JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Coarse-Grained HiRE-RNA Model for ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pairings

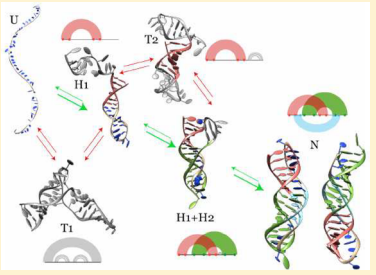
Tristan Cragnolini,[†] Yoann Laurin,[†] Philippe Derreumaux,[†,‡] and Samuela Pasquali*[,†]

[†]Laboratoire de Biochimie Théorique UPR 9080 CNRS, Université Paris Diderot, Sorbonne, Paris Cité, IBPC 13 rue Pierre et Marie Curie, 75005 Paris, France

[‡]Institut Universitaire de France, Boulevard Saint-Michel, 75005 Paris, France

Supporting Information

ABSTRACT: HiRE-RNA is a coarse-grained model for RNA structure prediction and the dynamical study of RNA folding. Using a reduced set of particles and detailed interactions accounting for base-pairing and stacking, we show that noncanonical and multiple base interactions are necessary to capture the full physical behavior of complex RNAs. In this paper, we give a full account of the model and present results on the folding, stability, and free energy surfaces of 16 systems with 12 to 76 nucleotides of increasingly complex architectures, ranging from monomers to dimers, using a total of 850 μs of simulation time.



1. INTRODUCTION

RNA molecules are essential cellular machines that perform a wide variety of functions. Aside from their well-known roles as genetic information carriers (mRNA) and amino acid recruiters (tRNA), they play many functions, notably in regulating gene expression through post-transcriptional processes (miRNA), gene silencing (RNAi), and catalytic activities (ribozymes).[1−3] Their sizes vary from a few dozen nucleotides for miRNAs and RNAi's, a hundred nucleotides for ribozymes, to several thousand nucleotides for rRNAs constituting the ribosome together with proteins. In all of their diversity, RNA molecules share the common feature of adopting specific three-dimensional (3D) structures to be functional[4] in the same way proteins adopt well-defined 3D organization for their biological activities, posing the question of RNA folding, that is, understanding how a linear RNA molecule adopts its characteristic 3D structure. RNA functionality depends crucially on their equilibrium structures and dynamical behavior[5,6] with distinct biologically active conformations under different conditions.[7]

With most genomic DNA identified as "noncoding" and possibly coding for RNA molecules, being able to determine RNA structures from sequences is essential for our understanding of the cellular machinery. However, obtaining high-resolution 3D structures through X-ray crystallography and NMR is a challenging task as shown by the small number of structures in the Nucleic Acid Data Bank (NDB) and by the scarcity of structures with substantially different architectures. Furthermore, low-resolution techniques, such as SAXS and Cryo-EM, require extensive modeling to propose a well-defined structure.

Computational methods have recently been developed to complement experimental studies for 3D structure predictions, and they follow different strategies. Bioinformatic algorithms based on sequence homology, fragment assembly, and secondary structure predictions[8−13] are successful for systems similar to those already present in structural databases. They provide a static view of the structure and, sometimes, partial thermodynamical information based on secondary structure predictions, but they are not suited for studyng the dynamical and global thermodynamical properties of RNA in three dimensions. These methods base much of their results on the prediction of a secondary structure first through more or less refined two dimensional (2D) prediction algorithms.[14−16] However, RNA structures are often intricate, giving rise to complex pseudoknots, triple or quadruple base pairings, noncanonical (non-Watson−Crick) pairings involving the base's sugar, and Hoogsteen edges,[17] which are not accounted for in secondary structure prediction methods developed to address nested structures (tree-like structures) or simple pseudoknots.

Physical models, considering the interactions of the system's particles in three-dimensions, are better suited to study RNA structures in all of their complexity. As opposed to secondary structure prediction methods, physical models do not have a specific term for pseudoknot formation. A pseudoknot results from the minimization of the free energy and is not encoded as a separate term in the potential energy: pseudoknots arise from a different organization of base pairings, but the interactions are

Received: March 2, 2015

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

the same that go into generating a hairpin or any nested structure.

All-atom simulations have successfully folded RNA of 12 nucleotides (nt)[18−20] but are limited to small systems even when using an implicit solvent representation.[21] Unfolding atomistic simulations have been performed on structures up to ∼40 nt.[22]

To overcome the limitations imposed by the size of the molecule, and follow the large scale rearrangements occurring in folding, one can resort to a simplification of the system through coarse graining. The challenge of this approach is to design a force field able to capture all of the subtle interactions that give rise to folding while maintaining a sufficiently simple description of the system for efficient simulation. In recent years, coarse-grained models have been proposed for nucleic acids with different levels of resolution and force field complexities. At a scale appropriate to describe the interactions leading to folding, Maciejczyk and Sheraga developed NARES-2P,[23] a 2-particle nucleic acid representation characterized by dipolar interactions, which are shown to be sufficient to drive the formation of double helices from unpaired single strands. Hyeon and Thirumalai developed the Go-like model TIS (Three Interaction Site)[24,25] to study the mechanical unfolding of hairpins and the stability of some pseudoknots, observing in particular the dependence of folding pathways and stability upon minor nucleotide sequence variations.[26] Dokholyan's group developed iFoldRNA, a 3-bead representation coupled to discrete molecular dynamics and used to study folding pathways and folding hierarchy. Plotkin's group developed a 3-particle model for DNA in which bases are treated as ellipsoids.[27] The model, shown to correctly predict persistence lengths of both single stranded and double stranded DNA, has been used to study temperature dependence of twisting and stacking of double helices. Doye's group recently developed a rigid model with 5 interaction sites for both DNA and RNA optimized on thermodynamic properties.[28] The model was shown to be suited for the study of the folding of a small pseudoknot, melting of a kissing complex, dynamics of double-helical nanoring, and hairpin unzipping under pulling of the extremities. Xia and co-workers developed a 5-particle RNA model capable of correctly folding structures of <30 nucleotides free of any experimental information, including hairpins, duplexes, and pseudoknots,[29] and of ∼120 nucleotides when coupled to experimental data. We refer to ref 30 for an extensive discussion of the challenges of developing a coarse-grained model suitable for RNA structure predictions and simulations.

These models are all able to investigate the formation of helical regions characterized by Watson−Crick pairings, but because they lack a detailed description of the realm of pairings occurring in RNA molecules, they have limited success for more intricate 3D structures. Any model aimed at predicting large-scale 3D rearrangements needs to give an accurate description of base-pairing and stacking, including the formation of noncanonical pairs and simultaneous pairings of three or four bases.[31] However, designing a force field properly accounting for these interactions is a challenging task as shown by the difficulties of otherwise very successful models. For atomistic simulations, Chen et al. had to reparameterize the AMBER force field to correctly model three tetraloops of 12 nucleotides.[18] The 3D ab initio FARFAR procedure, close in spirit to the ROSETTA approach used for proteins and based on sampling using a coarse-grained model followed by full-atom refinement, cannot reproduce any of the hydrogen bonds or stacking patterns within the UUGC tetraloop[32] and is of limited accuracy for RNAs of 12−20 nucleotides despite including noncanonical pairings. Interesting insights on folding mechanisms of structures, such as the 49 nt telomerase pseudoknot,[33] were obtained using Go-like potentials, allowing for the formation of native interactions only. These simulations were performed with a strong bias toward the already known experimental structure, and even though they provide important information on the overall folding process, they are not suited for the prediction of structures associated with a given sequence, nor for the study of the plurality of possible states that the molecule might explore during its lifetime.

Over the past several years, we have aimed to develop an RNA coarse-grained model with sufficient resolution to still capture the features of RNA molecules essential for their large-scale structural arrangements. HiRE-RNA[34,35] is an effective theory developed to fold any RNA architecture and study the structural dynamics of RNA molecules. Through the representation of 6 or 7 beads per nucleotide (1), the HiRE-RNA v3 force field, with specifically designed energy terms for stacking and base-pairing, including noncanonical and multiple pairs, points to the essential physics involved in folding. As with any modeling process, a choice is made as to which physical interactions are explicitly represented and which are not. Wanting to place the accent on the large scale rearrangements leading to folding from a completely open configuration, we explicitly represent the forces that contribute the most to motion on that scale, namely, electrostatics, base-pairing, and base-stacking. This leaves out terms that are present in nature but that are active on processes at different scales or phenomena. In particular, we do not explicitly represent water molecules even though their role is undoubtedly important for a number of enzymatic reactions and, in some cases, are responsible for specificities in the local structure. Similarly, we do not include at this stage base−sugar, base−phosphate, or sugar−sugar interactions, which are present in many RNA structures and are usually involved in local and non-large-scale structural arrangements. Detailed atomistic interactions can be recovered when converting back from the coarse-grained model to the full atomistic description. We also do not include explicit ions, and at this stage, they are indirectly

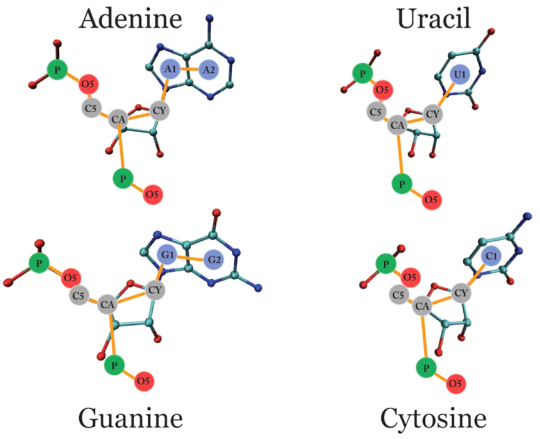


Figure 1. Atomistic and coarse-grained representation of the four bases, also showing the connection to the following base, represented in our model as a bond between particle CA of nucleotide *i* and particle P of nucleotide *i*+1.

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

accounted for by the electrostatic potential. Work is ongoing to develop a version of the model that better accounts for their presence.[30]

Similar to other coarse-grained models, two previous versions of HiRE-RNA, with a less sophisticated force field in the treatment of the bases, were useful for describing small helical stems and duplexes but failed to capture the physics of more complex molecules. The latest version of the model, HiRE-RNA v3, allows for folding of complex systems, such as multiple helices and pseudoknots, and provides realistic energy landscapes. In contrast to other structure prediction methods, such as iFoldRNA,[36] RNA2D3D,[37] and MC-Fold/MC-Sym,[9] which are limited to the study of one single RNA chain, HiRE-RNA can be used to study duplexes, quadruplexes, and any multiplexes.

We present here the completely redesigned force field of HiRE-RNA v3 and its results on 16 systems spanning from 12 to 76 nucleotides in length for a total simulation time of 856 μs, including several pseudoknots and systems with complex topologies. For these molecules, our model allows us to extract structural information and folding pathways in agreement with experimental data. Our work highlights the importance of an accurate description of base-pairing in the physical theory, including the plurality of possible base-pairs, in order to address all the fine structural details playing a key role in the folding process.

2. MODEL

The topology of the model remains the same as in previous versions (1).[34,35] Particle P, O5, C5, CA, and CY coincide with the phosphate, O5′, C5′, C4′, and C1′ atoms of a nucleotide, respectively. Bases are represented by beads positioned at the center of mass of each aromatic cycle. Purines are represented by two beads (A1 and A2 for adenine, G1 and G2 for guanine), whereas the pyrimidines are represented by one bead (C1 for cytosine, U1 for uracil). Despite the important reduction in the number of particles (70% less than an atomistic description without hydrogens), the occupied volume is preserved by an appropriate choice of the volume of the beads and the intrinsic rigidity of the molecule. In particular, that of the sugar ring not represented in our model is assured by the choice of the strength of the local interactions, and dihedral angles in particular, as has been shown in previous work, where calibration of these terms was carried out on a Poly-A chain first and then tested on some benchmark hairpins.[34]

HiRE-RNA v3 interaction potential is given by the sum of covalent bond interactions E_{lc}, excluded volume E_{ev}, electrostatics E_{el}, stacking E_{st}, and base-pairing E_{bp}.

As for the previous version, we take a top-down approach to build the force field. We define an effective theory in which the phenomenological interactions between particles in the system are given a priori with the aim of reproducing physical behaviors known from a variety of sources, including both experimental and theoretical. This approach allows us to capture behaviors that are consequences of atomistic interactions but also of quantum mechanical nature. This is the case in the treatment of base stacking, where the intrinsic nature of the interaction comes from the interplay of the orbitals of the aromatic rings and from dispersion forces[38,39] that elude a classical atomistic description.

In our model, we make a distinction between local interactions for covalently bound particles, for which we adopt a standard description typical of most atomistic force fields, and nonlocal interactions, which we further divide into short- and long-range. For long-range interactions, the correct distance dependence is crucial, as it determines the large-scale behavior of the molecules. We therefore adopt a realistic physical description, respecting the correct functional forms coming from physics theories. This is the case for the electrostatic potential, which we describe through a screened Coulomb potential. For short-ranged interactions, such as hydrogen-bonding and stacking, we have less information on the functional dependence of the potentials upon position because most data comes from QM calculations in the gas phase and does not directly apply to the conditions of the systems in solution. These potentials are therefore built in the spirit of reproducing features found in experimental structures, that is, that bases stack parallel to one another at short distances and with a vertical alignment, and that bases form pairs while on the same plane and at relatively well-defined angular orientations. We have chosen to adopt the analytically simple Gaussian form for all short-range potentials, where positions observed in experimental structures are narrowly distributed around a mean value, and angular functions that favor experimentally observed orientations. Given that these potentials act only locally, the precise dependence of the potential upon position is less crucial on the scale of the coarse-grained model than that of long-range potentials.

A reconstruction algorithm allows the program to go back to a full atomistic description at any time to recover details missed by the coarse-grained model, both structural and dynamical, be they the lack of some particle interactions (sugar−sugar, sugar−phosphate, base−phosphate) or the approximate distance dependence of short-range interactions. The reconstruction algorithm was presented in ref 35, where we have shown that atomistic simulations of double helical RNA and DNA were completely equivalent when started from the experimental configuration or from the configuration recovered by the coarse-grained Replica-Exchange Molecular Dynamics (REMD) simulations.

We present here the mathematical formulation of the force field with all of its details. All parameters of the model are given in the Supporting Information.

2.1. Local Interactions. The potential energy of the local interactions among particles in the molecule is the sum of three terms

$$E_{loc} = E_b + E_a + E_d$$

with E_b relates to the bond lengths between two bonded particles, E_a to the bond angles defined by three particles (two successive bonds), and E_d to the dihedral angles defined by quadruples of particles (three successive bonds).

The bond length potential between particles i and j is described by harmonic well

$$E_b(d;\, k_b,\, d_0) = \epsilon_b k_b (d - d_0)^2 \tag{1}$$

with $d \equiv |\vec{r}_{ij}| = |\vec{x}_i - \vec{x}_j|$, d_0 the equilibrium bond length, k_b the coupling constant specific for the pair in question, and ϵ_b a weight coefficient from the optimization.

Similarly, the potential for the bond angle θ, formed by particles i,j,k, is

$$E_a(\theta;\, k_a,\, \theta_0) = \epsilon_a k_a (\theta - \theta_0)^2 \tag{2}$$

where θ_0 is the equilibrium angle, k_a is the coupling, and ϵ_a is the optimized weight coefficient.

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

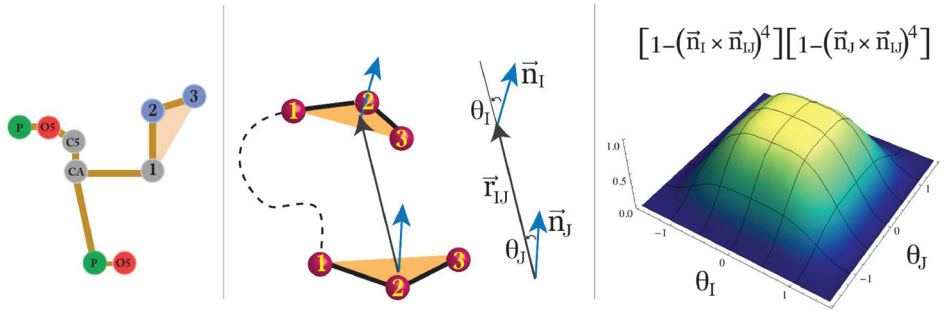


Figure 2. (left) Coarse-grained representation of a nucleotide highlighting the beads used to define the plane of the base. In the formulas in the text, these beads are generically named **1**, **2**, and **3**, and depending on the base type, correspond to CY-A1-A2, CY-G1,G2, CA-CY-C1, or CA-CY-U1. (center) Schematic representation of the stacking interaction through the definition of vectors normal to the planes. (right) View of the potential term regulating the dependence upon the vertical position of the two planes. It can be observed that the potential is flat at the top, allowing for a whole zone of possible positions centered around the perfectly vertically aligned planes, all with the same energetic contribution.

The torsion potential is given in terms of the dihedral angle φ as

$$E_d(\varphi; k_d, \varphi_0, m) = \epsilon_d k_d[1 + \cos(\varphi - \varphi_0)] \quad (3)$$

where k_d is the coupling constant, ϵ_d the weight coefficient, and φ_0 the equilibrium angle. The dihedral angle relative to four particles *i,j,k,l* is defined through the relation

$$\cos\varphi = \vec{n}_j \cdot \vec{n}_k \quad (4)$$

where $\vec{n}_j$ and $\vec{n}_k$ are the normals to the planes defined by particles *i,j,k* and *j,k,l*, respectively.

The form of the local potentials can give rise to both A- and B-form helices depending on the choice of the equilibrium parameters. This was shown in our previous study, where we addressed the assembly of both RNA and DNA double helices with RNA in the typical A-form helix and DNA in the typical B-form.[35] In the current study, where we analyze RNA molecules, A-form parameters were adopted (see Supporting Information for an exhaustive description of all parameters).

3. NONBONDED INTERACTIONS

The potentials contributing to the nonbonded interaction energy are the short-range excluded volume potential E_{ex}, long-range electrostatic potential E_{el}, stacking potential E_s, and base-pairing potential E_{bp}

$$E_{nb} = E_{ex} + E_{el} + E_s + E_{bp}$$

The excluded volume potential for nonbonded interactions is represented by an exponential repulsion

$$E_{ex}(d; d_0) = e^{-\kappa(d-d_0)} \quad (5)$$

with $d = |\vec{r}_{ij}| = |\vec{x}_i - \vec{x}_j|$, and d_0 a reference distance that also controls the coupling strength. The parameters are chosen in such a way that the potential increases rapidly when the two particles approach an interpenetration distance.

In the absence of explicit ions and water molecules, electrostatic repulsion of the phosphates and charge screening is represented by a Debye–Hückel potential between P particles, characterized by a screening length λ_D. The Debye–Hückel potential between charge q_i and charge q_j is given by

$$E_{DH}(d; q_i, q_j) = \epsilon_{el}\frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r}\frac{e^{-\frac{d}{\lambda_D}}}{d} \quad (6)$$

with $d = |\vec{r}_{ij}| = |\vec{r}_i - \vec{r}_j|$ being the distance between the charged particles.

Under physiological conditions in solution, the negatively charged nucleic acids are surrounded by positive ions. Ions have three different roles in their interplay with the highly charged nucleic acids. At long distance, they provide ionic screening for the charged phosphate groups. Close to the molecule, they create an ionic cloud that can at times provide an over screening and significantly alter the local electrostatic properties. The third kind are structural ions, occupying very specific locations in the architecture of some molecules. They can be considered to be bound to the molecule and part of the structure. Typically, removing such ions results in unfolding of the molecule in the region close to the ion and possibly of the whole molecule.

Ionic clouds and structural ions cannot be captured by our current description. Preliminary results of a version of HiRE-RNA, including explicit ions, was recently presented in ref 30 and is under extensive investigation. Results reported in this study involve molecules for which there are no structural ions in the experimental structures, and for which, therefore, the use of a model without explicit ions should be more legitimate. It remains to be noted that explicit modeling of ions in the presence of nucleic acids is also challenging for atomistic force fields,[40] and to our knowledge, there is not at this stage any coarse-grained model addressing the question of RNA folding that accounts for ions explicitly.

3.1. Base Interactions. RNA folding is driven by stacking and hydrogen bonding interactions.[41,42] These depend crucially on the relative position and orientation of the bases. To properly describe these interactions, we introduce the concept of base plane, identified through a vector $\vec{n}_i$ normal to the plane defined by the terminal beads of each base (Figure 2, left).

3.1.1. Stacking. Stacking occurs between any two bases in the system independently of their position along the chain, and it is calculated whenever the bases are in proximity. The interaction is modeled by the product of three terms, enforcing the relative positions of two stacked bases more widely observed in experimental structures. The first term defines the equilibrium distance r_{st}, chosen here to be 4.2 Å independent of the types of bases. The second term imposes that the planes of the bases assume a parallel orientation, whereas the third term restrains the potential to be active in only a limited portion of space where the two bases are vertically aligned, allowing some tolerance for lateral displacements (Figure 2, right). The stacking potential is defined as

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX–XXX

$$E_s(r_{IJ}, \vec{n}_I, \vec{n}_J) = -\epsilon_{st} e^{-\frac{(r_{IJ}-r_{st})^2}{\sigma_{st}}} (\vec{n}_I \cdot \vec{n}_J)^2 (1 - |\vec{n}_I \times \vec{n}_{IJ}|^4)(1 - |\vec{n}_J \times \vec{n}_{IJ}|^4)$$

where $\vec{r}_{IJ}$ is the distance between the center of mass of base I and J, $r_{IJ} = |\vec{r}_{IJ}|$, and $\vec{n}_I$ and $\vec{n}_J$ are the normals to the planes defined by base I and J, respectively

$$\vec{r}_{IJ} = \left| \sum_{k=1}^{3} \frac{1}{3}(\vec{r}_{I_k} - \vec{r}_{J_k}) \right|, \quad \vec{n}_I = \frac{\vec{n}_I}{|\vec{n}_I|}, \quad \vec{n}_J = \frac{\vec{n}_J}{|\vec{n}_J|}, \quad \vec{n}_{IJ} = \frac{\vec{r}_{IJ}}{|\vec{r}_{IJ}|} \tag{7}$$

with $\vec{N}_I = \vec{r}_{I_{12}} \times \vec{r}_{I_{32}}$, $\vec{r}_{I_{12}} = \vec{x}_{I(I)-2} - \vec{x}_{I(I)-1}$, and $\vec{r}_{I_{32}} = \vec{x}_{I(I)} - \vec{x}_{I(I)-1}$. Analogous definitions hold for base J. The parameters r_{st} and σ_{st} are the equilibrium distance and the Gaussian width, respectively.

The potential is fully symmetric in the two bases (I and J), reflecting decoupling between backbone configurations and base interactions that we try to respect in the model. Our stacking potential aims to describe the interactions between the aromatic rings of the bases independently of what happens to the phosphate-sugar chain. The final position of the nucleotides with their propensity to form a helix is the result of the combination of backbone local potentials and stacking of the aromatic rings. In our current version, we do not account for base specificity in stacking. To differentiate energetically between different pairs of stacked bases in our model, we need to define equilibrium distances and angular dependences for each possible pair of stacked bases, which requires a large analysis of existing stacking data to define all of the new parameters. The addition of this specification is ongoing.

As opposed to other coarse-grained models,[25,28,43] our stacking term is not temperature dependent. In principle, if the degrees of freedom relevant to the interaction are present in the model, there is no reason to add temperature dependence to the potential. (This is considered, for instance, in cases for atomistic potentials). Temperature dependence is necessary for coarse-grained models if the grains, because of their large size or of the simplified description of the particle interconnections, do not consider the source of the interaction explicitly enough. We believe our model, being high-resolution compared to all other models of similar approach, provides a sufficient description to adequately capture the degrees of freedom responsible for stacking. This assumption will tested in future work in which we plan to optimize the model against thermodynamic data.

3.1.2. Base Pairing. Base pairing occurs when two bases are side by side on the same plane and depends on the relative distance and angles of the particles forming the hydrogen bonds (Figure 3). We define the base-pairing potential, E_{bp}, as the product of E_{pl}, assuring planarity, and of E_{hb}, assuring the correct geometry of the interacting particles

$$E_{bp} = E_{pl} E_{hb} \tag{8}$$

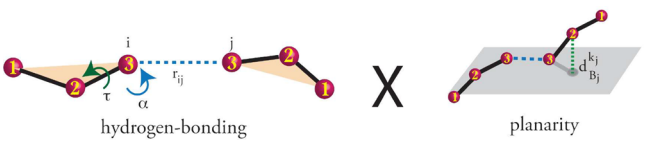


Figure 3. Schematic representation of base pairing composed of the product of a hydrogen-bonding potential and planarity.

The beads taking part in this interaction are the three terminal elements of the base, as indicated in Figure 2 (left).

Planarity is imposed by requiring that all particles of one base lie on the plane defined by the other base

$$E_{pl} = \epsilon_{pl} \left(\sum_{k_j=1}^{3} e^{-(d_{B_i}^{k_j}/\delta)^2} \right) \left(\sum_{k_i=1}^{3} e^{-(d_{B_j}^{k_i}/\delta)^2} \right) \tag{9}$$

where $d_{B_i}^{kj}$ is the distance of a particle k of base j with respect to the plane of base i, δ is the width of the interaction, and ϵ_{pl} is the adjustable strength parameter subject to optimization.

The hydrogen bond term, with coupling constant k_{hb} specific for each pair, reads

$$E_{hb}(\rho, \alpha_I, \alpha_J; k_{hb}, \rho_0) = -\epsilon_{hb} k_{hb} e^{\frac{-(\rho-\rho_0)^2}{\xi}} \in (\alpha_I) \in (\alpha_J) \tag{10}$$

It is a function of the inter-base distance $\rho = |\vec{\rho}|$ with $\vec{\rho} = \vec{x}_{I_3} - \vec{x}_{J_3}$ (or $\vec{\rho} = \vec{x}_{I(I)} - \vec{x}_{I(J)}$), the equilibrium distance ρ_0, and the angles α_I and α_J. The function $v(\alpha)$ is constructed such that only values of α close enough to the equilibrium angle α_0 contribute to the energy

$$\in(\alpha) = \begin{cases} \cos^6(\alpha - \alpha_0) & \cos(\alpha - \alpha_0) \geqslant 0 \\ 0 & \cos(\alpha - \alpha_0) < 0 \end{cases} \tag{11}$$

The α angles are defined through the relations

$$\cos \alpha_I = \vec{n}_{I\rho} \cdot \vec{n}_{I_{32}} \quad \text{and} \quad \cos \alpha_J = \vec{n}_{J\rho} \cdot \vec{n}_{J_{32}} \tag{12}$$

with

$$\vec{n}_{I\rho} = \frac{\vec{\rho}_I}{|\vec{\rho}_I|}, \quad \vec{\rho}_I = -\vec{\rho} + (\vec{\rho} \cdot \vec{n}_I)\vec{n}_I, \quad \vec{n}_\rho = \frac{\vec{\rho}}{|\vec{\rho}|}, \quad \vec{n}_{I_{32}} = \frac{\vec{r}_{I_{32}}}{|\vec{r}_{I_{32}}|} \tag{13}$$

and $\vec{r}_{I_{32}} = \vec{x}_{I_3} - \vec{x}_{I_2}$ (analogous definitions hold for base J). The torsional angle τ is used to discriminate between interaction minima at $+\alpha$ and $-\alpha$. To break the symmetry of the cosine function, which would give rise to interaction minima at both $+\alpha$ and $-\alpha$, we compute the dihedral angle τ between the particles defining the base plane of nucleotide i and the interacting bead of base j. If the cosine of the dihedral is negative, α is set to $-\alpha$

$$\alpha = \begin{cases} +\alpha & \text{if} \cos(\tau) > 0 \\ -\alpha & \text{otherwise} \end{cases} \tag{14}$$

In RNA complex architectures, it is typical to find noncanonical base pairs involving one or more of the three possible sides of the base: Watson−Crick, Hoogsteen, and Sugar. Bases can then form multiple simultaneous interactions, giving rise to triplets and quadruplets. For the wide variety of hydrogen bond patterns to be represented, HiRE-RNA v3 includes 22 different base-pairs occurring on all sides of the base (Figure 4), each associated with a specific set of distances, angles, and number of hydrogen bonds formed.[44]

The choice of 22 interactions is rather arbitrary and can be extended to any number of interactions as long as they are sufficiently distinct in their interaction centers. The pairs included in HiRE-RNA v3 have been chosen based on their abundance in the NDB, making sure to have at least two or three representatives for each letter pair. For some letter pairs,

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

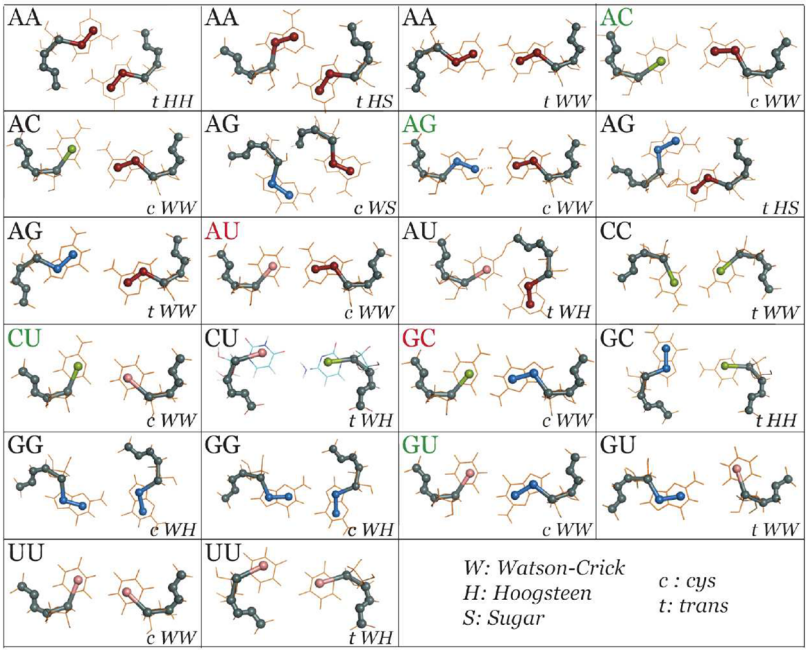


Figure 4. Set of 22 possible base-pairs in HiRE-RNA v3. In red are the canonical WC pairs considered by all CG models. In green are the pairs occurring on the WC sides of the bases also included in versions v1 and v2.

we can account for two distinct interaction sites occurring between the same sides at different geometric centers (see *cWW* A·C pairs). For any given letter pair, we consider all possible base-pairs from our list by summing over all possible E_{hb} terms. Because of the narrow distance dependence of E_{hb} and of the excluded volumes of the beads, there can effectively only be three interaction centers simultaneously present around a base, one on each side, thereby respecting the underlying atomistic condition that each side of the base can pair with at most one other base. The specification of base-paring used in the coarse-grained model are sufficiently detailed that, when going back to the atomistic description, hydrogen bonding acceptor and donor atoms are involved in at most one interaction at a time. In the coarse-grained model, in the absence of explicit donors and acceptors, the number of hydrogen bonds formed by each pairing is specified by the parameter k_{hb} ruling the strength of the interactions.

4. PARAMETRIZATION

Given our goal of accurate structural predictions, our model is parametrized based on structural information. For this reason, and in contrast to other models that approach the development of the force field on thermodynamical properties first,[25,28] we left a proper thermodynamical treatment for a later stage.

The model has geometric parameters whose values have been determined from distributions extracted from 200 NDB structures, including molecules of varying sizes, topologies, and overall energetic parameters, representing the relative weights of the different interaction terms that are subject to an optimization procedure. All values of the parameters are given in the Supporting Information. In this section, we describe the optimization procedure of the parameters ϵ, giving the relative strengths of the various potentials. Optimization of energy scales is done through a genetic algorithm to find the parameters that better distinguish native structures from decoys following the procedure already adopted for the optimization of the protein force field OPEP.[45] For a genetic algorithm to be used for optimization, the two sensitive issues to address are the algorithm itself and the choice of the training set.

4.1. Genetic Algorithm. A genetic algorithm is inspired by an evolutionary process. A "chromosome" is defined as a vector containing the parameters to be optimized. Multiple copies of the chromosome are generated at the beginning of the optimization cycle. Mutations and recombinations make the chromosomes evolve, modifying the vectors by random changes and permutations. The chromosomes are subjected to an evolutionary fit pressure to be able to distinguish between native structures and decoys in terms of the energy gap between experimentally existing structures and artificial, non-native structures. Only the best fitting chromosomes can reproduce, generating a new set of parameters, whereas the low-fitting chromosomes are discarded. The procedure is cycled until a desired accuracy is achieved in terms of a score, measuring the number of non-native structures having an energy greater than the corresponding native structure.

For the optimization, we used a set composed of 16 native structures of various topologies (as detailed in the next section) with 20 decoys for each for a total of 336 different structures. We initially generated 200 chromosomes that evolved in 5000 cycles of mutations and recombinations with a frequency of mutation of 20%. The algorithm stops if a chromosome is found that is stable for at least 250 cycles.

4.2. Training Set. The optimization with a genetic algorithm requires a robust database for training, with sufficient diversity to represent the plurality of possible structures, and balanced in different topologies to avoid introducing biases toward specific conformations. Because of the great abundance of double helical structures in the NDB over all other possible configurations, a simple selection from the NDB would result in a training set heavily biased toward double helices and in a resulting parameter set favoring the formation of double helical structures. This is not what we desire in a parameter set addressing the question of folding of a single stranded molecule, where the peculiarity of the structures is strongly

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

dependent on interactions, and departing significantly from the double helical conformation.

Given that we are interested in exploring the variety of possible structures adopted by a molecule corresponding to complete rearrangements in space and secondary structures, we built our training set based on the concept of a molecule's topology, with the schematic view of the molecule's secondary structure as a succession of stems and loops and their interconnections. This is a feature that is captured well by RNA dual graphs (Figure 5).[46,47] From the RAG (RNA-as-

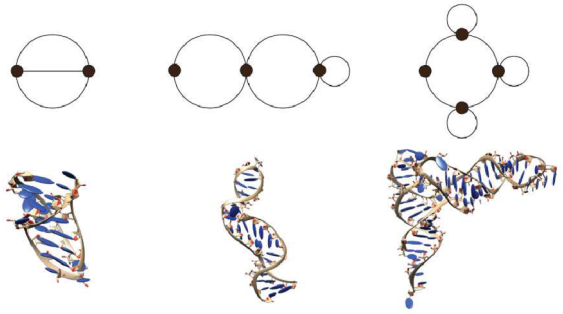


Figure 5. Dual graph representation of structures with different topologies: simple pseudoknot (left), helices and hairpin (center), and 4-way junction (right). The 3D structure of an example of each topology is given under the dual graph (PDB ID: 2A43, 1A9L, 1L9V).

graphs) database, we have selected 16 different topologies, and we have taken a representative of the smallest populated graphs (up to 6 nodes), corresponding to structures from 28 to 82 nucleotides, including nested structures (hairpin, loops, and bulges) and pseudoknots of various complexity.

To generate the decoys, we have run simulations using the very first version of HiRE-RNA[34] with two different parameter sets to generate decorrelated trajectories. Each simulation was run for 100 ns at a temperature of 300 K. For each native structure, we selected 20 decoys based on their RMSD value and the number of native base-pairs, that is, of base-pairs present in the native structure that are also present in the structure under consideration. To include in our training set the most diverse decoys, we have chosen the structures from four different scenarios (Figure 6): native-like structures (low RMSD and high percentage of native base-pairs), highly non-native structures (high RMSD and low percentage of native base-pairs), similar shape (low RMSD and low percentage of native base-pairs), and similar 2D structure (high RMSD and high percentage of native base-pairs).

4.3. New Parameters. Using the training set chosen as described above, we ran the genetic algorithm from an initial vector taken from version 2 of HiRE-RNA,[35] which was therefore already tested to give good results even if the parameters had been selected by hand. The new parameters were then tested in long stability MD simulations (from the native structure) on three molecules not included in the training set and compared with simulations ran with the initial, nonoptimized parameter set. The simulations with the optimized parameters show a more pronounced stability both in terms of RMSD, with average RMSD scores with respect to the native structures being lower by 3−4 Å and with fewer fluctuations, and in terms of preserved native base-pairs that exhibit greater stability.

Since this original validation, these parameters have been extensively employed in a multitude of simulations of various

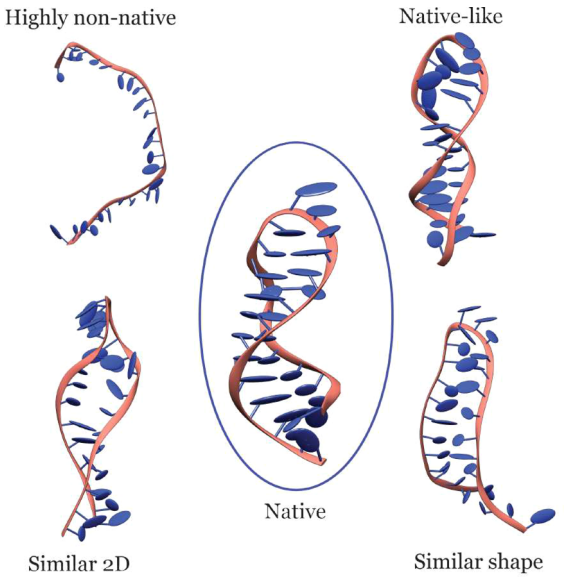


Figure 6. Example of the choice of decoys according to the four criteria.

systems (beyond that presented in this work) and have been shown to be general and robust for both molecular dynamics simulations of various kinds (MD, REMD, simulated tempering (ST)) and for Monte Carlo-based simulations (work in progress). The parameter set is sufficiently general to be able to address folding of a variety of different topologies with no specific adaptation from one class to another from hairpins to pseudoknots to multiple strand helices. Small variations in the parameter set do not significantly affect the simulation results as long as the general strength ratios between the various interactions are respected.

The parameters were not optimized for thermodynamics; however, REMD also gives access to thermodynamic information, allowing for the computation of specific heat curves and melting temperatures. With all three versions of the code, we have computed melting curves for a variety of systems and verified that the melting temperatures and shapes of the melting curves we obtained are at least in qualitative agreement with experimental data or with theories based on a thermodynamical description. In particular, we monitored as a reference the thermodynamical behavior of duplexes for which melting temperatures can be accurately computed from the sequence based on Turners parameters.[48,49] This validation a posteriori justifies the use of melting curves from our REMD to extract qualitative features of the thermodynamical behavior, such as the presence of more than one melting temperature, or the comparison of melting temperatures of simulated systems, even though it is not sufficient to allow quantitative comparisons with experimental data.

5. RESULTS

Table 1 reports the folding and stability results of 16 topologically different RNA molecules of 12−76 nts. For each system, we performed either molecular dynamics (MD) simulations at 300 K, REMD, 64 replicas with temperatures from 200 to 400 K and 500−1200 ns per replica, or simulated tempering (ST) with temperatures from 300 to 500 K, all with an integration time-step of 4 fs and a Langevin thermostat.[50,51] For analysis, we monitored the total number of stable base pairs formed during the simulations, the native base-pairs, and the

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 1. Summary of the systems studied, including PDB entry, topology (hp, hairpin; pk, pseudoknot; bl, bulge; c, complex; hx, helix; dp, duplex; G-quad, G-quadruplex), number of nucleotides (nb), simulation method, RMSD of the center of the most populated cluster at 300 K using REMD/ST or of the structure after several hundred nanoseconds, total number of base pairs (bp) in the predicted structure, number of base pairs present in the experimental structure also present in the predicted (native base pairs), and total simulation time[a]

PDB	topology	nb (nt)	method	RMSD (Å)	total bp (obs/nat)	time (μs)
1F9L	hp	22	ST	3.2	10 (9/9)	3
1L2X*	pk	27	REMD	5.8	11 (7/8)	134
1N8X	hp/bl	36	REMD	3.8	15 (11/15)	26
1RNG	hp	12	ST	2.7	5 (5/5)	3
1ZIH	hp	12	ST	1.9	5 (5/5)	3
1F7Y	hp	12	ST	2.0	5 (5/5)	3
1Y26*	c	71	REMD	8.1	32 (15/29)	96
2G1W	pk	22	REMD	4.3	9 (7/7)	153
2G1W	pk	22	ST	4.4	8 (7/7)	3
2K96	pk-3hx	47	REMD	4.3	23 (17/22)	120
480D	hp	27	REMD	5.5	12 (9/9)	180
405D	dp	2 × 16	REMD	3.6	16 (12/16)	64
433D	dp	2 × 14	REMD	4.0	14 (14/14)	64
3L0U*	tRNA	73	MD	~8	32 (23/30)	0.5
6TNA*	tRNA	76	MD	~8	31 (23/29)	0.5
1KF1*	G-quad	22	MD	~4	14 (12/12)	3

[a]The top simulations started from a fully extended structure, whereas the last three simulations give the results of MD stability starting from the NMR structures. Structures with a * are experimentally determined in conditions that cannot be fully include in our simulations. 1Y26 contains the adenine ligand. 3LOU, 1L2X, and 1KF1 contain structural ions. 6TNA contains modified bases.

root-mean-square deviation (RMSD) from the native structure using all particles.

With HiRE-RNA v3, we have been able to fold 13 structures from fully extended configurations within a few RMSD of the experimental structure and reproduce most of the native base-pairing network. To test the extent of the validity of our force field, we also performed long MD stability simulations on three systems for which complete folding from fully extended states is at the moment out of reach because of the presence of structural ions or of modified bases, which we cannot account for in our model at this stage. The goal of these simulation was to test that the new force field was able to correctly maintain the experimental structure over long time periods. We have analyzed G-quadruplexes, found in telomere repeats and of importance for cancer regulation, bound to several intercalating ions, and two tRNAs with either modified or unmodified bases and structural ions. Despite the absence of specific interactions for structural ions and modified bases, the three systems do not depart significantly from their NMR structures over several hundred nanoseconds. The quadruplex remains at 4.0 Å for over 3 μs and the two tRNAs of 73 and 76 nts remain at 8.0 Å with the native architecture preserved and most native base pairs formed (Figure 7).

To compare the accuracy of HiRE-RNA v3 to atomistic approaches, we first examined three tetraloops studied by Garcia's group in 2013 using extensive all atom REMD simulations in explicit solvent.[18] To achieve a high-accuracy folded structure, Garcia et al. had to reparametrize the AMBER force field. Using REMD and ST simulations, we fold the tetraloops with the same RMSD accuracy (1RNG: 3.1 Å Garcia vs 2.7 Å here. 1ZIH: 1.3 Å Garcia vs 1.9 Å here. 1F7Y: 0.8 Å Garcia vs 2.0 Å here) but with a much smaller computational cost (only 60 CPU hours per run for 3 μs of simulation time). Importantly for statistics, each ST simulation displays many folding/unfolding events. The importance of understanding tetraloop formation is reported in a recent article by Wales' group,[21] where the discrete path sampling method was used to determine the folding mechanisms and kinetics of three RNA tetraloops.

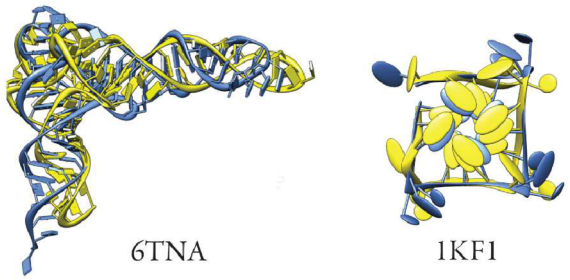


Figure 7. Superposition of the native structure (yellow) and a structure from the MD stability simulation (blue) after several hundred nanoseconds of simulation time.

5.1. Small Pseudoknots. Predicting small pseudoknots is particularly challenging because these molecules fold back on themselves forming tight bends and are stabilized extensively by stacking interactions as well as base-pairings. Whereas our v1 and v2 models lacked a detailed stacking term and could not predict such structures, HiRE-RNA v3 can fold the 22 nt 2G1W[52] and 28 nt 1L2X[53] pseudoknots with the experimental fold as the most stable structure at 300 K.

To study the impact of noncanonical base pairing on the system's behavior, for 2G1W, we also performed simulations considering only base pairs on the WC side of the base (red and green base-pairs in Figure 4). Although the molecule is still able to reach the native state, the absence of noncanonical pairs involving Hoogsteen and Sugar edges alters the energy landscape (Figure 8B).

With the full set of base-pairs, more states are populated as partial intermediates, creating a continuous path from the misfolded to folded state. Physically, this is an important difference as it allows the molecule to more easily interconvert

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

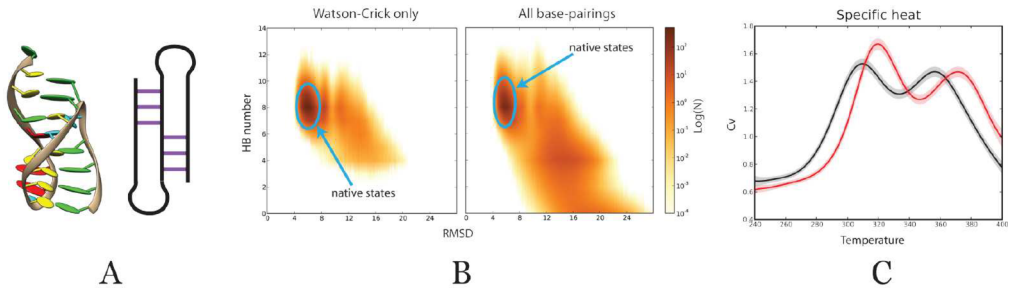


Figure 8. (A) Predicted 3D structure of 2G1W and 2D representation of the pseudoknot. (B) PMF (RMSD vs bp) at 300 K of 2G1W considering only WC base-pairs (left) and the full set of possible base pairs (right). The pseudoknot structure (blue insert) is the most populated state in both cases, but when considering the full set of base-pairs, a plurality of partially folded/misfolded states is also present. (C) Specific heat curves for 2G1W with WC-only base pairs (red) and with the full set of noncanonical pairs (black). The lowest peak corresponds to the transition to the native state, whereas the highest peak corresponds to the transition from a variety of partially folded states to the free chain. Both systems exhibit similar behavior, but melting temperatures are lower when noncanonical pairs are considered. Cv curves are computed using the MBAR algorithm;[54] shaded areas represent error bars.

between different states compared to when only WC base pairs are considered. This aspect is crucial for RNAs that are known to adopt alternative architectures at various stages of their biological activity. The importance of including noncanonical pairs also manifests in the qualitative thermodynamic behavior. From our REMD simulations, the estimated melting temperature of the model with all possible pairs is lower than that of the model including only WC base pairs (Figure 8 C). This is indeed expected as the presence of noncanonical pairs opens the possibility of new folding pathways, which is absent when only canonical base-pairs are considered, and renders the ground state more entropically accessible, lowering the melting temperature. By shifting the temperature of one of the two curves to superpose the melting peaks, we can observe that the low and high temperature behavior of the two systems is the same, but that the lowest peak for the WC-only system is narrower and taller than the corresponding peak for the full base-pairing set, indicating a stiffer transition when noncanonical pairings are turned off. The computed average energies at all temperatures are the same within error bars, whereas the fluctuation behavior of the two systems is different with the system including noncanonical pairs subject to larger fluctuations than the system with WC-only interactions.

The presence of two peaks in the specific heat curve is in good agreement with the experimental observations on the MMTV pseudoknot,[55] a 32-nt RNA with the same topology as 2G1W, and seems to be a common feature of pseudoknot folding, where melting of two separate stems is involved. The specific heat curve of 1L2X also displays two peaks (Figure 9). 1L2X is a pseudoknot composed of a longer stem (8 base pairs) and a shorter stem (3 base pairs). The lower-temperature peak at 330 K (T_{m_1}) corresponds to melting of the shorter stem and the higher-temperature peak at 370 K (T_{m_2}) corresponds to complete unfolding. As expected, T_{m_1} is higher than the corresponding melting temperature of 2G1W (320 K) given the higher number of base pairs breaking in the melting transition in 1L2X over 2G1W. The calculated difference between T_{m_1} and T_{m_2} at 40 K is consistent with experimental observations for similar pseudoknots for which the difference of $T_{m_2} - T_{m_1}$, measured by UV absorbance, was reported to be 20 K, 24 K and 35 K for the MMTV pseudoknot, wild-type T4 pseudoknot, and C8U BWYV pseudoknot, respectively[55,56].

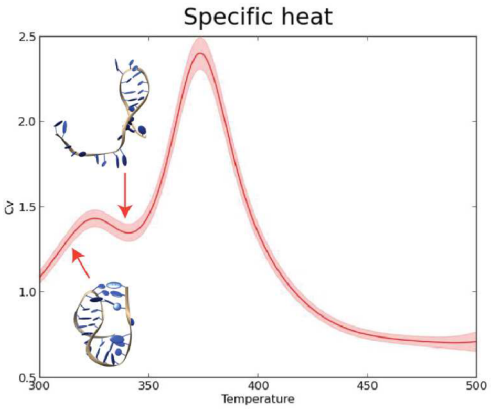


Figure 9. Specific heat curve of 1L2X exhibiting two peaks corresponding to melting of the two stems composing the pseudoknot. The shorter stem melts at a temperature that is lower than that for the longer stem.

5.2. Triple Helix Pseudoknot. We then studied larger and more complex systems. Starting from fully extended conformations, REMD simulations were able to fold the triple helix of the pseudoknot of the human telomerase (PDB ID: 2K96) of 47 nt,[57] for which the native structure is characterized by a 6 base-pair WC helix and an A-rich dangling strand inserting into the WC helix groove and forming several stacked triplets. Overall, 22 base pairs, canonical and noncanonical, stabilize the native structure. In the simulation, we can distinguish a short phase to form the WC helix and a longer phase in which the other contacts form, generating the full triple helix (Figure 10). After 1.2 μs of REMD time, a structure was reached with an RMSD of 4.3 Å and stabilized by 17 native base pairs (Figure 11, N). To our knowledge, this is the first time anyone has folded an RNA of such complexity solely from the sequence, an achievement possible only if the relevant physics is correctly taken into account by the model.

Although individual ST and REMD individual trajectories rapidly swap temperatures, and the observed pathways may not be identical to the pathways observed at constant temperature, significant insights can be obtained from the many unfolding/folding events we observe in enhanced sampling simulations. We have therefore further analyzed the folding path of the triple helix 2K96 (Figure 11). In the NDB structure of 2K96, the paired regions are the following: 1−6 paired with 24−29 (H1), 15−23 paired with 37−47 (H2), and 6−10 paired with

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

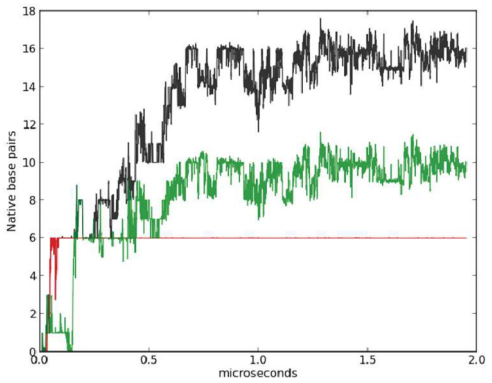


Figure 10. Formation of native contacts for triple helix folding. Contacts of the WC helix (red) form first whereas contacts stabilizing the triple helix (green) form on a much longer time scale. The full set of native contacts, including those of the triple helix, is shown in black.

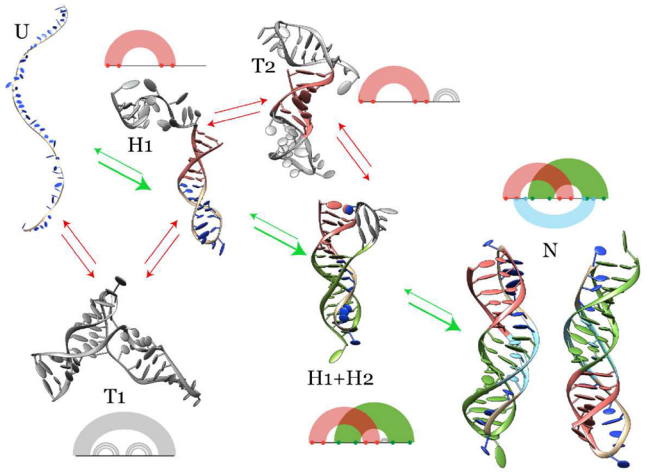


Figure 11. Folding pathway of the triple helix 2K96 extracted from REMD simulations. Green arrows indicate the reversible transitions between intermediate folds, leading to the native structure (N), red arrows indicate transitions leading to misfolded states (T1 and T2). Next to each structure, we give the schematic representation of base pairing with arc diagrams color coded according to the formation of H1 (red), H2 (green), and H3 (blue). Structures in gray correspond to non-native secondary elements.

36−40 (H3). Notice that there is an overlap of two paired regions with bases 37, 38, 39, and 40 forming triple contacts. In REMD simulations, we observe three separate folding steps corresponding to the successive formation of each one of the stems (H1-red first, H2-green second, and H3-light blue third), with different time scales. The system can remain trapped in misfolded states with structures exhibiting base-pairings that differ from the native state (T1 and T2). Our folding path (H1 → H1 + H2 → N) is in agreement with experimental studies on 2K96.[58] On the basis of the UV melting curves, it was proposed that the three melting transitions of increasing temperature correlate with the loss of tertiary structure, followed by melting of the AU-rich stem 2, and eventually the loss of the structure in G-C-rich stem 1. Our results are also in agreement with results obtained by Langevin simulations and a coarse-grained model with Go-like properties (TIS), where the formation of the stems and the assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures: if the secondary structural elements have comparable stability, then there are multiple routes to the native state; otherwise, there is one dominant path.[26] In the case of 2K96, H1 consists of G·C WC interactions that are more stabilizing than the A·U interactions of H2. Our predictions on the folding pathway extend those of the Go-like model study, which did not take into account the formation of noncanonical and triple pairings.

Given the challenges of folding a triple helix, for 2K96, we performed an extensive comparison of our results with seven secondary structure prediction methods that are available online, all of which allow the formation of pseudoknots. We tested the widely used MCFold,[9] Kinefold,[59] RNAstructure,[60] Vfold2D with Turner's parameters or MFOLD2.3[15] (considered to be the best performing algorithm to date), pknot,[16] pKiss,[61] and CyloFold.[62] As can be seen in Figure 12, none of these methods predict the triple helix. MCFold predicts two disjointed hairpins, and all of the other methods, considering the optimal or suboptimal solutions, predict a simple pseudoknot (pseudoknot H) when explicitly instructed to look for a pseudoknot. In principle, Vfold2D can predict structures that include base triplets, but it does not give the correct result for 2K96 despite the fact that it was shown to be able to produce the correct 2D structure, including triple contacts, for a similar system.[63]

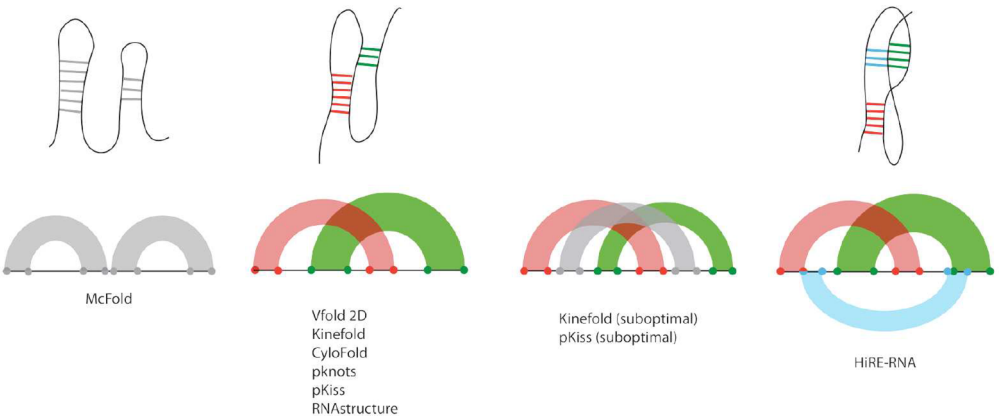


Figure 12. Results of secondary structure prediction algorithms for the triple helix pseudoknot. Under each topology, represented as arc graphs and sketched as secondary structure elements, we list the names of the algorithms that propose that result as an optimal or suboptimal prediction.

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

5.3. Large Systems. Despite the already substantial reduction in degrees of freedom of our theory, folding large structures remains challenging because of the long times needed for accurate sampling. Folding times can be reduced by adding partial experimental information, such as a few base-pairs from NMR or SHAPE.[64] This is the strategy we adopted for the riboswitch 1Y26 of 76 nt, starting from a fully extended state.[65] In its NMR state with an adenine ligand, 1Y26 adopts a Y shape with the two upper stems binding through kissing loops. Imposing three base pairs restraints (one WC pair on each helix) taken from VFold2D predictions, both simple MD (at 300 K) and REMD recover the overall organization of the kissing loops with an RMSD of 7−8 Å (Figure 13). The major discrepancy between our prediction and the NMR structures is at the junction where the adenine ligand, absent in our simulations, should sit. Our results for 1Y26 are comparable in quality to a prior prediction obtained with the automated 3DRNA program based on secondary structure reconstruction.[66] In this program, the smallest secondary elements (SSEs) are assembled into hairpins, hairpin loop, internal loop, bulge loop, pseudoknot loop, and junction, and then a network representation of the secondary structure is used to describe the locations and connectivity of the SSEs. Using the whole secondary structure extracted from the experimental structure (i.e., much more experimental information than the 3 base-pairing constraints we impose with HiRE-RNA), 3DRNA recovers the structure of 1Y26 with an RMSD of 6.7 Å.

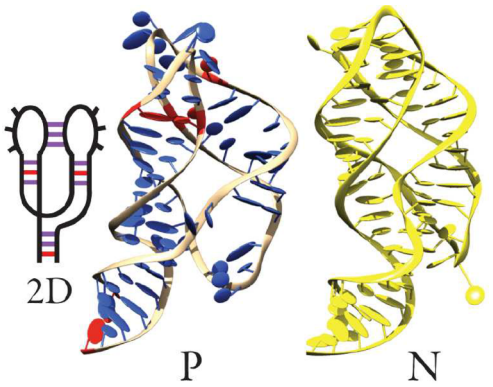


Figure 13. 1Y26 constrained prediction at 7.1 Å (P) and native structure in yellow (N). The three local constraints are shown in red in 2D.

6. DISCUSSION

We have presented here an effective theory for RNA folding based on a detailed and new description of the two main driving forces, stacking and base-pairing interactions, that are able to fold structurally diverse RNAs into their native states when coupled to REMD or ST. As the need to properly consider base pairing and stacking is clear to anyone working on RNA, how to actually define functions to describe these interactions in a simplified representation, allowing for the study of large-scale rearrangements, is less than straightforward. The force field we define is not close to any other existing coarse-grained model. The HiRE-RNA v3 force field shares with previous versions the particle description and functional forms of local interactions, but it is completely redesigned for all other terms, including new analytic energy functions for base pairing and base stacking that allow accurate prediction of equilibrium configurations of 13 systems with 12−76 nucleotides starting from fully extended states. The model makes use of a large number of parameters that have now been optimized following a rigorous procedure, a particularly complex task when the possible interactions considered go beyond simple Watson−Crick pairing and address a multitude of possible states, including noncanonical and multiple base-pairs.[67]

In spite of its simplicity, similar to Garcia's atomistic simulations, our model predicts the high resolution structure of tetraloops, which are not recovered by FARFAR/FARNA by fragment reconstructions, including noncanonical pairs. On larger structures, the prediction capabilities of HiRE-RNA are comparable to those of the most advanced methods currently used for RNA structure prediction. However, whereas we can obtain most of our results with no or minimal external information other than the sequence, these methods typically require substantial input of experimental evidence. As shown in the 2012 RNA-puzzle competition,[68] the best performing prediction algorithms so far are those based on fragment reconstruction, giving access to 3D structures, but not giving information on other aspects, such as the folding pathways and energy landscapes. In the competition, a riboswitch of 86 nts was predicted by eight research groups with RMSDs ranging from 7.2 to 23.0 Å. The model of lowest RMSD (7.2 Å) used a multiscale approach based on 2D structure prediction methods and self-assembly of fragments selected from the NDB.[11] Coarse-grained ab initio methods performed much poorer with the best RMSD at 11.5 Å even when relying heavily on experimental constraints.[36] More recently, Xia et al., with their coarse-grained model, obtained a structure at 7.6 Å using secondary structure prediction tools and 14 constraints.[69] As shown by the folding of the triple helix in the analysis of folding pathways, our model goes beyond what was done through Go-like models,[33,70] given that we do not introduce any bias toward the native structure and account for multiple pairings.

A key feature of our model, distinguishing it from most other methods, is the possibility of forming noncanonical and multiple base-pairs. Our results show the importance of considering noncanonical base pairs. Indeed, they are essential to fold complex molecules and should not be neglected, even for RNAs whose experimental structures contain only canonical pairs, as they have a significant impact on the free energy profile of the system, its thermodynamics, and possibly on folding pathways. The presence of noncanonical pairs gives rise to an increase of transition states that can favor interconversion between different configurations, a behavior that is observed for many biologically active RNAs[71] and that has been hypothesized for the switch within the pseudoknot domain of human telomerase RNA between the pseudoknot and hairpin conformations.[58] We are currently investigating these aspects in more detail through disconnectivity graphs[72] on several pseudoknots.

We also expect that RNA structures stabilized by interactions that are not currently represented in our model, such as ribose zippers, sugar-base hydrogen bonds, and base-phosphate interactions, will not be correctly treated by our model and will eventually need to be added in our representation. As the current status of HiRE-RNA v3 already represents a significant advancement in the study of RNA molecules both in terms of prediction capabilities and in the possibility of addressing questions concerning dynamical and thermodynamical behavior, we are extending our work to include other interaction

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

terms, such as the presence of ions and ligands, in order to be closer to the experimental conditions. Combined with the coarse-grained OPEP force field,[51] HiRE-RNA should help to understand the interplay between proteins and nucleic acids.

■ ASSOCIATED CONTENT

Supporting Information

Remaining parameters for the force field. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00200.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: samuela.pasquali@ibpc.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by the "Initiative d'Excellence" program from the French State (Grant "DYNAMO", ANR-11-LABX-0011-01) and IUF.

■ REFERENCES

(1) Ponting, C. P.; Oliver, P. L.; Reik, W. *Cell* **2009**, *136*, 629−641.
(2) Nagano, T.; Fraser, P. *Cell* **2011**, *145*, 178−181.
(3) Siomi, M. C.; Sato, K.; Pezic, D.; Aravin, A. a. *Nat. Rev. Mol. Cell Biol.* **2011**, *12*, 246−258.
(4) Cruz, J. A.; Westhof, E. *Cell* **2009**, *136*, 604−609.
(5) Holbrook, S. R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 302−308.
(6) Strobel, E.; Seeling, K.; Tebbe, C. C. *Env. Microbiol* **2008**, *10*, 483−496.
(7) Scott, W. G. *Curr. Opin. Struct. Biol.* **2007**, *17*, 280−286.
(8) Flores, S. C.; Altman, R. B. *RNA* **2010**, *16*, 1769−1778.
(9) Parisien, M.; Major, F. *Nature* **2008**, *452*, 51−55.
(10) Das, R.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 14664−14669.
(11) Cao, S.; Chen, S.-J. J. *J. Phys. Chem. B* **2011**, *115*, 4216−4226.
(12) Laing, C.; Jung, S.; Kim, N.; Elmetwaly, S.; Zahran, M.; Schlick, T. *PLoS One* **2013**, *8*, e71947.
(13) Shapiro, B. A.; Yingling, Y. G.; Kasprzak, W.; Bindewald, E. *Curr. Opin. Struct. Biol.* **2007**, *17*, 157−165.
(14) Zuker, M. *Nucleic Acids Res.* **2003**, *31*, 3406−3415.
(15) Xu, X.; Zhao, P.; Chen, S.-J. *PLoS One* **2014**, *9*, e107504.
(16) Rivas, E.; Eddy, S. R. *J. Mol. Biol.* **1999**, *285*, 2053−2068.
(17) Leontis, N. B.; Stombaugh, J.; Westhof, E. *Nucleic Acids Res.* **2002**, *30*, 3497−3531.
(18) Chen, A. A.; García, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 16820−16825.
(19) Bowman, G. R.; Huang, X.; Yao, Y.; Sun, J.; Carlsson, G.; Guibas, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2008**, *130*, 9676−9687.
(20) Zhuang, Z.; Jaeger, L.; Shea, J.-E. *Nucleic Acids Res.* **2007**, *35*, 6995−7002.
(21) Chakraborty, D.; Collepardo-Guevara, R.; Wales, D. J. *J. Am. Chem. Soc.* **2014**, *136*, 18052−18061.
(22) Zhang, Y.; Zhang, J.; Wang, W. *J. Am. Chem. Soc.* **2011**, *133*, 6882−6885.
(23) He, Y.; Maciejczyk, M.; M.Odziej, S.; Scheraga, H. A.; Liwo, A.; Maciejczyk, M.; Ołdziej, S. *Phys. Rev. Lett.* **2013**, *110*, 098101.
(24) Hyeon, C. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6789−6794.
(25) Denesyuk, N. a.; Thirumalai, D. *J. Phys. Chem. B* **2013**, *117*, 4901−4911.
(26) Cho, S. S.; Pincus, D. L.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 17349−17354.
(27) Morriss-Andrews, A.; Rottler, J.; Plotkin, S. S. *J. Chem. Phys.* **2010**, *132*, 35105.
(28) Sulc, P.; Romano, F.; Ouldridge, T. E.; Doye, J. P. K.; Louis, A. A. *J. Chem. Phys.* **2014**, *140*, 235102.
(29) Xia, Z.; Gardner, D. P.; Gutell, R. R.; Ren, P. *J. Phys. Chem. B* **2010**, *114*, 13497−13506.
(30) Cragnolini, T.; Derreumaux, P.; Pasquali, S. *J. Phys.: Condens. Matter* **2015**, *27*, 233102.
(31) Chen, G.; Chang, K.-Y.; Chou, M.-Y.; Bustamante, C.; Tinoco, I.; Tinoco, I., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12706−12711.
(32) Das, R. *PLoS One* **2011**, *6*, e20044.
(33) Biyun, S.; Cho, S. S.; Thirumalai, D. *J. Am. Chem. Soc.* **2011**, *133*, 20634−20643.
(34) Pasquali, S.; Derreumaux, P. *J. Phys. Chem. B* **2010**, *114*, 11957−11966.
(35) Cragnolini, T.; Derreumaux, P.; Pasquali, S. *J. Phys. Chem. B* **2013**, *117*, 8047−8060.
(36) Ding, F.; Sharma, S.; Chalasani, P.; Demidov, V. V.; Broude, N. E.; Dokholyan, N. V. *RNA* **2008**, *14*, 1164−1173.
(37) Martinez, H. M.; Maizel, J. V.; Shapiro, B. A. *J. Biomol. Struct. Dyn.* **2008**, *25*, 669−683.
(38) Sponer, J.; Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595−2610.
(39) Svozil, D.; Hobza, P.; Sponer, J. *J. Phys. Chem. B* **2010**, *114*, 1191−1203.
(40) Bowman, J. C.; Lenz, T. K.; Hud, N. V.; Williams, L. D. *Curr. Opin. Struct. Biol.* **2012**, *22*, 262−272.
(41) Li, P. T. X.; Vieregg, J.; Tinoco, I. *Annu. Rev. Biochem.* **2008**, *77*, 77−100.
(42) Sosnick, T. R.; Pan, T. *Curr. Opin. Struct. Biol.* **2003**, *13*, 309−316.
(43) Ouldridge, T. E.; Louis, A. a.; Doye, J. P. K. *J. Chem. Phys.* **2011**, *134*, 085101.
(44) Lemieux, S.; Major, F. *Nucleic Acids Res.* **2002**, *30*, 4250−4263.
(45) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins* **2007**, *69*, 394−408.
(46) Gan, H. H.; Pasquali, S.; Schlick, T. *Nucleic Acids Res.* **2003**, *31*, 2926−2943.
(47) Pasquali, S.; Gan, H. H.; Schlick, T. *Nucleic Acids Res.* **2005**, *33*, 1384−1398.
(48) Serra, M. J.; Axenson, T. J.; Turner, D. H. *Biochemistry* **1994**, *33*, 14289−14296.
(49) Lu, Z. J.; Turner, D. H.; Mathews, D. H. *Nucleic Acids Res.* **2006**, *34*, 4912−4924.
(50) Spill, Y. G.; Pasquali, S.; Derreumaux, P. *J. Chem. Theory Comput.* **2011**, *7*, 1502−1510.
(51) Sterpone, F.; Melchionna, S.; Tuffery, P.; Pasquali, S.; Mousseau, N.; Cragnolini, T.; Chebaro, Y.; St-Pierre, J.-F.; Kalimeri, M.; Barducci, A.; Laurin, Y.; Tek, A.; Baaden, M.; Nguyen, P. H.; Derreumaux, P. *Chem. Soc. Rev.* **2014**, *43*, 4871−4893.
(52) Nonin-Lecomte, S.; Felden, B.; Dardel, F. *Nucleic Acids Res.* **2006**, *34*, 1847−1853.
(53) Egli, M.; Minasov, G.; Su, L.; Rich, A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4302−4307.
(54) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 26−41.
(55) Theimer, C. A.; Giedroc, D. P. *RNA* **2000**, *6*, 409−421.
(56) Theimer, C. A.; Wang, Y.; Hoffman, D. W.; Krisch, H. M.; Giedroc, D. P. *J. Mol. Biol.* **1998**, *279*, 545−564.
(57) Kim, N.-K.; Zhang, Q.; Zhou, J.; Theimer, C. A.; Peterson, R. D.; Feigon, J. *J. Mol. Biol.* **2008**, *384*, 1249−1261.
(58) Theimer, C. A.; Blois, C. A.; Feigon, J. *Mol. Cell* **2005**, *17*, 671−682.
(59) Xayaphoummine, A.; Bucher, T.; Thalmann, F.; Isambert, H. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 15310−15315.
(60) Bellaousov, S.; Reuter, J. S.; Seetin, M. G.; Mathews, D. H. *Nucleic Acids Res.* **2013**, *41*, W471−W474.
(61) Theis, C.; Janssen, S.; Giegerich, R. Lecture Notes in Computer Science. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*; Moulton, V., Singh, M., Eds.; Springer: Berlin Heidelberg, 2010; Vol. *6293* LNBI, pp 52−64.

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(62) Bindewald, E.; Kluth, T.; Shapiro, B. A. *Nucleic Acids Res.* **2010**, *38*, W368−W372.
(63) Cao, S.; Giedroc, D. P.; Chen, S.-J. *RNA* **2010**, *16*, 538−552.
(64) Weeks, K. M. *Curr. Opin. Struct. Biol.* **2010**, *20*, 295−304.
(65) Serganov, A.; Yuan, Y.-R.; Pikovskaya, O.; Polonskaia, A.; Malinina, L.; Phan, A. T.; Hobartner, C.; Micura, R.; Breaker, R. R.; Patel, D. J. *Chem. Biol.* **2004**, *11*, 1729−1741.
(66) Zhao, Y.; Huang, Y.; Gong, Z.; Wang, Y.; Man, J.; Xiao, Y. *Sci. Rep* **2012**, *2*, 734.
(67) Bottaro, S.; Di Palma, F.; Bussi, G. *Nucleic Acids Res.* **2014**, *42*, 13306−13314.
(68) Cruz, J. A.; Blanchet, M.-F.; Boniecki, M.; Bujnicki, J. M.; Chen, S.-J.; Cao, S.; Das, R.; Ding, F.; Dokholyan, N. V.; Flores, S. C.; Huang, L.; Lavender, C. a.; Lisi, V.; Major, F.; Mikolajczak, K.; Patel, D. J.; Philips, A.; Puton, T.; Santalucia, J.; Sijenyi, F.; Hermann, T.; Rother, K.; Rother, M.; Serganov, A.; Skorupski, M.; Soltysinski, T.; Sripakdeevong, P.; Tuszynska, I.; Weeks, K. M.; Waldsich, C.; Wildauer, M.; Leontis, N. B.; Westhof, E. *RNA* **2012**, *18*, 610−625.
(69) Xia, Z.; Bell, D. R.; Shi, Y.; Ren, P. *J. Phys. Chem. B* **2013**, *117*, 3135−3144.
(70) Feng, J.; Walter, N. G.; Brooks, C. L. *J. Am. Chem. Soc.* **2011**, *133*, 4196−4199.
(71) Fürtig, B.; Richter, C.; Schell, P.; Wenter, P.; Pitsch, S.; Schwalbe, H. *RNA Biol.* **2008**, *5*, 41−48.
(72) Wales, D. *J. Mol. Phys.* **2002**, *100*, 3285.

DOI: 10.1021/acs.jctc.5b00200
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX