Ab initio determination of coarse-grained interactions in double-stranded DNA

Chia Wei Hsu, Maria Fyta, Greg Lakatos, Simone Melchionna, and Efthimios Kaxiras

Citation: J. Chem. Phys. **137**, 105102 (2012); doi: 10.1063/1.4748105
View online: http://dx.doi.org/10.1063/1.4748105
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v137/i10
Published by the American Institute of Physics.

Ab initio determination of coarse-grained interactions in double-stranded DNA

Chia Wei Hsu,[1] Maria Fyta,[1,2] Greg Lakatos,[1,3] Simone Melchionna,[3,4] and Efthimios Kaxiras[1,3,a)]

[1]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*Department of Physics, Technical University of Munich, Garching 85748, Germany*
[3]*School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[4]*IPCF-CNR, Istituto Processi Chimico-Fisici, Consiglio Nazionale delle Ricerche, Università La Sapienza, P.le A. Moro 2, 00185 Rome, Italy*

(Received 12 December 2011; accepted 13 August 2012; published online 12 September 2012)

We derive the coarse-grained interactions between DNA nucleotides from *ab initio* total-energy calculations based on density functional theory (DFT). The interactions take into account base and sequence specificity, and are decomposed into physically distinct contributions that include hydrogen bonding, stacking interactions, backbone, and backbone-base interactions. The interaction energies of each contribution are calculated from DFT for a wide range of configurations and are fitted by simple analytical expressions for use in the coarse-grained model, which reduces each nucleotide into two sites. This model is not derived from experimental data, yet it successfully reproduces the stable B-DNA structure and gives good predictions for the persistence length. It may be used to realistically probe dynamics of DNA strands in various environments at the μs time scale and the μm length scale. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4748105]

I. INTRODUCTION

Biological systems exhibit high degrees of complexity that are essential to the functions they perform. The DNA double helix is one such example: the properties of this macromolecule are directly influenced by its conformational variability as well as by environmental factors that include counterions, impurities, and temperature, as it performs a wide variety of vital cellular functions such as transcription and replication.[1] A full account of such biological functions must rely on a realistic description of the physical processes that underlie them. Fine-grained calculations of DNA at the atomic level[2,3] can provide this level of detailed description, but they are restricted to very small systems of order tens of base pairs and time scales of order ns, whereas most biological processes involve DNA behavior at the scale of more than a hundred base-pairs and take place at the μs time scale and beyond. To probe these biologically relevant processes, a realistic and efficient coarse-grained model of DNA is necessary. Examples of crucial functions under current investigation that could benefit from a coarse-grained model of DNA include the translocation of DNA through nanopores[4,5] in the context of ultra-fast electronic sequencing, DNA capture in and ejection from nanoscale capsules/wells, and the study of the interplay between histones and DNA during mitosis.[6]

The main theoretical challenge in biological systems is to bridge the scales between the atomistic and the macroscopic without wasting computational resources on uninteresting behaviors, such as the internal dynamics within a base which practically never changes shape, or the motion of solvent molecules far from the biomolecule. Computational methods of this type have been used before, for example, in the study of protein dynamics.[7] Multiscale simulations have also been successfully applied to the study of the electronic behavior and electron localization in stretched dry DNA.[8] Our ultimate goal is to enable the study of variations in the DNA structure using multiscale approaches that do not sacrifice accuracy while achieving high efficiency. In the present work we focus on the first stage toward this goal, namely, the development of a coarse-grained model capable of accurately reproducing the structure of double-stranded DNA (ds-DNA) in solution, and simple enough to be efficiently combined with multiscale simulation techniques.

Many coarse-grained models of DNA have been proposed in the past few years.[9–21] Most of them are constructed in a "top-down" fashion,[9–17] where the interaction potentials are chosen to reproduce certain sets of experimental data. Bead-string models[10] have been used to study diffusion and structural relaxation of single strands. Rigid base-pair models[11] have been used to describe the elastic properties of DNA.[22,23] The three-site-per-nucleotide model by Knotts and coworkers[14] captures the salt-dependent melting of DNA, and has been extended to include solvent-induced attraction between DNA strands[24] that helped to gain insights on hybridization[25–27] and on certain sequence-dependent effects;[28] this model has also been adapted to describe the mechanical denaturation of long DNA[29] and to include explicit solvent molecules.[30] Starr and coworkers proposed a simple model that captures the basics of hybridization,[13] and this model has been used to study Holliday junctions[31] and the self-assembly of DNA-linked nanoparticles.[32] Ouldridge and coworkers proposed another model that is sequence independent, but can reproduce several structural, mechanical, and thermodynamic properties

a)Electronic mail: kaxiras@physics.harvard.edu.

DNA.[16,33] Another family of models starts from all-atom empirical force fields, and construct the coarse-grained model potentials from bottom-up.[18–21] The model by Savelyev and coworkers[18,34] was parametrized by matching moments of observables in the Hamiltonian. Other approaches to a bottom-up construction include minimizing difference between the all-atom and coarse-grained potentials,[19] imposing molecular bonding geometry constraints,[20] and inverting the Boltzmann function to get the coarse-grained potentials.[21]

Each of these models has its regime of validity, depending on what experimental data were used for the model derivation. In this work, we develop a minimal model of ds-DNA that incorporates sequence specificity and has realistic mechanical robustness to bending and untwisting forces. We seek a model that is chemically accurate, yet not based on empirical observations. For these purposes, we take a bottom-up approach, and construct the potentials of the coarse-grained system directly from first-principles calculations. We divide the interaction potentials into independent parts that come from physically distinct contributions, and for each contribution we carry out *ab initio* calculations to find the functional forms of the potentials. We impose no *a priori* assumption on the functional forms of the potentials—these are determined based on results of the *ab initio* calculations. The present form of the model has its limitations that we will discuss, but it is flexible for future improvements.

This paper is organized as follows: in Sec. II we describe the methodology employed for the first-principles calculations. In Sec. III we present the *ab initio* data for each energy contribution and the analytical forms of the corresponding potentials. The implementation and performance of the coarse-grained model is described in Sec. IV. Finally, we present validations of the model in Sec. V, and conclude in Sec. VI.

II. METHODOLOGY OF *AB INITIO* CALCULATIONS

We carry out the first-principles calculations using density functional theory (DFT).[35] In our DFT calculations, we do not deal with environmental factors such as solvent molecules and ions, and the calculations are carried out at the ground state (zero temperature). The environment and temperature factors are added *a posteriori* in the coarse-grained model through electric field screening and through Brownian dynamics. A more accurate approach will include temperature dependence in the coarse-graining procedure, but this is a challenging task and is not yet taken into account in the current work. Presence of the water molecules and ions may also affect the energetics, but their effects are not investigated in this work.

We use SIESTA,[36] an electronic structure code based on atomic-like orbitals, to carry out the DFT calculations. This approach has been previously applied in similar studies of gas-phase DNA bases[37] and successfully reproduces properties such as optical response in comparison to available experimental data.[38] We use the Troullier-Martins scheme[39] to obtain pseudopotentials to eliminate the core electrons from the calculation and to produce a smoother valence charge density. We use a basis of double-ζ polarized atomic orbitals for all the atoms involved (13 orbitals for C, N, O, and P; 5 orbitals for H). An auxiliary real space grid equivalent to a plane-wave cutoff of 100 Ry is used for the calculation of the Hartree and exchange-correlation energies. For geometry optimization, a structure is considered fully relaxed when the magnitude of force on every atom is smaller than 0.04 eV/Å.

We use the generalized gradient approximation (GGA) with the PBE exchange-correlation functional[40] to describe the backbone-base and the inter-base-pair interactions, as this functional is known to describe covalent and hydrogen bonds well.[41,42] Interactions within the backbone are treated with the local-density approximation (LDA),[43] which is adequate for describing small radial and angular deviations of covalent bonds from their equilibrium values. The interaction between stacked base-pairs has a large contribution from long-ranged van der Waals (vdW) forces and exhibits an elaborate energy landscape that depends sensitively on the geometry.[44,45] Local or semi-local exchange-correlation functionals cannot describe such long-range effects:[46] they do not reproduce the $\sim r^{-6}$ behavior at large separation r that is characteristic of vdW interactions, and usually underestimate the stacking distance between two planar structures. There exists empirical corrections that add the vdW effects to the energies obtained from DFT calculations,[47] but we find that such correction still leads to underestimation of the stacking distance. Therefore, we employ a non-empirical long-range vdW density functional developed by Dion *et al.*[48] to carry out calculations for the stacked base-pairs interactions.

The calculated interaction energy between two constituents may be susceptible to the basis set superposition error (BSSE), which results in unphysical lowering of the interaction energy when the two constituents come close to each other. To correct for BSSE, we take the full counterpoise approach:[49] at each separation r, we optimize the dimer geometry and carry out four additional calculations (constituent A along and constituent B along, with and without ghost orbitals), and obtain the BSSE-corrected interaction energy as

$$E(r) = E_{\mathrm{AB}}(r) - E'_{\mathrm{A}}(r) - E'_{\mathrm{B}}(r) + E_{\mathrm{A}}(r) + E_{\mathrm{B}}(r) - E_{\mathrm{A}^*} - E_{\mathrm{B}^*}, \quad (1)$$

where the subscript denotes the geometry: AB is the optimized dimer geometry, A and B are its constituent geometries, and A* and B* are the individually optimized constituent geometries; the prime indicates that ghost orbitals of the other constituent are used in the energy calculation.

III. CONSTRUCTION OF MODEL POTENTIALS

We coarse-grain each nucleotide into two interaction sites: one for the base and one for the sugar-phosphate backbone. The base site is identified with the position of the nitrogen atom (N1 for pyrimidines; N9 for purines) that connects the base to the sugar, and the backbone site is identified with the position of the sugar C1′ atom. This representation is illustrated in Fig. 1. This choice of coarse-grained site coordinates allows for unambiguous determination of bonding distance, bonding angles, etc., that enter the model variables. It also

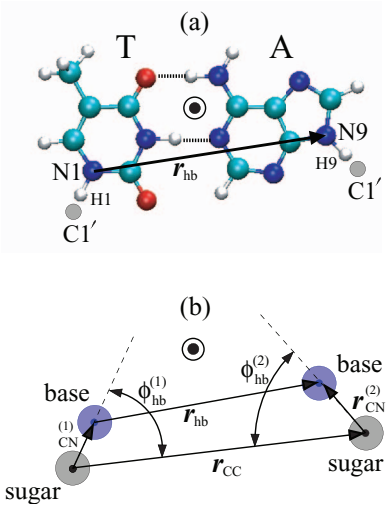

FIG. 1. Schematic of an adenine-thymine (AT) base-pair showing the geometric variables in the hydrogen bond potential in (a) the atomic structure calculations and (b) the coarse-grained model. The distance r_{hb}, the two flip-angles $\phi_{hb}^{(i)}$, and the dihedral angle θ_d (angle between the two base planes, not shown) are used to describe interaction between the two bases. The vectors $\mathbf{r}_{CC}$ and $\mathbf{r}_{CN}^{(i)}$ are used to define the normal vectors of the two bases and the dihedral angle. In this figure and Figs. 4 and 7, the colored spheres represent C (cyan), H (white), N (blue), O (red), and P (brown) atoms.

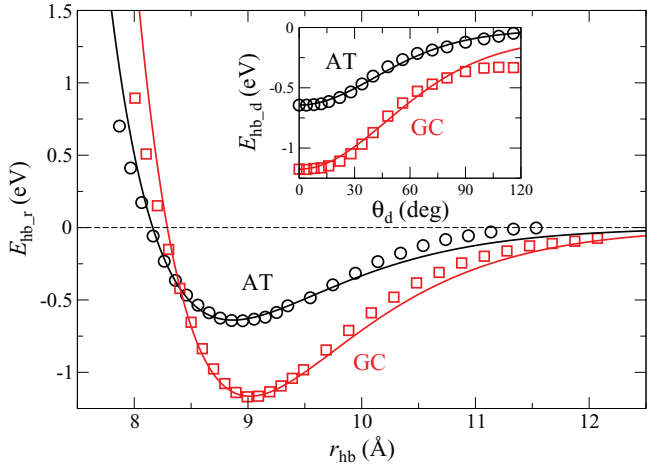

FIG. 2. Hydrogen bonding energy versus distance r_{hb} between two complimentary bases. Inset shows the dependence on the dihedral angle θ_d between the two bases. Symbols are BSSE-corrected results from DFT calculations, and lines are fittings to Eqs. (3) and (4) with parameters given in Table I.

facilitates direct comparison to experimental structures during validation of the model.

To derive the effective interaction between these coarse-grained sites, we follow earlier work[50] and decompose the total interaction energy into contributions that have distinct physical meanings: the hydrogen bond between complementary bases E_{hb}, the stacking energy between neighboring base-pairs E_{st}, contributions from the backbone E_{bb}, and electrostatic interaction between the charged phosphate groups E_{el}:

$$E_{total} = E_{hb} + E_{st} + E_{bb} + E_{el}. \quad (2)$$

These are physically distinct contributions, and in our model they are treated as independent and additive. Each contribution depends on several variables, and the effect of these variables are not taken as independent.

The choice of the coarse-grained sites and the decomposition of the interaction potentials has determined the structure of the model. The remaining construction of the model is to find explicit functional forms for each contribution, which we address in the following.

A. Hydrogen bonding

We consider first the interaction between two complementary bases: adenine-thymine (AT) or guanine-cytosine (GC), which comes from hydrogen bonds. We consider its dependence on the base-to-base separation and on the relative angles between the planes of the two bases. The angular dependence keeps the two bases coplanar and maintains the correct base-pair geometry.

We calculate the interaction energy as a function of the distance r_{hb} between the pyrimidine N1 atom and the purine N9 atom (see Fig. 1), by starting from the energy-minimized geometry and varying r_{hb} by translating the two bases parallel to the direction of the hydrogen bonds. At each r_{hb} value, we optimize the geometry while fixing the four atoms that correspond to the coarse-grained sites (N1 and H1 of the pyrimidine, and N9 and H9 of the purine) to preserve r_{hb} and the relative angles. At small r_{hb} values, the atoms are also constrained on the base-pair plane to prevent the two bases from rotating out of plane. The effect of flipping angles and of non-planarity will be examined separately.

Figure 2 shows the calculated interaction energy, which can be described by the universal binding energy relation (UBER)[51]

$$E_{hb\_r}(r_{hb}) = E_0(1 + a^*)e^{-a^*}, \quad a^* = (r_{hb} - r_0)/l, \quad (3)$$

with its minimum at $r_{hb} = r_0$, $E_{hb\_r} = E_0$. The parameters E_0, r_0, and l are given in Table I. The energy at the minimum is -0.64 eV for AT and -1.17 eV for GC, in agreement with the values -0.60 eV and -1.17 eV, respectively, obtained in previous calculations of the DNA hydrogen bonds.[42] We do not parametrize the base-to-base interaction between mismatched base-pairs, but doing so will be a straightforward extension of the current work.

Next we examine the effect of non-planarity of the bases: when the two planes of the bases are not aligned, the hydrogen bond weakens. We measure non-planarity by the dihedral angle θ_d between the planes of the two bases. Starting from the energy-minimized geometry ($\theta_d = 0$), we vary θ_d by rotating the two bases in opposite directions around the line from the pyrimidine N1 atom to the purine N9 atom. To keep θ_d fixed, we optimize the geometry while holding the 6-fold ring of the

TABLE I. Fitting parameters for the radial dependence of the hydrogen bond interaction in Eq. (3) and the dihedral angle dependence in Eq. (4).

bp	E_0 (eV)	r_0 (Å)	l (Å)	k_d
AT	−0.639	8.866	0.703	1.800
GC	−1.165	9.018	0.727	1.288

TABLE II. Fitting parameters for the flip-angle interaction in Eqs. (6) and (7), and the backbone base-sugar-sugar angle potential in Eq. (18). $k_{\rm bss}$ is in 10^{-4} eV/deg^2, all angles are in degrees.

Base	$\phi_{\rm hb}^{(i,0)}$	$\sigma^{(i)}$	$k_{\rm bss}$	$\theta_{3'}^{(0)}$	$\theta_{5'}^{(0)}$
A	54.53	18.67	3.489	94.17	61.36
T	55.93	16.82	4.689	92.67	68.40
G	52.69	25.57	4.165	90.30	63.69
C	54.87	22.43	6.178	91.55	66.37

two bases fixed. The resulting energies are shown in the inset of Fig. 2. We fit the interaction energy with the expression

$$E_{\rm hb\_d}(\theta_{\rm d}) = E_0 e^{k_{\rm d}(\cos\theta_{\rm d}-1)}, \tag{4}$$

where E_0 is the same as in Eq. (3); values of the parameter $k_{\rm d}$ are given in Table I.

The interaction between the two bases also depends on the in-plane angles of the two bases. In our coarse-grained model, this is described by the flip-angle $\phi_{\rm hb}^{(i)}$ for the two bases ($i = 1, 2$), defined as the base1-sugar1-sugar2 angle for $i = 1$ and the base2-sugar2-sugar1 angle for $i = 2$ (see Fig. 1(b)). The flip-angles in the ground-state configuration, $\phi_{\rm hb}^{(i,0)}$, are listed in Table II. Starting from the energy-minimized geometry ($\phi_{\rm hb}^{(i)} = \phi_{\rm hb}^{(i,0)}$), we consider either rotating the pyrimidine around the H1 atom, or rotating the purine around the H9 atom. The anchoring hydrogen atom mimics the backbone, which in physiological conditions remains relatively stationary when base flipping occurs. To keep $\phi_{\rm hb}^{(i)}$ fixed at each rotated configuration, we optimize the geometry while fixing the pyrimidine atoms N1 and H1, and the purine atoms N9 and H9. For rotation inward, the atoms are constrained on the base-pair plane to prevent the two bases from rotating out of plane.

The resulting energies are plotted as a function of $\Delta\phi_{\rm hb}^{(i)} = \phi_{\rm hb}^{(i)} - \phi_{\rm hb}^{(i,0)}$ in Fig. 3. When $\Delta\phi_{\rm hb}^{(i)}$ is positive, the energy is attractive and decays to zero. When $\Delta\phi_{\rm hb}^{(i)}$ is negative, the two bases repel each other. We describe these two types of behavior separately, and fit the base-flipping interaction energy to the expression

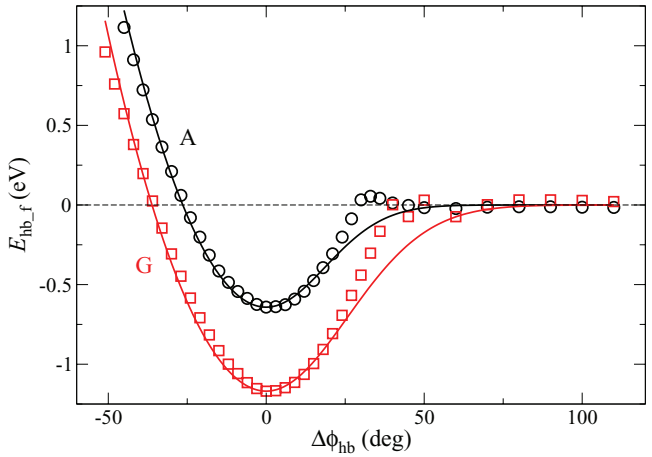

FIG. 3. Hydrogen bonding energy versus flip-angle $\phi_{\rm hb}$ of A and G (results for T and C are similar and are not shown). Symbols are BSSE-corrected results from DFT calculations, and lines are fitting to Eq. (5) with parameters given in Table II.

$$E_{\rm hb\_f}^{(i)}\left(\phi_{\rm hb}^{(i)}\right) = E_{\rm hb\_f}^{(a,i)}\left(\Delta\phi_{\rm hb}^{(i)}\right) + E_{\rm hb\_f}^{(b,i)}\left(\Delta\phi_{\rm hb}^{(i)}\right) \quad (i = 1, 2), \tag{5}$$

where $E_{\rm hb\_f}^{(a,i)}$ is described by an exponential and $E_{\rm hb\_f}^{(b,i)}$ is described by a harmonic spring

$$E_{\rm hb\_f}^{(a,i)} = E_0\left[\exp\left(-\frac{\left(\Delta\phi_{\rm hb}^{(i)}\right)^2}{2\sigma^2}\right)\Theta\left(\Delta\phi_{\rm hb}^{(i)}\right) + \Theta\left(-\Delta\phi_{\rm hb}^{(i)}\right)\right], \tag{6}$$

$$E_{\rm hb\_f}^{(b,i)} = E_0\left(-\frac{\left(\Delta\phi_{\rm hb}^{(i)}\right)^2}{2\sigma^2}\right)\Theta\left(-\Delta\phi_{\rm hb}^{(i)}\right), \tag{7}$$

and Θ is the step function. Again, E_0 is the same as in Eq. (3). The values of the parameter σ are listed in Table II.

Up to this point we have the interaction potential between the two complementary bases as a function of $r_{\rm hb}$, $\theta_{\rm d}$, $\phi_{\rm hb}^{(1)}$, and $\phi_{\rm hb}^{(2)}$ individually while keeping other variables fixed at the equilibrium values. Here we assume that the interaction depends only on these four variables; even in this case, the general dependence is not trivial, as the four variables are not independent. For example, when $\theta_{\rm d}$ or $\Delta\phi_{\rm hb}^{(i)}$ is large, the hydrogen bond is basically broken, and there should no longer be a strong dependence on $r_{\rm hb}$. To capture the interdependences between the variables while keeping a reasonably simple form for the interaction, we define the following functions

$$\tau_{\rm hb\_d}(\theta_{\rm d}) = E_{\rm hb\_d}(\theta_{\rm d})/E_0,$$
$$\tau_{\rm hb\_f}^{(i)}\left(\phi_{\rm hb}^{(i)}\right) = E_{\rm hb\_f}^{(a,i)}\left(\Delta\phi_{\rm hb}^{(i)}\right)/E_0 \quad (i = 1, 2) \tag{8}$$

and take the final expression of the hydrogen bonding energy to be

$$E_{\rm hb}\left(r_{\rm hb}, \theta_{\rm d}, \phi_{\rm hb}^{(1)}, \phi_{\rm hb}^{(2)}\right) = E_{\rm hb\_r}(r_{\rm hb})\tau_{\rm hb\_d}(\theta_{\rm d})\prod_{i=1}^{2}\tau_{\rm hb\_f}^{(i)}\left(\phi_{\rm hb}^{(i)}\right) + \sum_{i=1}^{2}E_{\rm hb\_f}^{(b,i)}\left(\phi_{\rm hb}^{(i)}\right). \tag{9}$$

Note that the repulsive part of the flipping interaction $E_{\rm hb\_f}^{(b,i)}$ is treated as additive since it does not serve to weaken the bond; this also ensures that the modulation functions $\tau_{\rm hb\_f}^{(i)}$ take values strictly between 0 and 1. Equation (9) is reduced to Eqs. (3)–(5) for close-to-minimum geometries, i.e.,

$$E_{\rm hb}\left(r_{\rm hb}, 0, \phi_{\rm hb}^{(1,0)}, \phi_{\rm hb}^{(2,0)}\right) = E_{\rm hb\_r}(r_{\rm hb}), \tag{10a}$$

$$E_{\rm hb}\left(r_0, \theta_{\rm d}, \phi_{\rm hb}^{(1,0)}, \phi_{\rm hb}^{(2,0)}\right) = E_{\rm hb\_d}(\theta_{\rm d}), \tag{10b}$$

$$E_{\rm hb}\left(r_0, 0, \phi_{\rm hb}^{(1)}, \phi_{\rm hb}^{(2,0)}\right) = E_{\rm flip}\left(\phi_{\rm hb}^{(1)}\right). \tag{10c}$$

Therefore, we expect that close to equilibrium, Eq. (9) serves as a good approximation to the interaction energy between the two bases.

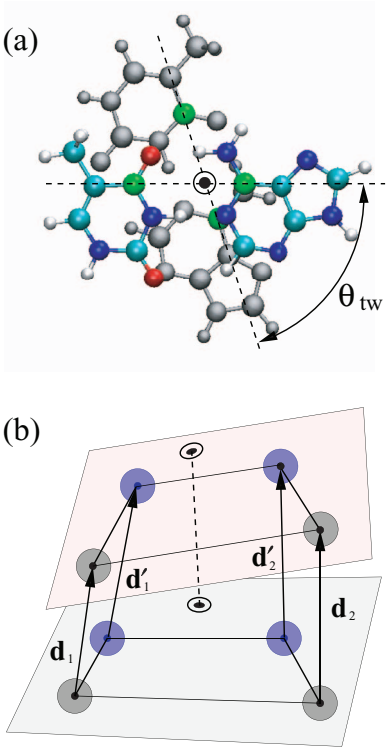

FIG. 4. Schematic showing two stacked base-pairs in (a) the atomic structure calculation (top view), and (b) the coarse-grained model (side view). In (a), the bottom base-pair is shown in gray for clarity. The atoms used to define the twist angle θ_{tw} are colored in green, and the axis of the double helix (shown as a dotted circle) is defined as the 1:2 weighted center of the green atoms. In (b), the planes of the two base-pairs are shown. The twist angle is reduced for clarity of illustration. The four vectors $\mathbf{d}_1$, $\mathbf{d}'_1$, $\mathbf{d}_2$, and $\mathbf{d}'_2$ as indicated are used to define the stacking distance r_{st} in the coarse grained representation. See main text for description of the other geometric variables involved in stacking.

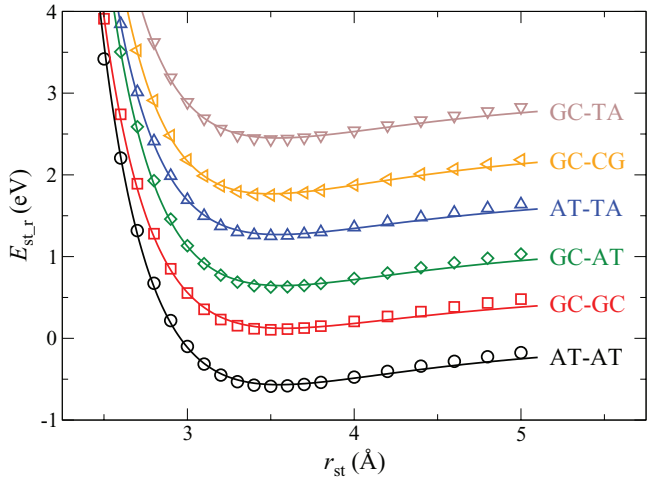

FIG. 5. Stacking energy between two neighboring base-pairs (results for AT-GC, CG-GC, TA-GC, and TA-AT stacking are similar and are not shown). Symbols are BSSE-corrected results from DFT calculations using vdW density functionals, and lines are fitting to Eq. (11) with parameters given in Table III. For clarity, curves are shifted upward by increments of 0.6 eV; the curve for AT-AT stacking is not shifted. In our notation, GC-AT indicates the stacking of GC base-pair and AT base-pair with GA and TC following the 3′–5′ direction.

B. Stacking interactions

We turn next to the interaction between two stacked base-pairs. We consider the dependence on the stacking distance r_{st} between the two base-pairs first. The two base-pairs are relaxed and stacked in parallel. Following usual conventions,[1] we define the axial point of each base-pair to be the 1:2 weighted center of the pyrimidine C4 atom and the purine C6 atom, and define the twist angle θ_{tw} to be the angle made by the C4–C6 vector of the two base-pairs. This is illustrated in Fig. 4(a). The two base-pairs are stacked so that their axial points align at a distance r_{st} apart in the direction normal to the base-pair plane, and so that the twist angle θ_{tw} is 36°. The precise definition of r_{st} and θ_{tw} in the coarse-grained model will be described in the Sec. IV.

There are ten different combinations of stacking, known as the ten Watson-Crick nearest-neighbors. For each of these ten combinations, we vary r_{st} from 2.5 Å to 5.0 Å and calculate the interaction energy without further relaxation. The results are shown in Fig. 5. At this range of r_{st}, the energy variation is well described by the expression

$$E_{st\_r}(r_{st}) = -\epsilon\left[5\left(\frac{r_m}{r_{st}}\right)^6 - 6\left(\frac{r_m}{r_{st}}\right)^5\right], \qquad (11)$$

which has a minimum at $r_{st} = r_m$, $E_{st\_r} = \epsilon$. The resulting values of the parameters from fitting the *ab initio* values with this expression are listed in Table III. It may seem surprising that the attractive part of the interaction behaves as r_{st}^{-5}. This is no coincidence: the attraction between the two base-pairs is due to the vdW force, which behaves as r^{-6} for two point particles and as r^{-4} for two thin sheets (assuming additivity of the vdW interaction). The range of r_{st} being considered here is comparable to the radius of the base-pairs (about 4 Å), so we expect a power in between the two, i.e., r_{st}^{-5}. We find this is indeed the case; expressions with any other integer power of r_{st} lead to poor fitting. Again, we do not parametrize the stacking interaction for mismatched base-pairs, which can be a straightforward extension to the current work.

Another dependence of the stacking interaction comes from the twist angle θ_{tw}. To examine this dependence, we fix r_{st} at 3.4 Å and vary θ_{tw} from 0° to 360°. Again, the two base-pairs are parallel and have their axial points aligned. The

TABLE III. Fitting parameters of the stacking interaction in Eq. (11) for the ten Watson-Crick nearest-neighbors. In our notation, GC-AT indicates the stacking of GC base-pair and AT base-pair with GA and TC following the 3′–5′ direction.

bps	r_m (Å)	ε (eV)
AT-AT	3.550	−0.567
GC-GC	3.580	−0.477
GC-AT	3.566	−0.555
AT-TA	3.546	−0.530
GC-CG	3.498	−0.632
GC-TA	3.543	−0.549
AT-GC	3.535	−0.538
CG-GC	3.537	−0.563
TA-GC	3.613	−0.529
TA-AT	3.668	−0.513

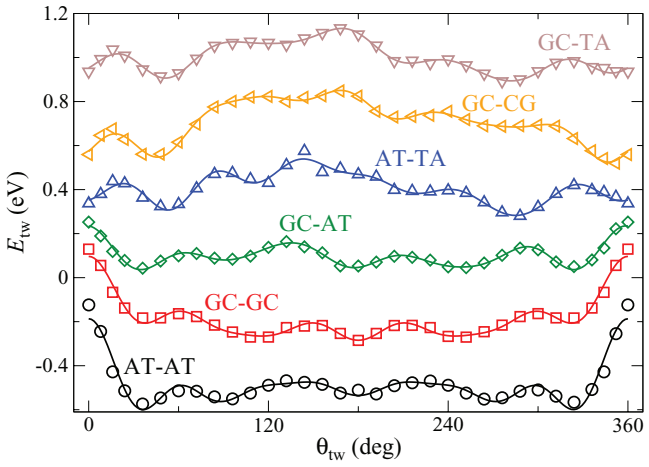

FIG. 6. Twisting energy between two neighboring base-pairs. Symbols are BSSE-corrected results from DFT calculations using vdW density functionals, and lines are fitting to Eq. (12) with parameters given in Table IV. For clarity, curves are shifted upward by increments of 0.3 eV; the curve for AT-AT stacking is not shifted. Notation is same as in Fig. 5. Note that, although there are 10 unique stacking combinations, only 6 are necessary in the evaluation of E_{tw}. For example, $E_{tw}(\theta_{tw})$ of AT-GC is given by $E_{tw}(360^\circ - \theta_{tw})$ of GC-AT.

resulting energies E_{tw} are shown in Fig. 6. This energy term has a complex dependence on θ_{tw}, defying a simple analytical expression. Given its periodicity, we fit E_{tw} with a Fourier series:

$$E_{tw}(\theta_{tw}) = a_0 + \sum_{n=1}^{7} [a_n \cos(n\theta_{tw}) + b_n \sin(n\theta_{tw})]. \quad (12)$$

The resulting values of the coefficients in this expansion are given in Table IV. For AT-AT and GC-GC stacking, $E_{tw}(-\theta_{tw}) = E_{tw}(\theta_{tw})$ by symmetry, and so the coefficients b_n are zero. We note that although Eq. (12) involves many trigonometric functions, the higher terms can be obtained from trigonometric addition rules or from the Chebyshev method. Therefore, its computational cost is similar to that of Eq. (11).

The stacking interaction also depends on the tilt-angle θ_{tl}, which we define as the angle between the normal vectors of the two base-pair planes. The two base-pairs can tilt in a variety of ways, and so this dependence can be complex. We find that close to the optimal stacked configuration (at small θ_{tl}, and $r_{st} = r_m$, $\theta_{tw} = 36^\circ$), the interaction energy dependence can be approximately described by

$$E_{tl}(\theta_{tl}) = \epsilon \cos^2(\theta_{tl}), \quad (13)$$

where ϵ is the same as in Eq. (11). This expression also has the physical meaning that the energy is at a minimum in the case of parallel ($\theta_{tl} = 0$) and anti-parallel ($\theta_{tl} = \pi$) stacking, and is zero when the two base-pairs are perpendicular to each other ($\theta_{tl} = \pm\pi/2$).

The interaction between two stacked base-pairs depends on r_{st}, θ_{tw}, and θ_{tl}. To capture all three dependences and their correlations with a reasonably simple form, we define the following functions:

$$\tau_{tw}(\theta_{tw}) = E_{tw}(\theta_{tw})/\epsilon, \quad \tau_{tl}(\theta_{tl}) = E_{tl}(\theta_{tl})/\epsilon \quad (14)$$

and take the final expression of the stacking interaction to be

$$E_{st}(r_{st}, \theta_{tw}, \theta_{tl}) = E_{st\_r}(r_{st})\tau_{tw}(\theta_{tw})\tau_{tl}(\theta_{tl}), \quad (15)$$

similar to our treatment of the hydrogen bond. Equation (15) involves much simplification, as the true interaction energy may depend on more than these three variables, and the dependence on r_{st}, θ_{tw}, and θ_{tl} may not factorize. By the same argument as for the hydrogen bonds, though, we expect Eq. (15) to be a good approximation close to equilibrium.

The stacking potential here is formulated as interaction between neighboring base-pairs, rather than between neighboring bases. This approach is sufficient when dealing with ds-DNA structures. However, we note that it may not be appropriate in processes like melting or hybridization that involve single-stranded DNA or broken base-pairs. In such situations, the stacking interaction should be reformulated as interaction between bases instead and be constructed in a similar fashion.

C. Backbone contribution

We next consider contributions from the backbone. First, to extract interactions within the sugar-phosphate backbone as distinct from any electrostatic or stacking interactions, we take the phosphate groups to be protonated and carry no charge, and replace the bases with terminating hydrogens. Starting from the energy-minimized geometry, we uniformly stretch or compress the backbone along the helical direction, as illustrated in Fig. 7(a), and allow the phosphate units to relax. The resulting energy E_{bb_r} is shown in Fig. 8 as a function of r_{ss}, the distance between the C1′ atoms of neighboring sugars. The interaction energy per sugar-phosphate unit can be described by the expression

$$E_{bb\_r}(r_{ss}) = c_2(r_{ss} - r_{ss\_0})^2 + c_4(r_{ss} - r_{ss\_0})^4 \quad (16)$$

with $c_2 = 18.773$ eV/Å^2, $c_4 = 0.333$ eV/Å^4, and $r_{ss\_0} = 4.976$ Å.

TABLE IV. Fitting parameters of the twisting interaction in Eq. (12). All parameters in units of 0.01 eV. Notation is same as in Table III.

bps	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	b_1	b_2	b_3	b_4	b_5	b_6	b_7
AT-AT	−49.2	2.89	5.52	5.88	3.07	5.93	4.55	2.73							
GC-GC	−49.0	8.84	5.29	2.65	2.33	5.43	1.96	2.13							
GC-AT	−50.1	1.56	1.18	2.29	1.07	4.48	1.87	1.09	1.13	−2.00	0.37	0.97	−1.02	−0.02	−0.34
AT-TA	−49.4	−5.01	2.62	1.57	−0.16	−1.19	−2.01	−1.20	4.70	−2.71	−0.67	−1.41	2.21	2.39	−0.82
GC-CG	−49.2	−11.0	−2.39	0.21	1.12	−0.72	0.25	−0.94	2.43	−3.57	0.05	1.14	3.85	1.34	1.34
GC-TA	−50.1	−5.73	2.53	0.56	1.06	−1.86	−0.37	−0.86	4.72	−2.73	−0.46	−0.34	3.16	1.16	0.48

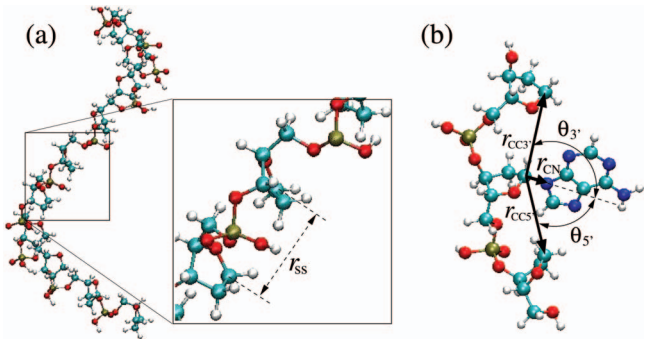

FIG. 7. (a) A single backbone strand in its relaxed B-DNA form, used to evaluate the backbone interaction E_{bb_r}. The variable r_{ss} is the distance between the C1′ atoms of neighboring sugars, and is shown in the magnified part. (b) Structure used to evaluate the base-sugar-sugar interaction E_{bb_b}. The vectors used to define the angles $\theta_{3'}$ and $\theta_{5'}$ are shown.

The base is covalently bonded to the backbone. For this interaction, we consider only a base and a sugar, with the phosphate groups replaced by hydrogen atoms. Starting from the energy-minimized geometry, we translate the base and the sugar groups in opposite directions along the vector connecting the two. Then we hold the two atoms of the base-sugar covalent bond fixed, optimize the geometry, and calculate the interaction energy. We find results for the four bases to be nearly identical, so only results for guanine will be discussed. These are shown in Fig. 9, with a fit to the UBER expression that includes two additional terms

$$E_{bb\_c}(r_{CN}) = E_0(1 + a^* + f_2 a^{*2} + f_3 a^{*3})e^{-a^*},$$
$$a^* = (r_{CN} - r_0)/l, \tag{17}$$

where r_{CN} is the distance between the sugar C1′ atom and the base N atoms (N9 of purines or N1 of pyrimidines). The values of the parameters are: $E_0 = -3.542$ eV, $r_0 = 1.455$ Å, $l = 0.400$ Å, $f_2 = -0.132$, and $f_3 = 0.215$.

The base also interacts with neighboring backbone groups, and this gives rise to the 5′/3′ asymmetry of the double-helix. We characterize this interaction using the angle between the backbone strand and the base. We consider a segment of single-stranded DNA, with three sugar groups connected by two protonated phosphate groups. The first and third bases are replaced with terminating hydrogen atoms, and the middle base is either A, C, G, or T (see Fig. 7(b)). From the energy-minimized geometry, we rotate the middle base around the vector $\mathbf{r}_{CC5'} \times \mathbf{r}_{CN}$ (both are shown in Fig. 7(b)), with the pivot point at C1′ of the middle sugar. For each rotated configuration, we relax the system while holding the connecting N atom of the base (N9 of purines or N1 of pyrimidines) and the C1′ atoms of the three sugars fixed. The resulting energy is shown in the inset of Fig. 9. We consider this energy to be a function of the two angles $\theta_{3'}$ and $\theta_{5'}$, where $\theta_{3'}$ is the angle between $\mathbf{r}_{CN}$ and $\mathbf{r}_{CC3'}$, and $\theta_{5'}$ is the angle between $\mathbf{r}_{CN}$ and $\mathbf{r}_{CC5'}$ (see Fig. 7(b)). We fit this energy term with the function

$$E_{bb\_b} = \frac{k_{bss}}{2}\left[\left(\theta_{3'} - \theta_{3'}^{(0)}\right)^2 + \left(\theta_{5'} - \theta_{5'}^{(0)}\right)^2\right], \tag{18}$$

where $\theta_{3'}^{(0)}$ and $\theta_{5'}^{(0)}$ are the angles in the initial configuration relaxed without constraint, and k_{bss} is obtained from fitting. The resulting values for these parameters are listed in Table II.

We treat Eqs. (16)–(18) as independent interactions, and so the total contribution of the backbone and its interaction with the base-pairs is given by

$$E_{bb} = E_{bb\_r} + E_{bb\_c} + E_{bb\_b}. \tag{19}$$

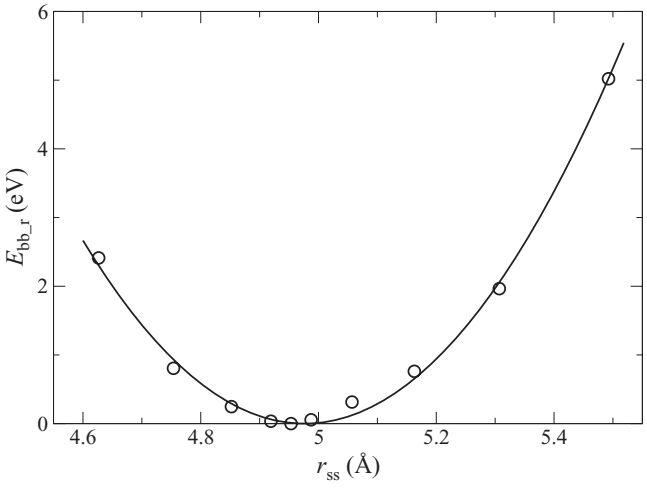

FIG. 8. Backbone energy per sugar-phosphate unit as a function of r_{ss}, the distance between adjacent C1′ atoms of the sugars. Solid line is fitting to Eq. (16).

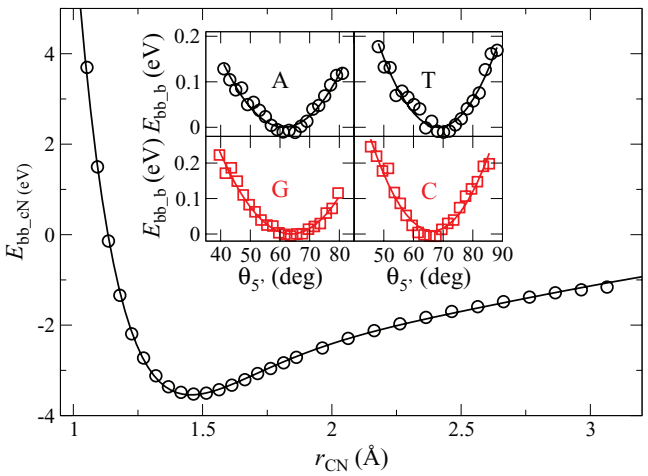

FIG. 9. Interaction energy between the backbone and the base. Main plot shows the covalent bond energy between the sugar and the base, as a function of the distance between the sugar C1′ atom and the base N1 or N9 atom to which it is bonded; solid line is fitting to Eq. (17). Inset shows the energy as a function of the sugar-sugar-base angle $\theta_{5'}$ defined in Fig. 7(b), and lines are fits to Eq. (18) with parameters given in Table II.

D. Electrostatics

All the potentials considered so far represent bonded or short-ranged interactions between components of DNA. The DFT calculations considered DNA as being composed of strictly neutral components in vacuum. However, under

physiological conditions, DNA is solvated in an aqueous electrolyte, and the phosphate groups in DNA are deprotonated. Therefore, we place a charge of $-e$ on each sugar-phosphate group of the coarse grained model, and include a Coulomb potential between these groups

$$E_{\rm el}(r) = \frac{e^2}{4\pi\varepsilon_0\varepsilon(r)r} \tag{20}$$

with a distance-dependent dielectric function $\varepsilon(r)$, r being the distance between the charges and ε_0 the dielectric constant in vacuum. This $\varepsilon(r)$ plays two roles: (i) it incorporates the effects of ionic screening, and (ii) it accounts for the fact that closely spaced interacting groups are only partially solvated by the surrounding electrolyte. We use the following expression, closely related to the formulation of Ref. 52, for the dielectric function:

$$\varepsilon(r) = \begin{cases} \varepsilon_{\rm int}, & \text{for } r < r_0 \\ \varepsilon_{\rm int} e^{\alpha(r-r_0)}, & \text{for } r_0 < r < r_1 \\ \varepsilon_\infty e^{\kappa r}, & \text{for } r > r_1 \end{cases}, \tag{21}$$

where $\varepsilon_\infty = 78$ is the dielectric constant of water, and $\varepsilon_{\rm int}$ is the dielectric constant in the *interior* of the DNA helix, which we take to be 3.[52] The Debye length is κ^{-1}, related to the ionic strength I through

$$\kappa^{-1} = \sqrt{\frac{\varepsilon_0\varepsilon_\infty k_{\rm B}T}{2N_A e^2 I}}, \tag{22}$$

where k_BT is the thermal energy and N_A is Avogadro's number. For example, at $[{\rm Na}^+] = 0.1$ M, the Debye length is 9.6 Å. The values r_0 and r_1 determine the boundary between unscreened and screened electrostatic interactions. The characteristic sizes of the chemical groups represented by our coarse-grained units range from about 2 Å to about 5 Å. Consequently we set $r_0 = 4$ Å. Similarly, we choose $r_1 = 13$ Å, approximately five times the mean water oxygen–water oxygen distance in bulk water,[53] as the distance where the effective dielectric constant recovers the value predicted from the Debye-Hückel theory of screening in a bulk electrolyte. The value of α is then chosen such that $\varepsilon(r)$ is continuous. Between the two charged groups within the same base-pair, $\varepsilon(r) = \varepsilon_\infty$ is used. Finally, since the electrostatic interaction decays exponentially at large distance, we truncate this interaction for distances above five times the Debye length. The simple electrostatics approach here is a first approximation, and is not meant to capture details such as the dependence of ion condensation on conformation,[54] which will require explicit solvent molecules.

IV. IMPLEMENTATION AND PERFORMANCE OF THE COARSE-GRAINED MODEL

In Sec. III we derived the interactions between the coarse-grained sites. The total interaction energy is given by Eq. (2), and the different contributions include summations over all relevant interacting units. Specifically, $E_{\rm hb}$ includes a summation over all base-pairs, and $E_{\rm st}$ includes a summation over all pairs of neighboring base-pairs. Of the three terms that comprise the $E_{\rm bb}$ contribution, $E_{\rm bb\_r}$ includes a summation over all pairs of neighboring backbone sites, $E_{\rm bb\_c}$ includes a summation over all pairs of connected backbone and base sites, and $E_{\rm bb\_b}$ includes a summation over all bases. Finally, $E_{\rm el}$ includes a summation over all pairs of backbone sites.

Next, we relate the geometrical variables to the positions of the coarse-grained sites. In the coarse-grained representation, the base site is identified with be the pyrimidine N1 atom or the purine N9 atom, and the backbone site is identified with the sugar C1′ atom. Then the distance and angle variables $r_{\rm hb}$, $\phi_{\rm hb}^{(1)}$, $\phi_{\rm hb}^{(2)}$, $r_{\rm ss}$, $r_{\rm CN}$, $\theta_{3'}$, and $\theta_{5'}$ follow from the position of the coarse-grained sites. We define the remaining geometrical variables as follows. For each base-pair, we define the normal vector of base i to be $\mathbf{n}^{(i)} = \mathbf{r}_{\rm CC} \times \mathbf{r}_{\rm CN}^{(i)}$ with $i = 1, 2$ corresponding to the two bases, and with $\mathbf{r}_{\rm CC}$ and $\mathbf{r}_{\rm CN}^{(i)}$ defined in Fig. 1(b). The dihedral angle $\theta_{\rm d}$ of this base-pair is given by $\cos\theta_{\rm d} = \hat{\mathbf{n}}^{(1)} \cdot \hat{\mathbf{n}}^{(2)}$, with the hat denoting unit vectors. We define the average normal of this base-pair to be $\mathbf{n} = \mathbf{n}^{(1)} + \mathbf{n}^{(2)}$, and let the tilt-angle $\theta_{\rm tl}$ between base-pair j and base-pair $j+1$ be given by $\cos\theta_{\rm tl} = \hat{\mathbf{n}}_j \cdot \hat{\mathbf{n}}_{j+1}$. As for the stacking distance $r_{\rm st}$, we take the average distance of the four sites in base-pair $j+1$ to the plane of base-pair j. Specifically, we compute the four vectors $\mathbf{d}_1$, $\mathbf{d}_1'$, $\mathbf{d}_2$, and $\mathbf{d}_2'$ shown in Fig. 4(b), project them onto the two normal vectors as $z_i = \mathbf{d}_i \cdot \hat{\mathbf{n}}^{(i)}$ and $z_i' = \mathbf{d}_i' \cdot \hat{\mathbf{n}}^{(i)}$ for $i = 1, 2$, and take $r_{\rm st} = (z_1 + z_1' + z_2 + z_2')/4$. Finally, the twist angle $\theta_{\rm tw}$ is defined as the angle between $\mathbf{r}_{{\rm CC},j}$ and $\mathbf{r}_{{\rm CC},j+1} - (\mathbf{r}_{{\rm CC},j+1} \cdot \hat{\mathbf{n}}_j^{(1)})\hat{\mathbf{n}}_j^{(1)}$, which is the projection of $\mathbf{r}_{{\rm CC},j+1}$ onto the plane of base-pair j.

We implement this model for calculations in the microcanonical ensemble (constant number of particles N, volume V, and energy E) and the canonical ensemble (constant N, V, and temperature T) with implicit solvent Brownian dynamics. We choose the mass of each site to be the total mass of the group of atoms that it represents: 178 amu for the backbone site, 134 amu for base A, 125 amu for T, 150 amu for G, and 110 amu for C. Multiple-time-step integrators are used for time propagation: the RESPA algorithm[55] is used for the NVE ensemble, and the multiple-time-step stochastic integrator[56] is used for Brownian dynamics. The Coulomb interaction between different base-pairs is treated as the slow-varying component of the Hamiltonian, integrated with a time-step of 20 fs, while all the other forces are categorized into the fast-varying component of the Hamiltonian, integrated with a time-step of 6.67 fs (i.e., three divisions per large time-step). This choice of time-steps ensures high stability: when running in the NVE ensemble at 300 K, the total energy is conserved to within 5×10^{-4} eV per base-pair for arbitrary sequences at several conditions tested.

As a measure of the performance of this coarse-grained model, a 20 fs step of the stochastic integrator for a DNA molecule consisting of 250 base-pairs at 0.1 M salt concentration takes 2.4 ms on a single Intel Xeon X5560 processor (2.80 GHz). A 100-ns simulation of such a strand takes 3 h 20 min. This performance exceeds the speed of all-atom simulation methodologies by many orders of magnitude, and is comparable to other coarse-grained models with similar complexity. The computation time scales linearly with the number

of base pairs, since the electrostatic interactions are damped. This performance makes simulations of ds-DNA in the μm length scale and μs time scale feasible.

V. MODEL VALIDATION

The different contributions in the model potentials have been derived from calculations of isolated units. Here we test the validity of the combined potential (additivity of the individual contributions), and its performance in larger systems (transferability of the potentials). Also, our *a posteriori* inclusions of the ionic and temperature effects are crude, and their consequences will also be assessed.

A. Equilibrium structure

First, we test the structure prediction of the model with poly-AT DNA consisting of 250 base-pairs (bp) at temperature 300 K and salt concentration 0.1 M. When initialized with coordinates of the B-form DNA,[57] the B-DNA structure is maintained. To test the robustness of our model potential, we also carried out simulations starting with a highly off-equilibrium structure—an elongated and uncoiled double strand—and followed the evolution. Figure 10 shows that in such case, the system is able to coil back to the B-DNA structure. Signatures of the B-DNA structure such as the major groove, minor groove, and the 10 bps-per-turn period are well reproduced. This demonstrates the model's ability to predict the stable double helix structure, without resorting to Go-like potentials that are defined relative to a specific reference structure as in Ref. 14.

To be quantitative, we calculate the averages of many structural properties after the system equilibrates. These quantities are listed in Table V, with comparison to reference values derived from recent crystallographic data of a naturally occurring 16 bps oligomer.[58] The average deviation of our model from experiment is about 2%. Given that no experimental data were used in the model construction, this close agreement with experiment is remarkable.

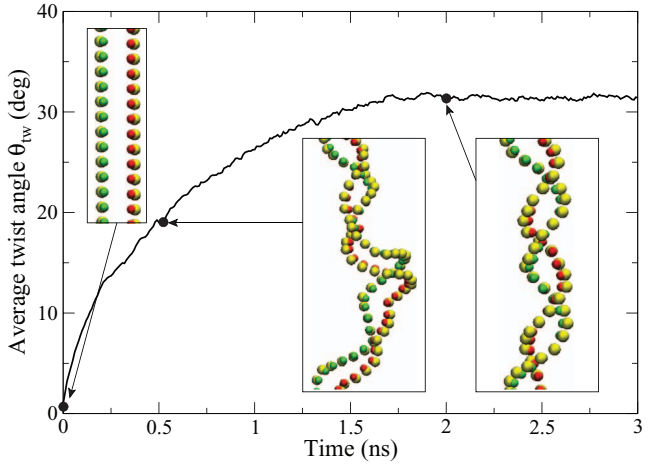

FIG. 10. Coiling of a 250-bp poly-AT strand in the coarse-grained model. The system starts in a completely uncoiled state and evolves at temperature 300 K. Insets show the middle part of the strand at time 0 ns, 0.5 ns, and 2.0 ns. Backbone sites are in yellow, A's in red, and T's in green.

TABLE V. Average structural properties of a 250 base-pair poly-AT strand at 300 K and salt concentration 0.1 M. The reference experimental values are derived from the positions of the C1′ atoms of the sugars, the N1 atoms of the pyrimidines, and the N9 atoms of the purines in the crystal structure of Ref. 58.

Quantity		Simulation	Experiment[58]	Difference (%)
H bond distance	r_{hb} (Å)	8.96	8.93	0.3
Stacking distance	r_{st} (Å)	3.68	3.54	3.9
Backbone distance	r_{ss} (Å)	4.99	4.91	1.5
Sugar-base distance	r_{CN} (Å)	1.46	1.47	0.5
Flip-angle	ϕ_{hb}	54.1°	55.6°	2.7
Twist-angle	θ_{tw}	31.4°	33.5°	6.2
Tilt-angle	θ_{tl}	7.81°	8.00°	2.4
Backbone angle	$\theta_{3'}$	96.5°	94.3°	2.3
Backbone angle	$\theta_{5'}$	62.3°	63.2°	1.5

B. Persistence length

Next, we check the model's ability to predict mechanical properties at long lengths by examining the persistence length (l_p), which gives an estimate of the bending rigidity of DNA. The persistence length can be extracted from the decay of the orientational correlation function

$$\langle \hat{r}_i \cdot \hat{r}_j \rangle = e^{-s_{ij}/l_p}, \tag{23}$$

where $\hat{r}_i$ is the unit tangent vector at base-pair i, and s_{ij} is the arc length from base-pair i to base-pair j. The tangent vectors and the arc length are evaluated along the axis of the double-helix of DNA. As the DNA axis is not an explicit interaction site in our coarse-grained model, we extrapolate its location by estimating the direction of the local helical axis and averaging the projection of the backbone sites onto the local axial direction. We consider 250-bp DNA with poly-AT sequence, poly-GC sequence, and a segment of the enterobacteria phage-λ genome (with GC content 0.47). Simulations at 300 K and 0.15 M salt concentration give $l_p = 53$ nm, 47 nm, and 41 nm, respectively, for the poly-AT, poly-GC, and the phage oligomer strand. These values are in close agreement with the experimental value of $l_p \approx 50$ nm for double-stranded DNA.[59] The fittings used to obtain these values are shown in the inset of Fig. 11.

We also test the validity of the implicit salt treatment by considering the salt dependence of the persistence length. This dependence has been shown experimentally[59] to agree well with the nonlinear Poisson-Boltzmann prediction for uniformly charged cylinders[60]

$$l_p = l_0 + \frac{1}{4\kappa^2 l_B}, \tag{24}$$

where l_0 is the persistence length at infinite salt concentration, κ^{-1} is the Debye length given in Eq. (22), and l_B is the Bjerrum length. The second term is the electrostatic contribution, and is inversely proportional to the salt concentration.

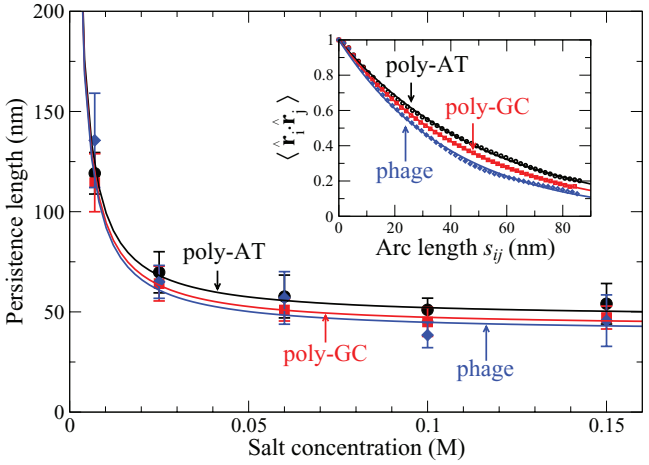

FIG. 11. Salt dependence of the persistence length at 300 K. Each point represents an average over 6 independent simulations, with the error bar showing standard deviation of l_p estimated from the individual runs. Lines show Eq. (24) with $l_0 = 47$ nm, 42 nm, and 39 nm, respectively, for poly-AT, poly-CG, and phage sequences. Inset shows the decay of the orientational correlation for a set of simulations at 0.15 M salt, where symbols are data and lines are exponential fits, Eq. (23).

Figure 11 shows that the salt dependence of l_p predicted by our model also agrees with Eq. (24).

C. Overstretching

To test the mechanical properties of the model under more extreme conditions, we perform numerical stretching experiments, using a 100-bp DNA from a segment of phage-λ, at temperature 300 K in 0.1 M salt. One end of the DNA is fixed to a surface with a harmonic potential, and a constant pulling force is applied onto the backbone site at the other end. It is known that at around 65 pN, the stretched DNA undergoes a sudden extension of about 70% known as superstretching.[61,62] Figure 12 shows the extension curve from our model for pulling on either the 3′ end or the 5′ end. Inset of Fig. 12 provides a typical image of the DNA before and after the sudden stretching and unwinding occur. The simulations capture the superstretching transition, although the critical force in our model is higher, and the amount of the

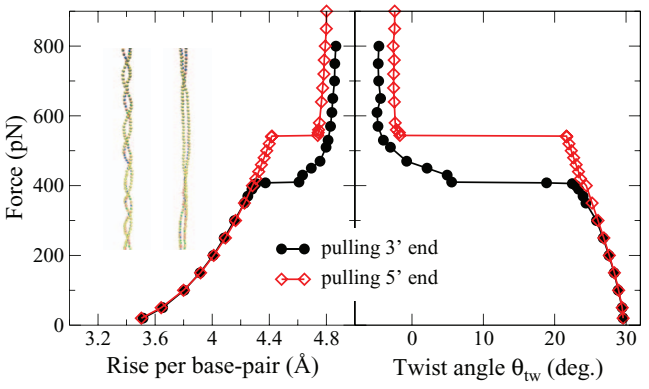

FIG. 12. Average of the rise per base-pair and the twist angle when a 100-bp phage-segment DNA is stretched. Inset shows snapshots before and after the sudden extension, for pulling with 640 pN on the 5′ end.

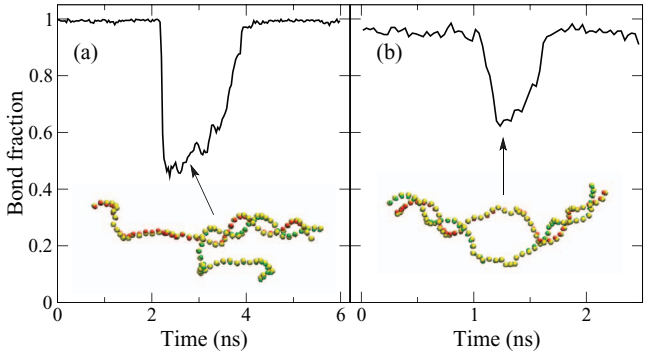

FIG. 13. A 50-bp poly-AT DNA in partially molten states: (a) unzipping from one end, and (b) formation of a bubble. Insets show snapshots of the molecule; the color schemes are the same as in Fig. 10.

sudden extension is less. The superstretching transition is a very stringent test for our computational model, as it involves strong departures from the equilibrium structure. Notwithstanding the neglect of solvent interactions and the athermal character, our coarse-grained potential still provides reasonable results under such critical conditions. Simulations also reveal that the superstretching is accompanied by unwinding of the double helix, as shown in Fig. 12. Above the critical force, our model shows that the DNA is twisted slightly in the left-handed direction.

D. Bubble formation

Although our potentials are formulated for ds-DNA, in some applications it is desirable to have a model that can account for broken hydrogen bonds. Here, we test the validity of our model when the double strand in a 50-bp poly-AT DNA is partially broken. The onset of melting can be observed in our model, but the melting temperature is overestimated. At temperature 550 K, the double-strand shows occasional unzipping; one instance is shown in Fig. 13(a). At 800 K, formation of bubbles can be observed; one instance is shown in Fig. 13(b). The overestimation of melting temperature may arise from several reasons. First, our stacking potential is not formulated for single strands or strands with broken hydrogen bonds. Second, the presence of water molecules may lower the energy of the broken-bond states, but this effect is not taken into account. Third, the coarse-grained potentials here are derived at zero temperature, which is stiffer than at finite temperature. However, the appearance of these intermediate molten states is still an indication that our model is stable beyond strictly ds-DNA structures.

VI. CONCLUSION

We have constructed the interaction potentials for a coarse-grained model of double-stranded DNA. The model potentials are derived from first-principles calculations for the DNA bases and base-pairs. Contributions to the potential include hydrogen bonding, base-stacking, backbone stretching, and interactions between bases and the backbone. Analytical functions are fitted to the computed energy, leading to a potential for DNA that is able to handle coarse-grained

configurations of random DNA sequences and can be used to model related biophysical processes. Unavoidably, some simplifying assumptions were used in the model formulation and construction. In order to use a minimal number of empirical terms, only the effect of ionic screening was added to the microscopically derived potential. At a later stage, such empirical term can be replaced by a more microscopic approach.[63]

A series of tests were performed, and they verified that the coarse-grained potential can reproduce the stable B-DNA structure, and that the predicted structure matches the known crystallographic structure of DNA to a few % in key structural parameters. The model produces persistence lengths in close agreement with experimental data, and the response of the persistence length to varying salt concentrations also agrees with the prediction for a linked-cylinder model. These tests suggest that the mechanical properties of the coarse-grained model will show realistic trends when subject to the complex electrokinetic environment in the cell or in artificial systems such as the interior of an nanochannel during translocation experiments.

The tests on overstretching and on bubble formation show that the model may capture the qualitative features in such states, but also reveal some limitations of the model. To describe such states accurately, we may need a more realistic account of the solvent molecules and of the temperature dependence of the coarse-grained potentials. This will be the subject of future work. Other possible and more straightforward extensions of the current model will be to formulate the stacking interaction to work with ss-DNA, and to parametrize interactions for mismatched base-pairs and for uracil (to extend the model to RNA). Also, the current model does not account for inter-strand interactions except for the electrostatic repulsion, and so situations like long DNA strands under confinement may not be appropriate for this model.

The approach described in this work is attractive for describing ds-DNA using a minimal of empirically derived parameters and with a good compromise between accuracy and computational efficiency. We suggest that this model could be a useful tool for simulating the behavior of ds-DNA in a variety of biologically relevant or device-related scenarios with modest computational resources.

ACKNOWLEDGMENTS

The authors thank Sheng Meng and Wei Li Wang for a critical reading of the manuscript, and acknowledge G.N. Patey for helpful comments and discussions. C.W.H. thanks Wei Li Wang for his kind helps and fruitful discussions. M.F. acknowledges support by Harvard's Nanoscale Science and Engineering Center, funded by the National Science Foundation, Award Number PHY-0117795. G.L. acknowledges support from the post-doctoral fellowship program of the Natural Science and Engineering Research Council of Canada.

[1]H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Brestscher, H. Ploegh, and P. Matsudaira, *Molecular Cell Biology*, 6th ed. (W. H. Freeman and Company, New York, 2007).
[2]A. D. MacKerell, J. Wiorkiewicz-Kuczera, and M. Karplus, J. Am. Chem. Soc. **117**, 11946 (1995).
[3]R. Lavery *et al.*, Nucleic Acids Res. **38**, 299 (2010).
[4]D. Branton *et al.*, Nat. Biotech. **26**, 1146 (2008).
[5]M. Fyta, S. Melchionna, and S. Succi, J. Polym. Sci., Part B: Polym. Phys. **49**, 985 (2011).
[6]H. L. Dormann, B. S. Tseng, C. D. Allis, H. Funabiki, and W. Fischle, Cell Cycle **5**, 2842 (2006).
[7]H. Y. Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans, and W. Yang, Proteins: Struct. Funct. Genet. **44**, 484 (2001).
[8]R. L. Barnett, P. Maragakis, A. Turner, M. Fyta, and E. Kaxiras, J. Mater. Sci. **42**, 8894 (2007).
[9]K. Drukker and G. C. Schatz, J. Phys. Chem. B **104**, 6108 (2000); K. Drukker, G. Wu, and G. C. Schatz, J. Chem. Phys. **114**, 579 (2001).
[10]R. M. Jendrejack, J. J. de Pablo, and M. D. Graham, J. Chem. Phys. **116**, 7752 (2002).
[11]B. D. Coleman, W. K. Olson, and D. Swigon, J. Chem. Phys. **118**, 7127 (2003).
[12]H. L. Tepper and G. A. Voth, J. Chem. Phys. **122**, 124906 (2005).
[13]F. W. Starr and F. Sciortino, J. Phys.: Condens. Matter **18**, L347 (2006).
[14]T. A. Knotts IV, N. Rathore, D. C. Schwartz, and J. J. de Pablo, J. Chem. Phys. **126**, 084901 (2007).
[15]P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano, J. Chem. Theory Comput. **6**, 1711 (2010).
[16]T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, Phys. Rev. Lett. **104**, 178101 (2010).
[17]M. C. Linak, R. Tourdot, and K. D. Dorfman, J. Chem. Phys. **135**, 205102 (2011).
[18]A. Savelyev and G. A. Papoian, Biophys. J. **96**, 4044 (2009).
[19]M. Maciejczyk, A. Spasic, A. Liwo, and H. A. Scheraga, J. Comput. Chem. **31**, 1644 (2010).
[20]S. M. Gopal, S. Mukherjee, Y. M. Cheng, and M. Feig, Proteins **78**, 1266 (2010).
[21]A. Morriss-Andrews, J. Rottler, and S. S. Plotkin, J. Chem. Phys. **132**, 035105 (2010).
[22]N. B. Becker and R. Everaers, Phys. Rev. E **76**, 021923 (2007).
[23]N. B. Becker and R. Everaers, J. Chem. Phys. **130**, 135102 (2009).
[24]E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo, Biophys. J. **96**, 1675 (2009).
[25]E. J. Sambriski, V. Ortiz, and J. J. de Pablo, J. Phys.: Condens. Matter **21**, 034105 (2009).
[26]E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo, Proc. Natl. Acad. Sci. U.S.A. **106**, 18125 (2009).
[27]M. J. Hoefert, E. J. Sambriski, and J. J. de Pablo, Soft Matter **7**, 560 (2010).
[28]V. Ortiz and J. de Pablo, Phys. Rev. Lett. **106**, 238107 (2011).
[29]A.-M. Florescu and M. Joyeux, J. Chem. Phys. **135**, 085105 (2011).
[30]R. C. DeMille, T. E. Cheatham III, and V. Molinero, J. Phys. Chem. B **115**, 132 (2011).
[31]T. E. Ouldridge, I. G. Johnston, A. A. Louis, and J. P. K. Doye, J. Chem. Phys. **130**, 065101 (2009).
[32]C. W. Hsu, J. Largo, F. Sciortino, and F. W. Starr, Proc. Natl. Acad. Sci. U.S.A. **105**, 13711 (2008); W. Dai, C. W. Hsu, F. Sciortino, and F. W. Starr, Langmuir **26**, 3601 (2010); C. W. Hsu, F. Sciortino, and F. W. Starr, Phys. Rev. Lett. **105**, 055502 (2010).
[33]T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, J. Chem. Phys. **134**, 085101 (2011).
[34]A. Savelyev and G. A. Papoian, Proc. Natl. Acad. Sci. U.S.A. **107**, 20340 (2010).
[35]P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. J. Sham, *ibid.* **140**, A1133 (1965).
[36]J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junguera, P. Ordejón, and D. Sánchez-Portal, J. Phys.: Condens. Matter **14**, 2745 (2002).
[37]A. Hübsch, R. G. Endres, D. L. Cox, and R. R. P. Singh, Phys. Rev. Lett. **94**, 178102 (2005); H. Wang, J. P. Lewis, and O. F. Sankey, *ibid.* **93**, 016401 (2004); **85**, 4992 (2000); R. E. A. Kelly and L. N. Kantorovich, J. Phys. Chem. C **111**, 3883 (2007); S. Meng, W. L. Wang, P. Maragakis, and E. Kaxiras, Nano Lett. **7**, 2312 (2007).
[38]A. Tsolakidis and E. Kaxiras, J. Phys. Chem. A **109**, 2373 (2005); D. Varsano, R. Di Felice, M. A. L. Marques, and A. Rubio, J. Phys. Chem. B **110**, 7129 (2006).
[39]N. Trouiller and J. L. Martins, Phys. Rev. B **43**, 8861 (1991).
[40]J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[41]J. Ireta, J. Neugebauer, and M. Scheffler, J. Phys, Chem. A **108**, 5692 (2004).
[42]T. van der Wijst, C. F. Guerra, M. Swart, and F. M. Bickelhaupt, Chem. Phys. Lett. **426**, 415 (2006).
[43]J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[44]J. Šponer, J. Leszczyński, and P. Hobza, J. Phys. Chem. **100**, 5590 (1996).
[45]P. Hobza, J. Šponer, and M. Polasek, J. Am. Chem. Soc. **117**, 792 (1995).
[46]E. R. Johnson, I. D. Mackie, and G. A. DiLabio, J. Phys. Org. Chem. **22**, 1127 (2009).
[47]Q. Wu and W. Yang, J. Chem. Phys. **116**, 515 (2002).
[48]M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, Phys. Rev. Lett. **92**, 246401 (2004).
[49]S. F. Boys and F. Bernandi, Mol. Phys. **19**, 553 (1970).
[50]W. T. M. Mooij, F. B. van Duijneveldt, J. G. C. M. van Duijneveldt-van de Rijdt, and B. P. van Eijck, J. Phys., Chem. A **103**, 9872 (1999).
[51]A. Banerjea and J. R. Smith, Phys. Rev. B **37**, 6632 (1988).
[52]J. Mazur and R. L. Jernigan, Biopolymers **31**, 1615 (1991).
[53]S. H. Lee and J. C. Rasaiah, J. Phys. Chem. **100**, 1420 (1996).
[54]G. L. Randall, L. Zechiedrich, and B. Montgomery Pettitt, Nucl. Acids Res. **37**(16), 5568 (2009).
[55]M. E. Tuckerman, B. J. Berne, and G. J. Martyna, J. Chem. Phys. **97**, 1990 (1992).
[56]S. Melchionna, J. Chem. Phys. **127**, 044108 (2007).
[57]S. Arnott, P. J. C. Smith, and R. Chandrasekaran, in *CRC Handbook of Biochemistry and Molecular Biology*, 3rd ed., edited by G. D. Fasman (CRC, Cleveland, 1976), Vol. 2, pp. 411–422.
[58]N. Narayana and M. A. Weiss, J. Mol. Biol. **385**, 469 (2009).
[59]C. G. Baumann, S. B. Smith, V. A. Boomfield, and C. Bustamante, Proc. Natl. Acad. Sci. U.S.A. **94**, 6185 (1997).
[60]J. Skolnick and M. Fixman, Macromolecules **10**, 944 (1977).
[61]P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J.-L. Viovy, D. Chatenay, and F. Caron, Science **271**, 792 (1996).
[62]S. B. Smith, Y. Cui, and C. Bustamante, Science **271**, 795 (1996).
[63]S. Melchionna and U. Marini Bettolo Marconi, Europhys. Lett. **95**, 44002 (2011).