Force Field Dependent DNA Breathing Dynamics: A Case Study of Hoogsteen Base Pairing in A6-DNA

Sharon Emily Stone,[†] Dhiman Ray,[†] and Ioan Andricioaei[*,‡,‡]

†*Department of Chemistry, University of California Irvine, Irvine CA 92697*

‡*Department of Physics and Astronomy, University of California Irvine, Irvine CA 92697*

E-mail: andricio@uci.edu

Abstract

The Hoogsteen (HG) base pairing (bp) conformation, commonly observed in damaged and mutated DNA helix, facilitates DNA repair mechanism and DNA recognition by proteins and small molecules. The free energy difference of HG and Watson-Crick (WC) base pairs has been computed in previous computational studies. But, the mechanism of the conformational transition is not well understood. A detailed understanding of the process of WC to HG base pair transition can provide deeper understanding of DNA repair and recognition. In an earlier study, we explored the free energy landscape for this process using extensive computer simulation with CHARMM36 force field for the nucleic acid. In this work, we study the impact of force field models in describing the WC to HG base pairing transition using the meta-eABF enhanced sampling, quasi-harmonic entropy calculation, and non-bonded energy analysis. The secondary structures of the both base pairing forms and the topology of the free energy landscapes were consistent over different force field models, although the relative free energy, entropy and the interaction energies tend to vary. The relative stability of WC and HG conformation is dictated by a delicate balance between the enthalpic stabilization and the reduced entropy of the structurally rigid HG structure. These findings provide a holistic view on the impact of force fields on the DNA base pair dynamics leading to the formation of HG conformation and will facilitate future computational investigations of DNA repair and recognition mechanism in physiologically relevant conditions.

Introduction

Based on the pioneering X-ray crystallography results of Rosalind Franklin[1] in 1953, Watson and Crick proposed the structure of duplex DNA,[2] a model later popularized as the Watson-Crick (WC) structure. By using the bases in their most plausible tautomeric form, they determined complementary base pairs (bp) between adenine and thymine (A−T), as well as guanine and cytosine (G−C).[2] When the following WC base parings are assembled in

a right-handed helix structure, the overall geometry is referred to as B-DNA. In relatively recent nuclear magnetic relaxation (NMR) $R_{1\rho}$ experiments of free B-DNA, Nikolova et al. showed that both A−T and protonated G−C (G−C$^+$) WC base pairs spontaneously formed the more uncommon Hoogsteen (HG) bp, with a population and lifetime of 0.1-1% and 0.3-1.5 ms, respectively.[3] Additionally, HG A−T bp are more abundant compared with G−C$^+$ bp under solution conditions.

The HG base pairing form was initially observed in the late 1950s using heavy atom X-ray crystallography.[4] HG structures form when the purine base (adenine or guanine) is flipped over 180 degrees with respect to the glycosidic bond, resulting in a *syn* conformation (as opposed to *anti* in WC) resulting in the shortening of thr helix diameter (Figure 1). These structural change creates a chemical environment different from canonical WC bp, and plays a key role in binding to proteins[5,6] and small molecule drugs.[7]

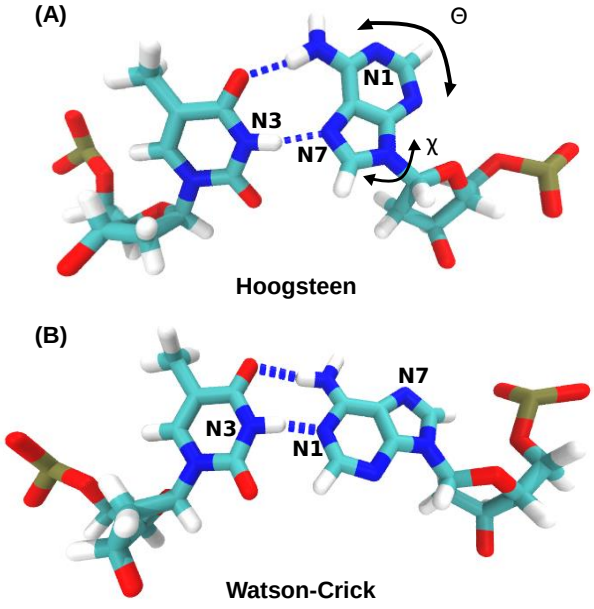

Figure 1: The conformation of (A) Hoogsteen and (B) Watson-Crick A-T base pairs in B-DNA. In the HG bp, the adenine base (on the right side) is flipped over 180 degrees with respect to the glycosidic bond, which results in hydrogen boding with the nitrogen in the five membered ring, and shorter helix diameter. The N1, N3, and N7 nitrogen atoms, referred to in the text, are also marked in the figure.

HG base pairs have also been observed in other scenarios, including damaged DNA,[8–10] DNA duplex conjugated with the DNA repair machinery,[5,11,12] and other diverse DNA-protein interactions.[3,5,13–15] When DNA repair processes does not go up to completion, it can lead to cytotoxic damage, potentially leading to cancer or cell death. For example, the methylation of DNA has been shown to cause blocking of DNA replication, leading to such cytotoxic damages. Recent work by Xu et al. showed that WC bp are vulnerable to methylation in double-stranded DNA, and the sensitivity to methylation increases for the HG bp.[6] Alternatively, a methylated adenine base is more prone to form a HG base pair as a result of stabilizing steric interactions[16] Additionally, HG base pairs have been hypothesized to play an important role in DNA repair mechanisms. For instance, the crystal structure of polymerase-ι (hPol-ι) revealed a replication mechanism dominated by HG base pairs.[17] The structure of hPol-ι encourages replication via the N^2-adducted G, which forms a syn conformation in DNA, and in turn facilitates a binding mechanism through the minor-groove adducts.[17] The HG bp have stabilizes the tumor suppressor p53-DNA complex and enhance its binding affinity.[11] Binding affinity of small molecules like intercalation agents and anticancer drugs such as echinomycine to nucleic acids is also increased in the vicinity of a Hoogsteen base pair.[18] These examples demonstrate the biochemical significance of HG configuration in nucleic acids and encourages detailed studies to understand the mechanism of formation of such non-canonical base pairing structures.

Molecular dynamics (MD) simulation can be used to understand the mechanism of biomolecular processes in atomistic detail. But, from experimental NMR relaxation studies, the timescale of the transition between WC and HG base pairing process has been reported to be 50-250 ms, which is beyond the reach of conventional MD simulation with currently available computing facilities. To this end, the enhanced sampling methods including umbrella sampling,[19] metadynamics,[20,21] meta-eABF,[22] and transition path sampling (TPS)[23] have been employed to sample the conformational space of the WC to HG bp transition. For example, Yang et al., computed the free energy landscape of the WC to HG transition

in A6-DNA, using umbrella sampling and multiple walker well-tempered metadynamics, and captured multiple transition pathways and observed the spontaneous changes in base flipping.[24,25] Similarly, studies from our group used a fast converging method called meta-eABF[22] to obtain the potential of mean force (PMF) for the same transition process with CHARMM36 force field parameters.[26] Our free energy landscape revealed a stable open base pair conformation adjacent to the WC base paired state. The mechanistic analysis using Markov State Modeling[27,28] also revealed a melted base pair state to be the primary intermediate in the transition.[26] This highly stable extra-helical conformation is inconsistent with the conventional idea about the stability of DNA, and provokes further discussion on the accuracy of force fields in describing the base pairing intermediates and transition pathways. Path sampling studies by Vreede et al.[29] and Hooft et al.[30] also revealed that partial opening of the A-T base pair is necessary for the conformational switching, but no stable open base paired intermediate conformation was observed.

The force field is an important component of MD simulations, since the accuracy of our results are dependent on the accuracy of the force field model. The force field is the parameter set used to represent the potential energy function of a biomolecule. Gradients of this potential energy function are used to propagate dynamical trajectories by solving Newton's equation of motion.[31] For nucleic acids, such as DNA and RNA, the commonly used non-polarizable force fields are the CHARMM (CHARMM27, CHARMM36) and AMBER (AMBERbsc0, AMBERbsc1) family models.[32] Different models have different refinements in order to improve the description of nucleic acid properties. For example, the CHARMM27 model has refined backbone dihedrals, sugar puckering, and glycosidic linkage to match the results of model compounds computed with *ab-initio* calculations.[33] But it can model the groove width, population of the BI backbone in DNA, and opening of WC bp with limited accuracy. For CHARMM36, improvements were focused on refining backbone dihedral angles, sugar puckering, and the 2'-hydroxyl dihedral for RNA parameters.[32] In the AMBER family, AMBERbsc0 was focused on refining the backbone conformations (α and γ dihedrals), in

order to lower overpopulated α gauche+ and γ trans conformations.[34,35] Additionally, the AMBERbsc1 model was constructed with AMBERbsc0, but with refined sugar puckering and χ, ϵ, and ζ dihedrals.[36] These refinements have become an essential process in order to obtain accurate results in the microsecond timescale. Comparative studies of the common nucleic acid force fields have been performed to evaluate the mesoscale properties of DNA,[37] and to reproduce NMR structures of the Drew-Dickerson dodecamer (DDD) system.[35] Minhas et al. found the CHARMM27 model to have the most stable trajectories and best agreement to the experimental data for describing the mesoscale properties of multi-microsecond simulations of long DNA fragments (40 base pairs) in a cubic solvation box of TIP3P water.[37] Additionally, the CHARMM27 model had the best agreement with X-ray diffraction and NMR results compared with CHARMM37, AMBERbsc0, and AMBERbsc1 models. While the study revealed both AMBER family models to have good agreement with the experimental data, the CHARMM36 model had unstable trajectories in the microsecond timescale and produced irreversible fluctuations.[37]

Whether the choice of force field can affect the pathways and free energy landscape of the conformational transition between the WC and HG base pairing forms is not yet clear. But such understanding is vital, considering the important physiological role of the HG base paring in DNA repair and replication. In the current work, we address this question by comparing the underlying free energy landscape of WC to HG transition, and the energetic and entropic stabilization of individual base pairing states for the four different force fields: CHARMM27, CHARMM36, AMBERbsc0 and AMBERbsc1. Although newer nucleic acid force fields are available,[35,38] our choice of force fields was inspired by the recent comparative study by Minhas et al.,[37] and molecular dynamics studies of HG base pair formation in the literature. The free energy surface and associated thermodynamic properties of the HG conformation of the A16−T9 base pair in A_6-DNA[3] are explored through enhanced sampling simulation using the advanced meta-eABF approach. The potentials of mean force (PMF) obtained from the meta-eABF simulations are used to compare the force field models in

terms of the minimum free energy pathway for the WC-HG transition. Presence of HG and WC base pairs is observed in all the four models, although the relative stability of each base pairing configuration can be somewhat different. To obtain a fundamental rationale behind this discrepancy, we computed the quasi-harmonic entropy and the root-mean-square fluctuations (RMSF) of the two base pairing conformations from equilibrium trajectories for all force field combinations. Based on our findings, we explain the main distinctions between each force fields in describing the molecular details of WC and HG base pairing conformations and compare our findings with experimental results.

Computational Methods

System Preparation and Equilibration

We performed our simulations on the A6-DNA fragment (PDB ID: 5UZF)[39] reported in the NMR relaxation study by Nikolova et al.[3] This structure includes six A-T base pairs in a 12 bp sequence (CGATTTTTTGGC). The topology files for the CHARMM and AMBER family force fields were prepared using Visual Molecular Dynamics (VMD)[40] and AmberTools (tleap)[41] respectively. The initial structure of A6-DNA was placed in a solvated box with TIP3P water[42] and 22 Na+ ions to neutralize the overall system. A water padding of 17Å was included in each direction. Each solvated system comprised of ∼26000 atoms. The structures were minimized for 50,000 steps using the conjugate gradient algorithm, followed by a slow heating to reach 298K temperature at a rate of 1 K/ps in the NVE ensemble with harmonic restraints of 3 kcal mol^{-1} $Å^{-2}$ applied on the heavy atoms. These restraints were then removed over a period of 1.2 ns, at a rate of 0.5 kcal mol^{-1} per 200 ps. Then, the system was placed in a NVE Ensemble for 3 ns with harmonic restraint with force constant of 0.1 kcal $mol^{-1}Å^{-2}$, followed by an unrestrained NPT simulation for 10 ns. The terminal base pairs were restrained with a force constant of 0.05 kcal $mol^{-1}Å^{-2}$ for all equilibration, production and enhanced sampling simulations. The temperature was kept constant at 298 K using the

Langvein thermostat with a coupling constant 1 pm^{-1}. The pressure was controlled to be at 1 atm, using the Nosè-Hoover Langvein piston with piston decay of 50 fs and oscillation period of 100 fs.[43,44] Identical protocol was used for all four force field models.

Enhanced Sampling Simulation

To accelerate the rare transitions between the WC and HG form, we performed enhanced sampling simulations using a combination of metadynamics[20] and extended-system adaptive biasing force (eABF),[45] commonly known as the meta-eABF approach.[22] Meat-eABF simulations were performed in NAMD 2.14 using the colvars module.[46] The 2-dimensional potential mean of force (PMF) was calculated along the following collective variables (CV): the glycosidic angle χ and a pseudo-dihedral angle Θ of the A16-T9 base pair.[24] The Θ angle characterizes the base flipping into the major and minor grooves of DNA, while the χ dihedral angle characterizes the rotation of the adenine base with respect to the deoxyribose sugar (O4',C1',N9, and C4). A pictorial representation of these two torsion angles are provided in the following reference 25. We started the meta-eABF simulations from the end points of NPT equilibration and propagated for ∼200-400 ns, until the PMFs were sufficiently converged. Gaussian hills of height 0.06 kcal/mol and width 3 CV unit (15°) were deposited every 2 ps for the metadynamics part. Additionally, the ABF bias was applied against the average force after 1000 samples were collected for the corresponding bin of width 5°. The MEPSA program[47] was used to obtain the minimum free energy pathway from the 2D energy landscape obtained from enhanced sampling simulation.

Equilibrium Simulations for Entropy and Energy Calculation

Structures of the WC and HG base paring conformations were sampled from the meta-eABF trajectories based on the hydrogen bond donor acceptor distances. Total 8 structures for four different force fields were used as starting points for 100 ns unbiased MD simulations with identical simulation setup as the meta-eABF simulation. The coordinates were saved every

10 ps. The hydrogen bonding donor distances between atoms N1(A)-N3(T) and N7(A)-N3(T), were monitored to ensure WC and HG conformations remained intact throughout the simulation (see Supporting Information). These trajectories were used to calculate the root-mean-square fluctuation (RMSF) and quasi-harmonic entropy.[48] Five 50 ns trajectory segments were extracted from the last 90 ns of each trajectory using overlapping time windows: 10 ns - 60 ns, 20 ns - 70 ns, 30 ns - 80 ns, 40 ns - 90 ns, 50 ns - 100 ns. The entropy and RMSF were computed from each segment, and the mean and 95% confidence intervals are reported for the 5 data points.

The RMSF of all heavy atoms were computed using GROMACS 2018.1 software[49] with the `gmx rmsf` tool. Details of the calculation of quasiharmonic entropy from the covariance matrix of atomic fluctuation is provided in the original paper.[48] In brief, the entropy is given by,

$$S = k_B \sum_{i}^{3N-6} \frac{\hbar\omega_i/k_BT}{\exp\left(\hbar\omega_i/k_BT\right) - 1} - \ln(1 - \exp\left(\hbar\omega_i/k_BT\right)), \tag{1}$$

where k_B is Boltzmann constant, $\hbar$ is $h/2\pi$ where h is Planck's constant and T is absolute temperature. The frequencies ω_i are obtained as $\omega_i = \sqrt{k_BT/\lambda_i}$, for $i = 1, 2, 3, ..., 3N - 6$ where λ_i are eigenvalues of the mass weighted covariance matrix σ obtained from MD simulation. The number of atoms involved in the calculation is N, which equals to the number of heavy atoms in our case. The entropy is dominated by slow vibrational modes, so we excluded the fast moving hydrogen degrees of freedom from our analysis. The covariance matrices and their eigenvalues were computed using the `gmx covar` module of GROMACS 2018.1 package.[49] The entropy was then evaluated using equation 1 by an in house python code. Non-bonded interaction energies between bases and nucleotides were calculated using the NAMDEnergy module of VMD.[50]

Results and Discussions

Free Energy Surfaces

Similar to our previous study,[26] the meta-eABF approach could produce the PMFs within a few hundreds of nanoseconds, at significantly less computational cost compared to other approaches,[24,25] facilitating the comparative study between multiple force fields. The free energy surfaces obtained for different models are shown in Figure 3. In all models, the Watson-Crick (WC) base paired configurations were present near $\Theta \sim 0°$ and $\chi \sim$ -100°, and the Hoogsteen (HG) form near $\Theta \sim 0°$ and $\chi \sim 60°$. The free energy difference between the WC and HG base pair ($\Delta G_{WC \rightarrow HG}$) was determined to be $\sim$ 8 kcal/mol for CHARMM27, $\sim$ 6 kcal/mol for CHARMM36, $\sim$ 1 kcal/mol for AMBERbsc1, and $\sim$ 0 kcal/mol for AMBERbsc0. When comparing the NMR value, 3.3 kcal/mol, the free energy difference is overestimated in the CHARMM family FFs and underestimated in both AMBER family models, and closest in agreement to the CHARMM36 and AMBERbsc1 force field. These discrepancies are significantly larger compared to the results obtained by Yang et al.[24,25] using more expensive umbrella sampling[19] and multiple-walker well tempered metadynamics[21] simulations and can possibly indicate an artifact of the meta-eABF approach which converges faster but only provides a qualitatively accurate free energy landscape for systems with high energy barrier.[51]

All the four free energy landscapes indicate that the opening of the base pairs through the major groove region is more energetically favorable compared to the minor groove region, especially when starting from a WC structure, which is the most prevalent form in solvated DNA duplex. This pathway is in agreement with the previous experimental[3] and simulation studies.[26,29] Except for CHARMM36, all three force fields show that the base paired structures (WC or HG) are more stable than the open base paired configurations. In CHARMM36 we can observe an open base pair state in the major groove region near the WC form which is lower in free energy in comparison to the HG state. This result is consistent

with the previous work, where a very stable open base paired intermediate state between WC and HG was captured using CHARMM36 force field and Markov State Modeling.[26]

Additionally, for the majority of models, the minimum free energy pathway reflected in the transition mechanism is a direct 180° flip-over of the adenine glycosidic angle whereas the CHARMM27 FF shows the minimum free energy pathway involves slight opening of the base pair indicating a possible extra-helical mechanism. From previous analysis of the T9-A4' base pair in sequence 5'-CGATTTTTTGGC-3', it was concluded that the extra-helical mechanism, where the purine base flips out of the DNA helix first, was preferred for the HG to WC base pair transition.[29] Although this mechanism was not the most energetically favorable in CHARMM36, AMBERbsc0, and AMBERbsc1, the metastable open base pair conformations observed in those FFs allows for the possibility of conformational switching between WC and HG form via an extra-helical intermediate. For all the force fields, shallow free energy regions are detectable in the major groove ($\Theta > 0$) and minor grooves ($\Theta < 0$). In the CHARMM36 and AMBERbsc1 force fields, the open base pair state, adjacent to the Waston-Crick minima, is clearly distinguishable.

In our previous study,[26] we suggested that the dihedral angle based reaction coordinates may not be able to uniquely distinguish between WC and HG states as some structures which does have the necessary hydrogen bonds for base pair formation can be also be assigned to either of these two categories just from the χ and θ torsion angle values. Here, we propose to use the hydrogen bond donor acceptor distances to identify the base paired forms. In WC base pairing, an N-H-N type hydrogen bond forms between the N1 nitrogen of adenine and the N3 nitrogen of thymine nucleotides (N1(A)-N3(T)). In the HG form, the N3(T) forms a hydrogen bond with N7(A) instead. So, by measuring the distance between the N1(A)-N3(T) atom pair and N7(A)-N3(T) atom pair the WC and HG states can be uniquely identified. This is reflected in the time series plots of dihedral angles and hydrogen bond donor acceptor distances in Figure 2. We projected our free energy landscapes from Figure 3 in the hydrogen bond donor acceptor distance space (Figure 4). All four models

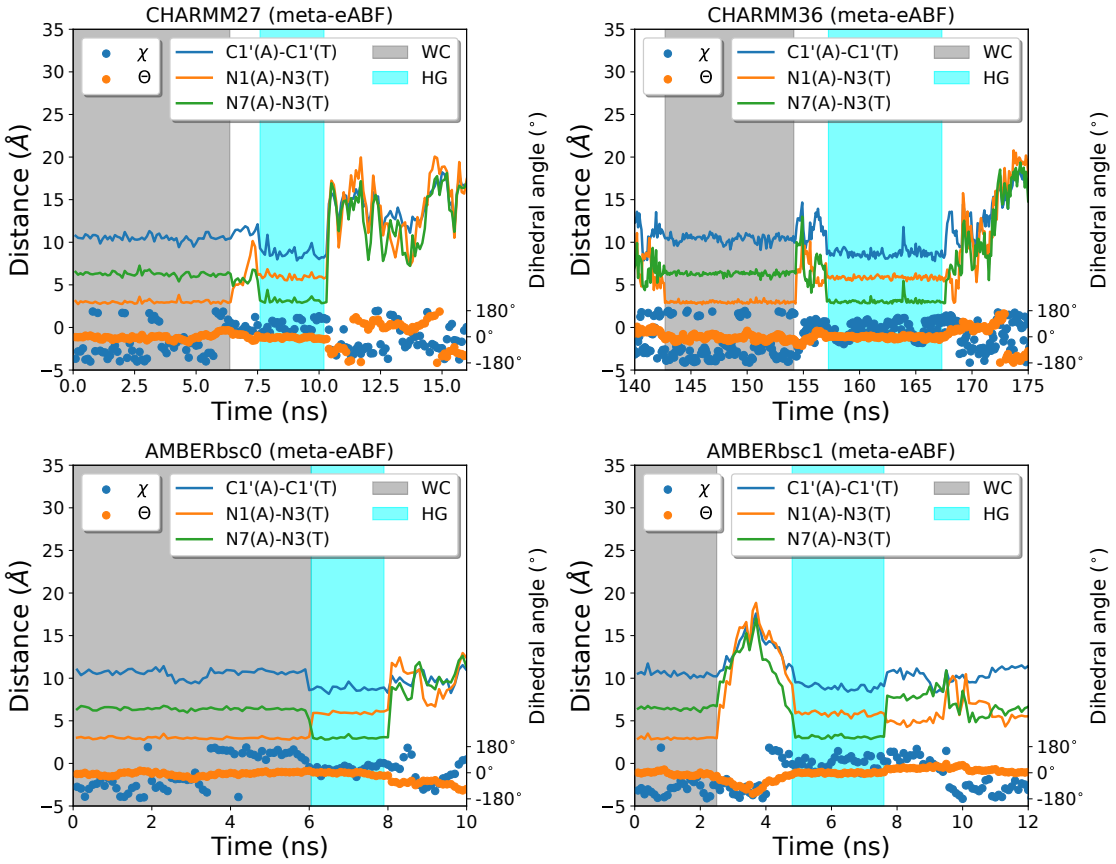

Figure 2: Representative regions of the meta-eABF trajectories showing transitions between WC and HG conformations. The hydrogen bond donor acceptor distance, the helix diameter measured as the C1'-C1' distance, and the two torsion angles χ and θ are depicted as a function of time.

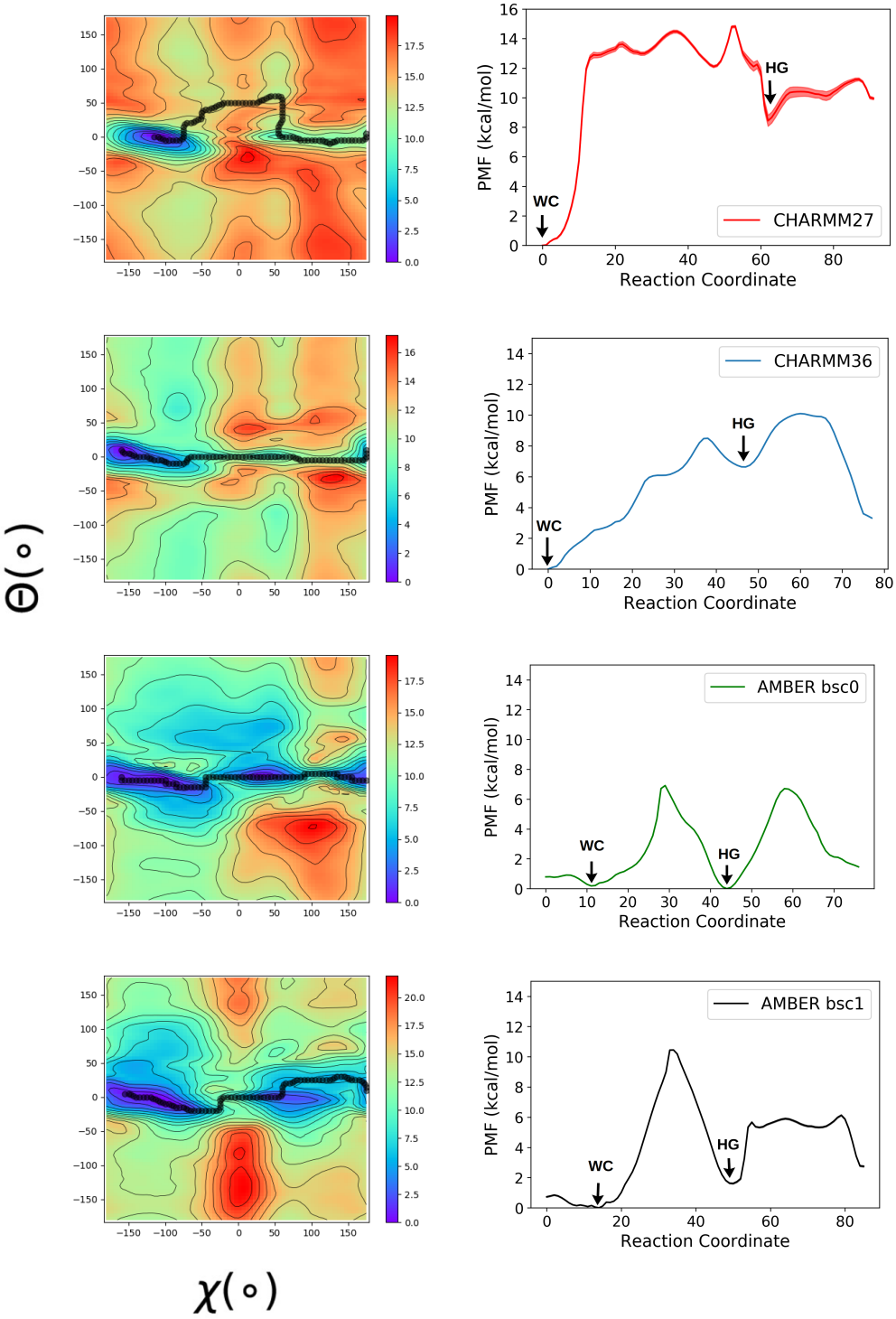

Figure 3: The free energy surface of the A16-T9 base pair in A6-DNA was obtained using meta-eABF method for different force field models.

clearly distinguish the WC and HG states as free energy minima in the configurational space. The relative free energies of the HG base pairs are in slightly better agreement with the experimental results (CHARMM27: ∼6 kcal/mol, AMBERbsc1: ∼2 kcal/mol, AMBERbsc0: ∼1 kcal/mol, CHARMM36: ∼6 kcal/mol), reinforcing our proposition that in χ-θ based representation the WC and HG states get contaminated by other states with higher/lower free energy and provides a poorer estimate of the relative free energy. One should note that these distance based coordinates may not be the optimal choice as collective variable for the application of the biasing force, as we indeed need to force the transition of the glycosidic torsion angle to go from one base pairing conformation to the other. The hydrogen-bond donor-acceptor distance based CVs are useful to distinguish between WC and HG conformations during the post-processing of the MD trajectories.

Root Mean Square Fluctuations and Absolute Entropy

The RMSF of heavy atoms are significantly different at the A16 base and the nearby base pairs when comparing both WC and HG conformations, Figure 5. For instance, in both the CHARMM family force fields, the A16 base experienced more fluctuations in the HG form than the WC conformation. The opposite pattern was observed in the AMBER family models, where the fluctuations at the A16 base were larger in the WC form. In all force fields, there are also fluctuations occurring near the A16 and T9 base, which indicate the impact of the Hoogsteen form on the entropy of the B-DNA. One common theme is apparent for all force fields: The base regions (particularly the six membered ring of the adenine base) becomes more flexible in the HG form than in the WC form, while the backbone region becomes slightly more rigid in the HG form. The effect is not very consistent for the Thymine base (Figure 6). This indicates that the apparent rigidity of the HG base pair is not equally contributed by all the atoms.

In Figure 7, the quasi-harmonic entropy of the DNA duplex involving one Hoosgteen conformation, is found to be lower than that of a full WC DNA duplex, for all four force

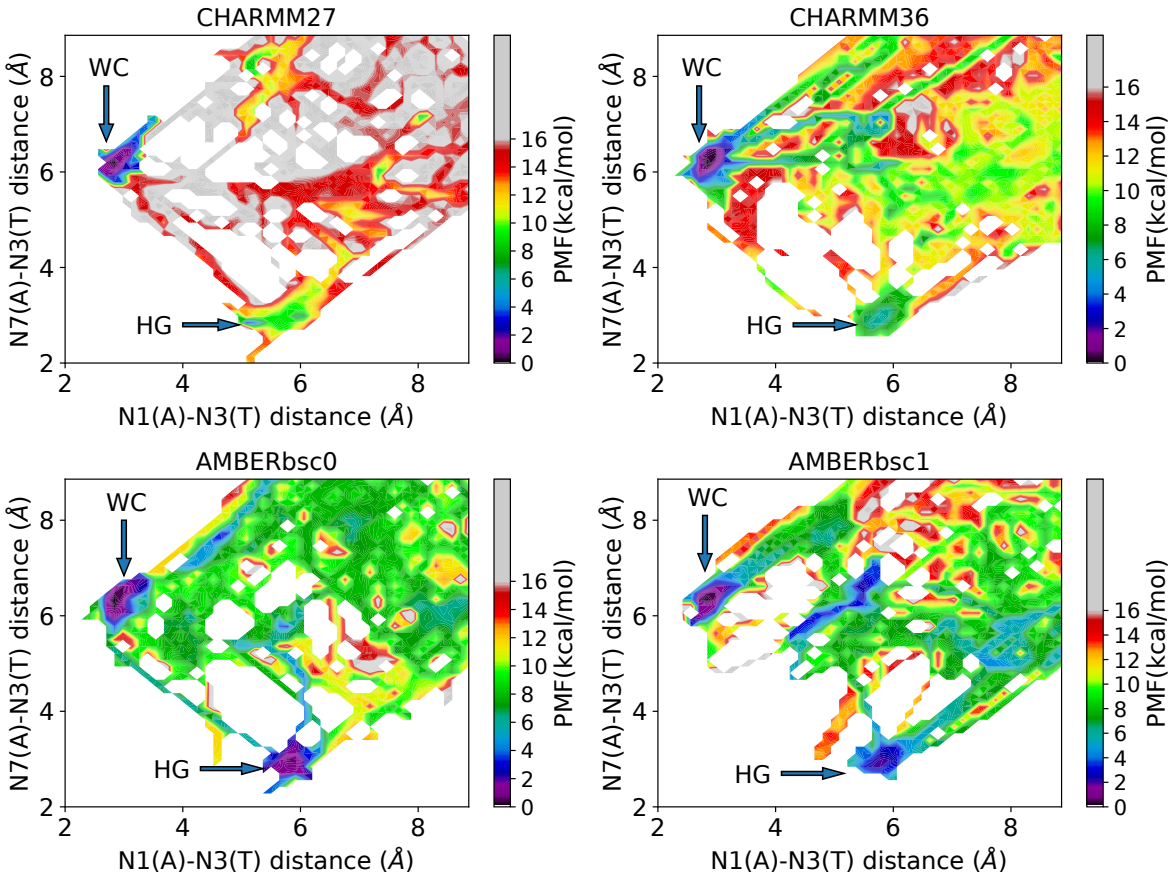

Figure 4: The free energy landscape of the conformational switching between WC and HG base pair configurations, projected onto the two hydrogen bond donor acceptor distances, for different force field models.

fields, possibly due to the increase of rigidity caused by the shrinkage of the helix diameter at the HG state. The extent by which the HG base pairing form is entropically destabilized varied between different force field models. The entropy difference of HG and WC base pairs is lowest in the CHARMM27 ff and is highest in AMBERbsc1, in contrast to the free energy results that indicate a lower free energy difference in AMBER ffs compared to CHARMM ffs. To understand the role of entropy and enthalpy in the stabilization of the WC base pairing form over the HG, we performed non-bonded interaction energy calculations. We found that the interaction energy between two bases are stronger for the HG base pairing form than the WC base pairing form, for all force field models. Although this seems counterintuitive, previous studies of Gould and Kollman show that in the crystalline state the HG base pairing form is energetically more stable than the WC form.[52] This stabilization is much higher for the AMBER family force fields than the CHARMM family force fields. When we look at the interaction energy of the base pair with the rest of the DNA duplex, the CHARMM family force fields show a positive (destabilizing) change in the interaction energy going from WC to HG, while a negative (stabilizing) change is observed for the AMBER family ffs. We also computed the relative base stacking energy between the two base pairing forms. The difference of base stacking energy between the WC and HG form in CHARMM family FFs is almost negligible. But the stacking energy for the AMBER family FFs is significantly stronger (more negative) in the HG form compared to the WC form. These results clearly suggests that the two types of force fields treat the HG base pairing very differently in terms of energy costs. The positive (or less negative) energy cost of forming a HG base pair in CHARMM FFs results in a higher free energy difference between the WC and HG base pairs despite a lower entropic cost compared to AMBER FFs.

Nevertheless, the structural properties of the WC and HG base pairs are fairly consistent throughout the four force field models. The two torsion angles and various inter-atomic distances between the A16 and T9 bases are reported in Table 1 and 2, respectively. Except for a wider fluctuation and a shifted location of the HG minima in the $\chi - \theta$ space for the

AMBERbsc0 force field, the results of different force fields are in agreement with each other. The time evolution of the H-bond donor acceptor distances, the helix diameter (measures as the distance between the C1' atoms), and the torsion angles for the 100 ns equilibrium trajectories are provided in the supporting information. It is clear from these plots that no WC to HG transition or base opening takes place during the simulation, confirming that these structures indeed represent minima in the free energy landscape.

Table 1: The values of the glycosidic angle χ and the base flip-out angle θ for the WC and HG conformation for the four different force fields. The results are reported as mean ± standard deviation.

Force Field	WC $\chi(^\circ)$	HG $\chi(^\circ)$	WC $\theta(^\circ)$	HG $\theta(^\circ)$
CHARMM27	-109.2±13.1	46.7±10.3	-2.2±3.9	0.0±2.4
CHARMM36	-113.0±14.8	45.2±10.7	-0.4±3.8	-0.2±2.3
AMBERbsc0	-113.3±20.1	26.8±22.8	-5.9±5.2	-0.3±2.2
AMBERbsc1	-110.3±15.7	61.0±9.3	-5.3±6.0	-1.3±2.0

Table 2: The distances between the hydrogen bond donor-acceptors (N1-N3 and N1-N7), and the C1' atoms of the A16 and T9 nucleotide, for the WC and HG conformation for the four different force fields. The results are reported as mean ± standard deviation.

Force Field	WC N1(A)-N3(T) (Å)	HG N1(A)-N3(T) (Å)	WC N7(A)-N3(T) (Å)	HG N7(A)-N3(T) (Å)	WC C1(A)-C1(A) (Å)	HG C1(A)-C1(A) (Å)
CHARMM27	2.96±0.13	5.90±0.25	6.31±0.15	3.12±0.30	10.60±0.28	8.88±0.51
CHARMM36	2.96±0.13	5.87±0.21	6.29±0.16	3.08±0.22	10.60±0.29	8.85±0.40
AMBERbsc0	2.96±0.12	5.84±0.17	6.48±0.15	3.02±0.16	10.56±0.29	9.03±0.31
AMBERbsc1	2.97±0.12	5.89±0.17	6.51±0.14	3.05±0.14	10.59±0.29	9.04±0.30

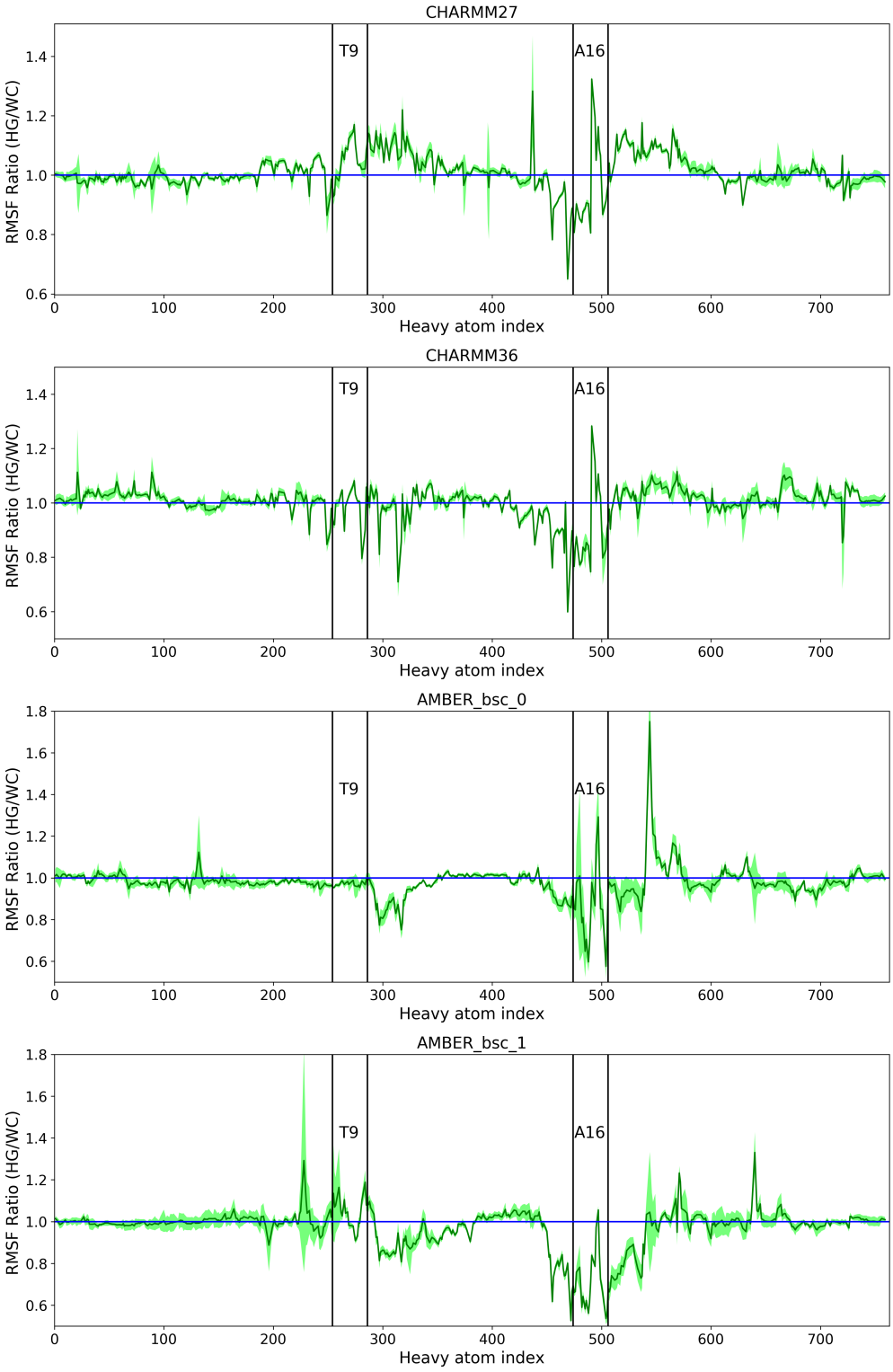

Figure 5: The RMSF ratio of the DNA duplexes involving the HG to WC form of the A16-T9 base pair. The location of the A16 and the T9 based are indicated by vertical lines.

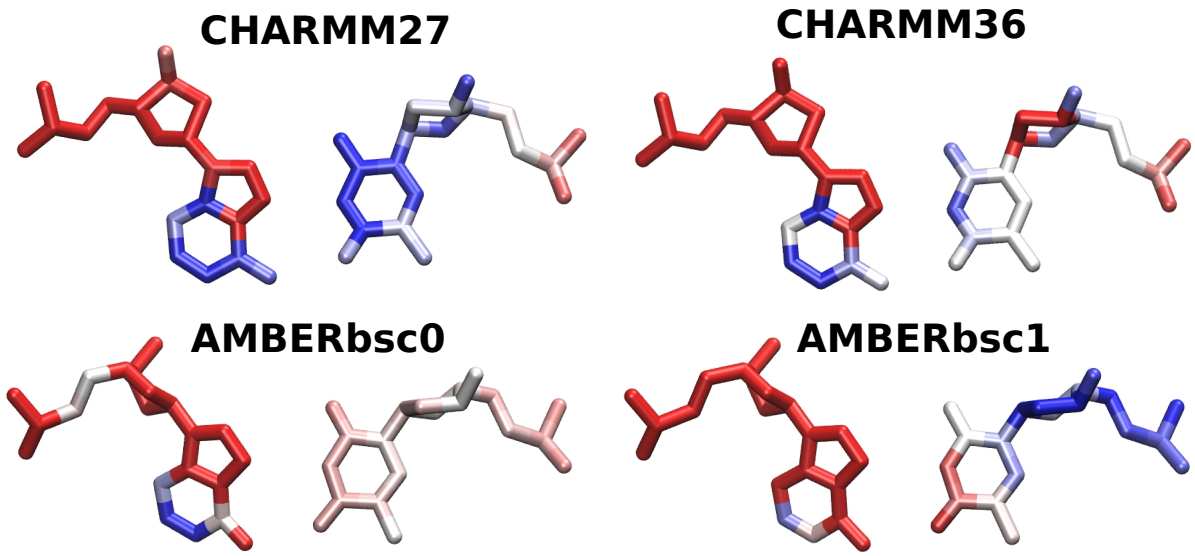

Figure 6: The atoms of the A16-T9 base pair colored according to the RMSF ratio between the WC and HG state. The blue color indicate RMSF ratio greater than 1 and red color indicate less than 1. The structures are NOT representative of the HG base pairing forms observed using each force field.

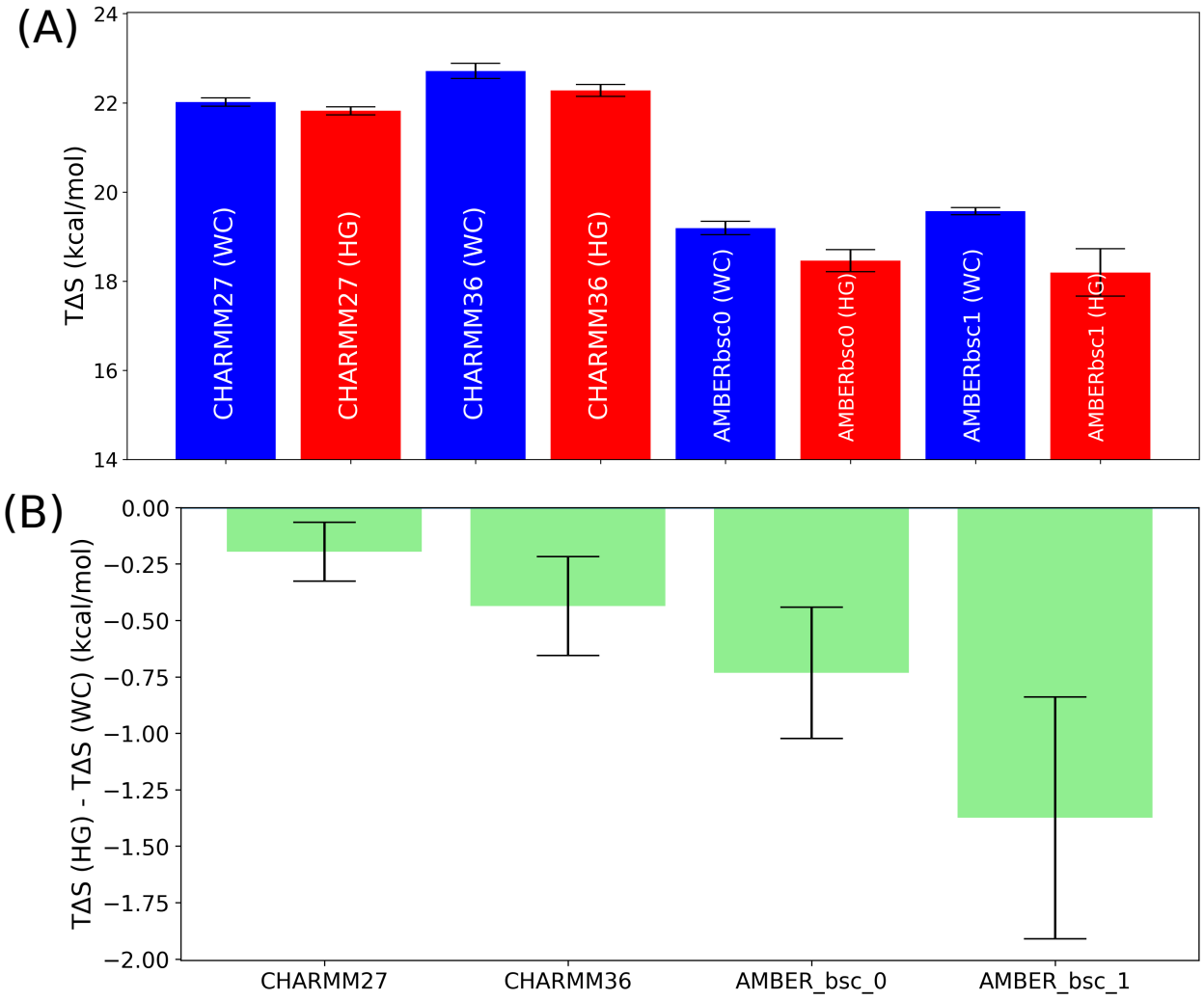

Figure 7: (A) The absolute value and (B) The relative value of the quasi-harmonic entropy of the HG base pair with respect to the WC base pair for different force fields.

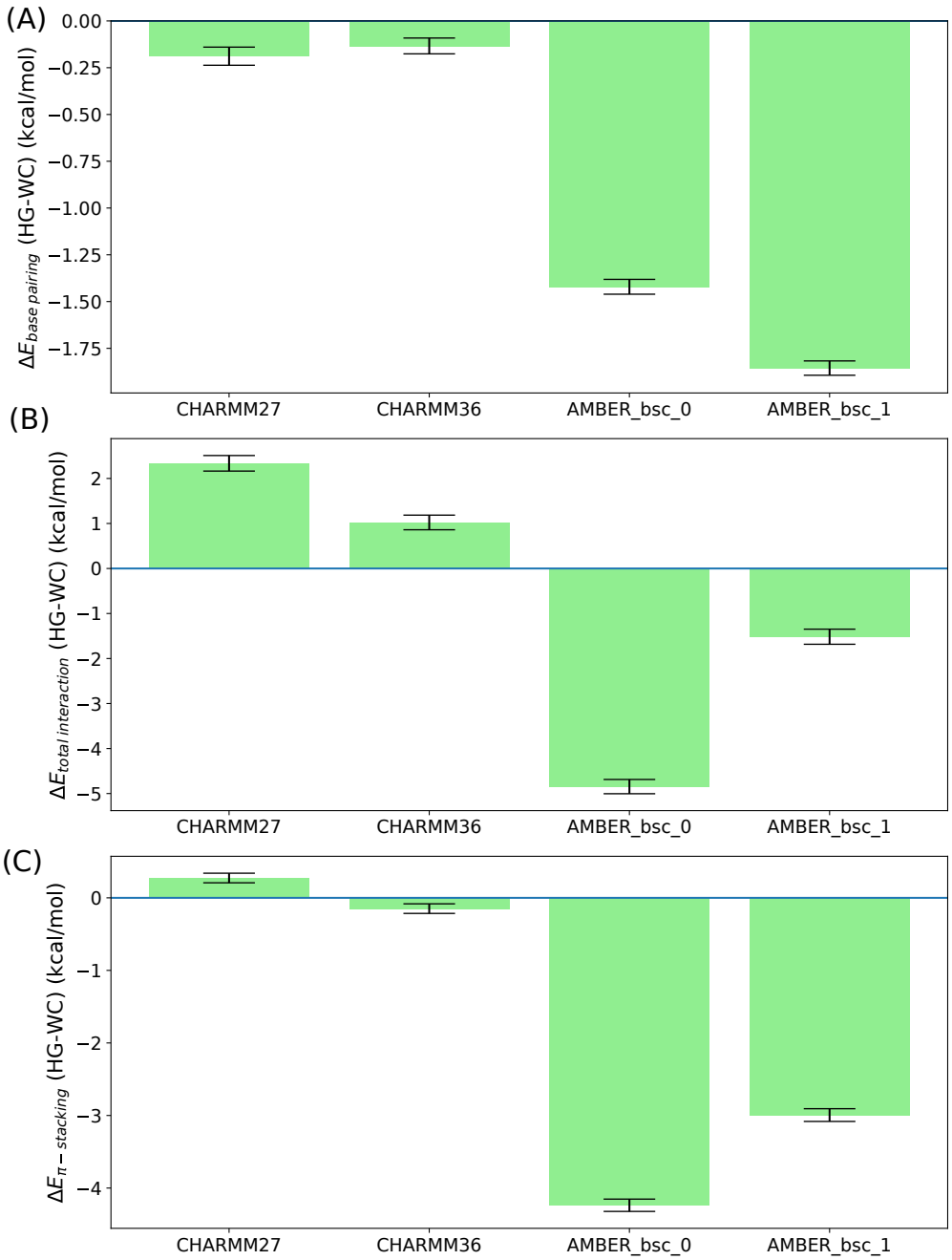

Figure 8: The non-bonded interaction energies in the HG base pair relative to the WC base pair, for the following interactions: (A) between the A16 and T9 bases, (B) between the base pair with the rest of the DNA duplex, and (C) the stacking energy defined as the interaction energy between the π conjugated region of the base pair with the two adjacent base pairs.

Conclusions

We performed enhanced sampling and equilibrium molecular dynamics simulations to understand the role of force field model in predicting the properties related to the formation of the Hoogsteen base pairing in the A6-DNA duplex. Although the structures of the HG and WC base pairing forms were consistent in the different force field models, the relative free energy, entropy and the interaction energies of the HG base pair changes significantly when a different force field model was used. The relative stability of the two base pairing forms are determined by a fine balance of the entropy cost of forming the structurally rigid HG base pair and the gain in interaction enthalpy over the WC structure. The topology of the underlying free energy landscape of the conformational transition between WC and HG a.k.a. "DNA breathing" is more or less consistent between the different force field models. But, the predicted free energy costs and the barriers varied significantly. Any quantitative comparison at the level of free energy should be performed with caution, owing to the approximate nature of the PMFs obtained from meta-eABF simulation. Multi-microsecond well tempered metadynamics simulations can produce more accurate free energy landscapes, but can become prohibitively expensive when trying to compare the results between different force fields. The collective variables used for biasing are not optimal for distinguishing the two conformations and more recent artificial intelligence driven order parameters including those obtained using spectral gap optimization[53] and harmonic linear discriminant analysis[54] may provide a better free energy difference. Despite the technical limitations, our study highlights the diversity of the predictions of molecular properties of WC and HG base pairs from different force fields, and facilitates the assessment of the reliability of future computer simulation studies on DNA breathing dynamics.

Note

The authors declare that they are committed to including, supporting and advancing women in Chemistry and scientific research in general. The authors are aware of the multitude of challenges faced by women in progressing their career in scientific research, particularly in academia, with only a small fraction of the leadership positions held by women. The lead author of this paper, SES, is a first generation college student from a historically underrepresented community, and has first-hand experience of the enormous difficulties involved. The research group of IA have hosted many female scientists and nourished their talents to help them become successful scientists. Yet, the authors are not complacent, but aware of the fact that much more is needed to be done, and all authors pledge to undertake conscious efforts in removing the barriers and create an inclusive and supportive environment for women in science.

Acknowledgement

This work was supported partially by the National Science Foundation (NSF) via grant MCB 2028443 awarded to IA. SES acknowledges support through Pfizer - La Jolla Academic Industrial Relations (AIR) Diversity Research Fellowship in Chemistry for Undergraduates. The authors thank the University of California Irvine High Performance Computing (HPC) facility and the Triton Shared Computing Cluster (TSCC) in the San Diego Supercomputer Center (SDSC) for providing the computational resources. The authors declare no competing financial interest.

Data and Software Availability

The open-source package NAMD 2.14[55] is used for all molecular dynamics simulations. The data, based on which the conclusions of this paper have been drawn, are provided in the main

manuscript and the supporting information. Additional data files and scripts are available from the corresponding author upon request.

Supporting Information Available

See supporting information for additional plots of the stability of the HG and WC structures, convergence of entropy calculations, and the absolute values of interaction energies.

References

(1) Franklin, R. E.; Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **1953**, *171*, 740–741.

(2) Watson, J. D.; Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171*, 737–738.

(3) Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **2011**, *470*, 498–502.

(4) Hoogsteen, K.; IUCr, The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallographica* **1959**, *12*, 822–823.

(5) Kitayner, M.; Rozenberg, H.; Rohs, R.; Suad, O.; Rabinovich, D.; Honig, B.; Shakked, Z. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nature Structural & Molecular Biology* **2010**, *17*, 423–429.

(6) Xu, Y.; Manghrani, A.; Liu, B.; Shi, H.; Pham, U.; Liu, A.; Al-Hashimi, H. M. Hoogsteen base pairs increase the susceptibility of double-stranded DNA to cytotoxic damage. *Journal of Biological Chemistry* **2020**, *295*, 15933–15947.

(7) Jiang, Y.-L.; Liu, Z.-P. Metallo-Organic G-Quadruplex Ligands in Anticancer Drug Design. *Mini-Reviews in Medicinal Chemistry* **2010**, *10*, 726–736.

(8) Patel, D. J.; Shapiro, L.; Kozlowski, S. A.; Gaffney, B. L.; Jones, R. A. Covalent carcinogenic O6-methylguanosine lesions in DNA structural studies of the O6meG·A and O6meG·G interactions in dodecanucleotide duplexes. *Journal of Molecular Biology* **1986**, *188*, 677–692.

(9) Singh, U. S.; Moe, J. G.; Reddy, G. R.; Weisenseel, J. P.; Marnett, L. J.; Stone, M. P. Proton NMR of an oligodeoxynucleotide containing a propanodeoxyguanosine adduct

positioned in a (CG)3 frameshift hotspot of Salmonella typhimurium hisD3052: Hoogsteen base-pairing at pH 5.8. *Chemical Research in Toxicology* **1993**, *6*, 825–836.

(10) Yang, H.; Zhan, Y.; Fenn, D.; Chi, L. M.; Lam, S. L. Effect of 1-methyladenine on double-helical DNA structures. *FEBS Letters* **2008**, *582*, 1629–1633.

(11) Golovenko, D.; Bräuning, B.; Vyas, P.; Haran, T. E.; Rozenberg, H.; Shakked, Z. New Insights into the Role of DNA Shape on Its Recognition by p53 Proteins. *Structure* **2018**, *26*, 1237–1250.

(12) Sathyamoorthy, B.; Shi, H.; Zhou, H.; Xue, Y.; Rangadurai, A.; Merriman, D. K.; Al-Hashimi, H. M. Insights into Watson–Crick/Hoogsteen breathing dynamics and damage repair from the solution structure and dynamic ensemble of DNA duplexes containing m1A. *Nucleic Acids Research* **2017**, *45*, 5586–5601.

(13) Rice, P. A.; Yang, S.-w.; Mizuuchi, K.; Nash, H. A. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. *Cell* **1996**, *87*, 1295–1306.

(14) Aishima, J.; Gitti, R. K.; Noah, J. E.; Gan, H. H.; Schlick, T.; Wolberger, C. A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Research* **2002**, *30*, 5244–5252.

(15) Bunting, K. A.; Roe, S. M.; Headley, A.; Brown, T.; Savva, R.; Pearl, L. H. Crystal structure of the Escherichia coli dcm very-short-patch DNA repair endonuclease bound to its reaction product-site in a DNA superhelix. *Nucleic Acids Research* **2003**, *31*, 1633–1639.

(16) Zhou, H.; Kimsey, I. J.; Nikolova, E. N.; Sathyamoorthy, B.; Grazioli, G.; McSally, J.; Bai, T.; Wunderlich, C. H.; Kreutz, C.; Andricioaei, I.; Al-Hashimi, H. M. m1A and m1G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs. *Nature Structural & Molecular Biology* **2016**, *23*, 803–810.

(17) Nair, D. T.; Johnson, R. E.; Prakash, S.; Prakash, L.; Aggarwal, A. K. Replication by human DNA polymerase-ι occurs by Hoogsteen base-pairing. *Nature* **2004**, *430*, 377–380.

(18) Xu, Y.; McSally, J.; Andricioaei, I.; Al-Hashimi, H. M. Modulation of Hoogsteen dynamics on DNA recognition. *Nature Communications* **2018**, *9*, 1473.

(19) Torrie, G.; Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187–199.

(20) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 12562–6.

(21) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **2008**, *100*, 020603.

(22) Fu, H.; Zhang, H.; Chen, H.; Shao, X.; Chipot, C.; Cai, W. Zooming across the Free-Energy Landscape: Shaving Barriers, and Flooding Valleys. *The Journal of Physical Chemistry Letters* **2018**, *9*, 4738–4745.

(23) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants. *The Journal of Chemical Physics* **1998**, *108*, 1964–1977.

(24) Yang, C.; Kim, E.; Pak, Y. Free energy landscape and transition pathways from Watson–Crick to Hoogsteen base pairing in free duplex DNA. *Nucleic Acids Research* **2015**, *43*, 7769–7778.

(25) Yang, C.; Kim, E.; Lim, M.; Pak, Y. Computational Probing of Watson–Crick/Hoogsteen Breathing in a DNA Duplex Containing N1-Methylated Adenine. *Journal of Chemical Theory and Computation* **2019**, *15*, 751–761.

(26) Ray, D.; Andricioaei, I. Free Energy Landscape and Conformational Kinetics of Hoogsteen Base Pairing in DNA vs. RNA. *Biophysical Journal* **2020**, *119*, 1568–1579.

(27) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **2010**, *52*, 99–105.

(28) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* **2018**, *140*, 2386–2396.

(29) Vreede, J.; Pérez de Alba Ortíz, A.; Bolhuis, P. G.; Swenson, D. W. H. Atomistic insight into the kinetic pathways for Watson–Crick to Hoogsteen transitions in DNA. *Nucleic Acids Research* **2019**,

(30) Hooft, F.; Pérez De Alba Ortíz, A.; Ensing, B. Discovering Collective Variables of Molecular Transitions via Genetic Algorithms and Neural Networks. *Journal of Chemical Theory and Computation* **2021**, *17*, 2294–2306.

(31) Vanommeslaeghe, K.; Mackerell, A. D. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2015**, *1850*, 861–871.

(32) Salsbury, A. M.; Lemkul, J. A. Recent developments in empirical atomistic force fields for nucleic acids and applications to studies of folding and dynamics. *Current Opinion in Structural Biology* **2021**, *67*, 9–17.

(33) Foloppe, N.; Mackerell, A. D. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data ALL-ATOM EMPIRICAL FORCE FIELD FOR NUCLEIC ACIDS. I. *Journal of Computational Chemistry* **2000**, *21*, 86–104.

(34) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.;

Orozco, M. Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophysical Journal* **2007**, *92*, 3817–3829.

(35) Galindo-Murillo, R.; Robertson, J. C.; Zgarbová, M.; Šponer, J.; Otyepka, M.; Jurečka, P.; Cheatham, T. E. Assessing the Current State of Amber Force Field Modifications for DNA. *Journal of Chemical Theory and Computation* **2016**, *12*, 4114–4127.

(36) Ivani, I. et al. Parmbsc1: a refined force field for DNA simulations. *Nature Methods 2016 13:1* **2015**, *13*, 55–58.

(37) Minhas, V.; Sun, T.; Mirzoev, A.; Korolev, N.; Lyubartsev, A. P.; Nordenskiöld, L. Modeling DNA Flexibility: Comparison of Force Fields from Atomistic to Multiscale Levels. *Journal of Physical Chemistry B* **2020**, *124*, 38–49.

(38) Liebl, K.; Zacharias, M. Tumuc1: A New Accurate DNA Force Field Consistent with High-Level Quantum Chemistry. *Journal of Chemical Theory and Computation* **2021**, *17*, 7096–7105.

(39) Mohideen, K.; Muhammad, R.; Davey, C. A. Perturbations in nucleosome structure from heavy metal association. *Nucleic Acids Research* **2010**, *38*, 6301–6311.

(40) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(41) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005**, *26*, 1668–1688.

(42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(43) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **1994**, *101*, 4177–4189.

(44) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics* **1995**, *103*, 4613–4621.

(45) Fu, H.; Shao, X.; Chipot, C.; Cai, W. Extended Adaptive Biasing Force Algorithm. An On-the-Fly Implementation for Accurate Free-Energy Calculations. *Journal of Chemical Theory and Computation* **2016**, *12*, 3506–3513.

(46) Fiorin, G.; Klein, M. L.; Hénin, J. Using collective variables to drive molecular dynamics simulations. *Molecular Physics* **2013**, *111*, 3345–3362.

(47) Marcos-Alcalde, I.; Setoain, J.; Mendieta-Moreno, J. I.; Mendieta, J.; Gómez-Puertas, P. MEPSA: minimum energy pathway analysis for energy landscapes. *Bioinformatics* **2015**, *31*, 3853–3855.

(48) Andricioaei, I.; Karplus, M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *The Journal of Chemical Physics* **2001**, *115*, 6289.

(49) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

(50) NAMD Energy Plugin, Version 1.4. *https://www.ks.uiuc.edu/Research/vmd/plugins/namdenergy/*

(51) Fu, H.; Chen, H.; Wang, X.; Chai, H.; Shao, X.; Cai, W.; Chipot, C. Finding an optimal pathway on a multidimensional free-energy landscape. *Journal of Chemical Information and Modeling* **2020**, *60*, 5366–5374.

(52) Gould, I. R.; Kollman, P. A. Theoretical Investigation of the Hydrogen Bond Strengths

in Guanine-Cytosine and Adenine-Thymine Base Pairs. *Journal of the American Chemical Society* **1994**, *116*, 2493–2499.

(53) Tiwary, P.; Berne, B. J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113*, 2839–2844.

(54) Mendels, D.; Piccini, G.; Parrinello, M. Collective Variables from Local Fluctuations. *Journal of Physical Chemistry Letters* **2018**, *9*, 2776–2781.

(55) Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics* **2020**, *153*, 044130.