

## ***Ab initio* predictions for 3D structure and stability of single- and double-stranded DNAs in ion solutions**

Zi-Chun Mu<sup>1,2,†</sup>, Ya-Lan Tan<sup>1,†</sup>, Ben-Gong Zhang<sup>1</sup>, Jie Liu<sup>1</sup>, and Ya-Zhou Shi<sup>1,\*</sup>

<sup>1</sup> *Research Center of Nonlinear Science, School of Mathematical & Physical Sciences, Wuhan Textile University, Wuhan, 430200, China*

<sup>2</sup> *School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, 430200, China*

### **Abstract**

The three-dimensional (3D) structure and stability of DNA are essential to understand/control their biological functions and aid the development of novel materials. In this work, we present a coarse-grained (CG) model for DNA based on the RNA CG model proposed by us, to predict 3D structures and stability for both dsDNA and ssDNA from the sequence. Combined with a Monte Carlo simulated annealing algorithm and CG force fields involving the sequence-dependent base-pairing/stacking interactions and an implicit electrostatic potential, the present model successfully folds 20 dsDNAs ( $\leq 52$ nt) and 20 ssDNAs ( $\leq 74$ nt) into the corresponding native-like structures just from their sequences, with an overall mean RMSD of  $3.4\text{\AA}$  from the experimental structures. For DNAs with various lengths and sequences, the present model can make reliable predictions on stability, e.g., for 27 dsDNAs with/without bulge/internal loops and 24 ssDNAs including pseudoknot, the mean deviation of predicted melting temperatures from the corresponding experimental data is only  $\sim 2.0^\circ\text{C}$ . Furthermore, the model also quantificationally predicts the effects of monovalent or divalent ions on the structure stability of ssDNAs/dsDNAs.

---

<sup>†</sup>The two authors contributed equally to the work.

<sup>\*</sup>To whom correspondence should be addressed. Email: [yzshi@wtu.edu.cn](mailto:yzshi@wtu.edu.cn)

## Author Summary

To determine 3D structures and quantify stability of single- (ss) and double-stranded (ds) DNAs is essential to unveil the mechanisms of their functions and to further guide the production and development of novel materials. Although many DNA models have been proposed to reproduce the basic structural, mechanical, or thermodynamic properties of dsDNAs based on the secondary structure information or preset constraints, there are very few models can be used to investigate the ssDNA folding or dsDNA assembly from the sequence. Furthermore, due to the polyanionic nature of DNAs, metal ions (e.g.,  $\text{Na}^+$  and  $\text{Mg}^{2+}$ ) in solutions can play an essential role in DNA folding and dynamics. Nevertheless, *ab initio* predictions for DNA folding in ion solutions are still an unresolved problem. In this work, we developed a novel coarse-grained model to predict 3D structures and thermodynamic stabilities for both ssDNAs and dsDNAs in monovalent/divalent ion solutions from their sequences. As compared with the extensive experimental data and available existing models, we showed that the present model can successfully fold simple DNAs into their native-like structures, and can also accurately reproduce the effects of sequence and monovalent/divalent ions on structure stability for ssDNAs including pseudoknot and dsDNAs with/without bulge/internal loops.

## Introduction

DNA can adopt many structures beyond the right-handed B-form double-helices, which takes it far beyond being the molecule that stores and transmits genetic information in biological systems (1,2). Some non-B-form DNAs within the human genes, such as hairpins, triplexes, Z-DNA, quadruplexes, and i-motifs, have been proposed to participate in several biologically important processes (e.g., regulation and evolution), leading to mutations, chromosomal translocations, deletions and amplifications in cancer and other human diseases (1-4). Furthermore, self-assembled functional DNA structures have proven to be excellent materials for designing and implementing a variety of nanoscale structures and devices, including interlocked, walkers, tweezers, shuttles, logic circuits, and origami, which have promising applications ranging from photonic devices to drug delivery (5-8). Since short double- and single-stranded DNA (dsDNA and ssDNA) structures (e.g., duplex, hairpins, pseudoknots, and junctions) are essential to build blocks for the construction of non-B-form DNAs and various nano-architectures, advancement in the knowledge of structures and

key properties (e.g., thermodynamics and mechanics) for these DNAs will be helpful to understand and ultimately control their biological functions and guide the production and development of novel materials (7-9).

Although several experimental methods such as cryo-electron microscopy, X-ray crystallography, NMR spectroscopy, and other single-molecule techniques (e.g., optical/magnetic tweezers and atomic force microscopy) can be used to determine three-dimensional (3D) structures or elastic properties of DNAs (10-15), there are still full of challenges (e.g., time-consuming and expensive) to experimentally provide insight into DNA folding/hybridization. Thus, the field of computer simulation is rapidly evolving to provide finer details on key features of DNA biophysics compared to experimental approaches (16-19). For example, all-atom molecular dynamics (MD) simulations based on force fields, such as CHARMM and AMBER, have been widely used to investigate dynamics, flexibility, mechanics, or form transition of dsDNA helices at the microscopic level (20-24). However, due to the innumerable degrees of freedom, the MD simulations are limited to small DNAs and to short times even with an advanced-sampling approach and parallel tempering scheme (16,24-26).

On the other hand, the simple continuum DNA models (e.g., worm-like chain model), which treat the double helix as a continuous elastic rod with bending and torsional stiffness, are commonly used to well describe mechanical behavior or elastic bending of dsDNA on long length-scales (27-30). Correspondingly, the nearest-neighbor model can predict secondary structures and melting profiles (e.g., free energy and melting temperature) for ssDNA and dsDNA through the combination of free energy minimization, partition function calculations, and stochastic sampling (9,31). However, these simple models are unable to provide any 3D structure information on DNAs.

Therefore, many coarse-grained (CG) DNA models, which represent DNA using a reduced number of interaction sites while striving to keep the important details, have been developed in recent years to model 3D structures or thermodynamic and structural properties of DNAs (32-39). For example, by mapping each nucleotide into six to seven CG beads, the Martini model combined with MD simulations opens the way to perform large-scale modeling of complex biomolecular systems involving DNA, such as DNA-DNA and DNA-protein interactions (40,41). Very recently, a three-bead CG model, MADna, was developed to reproduce the conformational and elastic features

of dsDNA, including the persistence length, stretching/torsion moduli, and twist–stretch couplings (42). However, since the two models need constraints (e.g., predefined elastic network or bonded interactions for paired bases) to maintain a double helix, they cannot be used to study DNA hybridization, melting, and hairpin formation (40-42).

Moreover, some other Go-like models, including 3SPN, oxDNA, and TIS, have been proposed to fill the gap (43-50). The 3SPN model, which reduces the complexity of a nucleotide to three interactions sites (i.e., phosphate, sugar, and base), can successfully capture DNA denaturation/renaturation and provide a reasonable description of other thermomechanical and structural properties for DNAs (e.g., persistence length, bubble formation, major and minor groove widths, and local curvature) by involving in base-stacking and base-pairing interactions (43-45). The oxDNA model uses three collinear sites and a vector normal to the base site to construct the angle-dependent potentials including coplanar base-stacking and linear hydrogen bonding interactions, which are parametrized to accurately describe basic structural, mechanical, and thermodynamic properties of ss/dsDNA (46-48). More significantly, with fine-tuned structural parameters, the model can also treat large DNA nanostructures, such as DNA origami and nanotweezers (48,49). The TIS-DNA is another robust three-interaction-site CG model, and using a set of nucleotide-specific stacking parameters obtained from thermodynamic properties of dimers, the model can reproduce the sequence-dependent mechanical, as well as thermodynamic properties of DNAs, covering the force-extension behavior and persistence lengths of poly(dA)/poly(dT) chains, elasticity of dsDNA and melting temperatures of hairpins (50-52). The use of Go-like interactions (e.g., non-bonded potentials to penalize deviations from a reference structure) can effectively restrict the range of conformations that may be sampled by the CG model, and simultaneously, it also limits the possibility of the model on structure prediction from the sequence.

Recently, the 3dRNA/DNA web server was further developed based on the 3dRNA to build three-dimensional (3D) structures of RNA and DNA from template segments with very high accuracy using sequence and secondary structure information (53,54). Similarly, a pipeline presented by Jreddi and Saiz can also be used to predict DNA hairpins by integrating the existed 2D and 3D structural tools (e.g., Mfold, Assemble, and MD) (55). However, the two structure prediction methods are dependent on secondary structures, while there is still a problem to exactly predict

secondary structures of DNAs (31). Fortunately, a minimal physics-based CG model of nucleic acids named NARES-2P was proposed to fold dsDNA from separate strands without any Go-like potentials and secondary structure information. Although the model was constructed using the bottom-up strategy, where each component of the energy function was fitted separately to the respective potential of mean force obtained from all-atom potential-energy surfaces, it can reproduce many properties of double-helix B-DNA, such as duplex formation, melting temperatures, and mechanical stability (56-58). Contrary to the oversimplification of NARES-2P (i.e., two sites per nucleotide), the HiRE-RNA is an empirical CG model for RNA and DNA, whose resolution is high enough (i.e., six or seven beads for each nucleotide) to preserve many important geometric details, such as base pairing and stacking. Without imposing preset pairings for the nucleotides, the HiRE-RNA can investigate both dynamical and thermodynamic aspects of dsDNA assemblies, as well as the effect of sequences on the melting curves of the duplexes (59,60). Despite the advances, the parameters of the two models may need further validation for quantifying thermodynamic and 3D structure to accord with experiments, especially for ssDNA.

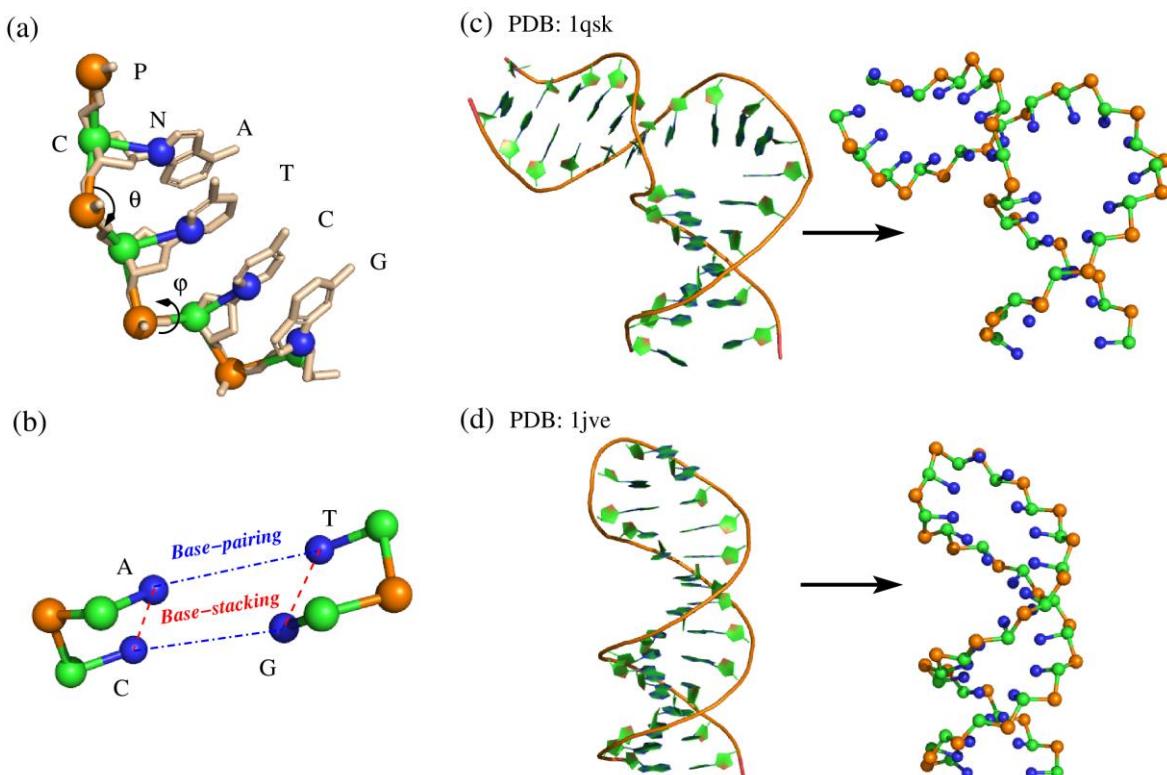
In addition, due to the polyanionic nature of DNAs, metal ions (e.g.,  $\text{Na}^+$  and  $\text{Mg}^{2+}$ ) in solutions can play an essential role in DNA folding and dynamics (12-14,61-64). Although several of the existing models such as 3SPN, oxDNA, TIS, and NARES-2P have taken the electrostatic interactions into account using the Debye-Hückel approximation or mean-field multipole–multipole potentials and reproduced some monovalent salt-dependent structural properties (e.g., persistence length, torsional stiffness or melting temperatures) of DNAs (45,48,50,57), quantitatively predicting the 3D structure and thermodynamic stability for DNA including both ssDNA and dsDNA in ion solutions (especially divalent ions) from the sequence is still an unresolved problem. Recently, we have proposed a three-bead CG model to simulate RNA folding from the sequence, and with an implicit electrostatic potential, the model can make reliable predictions on 3D structures and stability for RNA hairpins, pseudoknots, and kissing complexes in ion solutions (65-69). However, due to the differences in geometry, base stacking strength, and flexibility between DNA and RNA, the present model cannot be directly used to simulate DNA folding.

In this work, we further developed an *ab initio* CG model of DNA to predict the 3D structure, stability, and salt effect for both dsDNA and ssDNA. First, the bonded and nonbonded potentials

were parameterized based on the statistical analysis of known DNA 3D structures, as well as experimental thermodynamic parameters and melting data. Afterward, the model was validated through 3D structure and stability predictions for DNAs including double helices, hairpins, and pseudoknots with different lengths and sequences, as compared with the extensive experimental data. Furthermore, we also showed that the effects of monovalent and divalent ions on DNA structure stability predicted by the present model are in accordance with the corresponding experiments.

## Materials and methods

### CG structure representation for DNAs



**Figure 1.** The representations of all-atom and CG model for each deoxynucleotide and ss/dsDNA molecules, as well as the schematics of base-pairing and base-stacking in the present model. (a) Our coarse-grained representation of a DNA fragments including deoxynucleotides of A, T, G, and C superimposed on the all-atom representation. The three beads are located at phosphate (P), sugar (C4') and pyrimidine (N1) or purine (N9).  $\theta$  and  $\varphi$  are the schematics of CG bonded angle (CPC) and dihedral (CPCP), respectively. (b) The schematic representation of base-pairing (blue) and base-stacking (red) interaction. (c, d) The 3D structures of (c) a dsDNA with bulge loop (PDB:1qsk) and (d) an ssDNA hairpin (PDB: 1jve) in all-atomistic (left) and our CG representation (right). The 3D structures are shown with PyMol (<http://www.pymol.org>).

To be consistent with our previous RNA CG model (65), each nucleotide in DNA is also simplified into three beads: P, C, and N, to represent the phosphate group, sugar ring, and base plane, respectively. For simplicity, the three beads are placed at the positions of existing atoms (i.e., P, C4', and N1 for pyrimidine or N9 for purine) (Fig. 1) and are treated as van der Waals spheres with the radii of 1.9 Å, 1.7 Å and 2.2 Å, respectively (65,69). One unit negative charge ( $-e$ ) is placed on the center of P bead to describe the highly charged nature of DNA.

## Energy functions

The total energy  $U$  in the present DNA CG model is composed of the following eight components:

$$U = U_b + U_a + U_d + U_{exc} + U_{bp} + U_{bs} + U_{cs} + U_{el}. \quad (1)$$

The first three terms are bonded potentials for virtual bonds  $U_b$ , bond angles  $U_a$ , and dihedrals  $U_d$ , respectively, which are used to mimic the connectivity and local geometry of DNA chains. The function forms of these terms are listed in the Supporting information, which can also be found elsewhere (44,46,50,65).

The remaining terms of Eq. 1 describe various pairwise, nonbonded interactions. The  $U_{exc}$  represents the excluded volume interaction between the CG beads and it is modeled by a purely repulsive Lennard-Jones potential. The  $U_{bp}$  in Eq. 1 is an orientation-dependent base-pairing interaction for the possible canonical Watson-Crick base pairs (G-C and A-T). The formula of  $U_{bp}$  is similar to the form of hydrogen-bonding interaction used in the TIS model (50), and the backbone dihedrals are replaced by two simpler distances between CG beads in pairing nucleotides to describe the orientation of hydrogen-bonding interactions; see Eq. S6 in the Supporting information. The sequence-dependence base-stacking interaction  $U_{bs}$  between two nearest neighbour base pairs is given by

$$U_{bs} = \frac{1}{2} \sum_{i,j}^{N_{st}} |G_{i,i+1,j-1,j}| \left\{ \left[ 5 \left( \frac{\sigma_{st}}{r_{i,i+1}} \right)^{12} - 6 \left( \frac{\sigma_{st}}{r_{i,i+1}} \right)^{10} \right] + \left[ 5 \left( \frac{\sigma_{st}}{r_{j,j-1}} \right)^{12} - 6 \left( \frac{\sigma_{st}}{r_{j,j-1}} \right)^{10} \right] \right\}, \quad (2)$$

where  $\sigma_{st}$  is the optimum distance of two neighbour bases in the known DNA helix structures.  $G_{i,i+1,j-1,j}$  in Eq. 2 is the strength of base-stacking energy, and it can be calculated by  $G_{i,i+1,j-1,j} = \Delta H - T(\Delta S - \Delta S_c)$ . Here,  $T$  is the absolute temperature in Kelvin,  $\Delta H$  and  $\Delta S$  are the DNA thermodynamic parameters derived from experiments (9,70), and  $\Delta S_c$  is the conformational entropy

change that is naturally included in the Monte Carlo (MC) simulations, during the formation of one base pair stacking; see more details in Eq. S7 and Fig. S3 in the Supporting information as well as our previous works (65,69). In addition, the coaxial-stacking interaction  $U_{cs}$  between two discontinuous neighbour helices is also taken into account by the present model through calculating the stacking potential of the interfaced base-pairs, and the expression can be found in Eq. S10 in the Supporting information.

The last term  $U_{el}$  in Eq. 1 is used to calculate the electrostatic interactions between phosphates groups (e.g.,  $i$ -th and  $j$ -th P beads), and it is given by

$$U_{el} = \sum_{i < j}^{N_P} \frac{(Qe)^2}{4\pi\epsilon_0\epsilon(T)r_{ij}} e^{-\frac{r_{ij}}{l_D}}, \quad (3)$$

where  $e$  is the elementary charge,  $r_{ij}$  is the distance between  $i$ - and  $j$ -th P beads, and  $N_P$  is the total number of P beads in a DNA.  $l_D$  is Debye length, which defines the ionic screening at different solution ionic strengths.  $\epsilon_0$  and  $\epsilon(T)$  are the permittivity of vacuum and an effective temperature-dependent dielectric constant, respectively (50,65,66).  $Q$  is the reduced charge fraction derived based on Manning's counterion condensation theory and the tightly bound ion model (71-73); see Eq. S11 in the Supporting information. Due to the inclusion of the  $U_{el}$ , the present model can be used to study DNA folding in pure (e.g.,  $\text{Na}^+$ ) as well as mixed (e.g.,  $\text{Na}^+/\text{Mg}^{2+}$ ) ion solutions.

## Parametrization

The initial parameters of bonded potentials (i.e.,  $U_b$ ,  $U_a$ , and  $U_d$  in Eq. 1) were derived from the Boltzmann inversion of the corresponding CG atomistic distribution functions, obtained by the statistical analysis on experimental DNA structures in the Protein Data Bank (PDB) (<http://www.rcsb.org/pdb/-home/home.do>) (Fig. S1), and the PDB codes of 138 DNA structures used in this work are listed in Table S1 in the Supporting information. Note that these DNAs are with no overlap with ssDNAs/dsDNAs used for model validation on 3D structure prediction. Since the known DNA structures are generally double helices, the initial parameters from these structures could not be reasonable to describe DNA chains during folding processes. In our previous RNA model, two sets of parameters ( $\text{Para}_{\text{helical}}$  and  $\text{Para}_{\text{nonhelical}}$ ) were calculated from stems and loops in experimental structures, respectively (65,66), and the  $\text{Para}_{\text{nonhelical}}$  ones were used to successfully describe the folding of an RNA from a free chain. However, due to the limitation of the number of

loop regions in known DNA structures, obtaining suitable parameters for DNA-free chains directly from these structures is unrealistic. Although we also did MD simulations for unstructured ssDNA (e.g., polyA and polyT chains) and tried to extract the bonded parameters from the conformations (data not shown here), because of some differences in optimum values of several angles between experimental and MD simulated structures, we gave them up. Eventually, based on the distributions of bond length/angle/dihedral for nonhelical parts in RNA structures are just slightly broader than that of helical parts (65), we simply set the strengths of DNA bonded potentials in  $\text{Para}_{\text{nonhelical}}$  as one-half of that in  $\text{Para}_{\text{helical}}$ . Note that, the  $\text{Para}_{\text{nonhelical}}$  is used in the folding process, and the  $\text{Para}_{\text{helical}}$  is only used for stems during folded structure refinement. Whereafter, the initial parameters were further optimized through the comparisons between the simulated and experimental bond length/angle distributions (34,74), and in this process, there are only two dsDNAs (PDB codes: 1agh, 3bse) and two ssDNAs (PDB codes: 1ac7, 1jve) were used.

For nonbonded potentials, the geometric parameters in base-pairing/stacking functions were obtained from the known structures; see Fig. S2 in the Supporting information. The strength of base-stacking was estimated from the combination of the experimental thermodynamics parameters and MC simulations; see Eqs. S7-S9 and Fig. S3 in the Supporting information. The strength of base-pairing (i.e.,  $\varepsilon_{bp}$  in Eq. S6) was determined by comparing the predicted melting temperatures ( $T_m$ 's) of four ss-/dsDNAs with corresponding experiments. That is, for two ssDNA hairpins (sequences: GCGCTTTTGCGC and GGAGCTTTGCTGC; ion condition: 1M NaCl; see Table 1) and two dsDNAs (sequences: GCTAGC/GCTAGC and GGGACC/GGTCCC; strand concentration: 0.1mM and 0.4mM, respectively; ion condition: 1M NaCl; see Table 2), we used the present model to predict their  $T_m$ 's through continuously adjusting the  $\varepsilon_{bp}$  until the agreement between predicted and experimental data is satisfactory.

The detailed descriptions, as well as the parameters of all the potentials in Eq. 1, can be found in the Supporting information.

## Simulation procedure

During DNA folding from sequence without any preset constraints, it is easy to fall into a metastable state with local minimum energy. To effectively avoid that, the MC simulated annealing algorithm, whose capacity has been proved in protein/RNA folding, was used to sample

conformations for ssDNA or dsDNA (65,75,76). For each DNA, a random chain configuration is generated from its sequence, and for dsDNA, the two chains are separately placed in a cubic box, the size of which is determined by the concentration of a single strand. Afterward, the simulation of a DNA system with a given monovalent/divalent ion condition is performed from a high temperature (e.g., 120°C) to the target temperature (e.g., room/body temperature). At each temperature, conformational changes are accomplished via the translation and pivot moves, which have been demonstrated to be rather efficient in sampling conformations of polymers (77,78), and the changes are accepted or rejected according to the standard Metropolis algorithm (65,69). The equilibrium conformations at different temperatures during the cooling process are used to analyze the stability of the DNA. In structure prediction, the last conformation at the target temperature is taken as the initial predicted structure, which can be further refined to better capture the geometry of helical parts by introducing the bonded parameters of *Para<sub>helical</sub>* for consecutive base-pairing regions. After structure refinement, an ensemble of structures would be obtained, and the mean RMSD (the averaged value over the whole structure ensemble) and minimal RMSD (corresponding to the structure closest to the native one) calculated over CG beads from the corresponding atoms in the native PDB structure is used to evaluate the reliability of the present model on DNA 3D structure prediction.

### Calculation of melting temperature

At each temperature, the fractions of folded (F, consistent secondary structures with predicted at lowest temperature) and unfolded (U, no more than one base pair) states could be fitted to a two-state model through the following equations (9,65):

$$f_F(T) = \frac{1}{1 + e^{(T-T_{m1})/dT_1}}; \quad (4)$$

$$f_U(T) = \frac{1}{1 + e^{(T-T_{m2})/dT_2}}; \quad (5)$$

where  $T_{m1}$  and  $T_{m2}$  are the two melting temperatures of the corresponding transitions (folded state to possible intermediate state (I) and intermediate state to unfolded state), respectively.  $dT_1$  and  $dT_2$  are the corresponding adjustable parameters. Based on the  $f_F(T)$  and  $f_U(T)$ , the fraction of the number of denatured base pairs  $f(T)$  can be calculated by (68)

$$f(T) = 1 - [(1 - f_I) \cdot f_F(T) + f_I \cdot (1 - f_U(T))]. \quad (6)$$

Here,  $f_I$  is the fraction of the number of denatured base pairs when the fraction for the I state is maximum. And then, the  $df/dT$  (the first derivative of  $f$  with respect to temperature) profile can be calculated to compare with the corresponding experimental data. It should be noted that for simple hairpins and short duplexes used in this work, the I state almost never occurs, and the  $f_I$  in Eq. 6 could be set to 0, which means that  $f_U(T)$  is approximately equal to  $1 - f_F(T)$  and only one  $T_m$  can be obtained.

To improve the simulation efficiency for dsDNA with low strand concentrations  $c_s$  (e.g.,  $<0.1\text{mM}$ ), the MC simulations were performed at a relatively high strand concentration  $c_s^h$  (e.g.,  $1\text{mM}$ ), and the fraction ( $f(T; c_s)$ ) of denatured base pairs at lower  $c_s$  can be calculated by that ( $f(T; c_s^h)$ ) at  $c_s^h$  (69,79,80)

$$f(T; c_s) = \frac{\lambda f(T; c_s^h)}{1 + (\lambda - 1)f(T; c_s^h)}, \quad (7)$$

where  $\lambda = c_s^h/c_s$ . Furthermore, for a dsDNA with a two-state transition, the melting temperature  $T_m(c_s)$  at  $c_s$  can be directly obtained from  $T_m(c_s^h)$  based on Eqs. 4-7,

$$T_m(c_s) = T_m(c_s^h) - dT_1 \ln \lambda, \quad (8)$$

the derivation of which can be found in the Supporting information.

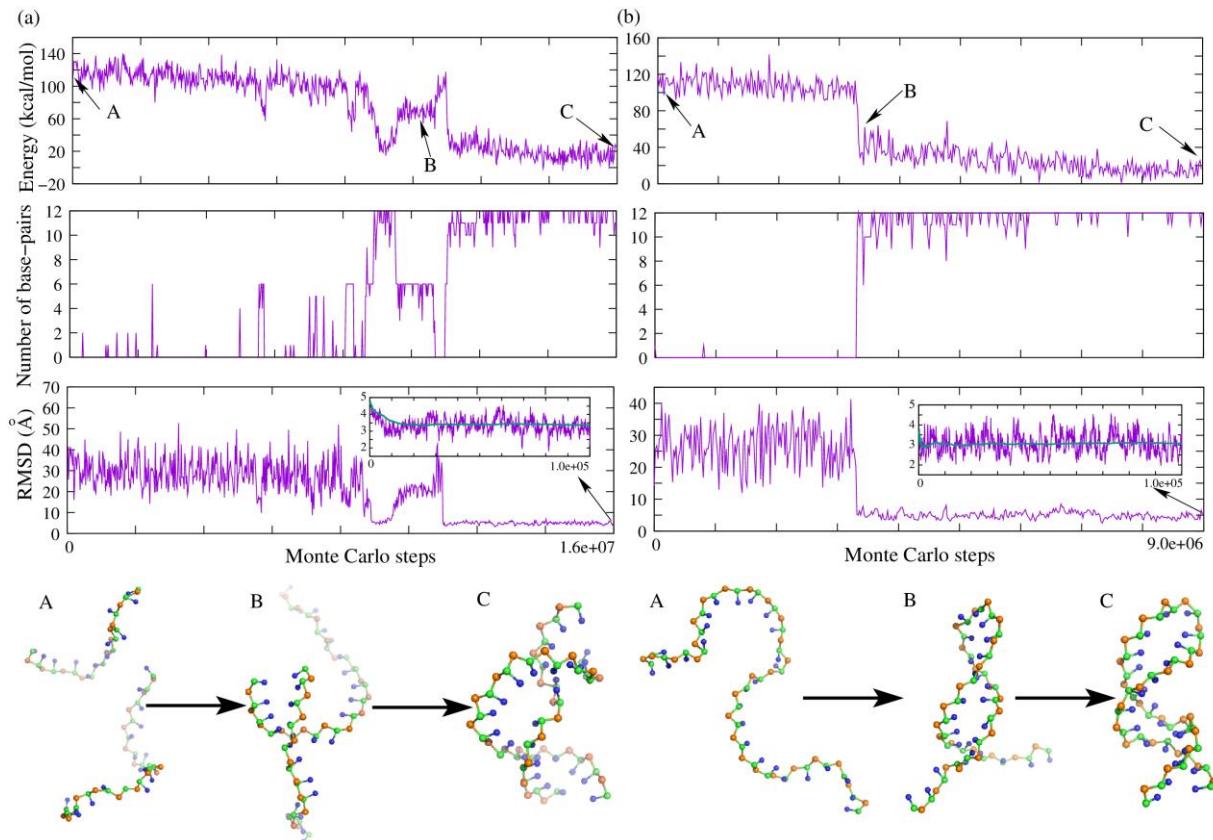
## Results

Based on the parameterized implicit-solvent/salt energy function and the MC simulated annealing algorithm, the present CG model can be used to predict 3D structures for dsDNA as well as ssDNA at different ion conditions and temperatures from the sequence. In this section, we tested the present model on the 3D structure and stability predictions for extensive DNAs with various lengths/sequences. As compared with the experimental structures and thermodynamics data, the present model can make overall reliable predictions.

### DNA 3D structure prediction from sequence

**For dsDNAs.** As described in the section of “Material and methods”, for each dsDNA, two random chains were generated from its sequence (e.g., structure A in Fig. 2a), which were further randomly placed in a cubic box, ensuring that there is no overlap. To guarantee no significant effect of the box on 3D structures, the strand concentration was set as  $1\text{mM}$  (i.e., box side length of  $149\text{\AA}$ ) for short

dsDNA (<10bp) and 0.1mM for longer ones. Due to the lack of the ion conditions for the experimental structures determined by X-ray crystallography, for simplicity, we only predicted the 3D structures for all DNAs at high ion concentrations (e.g., 1M NaCl), regardless of possible ion effects. As shown in Fig. 2a, for a dsDNA with a five-adenine bulge loop (PDB: 1qsk; 29nt, 12bp), the energy of the system reduces with the formation of base pairs as the temperature is gradually decreased from 120°C to 25°C, and the initial random chain folds into its native-like double-stranded structures (e.g., structure C in Fig. 2a). Following that, another MC simulation (e.g.,  $1 \times 10^5$  steps) is performed at target temperature based on the final structure predicted by the preceding annealing process, and the two sets of bonded potential parameters  $\text{Para}_{\text{nonhelical}}$  and  $\text{Para}_{\text{helical}}$  are employed respectively for the single-strands/loops and base-pairing regions to better capture the geometry of the helical part. As shown in the inset of the bottom panel of Fig. 2a, the mean and minimum RMSDs of the dsDNA between predicted structures and its native structure are  $\sim 3.2\text{\AA}$  and  $\sim 1.8\text{\AA}$ , respectively, and the corresponding predicted 3D structures are as shown in Fig. 3a.



**Figure 2. 3D structure prediction for the paradigm dsDNA/ssDNA in the present model.** (a, b) The time-evolution of system energy, number of base-pairs, RMSD from native structure, and typical 3D conformations

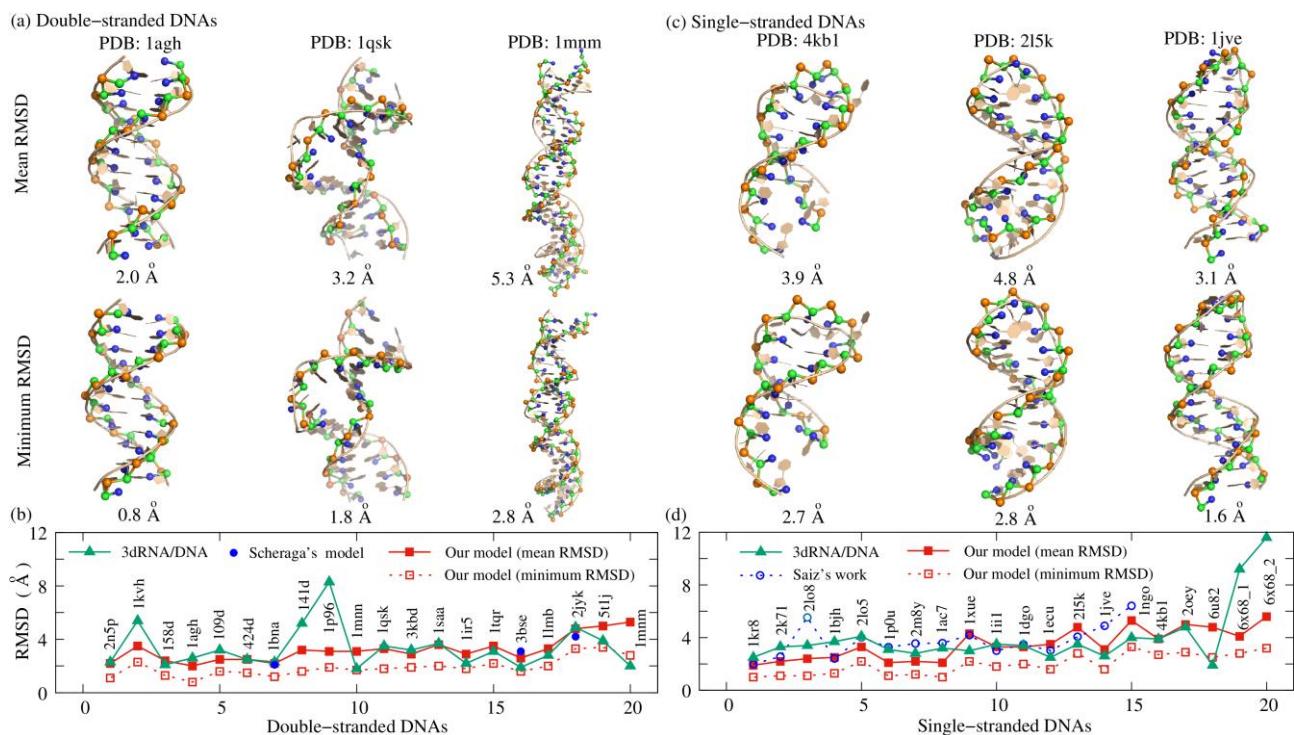
(from top to bottom, respectively) during the Monte Carlo simulated annealing simulation for (a) a dsDNA (PDB: 1qsk) and (b) an ssDNA (PDB: 1jve). The insets show the RMSDs of refined conformations calculated over all CG beads from the corresponding atoms in native structures. The 3D structures are shown with PyMol (<http://www.pymol.org>).

According to the above process, we employed the present model to predict the 3D structures of 20 dsDNAs (18nt-52nt) including helix with bulge loops, and the detailed descriptions (e.g., sequence, length, and structure feature) of these dsDNAs are listed in Table S3 in the Supporting information. For the 20 dsDNAs, the overall mean and minimum RMSD values are  $\sim 3.2$  Å and  $\sim 1.9$  Å, respectively; see Fig. 3a and Table S3 in the Supporting information, which suggest that the present model can make reliable predictions for 3D structures of dsDNA just from the sequence, despite a certain deviation (especially at the two ends) between the predicted and experimental structures for large dsDNA (e.g., PDB: 1mnm and 5t1j). Fig. 3a also shows the predicted 3D structures (ball-stick) with the mean and minimum RMSDs and the experimental structures (cartoon) for three typical dsDNAs with different lengths and sequences, intuitively indicating the ability of the model.

**For ssDNAs.** Compared with most of the existing models, the present model cannot only predict the double helix structure of dsDNA, but it can also make a prediction on the 3D structure for more flexible ssDNA. Similarly, a random chain generated from one ssDNA sequence can fold into native-like structures with temperature dropping; see Fig. 2b for an example of a DNA hairpin (PDB: 1jve; 27nt, 12bp), which could primarily benefit from the use of the soft parameters ( $\text{Para}_{\text{nonhelical}}$ ) of bonded potentials and sequence-dependent base-stacking interactions in the present model. As shown in Fig. 3b, for 20 ssDNAs (7nt-74nt) used in this work including hairpins with bulge/internal loops (Table S4 in the Supporting information), the overall mean (minimum) RMSD between the predicted and experimental structures is  $\sim 3.5$  Å ( $\sim 2.0$  Å), which strongly suggested that the present model can successfully predict 3D structures for simple ssDNA.

Since the structures of the largest hairpin (i.e., 6x68\_2) from the *piggyBac* DNA transposon (PDB: 6x68, a synaptic protein-DNA complex) has a significant bending possibly influenced by protein (81), our predictions without regard to protein has a certain deviation (mean RMSDs of 5.6 Å) from the experimental structure; see Fig. 3d. It is worth noting that beyond DNA hairpins, we also

tried to predict the 3D structure for a DNA three-way junction using the present model. As shown in Fig. S4 in the Supporting information, the structures (two hairpins at two ends) predicted from the sequence are pretty inconsistent with experimental ones. To find why, we further performed a MC simulation using the present model for the ssDNA starting from its PDB structure, and found that there is no significant difference in energies between predicted and simulated conformations (Fig. S4), which suggests that some tertiary interactions including the noncanonical base-pairing and base-backbone hydrogen bonding (82) and a more efficient algorithm (e.g., replica-exchanged MC) should be further taken into account in the model.



**Figure 3. The display of typical predicted 3D structures, and comparisons of RMSDs between the present model and other models.** (a, c) The predicted 3D structures (ball-stick) with the mean (top) and minimum (bottom) RMSDs for (a) three sample dsDNAs (PDBs: 1agh, 1qsk, and 1mmn) and (c) three ssDNAs (PDBs: 4kbl, 2l5k, and 1jve) from their native structures (cartoon). (b, d) The comparison of the predicted structures between the present model and the existing models including the 3dRNA/DNA and the models from Scheraga's or Saiz's group for (b) 20 dsDNAs and (d) 20 ssDNAs. The results of 3dRNA/DNA are predicted through their online server based on the native secondary structures (53). The other data is taken from refs. 55 and 58. The 3D structures are shown with PyMol (<http://www.pymol.org>).

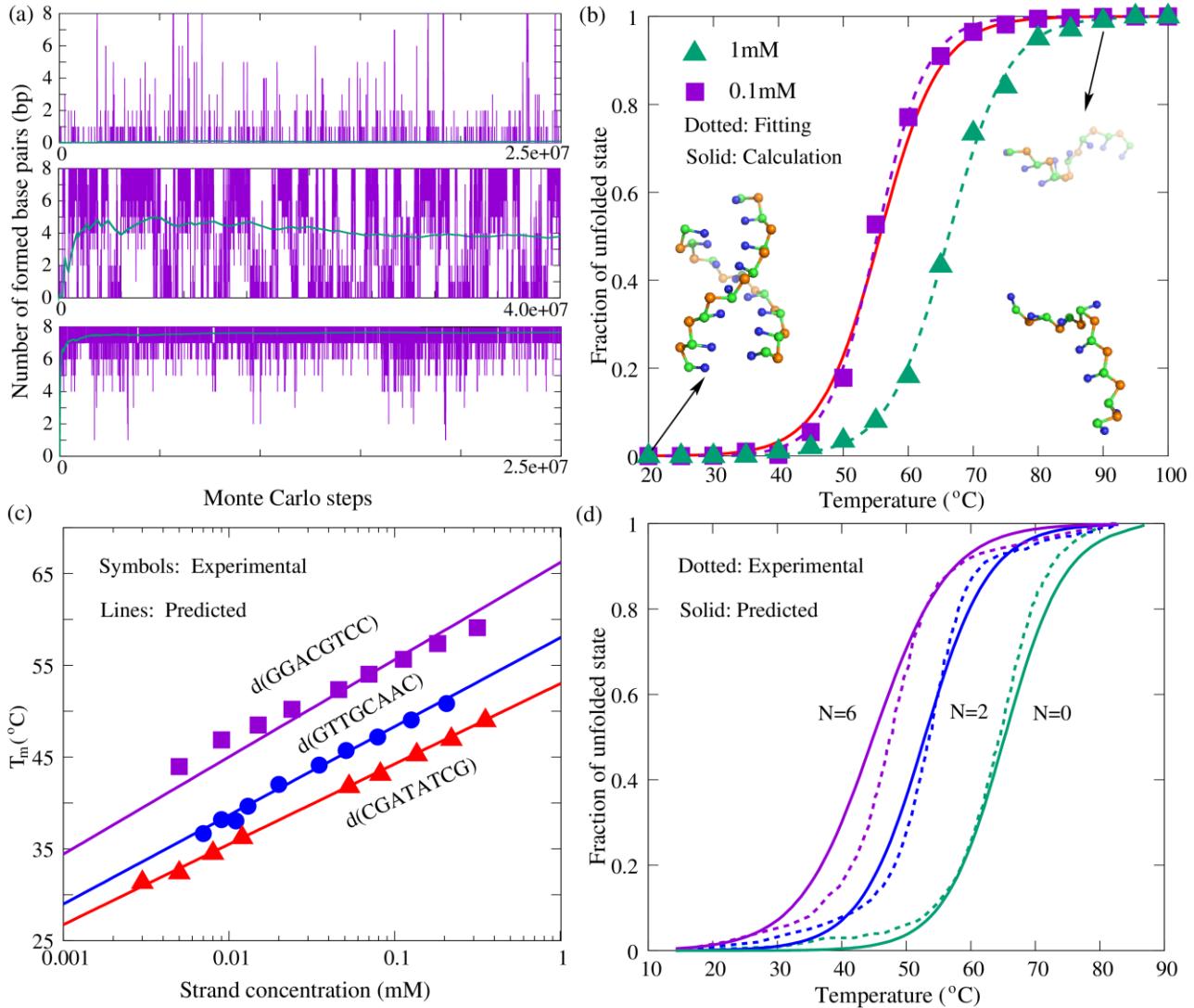
**Comparisons with other models.** To further examine the ability of the model on predicting 3D structures of DNAs (ssDNA and dsDNA), we also made comparisons with available results from the

existing models. First, we employed the 3dRNA/DNA web server (<http://biophy.hust.edu.cn/new/3dRNA/create>), which is an automatic, fast, and high-accuracy RNA and DNA tertiary structure prediction method (53,54), to predict 3D structures for all DNAs used in this work using the default options (e.g., Procedure: best; Loop Building Method: Bi-residue; # of Predictions: 5) and experimental secondary structures, and calculated the mean RMSD of returned conformations for each DNA over the atoms of P, C4' and N1/N9 from the corresponding atoms in the experimental structures. As shown in Fig. 3, for 20 dsDNAs, the overall mean RMSD ( $\sim 3.2 \text{ \AA}$ ) from the present model is not worse than that ( $\sim 3.3 \text{ \AA}$ ) from the 3dRNA/DNA, and for 20 ssDNAs, our prediction (overall mean RMSD:  $\sim 3.5 \text{ \AA}$ ) is slightly smaller than predicted result ( $\sim 4.0 \text{ \AA}$ ) from the 3dRNA/DNA.

Furthermore, we also made comparisons with the predictions from Refs. 55 and 58. Scheraga et al. also proposed a physics-based rigid-body CG model (3-bead) of DNA, and used it to successfully fold 3 dsDNAs (PDBs: 1bna, 3bse, and 2jyk) from complementary strands with only weak constraints between them (58). The all-bead RMSDs of the three lowest-energy predicted structures with respect to experimental references are  $2.1 \text{ \AA}$ ,  $3.1 \text{ \AA}$ , and  $4.2 \text{ \AA}$ , respectively, which are close to the mean RMSDs ( $2.2 \text{ \AA}$ ,  $2.6 \text{ \AA}$ , and  $4.8 \text{ \AA}$ , respectively) predicted from the present model (Fig. 3a). Jedd and Saiz presented a pipeline that integrates sequentially building ssDNA secondary structure from Mfold, constructing equivalent 3D ssRNA models by Assemble 2, transforming the 3D ssRNA models into ssDNA 3D structures, and refining the resulting ssDNA 3D structures through MD simulations (55). As shown in Fig. 3b, for 15 ssDNA hairpins, the average RMSD (over the sugar-phosphate backbone) for the best structures predicted by the pipeline is  $\sim 3.7 \text{ \AA}$ , a visibly larger value than the overall mean/minimum RMSD ( $\sim 3.2 \text{ \AA}$ / $\sim 2.2 \text{ \AA}$ ) from our predictions. Therefore, the comparisons with the other models fully show that the present model can successfully fold simple dsDNA/ssDNA from the sequence without the help of any secondary structure information.

## Stability of various DNAs

Beyond 3D structure predictions, the present model can also be used to predict the thermal stability for dsDNA and ssDNA in ion solutions. In order to verify the effect of the model, we further used it to predict the melt temperatures for extensive DNAs.



**Figure 4. The stability predictions for dsDNAs in the present model.** (a) The time-evolution of the number of base-pairs for a dsDNA (sequence:  $(GGACGTCC)_2$ ; strand concentration: 1mM) at different temperatures (90°C, 65°C, 40°C from top to bottom, respectively) in 1M NaCl solution. (b) The fractions of unfolded state  $f$  as functions of temperature for the dsDNA in (a). Green triangle: predictions at high strand concentration (1mM). Purple square: predictions at experimental strand concentration (0.1mM). Two dotted lines are the fitted melting curves to the corresponding predicted data. The solid line is calculated through Eq. 7. Ball-stick: the typical 3D structures predicted at low and high temperatures shown with PyMol (<http://www.pymol.org>). (c) The melting temperatures ( $T_m$ 's) as functions of strand concentration for three dsDNAs: purple,  $(GGACGTCC)_2$ , blue,  $(GTTGCAAC)_2$ , and red,  $(CGATATCG)_2$  at 1M  $[Na^+]$ . Symbols: experimental results (70). Lines: predictions from the present model. (d) The predicted (solid lines) and experimental (dotted lines) (86) melting curves for the dsDNA harboring symmetric internal loops with sequences of  $CTCGTC(T)_N CAGTGC/GCACTG(T)_N GACGAG$  in 1M NaCl solution. Green:  $N=0$ , i.e., the double helix without internal loop. Blue:  $N=2$ . Purple:  $N=6$ .

**Table 1.** The melting temperatures ( $T_m$ ) for dsDNAs at 1M  $[\text{Na}^+]$ .

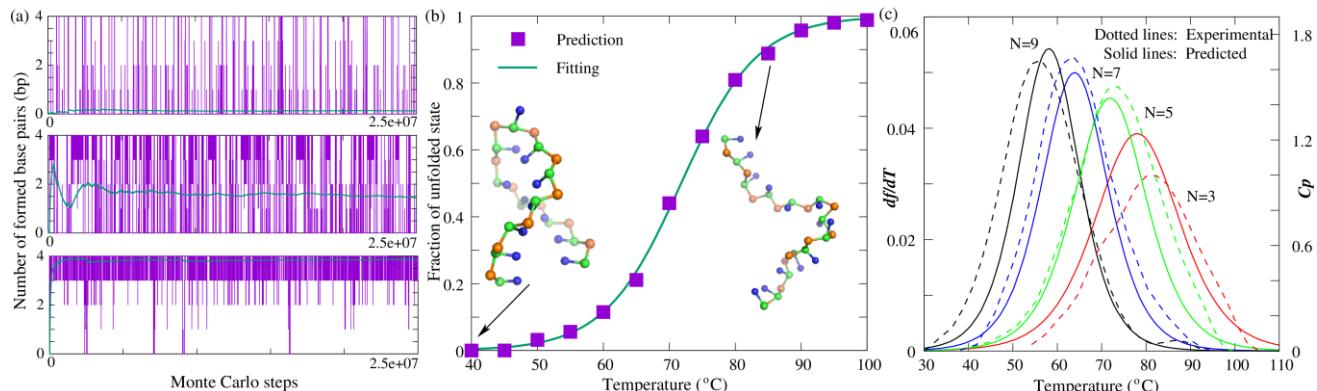
dsDNAs	Sequence <sup>a</sup>	Refs.	Length (nt)	$C_s$ (mM)	Expt. (°C)	Pred. (°C)	Deviation (°C)
DS1	(GCGC) <sub>2</sub>	70	8	0.1	27.5	26.6	-0.9
DS2	(CGCG) <sub>2</sub>	70	8	0.1	23.7	22.5	-1.2
DS3	(CGCGCG) <sub>2</sub>	70	12	0.1	55.7	54.0	-1.7
DS4	(GACGTC) <sub>2</sub>	70	12	0.1	36.1	36.3	0.2
DS5	(GGCGCC) <sub>2</sub>	70	12	0.1	53.9	53.0	-0.9
DS6	(CCGCGG) <sub>2</sub>	70	12	0.1	55.2	53.6	-1.6
DS7	(GCTAGC) <sub>2</sub>	70	12	0.1	34.3	36.6	2.3
DS8	(CGATATCG) <sub>2</sub>	70	16	0.1	44.1	47.0	2.9
DS9	(GGACGTCC) <sub>2</sub>	70	16	0.1	55.1	56.0	0.9
DS10	(CATATGGCCATATG) <sub>2</sub>	70	28	0.1	65.3	62.8	-2.5
DS11	GGGACC/GGTCCC	70	12	0.4	44.9	45.0	0.1
DS12	CAAAAAG/CTTTTG	70	14	0.4	31.5	33.0	1.5
DS13	CAAAAAAG/CTTTTTG	70	16	0.4	36.9	39.0	2.1
DS14	CAAACAAAG/CTTTGTTTG	70	18	0.4	47.3	49.1	1.8
DS15	CCATCGCTACC/GGTAGCGATGG	70	22	0.4	67.6	64.6	-3.0
DS16	GCGAAAAGCG/CGCTTTTCGC	70	20	0.4	65.2	66.4	1.2
DS17	(GCATGC) <sub>2</sub>	83	12	0.1	38.3	38.0	-0.3
DS18	GCGA <u>A</u> AGCG/CGCTTTTCGC	84	21	0.1	46.3	46.6	0.3
DS19	GCGAAAAGCG/CG <u>CTT</u> TCGC	84	21	0.1	44.6	46.4	1.8
DS22	(CG <u>ATT</u> ATCG) <sub>2</sub>	85	18	0.1	32.4	34.8	2.4
DS21	(GG <u>ACT</u> GTCC) <sub>2</sub>	85	18	0.1	44.5	46.2	1.7
DS22	CTCGTCCAGTGC/GCACTGGACGAG	86	24	0.01	66.0	65.2	-0.8
DS23	CTCGTCCAGTGC/GCA <u>CTG</u> ( <u>T</u> ) <sub>1</sub> GACGAG	86	25	0.01	59.2	62.0	2.8
DS24	CTCGTCCAGTGC/GCA <u>CTG</u> ( <u>T</u> ) <sub>6</sub> GACGAG	86	30	0.01	46.2	48.0	1.8
DS25	CTCGTC( <u>T</u> ) <sub>1</sub> CAGTGC/ GCA <u>CTG</u> ( <u>T</u> ) <sub>1</sub> GACGAG	86	26	0.01	59.7	57.9	-1.8
DS26	CTCGTC( <u>T</u> ) <sub>2</sub> CAGTGC/ GCA <u>CTG</u> ( <u>T</u> ) <sub>2</sub> GACGAG	86	28	0.01	54.9	56.6	1.7
DS27	CTCGTC( <u>T</u> ) <sub>6</sub> CAGTGC/ GCA <u>CTG</u> ( <u>T</u> ) <sub>6</sub> GACGAG	86	36	0.01	46.3	44.0	-2.3

<sup>a</sup>The subscript of “2” represents the self-complementary sequences, and for non-self-complementary sequences, the two strands are separated by ‘/’ and both start from 5’. The loop/unpaired nucleotides are underlined, and the “(T)<sub>N</sub>” means that the nucleotide T repeats N times.

**For dsDNA with various lengths/sequences.** The melting temperature ( $T_m$ ) of each dsDNA can be calculated by the present model based on 3D structures predicted at different temperatures; see Fig. 4a and the section “Material and methods”. For example, for the sequence (GGACGTCC)<sub>2</sub> at 1M  $[\text{Na}^+]$ , the melting curve of the dsDNA with a high strand concentration of 1 mM was predicted according to the fractions of unfolded state at different temperatures (Eqs. 5-7), and the melting curve, as well as the  $T_m$  of the dsDNA at low experimental strand concentration (0.1 mM), can be obtained through Eq. 8; see Figs. 4a and b. The predicted  $T_m$  of the sample sequence at  $c_s = 0.1\text{mM}$  is  $\sim 56.0^\circ\text{C}$ , which is only  $0.9^\circ\text{C}$  higher than the corresponding experimental value ( $\sim 55.1^\circ\text{C}$ ) (70). We further performed simulations for the dsDNA at  $c_s = 0.1\text{mM}$  to directly predict its  $T_m$  at experimental strand concentration, and found that there is no significant difference between two

melting curves/temperatures (Fig. 4b). In addition, as shown in Fig. 4c, the predicted  $T_m$ 's for three different dsDNAs at different strand concentrations are also in good accordance with the experiments (70), proving that it is feasible to infer the  $T_m$  at low  $c_s$  from the high ones ( $c_s^h$ ) (79,80).

To examine the sequence effect, 27 dsDNAs (8-36nt) with different sequences have been studied with the present model. The sequences, strand concentrations, and the predicted/experimental melting temperatures are listed in Table 1. Here, all dsDNAs are assumed at 1M  $[\text{Na}^+]$  to make comparisons with corresponding experimental data. As shown in Table 1, the  $T_m$  values of extensive dsDNAs from the present model are very close to the experimental measurements with a mean deviation of 1.5°C and maximal deviations  $< 3.0^\circ\text{C}$ , which indicates that the present model with the sequence-dependent base-stacking/pairing potential can make successful predictions on the stability for dsDNA of extensive sequences/lengths. Furthermore, due to the involvement of coaxial stacking potential, the present model can also provide reliable stability for dsDNA with bulge/internal loops. For example, for 4 dsDNAs with bulge loops and 5 dsDNAs with internal loops, the mean deviation of predicted  $T_m$ 's from the experiments is only 1.8°C; see Table 1, and the predicted melting curves for the dsDNAs with internal loops of different lengths are just slightly broader than the experiments (86) (Fig. 4d).



**Figure 5. The stability predictions for ssDNA hairpins in the present model.** (a) The time-evolution of the number of base-pairs for a simple hairpin (GCGC(T)<sub>5</sub>GCGC) at different temperatures (90°C, 70°C, 50°C from top to bottom, respectively) in 0.1M NaCl solution. (b) The fraction of unfolded state  $f$  as a function of temperature for the hairpin in (a). Symbols: predictions from the present model. Line: fitted melting curve to the predicted data through Eqs. 4-6. Ball-stick: the typical 3D structures predicted at low and high temperatures shown with PyMol (<http://www.pymol.org>). (c) The comparisons between predictions (solid lines) and experiments (dotted lines) for four DNA hairpins (GCGC(T)<sub>N</sub> GCGC) with different loop lengths at 0.1M  $[\text{Na}^+]$ . Red: N=3. Green: N=5. Blue: N=7. Black: N=9.  $df/dT$ : the first derivative of predicted  $f$  with the temperature.  $C_p$ : the heat capacity from experiment (87).

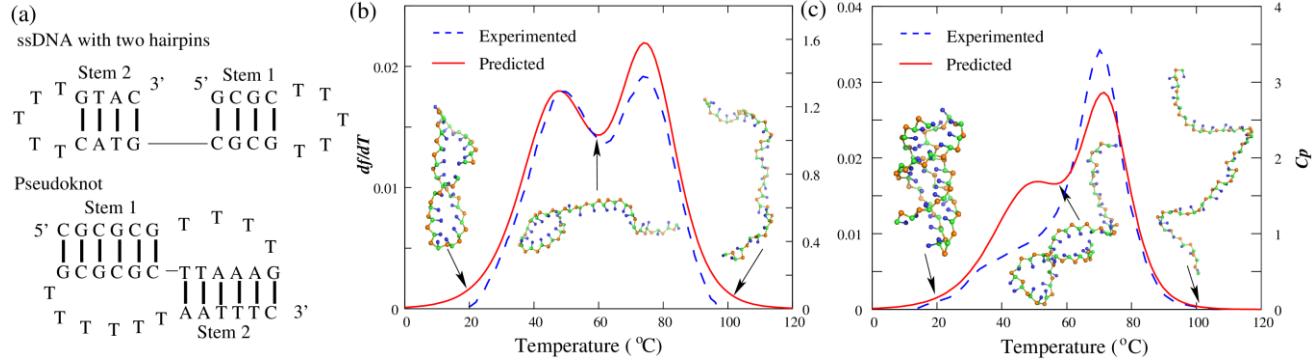
**Table 2.** The melting temperatures ( $T_m$ ) for single-stranded DNAs at given ion conditions

ssDNAs	Sequence <sup>a</sup>	Refs.	Length (nt)	[Na <sup>+</sup> ] (M)	Expt. (°C)	Pred. (°C)	Deviation (°C)
SS1	GCGC <u>(T)<sub>5</sub></u> GCGC	88	13	1.0	76.4	78.4	2.0
SS2	GCCGC <u>(T)<sub>5</sub></u> GCGGC	88	15	1.0	81.6	84.5	2.9
SS3	GCAGC <u>(T)<sub>5</sub></u> GCTGC	88	15	1.0	74.9	75.8	0.9
SS4	GCTGC <u>(T)<sub>5</sub></u> GCGC	88	15	1.0	58.0	59.0	1.0
SS5	GCTGC <u>(T)<sub>5</sub></u> GCTGC	88	15	1.0	56.9	55.1	-1.8
SS6	GCGC <u>(T)<sub>3</sub></u> GCGC	87	11	0.1	80.7	78.0	-2.7
SS7	GCGC <u>(T)<sub>5</sub></u> GCGC	87	13	0.1	73.3	71.9	-1.4
SS8	GCGC <u>(T)<sub>7</sub></u> GCGC	87	15	0.1	62.8	63.9	1.1
SS9	GCGC <u>(T)<sub>9</sub></u> GCGC	87	17	0.1	56.2	58.1	1.9
SS10	GTAC <u>(T)<sub>5</sub></u> GTAC	89	13	0.1	33.4	32.2	-1.2
SS11	ATCCTA <u>(T)<sub>5</sub></u> TAGGAT	90	17	0.2	51.0	54.0	3.0
SS12	ATCCTA <u>(T)<sub>6</sub></u> TAGGAT	90	18	0.2	48.1	51.6	3.5
SS13	ATCCTA <u>(T)<sub>7</sub></u> TAGGAT	90	19	0.2	43.9	47.1	3.2
SS14	GAATTC <u>(T)<sub>5</sub></u> GAATTC	91	17	0.2	55.3	53.1	-2.2
SS15	CCCAA <u>(T)<sub>12</sub></u> TTGGGG	92	22	0.25	57.9	59.2	1.1
SS16	CCCAA <u>(T)<sub>16</sub></u> TTGGGG	92	26	0.25	51.7	55.5	3.8
SS17	CGGATAA <u>(T)<sub>4</sub></u> TTATCCG	93	18	0.1	63.4	66.0	2.6
SS18	CGGATAA <u>(T)<sub>8</sub></u> TTATCCG	93	22	0.1	50.5	52.5	2.0
SS19	CGGATAA <u>(T)<sub>12</sub></u> TTATCCG	93	26	0.1	41.1	42.0	0.9
SS20	CGGATAA <u>(T)<sub>16</sub></u> TTATCCG	93	30	0.1	35.1	34.1	-1.0
SS21	CGGATAA <u>(T)<sub>20</sub></u> TTATCCG	93	34	0.1	29.6	26.8	-2.8
SS22	AAAAAAA <u>(C)<sub>5</sub></u> TTTTTTT	36	19	1.0	55.0	58.5	3.5
SS23	CGCGCG <u>(T)<sub>4</sub></u> GAAATTGCGCG <u>(T)<sub>6</sub></u> AATTTC	87	32	0.1	52.6/ 70.7	48.8/ 72.0	-3.8/ 1.3
SS24	GCGC <u>(T)<sub>5</sub></u> GCGCGTAC <u>(T)<sub>5</sub></u> GTAC	89	26	0.1	50.1/ 76.1	46.9/ 75.0	-2.1/ -1.1

<sup>a</sup>The sequences start from 5', the loop nucleotides are underlined, and the "(T)<sub>N</sub>" means that the nucleotide T repeats N times.

**For ssDNA with various lengths/sequences.** Beyond the dsDNA, the stability of ssDNA can also be captured by the present model. As shown in Fig. 5, for DNA hairpins (GCGC(T)<sub>N</sub>GCGC) with different loop lengths (N=3-9), the predicted thermal unfolding curves at 0.1M [Na<sup>+</sup>] agree reasonably with the experiments, despite that the predicted  $T_m$  (~78°C) for the hairpin with a small loop (e.g., N=3) is rather lower than the experimental value (~80.7°C), while it is a little higher (~58.1°C vs ~56.2°C) for large hairpin loops (e.g., N=9) (89). Moreover, 24 ssDNAs including pseudoknot are used to verify the ability of the present model for sequence effect on stability; see Table 2. In order to compare with experiments (36,87-93), all these predictions are at corresponding experimental ion conditions. As shown in Table 2, for 24 ssDNAs with different sequences and lengths (11-34nt), the mean/maximal deviation of  $T_m$  between predicted and experimental is ~2.1°C/~3.8°C, which suggests that the effect of sequence on ssDNA stability can also be well described by the present model. It is worth noting that due to the lack of stacking interactions

between unpaired bases, the present model cannot distinguish the stability of DNAs with same stems but different loop sequences (e.g., GCGC(T)<sub>5</sub>GCGC vs GCGC(A)<sub>5</sub>GCGC), and yet the stability of them generally differ somewhat (9,88,92).



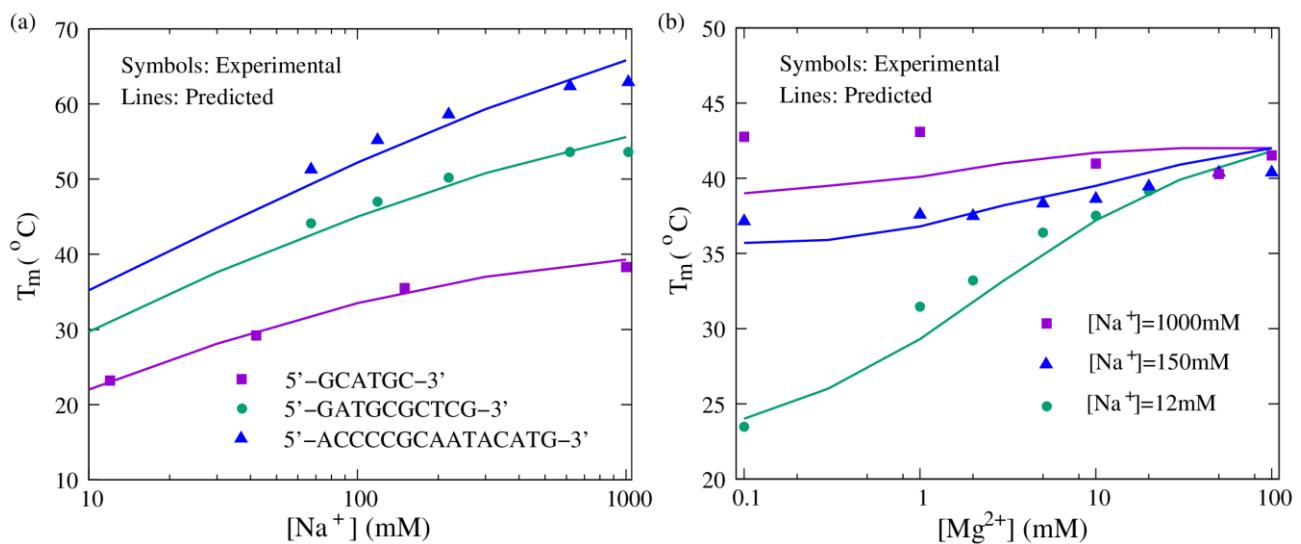
**Figure 6. The stability prediction for an ssDNA with two hairpins and a DNA pseudoknot in the present model.** (a) The schematics of secondary structure for the ssDNA with two hairpins (top) and the DNA pseudoknot (bottom). (b, c) The comparisons between predictions (solid lines) and experiments (dotted lines) for the two ssDNAs in (a).  $df/dT$ : the first derivative of predicted  $f$  with the temperature.  $C_p$ : the heat capacity from experiments (87,89). Ball-stick: the typical 3D structures predicted at different temperatures shown with PyMol (<http://www.pymol.org>).

Specifically, we made additional predictions for the stability of two more complex ssDNAs: a pseudoknot and a chain with two hairpins at two ends; see Fig. 6a. As shown in Table 2 and Fig. 6, for the ssDNA with two hairpins, two melting temperatures ( $T_{m1}$  and  $T_{m2}$ ) of the corresponding transitions are successfully predicted by the present model, with the deviations of  $\sim 2.1^\circ\text{C}$  and  $\sim 1.1^\circ\text{C}$  from experimental data, respectively. Since the hairpin at 3' end contains fewer G-C pairs than the other (Fig. 6a), it melts at a significantly low temperature in comparison to the 5' end hairpin (89). For the DNA pseudoknot at 0.1M  $[\text{Na}^+]$ , the predicted  $T_{m1}$  and  $T_{m2}$  are  $\sim 48.8^\circ\text{C}$  and  $\sim 72.0^\circ\text{C}$ , respectively, which also agree well with the experimental data ( $\sim 52.6^\circ\text{C}$  and  $\sim 70.7^\circ\text{C}$ ) (87); see Table 2, and the comparison between predicted and experimental thermal unfolding curves can be found in Fig. 6c. In the predicted curve, the first transition that is from folded pseudoknot state to intermediate hairpin state is more significant than that from experiment. One possible reason is that noncanonical interactions such as triple base interactions between loops and stems and self-stacking in loop nucleotides, which are common in RNA/DNA pseudoknots (68,87), are neglected by the present model, leading to a relatively simple unfolding energy surface. Even so, the comparison with the

experiment still suggests that the present model can be reliable in predicting thermal stability for DNA pseudoknots in monovalent ion solutions, and it is noted that the present model can also provide 3D structures for the pseudoknot at different temperatures from the sequences.

### Monovalent/divalent ion effects on stability of dsDNA/ssDNA

Due to the high density of negative charges on the backbone, DNA stability is sensitive to the ionic condition of the solution, while the effect of ions, especially divalent ions (e.g.,  $Mg^{2+}$ ), is generally ignored in the existing DNA CG models (43-50). Here, we employed the present model to examine the monovalent/divalent ion effects on the thermal stability of dsDNA and ssDNA.



**Figure 7. The comparisons of stability between predictions (lines) and experiments (symbols) for dsDNAs in monovalent/divalent ion solutions.** (a) The melting temperatures ( $T_m$ 's) as functions of  $[Na^+]$  for three dsDNAs with sequences of  $(GCATGC)_2$ ,  $(GATGCGCTCG)_2$ ,  $(ACCCCGCAATACATG)_2$  from bottom to top, respectively. Symbols: experimental data (84,94). Lines: predictions from the present model. (b) The melting temperatures ( $T_m$ 's) as functions of  $[Mg^{2+}]$  for the dsDNA with sequence of  $(GCATGC)_2$  at different  $[Na^+]$ 's: 0.012M, 0.15M, and 1M from bottom to top, respectively. Symbols: experimental data (84). Lines: predictions from the present model.

**Monovalent ion effect.** For each of the three dsDNAs with different lengths (6bp, 10bp, and 15bp), we performed simulations over a broad range of monovalent ion concentrations ( $[Na^+]$ : 0.01M-1.0M), and calculated the melting temperatures at different  $[Na^+]$ 's. As shown in Fig. 7a, the increase of  $[Na^+]$  enhances the dsDNA folding stability due to the stronger ion neutralization (61,62), and the predicted melting temperatures for the three dsDNAs are well in accordance with the experiment results (84,94), with the mean deviation of  $\sim 1.4^\circ C$ . Fig. 7a also shows that the  $[Na^+]$ -dependence of  $T_m$  is stronger for longer dsDNA, which could be caused by the larger buildup of negative charges

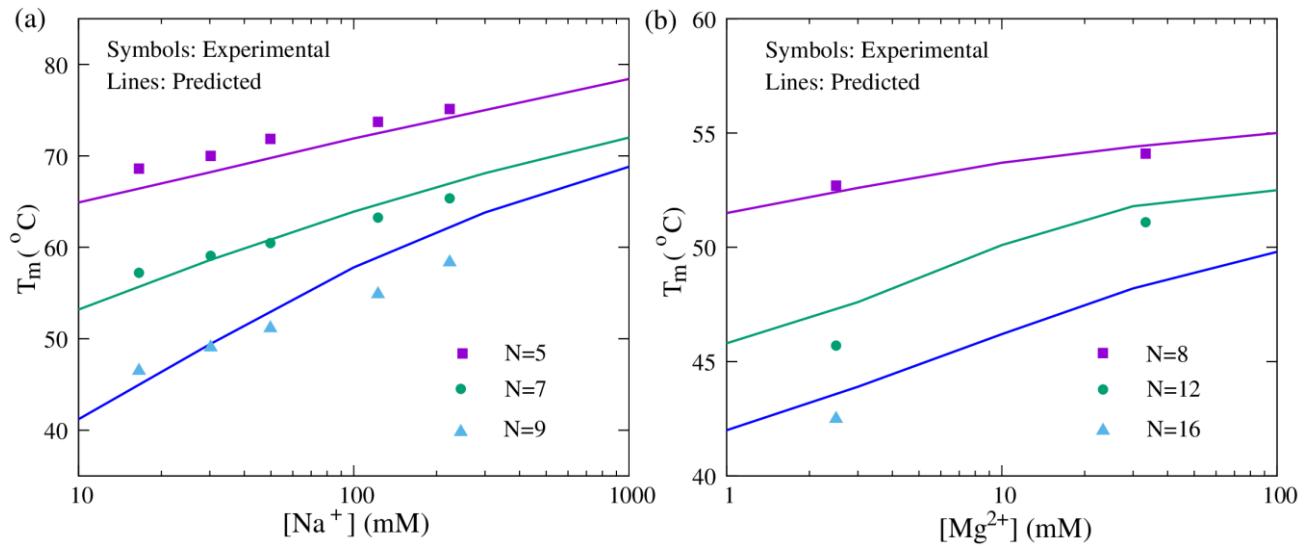
during base pair formation of longer dsDNA (62,73).

Although Table 2 has indicated that the present model can make reliable predictions for ssDNA stability at various  $[Na^+]$ 's, we further used a simple DNA hairpin (GCGC(T)<sub>N</sub>GC<sub>GC</sub>) with different loop lengths (N=5, 7, and 9) to test monovalent ion effect on stability in the present model. As shown in Fig. 8a, for the hairpin with small loops (e.g., N=5 and 7), the difference of predicted  $T_m$  from the experiments over a wide range of  $[Na^+]$ 's is very small (e.g., mean/maximal deviation of  $\sim 1.5^\circ C/\sim 1.0^\circ C$ ), and for the loop length of 9, our predictions are slightly larger than the experimental data only at high  $[Na^+]$ 's; e.g.,  $\sim 4.0^\circ C$  higher at  $\sim 0.1M$   $[Na^+]$  (87). The results on the stability of ssDNA and dsDNA in monovalent ion solutions reveal that it is a very effective way of involving the electrostatic interaction for DNA in the present model through the combination of the Debye-Huckel approximation and the concept of counterion condensation, which has also been validated by the TIS model (50,51).

**Divalent ion effect.** Remarkably, one important feature of the present model is that combining the counterion condensation theory and the results from the TBI model; see Eq. 3, it can also be used to simulate DNA folding in mixed monovalent/divalent ion solutions. For one dsDNA with the sequence of (GCATGC)<sub>2</sub> and one ssDNA hairpin with various lengths of the loop (CGGATAA(T)<sub>N</sub>TTATCCG), we made massive predictions in mixed  $Na^+/Mg^{2+}$  solutions and compared the melting temperatures with the corresponding experimental results (84,93); see Figs. 7b and 8b. The comparisons of  $T_m$ 's over a wide range of  $[Mg^{2+}]$  are in line with the experiments, whether for the dsDNA at different  $[Na^+]$ 's (i.e., 0.012M, 0.15M, and 1M) or for the ssDNA with different lengths of the loop, which suggests that the present model can nearly make quantitative predictions for the stability of DNAs in mixed ion solutions from their sequences, even though the ion effect is involved implicitly.

Furthermore, the competition between  $Na^+$  and  $Mg^{2+}$  on DNA stability can also be captured by the present model. For example, for dsDNA at 0.012M  $[Na^+]$  (Fig. 7b), when  $[Mg^{2+}]$  is very low (e.g.,  $<0.3mM$ ),  $Na^+$  dominates the stability of the dsDNA, while the increase of  $[Mg^{2+}]$  enhances the stability significantly. This is because the bindings of  $Na^+$  and  $Mg^{2+}$  are generally anti-cooperative and  $Mg^{2+}$ -binding is more efficient in stabilizing DNA structures (62,72). Naturally, as  $[Na^+]$  increases, the negative charge on DNA is strongly neutralized, and consequently, the effect of  $Mg^{2+}$

appears weak.



**Figure 8. The comparisons of stability between predictions (lines) and experiments (symbols) for ssDNAs in monovalent/divalent ion solutions.** (a) The melting temperatures ( $T_m$ 's) as functions of  $[Na^+]$  for the DNA hairpins (GCGC(T)<sub>N</sub>GCGC) with different loop lengths: 5, 7, and 9 from top to bottom, respectively. Symbols: experimental data (87). Lines: predictions from the present model. (b) The melting temperatures ( $T_m$ 's) as functions of  $[Mg^{2+}]$  for the DNA hairpins (CGGATAA(T)<sub>N</sub>TTATCCG) with different loop lengths: 8, 12, and 16 from top to bottom, respectively. Symbols: experimental data (93). Lines: predictions from the present model.

## Discussion

In this work, we have proposed a novel three-bead CG model to predict 3D structure and stability for both ssDNA and dsDNA in ion solutions only from the sequence. As compared with the extensive experiments, we have demonstrated that, (1) The present model can successfully predict the native-like 3D structures for ssDNAs and dsDNAs with an overall mean (minimum) RMSD of  $\sim 3.4\text{\AA}$  ( $\sim 1.9\text{\AA}$ ) from corresponding experimental structures, and the overall prediction accuracy of the present model is slightly higher than the existing models; (2) The present model can make reliable predictions on stability for dsDNAs with/without bulge loops and ssDNAs including pseudoknots, and for 51 DNAs with various lengths and sequences, the predicted melting temperatures are in good accordance with extensive experiment data (i.e., mean deviation of  $\sim 2.0^\circ\text{C}$ ); (3) The present model with implicit electrostatic potential can also reproduce the stability for ssDNAs/dsDNAs at extensive monovalent or mixed monovalent/divalent ion conditions, with the predicted melting temperatures consistent with the available experiments.

Nonetheless, the present model has several limitations that should be overcome in future model

development. For example, the present model failed to predict native-like structures for more complex DNAs such as that with three-way junction and cannot distinguish the stability for DNAs with different loop sequences, which suggest that possible noncanonical interactions (e.g., base triple interactions between loops and stems, self-stacking in loop nucleotides and special hydrogen bonds involving phosphates and sugars should be further taken into account (2,50,82). Furthermore, a more efficient sampling algorithm such as replica-exchanged MC or MC with umbrella sampling, as well as suitable structure constraints should be introduced to the model assembly for large DNAs (e.g., nano-architectures) (46-49,50), and accordingly, an accurate score function like statistical potential used for RNA and protein could be required to evaluate predicted DNA candidate structures (95-100). In addition, the 3D structure predicted by the present model is at the CG level, and it is still necessary to reconstruct all-atomistic structures based on the CG structures for further practical applications. After these further developments, a user-friendly web server would be further freely available, allowing users to predict 3D structure and stability for DNAs in ion solutions from sequence or given constraints.

## **Data Availability Statement**

All relevant computational data are within the paper and its Supporting information files. Experimental data are publicly available from published papers cited in the work. The 3dRNA/DNA web server used in this work is available by <http://biophy.hust.edu.cn/new/3dRNA/create>.

## **Supporting information**

**S1 Text.** The force field of the present model, the melting temperature calculation for dsDNAs at low strand concentrations, and the additional figures.

## **Acknowledgments**

We are grateful to Profs. Zhi-Jie Tan (Wuhan University), Wenbing Zhang (Wuhan University), and Yaoqi Zhou (Shenzhen Bay Laboratory) for valuable discussions, and we would like to acknowledge computing resources from the Shenzhen Bay Laboratory Supercomputing Center.

## **Author Contributions**

Conceptualization: Ya-Zhou Shi, Zi-Chun Mu, Ya-Lan Tan.

Data curation: Zi-Chun Mu, Ya-Lan Tan.

Formal analysis: Ya-Zhou Shi, Zi-Chun Mu.

Funding acquisition: Ya-Zhou Shi, Ben-Gong Zhang.

Investigation: Zi-Chun Mu, Ya-Zhou Shi, Jie Liu.

Methodology: Ya-Zhou Shi, Zi-Chun Mu.

Supervision: Ya-Zhou Shi.

Validation: Zi-Chun Mu, Ya-Zhou Shi, Ya-Lan Tan.

Writing - original draft: Zi-Chun Mu, Ya-Lan Tan, Ya-Zhou Shi.

Writing - review & editing: Zi-Chun Mu, Ya-Zhou Shi, Jie Liu.

## FUNDING

This work was supported by the Grants from the National Science Foundation of China (11971367 and 11605125).

## REFERENCES

1. Guiblet, W.M., Cremona, M.A., Harris, R.S., Chen, D., Eckert, K.A., Chiaromonte, F., Huang, Y.F. and Makova, K.D. (2021) Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Res.*, **49**, 1497-1516.
2. Ferry, G. (2019) The structure of DNA. *Nature*, **575**, 35-36.
3. Robinson, J., Raguseo, F., Nuccio, S.P., Liano, D. and Di Antonio, M. (2021) DNA G-quadruplex structures: more than simple roadblocks to transcription? *Nucleic Acids Res.*, **49**, 8419-8431.
4. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. and Balasubramanian, S. (2020) The regulation and functions of DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol*, **21**, 459-474.
5. Seeman, N., Sleiman, H. (2018) DNA nanotechnology. *Nat. Rev. Mater.*, **3**, 17068.
6. Hu, Q., Li, H., Wang, L., Gu, H. and Fan, C. (2019) DNA nanotechnology-enabled drug delivery systems. *Chem. Rev.*, **119**, 6459-6506.
7. Dey, S., Fan, C., Gothelf, K.V. *et al.* (2021) DNA origami. *Nat. Rev. Methods Primers*, **1**, 13.
8. Ma, W., Zhan, Y., Zhang, Y., Mao, C., Xie, X. and Lin, Y. (2021) The biological applications of DNA nanomaterials: current challenges and future directions. *Signal Transduct. Target. Ther.*, **6**, 351.
9. SantaLucia, J., Jr. and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415-440.
10. Bryant, Z., Oberstrass, F.C. and Basu, A. (2012) Recent developments in single-molecule DNA mechanics. *Curr. Opin. Struct. Biol.*, **22**, 304-312.
11. Kriegel, F., Ermann, N. and Lipfert, J. (2017) Probing the mechanical properties, conformational changes, and

interactions of nucleic acids with magnetic tweezers. *J. Struct. Biol.*, **197**, 26-36.

- 12. Cruz-Leon, S., Vanderlinde, W., Muller, P., Forster, T., Staudt, G., Lin, Y.Y., Lipfert, J. and Schwierz, N. (2022) Twisting DNA by salt. *Nucleic Acids Res.*, **50**, 5726-5738.
- 13. Fu, H., Zhang, C., Qiang, X.W., Yang, Y.J., Dai, L., Tan, Z.J. and Zhang, X.H. (2020) Opposite effects of high-valent cations on the elasticities of DNA and RNA duplexes revealed by magnetic tweezers. *Phys. Rev. Lett.*, **124**, 058101.
- 14. Zhang, Y., Zhou, H. and Ou-Yang, Z.C. (2001) Stretching single-stranded DNA: interplay of electrostatic, base-pairing, and base-pair stacking interactions. *Biophys. J.*, **81**, 1133-1143.
- 15. Di, W., Gao, X., Huang, W., Sun, Y., Lei, H., Liu, Y., Li, W., Li, Y., Wang, X., Qin, M. et al. (2019) Direct measurement of length scale dependence of the hydrophobic free energy of a single collapsed polymer nanosphere. *Phys. Rev. Lett.*, **122**, 047801.
- 16. Minhas, V., Sun, T., Mirzoev, A., Korolev, N., Lyubartsev, A.P. and Nordenskiold, L. (2020) Modeling DNA flexibility: comparison of force fields from atomistic to multiscale levels. *J. Phys. Chem. B*, **124**, 38-49.
- 17. Liebl, K. and Zacharias, M. (2021) Accurate modeling of DNA conformational flexibility by a multivariate Ising model. *Proc. Natl. Acad. Sci. U.S.A.*, **118**.
- 18. Jones, M.S., Ashwood, B., Tokmakoff, A. and Ferguson, A.L. (2021) Determining sequence-dependent DNA oligonucleotide hybridization and dehybridization mechanisms using coarse-grained molecular simulation, markov state models, and infrared spectroscopy. *J. Am. Chem. Soc.*, **143**, 17395-17411.
- 19. He, J., Wang, J., Tao, H., Xiao, Y. and Huang, S.Y. (2019) HNADOCK: a nucleic acid docking server for modeling RNA/DNA-RNA/DNA 3D complex structures. *Nucleic Acids Res.*, **47**, W35-W42.
- 20. Bao, L., Zhang, X., Shi, Y.Z., Wu, Y.Y. and Tan, Z.J. (2017) Understanding the relative flexibility of RNA and DNA duplexes: stretching and twist-stretch coupling. *Biophys. J.*, **112**, 1094-1104.
- 21. Wu, Y.Y., Bao, L., Zhang, X. and Tan, Z.J. (2015) Flexibility of short DNA helices with finite-length effect: From base pairs to tens of base pairs. *J. Chem. Phys.*, **142**, 125103.
- 22. Wang, Y., Gong, S., Wang, Z. and Zhang, W. (2016) The thermodynamics and kinetics of a nucleotide base pair. *J. Chem. Phys.*, **144**, 115101.
- 23. Liu, T., Yu, T., Zhang, S., Wang, Y. and Zhang, W. (2021) Thermodynamic and kinetic properties of a single base pair in A-DNA and B-DNA. *Phys. Rev. E*, **103**, 042409.
- 24. Kriegel, F., Matek, C., Drsata, T., Kulenkampff, K., Tschirpke, S., Zacharias, M., Lankas, F. and Lipfert, J. (2018) The temperature dependence of the helical twist of DNA. *Nucleic Acids Res.*, **46**, 7998-8009.
- 25. Zerze, G.H., Stillinger, F.H. and Debenedetti, P.G. (2021) Thermodynamics of DNA hybridization from atomistic simulations. *J. Phys. Chem. B*, **125**, 771-779.
- 26. Upadhyaya, A. and Kumar, S. (2021) Effect of loop sequence on unzipping of short DNA hairpins. *Phys. Rev. E*, **103**, 062411.
- 27. Nomidis, S.K., Kriegel, F., Vanderlinde, W., Lipfert, J. and Carlon, E. (2017) Twist-bend coupling and the torsional response of double-stranded DNA. *Phys. Rev. Lett.*, **118**, 217801.
- 28. Marko, J.F. and Siggia, E.D. (1994) Fluctuations and supercoiling of DNA. *Science*, **265**, 506-508.

29. Nomidis, S.K., Skoruppa, E., Carlon, E. and Marko, J.F. (2019) Twist-bend coupling and the statistical mechanics of the twistable wormlike-chain model of DNA: Perturbation theory and beyond. *Phys. Rev. E*, **99**, 032414.
30. Toan, N.M. and Thirumalai, D. (2012) On the origin of the unusual behavior in the stretching of single-stranded DNA. *J. Chem. Phys.*, **136**, 235103.
31. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406-3415.
32. Ingolfsson, H.I., Lopez, C.A., Uusitalo, J.J., de Jong, D.H., Gopal, S.M., Periole, X. and Marrink, S.J. (2014) The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **4**, 225-248.
33. Dans, P.D., Walther, J., Gomez, H. and Orozco, M. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29-45.
34. Sun, T., Minhas, V., Korolev, N., Mirzoev, A., Lyubartsev, A.P. and Nordenskiold, L. (2021) Bottom-up coarse-grained modeling of DNA. *Front. Mol. Biosci.*, **8**, 645527.
35. Reshetnikov, R.V., Stolyarova, A.V., Zalevsky, A.O., Panteleev, D.Y., Pavlova, G.V., Klinov, D.V., Golovin, A.V. and Protopopova, A.D. (2018) A coarse-grained model for DNA origami. *Nucleic Acids Res.*, **46**, 1102-1112.
36. Linak, M.C. and Dorfman, K.D. (2010) Analysis of a DNA simulation model through hairpin melting experiments. *J. Chem. Phys.*, **133**, 125101.
37. Walther, J., Dans, P.D., Balaceanu, A., Hospital, A., Bayarri, G. and Orozco, M. (2020) A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res.*, **48**, e29.
38. Maffeo, C., Ngo, T.T., Ha, T. and Aksimentiev, A. (2014) A coarse-grained model of unstructured single-stranded DNA derived from atomistic simulation and single-molecule experiment. *J. Chem. Theory Comput.*, **10**, 2891-2896.
39. Dans, P.D., Zeida, A., Machado, M.R. and Pantano, S. (2010) A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J. Chem. Theory Comput.*, **6**, 1711-1725.
40. Uusitalo, J.J., Ingolfsson, H.I., Akhshi, P., Tieleman, D.P. and Marrink, S.J. (2015) Martini coarse-grained force field: extension to DNA. *J. Chem. Theory Comput.*, **11**, 3932-3945.
41. Souza, P.C.T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grunewald, F., Patmanidis, I., Abdizadeh, H., Bruininks, B.M.H., Wassenaar, T.A. *et al.* (2021) Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat. Methods*, **18**, 382-388.
42. Assenza, S. and Perez, R. (2022) Accurate sequence-dependent coarse-grained model for conformational and elastic properties of double-stranded DNA. *J. Chem. Theory Comput.*, **18**, 3239-3256.
43. Hinckley, D.M., Freeman, G.S., Whitmer, J.K. and de Pablo, J.J. (2013) An experimentally-informed coarse-grained 3-Site-Per-Nucleotide model of DNA: structure, thermodynamics, and dynamics of hybridization. *J. Chem. Phys.*, **139**, 144903.
44. Knotts, T.A.t., Rathore, N., Schwartz, D.C. and de Pablo, J.J. (2007) A coarse grain model for DNA. *J. Chem. Phys.*, **126**, 084901.
45. Freeman, G.S., Hinckley, D.M., Lequieu, J.P., Whitmer, J.K. and de Pablo, J.J. (2014) Coarse-grained modeling of DNA curvature. *J. Chem. Phys.*, **141**, 165103.
46. Ouldridge, T.E., Louis, A.A. and Doye, J.P. (2011) Structural, mechanical, and thermodynamic properties of a

coarse-grained DNA model. *J. Chem. Phys.*, **134**, 085101.

- 47. Hong, F., Schreck, J.S. and Sulc, P. (2020) Understanding DNA interactions in crowded environments with a coarse-grained model. *Nucleic Acids Res.*, **48**, 10726-10738.
- 48. Poppleton, E., Romero, R., Mallya, A., Rovigatti, L. and Sulc, P. (2021) OxDNA.org: a public webserver for coarse-grained simulations of DNA and RNA nanostructures. *Nucleic Acids Res.*, **49**, W491-W498.
- 49. Ouldridge, T. E., Louis, A. A., Doye, J. P. (2010) DNA nanotweezers studied with a coarse-grained model of DNA. *Phys. Rev. Lett.*, **104**, 178101.
- 50. Chakraborty, D., Hori, N. and Thirumalai, D. (2018) Sequence-dependent three interaction site model for single- and double-stranded DNA. *J. Chem. Theory Comput.*, **14**, 3763-3779.
- 51. Denesyuk, N.A. and Thirumalai, D. (2013) Coarse-grained model for predicting RNA folding thermodynamics. *J. Phys. Chem. B*, **117**, 4901-4911.
- 52. Hyeon, C. and Thirumalai, D. (2011) Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.*, **2**, 487.
- 53. Zhang, Y., Wang, J. and Xiao, Y. (2022) 3dRNA: 3D Structure Prediction from Linear to Circular RNAs. *J. Mol. Biol.*, **434**, 167452.
- 54. Zhao, Y., Huang, Y., Gong, Z., Wang, Y., Man, J. and Xiao, Y. (2012) Automated and fast building of three-dimensional RNA structures. *Sci. Rep.*, **2**, 734.
- 55. Jeddi, I. and Saiz, L. (2017) Three-dimensional modeling of single stranded DNA hairpins for aptamer-based biosensors. *Sci. Rep.*, **7**, 1178.
- 56. He, Y., Maciejczyk, M., Oldziej, S., Scheraga, H.A. and Liwo, A. (2013) Mean-field interactions between nucleic-acid-base dipoles can drive the formation of a double helix. *Phys. Rev. Lett.*, **110**, 098101.
- 57. He, Y., Liwo, A. and Scheraga, H.A. (2015) Optimization of a nucleic acids united-RESidue 2-Point model (NARES-2P) with a maximum-likelihood approach. *J. Chem. Phys.*, **143**, 243111.
- 58. Maciejczyk, M., Spasic, A., Liwo, A. and Scheraga, H.A. (2014) DNA duplex formation with a coarse-grained model. *J. Chem. Theory Comput.*, **10**, 5020-5035.
- 59. Cragnolini, T., Derreumaux, P. and Pasquali, S. (2013) Coarse-grained simulations of RNA and DNA duplexes. *J. Phys. Chem. B*, **117**, 8047-8060.
- 60. Pasquali, S. and Derreumaux, P. (2010) HiRE-RNA: a high resolution coarse-grained energy model for RNA. *J. Phys. Chem. B*, **114**, 11957-11966.
- 61. Lipfert, J., Doniach, S., Das, R. and Herschlag, D. (2014) Understanding nucleic acid-ion interactions. *Annu. Rev. Biochem.*, **83**, 813-841.
- 62. Tan, Z.J. and Chen, S.J. (2006) Electrostatic free energy landscapes for nucleic acid helix assembly. *Nucleic Acids Res.*, **34**, 6629-6639.
- 63. Wu, Y.Y., Zhang, Z.L., Zhang, J.S., Zhu, X.L. and Tan, Z.J. (2015) Multivalent ion-mediated nucleic acid helix-helix interactions: RNA versus DNA. *Nucleic Acids Res.*, **43**, 6156-6165.
- 64. Zhang, C., Tian, F., Lu, Y., Yuan, B., Tan, Z.J., Zhang, X.H. and Dai, L. (2022) Twist-diameter coupling drives DNA twist changes with salt and temperature. *Sci. Adv.*, **8**, eabn1384.

65. Shi, Y.Z., Wang, F.H., Wu, Y.Y. and Tan, Z.J. (2014) A coarse-grained model with implicit salt for RNAs: predicting 3D structure, stability and salt effect. *J. Chem. Phys.*, **141**, 105102.
66. Shi, Y.Z., Jin, L., Wang, F.H., Zhu, X.L. and Tan, Z.J. (2015) Predicting 3D structure, flexibility, and stability of RNA hairpins in monovalent and divalent ion solutions. *Biophys. J.*, **109**, 2654-2665.
67. Jin, L., Tan, Y.L., Wu, Y., Wang, X., Shi, Y.Z. and Tan, Z.J. (2019) Structure folding of RNA kissing complexes in salt solutions: predicting 3D structure, stability, and folding pathway. *RNA*, **25**, 1532-1548.
68. Shi, Y.Z., Jin, L., Feng, C.J., Tan, Y.L. and Tan, Z.J. (2018) Predicting 3D structure and stability of RNA pseudoknots in monovalent and divalent ion solutions. *PLoS Comput. Biol.*, **14**, e1006222.
69. Jin, L., Shi, Y.Z., Feng, C.J., Tan, Y.L. and Tan, Z.J. (2018) Modeling structure, stability, and flexibility of double-stranded RNAs in salt solutions. *Biophys. J.*, **115**, 1403-1416.
70. SantaLucia, J., Jr., Allawi, H.T. and Seneviratne, P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555-3562.
71. Manning, G.S. (1978) The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Q. Rev. Biophys.*, **11**, 179-246.
72. Tan, Z.J. and Chen, S.J. (2005) Electrostatic correlations and fluctuations for ion binding to a finite length polyelectrolyte. *J. Chem. Phys.*, **122**, 44903.
73. Tan, Z., Zhang, W., Shi, Y. and Wang, F. (2015) RNA folding: structure prediction, folding kinetics and ion electrostatics. *Adv. Exp. Med. Biol.*, **827**, 143-183.
74. Leonarski, F., Trovato, F., Tozzini, V., Les, A. and Trylska, J. (2013) Evolutionary algorithm in the optimization of a coarse-grained force field. *J. Chem. Theory Comput.*, **9**, 4874-4889.
75. S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.
76. Li, J., Zhang, J., Wang, J., Li, W. and Wang, W. (2016) Structure prediction of RNA loops with a probabilistic approach. *PLoS Comput. Biol.*, **12**, e1005032.
77. Madras, N., Sokal, A.D. (1988) The pivot algorithm: A highly efficient monte carlo method for the self-avoiding walk. *J. Stat. Phys.*, **50**, 109–186.
78. Mavrantzas, V. G. (2021) Using Monte Carlo to simulate complex polymer systems: recent progress and outlook. *Front. Phys.*, **9**, 661367.
79. Ouldridge, T.E., Louis, A.A. and Doye, J.P. (2010) Extracting bulk properties of self-assembling systems from small simulations. *J. Phys. Condens. Matter*, **22**, 104102.
80. Privalov, P.L. and Crane-Robinson, C. (2018) Translational entropy and DNA duplex stability. *Biophys. J.*, **114**, 15-20.
81. Chen, Q., Luo, W., Veach, R.A., Hickman, A.B., Wilson, M.H. and Dyda, F. (2020) Structural basis of seamless excision and specific targeting by piggyBac transposase. *Nat. Commun.*, **11**, 3446.
82. Wu, B., Girard, F., van Buuren, B., Schleucher, J., Tessari, M. and Wijmenga, S. (2004) Global structure of a DNA three-way junction by solution NMR: towards prediction of 3H fold. *Nucleic Acids Res.*, **32**, 3228-3239.
83. LeBlanc, D.A. and Morden, K.M. (1991) Thermodynamic characterization of deoxyribooligonucleotide duplexes

containing bulges. *Biochemistry*, **30**, 4042-4047.

84. Williams, A.P., Longfellow, C.E., Freier, S.M., Kierzek, R. and Turner, D.H. (1989) Laser temperature-jump, spectroscopic, and thermodynamic study of salt effects on duplex formation by dGCATGC. *Biochemistry*, **28**, 4283-4291.
85. Peyret, N., Seneviratne, P.A., Allawi, H.T. and SantaLucia, J., Jr. (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry*, **38**, 3468-3477.
86. Tran, T. and Cannon, B. (2017) Differential effects of strand asymmetry on the energetics and structural flexibility of DNA internal loops. *Biochemistry*, **56**, 6448-6459.
87. Reiling, C., Khutsishvili, I., Huang, K. and Marky, L.A. (2015) Loop contributions to the folding thermodynamics of DNA straight hairpin loops and pseudoknots. *J. Phys. Chem. B*, **119**, 1939-1946.
88. Rentzepis, D., Shikiya, R., Maiti, S., Ho, J. and Marky, L.A. (2002) Folding of intramolecular DNA hairpin loops: enthalpy-entropy compensations and hydration contributions and hydration contributions. *J. Phys. Chem. B*, **106**, 9945-9950.
89. Rentzepis, D., Kharakoz, D. P., Marky, L. A. (1991) Coupling of sequential transitions in a DNA double hairpin: energetics, ion binding, and hydration. *Biochemistry*, **30**, 6276-6283.
90. Hilbers, C. W., Haasnoot, C. A., de Bruin, S. H., Joordens, J. J., van der Marel, G. A., van Boom, J. H. (1985) Hairpin formation in synthetic oligonucleotides. *Biochimie*, **67**, 685-695.
91. Germann, M. W., Kalisch, B. W., Lundberg, P., Vogel, H. J., & van de Sande, J. H. (1990) Perturbation of DNA hairpins containing the EcoRI recognition site by hairpin loops of varying size and composition: physical (NMR and UV) and enzymatic (EcoRI) studies. *Nucleic Acids Res.*, **18**, 1489-1498.
92. Goddard, N. L., Bonnet, G., Krichevsky, O., Libchaber, A. (2000) Sequence dependent rigidity of single stranded DNA. *Phys. Rev. Lett.*, **85**, 2400-2403.
93. Kuznetsov, S. V., Ren, C. C., Woodson, S. A., Ansari, A. (2008) Loop dependence of the stability and dynamics of nucleic acid hairpins. *Nucleic Acids Res.*, **36**, 1098-1112.
94. Owczarzy, R., You, Y., Moreira, B.G., Manthey, J.A., Huang, L., Behlke, M.A. and Walder, J.A. (2004) Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry*, **43**, 3537-54.
95. Stefaniak, F. and Bujnicki, J.M. (2021) AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses. *PLoS Comput. Biol.*, **17**, e1008309.
96. Wang, J., Zhao, Y., Zhu, C. and Xiao, Y. (2015) 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Res.*, **43**, e63.
97. Xiong, P., Wu, R., Zhan, J., and Zhou, Y. (2021) Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement. *Nat. Commun.*, **12**, 2777.
98. Zhao, H., Yang, Y., and Zhou, Y. (2010) Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. *Bioinformatics*, **26**, 1857-1863.
99. Tan, Y.L., Wang, X., Shi, Y.Z., Zhang, W. and Tan, Z.J. (2022) rsRNAsP: A residue-separation-based statistical potential for RNA 3D structure evaluation. *Biophys. J.*, **121**, 142-156.
100. Li, J., Zhu, W., Wang, J., Li, W., Gong, S., Zhang, J. and Wang, W. (2018) RNA3DCNN: Local and global quality

assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLoS Comput. Biol.*, **14**, e1006514.