 0 and 1 to avoid exceedingly large E and control the acceleration during simulations. Physically, $V_{max} + \eta * |V_{max}|$ represents the upper limit of the reference energy E.

$$\begin{gathered} E = V_{min} + \frac{V_{max} - V_{min}}{k_0} \\ if\ E > V_{max} + \eta * |V_{max}|,\quad then:\ E = V_{max} \end{gathered} \tag{3}$$

Therefore, the boost potential can be rewritten as:

$$\Delta V(r) = \begin{cases} \frac{k_0}{2(V_{max} - V_{\min})}(E - V(r))^2, \; V(r) < E \\ 0, \; V(r) \geq E. \end{cases} \tag{4}$$

To characterize the extent to which ΔV follows a Gaussian distribution, its distribution anharmonicity γ is calculated as:

$$\gamma = S_{max} - S_{\Delta V} = \frac{1}{2}\ln\left(2\pi e\sigma_{\Delta V}^2\right) + \int_0^\infty p(\Delta V)\ln(p(\Delta V))d\Delta V, \tag{5}$$

where ΔV is dimensionless as divided by $k_B T$ with k_B and T being the Boltzmann constant and system temperature, respectively, and $S_{max} = \frac{1}{2}\ln\left(2\pi e\sigma_{\Delta V}^2\right)$ is the maximum entropy of ΔV[8]. When γ is zero, ΔV follows exact Gaussian distribution with sufficient sampling. Reweighting by approximating the ensembled-averaged Boltzmann factor with cumulant expansion to the 2nd order ("Gaussian approximation") can accurately recover the original free energy landscape[7, 23]. As γ increases, the ΔV distribution becomes less harmonic, and the reweighted free energy profile obtained from cumulant expansion to the 2nd order would deviate from the original[7]. The anharmonicity of ΔV distribution serves as an indicator of the enhanced sampling convergence and accuracy of the reweighted free energy[7].

Deep Learning of Potential Energies

In DBMD, the probabilistic Bayesian neural network model, developed based on Bayes' theorem[24], within the *TensorFlow Probability*[25] module was applied to minimize the anharmonicity of boost potentials ΔV. The probabilistic model was initiated with the definition of a prior distribution. A standard normal distribution was adopted as the prior distribution since the central limit theorem asserts that a properly normalized sum of samples will approximate a normal distribution[26, 27]. Here, a multivariate normal distribution with a diagonal covariance matrix was used, with the mean values initialized to zero and the variances σ_i^2 to one[26].

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \ldots \\ 0 & \sigma_2^2 & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}. \tag{6}$$

The posterior distribution was also set to be a multivariate Gaussian distribution, but the off-diagonal elements in the covariance matrix were allowed to be non-zero. This was achieved with a lower-triangular matrix $\mathbf{L}$ with positive-valued diagonal entries such that $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, and the triangular matrix can be obtained through Cholesky decomposition of the covariance matrix[26].

$$\mathbf{L} = \begin{pmatrix} L_{11} & 0 & \ldots \\ L_{21} & L_{22} & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}. \tag{7}$$

Finally, the probabilistic layers were defined using the *DenseVariational* function of the *TensorFlow Probability* module[25–27]. Our Bayesian neural network model consisted of two or four dense variational layers of two different types, namely *L1* and *L2*. The first dense variational layer *L1* had 64 filters, with a sigmoid activation function to enable the fitting of non-linear data[26, 27]. The second dense variational layer *L2* used the *IndependentNormal*[25] function to parameterize a normal distribution and capture aleatoric uncertainty, with an event shape equal to one[26, 27]. The prior and posterior distributions used in both *L1* and *L2* were specified above. Testing simulations have showed us that the number of the second dense variational layer *L2* could significantly affect the average and standard deviation of the output boost potentials after DL. Overall, the lower the numbers of *L2*, the wider the distributions and the higher the average boost potentials. Therefore, to balance between the stability and sampling of the simulations as well as the learning speed, we included one *L2* layer in the DL model for explicit-solvent simulations and three *L2* layers for implicit-solvent simulations. The input and output shape were set to one since both the potential energies and boost potentials were scalars.

Workflow of DBMD

The workflow of DBMD is shown in Figure 1. First, a short cMD was performed on the biological system of interest, and the potential statistics (V_{min} and V_{max}) were collected as parameters for pre-equilibration of DBMD simulation. During the pre-equilibration, the effective harmonic force constants (k_{0P} and k_{0D}) were kept fixed at (1.0, 1.0) for explicit-solvent simulations and (0.05, 1.0) for implicit-solvent simulations. The boost potentials were calculated based on equation (4), and the potential statistics (V_{min} and V_{max}) were updated during pre-equilibration. The system total and dihedral potential energies from the pre-equilibration were then collected (Figure 1a), which served as the X inputs for the probabilistic Bayesian DL models[25, 26] (Figure 1b). Initial boost potentials were randomly generated from the system potential energies and randomly assigned k_0 using equations (3–4) and used as the Y inputs for DL (Figure 1c). DL was carried out in multiple iterations until the output boost potentials followed Gaussian distribution with anharmonicity $\gamma < 0.01$ (Figure 1d). If $\gamma \geq 0.01$, the generated boost potentials were used as Y inputs to retrain the DL model until $\gamma < 0.01$ (Figure 1e). Based on the potential statistics learnt until the last frame of the pre-equilibration (V_{min}, V_{max}, V, and ΔV), the effective harmonic force constants were calculated as following:

$$\begin{aligned} &k_0 = \left(\frac{\sqrt{2\Delta V(V_{max} - V_{min})} - \sqrt{2\Delta V(V_{max} - V_{min}) - 4(V_{min} - V)(V_{max} - V_{min})}}{2(V_{min} - V)} \right)^2 \\ &if\ k_0 > 1\ or\ E > V_{max} + \eta * |V_{max}|,\ \ then: \\ &k_0 = \frac{2\Delta V(V_{max} - V_{min})}{(E - V)^2} \\ &k_0 = \min(1.0,\ k_0)\,. \end{aligned} \tag{8}$$

and used as input alongside V_{min} and V_{max} to equilibrate the simulation system (Figure 1f). The equilibration usually consisted of multiple rounds, with the effective harmonic force constants (k_{0P} and k_{0D}) kept fixed and potential statistics (V_{min} and V_{max}) updated in each round. DL was carried out at the end of each round using the updated potential energies as inputs, with the same DL model as obtained at the end of the pre-equilibration (Figure 1). This multi-round equilibration approach also allows users to select a simulation checkpoint that suits their simulation goals the most to go into the production simulations, giving users some control over their simulation outcomes. Finally, the effective harmonic force constants (k_{0P} and k_{0D}) and potential statistics (V_{min} and V_{max}) taken from the last round of the equilibration were used as input parameters for DBMD production simulations (Figure 1f), during which the effective harmonic force constants and potential statistics were kept fixed, and boost potentials were calculated based on equation (4).

System Setup and Simulation Protocols

Simulations of the alanine dipeptide and chignolin were performed using the AMBER ff99SB force field parameter set[28]. The LEaP module in the AmberTools package[28] were used to build the simulation systems. For the DBMD simulations in explicit solvent, alanine dipeptide was solvated in a TIP3P[29] water box that extended ~8 Å from the solute surface. The unfolded chignolin with a sequence of 10 residues (GYDPETGTWG)[16] was solvated in a TIP3P[29] water box that extended ~10 Å from the solute surface. The final system for alanine dipeptide in explicit solvent, alanine dipeptide in implicit solvent, and chignolin in explicit solvent contained 1912, 22, and 6773 atoms, respectively.

Simulations of the hairpin RNAs with the GCAA, GAAA, and UUCG tetraloops were carried out using the AMBER Shaw force field parameter set[21], starting from their unfolded states. The sequences of the hairpin RNAs with GCAA, GAAA, and UUCG tetraloops were GGGCGCAAGCCU (12 nucleotides)[17], CGGGGAAACUUG (12 nucleotides)[18], and GGCACUUCGGUGCC (14 nucleotides)[19], respectively. The final systems of the hairpin RNAs with GCAA, GAAA, and UUCG tetraloops in implicit solvent contained 389, 390, and 447 atoms, respectively. All simulations were carried out at 300K temperature.

For the explicit-solvent simulations, periodic boundary conditions were applied, and bonds containing hydrogen atoms were restrained with the SHAKE[30] algorithm. Weak coupling to an external temperature and pressure bath was necessary to control both temperature and pressure[31]. The electrostatic interactions were calculated using the particle mesh Ewald (PME) summation[32] with a cutoff of 8.0–9.0 Å for long-range interactions. For the implicit-solvent simulations, the generalized Born solvent model 2 (GBn2)[33] parameters were used. No nonbonded cutoff was set and no periodic boundary condition was used in the implicit-solvent simulations. The solute and solvent dielectric constants were set to 1.0 and 78.5, respectively, and the effect of a non-zero salt concentration was achieved by setting the Debye-Huckel screening parameter[34] to 1.0/nm. A 2-fs timestep with the SHAKE[30] algorithm applied was used in all simulations.

For alanine dipeptide, the simulations consisted of a 2ns short cMD, followed by a 2ns DBMD pre-equilibration, one round of 2ns DBMD equilibration, and three independent 30ns DBMD production simulations. The reference energy factors were set to zero for both total and dihedral potential energy (η_P and η_D), i.e., $E = V_{max}$. For chignolin, the simulation involved a 5ns cMD, a 2ns DBMD pre-equilibration, two rounds of 5ns DBMD equilibration, and three independent 300ns DBMD production simulations, with η_P and η_D both set to 0.05. For the hairpin RNAs with GCAA, GAAA, and UUCG tetraloops, the simulations consisted of a 20ns cMD, followed by a 5ns DBMD pre-equilibration, three rounds of 5ns DBMD equilibration, and three-four independent 2µs DBMD production simulations. η_P and η_D were set to 0.05 and 0.05 for GCAA, 0.05 and 0.0 for GAAA, and 0.0 and 0.0 for UUCG RNA tetraloops. The simulation frames were saved every 0.1 ps. The CPPTRAJ[35] tool was used for simulation trajectory analysis.

Finally, the PyReweighting toolkit[8] was used to compute the potential of mean force (PMF) profiles of the backbone dihedrals Phi and Psi (Φ and Ψ) in the alanine dipeptide (Figure 2a). The C_α-atom root-mean-square deviation (RMSD) of residues Y2-W9 of chignolin relative to the 1UAO[16] PDB and C_α-atom radius of gyration (Rg) of residues Y2-W9 were selected as RCs to calculate the PMF profiles in the simulations of chignolin folding. The heavy-atom RMSD of the whole hairpin RNAs with tetraloops relative to respective PDB structures (1ZIH[17] for GCAA, 2ADT[18] for GAAA, and 2KOC[19] for UUCG) and the G1-U12, C1-G12, and G1-C14 center-of-mass (COM) distances were used as RCs to calculate the PMF profiles in the simulations of hairpin RNAs with tetraloops. A bin size of 6°, 1.0 Å, and 1.0–2.0 Å and cutoff of 10, 100, and 100–500 in one bin were used for reweighting of DBMD simulations of alanine dipeptide, chignolin, and hairpin RNAs with tetraloops, respectively.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Matthew Copeland and Dr. Jinan Wang for the valuable discussions. This work used supercomputing resources with allocation award TG-MCB180049 through the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296, 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 and BigJay Cluster funded through NSF Grant MRI-2117449 at the University of Kansas. This work was supported by the National Institutes of Health (R01GM132572).

References

(1). Karplus M; McCammon JA Molecular dynamics simulations of biomolecules. Nature Structural Biology 2002, 9 (9), 646–652. DOI: Doi 10.1038/Nsb0902-646. [PubMed: 12198485]

(2). Hollingsworth S; Dror R Molecular dynamics simulation for all. Neuron 2018, 99, 1129–1143. [PubMed: 30236283]

(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. DOI: 10.1002/wcms.1521. [PubMed: 34899998]

(4). Harvey MJ; Giupponi G; Fabritiis GD ACEMD: accelerating biomolecular dynamics in the microsecond time scale. Journal of Chemical Theory and Computation 2009, 5, 1632–1639. [PubMed: 26609855] Shaw D. E. e. a. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation Before Lunch. SC'21: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis 2021.

(5). Henzler-Wildman K; Kern D Dynamic personalities of proteins. Nature 2007, 450, 964–972. [PubMed: 18075575]

(6). Spiwok V; Sucur Z; Hosek P Enhanced sampling techniques in biomolecular simulations. Biotechnology Advances 2015, 33, 1130–1140. [PubMed: 25482668] Gao YQ; Yang LJ; Fan YB; Shao Q Thermodynamics and kinetics simulations of multi-timescale processes for complex systems. International Reviews in Physical Chemistry 2008, 27, 201–227.Liwo A; Czaplewski C; Oldziej S; Scheraga HA Computational techniques for efficient conformational sampling of proteins. Current Opinion in Structural Biology 2008, 18, 134–139. [PubMed: 18215513] 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, 157–166. [PubMed: 17570138] Miao Y; McCammon JA Unconstrained enhanced sampling for free energy calculations of biomolecules: a review. Molecular Simulation 2016, 42, 1046–1055. [PubMed: 27453631]

(7). Miao Y; Feher VA; McCammon JA Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. Journal of Chemical Theory and Computation 2015, 11, 3584–3595. [PubMed: 26300708]

(8). Miao Y; Sinko W; Pierce L; Bucher D; Walker RC; McCammon JA Improved reweighting of accelerated molecular dynamics simulations for free energy calculation. Journal of Chemical Theory and Computation 2014, 10, 2677–2689. [PubMed: 25061441]

(9). Pang Y; Miao Y; McCammon JA Gaussian accelerated molecular dynamics in NAMD. Journal of Chemical Theory and Computation 2017, 13, 9–19. [PubMed: 28034310]

(10). Copeland MC; Do HN; Votapka L; Joshi K; Wang J; Amaro RE; Miao Y Gaussian accelerated molecular dynamics in OpenMM. The Journal of Physical Chemistry B 2022, 126 (31), 5810–5820. [PubMed: 35895977]

(11). 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, 5199–5208. [PubMed: 31539245]

(12). Celerse F; Inizan TJ; 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. [PubMed: 35080892]

(13). Lee H; Turilli M; Jha S; Bhowmik D; Ma H; Ramanathan A DeepDriveMD: Deep-Learning Driven Adaptive Molecular Simulations for Protein Folding. 2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS) 2019, 12–19. DOI: 10.1109/DLS49591.2019.00007.

(14). Mehdi S; Wang D; Pant S; Tiwary P Accelerating All-Atom Simulations and Gaining Mechanistic Understanding of Biophysical Systems through State Predictive Information Bottleneck. Journal of Chemical Theory and Computation 2022, 18, 3231–3238. [PubMed: 35384668]

(15). Wang Y; Herron L; Tiwary P From data to noise to data for mixing physics across temperatures with generative artificial intelligence. PNAS 2022, 119 (32), e2203656119. [PubMed: 35925885]

(16). Honda S; Yamasaki K; Sawada Y; Morii H 10 Residue Folded Peptide Designed by Segment Statistics. Structure 2004, 12, 1507–1518. [PubMed: 15296744]

(17). Jucker FM; Heus HA; Yip PF; Moors EH; Pardi A A network of heterogeneous hydrogen bonds in GNRA tetraloops. Journal of Molecular Biology 1996, 264 (5), 968–980. [PubMed: 9000624]

(18). Davis JH; Tonelli M; Scott LG; Jaeger L; Williamson JR; Butcher SE RNA Helical Packing in Solution: NMR Structure of a 30 kDa GAAA Tetraloop-Receptor Complex. Journal of Molecular Biology 2005, 351, 371–382. [PubMed: 16002091]

(19). Nozinovic S; Furtig B; Jonker HRA; Richter C; Shawalbe H High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. Nucleic Acids Research 2010, 38 (2), 683–694. [PubMed: 19906714]

(20). Miao Y; Feixas F; Eun C; McCammon JA Accelerated Molecular Dynamics Simulations of Protein Folding. Journal of Computational Chemistry 2015, 36, 1536–1549. [PubMed: 26096263]

(21). Tan D; Piana S; Dirks RM; Shaw DE RNA force field with accuracy comparable to state-of-the-art protein force fields. Proc Natl Acad Sci USA 2017, 115 (7), E1346–E1355.

(22). Chen AA; Garcia AE High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. PNAS 2013, 110 (42), 16820–16825. [PubMed: 24043821]

(23). 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. [PubMed: 35200019]

(24). Bayes T An Essay Toward Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London 1764, 53, 370–418.

(25). Dillon JV; Langmore I; Tran D; Brevdo E; Vasudevan V; Moore D; Patton B; Alemi A; Hoffman M; Saurous RA Tensorflow Distributions. arXiv preprint 2017, arXiv:1711.10604.

(26). Kamperis S Probabilistic regression with Tensorflow. GitHub 2021.

(27). Blundell C; Cornebise J; Kavukcuoglu K; Wierstra D Weight Uncertainty in Neural Networks. arXiv preprint 2015. DOI: https://arxiv.org/abs/1505.05424.

(28). Case DA; Belfon K; Ben-Shalom IY; Brozell SR; Cerutti DS; Cheatham TE; Cruzeiro VWD; Darden T; Duke RE; Giambasu G AMBER 2020. 2020.Gotz AW; Williamson MJ; Xu D; Poole D; Le Grand S; Walker RC Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. Journal of Chemical Theory and Computation 2012, 8 (5), 1542–1555. [PubMed: 22582031] Salomon-Ferrer R; Gotz AW; Poole D; Le Grand S; Walker RC 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. [PubMed: 26592383] Salomon-Ferrer R; Case DA; Walker RC An overview of the Amber biomolecular simulation package. Wiley Interdiscip Rev Comput Mol Sci 2013, 3 (2), 198–210.

(29). 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.

(30). Ryckaert J-P; Ciccotti G; Berendsen HJC 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.

(31). Berendsen HJC; Postma JPM; Vangunsteren WF; Dinola A; Haak JR Molecular Dynamics with Coupling to an External Bath. Journal of Chemical Physics 1984, 81 (8), 3684–3690.

(32). Essmann U; Perera L; Berkowitz ML; Darden T; Lee H; Pedersen LG A Smooth Particle Mesh Ewald Method. Journal of Chemical Physics 1995, 103 (19).

(33). Nguyen H; Roe DR; Simmerling C Improved generalized Born solvent model parameters for protein simulations. Journal of Chemical Theory and Computation 2013, 9 (4), 2020–2034. [PubMed: 25788871]

(34). Srinivasan J; Trevathan MW; Beroza P; Case DA Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. Theor. Chem. Acc. 1999, 101, 426–434.

(35). Roe DR; Cheatham ITE PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. Journal of Chemical Theory and Computation 2013, 9, 3084–3095. [PubMed: 26583988]

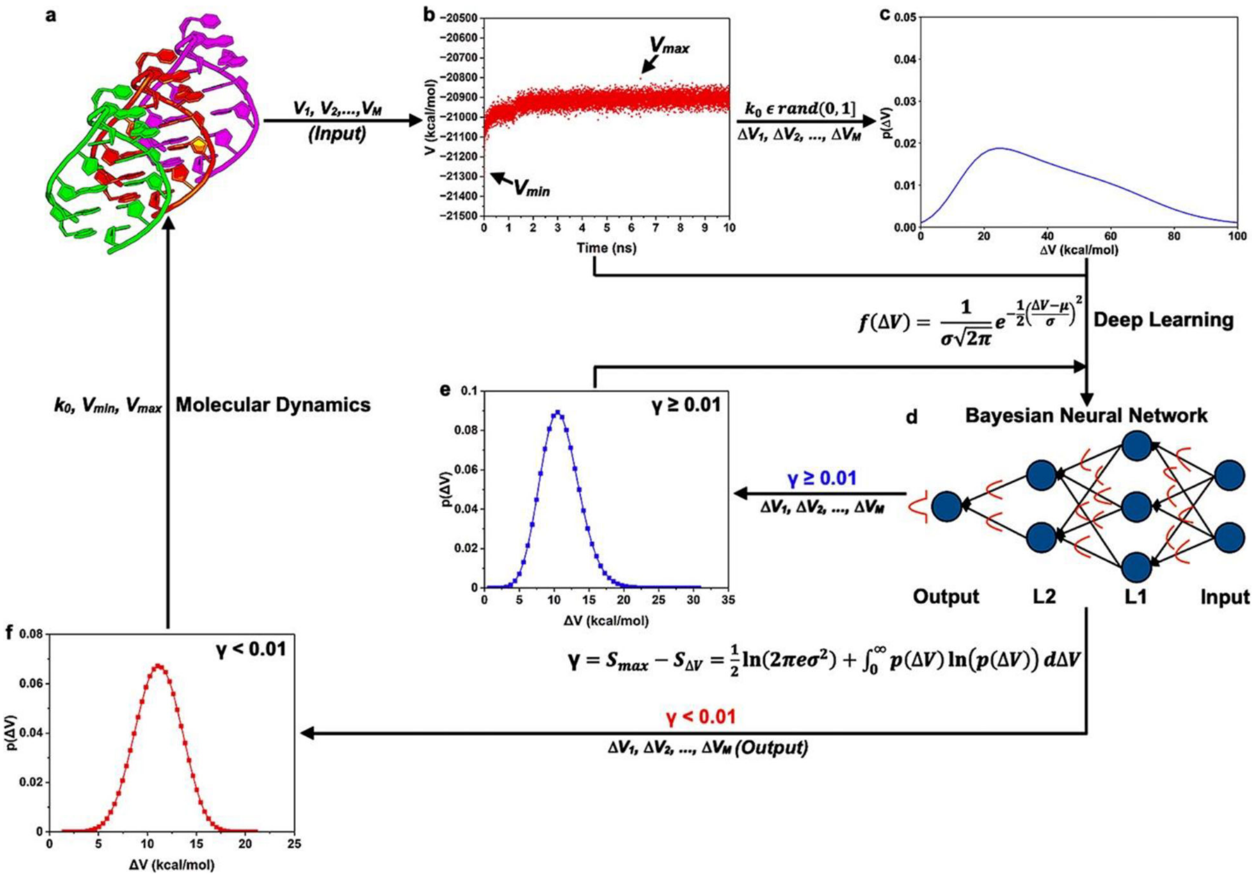


Figure 1. Summary of Deep Boosted Molecular Dynamics (DBMD).

(a) First, molecular dynamics (MD) simulation is performed on the system of interest. **(b)** The system potential energies from finished simulation frames $(V_1, V_2, \ldots, V_M)$ are collected as the X inputs for the probabilistic Bayesian Deep Learning (DL) model. **(c)** Reference boost potentials $(\Delta V_1, \Delta V_2, \ \ldots, \ \Delta V_M)$ were generated from the collected system potential energies and randomized effective harmonic force constants k_0 to serve as the Y inputs for the DL. **(d)** The probabilistic Bayesian neural network was trained to generate boost potentials that follow Gaussian distribution with the probability density function $f(\Delta V)$. Here, ΔV is boost potential, and μ and σ are the average and standard deviation of the boost potentials. DL is carried out in multiple iterations until the anharmonicity of output boost potentials $\gamma < 0.01$. **(e)** If the anharmonicity of output boost potential γ is ≥ 0.01, the generated boost potentials are used as Y inputs to retrain the DL model until $\gamma < 0.01$. **(f)** Finally, the effective harmonic force constants k_0 are calculated from the system potential energy (V_M) and used as input alongside the minimum and maximum of potential energy (V_{min} and V_{max}) **(b)** for the next round of enhanced sampling simulation.

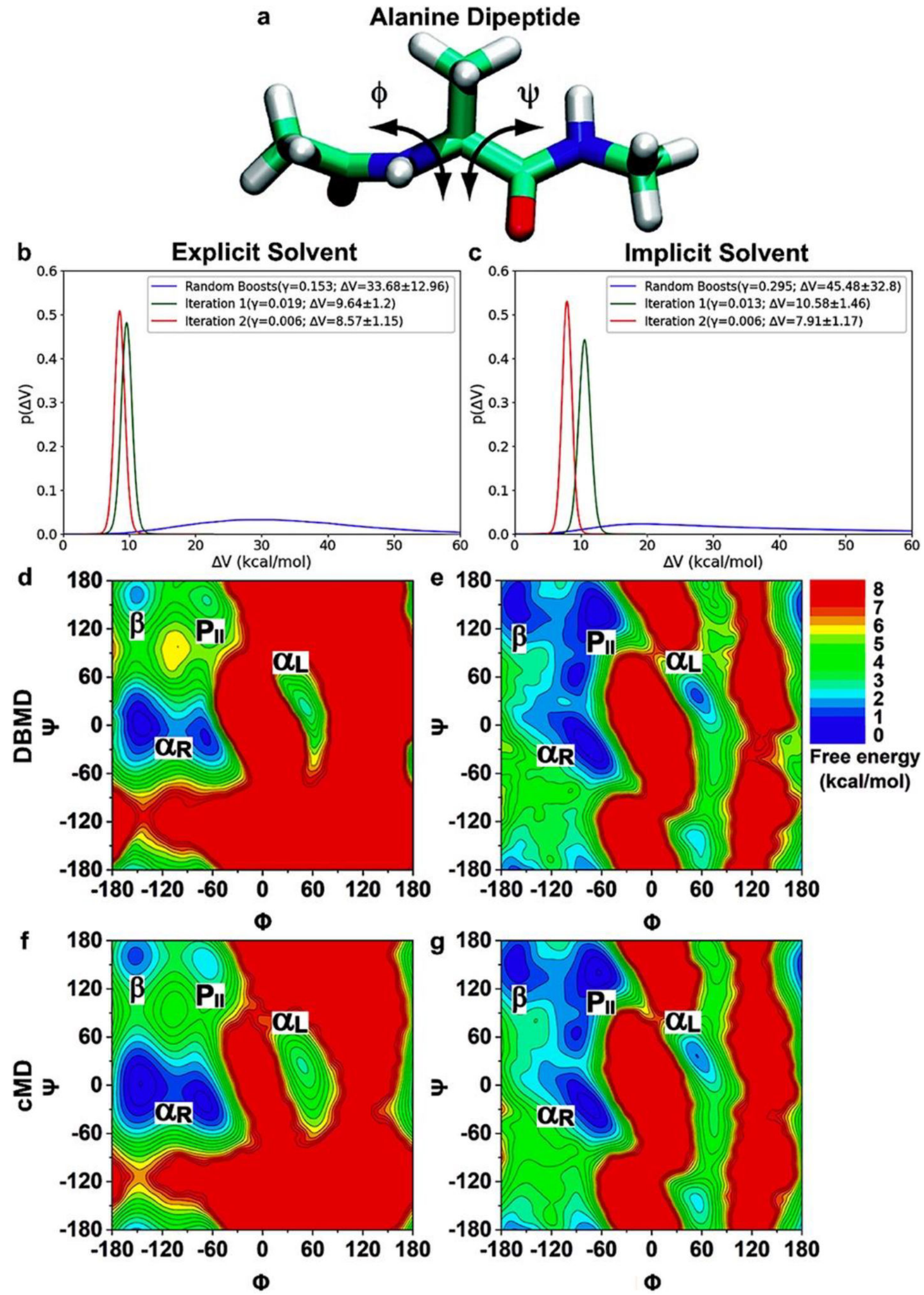


Figure 2. DBMD simulations of alanine dipeptide.

(a) Schematic representation of backbone dihedrals Phi (Φ) and Psi (Ψ) dihedrals of alanine dipeptide. **(b-c)** Representative distributions of randomly generated dual boost potentials and DL-generated boost potentials iterated until $\gamma < 0.01$ from the potential energies collected from the pre-equilibration of the alanine dipeptide in explicit solvent **(b)** and implicit solvent **(c)**. The legends include the anharmonicity and average ± standard deviation of the dual boost potentials. **(d-g)** 2D Potential of mean force (PMF) free energy profile of backbone dihedrals (Φ, Ψ) of alanine dipeptide calculated from three 30ns DBMD simulations **(d-e)** compared to 1μs cMD simulations **(f-g)** in explicit solvent **(d, f)** and implicit solvent **(e, g)**.

The low-energy states 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.

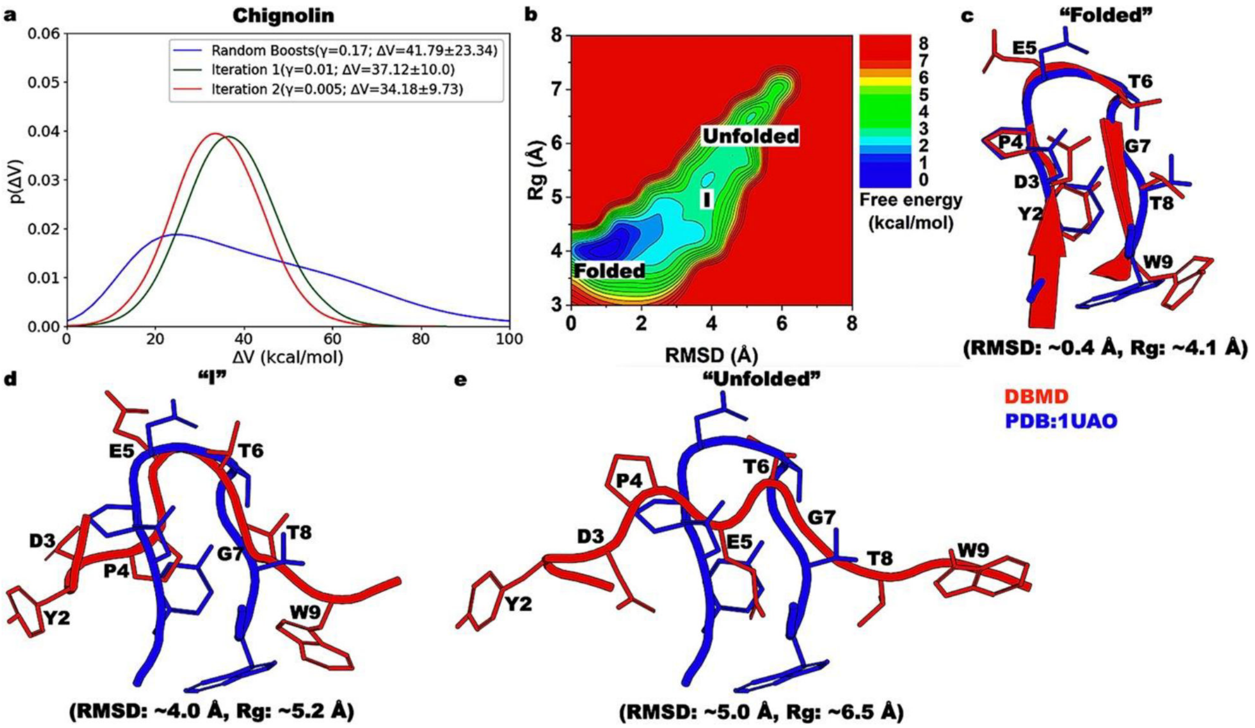


Figure 3. Folding of chignolin in explicit solvent captured by DBMD.

(a) Representative distributions of randomly generated dual boost potentials and DL-generated boost potentials iterated until $\gamma < 0.01$ from the potential energies collected from the pre-equilibration of chignolin. The legends include the anharmonicity and average ± standard deviation of the dual boost potentials. **(b)** 2D PMF free energy profile of the C_α-atom root-mean-square deviation (RMSD) of residues Y2-W9 of chignolin relative to the 1UAO PDB and C_α-atom radius of gyration (Rg) of residues Y2-W9. The low-energy conformational states are labeled "Folded", "I", and "Unfolded". **(c)** The "Folded" low-energy conformational state compared to the 1UAO PDB structure, for which the RMSD is ~0.4 Å and the Rg is ~4.1 Å. **(d)** The intermediate "I" low-energy conformational state compared to the 1UAO PDB structure, for which the RMSD is ~4.0 Å and the Rg is ~5.2 Å. **(e)** The "Unfolded" low-energy conformational state compared to the 1UAO PDB structure, for which the RMSD is ~5.0 Å and the Rg is ~6.5 Å. The low-energy conformational states are colored red, and the 1UAO PDB structure is colored blue.

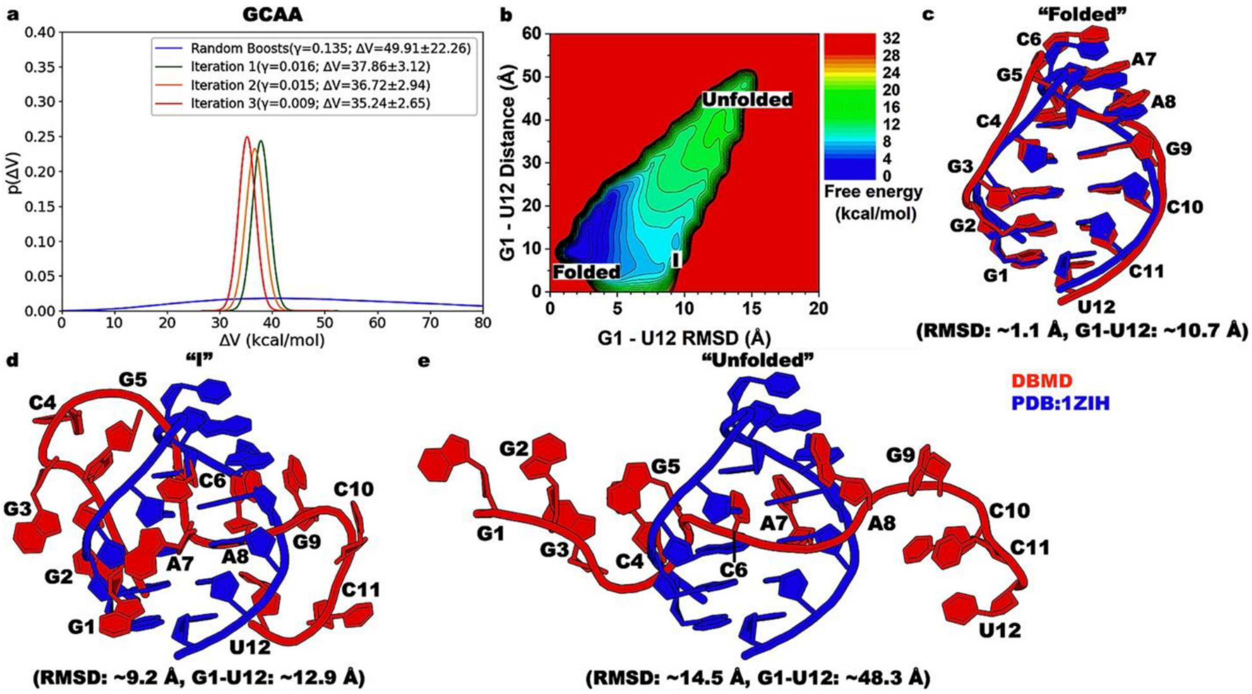


Figure 4. Folding of the 12-mer hairpin RNA with GCAA tetraloop in implicit solvent captured by DBMD.

(a) Representative distributions of randomly generated dual boost potentials and DL-generated boost potentials iterated until $\gamma < 0.01$ from the potential energies collected from the pre-equilibration of the 12-mer hairpin RNA with GCAA tetraloop. The legends include the anharmonicity and average ± standard deviation of the dual boost potentials. **(b)** 2D PMF free energy profile of the heavy-atom RMSD of the 12-mer hairpin RNA relative to the 1ZIH PDB and the center of mass (COM) distance between terminal nucleotides G1 and U12. The low-energy conformational states are labeled "Folded", "I", and "Unfolded". **(c)** The "Folded" low-energy conformational state compared to the 1ZIH PDB structure, for which the RMSD is ~1.1 Å and the G1-U12 distance is ~10.7 Å. **(d)** The intermediate "I" low-energy conformational state compared to the 1ZIH PDB structure, for which the RMSD is ~9.2 Å and the G1-U12 distance is ~12.9 Å. **(e)** The "Unfolded" low-energy conformational state compared to the 1ZIH PDB structure, for which the RMSD is ~14.5 Å and the G1-U12 distance is ~48.3 Å. The low-energy conformational states are colored red, and the 1ZIH PDB structure is colored blue.

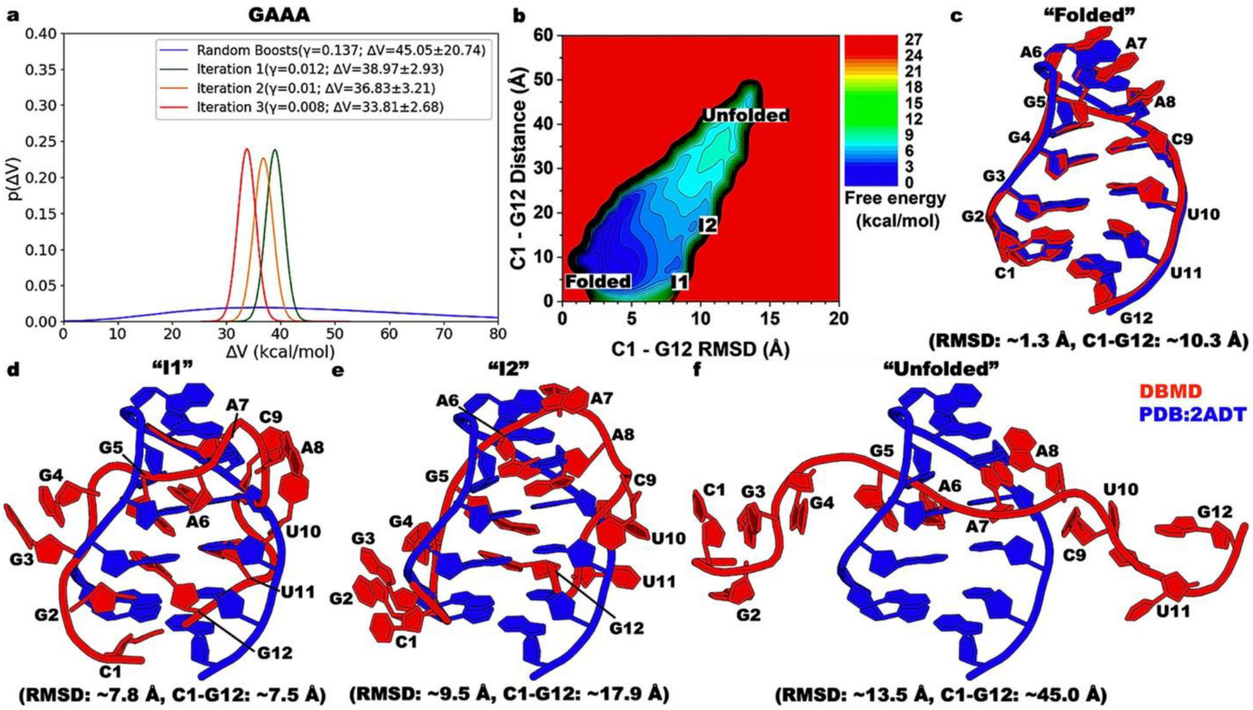


Figure 5. Folding of the 12-mer hairpin RNA with GAAA tetraloop in implicit solvent captured by DBMD.

(a) Representative distributions of randomly generated dual boost potentials and DL-generated boost potentials iterated until $\gamma < 0.01$ from the potential energies collected from the pre-equilibration of the 12-mer hairpin RNA with GAAA tetraloop. The legends include the anharmonicity and average ± standard deviation of the dual boost potentials. **(b)** 2D PMF free energy profile of the heavy-atom RMSD of the 12-mer hairpin RNA relative to the 2ADT PDB and the COM distance between terminal nucleotides C1 and G12. The low-energy conformational states are labeled "Folded", "I1", "I2", and "Unfolded". **(c)** The "Folded" low-energy conformational state compared to the 2ADT PDB structure, for which the RMSD is ~1.3 Å and the C1-G12 distance is ~10.3 Å. **(d)** The intermediate "I1" low-energy conformational state compared to the 2ADT PDB structure, for which the RMSD is ~7.8 Å and the C1-G12 distance is ~7.5 Å. **(e)** The intermediate "I2" low-energy conformational state compared to the 2ADT PDB structure, for which the RMSD is ~9.5 Å and the C1-G12 distance is ~17.9 Å. **(f)** The "Unfolded" low-energy conformational state compared to the 2ADT PDB structure, for which the RMSD is ~13.5 Å and the C1-G12 distance is ~45.0 Å. The low-energy conformational states are colored red, and the 2ADT PDB structure is colored blue.

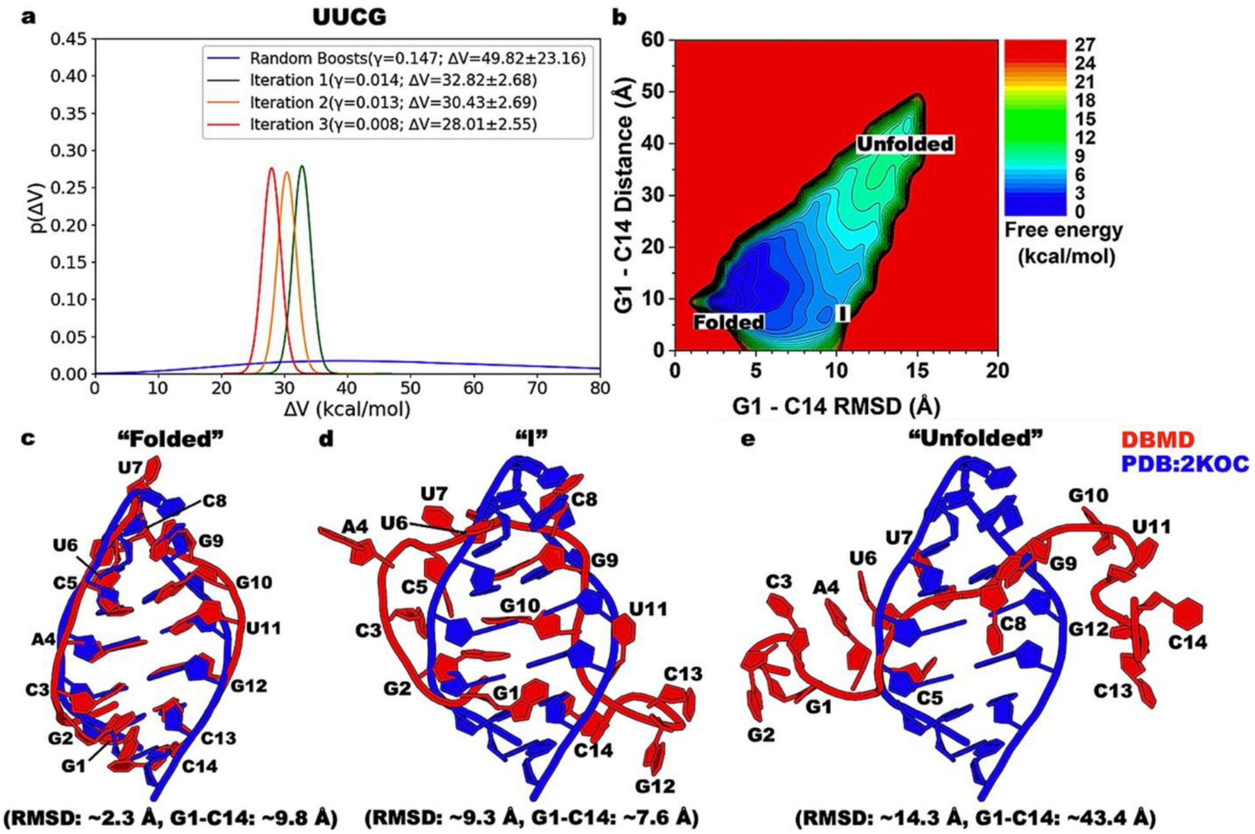


Figure 6. Folding of the 14-mer hairpin RNA with UUCG tetraloop in implicit solvent captured by DBMD.

(a) Representative distributions of randomly generated dual boost potentials and DL-generated boost potentials iterated until $\gamma < 0.01$ from the potential energies collected from the pre-equilibration of the 14-mer hairpin RNA with UUCG tetraloop. The legends include the anharmonicity and average ± standard deviation of the dual boost potentials. **(b)** 2D PMF free energy profile of the heavy-atom RMSD of the 14-mer hairpin RNA relative to the 2KOC PDB and the COM distance between terminal nucleotides G1 and C14. The low-energy conformational states are labeled "Folded", "I", and "Unfolded". **(c)** The "Folded" low-energy conformational state compared to the 2KOC PDB structure, for which the RMSD is ~2.3 Å and the G1-C14 distance is ~9.8 Å. **(d)** The intermediate "I" low-energy conformational state compared to the 2KOC PDB structure, for which the RMSD is ~9.3 Å and the G1-C14 distance is ~7.6 Å. **(e)** The "Unfolded" low-energy conformational state compared to the 2KOC PDB structure, for which the RMSD is ~14.3 Å and the G1-C14 distance is ~43.4 Å. The low-energy conformational states are colored red, and the 2KOC PDB structure is colored blue.