Biophysical Journal
Article

Biophysical Society

A Polymer Physics Framework for the Entropy of Arbitrary Pseudoknots

Ofer Kimchi,[1,*] Tristan Cragnolini,[2] Michael P. Brenner,[3,4] and Lucy J. Colwell[2,*]
[1]Harvard Graduate Program in Biophysics, Harvard University, Cambridge, Massachusetts; [2]Department of Chemistry, University of Cambridge, Cambridge, United Kingdom; [3]School of Engineering and Applied Sciences and [4]Kavli Institute for Bionano Science and Technology, Harvard University, Cambridge, Massachusetts

ABSTRACT The accurate prediction of RNA secondary structure from primary sequence has had enormous impact on research from the past 40 years. Although many algorithms are available to make these predictions, the inclusion of non-nested loops, termed pseudoknots, still poses challenges arising from two main factors: 1) no physical model exists to estimate the loop entropies of complex intramolecular pseudoknots, and 2) their NP-complete enumeration has impeded their study. Here, we address both challenges. First, we develop a polymer physics model that can address arbitrarily complex pseudoknots using only two parameters corresponding to concrete physical quantities—over an order of magnitude fewer than the sparsest state-of-the-art phenomenological methods. Second, by coupling this model to exhaustive enumeration of the set of possible structures, we compute the entire free energy landscape of secondary structures resulting from a primary RNA sequence. We demonstrate that for RNA structures of ~80 nucleotides, with minimal heuristics, the complete enumeration of possible secondary structures can be accomplished quickly despite the NP-complete nature of the problem. We further show that despite our loop entropy model's parametric sparsity, it performs better than or on par with previously published methods in predicting both pseudoknotted and non-pseudoknotted structures on a benchmark data set of RNA structures of $\leq$80 nucleotides. We suggest ways in which the accuracy of the model can be further improved.

SIGNIFICANCE The functions and properties of RNA molecules are closely tied to the set of structures they can fold into and their free energies. However, complex structures termed pseudoknots are not well predicted by current tools despite their prevalence. Here, we describe a method to analytically calculate the entropies of arbitrarily complex pseudoknots using only two parameters corresponding to concrete physical quantities. This approach represents an order-of-magnitude reduction in parameters compared to even the sparsest state-of-the-art tools. We employ this method alongside an exhaustive enumeration of the set of possible structures to predict the entire free energy landscape of short RNA molecules, given their sequence. Finally, we show that despite its parametric sparsity, our algorithm outperforms current state-of-the-art methods in pseudoknot prediction.

INTRODUCTION

RNA molecules play physiological roles that extend far beyond translation. In human cells, most RNA molecules are not translated (1). Noncoding RNAs interact functionally with messenger RNA (2), DNA (3), and proteins (4) and can be as large as thousands of nucleotides (nts) (5,6). However, a substantial fraction are <40 nts in length, including microRNAs and small interfering RNAs, which serve as regulators for the translation of messenger RNA (2,7), and piwi-interacting RNAs, which form RNA-protein complexes to regulate the germlines of mammals (8). The in vitro evolution of RNA, especially through systematic evolution of ligands by exponential enrichment (9–11), has led to an explosion of applications for short RNA molecules because of their ability to tightly and specifically bind to a remarkable range of target ligands (12).

Overwhelmingly, the properties of short noncoding RNA molecules are tied to their structures (13–15). Such structures are formed because of the energetic favorability of bonds between complementary nts (primarily A to U, C to G, and the wobble pair G to U). However, these bonds impose an entropic cost. Therefore, the conformations most frequently adopted balance the energetic gain of

Submitted November 8, 2018, and accepted for publication June 27, 2019.
*Correspondence: okimchi@g.harvard.edu or ljc37@cam.ac.uk
Editor: Tamar Schlick.
https://doi.org/10.1016/j.bpj.2019.06.037

maximal basepairing with the entropic cost of structural constraints. In equilibrium, the RNA adopts each possible structure with Boltzmann weighted probabilities.

Because of the relevance of RNA structure to function (16,17), current research aims to predict the minimum free energy (MFE) structures given the sequence. Algorithms typically predict "secondary structure," a list of the basepairings (18,19). The early Pipas-McMahon RNA structure prediction algorithm sought to completely enumerate and evaluate the free energy of all possible secondary structures, thereby constructing the entire energy landscape (20). More recent algorithms have made progress in making similar enumerations less computationally intensive (21), the most successful of which are the TT2NE algorithm and its stochastic version, McGenus (22,23). The complete landscape enumeration approach including all secondary structures has so far been limited to short ($<$30 nt) RNA molecules (24,25), and the field has instead almost entirely been dominated by dynamic programming approaches (26–30). Such algorithms efficiently consider an enormous number of structures without explicitly generating them by iteratively finding the optimal structure for subsequences (18).

Despite the substantial success of dynamic programming, these algorithms have difficulty predicting RNA secondary structures that include pseudoknots (i.e., structural elements with at least two non-nested basepairs) (see Fig. S1 *A* for an example) that make up roughly 1.4% of basepairs (18) and are overrepresented in functionally important regions of RNA (31). Pseudoknots are disallowed from the most popular RNA structure prediction algorithms (32,33) because of computational cost; indeed, enumerating all pseudoknotted structures a given RNA molecule can fold into has been shown to be NP-complete (34–36). Significant advances have been made with heuristics, which do not guarantee finding the MFE structure (23,37–43), and by disallowing all but a limited class of pseudoknots (44–51).

A further major challenge for predicting pseudoknotted structures is the relative lack of experimental data or physical models to estimate their entropies (52,53). An important caveat is the simple "H-type" pseudoknot for which both experimental data (54–57) and physical models (37,50,51,58–60) are available. However, for more complex single-molecule pseudoknots, even those which can be enumerated by current dynamic programming algorithms (47), entropy estimates have been limited to phenomenological extensions of the non-pseudoknotted and H-type pseudoknot models (43,44,61), and few experimental studies are available (62). A recent strategy uses machine learning of large experimental data sets (50,63); although these approaches can be useful, they come with the disadvantages of compounding possible experimental errors and often using an enormous number of parameters, which can impact generalizability. A sketch of a theoretical description of simple pseudoknot entropies based on polymer physics was developed by Isambert and Siggia (37,60); however, their derivations have not been published. Given the relative lack of experimental data to validate current simple phenomenological approaches on complex pseudoknots, the lack of a physical model for such structures is a pressing concern.

In this study, we develop a physical model to calculate the entropies of arbitrarily complex pseudoknots. We combine our model with complete enumeration of the secondary structure landscape, demonstrating that we can exactly solve for the probabilities of the RNA folding into each of the possible structures, including those with pseudoknots (Fig. 1). We demonstrate that this approach is feasible, not only for short RNA molecules of $\sim$25 nts that have been examined in previous studies (25) but even for biologically relevant RNA sequences of $\sim$80 nts in length.

Our approach combines a method based on the work of Isambert and Siggia with a, to our knowledge, novel graph-theoretical depiction of the RNA, allowing us to calculate the entropy of any arbitrary RNA structure. We demonstrate the generality of our formalism using the H-type and kissing hairpin pseudoknots as examples. Despite this generality, our loop entropy model uses only two parameters corresponding to experimentally derived physical quantities: the persistence length of single-stranded RNA, and the volume within which two RNA nts are considered bound. This represents an enormous parameter reduction compared to state-of-the-art algorithms; for example, the phenomenological Dirks-Pierce model has 11 parameters for the loop entropy of pseudoknots and $\sim$18 parameters for non-pseudoknotted loops (63).

We test our model predictions on molecules from the RNA STRAND (64), PseudoBase++ (65), and CompaRNA (66) databases and find good agreement with experimental results. We find that a significant heterogeneity in pseudoknot types exists even for sequences $\leq$80 nts in length, based on the polymer model representing their entropies. This heterogeneity is found to result in systematic errors of heuristic models' estimates of the entropies of complex pseudoknots, motivating the generality of the entropic model derived here, which can correct such errors. Although we fit our entropy model only to data from non-pseudoknotted structures, we find that our model performs as well or better than previously published methods in predicting pseudoknots while performing on par with current methods in the prediction of non-pseudoknotted structures. Given the success of the model alongside its parametric sparsity, future work should build upon it to include further biological considerations neglected in the current treatment, and we give suggestions for where such improvements can be made.

METHODS

Calculating free energies

The probability of the RNA sequence folding into a given equilibrium structure σ is given by the Boltzmann factor:

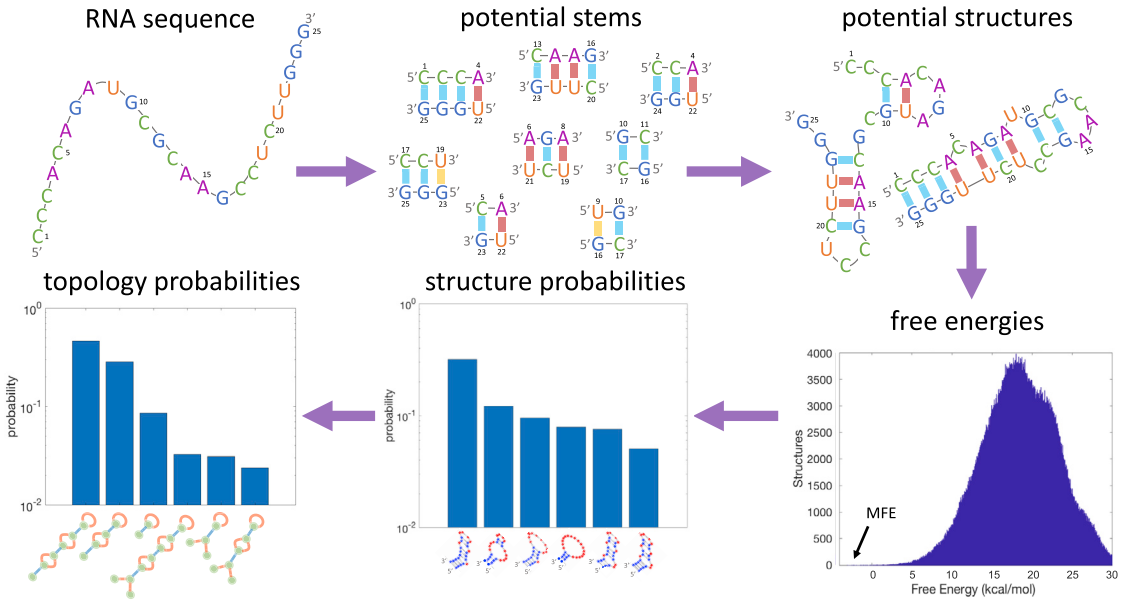

FIGURE 1 Schematic overview of the algorithm. Given an RNA sequence, the algorithm first enumerates all potential stems (sequences of consecutive basepairs) that can form. It then searches for all possible combinations of stems such that no nt is paired with more than one other, thus forming all possible secondary structures. For each structure, it calculates the free energy, which is comprised of a bond free energy term and a loop entropy term. In this work, we describe a polymer physics model to calculate this loop entropy term for arbitrarily complex pseudoknotted structures using only two parameters. The histogram of free energies for the sequence shown is plotted with an arrow pointing to the minimum free energy (MFE). Given the entire free energy landscape, the algorithm calculates the probability of any arbitrary secondary structure of forming in equilibrium. Finally, we coarse grain over similar structures described by the same topology, arriving at a probability distribution for every possible topology forming in equilibrium. To see this figure in color, go online.

$$p(\sigma) = \exp(-\beta G_\sigma)/Z, \tag{1}$$

where $\beta = 1/k_B T$ (T is the temperature, and k_B is Boltzmann's constant), and the partition function, Z, is defined such that the probability distribution is normalized: $\sum p(\sigma) = 1$. Here G_σ, the Gibbs free energy of structure σ, is a function of the enthalpy H_σ and entropy S_σ of the structure:

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ, \tag{2}$$

where we drop the subscripts for notational convenience and introduce Δ to signify that free energies are measured with respect to the free chain. The superscripts implying standard conditions will be dropped from here on.

We separate the free energy calculation into two independent components: the free energy of consecutive basepairs (stems) and the free energy of loops. We make the simplifying assumption that ΔH is determined solely by the basepairs in the structure, ignoring higher order corrections, such that $\Delta H = \Delta H_{\text{stems}}$. For the entropy, we make no such assumption, and $\Delta S = \Delta S_{\text{stems}} + \Delta S_{\text{loops}}$, where the entropy of stems represents the entropy lost by basepaired nts, and the entropy of loops represents the entropy lost by the constraints those basepairs place on the rest of the molecule. To calculate the terms ΔH_{stems} and ΔS_{stems}, we consider nearest-neighbor interactions among basepairs following the Nearest Neighbor Database (67), assuming (with few exceptions tabulated in the database) independence of the free energy contributions of each stem. See further details in Supporting Materials and Methods.

Calculating loop entropies

The goal of this and the next section is to build up a theoretical framework to estimate the loop entropies of arbitrarily complex RNA pseudoknots. This calculation has a significant effect on the prediction results. In fact, the magnitude of the loop entropy is on average equal to that of the overall free energy at physiological temperatures (see Fig. S7). This is as expected intuitively; the difficulty in RNA structure prediction lies precisely in predicting the balance between the energy gain from basepair constraints and the entropy gain from unpaired nts.

Because the following calculation is somewhat involved, we will begin by clarifying explicitly the nature of the loop entropy. A free RNA chain has a large number of conformations available to it, which we will call Ω. The loop entropy is the quantification of the reduction in conformations available to the RNA molecule upon introducing constraints on the structure, such as that certain nts are paired (68).

Ω depends on the length x of the RNA, such that $\Omega(x_1)\Omega(x_2) = \Omega(x_1 + x_2)$; in other words, we assume (for the free chain) independence of the various subsections of the RNA. This is in principle only true in the limit x_1, $x_2 \gg b$, where $b \approx 2.4$ nts is the Kuhn length of single-stranded RNA, and further neglects self-avoidance of the RNA molecule. Throughout, we will consider regions of single-stranded RNA long enough such that $x \gg b$ but short enough such that we assume self-avoidance has negligible probability. We discuss how to systematically consider shorter RNA loops in Supporting Materials and Methods and will make some notes regarding self-avoidance later in this section. We will also make the approximation that Ω is independent of sequence.

When a loop is formed in RNA, that loop constrains the number of conformations available to the RNA. For example, an RNA molecule that has its first nts bonded to its last only has available to it a fraction of the conformations available to the free chain—namely, all those that have the first and last nts close enough to bind. We are interested not in absolute values of the entropy S, but in ΔS, where the free chain is our reference state with $\Delta S^{\text{free}}_{\text{loops}} = 0$. The entropy of a structured RNA of length x with ω_{struct} conformations available to it is given by $\Delta S_{\text{loops}} = k_B \log(\omega_{\text{struct}}/\Omega(x)) < 0$, where we have written the difference of logs as the log of the ratio. We can simplify this formula by writing $\omega_{\text{struct}} = \Omega(x) \times p$, where p is the fraction of conformations available to the free chain that are consistent with the structure being considered. We therefore have

$$\Delta S_{\text{loops}} = k_B \log p. \tag{3}$$

It is worth reiterating that the entropy of stems themselves was already taken into account in the term ΔS_{stems} and that ΔS_{loops} only measures the entropy lost because of loop closures (69). To avoid overcounting the entropy lost because of the constraints placed on basepaired nts, stems do not directly contribute to ΔS_{loops}. Therefore, a stem comprised of l basepairs (or $2l$ nts) should be treated—for the purposes of the ΔS_{loops} calculation—as if it has $\Omega(2l)$ available conformations; that it in reality has far fewer has already been quantitatively accounted for in the ΔS_{stems} term. Because of this, factors of Ω cancel out entirely in calculations of ΔS_{loops}.

We now turn to polymer physics to quantitatively describe how loop closure constraints affect p, the fraction of configurational space available to the molecule. We model a single-stranded region comprised of x unpaired nts as a random walk of $(x+1)/b$ steps, whereas before b is the Kuhn length of single-stranded RNA. We denote by $P_s(\vec{R})d\vec{R}$ the probability of a random walk of length s to have end-to-end vector $\vec{R}$:

$$P_s(\vec{R}) = \left(\frac{3}{2\pi sb}\right)^{3/2} \exp\left(-\frac{3R^2}{2sb}\right). \quad (4)$$

We have assumed $s \gg b$ to arrive at the Gaussian formula above through the central limit theorem. The mean of the Gaussian is zero by symmetry. To find the variance we first consider a single step of length b in three dimensions, which has variance in the $\hat{i}$, $\hat{j}$, and $\hat{k}$ coordinates of $b^2/3$ by symmetry. For a random walk of $N = s/b$ steps, by independence of subsequent steps, the total variance is equal to $Nb^2/3 = sb/3$, leading to Eq. 4.

Eq. 4 is accurate for non-self-avoiding random walks; self-avoiding random walks cannot be treated analytically in this way. However, for sufficiently short walks, the probability of self-interaction is low. As described in Supporting Materials and Methods, we can systematically consider higher order corrections to Eq. 4 while maintaining its Gaussian nature. Whereas the accuracy of the assumption $s \gg b$ does not always hold in the problems considered, we ultimately find very good agreement between results using Eq. 4 and experiment and that corrections to Eq. 4 as described in Supporting Materials and Methods, are negligible.

For a structure with n single-stranded regions of lengths s_i $(1 \le i \le n)$, the fraction of conformations consistent with the structure is given by the following:

$$p = \int' \prod_i P_{s_i}\left(\vec{R}_i\right) d\vec{R}_i, \quad (5)$$

where $\vec{R}_i$ is the end-to-end distance vector of the i^{th} single-stranded region, and the primed integral is taken only over those $\vec{R}_i$ consistent with the overall structure. We will describe how to address these integrals via a Feynman diagram-like approach in the next section.

To demonstrate how Eqs. 3, 4, and 5 are applied, we first consider the simple hairpin loop. We will call its entropy $\Delta S_{\text{closed-net-0}}$, neglecting the subscript of "loops" from here on. The notation follows (37,60), and the subscript references the number of stems enclosed by the loop (zero in this case; see Fig. S2 for other examples). Following Jacobson and Stockmayer (70), we allow that basepairing can occur as long as the two nts are within a small volume v_s of one another. We assume that the bond length r_s is small enough that for all $|\vec{R}| \le r_s$, $P_s(\vec{R}) \approx P_s(\vec{0})$. Therefore, $p = v_s P_s(\vec{0})$, and Eqs. 3, 4, and 5 yield

$$\Delta S_{\text{closed-net-0}} = k_B\left[\log(v_s) + \frac{3}{2}\log\left(\frac{3}{2\pi sb}\right)\right]. \quad (6)$$

We emphasize that within our model, this formula is applicable to hairpin loops, bulge loops, internal loops, and multiloops. We discuss in a later section how our model can be extended to break this equivalency.

We estimate v_s by fitting experimental measurements of the entropy of hairpin loops of variable lengths to Eq. 6. Although Eq. 6 implies that the entropy of a hairpin should increase monotonically as a function of its length, the experimental measurements are nonmonotonic, and their nonmonotonicity exceeds the error bars (71). This nonmonotonicity may be due to enthalpic effects (72), which were neglected in our analysis following (30). Nevertheless, Fig. 2 shows that Eq. 6 gives a reasonable fit to the experimental data with $v_s = 0.0201 \pm 0.0036$ nts^3. A more precise definition of v_s might include a dependence on the closing basepairs of the hairpin loop; we expect that the penalties placed on specific closing basepairs and first mismatches in (30,71) play a similar role, though such penalties were not included here. If one ignores all angular dependences of bond formation, our estimate of v_s leads to a naive underestimate of the length of a hydrogen bond of 0.56 Å, which nonetheless is well within an order of magnitude of the true length of hydrogen bonds.

Because we find b using previous experimental results and fit v_s based on data from non-pseudoknotted structures, our model is in truth a zero-parameter model when it comes to pseudoknots. No data from pseudoknotted structures were used to fit our model.

Pseudoknot loop entropies: RNA Feynman diagrams

Our goal in this section is to find Eq. 5 for arbitrary pseudoknots. In Eq. 5, the $P_s(R)$ terms are given by the single-stranded segments, whereas stems appear through the constraint on the integral. The persistence length of double-stranded RNA is extremely long (~200 nts (73)) compared to both single-stranded RNA and the length of any stem we will actually consider. Therefore, we will model stems as rigid rods with a fixed end-to-end distance given by the length of the stem. In other words, a stem in which nts i through $i+k$ are bound to j through $j-k$ constrains nts i and $i+k$ (as well as j and $j-k$) to be a fixed distance apart. As we will see, such constraints end up only affecting the value of the integral for pseudoknotted structures, as exemplified by Fig. 3 *c*.

To calculate the entropy of a pseudoknot of arbitrary complexity, we invent a, to our knowledge, novel graph formulation inspired by Feynman diagrams from quantum field theory. We build on previous work by Rivas and Eddy (44) and later by Orland and Zee (74) who developed innovative graphical decomposition methods for RNA structures for the purposes of pseudoknot enumeration; here, we use a related diagrammatic approach for the entropy calculation instead. First, the RNA structure being considered is translated into a graph. Nodes are used to represent the two end points of a stem, and two types of edges represent single- and double-stranded RNA.

Defined in this way, the graph of the RNA structure directly represents the integrals necessary to compute its entropy. The positions of the nodes,

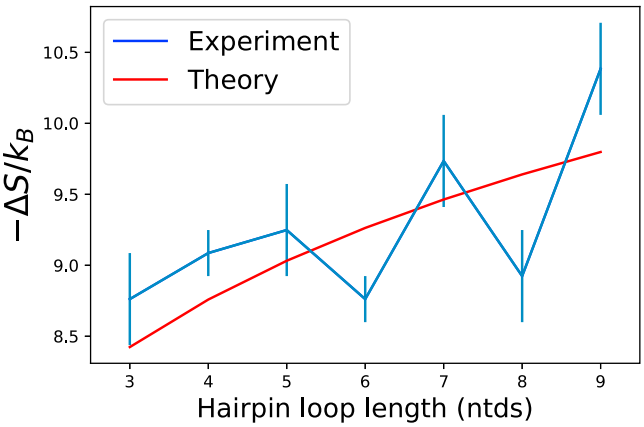

FIGURE 2 v_s estimated from experimental data. Experimental estimates for the free energy of hairpin loops of length s from Table 1 of (71) were converted to entropy estimates (*blue points* and error bars) by assuming $\Delta H = 0$ as in (30). These data were fit to Eq. 6, yielding an estimate of $v_s = 0.0201 \pm 0.0036$ nts^3. To see this figure in color, go online.

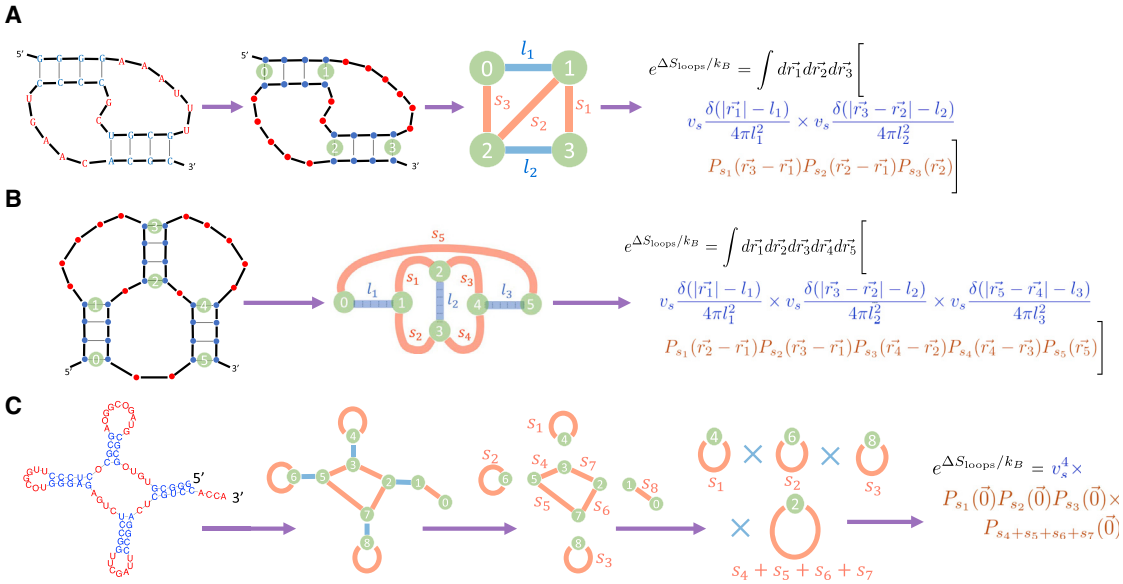

FIGURE 3 RNA Feynman diagrams. (*a*) An instance of the canonical H-type pseudoknot is shown (*the first panel*). Bold lines represent the RNA backbone; thin lines represent hydrogen bonds. The loop entropy of this structure can be calculated by first assuming sequence independence of the loop entropy (*second panel*) and then converting the structure to a graph (*third panel*). The nodes of the graph represent the first and last basepairs of each stem, and two types of edges represent single- and double-stranded RNA. The graph directly represents the integral in Eq. 7, reprinted in the fourth panel. The nodes are integrated over three-dimensional space, subject to constraints specified by the rigid double-stranded edges (*blue*), which correspond to delta functions. The integrand is given by the flexible single-stranded edges (*red*), which correspond to a $P_s(\vec{R})$ term. (*b*) The intramolecular kissing hairpin pseudoknot (*first panel*) is converted to a graph (*second panel*), representing the integrals necessary to compute its loop entropy (*third panel*). Although for this structure, the integrals are in general not analytically solvable, we numerically solve them (see Supporting Materials and Methods) as well as solve them analytically for the case $s_1/s_3 = s_2/s_4$ (Eq. 8). (*c*) The process of calculating the loop entropy of an RNA structure by converting it to a graph representing the entropy in integral form can be applied to any arbitrary structure. Separable integrals are represented by graphs which can be disconnected by the removal of any one edge. Thus, once appropriate factors of v_s are included (one for each stem in the original structure), the loop entropy of the example structure in question is simple to calculate and is given by four closed-nets-0 (originating from the three hairpins and multiloop). The four closed-net-0 loops contribute multiplicatively to the exponential of the loop entropy, meaning additively to the loop entropy itself. For non-pseudoknotted structures, all double-stranded edges (*blue*) can be removed in this way. To see this figure in color, go online.

$\vec{r}_i$, are integrated over all of space, while the constraints of the structure are included in the integrand: a double-stranded edge of length l between nodes i and j leads to a term $v_s\delta(|\vec{r}_i - \vec{r}_j| - l)/4\pi l^2$ (because of the rigid rod approximation of the stem), and a single-stranded edge of length s between these nodes leads to a term $P_s(\vec{r}_i - \vec{r}_j)$ in the integrand (as in Eq. 5). Note that two bonded nts in isolation are considered a stem of length $l \to 0$.

As a concrete example, we consider the canonical H-type pseudoknot, an instance of which is shown in Fig. 3 *a* (first panel). The loop entropy is sequence independent (second panel) and can be calculated by translating the structure into a graph (third panel) in which each node represents the edge of a stem, blue edges represent regions of double-stranded RNA of length l_i, and red edges represent regions of single-stranded RNA of length s_i. For the example in Fig. 3 *a*, $s_3 = 6$ nts, and $l_1 = 3$ nts. We set the origin of our coordinate system to node 0 and call the distance vector between node i and the origin $\vec{r}_i$. Integrating over the possible placements of nodes 1–3 (while including the constraints of the structure in the integrand as described previously) we obtain the following Gaussian integral formulation of the entropy:

$$e^{\Delta S_{\text{H-type}}/k_B} = v_s^2 \int d\vec{r}_1 \int d\vec{r}_2 \int d\vec{r}_3 \frac{\delta\left(\left|\vec{r}_1\right| - l_1\right)}{4\pi l_1^2} \times \frac{\delta\left(\left|\vec{r}_3 - \vec{r}_2\right| - l_2\right)}{4\pi l_2^2} P_{s_1}\left(\vec{r}_3 - \vec{r}_1\right) \times P_{s_2}\left(\vec{r}_2 - \vec{r}_1\right) P_{s_3}\left(\vec{r}_2\right), \quad (7)$$

where using the assumption $s \gg b$, we allow the integrals to extend over all of space. A more comprehensive derivation of this formula, including the origin of the v_s terms, can be found in Supporting Materials and Methods. This integral can be calculated analytically (Supporting Materials and Methods; (37)).

A complex pseudoknot involved in biological processes ranging from viral replication to antisense regulation is the intramolecular kissing hairpin pseudoknot (Fig. 3 *b*; (43,75–79)). Despite its biological prevalence, its entropy cannot be estimated using existing formalisms, necessitating the use of simple heuristic energy models (43). Our formalism on the other hand can readily address this pseudoknot by translating the structure to integrals as in Eq. 7. Although the integrals representing the entropy of the kissing hairpin are not in general analytically solvable, they are for the special case of $s_1/s_3 = s_2/s_4$. Rescaling the s to be s/γ with $\gamma = 3/2b$, we define the variables $s_c = s_5(s_1 + s_2)(s_3 + s_4)$ and $s_d = s_1 s_2(s_3 + s_4) + s_3 s_4(s_1 + s_2)$ along with

$$s_A = \frac{s_5(s_c + s_d)}{s_c}; s_B = \frac{s_3 s_4 (s_1 + s_2)^2}{s_d}; s_v = \sqrt{s_c + s_d}$$

to arrive at

$$e^{\Delta S_{\text{KH}}/k_B} = \frac{(v_s/s_v)^3}{2\pi^{9/2}} \frac{s_A}{l_1 l_3} e^{-\left(\frac{l_1^2 + l_3^2}{s_A}\right) - \frac{l_2^2}{s_B}} \sinh\left(\frac{2 l_1 l_3}{s_A}\right), \quad (8)$$

where sinh is the hyperbolic sin function.

The complete derivation of Eq. 8, along with a derivation of the numerically solvable general case, can be found in Supporting Materials and Methods. We have also provided an eight-dimensional table of the results of the numerical integration for different combinations of the s and l as Supporting Materials and Methods (Table S2).

We note that the intermolecular kissing hairpin complex, for which physical models have previously been developed (80), is simpler than the intramolecular structure in the context of our formalism, and its entropy calculation is shown in Fig. S2.

Our Feynman diagram-like graphical formalism allows intuitive manipulation of the integrals. Graphs that can be disconnected by the removal of any one edge correspond to separable integrals and thus to distinct motifs in the RNA structure. The decomposition of a structure into its component graphs is depicted in Fig. 3 *c* for a classical cloverleaf RNA (a second example, this one of a pseudoknotted RNA, is provided in Supporting Materials and Methods, Section 10). The RNA in question decomposes into four instances of closed-net-0 (originating from the three hairpins and multiloop) and one instance of an open-net-0, or free chain (which by definition does not affect the entropy). For non-pseudoknotted structures, once appropriate factors of v_s are included in the integrals (one for each double-stranded edge of the graph), all double-stranded edges can be removed through this graphical decomposition process. As shown in the figure, nodes that can be removed without changing the topology can be removed in the graph decomposition process. This is made possible by the property of $P_s(\vec{r})$ that $\int P_x(\vec{r_1})P_y(\vec{r_2}-\vec{r_1})d\vec{r_1} = P_{x+y}(\vec{r_2})$ (see Supporting Materials and Methods for further discussion).

In Fig. S2, we display all possible graphs of up to two stems and their respective RNA structures. As in Fig. 3, single-stranded edges are displayed with red, and double-stranded are displayed with blue. For each graph, the integral formulation of its entropy is displayed in the figure alongside what it evaluates. RNA sequences, even those of length $\leq$80 nts, form a wide array of pseudoknots more complex than those discussed in that figure, such as H-type pseudoknots with internal loops. Heuristics for treating such pseudoknots make systematic errors that our model can correct. See Supporting Materials and Methods for further discussion.

In Supporting Materials and Methods, we provide a full sample calculation for the free energy of a pseudoknotted structure.

Comparison of methodology to other physics-based pseudoknot entropy models

Although our model is able to address arbitrarily complex pseudoknots, prior physical models have been developed to address H-type pseudoknots in particular. The parametric sparsity of the model described above necessitates a neglect of several biological considerations, which have been considered by these previous models. Here, we will discuss how the framework developed above can be modified to include several factors considered in such models. The rationale for building atop our framework is provided in the next section. We demonstrate that despite the loop entropy model's apparent physical simplicity—it uses an order of magnitude fewer parameters than current tools while being general enough to apply to arbitrarily complex pseudoknots—it performs on par with state-of-the-art prediction software and therefore appears to succinctly capture the essential physics at play (see Results and Discussion).

An early model for the loop entropy of pseudoknots was developed by Gultyaev et al. (81). That model was based in large part on Jacobson and Stockmayer's derivation of the loop entropy of hairpins, which is rederived (Eq. 6) and then significantly extended by our formalism. To account for excluded volume, Gultyaev et al. replaced the factor of 3/2 in Eq. 4 with 1.75 (82). Such a change does not accurately account for excluded volume for the case of pseudoknots; we therefore did not make this replacement in our own article (in an effort for self-consistency), though it can easily be made. A more systematic treatment of how to include self-avoidance for the case of complex pseudoknots is still lacking.

The first pseudoknot models such as Gultyaev's did not consider interhelix loops for the H-type pseudoknot (i.e., they only considered those structures for which $s_2 = 1$ in the language of Fig. 3 *a*). The approximation made in our own work is in fact the opposite limit—that of $s_2 \gg b$—and our results should be most appropriate for long single-stranded regions. More precise treatment of short loops would forgo the simple ideal chain approximation of Eq. 4 in favor of the worm-like chain approximation. Although it would preclude analytic solutions of the integrals, numeric integration can easily be employed to make an effective look-up table as we demonstrated for the intramolecular kissing hairpin pseudoknot.

A similar complication is dealt with in Cao and Chen's Vfold model, which considers bond geometries explicitly using the diamond lattice (50). Although the enumeration procedure employed on the lattice is not computationally feasible for very large or complex pseudoknots, it is expected to capture the atomistic geometries more precisely than our own continuous three-dimensional space theory. Modifications can still be made within our framework, most directly by integrating only over a specific range of angles determined by the geometry. Such geometric considerations may also affect our treatment of non-pseudoknotted structures and, in particular, our equivalent treatment of hairpin, internal, bulge, and multiloops (83).

Perhaps most importantly, our model neglects the twists of the RNA helix. These twists may play a role in the nonmonotonicity of the experimental data in Fig. 2 and are likely significant. Isambert's KineFold model claims to effectively consider such twists by modification of the value of the double-stranded stem lengths l inputted to the pseudoknot formulae (60); however, as for the pseudoknot formulae themselves, the derivations of these modifications have not been published, and no physical basis for them was given. Finally, although we do not distinguish between the major and minor grooves of the RNA, accounting for the different grooves can explain asymmetries in physiological H-type pseudoknots (58). Aalberts and Nandagopal demonstrated that with the addition of a single experimentally measured parameter, $P_s(\vec{R})$ can be modified to account for this factor (84,85).

Enumerating RNA structures

In this section, we describe the process by which we exhaustively enumerate the secondary structures, including pseudoknots, into which an arbitrary given sequence can fold. This process was developed by Pipas and McMahon (20). The Pipas-McMahon algorithm first enumerates all possible secondary structures for a given sequence (sans pseudoknots) and then evaluates the free energy for each to construct the entire free energy landscape for non-pseudoknotted structures. A major shortcoming is the significant computer time required for long sequences. However, the exponential increase in computer power over the past 40 years, coupled with increased appreciation for the physiological and engineering relevance of short RNA strands, suggests revisiting this approach. This process is also employed by the TT2NE algorithm, with the caveat that rather than stems, that algorithm uses helipoints—defined as sets of stems separated by a bulge loop of size one or a 1 × 1 internal loop—as the backbone of the enumeration procedure, thus coarse graining over many similar structures (22).

We first number the nts in the RNA sequence from 1 to N from the 5′ end. We define an $N \times N$ symmetric matrix B, which describes which nts can bind to each other: $B_{i,j} = 1$ if nts i and j can bind to form a basepair (i.e., they belong to the set $\{(A,U)(C,G)(G,U)\}$) and 0 otherwise.

Next, we search for all possible stems (strings of consecutive basepairs) that could form. We define a parameter m to be the minimal allowed stem length ($m \geq 1$; we set $m = 1$ throughout unless otherwise specified). We also impose the physical constraint that hairpins (single-stranded region connecting one end of a stem) have a minimal length of three nts. We include not only the longest possible stems that can form but all contiguous subsets of those stems (86,87). We denote the number of stems found by N_{stems}.

We next define the $N_{\text{stems}} \times N_{\text{stems}}$ symmetric compatibility matrix C, where $C_{p,q} = 1$ if a structure could be made with both stems p and q ac($C_{q,q} = 1\ \forall\ q$). We impose the constraint that each nt may be paired with, at most, one other nt by setting $C_{p,q} = 0$ if stems p and q share at least one nt.

Finally, we explicitly enumerate the remaining possible secondary structures by identifying all compatible combinations of stems. Starting from a single stem s_1, we consider stems s_2 where $1 \leq s_1 \leq s_2 \leq N_{\text{stems}}$ and add the first stem for which $C_{s1,s2} = 1$. Then, we repeat the process, adding the first stem $s_3 > s_2$ compatible with both s_1 and s_2 and so forth, continuing until we can add no more stems. We add the resulting structure, composed of M stems, to the list of possible structures, remove the last stem added (to obtain the structure composed of stems s_1, s_2, …, s_{M-1}), and continue the process. This algorithm returns all possible secondary structures resulting from the primary sequence.

The algorithm described here was implemented in MATLAB (The MathWorks, Natick, MA), and all code is available on the GitHub repository (https://github.com/ofer-kimchi/RNA-FE-Landscape). The repository also includes a Python version of the code.

Once we completely enumerate the possible secondary structures, we calculate the probabilities that the RNA will fold into each of them by calculating their free energies as described in the previous sections.

RESULTS AND DISCUSSION

We use experimentally determined structures to compare the predictions of our model with other current methods; results are shown in Fig. 4. For sequences of length $\leq$80 nts from the RNA STRAND (64), PseudoBase++ (65), and CompaRNA (66) databases (186 non-pseudoknotted structures with 58 different topologies; 235 pseudoknotted structures with 52 different topologies), which had a sequence dissimilarity $\geq$0.2 (using Jukes-Cantor), we measured the number of basepairs correctly predicted by our algorithm's MFE structure compared to 14 other current algorithms. Seven of these cannot predict pseudoknots and serve as useful benchmarks for the non-pseudoknotted results (detailed methods in Supporting Materials and Methods; we have included the entire benchmark data set in Table S1). We also tested whether our algorithm's predictions are dependent on the accuracy of our loop entropy model by setting all loop entropies to zero (dark green). The poor performance of our algorithm in this case compared to the case in which loop entropies are considered demonstrates the success of the loop entropy model.

Although the entropy model presented here can give an integral expression for arbitrarily complex pseudoknots, the integral may need to be solved numerically for sufficiently complex structures. For this large-scale comparison, we disallowed pseudoknots more complex than those displayed in Fig. S2, and our algorithm therefore did not require any numerical integration. Fig. S6 demonstrates that even without this practical constraint, the complete enumeration of secondary structures including all possible pseudoknots is nonprohibitive. We similarly disallowed parallel stems, which can be stable in neutral and acidic pH conditions (88). We also set the minimal stem length for each sequence (m) to the minimal value it could take such that the total number of possible stems is less than $N_{\text{stems}}^{\max} = 150$. These choices were all made to speed up computation time; each sequence took between several seconds and an hour to run. Details of the computation time of our algorithm can be found in Figs. S4–S6.

Although these practical constraints were chosen to speed up the computation time, they also led to errors in the

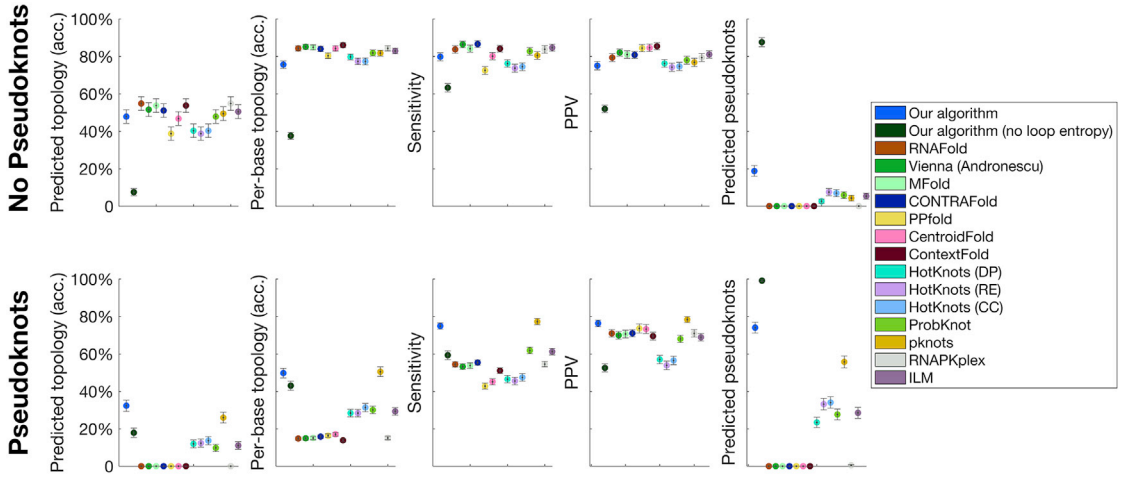

FIGURE 4 Summary statistics for comparison to other prediction tools. To assess the relative success of our algorithm, we compare its performance in predicting experimentally determined RNA structures to that of 14 other current prediction tools: RNAFold (33,118), ViennaRNA (Andronescu parameters) (119), Mfold (32), CONTRAfold (120), PPfold (121), CentroidFold (122), Context Fold (123), HotKnots (Dirks-Pierce parameters), HotKnots (Rivas-Eddy parameters), HotKnots (Cao-Chen parameters) (63), ProbKnot (40), PKNOTS (44), RNAPKplex (33,118), and iterated loop matching (ILM) (38). We measure sensitivity, PPV, the fraction of topologies predicted correctly by the MFE structure, the average per-base topology accuracy (defined in the main text), and the fraction of MFE structures containing a pseudoknot. We separate the results into sequences that experimentally form pseudoknots and those that do not. Error bars show the standard error. Despite the fact that our algorithm requires only two parameters to describe the entropy of any arbitrary secondary structure (at least an order of magnitude–and often several–fewer than the other algorithms tested against) and that the parameters were trained on non-pseudoknotted structures, our algorithm outperforms the other algorithms tested in predicting pseudoknotted structures and performs on par with them in predicting non-pseudoknotted structures. We also demonstrate that our algorithm's success is dependent on the accuracy of our loop entropy model because setting all loop entropies to zero (*dark green*) leads to poor performance (see main text for further discussion). To see this figure in color, go online.

algorithm's predictions. Of the tested pseudoknots, 64 were topologically more complex than any of those presented in Fig. S2. Furthermore, 33 of the non-pseudoknotted sequences tested (and eight of the pseudoknotted) include basepairs outside of those allowed by the algorithm (AU, GC, and GU). Removing such structures from our comparison analysis leads to our algorithm performing even better compared to current tools (see Fig. S3).

Further errors were due to our choice of m, which was not optimized and was too high compared to the length of the shortest stem in the experimental structure for 58 non-pseudoknotted cases and 54 pseudoknotted cases. By changing $N_{\text{stems}}^{\max}$ from 150 to 200, these numbers decreased to 46 for both pseudoknotted and non-pseudoknotted sequences, but the results for $N_{\text{stems}}^{\max} = 200$ were practically identical to the results of Fig. 4 (see full results in Table S1). For $N_{\text{stems}}^{\max} = 200$, the computation time was increased significantly (to several hours in the worst cases, though the majority of the computation time is spent on the Feynman diagram decomposition process, which has not been optimized in the current code). In addition to these sources of error, the nearest-neighbor parameters may need to be reexamined to be used most effectively with the loop entropy model presented here.

We considered the basepairs present in the experimental structure and in each algorithm's MFE structure. Basepairs present in both were labeled as true positives (*TP*), those present only in the predicted algorithm were labeled as false positives (*FP*), and those present in the experimental structure but not the predicted MFE structure were labeled as false negatives (*FN*). The sensitivity ($TP/TP + FN$) and the positive predictive value (PPV; $TP/TP + FP$) of our algorithm were measured to be 0.80 and 0.75 for the non-pseudoknotted cases and 0.75 and 0.76 for the pseudoknotted cases, respectively. Our algorithm performed better than or as well as all other prediction tools tested for the prediction of pseudoknots and on par with other tools in the prediction of non-pseudoknotted sequences. The full results can be found in Table S1.

Although sensitivity and PPV are the most common metrics used to establish the success of an RNA prediction algorithm (89), we sought to develop a test that measures success on the scale of the full RNA rather than on the scale of individual basepairs. To this end, we measured how frequently each algorithm was able to correctly predict the topology of the experimentally measured structure, in which the topology of a structure is defined by its graph. We found for our algorithm that the experimental topology is within the top 1, 5, and 10 topologies at frequencies of 49, 65, and 70% for non-pseudoknotted structures, and 34, 59, and 62% for pseudoknotted, demonstrating a sharp increase between top 1 and top 5 and a plateau between top 5 and top 10.

Considering whether an algorithm correctly predicts the full topology can lead to errors arising from small variations in structure. For example, the opening of a single bond on the edge of a stem can lead to a different topology as we have defined it, if that stem includes one of the ends of the molecule. To arrive at a per-base measure of topology, we consider for each bond along the RNA backbone to which of the minimal graphs of Fig. S2 it belongs. For example, the bond between the second and third nts of Fig. 3 *a* belong to a stem of an open-net-2a graph. We then measure for each sequence the fraction of correct per-base topology predictions made by each algorithm's predicted MFE structure. We find that our algorithm averages an 76% per-base topology prediction accuracy for non-pseudoknotted sequences and a 49% accuracy for pseudoknotted.

Finally, we compare how frequently each algorithm predicts an MFE structure containing a pseudoknot. Our algorithm correctly predicted 174/235 pseudoknots among the pseudoknotted cases, far more than any other algorithm tested. However, it also erroneously predicted 35/186 incorrect pseudoknots among the non-pseudoknotted cases.

For each of these metrics, the success of our algorithm is dependent on the loop entropy model. If we set all loop entropies to zero, our algorithm's predictive power plummets (see Fig. 4, dark green). This is especially true for the prediction of non-pseudoknotted structures because removing the loop entropy term leads the algorithm to erroneously predict that 88% of these would form pseudoknots.

Our algorithm also provides the probability of folding into a pseudoknotted structure for each sequence. These data for the 421 sequences tested are presented in Fig. 5. Each data point represents a different sequence and the total

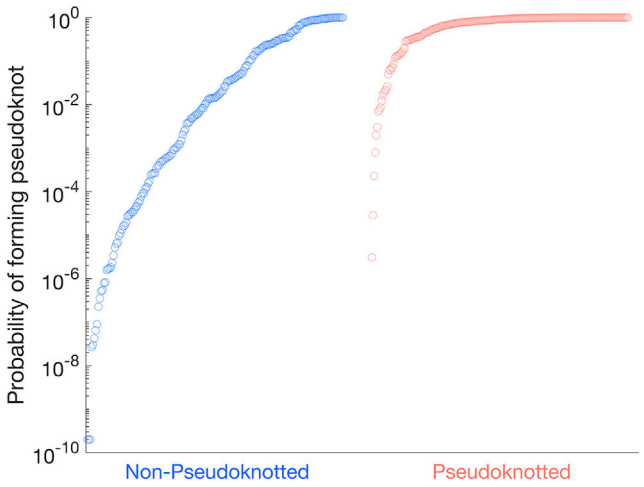

FIGURE 5 Probability of folding into a pseudoknot. The predicted probability of each of the 421 sequences tested folding into a pseudoknot is presented. Of these sequences, 186 were experimentally found not to form pseudoknots (*blue*) and 235 were found to form pseudoknots (*red*). Our algorithm successfully predicts pseudoknots forming in the latter category far more frequently than in the former. For figure clarity, a lower bound of pseudoknot probability was set at 2×10^{-10}. To see this figure in color, go online.

probability calculated of that sequence folding into a pseudoknotted structure. For figure clarity, a lower bound of pseudoknot probability was set at 2×10^{-10}.

The algorithm's predictions for the six longest RNA sequences less than 89 nts in length from the PseudoBase++ database are presented in Fig. 6. We considered only those sequences whose structure was directly supported by experiments and which could be decomposed into the minimal topologies shown in Fig. S2. We display the experimental structure (green background) alongside the MFE predicted structure (light blue background) and the top six predicted topologies (out of several hundred, depending on the sequence; dark blue) in which the experimental topology is highlighted (purple). RNA secondary structure was plotted using the PseudoViewer package (90). Our results demonstrate successful predictions even for long pseudoknotted sequences, especially in terms of the predicted topology. Detailed methods are provided in Supporting Materials and Methods.

CONCLUSIONS

The accurate prediction of the ensemble of secondary structures explored by an RNA or DNA molecule has played a major role in shaping modern molecular biology and DNA nanotechnology over the past several decades. In this work, we showed that the modern ubiquity of extremely powerful computers can be used alongside novel polymer physics techniques to completely enumerate and solve for the free energy landscape of an RNA molecule including complex pseudoknots. This exponential time algorithm can be used to tackle even relatively long (~80 nts) RNA sequences and, aside from the enumeration procedure (which is relatively fast compared to the free energy calculation for long sequences; see Figs. S4 and S6), is easily parallelizable.

Remarkably, the entropy model discussed in this work requires only two parameters—orders of magnitude fewer than other current algorithms—corresponding to clearly

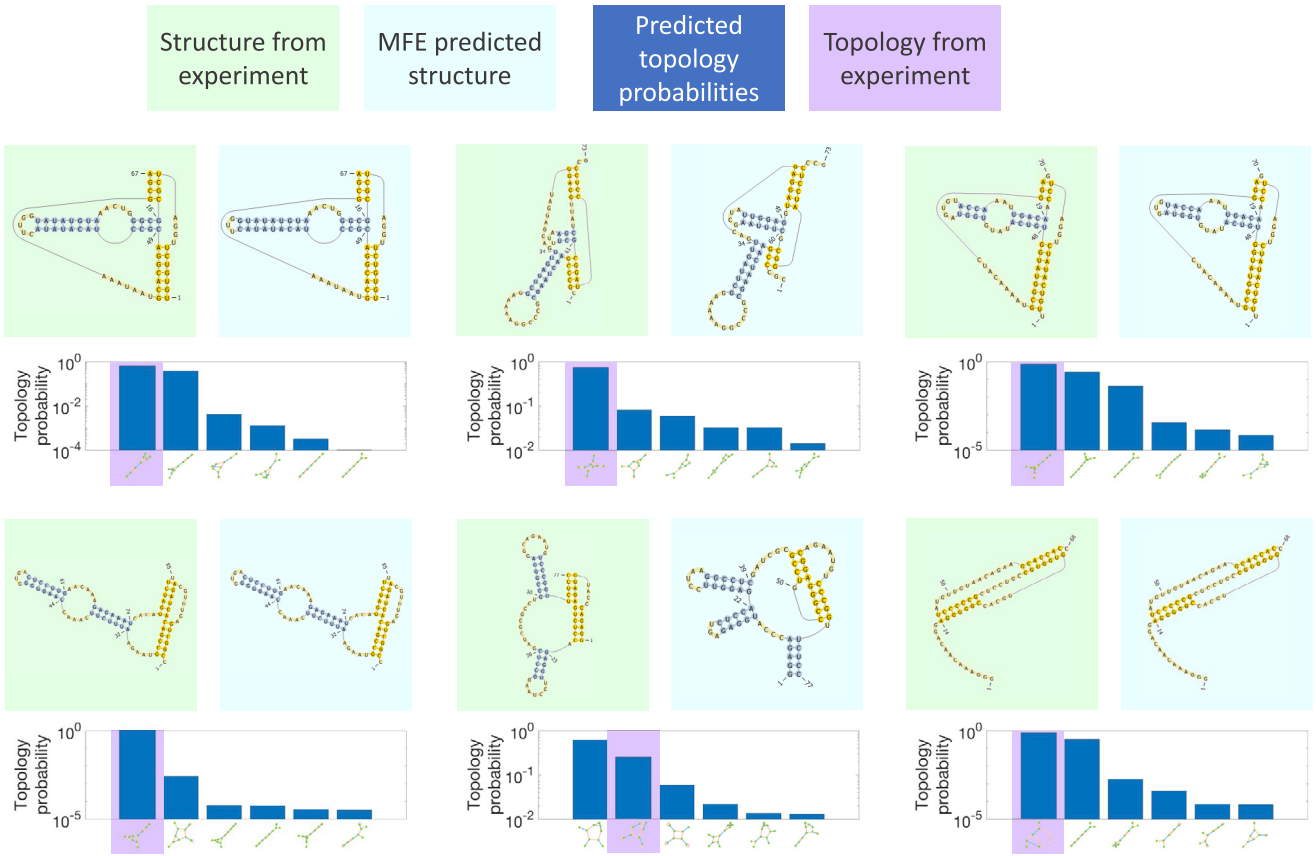

FIGURE 6 Comparison to experiments for long sequences. Six long sequences were chosen from the PseudoBase++ database as described in the main text. The sequences are fragments derived from the following (starting from the *top left* and moving across): tobacco mosaic virus (124–126), *Bacillus subtilis* (127), tobacco mild green mosaic virus (125,128), *Bacillus subtilis* (129), Giardiavirus (130), and Visna-Maedi virus (131). The experimental structures are supported by (numbering the sequences in the same order) sequence comparison (1–4,6), structure probing (1,3,5,6), mutagenesis (2,4–6), three-dimensional modeling (1), and NMR (6). We show the experimental structure (*green background*) and the MFE-predicted structure (*light blue background*) plotted using the PseudoViewer software (90). We also display the top six topologies (out of several hundred, depending on the particular sequence) and their respective predicted probabilities, with the topology corresponding to the experimental structure highlighted in purple. Overall, our results demonstrate successful predictions even for these long pseudoknotted sequences, especially in terms of the predicted topology. To see this figure in color, go online.

measurable physical quantities. Despite this and despite the fact that all parameters used in our model were derived using experiments on non-pseudoknotted RNA, our algorithm is more successful in predicting pseudoknotted structures than any of the other algorithms tested and on par with all predictors tested in predicting non-pseudoknotted structures on a benchmark data set of sequences of length $\leq$80 nts. The success of our algorithm is particularly notable given that the entropy model developed in this work can be used to address any RNA secondary structure, regardless of complexity. Given these results, we expect that more accurate entropy models can be formulated by building atop the framework presented here and have highlighted several avenues for improvement.

Although we have not done so in this work, we expect that our results can be further improved by optimizing the nearest-neighbor parameters, given the entropy model presented here.

The algorithm presented here can also be easily generalized to probe multiple interacting strands (see discussion in Supporting Materials and Methods). The sequences considered can be any combination of DNA and RNA; their identities affect the nearest-neighbor parameters of the model that have been previously tabulated (91) and to a lesser extent, the two entropy parameters (b and v_s).

Our finding that the integral formulation of the entropy of arbitrary complex RNA secondary structures can be represented graphically is reminiscent of Feynman diagrams in quantum field theory. The topologies defined by these graphs can also serve as useful biological constructs to group similar RNA structures together. The depiction of RNA structure as a graph has played an important role in the prediction of RNA secondary structure (22,74,92,93) as well as in the search for novel RNAs (94,95) and the description of similarity between RNA structures (96–99), which is especially useful in the study of the effects of mutations (100,101). A common approach among these graphical depictions of RNA has been to represent loops (e.g., hairpins, internal loops, etc.) as vertices and stems as edges (94,98,99). However, this depiction of RNA does not always distinguish between pseudoknotted and non-pseudoknotted structures (94). Our approach has a similar coarse-graining effect of grouping similar structures as the same graph but explicitly distinguishes between different topologies of secondary structure and may therefore be useful in the contexts described previously. Although our approach is in many ways similar to the planar digraphs of (94), it is able to address the ambiguity present in those graphs, particularly with regards to parallel stems (see Fig. 2 of (94)).

We expect that the complete free energy landscape prediction described in this work will be useful in understanding the kinetics of RNA and DNA structure transitions, including the interactions of multiple strands (24,25,102–108). In addition to the complete energy and entropy landscapes, a complete kinetics model only needs a definition of the transition state matrix. Such a matrix can be derived from the energy and entropy landscapes directly. For example, by defining neighboring states as secondary structures differing by the opening or closing of a single basepair, the transition rate of opening a basepair is expected to be exponential in the energy difference of the two states, whereas the rate of closing a basepair is exponential in the entropy difference (24,102,109). Even for transitions between two non-pseudoknotted structures, pseudoknots often play a significant role in the transition pathway (108,110–113). Predicting the kinetics of structure transitions using this framework and determining whether such kinetics can be accurately predicted for RNA molecules of the lengths considered here, using only secondary structure considerations, will be a subject for future work.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2019.06.037.

AUTHOR CONTRIBUTIONS

All the authors designed research. O.K. carried out theoretical calculations and wrote the code. O.K. and T.C. analyzed data. All the authors wrote the article.

ACKNOWLEDGMENTS

We thank Elena Rivas, Yohai Bar Sinai, and Carl Goodrich for fruitful discussions.

This work was supported by the National Science Foundation through the Harvard Materials Research Science and Engineering Center (grant numbers DMR-1420570, DMREF grant DMR-123869) and Office of Naval Research (grant N00014-17-1-3029). This research was conducted with Government support under and awarded by Department of Defense, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate Fellowship 32 CFR 168a (O.K.). M.P.B. is an investigator of the Simons Foundation.

SUPPORTING CITATIONS

References (114–117) appear in the Supporting Material.

REFERENCES

1. Kapranov, P., J. Cheng, …, T. R. Gingeras. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science.* 316:1484–1488.
2. Okada, Y., T. Muramatsu, …, T. Tanaka. 2016. Significant impact of miRNA-target gene networks on genetics of human complex traits. *Sci. Rep.* 6:22223.
3. Sridhar, B., M. Rivas-Astroza, …, S. Zhong. 2017. Systematic mapping of RNA-chromatin interactions in vivo. *Curr. Biol.* 27:602–609.
4. Butter, F., M. Scheibe, …, M. Mann. 2009. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc. Natl. Acad. Sci. USA.* 106:10626–10631.
5. Seemann, S. E., S. M. Sunkin, …, J. Gorodkin. 2012. Transcripts with in silico predicted RNA structure are enriched everywhere in the mouse brain. *BMC Genomics.* 13:214.

6. Mercer, T. R., M. E. Dinger, and J. S. Mattick. 2009. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10:155–159.
7. McManus, M. T., and P. A. Sharp. 2002. Gene silencing in mammals by small interfering RNAs. *Nat. Rev. Genet.* 3:737–747.
8. Juliano, C., J. Wang, and H. Lin. 2011. Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. *Annu. Rev. Genet.* 45:447–469.
9. Ellington, A. D., and J. W. Szostak. 1990. In vitro selection of RNA molecules that bind specific ligands. *Nature.* 346:818–822.
10. Tuerk, C., and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science.* 249:505–510.
11. Robertson, D. L., and G. F. Joyce. 1990. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature.* 344:467–468.
12. Olea, C., and G. F. Joyce. 2016. Real-time detection of a self-replicating RNA Enzyme. *Molecules.* 21:1–12.
13. Nowakowski, J., and I. Tinoco. 1997. RNA structure and stability. *Semin. Virol.* 8:153–165.
14. Batey, R. T., R. P. Rambo, and J. A. Doudna. 1999. Tertiary motifs in RNA structure and folding. *Angew. Chem. Int.Engl.* 38:2326–2343.
15. Montange, R. K., and R. T. Batey. 2008. Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* 37:117–133.
16. Ilyinskii, P. O., T. Schmidt, …, A. M. Shneider. 2009. Importance of mRNA secondary structural elements for the expression of influenza virus genes. *OMICS.* 13:421–430.
17. Poot, R. A., N. V. Tsareva, …, J. van Duin. 1997. RNA folding kinetics regulates translation of phage MS2 maturation gene. *Proc. Natl. Acad. Sci. USA.* 94:10110–10115.
18. Mathews, D. H., and D. H. Turner. 2006. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* 16:270–278.
19. Hofacker, I. L. 2014. 4. Energy-directed RNA structure prediction. *In* RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. J. Gorodkin and W. L. Ruzzo, eds. Humana Press, pp. 71–84.
20. Pipas, J. M., and J. E. McMahon. 1975. Method for predicting RNA secondary structure. *Proc. Natl. Acad. Sci. USA.* 72:2017–2021.
21. Bleckley, S., J. W. Stone, and S. J. Schroeder. 2012. Crumple: a method for complete enumeration of all possible pseudoknot-free RNA secondary structures. *PLoS One.* 7:e52414.
22. Bon, M., and H. Orland. 2011. TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res.* 39:e93.
23. Bon, M., C. Micheletti, and H. Orland. 2013. McGenus: a Monte Carlo algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res.* 41:1895–1900.
24. Zhang, W., and S. J. Chen. 2002. RNA hairpin-folding kinetics. *Proc. Natl. Acad. Sci. USA.* 99:1931–1936.
25. Cao, S., and S. J. Chen. 2007. Biphasic folding kinetics of RNA pseudoknots and telomerase RNA activity. *J. Mol. Biol.* 367:909–924.
26. Waterman, M. S. 1978. Secondary structure of single-stranded nucleic acidst. *In* Studies in Foundations and Combinatorics, Advances in Mathematics Supplementary Studies. G. C. Rota, ed. Academic Press, pp. 167–212.
27. Waterman, M. S., and T. F. Smith. 1986. Rapid dynamic programming algorithms for RNA secondary structure. *Adv. Appl. Math.* 7:455–464.
28. Nussinov, R., G. Pieczenik, …, D. J. Kleitman. 1978. Algorithms for loop matchings. *SIAM J. Appl. Math.* 35:68–82.
29. Zuker, M., and P. Stiegler. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9:133–148.
30. Serra, M. J., and D. H. Turner. 1995. Predicting thermodynamic properties of RNA. *Methods Enzymol.* 259:242–261.
31. Hajdin, C. E., S. Bellaousov, …, K. M. Weeks. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. USA.* 110:5498–5503.
32. Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406–3415.
33. Hofacker, I. L., W. Fontana, …, P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125:167–188.
34. Lyngsø, R. B., and C. N. S. Pedersen. 2000. Pseudoknots in RNA secondary structures. *In* Proceedings of the fourth annual international Conference on Computational Molecular Biology. ACM, pp. 201–209.
35. Lyngsø, R. B., and C. N. Pedersen. 2000. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* 7:409–427.
36. Liu, B., D. H. Mathews, and D. H. Turner. 2010. RNA pseudoknots: folding and finding. *F1000 Biol. Rep.* 2:8.
37. Isambert, H., and E. D. Siggia. 2000. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. USA.* 97:6515–6520.
38. Ruan, J., G. D. Stormo, and W. Zhang. 2004. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics.* 20:58–66.
39. Ren, J., B. Rastegari, …, H. H. Hoos. 2005. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA.* 11:1494–1504.
40. Bellaousov, S., and D. H. Mathews. 2010. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA.* 16:1870–1880.
41. Sato, K., Y. Kato, …, K. Asai. 2011. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics.* 27:i85–i93.
42. Jabbari, H., A. Condon, and S. Zhao. 2008. Novel and efficient RNA secondary structure prediction using hierarchical folding. *J. Comput. Biol.* 15:139–163.
43. Sperschneider, J., A. Datta, and M. J. Wise. 2011. Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins. *RNA.* 17:27–38.
44. Rivas, E., and S. R. Eddy. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285:2053–2068.
45. Uemura, Y., A. Hasegawa, …, T. Yokomori. 1999. Tree adjoining grammars for RNA structure prediction. *Theor. Comput. Sci.* 210:277–303.
46. Akutsu, T. 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.* 104:45–62.
47. Condon, A., B. Davy, …, F. Tarrant. 2004. Classifying RNA pseudoknotted structures. *Theor. Comput. Sci.* 320:35–50.
48. Dirks, R. M., and N. A. Pierce. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* 24:1664–1677.
49. Reeder, J., and R. Giegerich. 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics.* 5:104.
50. Cao, S., and S. J. Chen. 2006. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.* 34:2634–2652.
51. Cao, S., and S. J. Chen. 2009. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA.* 15:696–706.
52. Tinoco, I., Jr., and C. Bustamante. 1999. How RNA folds. *J. Mol. Biol.* 293:271–281.
53. van Batenburg, F. H., A. P. Gultyaev, …, J. Oliehoek. 2000. PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res.* 28:201–204.
54. Wyatt, J. R., J. D. Puglisi, and I. Tinoco, Jr. 1990. RNA pseudoknots. Stability and loop size requirements. *J. Mol. Biol.* 214:455–470.

55. Gluick, T. C., and D. E. Draper. 1994. Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.* 241:246–262.

56. Liu, B., N. Shankar, and D. H. Turner. 2010. Fluorescence competition assay measurements of free energy changes for RNA pseudoknots. *Biochemistry.* 49:623–634.

57. Qiu, H., K. Kaluarachchi, …, D. P. Giedroc. 1996. Thermodynamics of folding of the RNA pseudoknot of the T4 gene 32 autoregulatory messenger RNA. *Biochemistry.* 35:4176–4186.

58. Aalberts, D. P., and N. O. Hodas. 2005. Asymmetry in RNA pseudoknots: observation and theory. *Nucleic Acids Res.* 33:2210–2214.

59. Lucas, A., and K. A. Dill. 2003. Statistical mechanics of pseudoknot polymers. *J. Chem. Phys.* 119:2414–2421.

60. Xayaphoummine, A., T. Bucher, and H. Isambert. 2005. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.* 33:W605–W610.

61. Chen, H. L., A. Condon, and H. Jabbari. 2009. An O(n(5)) algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. *J. Comput. Biol.* 16:803–815.

62. Gregorian, R. S., Jr., and D. M. Crothers. 1995. Determinants of RNA hairpin loop-loop complex stability. *J. Mol. Biol.* 248:968–984.

63. Andronescu, M. S., C. Pop, and A. E. Condon. 2010. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA.* 16:26–42.

64. Andronescu, M., V. Bereg, …, A. Condon. 2008. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics.* 9:340.

65. Taufer, M., A. Licon, …, M. Y. Leung. 2009. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic Acids Res.* 37:D127–D135.

66. Puton, T., L. P. Kozlowski, …, J. M. Bujnicki. 2013. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.* 41:4307–4323.

67. Turner, D. H., and D. H. Mathews. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* 38:D280–D282.

68. Chirikjian, G. S. 2011. Modeling loop entropy. *Methods Enzymol.* 487:99–132.

69. Turner, D. H. 2000. 8. Conformational changes. *In* Nucleic Acids: Structures, Properties, and Functions. V. A. Bloomfield, D. M. Crothers, and I. Tinoco, eds. University Science Books, pp. 271–291.

70. Jacobson, H., and W. H. Stockmayer. 1950. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* 18:1600–1606.

71. Mathews, D. H., M. D. Disney, …, D. H. Turner. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA.* 101:7287–7292.

72. Lu, Z. J., D. H. Turner, and D. H. Mathews. 2006. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res.* 34:4912–4924.

73. Abels, J. A., F. Moreno-Herrero, …, N. H. Dekker. 2005. Single-molecule measurements of the persistence length of double-stranded RNA. *Biophys. J.* 88:2737–2744.

74. Orland, H., and A. Zee. 2002. RNA folding and large N matrix theory. *Nucl. Phys. B.* 620:456–476.

75. Gago, S., M. De la Peña, and R. Flores. 2005. A kissing-loop interaction in a hammerhead viroid RNA critical for its in vitro folding and in vivo viability. *RNA.* 11:1073–1083.

76. Chang, K. Y., and I. Tinoco, Jr. 1997. The structure of an RNA "kissing" hairpin complex of the HIV TAR hairpin loop and its complement. *J. Mol. Biol.* 269:52–66.

77. Melchers, W. J., J. G. Hoenderop, …, J. M. Galama. 1997. Kissing of the two predominant hairpin loops in the coxsackie B virus 3′ untranslated region is the essential structural feature of the origin of replication required for negative-strand RNA synthesis. *J. Virol.* 71:686–696.

78. Verheije, M. H., R. C. Olsthoorn, …, J. J. Meulenberg. 2002. Kissing interaction between 3′ noncoding and coding sequences is essential for porcine arterivirus RNA replication. *J. Virol.* 76:1521–1526.

79. Friebe, P., J. Boudet, …, R. Bartenschlager. 2005. Kissing-loop interaction in the 3′ end of the hepatitis C virus genome essential for RNA replication. *J. Virol.* 79:380–392.

80. Cao, S., and S. J. Chen. 2011. Structure and stability of RNA/RNA kissing complex: with application to HIV dimerization initiation signal. *RNA.* 17:2130–2143.

81. Gultyaev, A. P., F. H. van Batenburg, and C. W. Pleij. 1999. An approximation of loop free energy values of RNA H-pseudoknots. *RNA.* 5:609–617.

82. Fisher, M. E. 1966. Effect of excluded volume on phase transitions in biopolymers. *J. Chem. Phys.* 45:1469–1473.

83. Zhang, J., M. Lin, …, J. Liang. 2008. Discrete state model and accurate estimation of loop entropy of RNA secondary structures. *J. Chem. Phys.* 128:125107.

84. Aalberts, D. P., and N. Nandagopal. 2010. A two-length-scale polymer theory for RNA loop free energies and helix stacking. *RNA.* 16:1350–1355.

85. Aalberts, D. P. 2011. Loop entropy assists tertiary order: loopy stabilization of stacking motifs. *Entropy (Basel).* 13:1958–1966.

86. Studnicka, G. M., G. M. Rahn, …, W. A. Salser. 1978. Computer method for predicting the secondary structure of single-stranded RNA. *Nucleic Acids Res.* 5:3365–3387.

87. Zuker, M., and D. Sankoff. 1984. RNA secondary structures and their prediction. *Bull. Math. Biol.* 46:591–621.

88. Parvathy, V. R., S. R. Bhaumik, …, H. T. Miles. 2002. NMR structure of a parallel-stranded DNA duplex at atomic resolution. *Nucleic Acids Res.* 30:1500–1511.

89. Lu, Z. J., J. W. Gloor, and D. H. Mathews. 2009. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA.* 15:1805–1813.

90. Byun, Y., and K. Han. 2009. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics.* 25:1435–1437.

91. SantaLucia, J., Jr., and D. Hicks. 2004. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* 33:415–440.

92. Koessler, D. R., D. J. Knisley, …, T. Haynes. 2010. A predictive model for secondary RNA structure using graph theory and a neural network. *BMC Bioinformatics.* 11 (*Suppl 6*):S21.

93. Zhao, J., R. L. Malmberg, and L. Cai. 2008. Rapid ab initio prediction of RNA pseudoknots via graph tree decomposition. *J. Math. Biol.* 56:145–159.

94. Gan, H. H., S. Pasquali, and T. Schlick. 2003. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res.* 31:2926–2943.

95. Laing, C., and T. Schlick. 2011. Computational approaches to RNA structure prediction, analysis, and design. *Curr. Opin. Struct. Biol.* 21:306–318.

96. Haslinger, C., and P. F. Stadler. 1999. RNA structures with pseudoknots: graph-theoretical, combinatorial, and statistical properties. *Bull. Math. Biol.* 61:437–467.

97. Bermúdez, C. I., E. E. Daza, and E. Andrade. 1999. Characterization and comparison of Escherichia coli transfer RNAs by graph theory based on secondary structure. *J. Theor. Biol.* 197:193–205.

98. Benedetti, G., and S. Morosetti. 1996. A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. *Biophys. Chem.* 59:179–184.

99. Le, S. Y., R. Nussinov, and J. V. Maizel. 1989. Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.* 22:461–473.

100. Fontana, W., and P. Schuster. 1998. Continuity in evolution: on the nature of transitions. *Science.* 280:1451–1455.

101. Ancel, L. W., and W. Fontana. 2000. Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.* 288:242–283.

102. Zhang, W., and S. J. Chen. 2006. Exploring the complex folding kinetics of RNA hairpins: I. General folding kinetics analysis. *Biophys. J.* 90:765–777.

103. Flamm, C., W. Fontana, …, P. Schuster. 2000. RNA folding at elementary step resolution. *RNA.* 6:325–338.

104. Flamm, C., I. L. Hofacker, …, M. T. Wolfinger. 2002. Barrier trees of degenerate landscapes. *Z. Phys. Chem.* 216:155–173.

105. Thachuk, C., J. Manuch, …, A. Condon. 2010. An algorithm for the energy barrier problem without pseudoknots and temporary arcs. *Pac. Symp. Biocomput* https://doi.org/10.1142/9789814295291_0013.

106. Dotu, I., W. A. Lorenz, …, P. Clote. 2010. Computing folding pathways between RNA secondary structures. *Nucleic Acids Res.* 38:1711–1722.

107. Kucharík, M., I. L. Hofacker, …, J. Qin. 2014. Basin Hopping Graph: a computational framework to characterize RNA folding landscapes. *Bioinformatics.* 30:2009–2017.

108. Kucharík, M., I. L. Hofacker, …, J. Qin. 2016. Pseudoknots in RNA folding landscapes. *Bioinformatics.* 32:187–194.

109. Zhao, P., W. B. Zhang, and S. J. Chen. 2010. Predicting secondary structural folding kinetics for nucleic acids. *Biophys. J.* 98:1617–1625.

110. Xu, X., and S. J. Chen. 2012. Kinetic mechanism of conformational switch between bistable RNA hairpins. *J. Am. Chem. Soc.* 134:12499–12507.

111. Isambert, H. 2009. The jerky and knotty dynamics of RNA. *Methods.* 49:189–196.

112. Fürtig, B., P. Wenter, …, H. Schwalbe. 2007. Conformational dynamics of bistable RNAs studied by time-resolved NMR spectroscopy. *J. Am. Chem. Soc.* 129:16222–16229.

113. Höbartner, C., M. O. Ebert, …, R. Micura. 2002. RNA two-state conformation equilibria and the effect of nucleobase methylation. *Angew. Chem. Int. Ed.* 41:605–609.

114. Mammen, M., E. I. Shakhnovich, …, G. M. Whitesides. 1998. Estimating the entropic cost of self-assembly of multiparticle hydrogen-bonded aggregates based on the cyanuric acid-melamine lattice. *J. Org. Chem.* 63:3821–3830.

115. Zhou, H. X., and M. K. Gilson. 2009. Theory of free energy and entropy in noncovalent binding. *Chem. Rev.* 109:4092–4107.

116. Sugimoto, N., S. Nakano, …, M. Sasaki. 1995. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry.* 34:11211–11216.

117. Watkins, N. E., Jr., W. J. Kennelly, …, J. Santalucia, Jr. 2011. Thermodynamic contributions of single internal rA·dA, rC·dC, rG·dG and rU·dT mismatches in RNA/DNA duplexes. *Nucleic Acids Res.* 39:1894–1902.

118. Lorenz, R., S. H. Bernhart, …, I. L. Hofacker. 2011. ViennaRNA package 2.0. *Algorithms Mol. Biol.* 6:26.

119. Andronescu, M., A. Condon, …, K. P. Murphy. 2007. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics.* 23:i19–i28.

120. Do, C. B., D. A. Woods, and S. Batzoglou. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics.* 22:e90–e98.

121. Sükösd, Z., B. Knudsen, …, C. N. Pedersen. 2012. PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics.* 28:2691–2692.

122. Sato, K., M. Hamada, …, T. Mituyama. 2009. CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res.* 37:W277–W280.

123. Zakov, S., Y. Goldberg, …, M. Ziv-Ukelson. 2011. Rich parameterization improves RNA structure prediction. *J. Comput. Biol.* 18:1525–1542.

124. Rietveld, K., K. Linschooten, …, L. Bosch. 1984. The three-dimensional folding of the tRNA-like structure of tobacco mosaic virus RNA. A new building principle applied twice. *EMBO J.* 3:2613–2619.

125. Mans, R. M., C. W. Pleij, and L. Bosch. 1991. tRNA-like structures. Structure, function and evolutionary significance. *Eur. J. Biochem.* 201:303–324.

126. Felden, B., C. Florentz, …, E. Westhof. 1996. A central pseudoknotted three-way junction imposes tRNA-like mimicry and the orientation of three 5′ upstream pseudoknots in the 3′ terminus of tobacco mosaic virus RNA. *RNA.* 2:201–212.

127. Soukup, G. A. 2006. Core requirements for glmS ribozyme self-cleavage reveal a putative pseudoknot structure. *Nucleic Acids Res.* 34:968–975.

128. García-Arenal, F. 1988. Sequence and structure at the genome 3′ end of the U2-strain of tobacco mosaic virus, a histidine-accepting tobamovirus. *Virology.* 167:201–206.

129. Wilkinson, S. R., and M. D. Been. 2005. A pseudoknot in the 3′ noncore region of the glmS ribozyme enhances self-cleavage activity. *RNA.* 11:1788–1794.

130. Garlapati, S., and C. C. Wang. 2002. Identification of an essential pseudoknot in the putative downstream internal ribosome entry site in giardiavirus transcript. *RNA.* 8:601–611.

131. Pennell, S., E. Manktelow, …, I. Brierley. 2008. The stimulatory RNA of the Visna-Maedi retrovirus ribosomal frameshifting signal is an unusual pseudoknot with an interstem element. *RNA.* 14:1366–1377.

Biophysical Journal, Volume *117*

Supplemental Information

A Polymer Physics Framework for the Entropy of Arbitrary Pseudoknots

Ofer Kimchi, Tristan Cragnolini, Michael P. Brenner, and Lucy J. Colwell

A polymer physics framework for the entropy of arbitrary pseudoknots Supplementary Information

Ofer Kimchi,[1,*] Tristan Cragnolini,[2] Michael P. Brenner,[3,4] and Lucy J. Colwell[2,†]

[1]*Harvard Graduate Program in Biophysics, Harvard University, Cambridge, MA 02138*
[2]*Department of Chemistry, University of Cambridge, CB2 1EW, Cambridge, United Kingdom*
[3]*School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138*
[4]*Kavli Institute of Bionano Science and Technology, Harvard University, Cambridge, MA 02138*

The supplementary information is divided into several sections. In Sections S1 and S2 we detail our implementation of the nearest-neighbor parameters, as well as the methods used to compare our algorithm's performance to other current models. In Section S3 we discuss how our algorithm can be easily generalized to probe multiple interacting strands including any combination of DNA and RNA. In Section S4 and Fig. S1 we provide a more complete derivation of Eq. 7. In Section S5, we show how to analytically calculate the integrals in Eq. 7. In Section S6 we derive the higher-order corrections to Eq. 5.

In Fig. S2 we display all possible graphs of up to two stems and their respective RNA structures along with the integral formulation of their entropies and their evaluated forms. In Fig. S3 we discuss how our algorithm compares to state-of-the-art prediction tools (the analogue of Fig. 4) when restricting ourselves to structures allowed by the chosen constraints on our algorithm.

In Section S7A we discuss how our algorithm's properties scale with the length of the sequence for random sequences between 10 and 21 ntds in length, shown in Fig. S4. In Section S7B, we provide a mathematical discussion for why the average number of structures for a sequence of length n scales exponentially with n; the discussion corresponds to Fig. S5. We show running time and total number of secondary structure distributions for sequences in our dataset in Fig. S6, with a corresponding discussion in Section S7C.

In Fig. S7, we demonstrate that loop entropies are highly non-negligible; the magnitude of the predicted loop entropy is roughly equal to the magnitude of the total free energy of a structure. In sections S8 and S9 and in figures S8 - S10 we show the entropy calculation for pseudoknots more complex than those in Fig. S2; namely, the kissing hairpins pseudoknot and the most common pseudoknots found in our benchmark dataset. Finally, in Section S10 and Fig. S11, we demonstrate a sample free energy calculation and graph decomposition process.

S1. FURTHER METHOD DETAILS

A. Implementation of nearest-neighbor free energies

Our entropy model (described in the Materials and Methods section) was used in place of the entropies of hairpin, bulge, internal, and multibranch loops and we set the enthalpy terms of these loops (aside from nearest-neighbor interactions) to zero; we did not consider mismatch-mediated coaxial stacking, symmetry penalties or penalties for specific closures of stems; and we implemented coaxial stacking terms in place of terminal mismatches or dangling ends whenever two stems in multibranch loops are directly adjacent.

B. Comparison with other prediction tools

In order to compare the sensitivity and PPV of different prediction tools, we considered the base pairs present in the experimental structure and in each algorithm's MFE structure. Base pairs present in both were labeled as true positives (TP), base pairs present in the predicted algorithm were labeled as false positives (FP) and those present in the experimental structure but not the predicted MFE structure were labeled as false negatives (FN). In order to compare different metrics we use the summary statistics of sensitivity ($TP/TP+FN$) and PPV ($TP/TP+FP$). PPV is a more useful metric for RNA structure prediction algorithms than specificity because the definition of true negatives is unclear when considering base pairs.

*Electronic address: okimchi@g.harvard.edu
†Electronic address: ljc37@cam.ac.uk

The sequences tested were downloaded from the Pseudobase++, RNAStrand, and CompaRNA PDB databases. We constrained database searches to return results only for sequences of length ≤ 80 ntds. We further restricted the search of the RNAStrand database to only include sequences where all nucleotides were known, and to not include fragments, multiple strands, or duplicates. We removed all sequences that had hairpins of under 3 ntds. Finally, we compared the sequence similarity of the sequences derived and kept only sequences with ≥ 0.2 Jukes-Cantor sequence dissimilary measured using the MatLab command `seqpdist`, which aligns sequences using the Needleman-Wunsch algorithm with the *NUC44* scoring matrix. The Jukes-Cantor distance between two sequences is defined as

$$d_{\mathrm{JC}} = -\frac{3}{4}\log\left(1-\frac{4p}{3}\right) \tag{S1}$$

where p is the fraction of sites which differ between the sequences after they have been aligned. By imposing $d_{\mathrm{JC}} \geq 0.2$ we impose a constraint that $p > 0.17$.

We assumed $T = 300K$ for all predictions.

In order to speed up computation for longer sequences, we set the parameter m describing the minimum number of consecutive base pairs in a stem to the minimum value it can take such that the total number of possible stems is less than 150. This latter parameter was chosen arbitrarily and is likely not optimized; however, changing it to 200 had no significant effect (see data in Supplementary Table 1). Setting the maximum total number of possible stems to 150 resulted in $m = 1$ for 22% of the sequences, $m = 2$ for 33% of the sequences, $m = 3$ for 23%, $m = 4$ for 20%, and $m = 5$ for nine sequences. Changing the maximum total number of possible stems to 200 resulted in $m = 1$ for 34% sequences, $m = 2$ for 29% of sequences, $m = 3$ for 22%, $m = 4$ for 15%, and $m = 5$ for one sequence.

Our algorithm can enumerate and calculate the entropies of both parallel and antiparallel stems. (An antiparallel stem is a list of consecutive base pairs of the form $[i \cdot j,\ (i+1)\cdot(j-1),\ (i+2)\cdot(j-2)...]$, while a parallel stem has the form $[i \cdot j,\ (i+1)\cdot(j+1),\ (i+2)\cdot(j+2)...]$.) Parallel stems are disallowed in non-pseudoknotted structures, and are stabilized at certain pH levels. We disallowed parallel stems in our calculations.

As part of the enumeration procedure, we created a compatibility matrix $C_{p,q}$ detailing the compatibility of structures p and q (structures p and q are compatible if they do not share any nucleotides). In practice, since there are some structures whose entropies we have not analytically derived, we found it useful to also construct three- and four-dimensional matrices C_3 and C_4 which define three- and four-way compatibility, in order to exclude most such structures at this stage.

In order to compare topologies, we measure whether the eigenvalue spectra of the two matrices defining the bonds between each node are equal (two matrices are needed because there are two types of bonds). This method is guaranteed to correctly identify graph isomorphisms in all cases but may have false positives. We have found no evidence of false positives in all cases tested (compared against the MatLab `isisomorphic` command).

For the analysis in Fig. 6 we also set $m > 1$ to speed up computation. Starting from the top left and going across, we set $m = (4, 3, 3, 4, 4, 4)$. We also disallowed parallel stems in order to speed up the computation.

S2. PREDICTION TOOL PARAMETERS

To compare our results, we used the implementation of other prediction tools, when provided by the authors. In most cases, program options have been left to their default value. We list below some of the more important options.

- RNAFold: Temperature: 37 C
- Andronescu: Temperature: 37 C
- Mfold: Temperature: 37 C
- CONTRAFold: $\gamma = 6$
- PPfold: N/A
- Centroidfold: $\gamma = 6$
- ContextFold: Model: "trained/StHighCoHigh.model"
- HotKnots DP/RE/CC: energy model DP/RE/CC
- ProbKnot: 1 iteration

- pknots: N/A
- RNAPKplex: Temperature: 37 C
- ILM: N/A

S3. PROBING MULTIPLE INTERACTING STRANDS

The algorithm presented here can also be easily generalized to probe multiple interacting strands, using only one further parameter which has been previously studied to define the free energy cost of forming a duplex [1, 2]. Following Ref. [3] we concatenate the two (or more) sequences, separated by a number of inert nucleotides which serve as a placeholder and which are removed before free energy calculations are implemented.

The algorithm described here can be equally well-applied to DNA strands by using the parameter sets from the SantaLucia laboratory [4] . In addition, our algorithm can probe DNA-RNA bonds using the parameter sets from Refs. [5, 6], and interpolating between the DNA and RNA cases for those parameters that have not yet been tabulated from experimental data. The inclusion of DNA strands may require slight modification to the two entropy parameters (b and v_s) which are based on data from RNA experiments.

S4. DERIVING EQ. 7

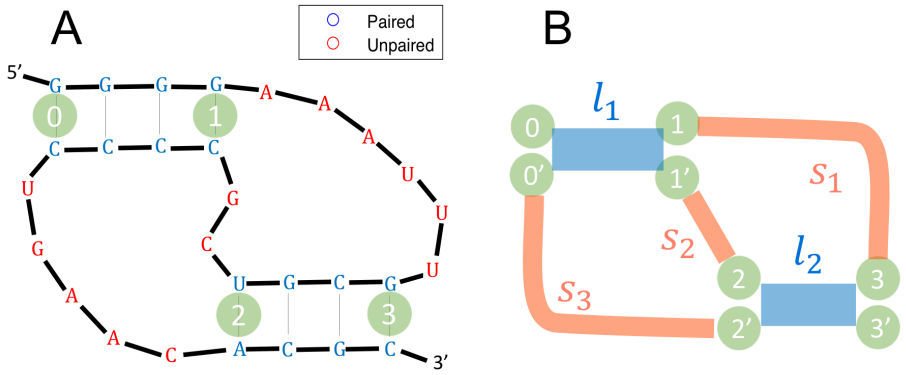

FIG. S1: **A preliminary description of an H-type pseudoknot. A:** An instance of the canonical H-type pseudoknot, reprinted from Fig. 3. **B:** A preliminary version of the graph representing its entropy. In Sec. S4 we demonstrate that this graph is equivalent to that shown in Fig. 3A.

In this section we more fully detail the steps leading to Eq. 7, the entropy of the RNA structure depicted in Fig. S1A.

We start by treating each nucleotide as its own node, subject to the constraint that the distance between nucleotides is given by $a = 0.33$ nm. Writing such an expression is cumbersome, but because of the property of $P_s(\vec{r})$ that $\int P_x(\vec{r_1})P_y(\vec{r_2} - \vec{r_1})d\vec{r_1} = P_{x+y}(\vec{r_2})$, we can simply integrate over all nodes not at the edges of stems.

The full expression for the entropy of this graph is thus given by

$$e^{\Delta S/k_B} = \int d\vec{r_{0'}} \int d\vec{r_1} \int d\vec{r_{1'}} \int d\vec{r_2} \int d\vec{r_{2'}} \int d\vec{r_3} \int d\vec{r_{3'}}\; q(\vec{r_{0'}})\; q(\vec{r_2} - \vec{r_{2'}}) \times$$
$$\delta^3(|\vec{r_1}| - l_1)\; \delta^3(|\vec{r_1} - (\vec{r_{1'}} - \vec{r_{0'}})|)\; \delta^3(|\vec{r_3} - \vec{r_2}| - l_2)\; \delta^3(|\vec{r_3} - \vec{r_2} - (\vec{r_{3'}} - \vec{r_{2'}})|)\; P_{s_1}(\vec{r_3} - \vec{r_1})\; P_{s_2}(\vec{r_2} - \vec{r_{1'}})\; P_{s_3}(\vec{r_{2'}} - \vec{r_{0'}})$$

which is depicted graphically in Fig. S1B. We are using $\delta^3(|x| - a)$ to signify

$$\delta^3(|\vec{x}| - a) = \frac{\delta(|\vec{x}| - a)}{4\pi a^2}; \qquad \int d\vec{x}\; \delta^3(|\vec{x}| - a) = 1. \tag{S2}$$

$\delta^3(|x| - a)$, like $P_s(\vec{r})$, has units of inverse volume.

Vectors are defined relative to the origin where node 0 is placed (i.e. $|\vec{r_0}| = 0$). There is no integration over $\vec{r_0}$ because such an integral would cancel out with the corresponding term in S_{free}, and thus disappear in the formula for ΔS.

$q(\vec{r})$ is defined as the probability of a nucleotide located a vector $\vec{r}$ from the origin to be bonded to a nucleotide located at the origin (assuming the two nucleotides are complementary). If following Ref. [7] we wish to include an upper bound for the bond length, r_s, $q(\vec{r})$ becomes a Heaviside Θ function. Integration over q leads to the definition of v_s: $v_s = \int d\vec{r} q(\vec{r})$.

Only two factors of q are present, as opposed to one factor for each base pair in the structure, because we take the entropy of stems into account separately. For this expression, we treat stems as rigid rods; while the rods have variable and finite width (corresponding to the property that nucleotides do not need to be at a precise separation in order to bond), they cannot be thicker on one end than the other, since including such possibilities would overcount the entropy of the stem. Our expression thereby has the property that it is invariant if we also integrate over two nodes representing two arbitrary base pairs (say, one on the stem between node 0 and node 1, and one between nodes $0'$ and $1'$). The choice of which bonded nodes on each stem to put in the argument of q is arbitrary, but there is only one bonded node (and therefore one q term) for each stem.

We make progress by assuming that because of the q terms and delta functions, nodes representing nucleotides which are bonded are located close enough that the vector $\vec{r}$ between them can be approximated as having zero length within the context of the terms $P_s(\vec{r})$.

We therefore approximate our formula as

$$e^{\Delta S/k_B} = \int d\vec{r_{0'}} \int d\vec{r_1} \int d\vec{r_{1'}} \int d\vec{r_2} \int d\vec{r_{2'}} \int d\vec{r_3} \int d\vec{r_{3'}}\ q(\vec{r_{0'}})\ q(\vec{r_2} - \vec{r_{2'}}) \delta^3(|\vec{r_1} - (\vec{r_{1'}} - \vec{r_{0'}})|) \times$$
$$\delta^3(|\vec{r_3} - \vec{r_2} - (\vec{r_{3'}} - \vec{r_{2'}})|)\ \delta^3(|\vec{r_1}| - l_1)\ \delta^3(|\vec{r_3} - \vec{r_2}| - l_2)\ P_{s_1}(\vec{r_3} - \vec{r_1})\ P_{s_2}(\vec{r_2} - \vec{r_1})\ P_{s_3}(\vec{r_2})$$

By employing transformations as in Section S5 (e.g. $\vec{r'} \equiv \vec{r_{1'}} - \vec{r_{0'}}$), the four integrals over the primed nodes become two integrals over delta functions (which give unity) and two over the q terms. The latter two become two factors of v_s, and we arrive at Eq. 7.

S5. PERFORMING THE GAUSSIAN INTEGRALS

The method of performing the Gaussian integrals of Eq. 7 can be generally applied to the calculation of the entropies of other pseudoknots, and so we describe it in detail here.

Eq. 7 is given by

$$e^{\Delta S/k_B} = v_s^2 \int d\vec{r_1} \int d\vec{r_2} \int d\vec{r_3} \frac{\delta(|\vec{r_1}| - l_1)}{4\pi l_1^2} \frac{\delta(|\vec{r_3} - \vec{r_2}| - l_2)}{4\pi l_2^2} P_{s_1}(\vec{r_3} - \vec{r_1}) P_{s_2}(\vec{r_2} - \vec{r_1}) P_{s_3}(\vec{r_2})$$

We start by utilizing our approximation that the integrals extend over all of space to rewrite $d\vec{r_2}d\vec{r_3}$ as $d\vec{r_2}d(\vec{r_3} - \vec{r_2})$, and we rewrite all instances of $\vec{r_3}$ as $(\vec{r_3} - \vec{r_2}) + \vec{r_2}$.

$$e^{\Delta S/k_B} = v_s^2 \prod_{i=1}^{3} \left(\frac{\gamma}{\pi s_i}\right)^{3/2} \int d\vec{r_1} \frac{\delta(|\vec{r_1}| - l_1)}{4\pi l_1^2} \int d\vec{r_2} \int d(\vec{r_3} - \vec{r_2}) \frac{\delta(|\vec{r_3} - \vec{r_2}| - l_2)}{4\pi l_2^2} \times e^{\gamma\left[-\left(\frac{(\vec{r_3} - \vec{r_2})^2}{s_1}\right) - (\vec{r_2} - \vec{r_1})^2\left(\frac{1}{s_1} + \frac{1}{s_2}\right) - \frac{r_2^2}{s_3} - \frac{2}{s_1}(\vec{r_3} - \vec{r_2}) \cdot (\vec{r_2} - \vec{r_1})\right]},$$

where for notational convenience have defined a parameter $\gamma = 3/2b$.[1].

To do the $(\vec{r_3} - \vec{r_2})$ integral, we convert to polar coordinates such that $(\vec{r_3} - \vec{r_2}) \cdot (\vec{r_2} - \vec{r_1}) = |\vec{r_3} - \vec{r_2}||\vec{r_2} - \vec{r_1}| \cos\theta$. Performing the integral yields

$$e^{\Delta S/k_B} = v_s^2 \prod_{i=1}^{3} \left(\frac{\gamma}{\pi s_i}\right)^{3/2} \frac{e^{-\gamma l_2^2/s_1}}{2} \int d\vec{r_1} \frac{\delta(|\vec{r_1}| - l_1)}{4\pi l_1^2} \int d\vec{r_2}\ e^{\gamma\left[-\frac{r_2^2}{s_3} - (\vec{r_2} - \vec{r_1})^2\left(\frac{1}{s_1} + \frac{1}{s_2}\right)\right]} \left(\frac{e^{(2\gamma l_2 |\vec{r_2} - \vec{r_1}|/s_1)} - e^{(-2\gamma l_2 |\vec{r_2} - \vec{r_1}|/s_1)}}{2\gamma l_2 |\vec{r_2} - \vec{r_1}|/s_1}\right).$$

We now use the same trick from before to rewrite $d\vec{r_2}$ as $d(\vec{r_2} - \vec{r_1})$, and rewrite each instance of $\vec{r_2}$ as $(\vec{r_2} - \vec{r_1}) + \vec{r_1}$. As before, $(\vec{r_2} - \vec{r_1}) \cdot \vec{r_1}$ becomes $|\vec{r_2} - \vec{r_1}||\vec{r_1}| \cos\theta$. Denoting $(\vec{r_2} - \vec{r_1})$ as $\vec{r}$ and doing the integral over r_1 after performing this transformation yields

$$e^{\Delta S/k_B} = v_s^2 \prod_{i=1}^{3} \left(\frac{\gamma}{\pi s_i}\right)^{3/2} \frac{e^{-\gamma\left(\frac{l_2^2}{s_1} + \frac{l_1^2}{s_3}\right)}}{2} \int_0^\infty dr\ r^2\ e^{-\gamma r^2\left(\frac{1}{s_1} + \frac{1}{s_2} + \frac{1}{s_3}\right)} \left(\frac{e^{(2\gamma l_2 r/s_1)} - e^{(-2\gamma l_2 r/s_1)}}{2\gamma l_2 r/s_1}\right) \int_{-1}^{1} d\cos(\theta) e^{-2\gamma \frac{l_1 r}{s_3} \cos(\theta)}.$$

Finally, we perform the integrals remaining to arrive at

$$e^{\Delta S/k_B} = \frac{v_s^2 \gamma^2 \exp\left(-\frac{\gamma\left(l_1^2(s_1+s_2) + l_2^2(s_2+s_3)\right)}{s_1 s_2 + s_1 s_3 + s_2 s_3}\right)}{2\pi^3 l_1 l_2 s_2 \sqrt{s_1 s_2 + s_1 s_3 + s_2 s_3}} \times \sinh\left(\frac{2\gamma l_1 l_2 s_2}{s_1 s_2 + s_1 s_3 + s_2 s_3}\right)$$

where sinh is the hyperbolic sine function. This formula is equivalent to the one presented without proof in Ref. [8].

It can be easily verified (see Fig. S2) that the entropy of an open net can be calculated given the formula for the corresponding closed net, which has an extra single-bond of length s_i, through multiplication by $(\gamma/\pi s_i)^{-3/2}$ and taking the limit $s_i \to \infty$. The formula for the "very open net 2", which is identical to any of the open nets that have two stems after removing the edge corresponding to s_2, can thus be calculated to be

$$e^{\Delta S_{\text{very-open-net-2}}/k_B} = \frac{v_s^2 \gamma^{1/2}}{2\pi^{3/2} l_1 l_2 \sqrt{s_1 + s_2}} \sinh\left(\frac{2\gamma l_1 l_2}{s_1 + s_2}\right) \exp\left(-\gamma \frac{l_1^2 + l_2^2}{s_1 + s_2}\right)$$

where we've labeled the two single-stranded edges' lengths to be s_1 and s_2. This net can form only from two strands binding to one another, as opposed to some of the other nets shown in Fig. S2 which describe two strands bound or one strand with parallel stems.

---

[1] The parameter γ was called β in Refs. [8] and [9]

S6. HIGHER-ORDER CORRECTIONS TO ENTROPY

Eq. 5 , which gives the probability of a random walk of length s to have end-to-end distance $\vec{R}$, is valid only in the limit of $R \gg b$ (where we've denoted $R \equiv |\vec{R}|$). For shorter walks, the Central Limit Theorem no longer holds. In this section, we show a systematic approach to deriving higher-order corrections to the probability distribution given by Eq. 5. The approach taken here is based on a textbook by Ariel Amir (to be published).

We consider n steps in three dimensions, where each step is taken to be of length b with equal probabilities in all directions. Thus, $s = nb$. The probability distribution for where a walker will be after $n = 1$ steps is given by $P_{n=1}(\vec{R}) \equiv \delta(|R| - b)/4\pi b^2$. After two steps, the probability distribution for where the walker will be is given by

$$P_2(\vec{R}) = \int d\vec{R_1} P_1(\vec{R_1}) P_1(\vec{R} - \vec{R_1}). \tag{S3}$$

The form of Eq. S3 is that of a convolution of $P_1(\vec{R})$ with itself. In order to iterate many convolutions easily, we move to Fourier space, since the Fourier transform of a convolution is the product of Fourier transforms. Fourier transforming $P_1(\vec{R})$ yields its characteristic function: $\hat{p}_1(\vec{\omega}) = \int\int\int_{-\infty}^{\infty} d\vec{R}\ P_1(\vec{R}) e^{i\vec{\omega}\cdot\vec{R}}$, which simplifies to

$$\hat{p}_1(\omega) = \frac{\sin(\omega b)}{\omega b} \tag{S4}$$

which only depends on $\omega \equiv |\vec{\omega}|$.

In order to iterate n convolutions in real space, we can simply take the n^{th} power of the Fourier transform, finding

$$\hat{p}_n(\omega) = \left(\sin(\omega b)/\omega b\right)^n . \tag{S5}$$

Taking the inverse Fourier transform, we find

$$P_n(\vec{R}) = \frac{2}{(2\pi)^2} \int_0^{\infty} d\omega\ \omega^2 \left(\frac{\sin(\omega b)}{\omega b}\right)^n \frac{\sin(\omega R)}{\omega R}. \tag{S6}$$

At this point, we use our assumption that n is large. This formula tends to zero for large values of ωb, and we therefore Taylor expand the sin function for small ωb. If we take only the first two terms of this series, we would arrive at Eq. 5 ; we therefore take the first three terms to get the first correction to Eq. 5 . Higher-order corrections can be found by simply taking more terms of the series. Eq. S6 thus becomes

$$P_n(\vec{R}) = \frac{2}{(2\pi)^2} \int_0^{\infty} d\omega\ \omega^2 e^{n \log\left(1 - \frac{(\omega b)^2}{6} + \frac{(\omega b)^4}{120} + \mathcal{O}(\omega b)^6\right)} \frac{\sin(\omega R)}{\omega R}$$

Next, we Taylor expand the logarithm and write the sin as a sum of exponentials. Since the two terms in the sum are identical under the exchange $\omega \to -\omega$, we combine them into one term by changing the lower limit of integration to $-\infty$.

$$P_n(\vec{R}) = \frac{1}{(2\pi)^2 iR} \int_{-\infty}^{\infty} d\omega\ \omega e^{-n\left[\frac{(\omega b)^2}{6} + \frac{(\omega b)^4}{180} + \mathcal{O}(\omega b)^6\right] + i\omega R}. \tag{S7}$$

If we didn't have the quartic term, this integral would be Gaussian and would result in Eq. 5. However, if we keep this term, the integral is no longer solvable analytically. We proceed by setting

$$e^{-n\left[\frac{(\omega b)^4}{180}\right]} = 1 - \frac{n(\omega b)^4}{180} + \mathcal{O}(\omega b)^8. \tag{S8}$$

As is apparent, the finite truncation of this series results in corrections of higher order than the truncation of the series for $\sin(\omega b)$ or of the logarithm above.

Using this series expansion, Eq. S7 becomes a Gaussian integral, which can be solved analytically to yield

$$P_n(\vec{R}) = \left(\frac{3}{2\pi sb}\right)^{3/2} e^{\left(-\frac{3R^2}{2sb}\right)} \left[1 - \frac{3\left(5s^2b^2 - 10sbR^2 + 3R^4\right)}{20s^3b}\right]. \tag{S9}$$

where we've replaced n by s/b.

One of the essential properties of $P_n(\vec{R})$ for our formalism to function is that $\int P_{n_1}(\vec{R}_1)P_{n_2}(\vec{R}_2 - \vec{R}_1)d\vec{R}_1 = P_{n_1+n_2}(\vec{R}_2)$. One can check directly that this holds for Eq. S9. Keeping only first-order correction terms, and defining $\vec{R}_{21} = \vec{R}_2 - \vec{R}_1$,

$$\begin{aligned}
&\int P_{n_1}(\vec{R}_1)P_{n_2}(\vec{R}_2 - \vec{R}_1)d\vec{R}_1 \\
&= \int d\vec{R}_1 \left(\frac{3^2}{2^2\pi s_1 s_2 b^2}\right)^{3/2} e^{\left[-\frac{3}{2b}\left(\frac{R_1^2}{s_1} + \frac{\vec{R}_{21}^2}{s_2}\right)\right]} \left[1 - \frac{3\left(5s_1^2b^2 - 10s_1bR_1^2 + 3R_1^4\right)}{20s_1^3b} - \frac{3\left(5s_2^2b^2 - 10s_2b\vec{R}_{21}^2 + 3\vec{R}_{21}^4\right)}{20s_2^3b}\right] \\
&= \left(\frac{3}{2\pi(s_1+s_2)b}\right)^{3/2} e^{\left(-\frac{3R_2^2}{2(s_1+s_2)b}\right)} \left[1 - \frac{3\left(5(s_1+s_2)^2b^2 - 10(s_1+s_2)bR_2^2 + 3R_2^4\right)}{20(s_1+s_2)^3b}\right] = P_{n_1+n_2}(\vec{R}_2).
\end{aligned}$$

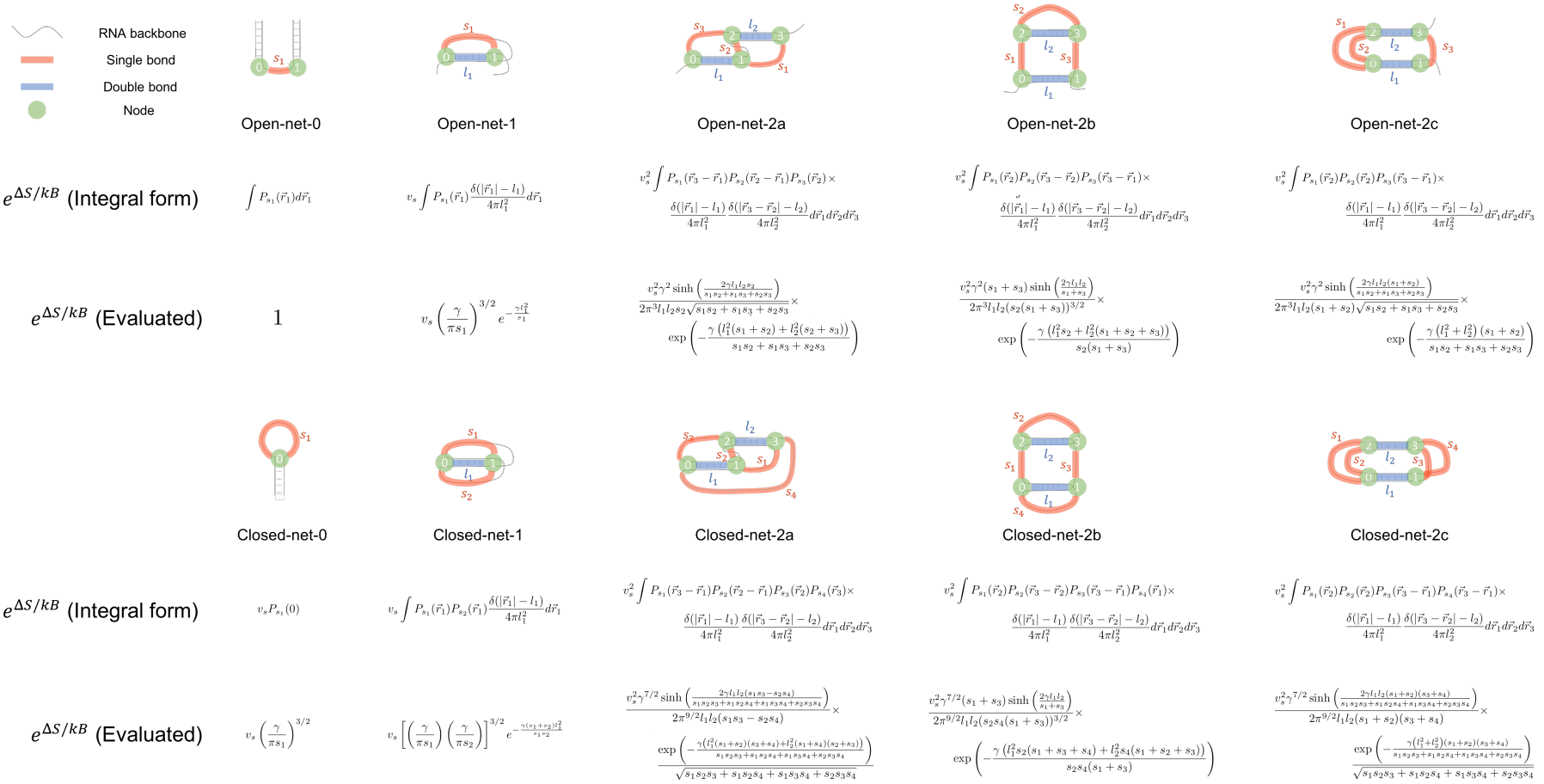

FIG. S2: **Graphs of simple RNA structures.** The 10 graphs with at most two regions of double-stranded RNA and their corresponding RNA backbones are displayed alongside integral and evaluated expressions for the entropy of each graph. Note that stems shown as parallel could be antiparallel if the system considered is comprised of more than one strand. For example, closed-net-1 is an *inter*molecular kissing hairpin complex. We do not include the "very open net 2", a bi-molecular structure that can be created by taking any of the open nets with two stems and removing the edge corresponding to s_2; the relevant calculation is described in the main text. See Fig. 2 of Ref. [9] for comparison. See Section S5 for a description of how to perform the integrals, and for a discussion of how to easily calculate the entropy of an open-net given that of the corresponding closed-net.

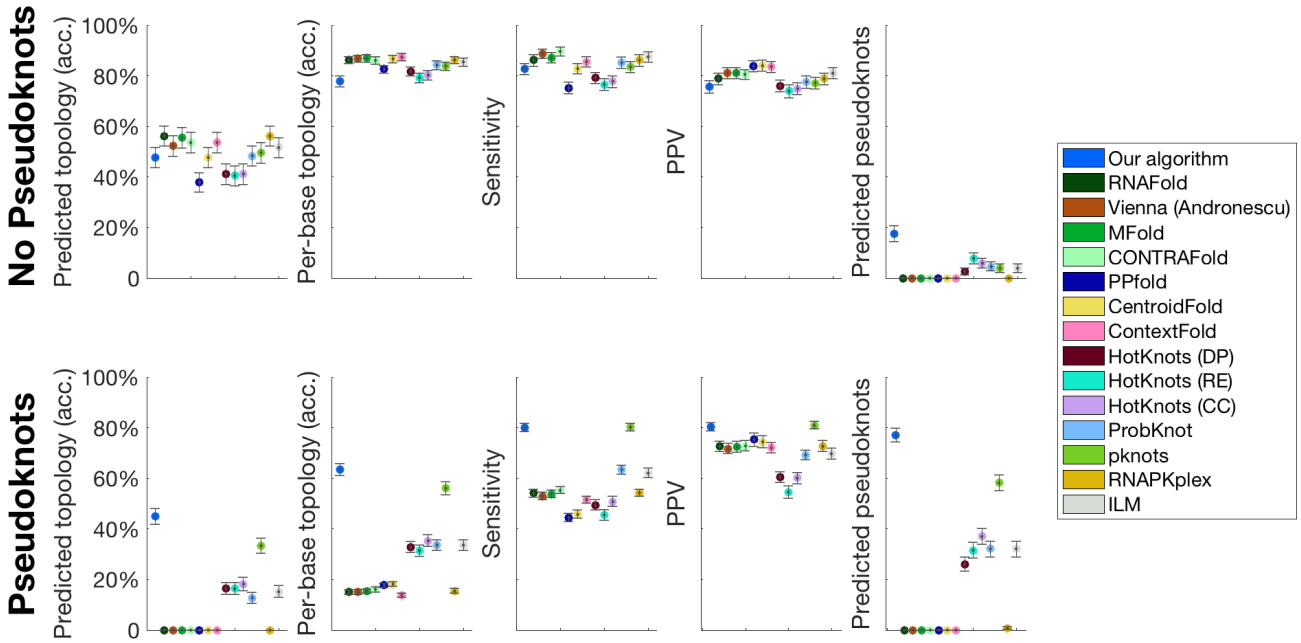

FIG. S3: **Results only including sequences whose structure our algorithm could have predicted.** We consider only the 153 non-pseudoknotted and 165 pseudoknotted sequences whose structures do not include base pairs or topologies disallowed by our algorithm. In this case, we predict the correct topology with 49% (47%) accuracy for non-pseudoknotted (pseudoknotted) structures. This number increases to 62% (82%) and 67% (85%) for top-5 and top-10 accuracy. Surprisingly, we therefore find that our algorithm actually performs better in predicting the pseudoknotted structures in the databases used than the non-pseudoknotted structures. The main results are the same for this dataset as for the full dataset plotted in Fig. 4: our algorithm outperforms all 14 algorithms tested against in predicting pseudoknotted structures, and performs on par with the other algorithms in predicting non-pseudoknotted structures, even though it uses orders of magnitude fewer entropic parameters than the other algorithms tested against.

The constraint placed on allowed sequences in this figure allows us to address to what extent the polymer physics entropy model developed in this work is responsible for our good results, rather than the enumeration scheme. This figure represents a control of the enumeration procedure; pknots, which comes closest to our algorithm's success, only predicts seven sequences included in this dataset to fold into a structure more complex than our algorithm's chosen constraints allow. Removing these seven sequences (in addition to those already removed) does not have a significant effect on the results presented in this figure. (The largest effect is in the accuracy of the predicted topology which increases for pknots from 0.33 to 0.35). We conclude that the difference between our novel entropy model and pknots' (or other algorithms') phenomenological model, rather than the difference in the enumeration procedure, is primarily responsible for the success of our algorithm compared to current metrics.

S7. SCALING OF THE ALGORITHM PROPERTIES FOR RANDOM SEQUENCES AND DISTRIBUTIONS OF ALGORITHM PROPERTIES FOR SEQUENCES IN THE BENCHMARK DATASET

A. Scaling for random sequences

In order to test the scaling properties of the algorithm, we input 100 random sequences for each length between 10 and 21 nucleotides, and set $m = 1$. We plot various properties of the results as a function of the length of the sequence in Fig. S4A. Blue circles are datapoints for each of the 100 sequences in each column. Purple points show the mean. The number of secondary structures grows exponentially with the length of the sequence, as expected due to the brute-force nature of the algorithm, though the number of possible stems grows sub-exponentially. These results are explained later in this section. Similarly, the number of topologies grows exponentially. The probability of forming a pseudoknot appears to plateau at around 10%.

In Fig. S4B, we show that the time the algorithm takes to calculate free energies (the rate limiting step for sequences of any substantial length) grows approximately linearly with the number of possible secondary structures. This is precisely as expected, since the algorithm independently calculates the free energy of each structure, in a process that is easily parallelizable. Deviations from linearity are presumably due to memory constraints which lead to increased computational time for sequences for which many structures need to be stored. While it is customary to plot the time taken as a function of sequence length, as shown in panel A there is a wide variability for each sequence length in the total number of structures, and therefore a similarly wide variability in the time taken. The time taken by the algorithm for a given sequence is better-predicted by the total number of structures enumerated for that sequence than by its length. As shown in panel A (top left) and explained below, the average number of structures for a given sequence grows exponentially with the sequence length, and therefore, the total time taken by the algorithm also grows exponentially with sequence length.

In panel C, we show that for large numbers of stems, the number of possible secondary structures grows as a power law with the number of possible stems. This sub-exponential behavior is due to the fact that some stems cannot coexist in the same structure (if they share any of the same nucleotides or if their coexistence leads to a topology more complex than those in Fig. S2).

B. Scaling of number of structures with sequence length

One main result of the above analysis is that the algorithm runtime is dominated by the scaling properties of the number of structures with the length of the sequence. We therefore sought to better understand this scaling, especially for longer sequences which are not examined in Fig. S4.

A first-order estimate ignores the steric effects of pairing (such as the constraint that if two nucleotides are within a certain linear distance in sequence space, they cannot pair to one another as doing so would create a hairpin that is too small). We make this approximation, and only consider that two nucleotides can pair if they are complementary, and importantly, cannot pair to more than one partner within the same structure. The neglected effect is of course important, though it is expected to give only a higher order correction (i.e. it will not be the dominant effect for purposes of examining scaling behavior for long sequences). The exception of course is for short sequences for which steric effects will be significant – and for which we have enumerated a representative sample of possible structures in Fig. S4. If sterics have any significant effect, it will be to decrease the number of possible structures for short sequences especially, and the effect will be less pronounced for longer sequences which are those we are concerned with here.

For each sequence of length n, we can therefore make structures that include up to $j_{\max} = \text{floor}(n/2)$ base pairs. We can enumerate the number of structures with j base pairs, which we call $N^j_{\text{structures}}$, and then sum this function up for j values from 1 to $j_{\max}$. In other words

$$N^{\text{total}}_{\text{structures}} = \sum_{j=1}^{j_{\max}} N^j_{\text{structures}}$$

We calculate $N^j_{\text{structures}}$ by going base pair by base pair. For the first base pair, there are n first nucleotides to choose from, and on average $3(n-1)/8$ complementary nucleotides. Since we could flip which nt is chosen first and which second, we also multiply by a factor of 1/2. Once the first base pair is chosen, there are $n-2$ nts remaining, which form a sequence of length $n-2$ nts which can be analyzed just as the previous sequence of length n (in other

words, we're describing a mathematically recursive process). Finally, for a structure comprised of j base pairs, there are $j!$ possible (equivalent) orderings of base pairs. Therefore,

$$N^{j}_{\text{structures}} = \frac{1}{j!}\prod_{i=0}^{j-1}\frac{3}{16}(n-2i)(n-2i-1). = \frac{1}{j!}\left(\frac{3}{16}\right)^{j}\frac{n!}{(n-2j)!}.$$

Simplifying, we find that the average total number of structures for a sequence of length n, ignoring steric constraints, is

$$N^{\text{total}}_{\text{structures}} = \sum_{j=1}^{\text{floor}(n/2)}\frac{1}{j!}\left(\frac{3}{16}\right)^{j}\frac{n!}{(n-2j)!}.$$

We tested this equation by explicitly enumerating all possible sequences of length up to 20 nucleotides (see Python code posted to GitHub) and finding perfect agreement with the equation.

This result demonstrates that the total number of structures grows approximately exponentially with the length of the sequence, even for sequences much longer than those examined in Fig. S4 (for which this exponential scaling was also apparent). We plot the result for sequences up to length 400 in Fig. S5. Despite slight curvature for short sequences (for which this naive scaling estimate will not be accurate since steric constraints will be dominant), the result shows exponential growth of the total number of possible structures with the length of the sequence.

As the figure makes clear, the number of possible structures places a significant limit on the length of sequences one can consider by complete landscape enumeration. However, the limit is not nearly as bad as what is suggested by the figure, since the steric considerations ignored to produce it eliminate many structures. Furthermore, by considering stems of length m rather than single base pairs, we can reach sequences up to around 90 nts.

C. Distribution of algorithm scaling properties for benchmark dataset sequences

In Fig. S6 we describe running time and secondary structure count distributions for sequences in the benchmark dataset. We show the histogram of the total time taken to run the algorithm with $N^{\text{max}}_{\text{stems}} = 150$ in panel A (left), finding that the longest time taken was 25 minutes for one sequence. In panel A (middle) we show a histogram of the total number of secondary structures enumerated by the algorithm, finding a wide distribution spanning several orders of magnitude. We also demonstrate that we are in the regime where parallelization will strongly affect the runtime of the algorithm by showing (panel A, right) that the free energy calculation took several times longer than the enumeration procedure. We note however that the details of this calculation (especially the graph decomposition procedure which takes the bulk of the time) have likely not been optimized.

In panel B we show similar plots for the case when no constraints on the types of pseudoknots possible were included (i.e. pseudoknots more complex than those shown in Fig. S2 were also enumerated). We show that including these pseudoknots increases the time it takes to enumerate the structures significantly; the maximum time for a single sequence using $N^{\text{max}}_{\text{stems}} = 150$ increases to 11 hours (panel B, left). While the total number of enumerated secondary structures also increases dramatically (panel B, middle) by leveraging the parallelizability of the algorithm we remain well within the realm of feasibility given the rapid recent growth of available computing power. We also demonstrate (panel B, right) that by decreasing $N^{\text{max}}_{\text{stems}}$ even to 100, orders of magnitude fewer structures are enumerated. The time taken to enumerate the structures also decreases significantly (the maximum is 9 minutes). Our results demonstrate that even exponential-time algorithms such as this complete enumeration are not prohibitive.

In Fig. S7 we examine the loop entropies for the MFE structures predicted by our algorithm for the sequences in the benchmark dataset. We find that the loop entropy (multiplied by temperature) ranges from 5-35 kcal/mol, and is in particular higher for pseudoknotted structures. We further find that the magnitude of the loop entropy is on average slightly over half that of the stem free energy, but represents a higher fraction for pseudoknotted structures. Since the loop entropies contribute in opposite sign to the stem free energies, this demonstrates that as a general rule, the magnitude of the predicted loop entropy is roughly equal to the magnitude of the total free energy of a structure. The accuracy of the loop entropy model is therefore highly significant.

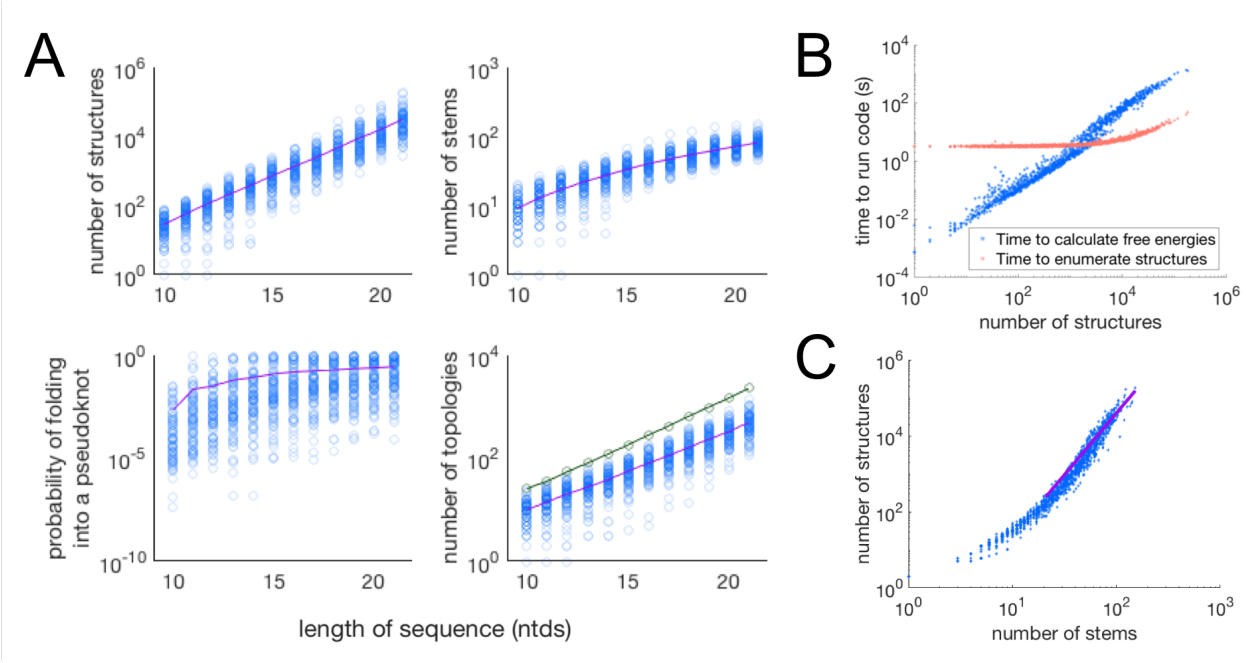

FIG. S4: **Scaling of the algorithm properties with length of sequence.** We input 100 random sequences for each length between 10 and 21 nucleotides into the algorithm. **(A)** Various properties of the results are plotted as a function of the length of the sequence. Blue circles are datapoints for each of the 100 sequences in each column. Purple points show the mean. The number of secondary structures grows exponentially with the length of the sequence, as expected due to the brute-force nature of the algorithm, though the number of possible stems grows sub-exponentially. The probability of forming a pseudoknot appears to plateau at around 10%. The number of topologies grows exponentially (we exclude topologies more complex than those shown in Fig. S2 and the structures leading to them). The green line shows the total number of different topologies over all 100 sequences of a given length. We disallowed parallel stems for this analysis. **(B)** The time the algorithm takes to calculate free energies grows approximately linearly with the number of possible secondary structures, and therefore exponentially with sequence length (see panel A, top left). The data is well-fit to a power law $y = ax^b$ with parameters $a = (3.8 \pm 0.3) * 10^{-4}$ and $b = 1.27 \pm 0.01$. The time taken to enumerate all the structures is constant for short sequences (when few structures are enumerated and the algorithm's overhead is the rate-limiting factor) and then grows as a power law. For sequences of any substantial length, the algorithm is rate-limited by the time it takes to compute free energies, rather than the time taken to enumerate structures. The MatLab program was run on a MacBook Pro 2012 laptop with a 2.3 GHz Intel Core i7 processor and 8 GB memory. **(C)** For large numbers of stems, the number of possible secondary structures grows as a power law with the number of possible stems. This sub-exponential behavior is because some stems cannot coexist in the same structure (if they share any of the same nucleotides or if their coexistence leads to a topology more complex than those in Fig. S2). The purple line shows a fit to the equation $y = ax^b$ with $R^2 = 0.81$. The best-fit values of a and b are found to be $a = 0.0129 \pm 0.0065$ and $b = 3.24 \pm 0.11$.

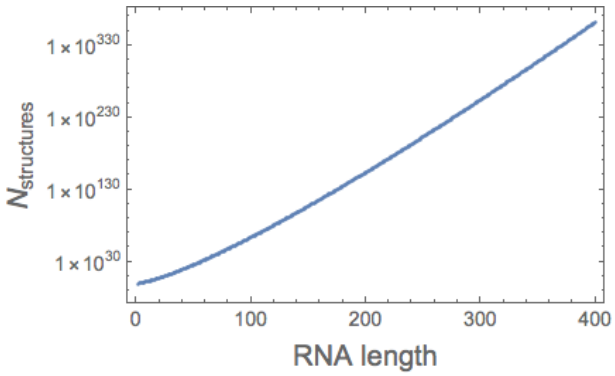

FIG. S5: **Simple scaling estimate for the total number of structures with sequence length.** In Section S7 we find an exact formula for the average number of possible structures as a function of sequence length, neglecting steric effects. Here we plot the results of that formula. We find that despite slight curvature for short sequences (for which this naive scaling estimate will not be accurate since steric constraints will be dominant), the result shows exponential growth of the total number of possible structures with the length of the sequence. Plot created using Mathematica

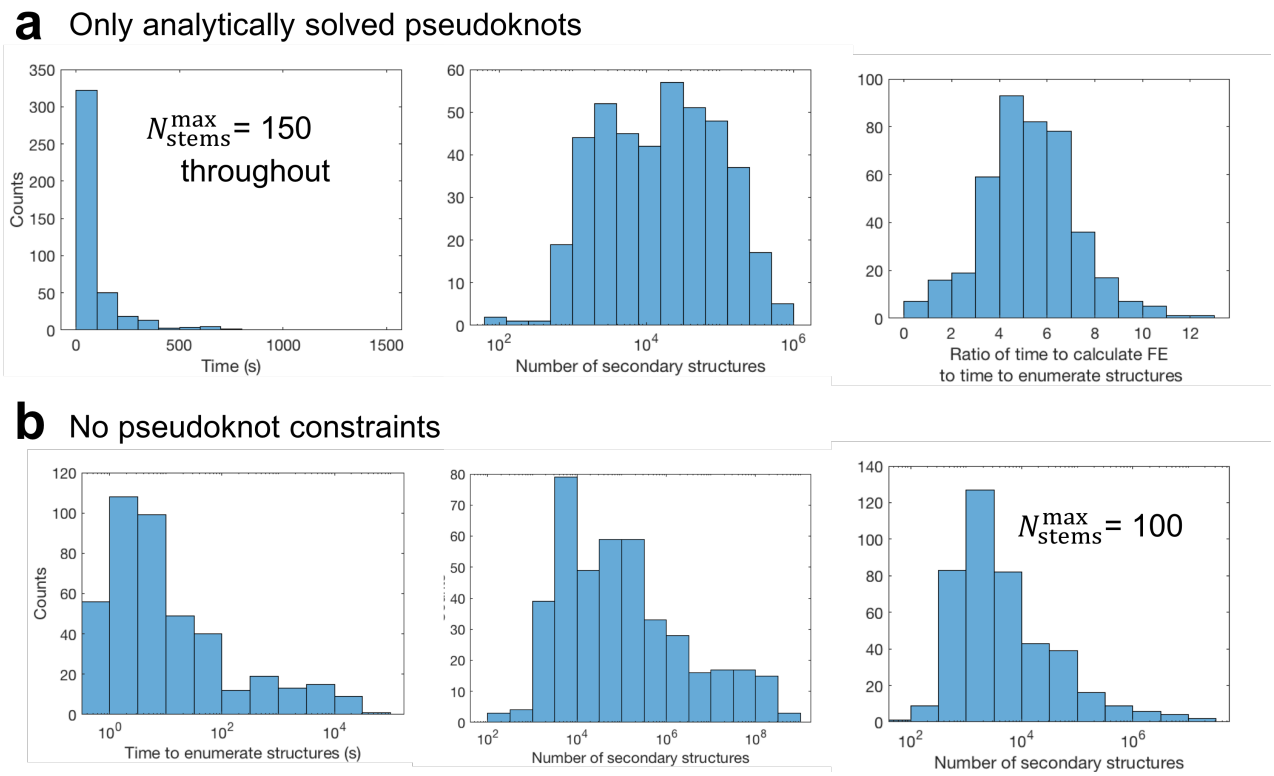

FIG. S6: **Running time and secondary structure count distributions for sequences in the benchmark dataset. a: Left:** Histogram of the total time taken to run the algorithm with $N_{\text{stems}}^{\max} = 150$ for sequences in the benchmark dataset. The longest time taken was 25 minutes for one sequence. Unlike Fig. S4, these results were calculated on a Macbook Pro 2016 laptop with a 3.1 GHz Intel Core i7 processor and 16 GB memory. **Middle:** A histogram of the total number of secondary structures enumerated by the algorithm. **Right:** Calculating the free energy (FE) took several times longer than the enumeration procedure, though the details of this calculation (especially the graph decomposition procedure which takes the bulk of the time) have likely not been optimized. **b:** Results when no constraints on the types of pseudoknots possible were included (i.e. pseudoknots more complex than those shown in Fig. S2 were also enumerated) **Left:** Including all types of pseudoknots increases the time it takes to enumerate the structures significantly; the maximum time for a single sequence using $N_{\text{stems}}^{\max} = 150$ increases to 11 hours. **Middle:** The total number of enumerated secondary structures also increases dramatically, but remains well within the realm of feasibility given the rapid recent growth of available computing power. **Right:** Orders of magnitude fewer structures are enumerated if $N_{\text{stems}}^{\max}$ is decreased even to 100. The time taken to enumerate the structures also decreases significantly (the maximum is 9 minutes). Our results demonstrate that even exponential-time algorithms such as this complete enumeration are not prohibitive.

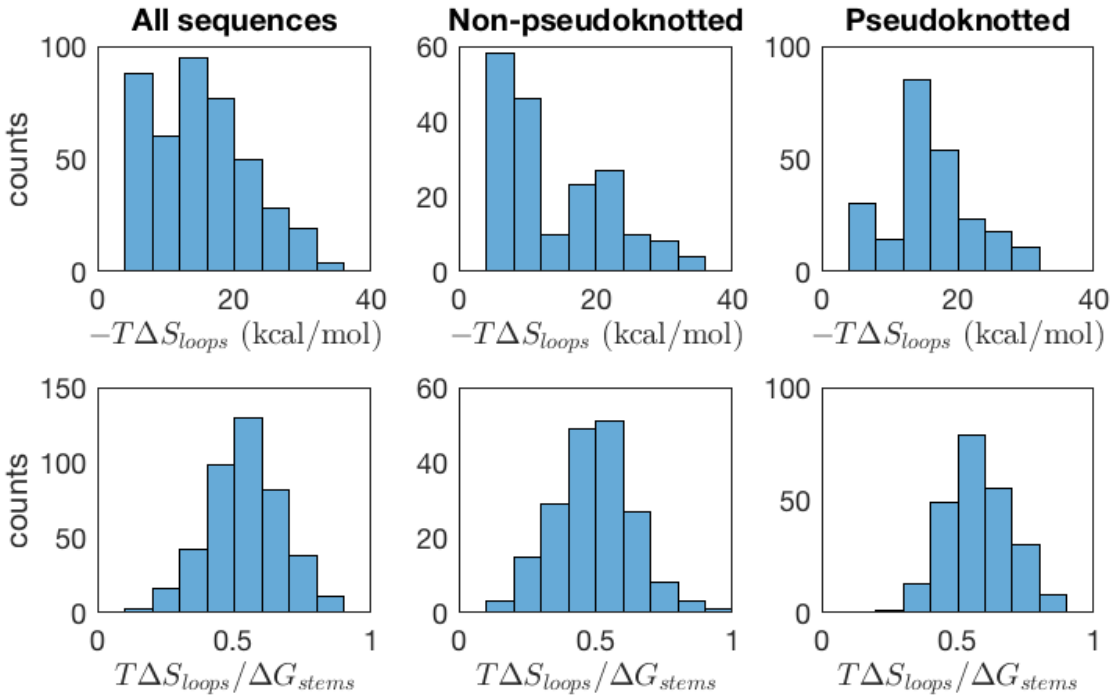

FIG. S7: **Loop entropy statistics.** We examine the loop entropies for the MFE structures predicted by our algorithm for the sequences in the benchmark dataset. We show the results for all sequences (first column), only non-pseudoknotted structures (second column), and only pseudoknotted structures (third column). We considered the predicted structures for the purposes of this classification, but the results don't change significantly if they are classified based on the experimental structures. **The first row** shows the magnitude of the predicted loop entropies. We find that the loop entropies range from 0 to ∼35 kcal/mol, and are in particular higher for pseudoknotted structures, as expected. **The second row** shows the ratio between the magnitude of the predicted loop entropies and the stem free energies $\Delta G_{\text{stems}} = \Delta H_{\text{stems}} - T\Delta S_{\text{stems}}$, which were calculated using the Turner parameters. We find that the magnitude of the loop entropy is on average half that of the stem free energy, but represents a higher fraction for pseudoknotted structures. Since the loop entropies contribute in opposite sign to the stem free energies, this demonstrates that as a general rule, the magnitude of the predicted loop entropy is roughly equal to the magnitude of the total free energy of a structure.

S8. APPLYING OUR FORMALISM TO KISSING HAIRPIN PSEUDOKNOTS

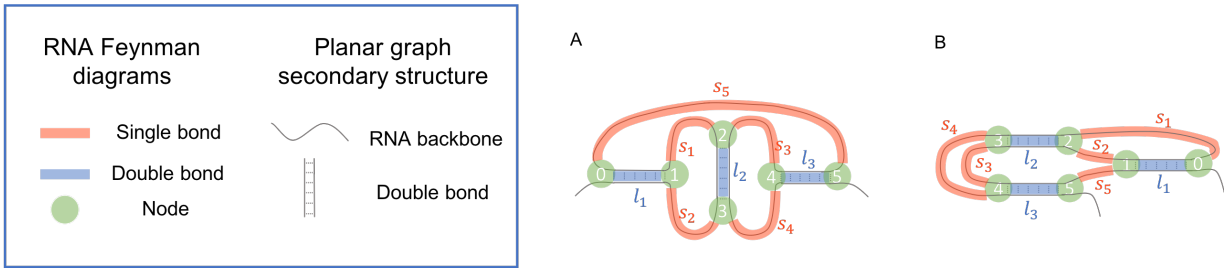

FIG. S8: **Examples of topologies whose entropies need to be solved numerically. A** A kissing hairpin pseudoknot. **B** The most common topology in the dataset which is more complex than those allowed by our chosen constraints. It is equivalent to an H-type pseudoknot with an internal loop in one of the stems.

A biologically common complex pseudoknot for which no entropy calculation has been available is the kissing-hairpin pseudoknot (Fig. S8A). Using our formalism, the entropy of this structure can be estimated by solving the integral

$$e^{\Delta S/k_B} = v_s^3 \int d\vec{r_1} \int d\vec{r_2} \int d\vec{r_3} \int d\vec{r_4} \int d\vec{r_5} \; \frac{\delta(|\vec{r_1}| - l_1)}{4\pi l_1^2} \frac{\delta(|\vec{r_3} - \vec{r_2}| - l_2)}{4\pi l_2^2} \frac{\delta(|\vec{r_5} - \vec{r_4}| - l_3)}{4\pi l_3^2} \times$$
$$P_{s_1}(\vec{r_5}) P_{s_2}(\vec{r_2} - \vec{r_1}) P_{s_3}(\vec{r_3} - \vec{r_1}) P_{s_4}(\vec{r_4} - \vec{r_2}) P_{s_5}(\vec{r_4} - \vec{r_3}). \quad \text{(S10)}$$

We describe the process by which this integral can be solved. First, we let $\gamma = 3/2b$ and call $s_i' = s_i/\gamma$, neglecting the primes from here on for notational convenience. We let $\alpha = \pi^{-15/2}(s_1 s_2 s_3 s_4 s_5)^{-3/2}(v_s/4\pi)^3 (l_1 l_2 l_3)^{-2}$. We let $\vec{r_{ij}} = \vec{r_i} - \vec{r_j}$. The main difficulty in solving these integrals is choosing the proper integration variables. The integral is

$$e^{\Delta S/k_B} = \alpha \int d\vec{r_{54}}\; d\vec{r_4}\; d\vec{r_{32}}\; d\vec{r_2}\; d\vec{r_1}\; \delta(|\vec{r_1}| - l_1)\; \delta(|\vec{r_{32}}| - l_2)\; \delta(|\vec{r_{54}}| - l_3) \times$$
$$\exp\left[-(\vec{r_{54}} + \vec{r_4})^2/s_5 - (\vec{r_2} - \vec{r_1})^2/s_1 - (\vec{r_{32}} + \vec{r_2} - \vec{r_1})^2/s_2 - (\vec{r_4} - \vec{r_2})^2/s_3 - (\vec{r_4} - \vec{r_{32}} - \vec{r_2})^2/s_4\right]. \quad \text{(S11)}$$

We can now proceed to first do the $\vec{r_{54}}$ integral, following the same procedure as in Section S5. $\vec{r_{54}} \cdot \vec{r_4}$ becomes $|\vec{r_{54}}||\vec{r_4}| \cos\theta$ where θ is the integral between the vectors $\vec{r_{54}}$ and $\vec{r_4}$. The integral over all terms containing $\vec{r_{54}}$ yields $\frac{\pi l_3 s_5}{r_4} e^{-l_3^2/s_5 - r_4^2/s_5}(e^{2l_3 r_4/s_5} - e^{-2l_3 r_4/s_5})$.

We can similarly do the $\vec{r_1}$ integral. In order to do so, we define a variable $x = \vec{r_2}(1/s_1 + 1/s_2) + \vec{r_{32}}/s_2$. Thus, $\vec{r_1}$ only appears in our integrals as r_1^2 and as $\vec{r_1} \cdot x$. In order to change the integration variable r_2 to x, we need to introduce the Jacobian $J = (s_1 s_2/s_1 + s_2)^3$. We also set $a = \frac{s_1}{s_3} - \frac{s_2}{s_4}$. After doing the integral, we can expand out the exponent to get

$$e^{\Delta S/k_B} = \alpha J \pi^2 l_1 l_3 s_5 e^{-l_3^2/s_5 - l_1^2(\frac{1}{s_1} + \frac{1}{s_2})} \int d\vec{r_4} \frac{1}{r_4} e^{-r_4^2(\frac{1}{s_3} + \frac{1}{s_4} + \frac{1}{s_5})} \left(e^{2l_3 r_4/s_5} - e^{-2l_3 r_4/s_5}\right) \times$$
$$\int d\vec{x} \frac{1}{x} e^{-x^2\left[\frac{s_1 s_2}{s_1+s_2} + (\frac{s_1 s_2}{s_1+s_2})^2(\frac{1}{s_3} + \frac{1}{s_4})\right]} \left(e^{2l_1 x} - e^{-2l_1 x}\right) e^{2\vec{x}\cdot\vec{r_4}\left(\frac{s_1 s_2}{s_1+s_2}(\frac{1}{s_3} + \frac{1}{s_4})\right)} \times$$
$$\int d\vec{r_{32}} \delta(|\vec{r_{32}}| - l_2) e^{-r_{32}^2\left(\frac{1}{s_1+s_2} + \frac{s_1^2/s_3 + s_2^2/s_4}{(s_1+s_2)^2}\right)} \; e^{2\vec{x}\cdot\vec{r_{32}}\frac{s_1 s_2 a}{(s_1+s_2)^2} - 2\vec{r_4}\cdot\vec{r_{32}}\frac{a}{s_1+s_2}}. \quad \text{(S12)}$$

As can be seen, if $a = 0$, meaning $\frac{s_1}{s_3} = \frac{s_2}{s_4}$, then $\vec{r_{32}}$ only enters our equations as r_{32}^2. In this case, integration over $\vec{r_{32}}$ simply yields $4\pi l_2^2 e^{-l_2^2\left(\frac{1}{s_1+s_2} + \frac{s_1^2/s_3 + s_2^2/s_4}{(s_1+s_2)^2}\right)}$. Setting θ to be the angle between $\vec{r_4}$ and $\vec{x}$, integration over θ proceeds as in previous cases. Integration over the remaining three angles gives $8\pi^2$. Thus,

$$e^{\Delta S/k_B}(a=0) = \alpha J 16\pi^5 l_1 l_2^2 l_3 s_5 \left(\frac{s_1+s_2}{s_1 s_2(\frac{1}{s_3}+\frac{1}{s_4})}\right) e^{-l_1^2(\frac{1}{s_1}+\frac{1}{s_2})-l_2^2\left(\frac{1}{s_1+s_2}+\frac{s_1^2/s_3+s_2^2/s_4}{(s_1+s_2)^2}\right)-l_3^2/s_5} \times$$
$$\int_0^\infty dr_4 e^{-r_4^2(\frac{1}{s_3}+\frac{1}{s_4}+\frac{1}{s_5})}\left(e^{2l_3 r_4/s_5}-e^{-2l_3 r_4/s_5}\right)\int_0^\infty dx e^{-x^2\left[\frac{s_1 s_2}{s_1+s_2}+(\frac{s_1 s_2}{s_1+s_2})^2(\frac{1}{s_3}+\frac{1}{s_4})\right]}\left(e^{2l_1 x}-e^{-2l_1 x}\right)\times$$
$$\left(e^{2xr_4\left(\frac{s_1 s_2}{s_1+s_2}(\frac{1}{s_3}+\frac{1}{s_4})\right)}-e^{-2xr_4\left(\frac{s_1 s_2}{s_1+s_2}(\frac{1}{s_3}+\frac{1}{s_4})\right)}\right). \quad \text{(S13)}$$

These integrals can be solved analytically (by completing the square in the exponent for each of the eight terms in the sum). The result is

$$e^{\Delta S/k_B}(a=0) = \alpha J 16\pi^5 l_1 l_2^2 l_3 s_5 \left(\frac{s_3 s_4(s_1+s_2)}{s_1 s_2(s_3+s_4)}\right) e^{-l_1^2(\frac{s_1+s_2}{s_1 s_2})-l_2^2\left(\frac{s_{1234}+s_3 s_4(s_1-s_2)a}{s_3 s_4(s_1+s_2)^2}\right)-l_3^2/s_5} \times$$
$$2\pi(s_1+s_2)\sqrt{\frac{s_3 s_4 s_5}{s_1 s_2(s_q+s_{1234})}}\, e^{\frac{l_1^2 s_3 s_4(s_1+s_2)^2}{s_1 s_2 s_{1234}}+\frac{l_1^2 s_q^2+l_3^2 s_{1234}^2}{s_5 s_{1234}(s_q+s_{1234})}} \sinh\left(\frac{2l_1 l_3 s_q}{s_5(s_q+s_{1234})}\right) \quad \text{(S14)}$$

where we've defined $s_q = s_5(s_1+s_2)(s_3+s_4)$ and $s_{1234} = s_1 s_2 s_3 + s_1 s_2 s_4 + s_1 s_3 s_4 + s_2 s_3 s_4$. We've written the solution so that the bottom line is the result of the integrals, and written the prefactor of l_2^2 in a way that clarifies how it simplifies.

We can simplify the final result by introducing the variables

$$s_A = \frac{s_5(s_q+s_{1234})}{s_q}; \quad s_B = \frac{s_3 s_4(s_1+s_2)^2}{s_{1234}}; \quad s_v = \sqrt{s_q+s_{1234}}.$$

yielding

$$e^{\Delta S/k_B}(a=0) = \frac{(v_s/s_v)^3}{2\pi^{9/2}}\frac{s_A}{l_1 l_3} e^{-\left(\frac{l_1^2+l_3^2}{s_A}\right)-\frac{l_2^2}{s_B}} \sinh\left(\frac{2l_1 l_3}{s_A}\right) \quad \text{(S15)}$$

One of the concrete predictions emerging from this calculation is that if $a = 0$, meaning that the pseudoknot is symmetric, that the entropy of the structure should depend on l_2 only as $\exp(-l_2^2/s_B)$ where s_B depends on the lengths of the various loops but is independent of l_1, l_3, and s_5.

We now return to the more general case of $a \neq 0$. In this case, we define a new variable $\vec{y}$ to be the total vector dotted with $\vec{r_{32}}$ in Eq. S12: $\vec{y} = \frac{s_1 s_2 a}{(s_1+s_2)^2}\vec{x} - \frac{a}{s_1+s_2}\vec{r_4}$. Integration over $\vec{r_{32}}$ then yields

$$e^{\Delta S/k_B}(a\neq 0) = \alpha J \pi^3 l_1 l_2 l_3 s_5 e^{-l_1^2(\frac{1}{s_1}+\frac{1}{s_2})-l_2^2\left(\frac{1}{s_1+s_2}+\frac{s_1^2/s_3+s_2^2/s_4}{(s_1+s_2)^2}\right)-l_3^2/s_5} \times$$
$$\int d\vec{r_4}\frac{1}{r_4}e^{-r_4^2(\frac{1}{s_3}+\frac{1}{s_4}+\frac{1}{s_5})}(e^{2l_3 r_4/s_5}-e^{-2l_3 r_4/s_5})\int d\vec{x}\frac{1}{x}e^{-x^2\left[\frac{s_1 s_2}{s_1+s_2}+(\frac{s_1 s_2}{s_1+s_2})^2(\frac{1}{s_3}+\frac{1}{s_4})\right]}(e^{2l_1 x}-e^{-2l_1 x})\times$$
$$e^{2\vec{x}\cdot\vec{r_4}\left(\frac{s_1 s_2}{s_1+s_2}(\frac{1}{s_3}+\frac{1}{s_4})\right)}\frac{1}{y}\left(e^{2l_2 y}-e^{-2l_2 y}\right). \quad \text{(S16)}$$

As before, we can perform three of the angle integrals to yield $8\pi^2$, and define θ to be the angle between $\vec{r_4}$ and $\vec{x}$. Then, $\vec{x}\cdot\vec{r_4}$ becomes $r_4 x\cos\theta$. We can then write y in terms of $\cos\theta$: $y = \sqrt{\vec{y}\cdot\vec{y}} = \frac{a}{s_1+s_2}\sqrt{\frac{s_1^2 s_2^2}{(s_1+s_2)^2}x^2 + r_4^2 - \frac{2s_1 s_2}{s_1+s_2}r_4 x\cos\theta}$. We can thus turn the integration over $\cos\theta$ into an integration over y (again, the Jacobian needs to be accounted for). Defining the limits of the integration to be $y_\pm = \sqrt{\frac{a^2}{(s_1+s_2)^2}\left(\frac{s_1 s_2}{(s_1+s_2)}x \pm r_4\right)^2}$, we have

$$e^{\Delta S/k_B}(a\neq 0)=\alpha J 8\pi^5 l_1 l_2 l_3 s_5\frac{(s_1+s_2)^3}{s_1 s_2 a^2}e^{-l_1^2(\frac{1}{s_1}+\frac{1}{s_2})-l_2^2\left(\frac{1}{s_1+s_2}+\frac{s_1^2/s_3+s_2^2/s_4}{(s_1+s_2)^2}\right)-l_3^2/s_5}\times$$
$$\int_0^\infty dr_4 e^{-r_4^2(\frac{1}{s_3}+\frac{1}{s_4}+\frac{1}{s_5})}(e^{2l_3 r_4/s_5}-e^{-2l_3 r_4/s_5})\int_0^\infty dx e^{-x^2\left(\frac{s_1 s_2 s_{1234}}{(s_1+s_2)^2 s_3 s_4}\right)}(e^{2l_1 x}-e^{-2l_1 x})\times$$
$$\int_{y_-}^{y_+} dy e^{\left(\frac{s_1^2 s_2^2}{(s_1+s_2)^2}x^2+r_4^2-\frac{(s_1+s_2)^2}{a^2}y^2\right)\left(\frac{1}{s_3}+\frac{1}{s_4}\right)}\left(e^{2l_2 y}-e^{-2l_2 y}\right). \quad \text{(S17)}$$

The y integral must be done first because its limits include the other two integration variables. However, this integral results in an error function which cannot be integrated analytically. While various limits might be taken to impose analyticity, given the speeds of programs like Mathematica in performing simple numerical integrals like this one, we prefer to solve the resulting integrals numerically.

There are eight parameters to be varied, and we display the results of the entropy calculation for single-parameter sweeps in Fig. S9. For this figure, we set $s_1=3$, $s_2=4$, $s_3=6$, $s_4=8$, $s_5=3$, $l_1=2$, $l_2=3$, $l_3=4$. Then, keeping all other parameters at those values, we take each parameter and measure the entropy as a function of varying that parameter.

The resulting plot contains eight different curves, which we've plotted in Fig. S9. As expected, for $s_1=s_5=3$, the blue and orange curves coincide. Varying the loop lengths (panel A) appears to give less dramatic changes than varying the stem lengths (panel B). The parameter l_2 was capped at seven because for values greater than that, one of the hairpins wouldn't be able to close ($s_1+s_2<l_2$). The asymmetry between the l_1 and l_3 curves is due to the asymmetry between the constant values of l_1 and l_3 chosen. We also verified that the result of the numerical integration for $a\neq 0$ approaches the result of the analytic solution ($a=0$) as a approaches zero.

We also give a more comprehensive result of the numerical integration. Since displays of eight-parameter tables are difficult to achieve, we give the results of this numerical integration for values of s_i and l_i ranging from 1 to 5 (or s_i' ranging from $1/\gamma$ to $5/\gamma$) as a .h5 file. These types of files can easily be imported using, for example, Python, with the following lines of code:

```
import h5py
import numpy as np
f = h5py.File('kissingHairpinsSuppFile.h5', 'r')
k = np.array(f[list(f.keys())[0]])
```

This code sets the variable `k` to be an eight-dimensional array, such that `k[a][b][c][d][e][f][g][h]` is the entropy (in units of k_B) of a kissing hairpin with $s_1'=(a+1)/\gamma$, $s_2'=(b+1)/\gamma$, ..., $l_1=f+1$, ..., $l_3=h+1$. The addition of 1 is included because Python begins indexing at 0.

We set the two loop entropy parameters to $b=2.4$ and $v_s=0.02$. As mentioned, the entropy is measured in units of Boltzmann's constant k_B.

We also considered the constraints that each hairpin must have ≥ 3 nts (so $s_1+s_2+l_2\geq 4$ and same for s_3 and s_4) and that the hairpins must be able to close (so $s_1+s_2\geq l_2$ and same for s_3 and s_4). We included these constraints by setting the table values to 0 if these constraints aren't satisfied; of course, if these constraints aren't satisfied the entropy should really be considered to be $-\infty$.

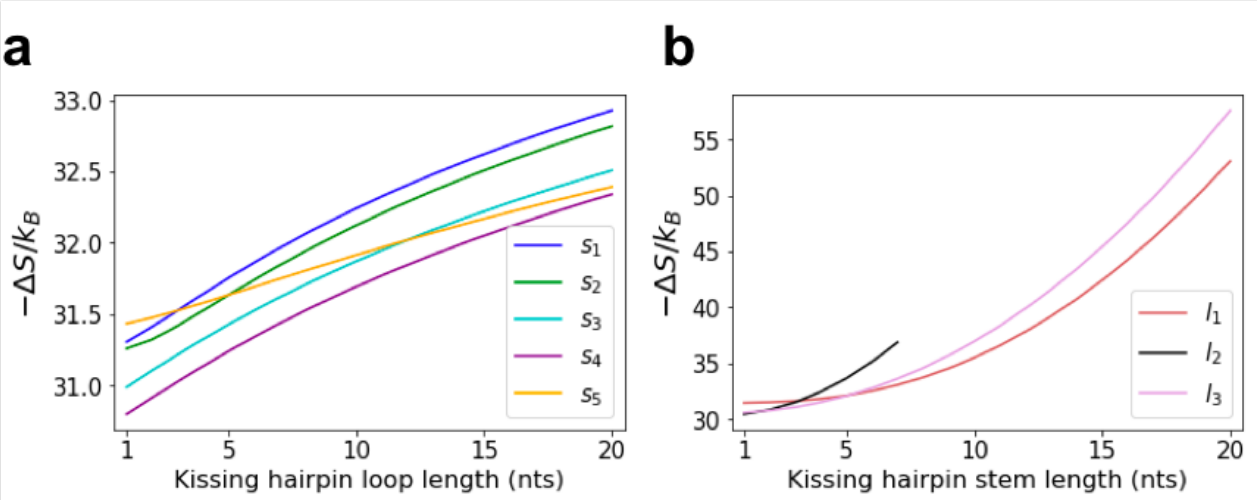

FIG. S9: **One dimensional parameter sweeps for the kissing hairpin pseudoknot entropy.** We set $s_1 = 3$, $s_2 = 4$, $s_3 = 6$, $s_4 = 8$, $s_5 = 3$, $l_1 = 2$, $l_2 = 3$, $l_3 = 4$. Then, keeping all other parameters at those values, we take each parameter and measure the entropy as a function of varying that parameter. See the text for detailed discussion.

S9. CONSIDERING OTHER COMPLEX PSEUDOKNOTS

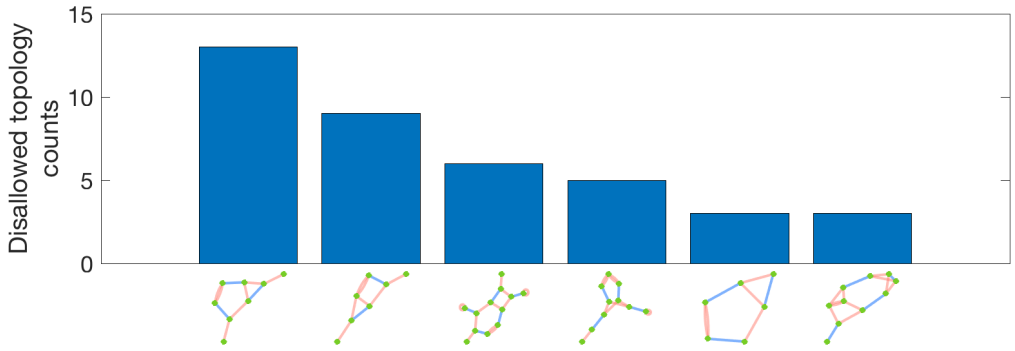

FIG. S10: **Common topologies disallowed by constraints chosen for our algorithm implementation.**

Similar approaches as in the previous section can be taken for other pseudoknots more complex than those shown in Fig. S2.

As discussed, there were 64 pseduoknotted sequences in the experimental datasets used which were found to fold into topologies more complex than those allowed by the constraints we chose to place on the algorithm. Of these 64 sequences, we sought to determine the topologies they shared in common. The six most common topologies and the number of sequences folding into them are plotted in Fig. S10. The most common topology (shown in large in Fig. S8B) is equivalent to an H-type pseudoknot with an internal loop in one stem. As can be seen from Fig. S10, the second and fifth most common topologies are only slight variations on the first: the second is identical to the first with one of the stem lengths set to zero (i.e. the stem is made up of a single base pair) and the fifth is identical to the first with the dangling unpaired regions on the 3' and 5' ends removed.

The entropy of the most common disallowed topology, displayed in large in Fig. S8B, is given by

$$e^{\Delta S/k_B} = v_s^3 \int d\vec{r_1} \int d\vec{r_2} \int d\vec{r_3} \int d\vec{r_4} \int d\vec{r_5} \, \frac{\delta(|\vec{r_1}| - l_1)}{4\pi l_1^2} \frac{\delta(|\vec{r_3} - \vec{r_2}| - l_2)}{4\pi l_2^2} \frac{\delta(|\vec{r_5} - \vec{r_4}| - l_3)}{4\pi l_3^2} \times P_{s_1}(\vec{r_2}) P_{s_2}(\vec{r_2} - \vec{r_1}) P_{s_3}(\vec{r_4} - \vec{r_3}) P_{s_4}(\vec{r_4} - \vec{r_3}) P_{s_5}(\vec{r_5} - \vec{r_1}). \quad \text{(S18)}$$

After changing our integration variables to be $\vec{r_1}$, $\vec{r_{21}}$, $\vec{r_{32}}$, $\vec{r_{45}}$, and $\vec{r_{53}}$, we follow the same formula as for the kissing hairpin pseudoknot to get a similar expression:

$$e^{\Delta S/k_B} = \alpha 8\pi^5 l_1 l_2 l_3 \frac{s_1 s_3 s_4 s_5}{s_3 + s_4} e^{-l_1^2/s_1 - l_2^2/s_5 - l_3^2\left(\frac{1}{s_3}+\frac{1}{s_4}\right)} \int_0^\infty dr_{21} e^{-r_{21}^2\left(\frac{1}{s_1}+\frac{1}{s_2}\right)} \left(e^{2l_1 r_{21}/s_1} - e^{-2l_1 r_{21}/s_1}\right) \times \int_0^\infty dr_{53} e^{-r_{53}^2\left(\frac{1}{s_3}+\frac{1}{s_4}\right)} \left(e^{2\left(\frac{1}{s_3}+\frac{1}{s_4}\right) l_3 r_{53}} - e^{-2\left(\frac{1}{s_3}+\frac{1}{s_4}\right) l_3 r_{53}}\right) \int_{y_-}^{y_+} dy e^{-y^2/s_5} \left(e^{2l_2 y/s_5} - e^{-2l_2 y/s_5}\right) \quad \text{(S19)}$$

where $y_\pm = \sqrt{(r_{53} \pm r_{21})^2}$.

Using this formula, we find that if one instead considers the entropy of the H-type pseudoknot with an internal loop to be comprised of the sum of the entropies of the H-type pseudoknot and the internal loop, this leads to an overestimate of the entropy cost of at least 1 kcal/mol over nearly all parameter values at 37°C. This overestimate is significantly higher for some parameters; a representative example is the case of $l_1 = 2$, $l2 = 4$, $l_3 = 4$, $s_1 = 3$, $s_2 = 3$, (the results are fairly insensitive to s_3, s_4, s_5) which yields an entropy difference of 3.3 kcal/mol, or a 23% error. Changing these parameters can both increase or decrease this error, but there is a very wide parameter regime in which the error due to not taking into account the nestedness of the internal loop is significant.

S10. SAMPLE FREE ENERGY CALCULATION AND GRAPH DECOMPOSITION PROCESS

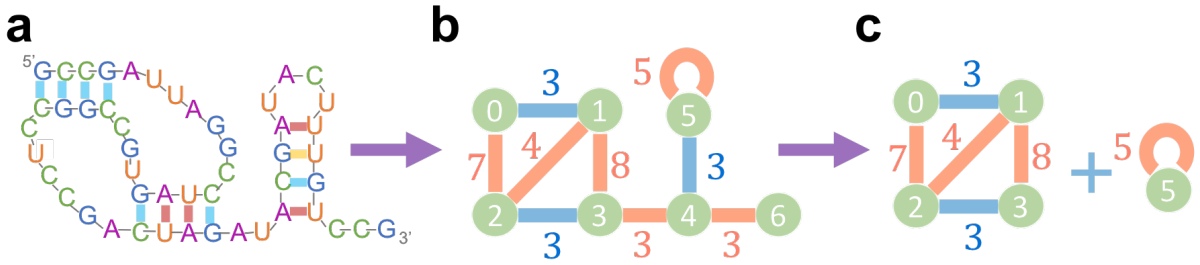

FIG. S11: **Sample structure.** **A** structure under consideration; **B** graph representing the structure; **C** fully decomposed graph. The loop entropy of the structure is the sum of the loop entropies of the graphs in this panel.

Here we describe the graph decomposition process – the basis for the loop entropy calculation in practice – in some more detail and provide a sample calculation of the free energy as an example.

Given a structure (graph) we test each possible edge for whether removing that edge leads to a disconnected graph. If so, we remove it, and the two resulting graphs represent two different motifs. We repeat, and compare the final graphs (that cannot be decomposed further) to our tabulated list (Fig. S2); some of these graphs may represent pseudoknots, while others represent hairpins. Thus, using our tabulated or analytically calculated results for the loop entropy of each possible graph, we calculate the loop entropy of each motif in the RNA structure, and sum them to find the total loop entropy.

As an example, let's consider the structure shown in Fig. S11A. We'd like to calculate the free energy of this structure. First, we calculate the enthalpy terms using the Turner parameters. These include a dangling end, as well as stacking terms and terminal mismatches:

G . G C . C C . C G . G A
C C ' C G ' G G ' G C ' C C

C C . C U . U A . A G . G A
A G ' G A ' A U ' U C ' C **C**

U A . A C . C G . G A . A U
C U ' U G ' G U ' U U ' U U

where the top line goes from 5' to 3' and the bottom line is antiparallel. The bolded C (last in second row) represents the approximation in the Turner rules that if two base pairs can bind but are unbound in the structure, the purine is replaced with A and the pyrimidine with C. Each of these terms has an associated enthalpy and entropy from the tabulated Turner parameters.

Once these terms have been added up, the remaining step is calculating the free energy of the loops. First, we convert the structure to a graph (Fig. S11B) by placing nodes at the edges of stems (here we also place nodes at the ends of the sequence). These nodes are connected by double-stranded (blue) or single-stranded (red) edges. In fact, since the stems have at least length 1, each node (except for perhaps the ones representing the edges of the molecule) must be connected to one double-stranded and two single-stranded edges; the hairpin loop counts as two edges for this purpose. The lengths of the various edges are provided in the figure.

Now, we perform the graph decomposition process. We test each possible edge for whether removing that edge leads to a disconnected graph. The first edge for which this is true is that connecting nodes three and four. We therefore remove that edge, and the two resulting graphs represent two different motifs. We repeat, finding that removing the edge between nodes four and five similarly disconnects the graphs, and same for the edge between nodes four and six. Finally, finding that nodes four and six are not connected to any edges we remove those. We compare the final graphs that cannot be decomposed further – Fig. S11C – to our tabulated list (Fig. S2). We find here that we have one instance of an open-net-2a ($l_1 = 3$; $l_2 = 3$; $s_1 = 7$; $s_2 = 4$; $s_3 = 8$) and a closed-net-0 ($s_1 = 5$). This gives us the loop entropy resulting from this structure, which we add to the bond entropy found using the Turner parameters to

get the total free energy of the structure.

---

[1] Mathai Mammen, Eugene I. Shakhnovich, John M. Deutch, and George M. Whitesides. Estimating the Entropic Cost of Self-Assembly of Multiparticle Hydrogen-Bonded Aggregates Based on the Cyanuric Acid-Melamine Lattice. *Journal of Organic Chemistry*, 63(12):3821–3830, 1998.

[2] Huan-xiang Zhou and Michael K Gilson. Theory of Free Energy and Entropy in Noncovalent Binding. *Chemical Reviews*, 109(9):4092–4107, 2009.

[3] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 2003.

[4] John SantaLucia and Donald Hicks. The Thermodynamics of DNA Structural Motifs. *Annual Review of Biophysics and Biomolecular Structure*, 33(1):415–440, 2004.

[5] Naoki Sugimoto, Shu ichi Nakano, Misa Katoh, Akiko Matsumura, Hiroyuki Nakamuta, Tatsuo Ohmichi, Mari Yoneyama, and Muneo Sasaki. Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. *Biochemistry*, 34(35):11211–11216, 1995.

[6] Norman E. Watkins, William J. Kennelly, Mike J. Tsay, Astrid Tuin, Lara Swenson, Hyung Ran Lee, Svetlana Morosyuk, Donald A. Hicks, and John SantaLucia. Thermodynamic contributions of single internal rA·dA, rC·dC, rG·dG and rU·dT mismatches in RNA/DNA duplexes. *Nucleic Acids Research*, 39(5):1894–1902, 2011.

[7] Homer Jacobson and Walter H. Stockmayer. Intramolecular reaction in polycondensations. I. The theory of linear systems. *The Journal of Chemical Physics*, 18(12):1600–1606, 1950.

[8] H. Isambert and E. D. Siggia. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. *Proceedings of the National Academy of Sciences*, 97(12):6515–6520, 2000.

[9] A. Xayaphoummine, T. Bucher, and Herve Isambert. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, 33(SUPPL. 2):605–610, 2005.