# RNA structure prediction including pseudoknots through direct enumeration of states

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 accurate prediction of RNA secondary structure from primary sequence has had enormous impact on research from the past forty years. While many algorithms are available to make these predictions, the inclusion of non-nested loops, termed pseudoknots, still poses challenges. Here, we describe a new method to compute the entire free energy landscape of secondary structures of RNA resulting from a primary RNA sequence, by combining a polymer physics model for the entropy of pseudoknots with exhaustive enumeration of the set of possible structures. Our polymer physics model can address arbitrarily complex pseudoknots and has only two free loop entropy parameters that correspond to concrete physical quantities, over an order of magnitude fewer than even the sparsest state-of-the-art algorithms. Our model outperforms previously published methods in predicting pseudoknots, while performing on par with current methods in the prediction of non-pseudoknotted structures. For RNA sequences of $\sim 45$ nucleotides, or $\sim 90$ with minimal heuristics, the complete enumeration of possible secondary structures can be accomplished quickly despite the NP-complete nature of the problem.

RNA molecules play physiological roles that extend far beyond translation. In human cells, most RNA molecules are not translated [1]. Non-coding RNAs interact functionally with mRNA [2], DNA [3], and proteins [4], and can be as large as $> 200$ nucleotides (ntds) [5, 6]. However, a substantial fraction are $< 40$ ntds in length, including miRNAs and siRNAs, which serve as regulators for the translation of mRNA [2, 7], and piRNAs which form RNA-protein complexes to regulate the germlines of mammals [8]. The *in vitro* evolution of RNA, especially through SELEX [9–11], has led to an explosion of applications for short RNA molecules, due their ability to tightly and specifically bind to a remarkable range of target ligands [12].

Overwhelmingly, the properties of short non-coding RNA molecules are tied to their three-dimensional, or tertiary, structures [5, 13–16]. Such structures are formed because of the energetic favorability of bonds between complementary nucleotides (primarily A to U, C to G, and G to U). However, these bonds impose an entropic cost; therefore, the conformations most frequently adopted balance the energetic gain of maximal base-pairing 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 [17, 18], current research aims to predict the minimum free energy structures given the sequence. Algorithms typically predict "secondary structure", a list of the base pairings [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]. This NP-complete approach was quickly supplanted by dynamic programming, which has since dominated RNA structure prediction [21–25]. These algorithms efficiently consider an enormous number of structures without explicitly generating them, by iteratively finding the optimal structure for subsequences [26].

However, such algorithms have difficulty predicting RNA secondary structures that include pseudoknots, i.e. structural elements with at least two non-nested base pairs (see Fig. S1A for an example) that make up roughly 1.4% of base pairs [26] and are overrepresented in functionally important regions [27] of RNA. Pseudoknots are disallowed from the most popular RNA structure prediction algorithms (e.g. Refs. [28–30]) due to computational cost; indeed, structural prediction including all pseudoknots has been shown to be NP-complete [31–33]. Significant advances have been made with heuristics, which do not guarantee finding the minimum free energy structure [34–38], and by disallowing all but a narrow class of pseudoknots [39–46].

A major challenge for predicting pseudoknotted structures is the relative lack of experimental data [47]. Thus, up until recently, theoretical approaches have largely been limited to simple H-type pseudoknots [39, 45, 48, 49]. A recent strategy uses machine-learning of large experimental datasets [45, 50, 51]. 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 hamper generalizability. A sketch of a theoretical description of pseudoknot entropies based on polymer physics was developed by Isambert and Siggia [34, 52]; however, their derivations have not been published.

In this study, we demonstrate that for short RNA sequences, it is possible to exactly solve for the probability that the RNA will fold into any given structure, in-

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

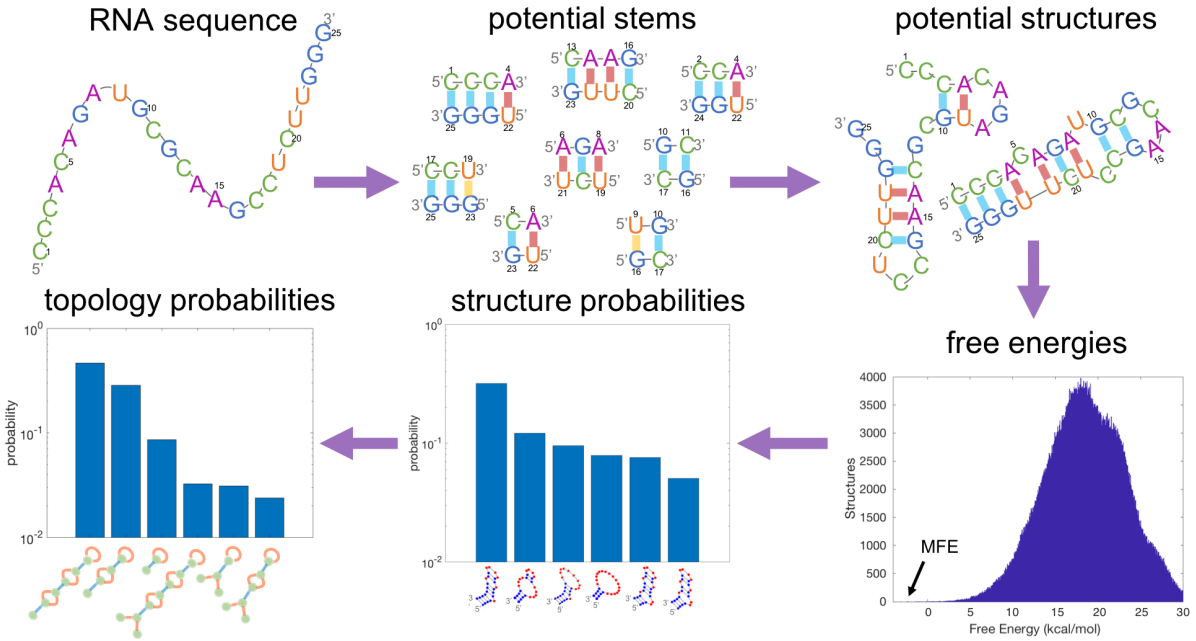


FIG. 1: **Schematic overview of the algorithm.** Given an RNA sequence, the algorithm first enumerates all potential stems (sequences of base pairs) which can form. It then searches for all possible combinations of stems, such that no nucleotide is paired with more than one other, thus forming all possible secondary structures. For each structure, it calculates the free energy, which is comprised of a bond energy term and an entropy term. 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 (described in Section III), arriving at a probability distribution for every possible topology forming in equilibrium.

cluding those with pseudoknots. Complete enumeration of the RNA structure landscape is feasible even for biologically relevant RNA sequences (Section I). Our approach combines a method based on the work of Isambert and Siggia (Section II) with a novel graph-theoretical depiction of the RNA (Section III) to exactly calculate the entropy of each structure, treating both pseudoknotted and non-pseudoknotted RNA structures equivalently. The entropies of structures of arbitrary complexity can be analytically computed with just two experimentally derived physical parameters: the persistence length of single-stranded RNA, and the volume within which two RNA nucelotides are considered bound. This represents an enormous parameter reduction compared to state-of-the-art algorithms like the Cao-Chen or Dirks-Pierce models, which have 258 and 11 parameters, respectively, for H-type pseudoknots alone, and $\sim 18$ parameters for non-pseudoknotted loops [51]. We test our model predictions on molecules from the RNAStrand [53], PseudoBase++ [54], and CompraRNA [55] databases and find good agreement with experimental results (Section IV). Although we fit our entropy model to data from non-pseudoknotted structures, we find that our model outperforms previously published methods in predicting pseudoknots, while performing on par with current methods in the prediction of non-pseudoknotted structures.

## I. ENUMERATING RNA STRUCTURES

The Pipas-McMahon algorithm [20] 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 forty years, coupled with increased appreciation for the physiological and engineering relevance of short RNA strands suggest revisiting this approach. In this section, we describe the process by which we exhaustively enumerate the secondary structures into which an arbitrary given sequence can fold. We first number the nucleotide sequence from 1 to $N$ from the 5' end. We define an $N \times N$ symmetric matrix $B$ which describes which nucleotides can bind to each other: $B_{i,j} = 1$ if nucleotides $i$ and $j$ can bind to make base pair $i \cdot j$ (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 base pairs) that could form. We define a parameter $m$ to be the minimum allowed stem length ($m \geq 1$; $m = 1$ throughout unless otherwise specified). We also impose the physical constraint that hairpins (single-

stranded region connecting one end of a stem) have a minimum length of 3 nucleotides. We include not only the longest possible stems that can form, but all contiguous subsets of those stems [56, 57]. We denote the number of stems found by $N_{\text{stems}}$.

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 $q$ ($C_{q,q} = 1 \ \forall \ q$). We impose the constraint that each nucleotide may be paired with, at most, one other nucleotide by setting $C_{p,q} = 0$ if stems $p$ and $q$ share at least one nucleotide.

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 $1 \leq s_1 < s_2 \leq N_{\text{stems}}$ and add the first stem for which $C_{s_1,s_2} = 1$. 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 say $M$ stems, to the list of possible structures, then 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 and all code is available on the GitHub repository `https://github.com/ofer-kimchi/RNA-FE-Landscape`.

Having completely enumerated the possible secondary structures, we calculate the probabilities that the RNA will fold into each of them by calculating their free energies.

## II. CALCULATING FREE ENERGIES

The probability of the RNA sequence folding into a given equilibrium structure $\sigma$ is given by the Boltzmann factor

$$p(\sigma) = \exp\left(-\beta F_\sigma\right)/Z \tag{1}$$

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: $\sum_\sigma p(\sigma) = 1$. Here $F_\sigma$, the free energy of structure $\sigma$, is a function of the energy $E_\sigma$ and entropy $S_\sigma$ of the structure:

$$\Delta F = \Delta E - T\Delta S \tag{2}$$

where we drop the subscripts for notational convenience and introduce $\Delta$s to signify that free energies are measured with respect to the free chain. We separate the free energy calculation into the free energy of stems and the free energy of loops.

### A. Calculating bond energies

We make the simplifying assumption that the energy $\Delta E$ in Eq. (2) is determined solely by the base pairs in the structure, ignoring higher order corrections to the energy. Thus, each stem, $s$, contributes an energy $\Delta E_s$ such that $\Delta E = \sum_s \Delta E_s$. To calculate the terms $\Delta E_s$, we consider nearest-neighbor interactions among base pairs [58]. Previous work has shown it reasonable to include (whenever appropriate) the contribution of unpaired nucleotides on both sides of each stem in the nearest-neighbor terms for the first and last base pairs of the stem [25]. Specifically, we used tabulated parameters for $\Delta H$ from Refs. [50, 59, 60], well documented by Turner and Mathews in the Nearest Neighbor Database [61]. Our entropy model (described below) 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 possible in multibranch loops.

### B. Calculating entropies

Entropies are calculated as being comprised of two independent parts: the entropic cost of forming stems and the entropic cost of forming loops, such that $\Delta S = \Delta S_{\text{loops}} + \sum_{\text{stems}} \Delta S_{\text{stem}}$.

The entropies of stems represent the entropy lost when an RNA forms base pairs. This entropy is considered in the same fashion as the energetic parameters (each energetic parameter has an accompanying entropic parameter). Therefore, as for the energies, the entropic parameters consider pairwise RNA base pair interactions, and $\Delta S_{\text{stem}}$ thus depends on the specific nucleotides comprising the stem. In contrast, we make the approximation that $\Delta S_{\text{loops}}$ is independent of the identities of the nucleotides comprising the single-stranded regions.

## III. CALCULATING LOOP ENTROPIES: RNA FEYNMAN DIAGRAMS

We model single-stranded regions comprised of $x$ unpaired nucleotides (ntds) as a random walk of $(x+1)/b$ steps, where $b \approx 2.4$ ntds is the Kuhn length of single stranded RNA [34, 62]. Since the entropic cost of forming base pairs has already been considered in $\Delta S_{\text{stem}}$, for the purposes of calculating $\Delta S_{\text{loops}}$ we consider stems as rigid rods. This approximation is justified because of the extremely long persistence length of double-stranded RNA ($\sim 200$ ntds [63]) compared to both single-stranded RNA and the length of any stem we consider.

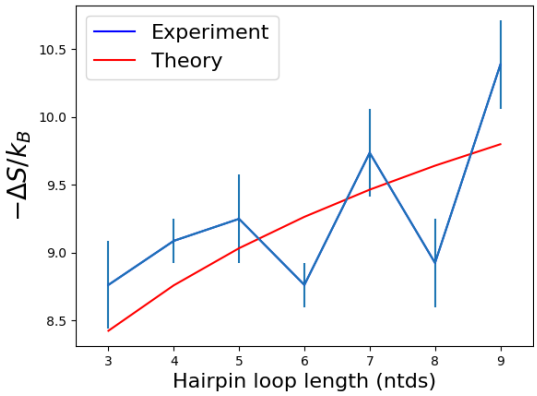


FIG. 2: $v_s$ **estimated from experimental data.** Experimental estimates for the free energy of hairpin loops of length $s$ from Table 1 of Ref. [64] were converted to entropy estimates (blue points and error bars) by assuming $\Delta H = 0$ as in Ref. [25]. These data were fit to Eq. (6), yielding an estimate of $v_s = 0.0201 \pm 0.0036$ ntds$^3$.

The entropy of a single-stranded region of length $s_i$ is given by $k_B \log \omega_i(s_i)$, where $\omega_i(s_i)$ is the number of ways of arranging the region consistent with the topology of the overall structure. Defining $\Omega(s)$ as the *total* number of conformations a random walk of length $s$ can take, for a free chain, $\omega = \Omega$. For structures which include constraints, $\omega(s_i) = \Omega(s_i) \times p(s_i)$, where $p(s_i)$ is the probability that the random walk of length $s_i$ will yield a conformation consistent with the topology of the overall structure being considered. Since free energies are measured relative to the free chain, factors of $\Omega$ cancel out in equations for $\Delta S_{\text{loops}}$ (see further discussion in Section S3). The entropy of the single-stranded regions in a given structure is thus given by

$$\Delta S_{\text{loops}} = \sum_i k_B \log p(s_i), \tag{3}$$

where $s_i$ is the number of nucleotides in the $i^{\text{th}}$ single-stranded region. The sum is generally over non-independent terms; we will describe how to address these sums via a Feynman diagram-like approach in this section.

As demonstrated in Eq. (3), the physics of the situation are held in $p(s)$, which is best calculated by considering the end-to-end vector of the random walk undergone by the single-stranded RNA, as

$$p(s) = \int_{\vec{R} \text{ consistent with overall structure}} P_s(\vec{R}) d\vec{R}, \tag{4}$$

where we define $P_s(\vec{R})d\vec{R}$ as the probability of a random walk of length $s$ to have end-to-end vector $\vec{R}$:

$$P_s(\vec{R}) = \left(\frac{3}{2\pi s b}\right)^{3/2} \exp\left(-\frac{3R^2}{2sb}\right). \tag{5}$$

We have assumed $s \gg b$ in order to arrive at the Gaussian formula above through the central limit theorem. The mean of the Gaussian is zero by symmetry. In order to find the variance we first consider a single step of length $b$ in three dimensions which has variance in the $x$, $y$, and $z$ 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. (5).

As described in Section S5 of the supplement, we can systematically consider higher order corrections to Eq. (5) while maintaining its Gaussian nature. Eq. (5) 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. While 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. (5) and experiment, and that corrections to Eq. (5) as described in Section S5 are negligible.

In order to demonstrate how Eqs. 3 - 5 are applied, we first consider the simple hairpin loop. Following Jacobson and Stockmayer [65], we allow that base pairing can occur as long as the two nucleotides are within a small volume $v_s = 4\pi \int_0^{r_s} R^2 dR$ of one another, where $r_s$ roughly corresponds to the bond length.[1] We assume that $r_s$ is small enough that for all $|\vec{R}| \leq r_s$, $P_s(\vec{R}) \approx P_s(\vec{0})$. Therefore, Eqs. 3 - 5 yield

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

We have called the LHS of the equation $S_{\text{closed-net-0}}$ (the zero references the lack of stems enclosed by the loop) following [34, 52] (rather than, say, $S_{\text{hairpin}}$) to emphasize that this formula is applicable to hairpin loops, bulge loops, internal loops, and multiloops – all of which can be thought of as closed loops of RNA. Aside from the appropriate inclusion of $v_s$ terms to account for the finite and variable width of RNA stems, RNA stems are treated as having negligible width by performing the approximation $P_s(|\vec{R}| < r_s) \approx P_s(0)$.

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

[1] More generally, we can define a probability $q(\vec{R})$ of a nucleotide at the origin being base paired with a nucleotide a vector $\vec{R}$ away. Then, $v_s$ is defined as $v_s = \int d\vec{R}\, q(\vec{R})$ and $r_s$ is the value of $|\vec{R}|$ for which $q(\vec{R})$ is non-negligible.

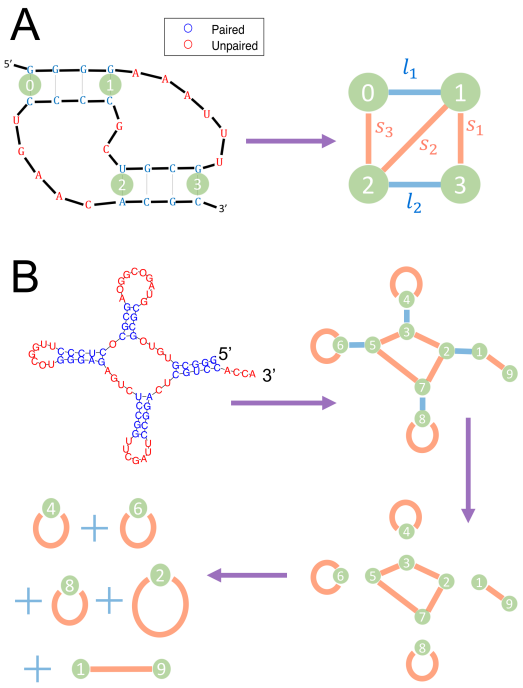


FIG. 3: **RNA Feynman Diagrams. (A): The Canonical Pseudoknot** An instance of the canonical H-type pseudoknot. Bold lines represent the RNA backbone; thin lines represent Hydrogen bonds. The entropy of this structure can be calculated by converting it to a graph format as shown in RHS of panel. The nodes of the graph represent the first and last base pairs of each stem, and two types of edges represent single- and double-stranded RNA. The graph directly represents the integral in Eq. (7). **(B): Graph Decomposition**. The entropy of a sample RNA structure (top left) can be computed by converting the structure to a graph as defined in the text (top right). The graph directly represents the integrals necessary to compute the entropy. Separable integrals are represented by graphs which can be disconnected by the removal of any one edge (bottom right). Thus, once appropriate factors of $v_s$ are included (one for each stem in the original structure), the entropy of the structure in question is given by (bottom left) the sum of four closed-nets-0 (originating from the three hairpins and multiloop) and one open-net-0.

of its length, the experimental measurements are non-monotonic, and their nonmonotonicity exceeds the error bars [64]. This non-monotonicity may be due to enthalpic effects [66] which were neglected in our analysis following Ref. [25]. Nevertheless, Fig. 2 shows that Eq. (6) gives a reasonable fit to the experimental data with $v_s = 0.0201 \pm 0.0036$ ntds$^3$.[2] If one ignores all angular dependences of bond formation, this 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.

Finally, we consider pseudoknots. To calculate the entropy of a pseudoknot of arbitrary complexity we invent a novel graph formulation inspired by Feynman diagrams from quantum field theory. 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 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 $v_s\delta(|\vec{r}_i-\vec{r}_j|-l)/4\pi l^2$, 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. Note that two bonded nucleotides 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. 3A (LHS). As we described, its conformational entropy can be calculated by translating the structure into a graph (Fig. 3A RHS), where each node represents the edge of a stem; blue edges represent regions of double-stranded RNA of length $l_i$; red edges represent regions of single-stranded RNA of length $s_i$. For example, here, $s_3 = 5$ ntds, and $l_1 = 3$ ntds. We set the origin of our coordinate system to node 0 and call the distance between node $i$ and the origin $r_i$. 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:

$$e^{\Delta S/k_B} = v_s^2 \int d\vec{r_1} \int d\vec{r_2} \int d\vec{r_3}\ \frac{\delta(|\vec{r_1}|-l_1)}{4\pi l_1^2} \times \frac{\delta(|\vec{r_3}-\vec{r_2}|-l_2)}{4\pi l_2^2} P_{s_1}(\vec{r_3}-\vec{r_1}) P_{s_2}(\vec{r_2}-\vec{r_1}) P_{s_3}(\vec{r_2}) \quad (7)$$

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 Section S4. This integral can be calculated analytically (Sec. S5) [34].

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. 3B for a classical cloverleaf RNA. The RNA in

[2] A more precise definition of $v_s$ might include a dependence on the closing base pairs of the hairpin loop; we expect that the penalties placed on specific closing base pairs and first mismatches in e.g. Refs. [64] and [25] play a similar role, though such penalties were not included here.

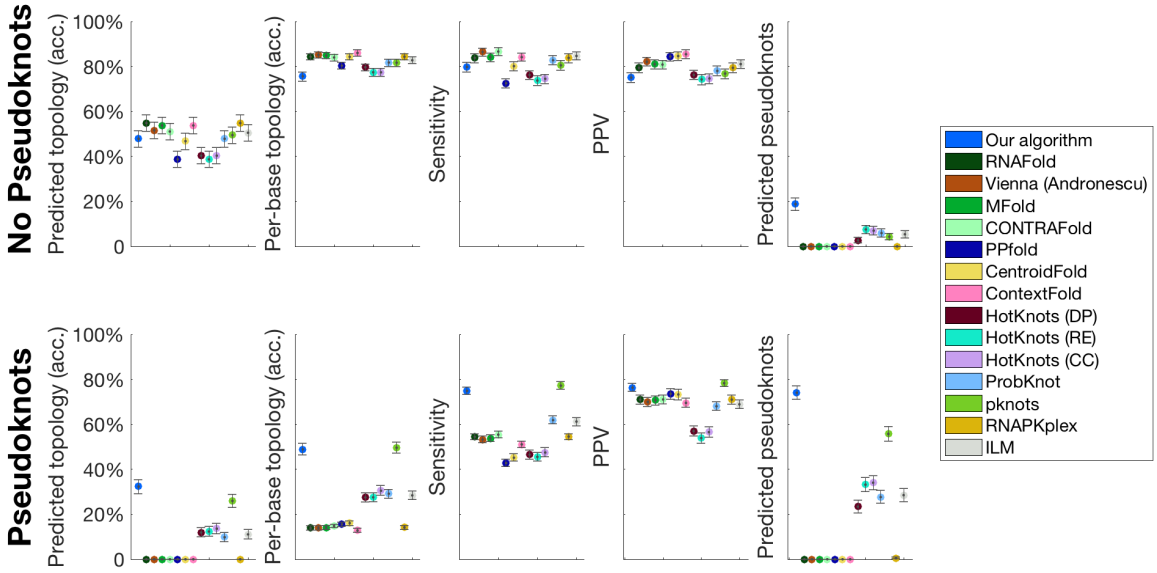


FIG. 4: **Summary statistics for comparison to other prediction tools**. To assess the relative success of our algorithm, we compare its performance to that of 14 other current prediction tools: RNAFold [29, 67], ViennaRNA (Andronescu parameters) [68], Mfold [28], CONTRAFold [69], PPfold [70], CentroidFold [71], ContextFold [72], HotKnots (Dirks-Pierce parameters), HotKnots (Rivas-Eddy parameters), HotKnots (Cao-Chen parameters) [51], ProbKnot [37], pknots [39], RNAPKplex [29, 67], and ILM [35]. 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 proportion of the time the MFE structure contains a pseudoknot. We separate the results for sequences which form into pseudoknots and those which don't. 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. See main text for further discussion.

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). As shown in the figure, once appropriate factors of $v_s$ are included in the integrals (one for each stem) the stems can be treated as having negligible width; thus, nodes which can be removed without changing the topology can be removed in the graph decomposition process. See Section S4 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; double-stranded with blue. For each graph, the integral formulation of its entropy is displayed in the figure alongside what it evaluates to.

## IV. COMPARISON WITH PUBLISHED TOOLS

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 ntds from the RNAStrand [53], PseudoBase++ [54], and CompraRNA [55] 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 base pairs correctly predicted by our algorithm's MFE structure compared to fourteen other current algorithms. Seven of these cannot predict pseudoknots and serve as useful benchmarks for the non-pseudoknotted results, (detailed methods in Section S1).

While 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. We similarly disallowed parallel stems which can be stable in neutral and acidic pH conditions [73]. We also set the minimum stem length for each sequence ($m$) to the minimum 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 $\sim$ an hour to run on a MacBook Pro 2012 laptop. Details of the computation time of our algorithm can be found in Fig. S4.

While these practical constraints were chosen to speed up the computation time, they also led to errors in the algorithm's predictions. 64 of the tested pseudoknots were topologically more complex than any of those presented in Fig. S2. Furthermore, 33 of the non-pseudoknotted sequences tested (and 8 of the pseudoknotted) include base pairs outside of those allowed by the algorithm (A·U, G·C, and G·U). 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. By changing $N_{\text{stems}}^{\text{max}}$ from 150 to 200, these numbers decreased to 46 for both pseudoknotted and non-pseudoknotted sequences, but the results for $N_{\text{stems}}^{\text{max}} = 200$ were practically identical to the results of Fig. 4 (see full results in Supplementary Table 1). For $N_{\text{stems}}^{\text{max}} = 200$, the computation time was increased significantly (to $\sim$ 17 hours for one sequence).

The sensitivity ($TP/TP + FN$) and 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 outperformed 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 Supplementary Table S1.

While sensitivity and PPV are the most common metrics used to establish the success of an RNA prediction algorithm [74], we sought to develop a test that measures success on the scale of the full RNA, rather than on the scale of individual base pairs. To this end, we measured how frequently each algorithm was able to correctly predict the topology of the experimentally measured structure, where the topology of a structure is defined by its graph (Section III). 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've defined it, if that stem includes one of the ends of the molecule. In order 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 nucleotides of Fig. 3A 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.

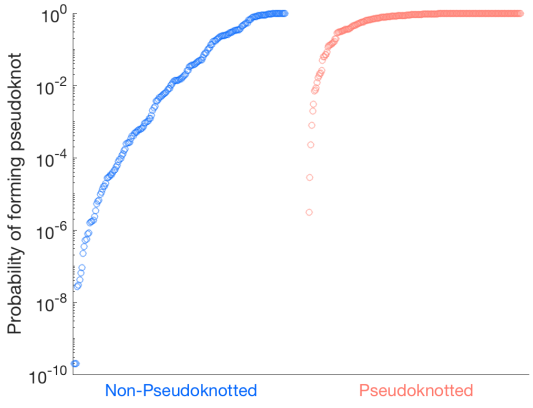


FIG. 5: **Probability of folding into a pseudoknot**. The predicted probability of each of the 421 sequences tested folding into a pseudoknot is presented. Of these sequences, 186 were experimentally found not to form pseudoknots (blue) and 235 were found to form pseudoknots (red). Our algorithm successfully 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 $2 \times 10^{-10}$.

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. We have found that the probability of predicting pseudoknots can be significantly decreased with minor changes in the Turner parameters energy function, and these parameters may need to be re-examined in order to be used most effectively with the entropy model presented here.

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 datapoint represents a different sequence and the total probability calculated of that sequence folding into a pseudoknotted structure. For figure clarity, a lower bound of pseudoknot probability was set at $2 \times 10^{-10}$.

The algorithm's predictions for the six longest RNA molecules less than 89 ntds 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)

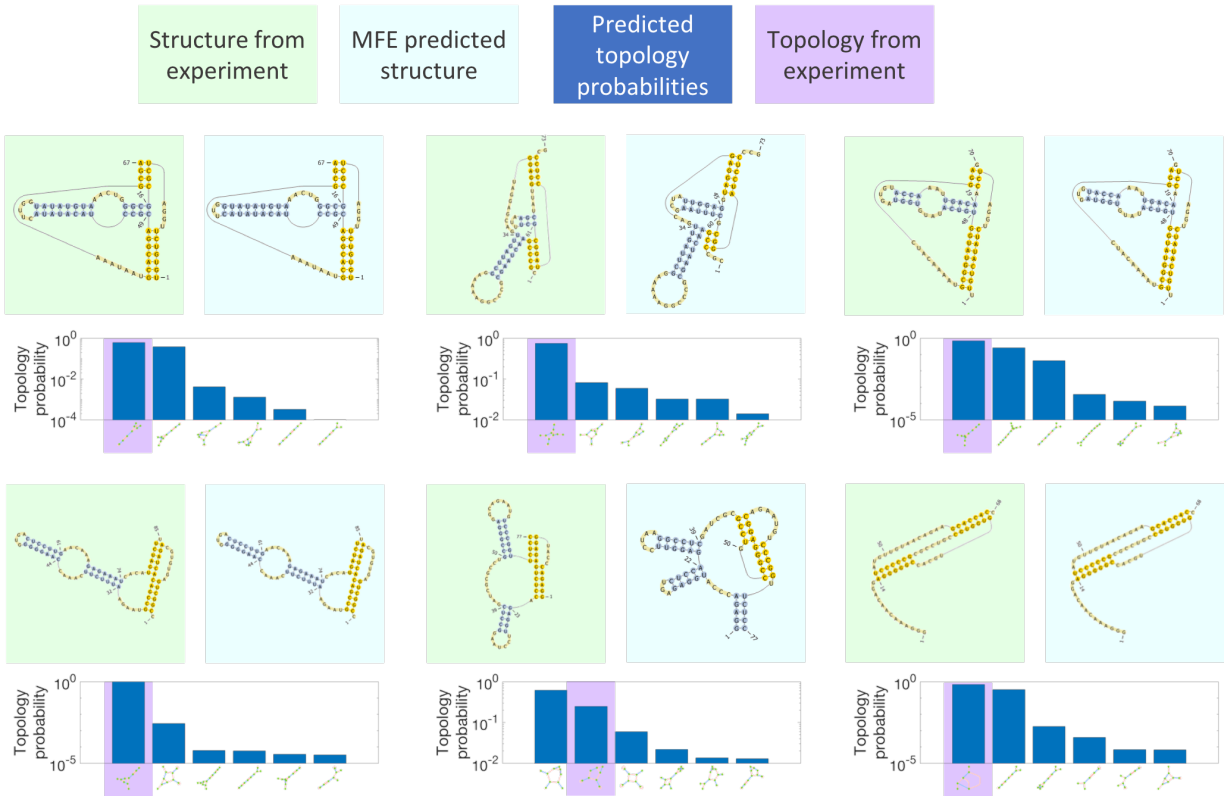


FIG. 6: **Comparison to experiments for long sequences**. Six long sequences were chosen from the Pseudobase++ database as described in the main text. The sequences are derived from (starting from the top left and moving across): tobacco mosaic virus [75–77], *Bacillus subtilis*, [78], tobacco mild green mosaic virus [76, 79], *Bacilis subtilis* [80], Giardiavirus [81], and Visna-Maedi virus [82]. We show the experimental structure (green background) and the MFE predicted structure (light blue background) plotted using the PseudoViewer software [83]. 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.

where the experimental topology is highlighted (purple). RNA secondary structure was plotted using the PseudoViewer package [83]. Our results demonstrate successful predictions even for long pseudoknotted sequences, especially in terms of the predicted topology. Detailed methods are provided in Section S1.

## V. DISCUSSION AND 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 the free energy landscape of an RNA molecule including complex pseudoknots. This NP-complete algorithm can be used to tackle even relatively long ($\sim$ 90 ntds) RNA sequences, and aside from the enumeration procedure (which is relatively fast for long sequences; see Fig. S4) 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 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. Although we have not done so in this work, we expect that our results can be even further improved by optimizing the energy function given the entropy model presented here. 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.

The algorithm presented here can also be easily generalized to probe multiple interacting strands (see discussion in supplement). The sequences considered can be any combination of DNA and RNA; their identities affects the energy parameters of the model which have

been previously tabulated, 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 [84–87], as well as in the search for novel RNAs [88, 89], and the description of similarity between RNA structures [90–93] which is especially useful in the study of the effects of mutations [94, 95]. A common approach among these graphical depictions of RNA has been to represent loops (e.g. hairpins, internal loops, etc.) as verticies and stems as edges [88, 92, 93]. However, this depiction of RNA does not always distinguish between pseudoknotted and non-pseudoknotted structures [88]. Other approaches have represented each nucleotide as a separate node and bonds (either hydrogen or covalent) as edges [89, 91]; while useful in many contexts (for example, secondary structure visualization), this approach does not have the benefit of coarse-graining to group similar structures as the same graph [90]. Our approach, described in Section III, can be viewed as a middle ground and may be useful in the contexts described previously.

## VI. ACKNOWLEDGEMENTS

We thank Elena Rivas and Yohai Bar Sinai for fruitful discussions. This research was funded by the National Science Foundation through the Harvard Materials Research Science and Engineering Center Grant DMR-1420570, DMREF Grant DMR-123869 and ONR Grant N00014-17-1-3029. OK acknowledges support from an NDSEG fellowship and Molecular Biophysics Training Grant NIH/NIGMS T32 GM008313 (PI: James M. Hogle). M.P.B. is an investigator of the Simons Foundation.

[1] Philipp Kapranov, Jill Cheng, Sujit Dike, David A. Nix, Radharani Duttagupta, Aarron T. Willingham, Peter F. Stadler, Jana Hertel, Jörg Hackermüller, Ivo L. Hofacker, Ian Bell, Evelyn Cheung, Jorg Drenkov, Erica Dumais, Sandeep Patel, Gregg Helt, Madhavan Ganesh, Srinka Ghosh, Antonio Piccolboni, Victor Sementchenko, Hari Tammana, and Thomas R. Gingeras. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, 316(5830):1484–1488, 2007.

[2] Yukinori Okada, Tomoki Muramatsu, Naomasa Suita, Masahiro Kanai, Eiryo Kawakami, Valentina Iotchkova, Nicole Soranzo, Johji Inazawa, and Toshihiro Tanaka. Significant impact of miRNA-target gene networks on genetics of human complex traits. *Scientific Reports*, 6:1–9, 2016.

[3] Bharat Sridhar, Marcelo Rivas-Astroza, Tri C. Nguyen, Weizhong Chen, Zhangming Yan, Xiaoyi Cao, Lucie Hebert, and Sheng Zhong. Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Current Biology*, 27(4):602–609, 2017.

[4] F. Butter, M. Scheibe, M. Morl, and M. Mann. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proceedings of the National Academy of Sciences*, 106(26):10626–10631, 2009.

[5] Stefan E Seemann, Susan M Sunkin, Michael J Hawrylycz, Walter L Ruzzo, and Jan Gorodkin. Transcripts with in silico predicted RNA structure are enriched everywhere in the mouse brain. *BMC Genomics*, 13(214), 2012.

[6] Tim R. Mercer, Marcel E. Dinger, and John S. Mattick. Long non-coding RNAs: Insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009.

[7] Michael T. McManus and Phillip A. Sharp. Gene silencing in mammals by small interfering RNAs. *Nature Reviews Genetics*, 3(10):737–747, 2002.

[8] Celina Juliano, Jianquan Wang, and Haifan Lin. Uniting Germline and Stem Cells: The Function of Piwi Proteins and the piRNA Pathway in Diverse Organisms. *Annual Review of Genetics*, 45(1):447–469, 2011.

[9] Andrew D. Ellington and Jack W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346:818–822, 1990.

[10] C Tuerk and L Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510, 1990.

[11] Debra L. Robertson and Gerald F. Joyce. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*, 344(6265):467–468, 1990.

[12] Charles Olea and Gerald F. Joyce. Real-Time detection of a self-replicating RNA Enzyme. *Molecules*, 21(10):1–12, 2016.

[13] Miriam H. Huntley, Arvind Murugan, and Michael P. Brenner. Information capacity of specific interactions. *Proceedings of the National Academy of Sciences*, 113(21):5841–5846, 2016.

[14] Vera Pancaldi and Jürg Bähler. In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic Acids Research*, 39(14):5826–5836, 2011.

[15] Jiamin Xiao, Yizhou Li, Kelong Wang, Zhining Wen, Menglong Li, Lifang Zhang, and Xuanmin Guang. In silico method for systematic analysis of feature importance in microRNA-mRNA interactions. *BMC Bioinformatics*, 10:1–13, 2009.

[16] Nancy Martínez-Montiel, Laura Morales-Lara, Julio M. Hernández-Pérez, and Rebeca D. Martínez-Contreras. In silico analysis of the structural and biochemical features of the NMD factor UPF1 in Ustilago maydis. *PLoS ONE*, 11(2):1–26, 2016.

[17] P O Ilyinskii, T Schmidt, D Lukashev, A B Meriin, G Thoidis, D Frishman, and A M Shneider. Impor-

tance of mRNA secondary structural elements for the expression of influenza virus genes. *Omics*, 13(5):421–430, 2009.

[18] R. A. Poot, N. V. Tsareva, I. V. Boni, and J. van Duin. RNA folding kinetics regulates translation of phage MS2 maturation gene. *Proceedings of the National Academy of Sciences*, 94(19):10110–10115, 1997.

[19] Maximillian H Bailor, Xiaoyan Sun, and Hashim M. Al-Hashimi. Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science*, 327(5962):202–206, 2010.

[20] J M Pipas and J E McMahon. Method for predicting RNA secondary structure. *Proceedings of the National Academy of Sciences*, 72(6):2017–2021, 1975.

[21] Michael S. Waterman. Secondary Structure of Single-Stranded Nucleic Acidst. *Studies in Foundations and Combinatorics, Advances in Mathematics Supplementary Studies*, 1:167–212, 1978.

[22] Michael S. Waterman and Temple F. Smith. Rapid dynamic programming algorithms for RNA secondary structure. *Advances in Applied Mathematics*, 7(4):455–464, 1986.

[23] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978.

[24] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

[25] Martin J. Serra and Douglas H. Turner. Predicting thermodynamic properties of RNA. *Methods in Enzymology*, 259:242–261, 1995.

[26] David H. Mathews and Douglas H. Turner. Prediction of RNA secondary structure by free energy minimization. *Current Opinion in Structural Biology*, 16(3):270–278, 2006.

[27] C. E. Hajdin, S. Bellaousov, W. Huggins, C. W. Leonard, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences*, 110(14):5498–5503, 2013.

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

[29] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, 125(2):167–188, 1994.

[30] Raheleh Salari, Chava Kimchi-Sarfaty, Michael M. Gottesman, and Teresa M. Przytycka. Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: Application to disease studies. *Nucleic Acids Research*, 41(1):44–53, 2013.

[31] Rune B. Lyngsø and Christian N. S. Pedersen. Pseudoknots in RNA Secondary Structures. *Proceedings of the fourth annual international Conference on Computational Molecular Biology,*, pages 201–209, 2000.

[32] Rune B. Lyngsø and Christian N. S. Pedersen. RNA Pseudoknot Prediction in Energy-Based Models. *Journal of Computational Biology*, 7(3-4):409–427, 2000.

[33] Biao Liu, David H Mathews, and Douglas H Turner. RNA pseudoknots: folding and finding. *F1000 Biology Reports*, 5(January):1–5, 2010.

[34] 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.

[35] Jianhua Ruan, Gary D. Stormo, and Weixiong Zhang. An Iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1):58–66, 2004.

[36] Jihong Ren, Baharak Rastegari, Anne Condon, and Holger H Hoos. HotKnots : Heuristic prediction of RNA secondary structures including pseudoknots HotKnots : Heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11(1):1494–1504, 2005.

[37] S. Bellaousov and D. H. Mathews. ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *RNA*, 16(10):1870–1880, 2010.

[38] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu, and Kiyoshi Asai. IPknot: Fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):85–93, 2011.

[39] Elena Rivas and Sean R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285(5):2053–2068, 1999.

[40] Yasuo Uemura, Aki Hasegawa, Satoshi Kobayashi, and Takashi Yokomori. Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science*, 210(2):277–303, 1999.

[41] Tatsuya Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1-3):45–62, 2000.

[42] Anne Condon, Beth Davy, Baharak Rastegari, Shelly Zhao, and Finbarr Tarrant. Classifying RNA pseudoknotted structures. *Theoretical Computer Science*, 320(1):35–50, 2004.

[43] Robert M Dirks and Niles A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24(13):1664–1677, 2003.

[44] Jens Reeder and Robert Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5:1–12, 2004.

[45] Song Cao and Shi Jie Chen. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Research*, 34(9):2634–2652, 2006.

[46] Song Cao and Shi-Jie Chen. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA*, 15(4):696–706, 2009.

[47] F H van Batenburg, a P Gultyaev, C W Pleij, J Ng, and J Oliehoek. PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Research*, 28(1):201–204, 2000.

[48] Daniel P. Aalberts and Nathan O. Hodas. Asymmetry in RNA pseudoknots: Observation and theory. *Nucleic Acids Research*, 33(7):2210–2214, 2005.

[49] Adam Lucas and Ken A. Dill. Statistical mechanics of pseudoknot polymers. *Journal of Chemical Physics*, 119(4):2414–2421, 2003.

[50] D H Mathews, J Sabina, M Zuker, and D H Turner. Expanded sequence dependence of thermodynamic paramenters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[51] Mirela S Andronescu, Cristina Pop, and Anne E Condon. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA*, 16(1):26–42, 2010.

[52] 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.

[53] Mirela Andronescu, Vera Bereg, Holger H. Hoos, and Anne Condon. RNA STRAND: The RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 9:1–10, 2008.

[54] Michela Taufer, Abel Licon, Roberto Araiza, David Mireles, F. H D van Batenburg, Alexander P. Gultyaev, and Ming Ying Leung. PseudoBase++: An extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic Acids Research*, 37(SUPPL. 1):127–135, 2009.

[55] Tomasz Puton, Lukasz P. Kozlowski, Kristian M. Rother, and Janusz M. Bujnicki. CompaRNA: A server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Research*, 41(7):4307–4323, 2013.

[56] Gary M Studnicka, Georgia M Rahn, Ian W Cummings, and Winston A Salser. Computer method for predicting the secondary structure of single-stranded RNA. *Nucleic Acids Research*, 5(9):3365–3388, 1978.

[57] Michael Zuker and D Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, 1984.

[58] William Bialek and Rama Ranganathan. Rediscovering the power of pairwise interactions. *arXiv*, 2007.

[59] Tianbing Xia, John SantaLucia, Mark E. Burkard, Ryszard Kierzek, Susan J. Schroeder, Xiaoqi Jiao, Christopher Cox, and Douglas H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson - Crick base pairs. *Biochemistry*, 37(42):14719–14735, 1998.

[60] Tianbing Xia, David H. Mathews, and Douglas H. Turner. Thermodynamics of RNA Secondary Structure Formation. In Dieter Soll, Susumu Nishimura, and Peter B. Moore, editors, *RNA*, chapter 2, pages 21–48. Pergamon, 1 edition, 2001.

[61] Douglas H. Turner and David H. Mathews. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38(SUPPL.1):2009–2011, 2009.

[62] S. B. Smith, Y. Cui, and C. Bustamante. Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules. *Science*, 271(5250):795–799, 1996.

[63] J. A. Abels, F. Moreno-Herrero, T. Van Der Heijden, C. Dekker, and Nynke H. Dekker. Single-molecule measurements of the persistence length of double-stranded RNA. *Biophysical Journal*, 88(4):2737–2744, 2005.

[64] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19):7287–7292, 2004.

[65] 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.

[66] Zhi John Lu, Douglas H. Turner, and David H. Mathews. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Research*, 34(17):4912–4924, 2006.

[67] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):1–14, 2011.

[68] Mirela Andronescu, Anne Condon, Holger H. Hoos, David H. Mathews, and Kevin P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13):19–28, 2007.

[69] Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):90–98, 2006.

[70] Zsuzsanna Sükösd, Bjarne Knudsen, Jorgen Kjems, and Christian N.S. Pedersen. PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics*, 28(20):2691–2692, 2012.

[71] Kengo Sato, Michiaki Hamada, Kiyoshi Asai, and Toutai Mituyama. CentroidFold: A web server for RNA secondary structure prediction. *Nucleic Acids Research*, 37(SUPPL. 2):277–280, 2009.

[72] S Zakov, Y Goldberg, M Elhadad, and M Ziv-Ukelson. Rich parameterization improves RNA structure prediction. *Journal of Computational Biology*, 18(11):1525–1542, 2011.

[73] V Rani Parvathy, Sukesh R Bhaumik, Kandala V R Chary, Girjesh Govil, Keliang Liu, Frank B Howard, and H Todd Miles. NMR structure of a parallel-stranded DNA duplex at atomic resolution. *Nucleic Acids Research*, 30(7):1500–1511, 2002.

[74] Z. J. Lu, J. W. Gloor, and D. H. Mathews. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, 15(10):1805–1813, 2009.

[75] K Rietveld, K Linschooten, C W Pleij, and L Bosch. The three-dimensional folding of the tRNA-like structure of tobacco mosaic virus RNA. A new building principle applied twice. *The EMBO journal*, 3(11):2613–9, 1984.

[76] Ruud M.W. Mans, Cornelis W.A. Pleij, and Leendert Bosch. tRNAlike structures: Structure, function and evolutionary significance. *European Journal of Biochemistry*, 201(2):303–324, 1991.

[77] B Felden, C Florentz, R Giegé, and E Westhof. 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(3):201–12, 1996.

[78] Garrett A. Soukup. Core requirements for glmS ribozyme self-cleavage reveal a putative pseudoknot structure. *Nucleic Acids Research*, 34(3):968–975, 2006.

[79] Fernando García-Arenal. Sequence and structure at the genome 3' end of the U2-strain of tobacco mosaic virus, a histidine-accepting tobamovirus. *Virology*, 167(1):201–206, 1988.

[80] S R Wilkinson and M D Been. A pseudoknot in the 3' non-core region of the glmS ribozyme enhances self-cleavage activity. *RNA*, 11(12):1788–1794, 2005.

[81] Srinivas Garlapati and Ching C. Wang. Identification

of an essential pseudoknot in the putative downstream internal ribosome entry site in giardiavirus transcript. *RNA*, 8(5):601–611, 2002.

[82] Simon Pennell, Emily Manktelow, Andrew Flatt, Geoff Kelly, Stephen J Smerdon, and Ian Brierley. The stimulatory RNA of the Visna-Maedi retrovirus ribosomal frameshifting signal is an unusual pseudoknot with an interstem element. *RNA*, 14(7):1366–77, 2008.

[83] Yanga Byun and Kyungsook Han. PseudoViewer3: Generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics*, 25(11):1435–1437, 2009.

[84] Denise R. Koessler, Debra J. Knisley, Jeff Knisley, and Teresa Haynes. A predictive model for secondary RNA structure using graph theory and a neural network. *BMC Bioinformatics*, 11(SUPPL. 6):1–10, 2010.

[85] Michaël Bon and Henri Orland. TT2NE: A novel algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Research*, 39(14), 2011.

[86] Henri Orland and A. Zee. RNA folding and large N matrix theory. *Nuclear Physics B*, 620(3):456–476, 2002.

[87] Jizhen Zhao, Russell L. Malmberg, and Liming Cai. Rapid ab initio prediction of RNA pseudoknots via graph tree decomposition. *Journal of Mathematical Biology*, 56(1-2):145–159, 2008.

[88] Hin Hark Gan, Samuela Pasquali, and Tamar Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Research*, 31(11):2926–2943, 2003.

[89] Christian Laing and Tamar Schlick. Computational approaches to RNA structure prediction, analysis, and design. *Current Opinion in Structural Biology*, 21(3):306–318, 2011.

[90] C. Haslinger and P. F. Stadler. RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bulletin of Mathematical Biology*, 61(3):437–467, 1999.

[91] Clara I. Bermúdez, Edgar E. Daza, and Eugenio Andrade. Characterization and comparison of Escherichia coli transfer RNAs by graph theory based on secondary structure. *Journal of Theoretical Biology*, 197(2):193–205, 1999.

[92] Giorgio Benedetti and Stefano Morosetti. A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. *Biophysical Chemistry*, 59(1-2):179–184, 1996.

[93] Shu Yun Le, Ruth Nussinov, and Jacob V. Maizel. Tree graphs of RNA secondary structures and their comparisons. *Computers and Biomedical Research*, 22(5):461–473, 1989.

[94] Walter Fontana and Peter Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280(5368):1451–1455, 1998.

[95] Lauren W Ancel and Walter Fontana. Plasticity, Evolability and Modularity in RNA. *Journal of Experimental Zoology*, 288(3):242–283, 2000.

[96] 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.

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

[98] Hatim T. Allawi and John SantaLucia. Thermodynamics and NMR of internal G·T mismatches in DNA. *Biochemistry*, 36(34):10581–10594, 1997.

[99] Hatim T. Allawi and John SantaLucia. Nearest neighbor thermodynamic parameters for internal G·A mismatches in DNA. *Biochemistry*, 37(8):2170–2179, 1998.

[100] Hatim T. Allawi and John SantaLucia. Thermodynamics of internal C·T mismatches in DNA. *Nucleic Acids Research*, 26(11):2694–2701, 1998.

[101] Hatim T. Allawi and John SantaLucia. Nearest-neighbor thermodynamics of internal A·C mismatches in DNA: Sequence dependence and pH effects. *Biochemistry*, 37(26):9435–9444, 1998.

[102] Nicolas Peyret, P. Ananda Seneviratne, Hatim T. Allawi, and John SantaLucia. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A·A, C·C, G·G, and T·T mismatches. *Biochemistry*, 38(12):3468–3477, 1999.

[103] 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.

[104] 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.

**Supporting Information**

The supporting information is divided into several sections. In Section S1 we detail the methods used to compare our algorithm's performance to other current models. In Section S2 we discuss how our algorithm can be easily generalized to probe multiple interacting strands including any combination of DNA and RNA. In Section S3 we give a further discussion of the $\Omega(s)$ term defined in Section II. 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. Finally, in Fig. S4 we show how our algorithm's properties scale with the length of the sequence for random sequences between 10 and 21 ntds in length.

## S1. DETAILED METHODS FOR 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.

The sequences tested were downloaded from the Pseudobase++, RNAStrand, and CompraRNA PDB databases. We constrained database searches to return results only for sequences of length $\leq 80$ 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 $\geq 0.2$ Jukes-Cantor sequence dissimilary measured using the MatLab command `seqpdist`. 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). This resulted in $m = 1$ for 22% sequences, $m = 2$ for 33% of 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. 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 [96, 97]. Following Ref. [28] 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 [98–102]. In addition, our algorithm can probe DNA-RNA bonds using the parameter sets from Refs. [103, 104], 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.

## S3. DISCUSSION OF $\Omega(s)$

In Section II, we defined a parameter $\Omega(s)$ to be the total number of conformations a random walk of length $s$ can take. $\Omega(s)$ has the property that $\Omega(s_1)\Omega(s_2) = \Omega(s_1 + s_2)$.

For a free chain of length $s$, $\vec{R}$ can take on any value, as long as $R \equiv |\vec{R}| \leq s$. Taking the limit $s \gg b$ (so that the integral extends over all of space), and using the normalization of $P_s(\vec{R})$, we find $S_{\text{free}} = k_B \log \Omega(s)$, as expected from the definition of $\Omega(s)$. Therefore, in order to calculate changes in free energy compared to $S_{\text{free}}$, we simply omit $\Omega(s)$ terms from our formulae.

To be more precise, this argument directly demonstrates only why $\Omega(s)$ terms should cancel out for non-base-paired RNA. However, it motivates us to consider the experimentally measured entropy of each base pair as being multiplied by a factor of $\Omega(2)$ for the two nucleotides comprising it. Including such terms, all factors of $\Omega$ drop out of expressions representing physical results. We therefore can compute expressions for $\Delta S$, the difference in entropy between a given structure and a free chain, by omitting factors of $\Omega$ from the relevant formulae.

## S4. DERIVING EQ. 7

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

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

$\delta^3(|x| - a)$, 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 $S_{\text{free}}$, and thus disappear in the formula for $\Delta S$.

One can check that introducing a new node representing any nucleotide in the structure (say a node on the edge between nodes 0 and 3) does not affect the result.

$q(\vec{r})$ 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. [65] we wish to include an upper bound for the bond length, $r_s$, $q(\vec{r})$ becomes a Heaviside $\Theta$ 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).

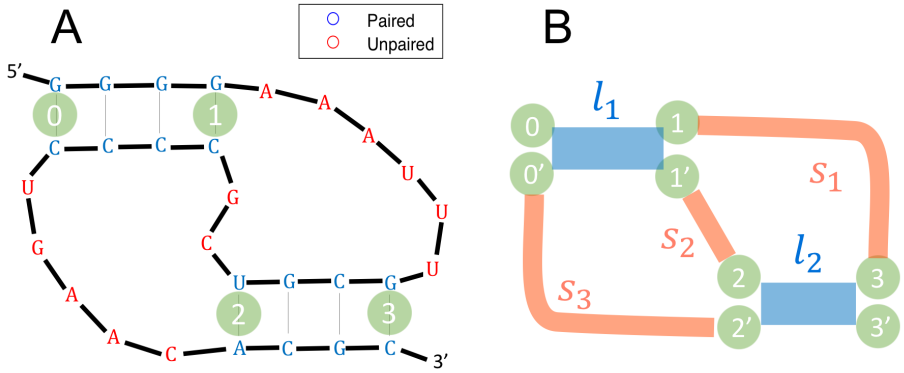


FIG. S1: **A preliminary description of an H-type pseudoknot. A:** An instance of the canonical H-type pseudoknot, reprinted from Fig. 3. **B:** A preliminary version of the graph representing its entropy. In Sec. S4 we demonstrate that this graph is equivalent to that shown in Fig. 3A.

## 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 $(\vec{r_3} - \vec{r_2}) + \vec{r_2}$.

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

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

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

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

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

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

Finally, we perform the integrals remaining to arrive at

$$e^{\Delta S/k_B} = \frac{v_s^2\gamma^2 \exp\left(-\frac{\gamma\left(l_1^2(s_1+s_2)+l_2^2(s_2+s_3)\right)}{s_1s_2+s_1s_3+s_2s_3}\right)}{2\pi^3 l_1 l_2 s_2\sqrt{s_1s_2+s_1s_3+s_2s_3}} \times \sinh\left(\frac{2\gamma l_1 l_2 s_2}{s_1s_2+s_1s_3+s_2s_3}\right)$$

where sinh is the hyperbolic sine function. This formula is equivalent to the one presented without proof in Ref. [34].

## 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

$$P_2(\vec{R}) = \int d\vec{R}_1 P_1(\vec{R}_1) P_1(\vec{R} - \vec{R}_1). \tag{S2}$$

The form of Eq. (S2) 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

$$\hat{p}_1(\omega) = \frac{\sin(\omega b)}{\omega b} \tag{S3}$$

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

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

$$\hat{p}_n(\omega) = (\sin(\omega b)/\omega b)^n\,. \tag{S4}$$

Taking the inverse Fourier transform, we find

$$P_n(\vec{R}) = \frac{2}{(2\pi)^2}\int_0^\infty d\omega\ \omega^2 \left(\frac{\sin(\omega b)}{\omega b}\right)^n \frac{\sin(\omega R)}{\omega R}. \tag{S5}$$

At this point, we use our assumption that $n$ is large. This formula tends to zero for large values of $\omega b$, and we therefore Taylor expand the sin function for small $\omega b$. 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. (S5) 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 $-\infty$.

$$P_n(\vec{R}) = \frac{1}{(2\pi)^2 iR}\int_{-\infty}^{\infty} d\omega\ \omega e^{-n\left[\frac{(\omega b)^2}{6}+\frac{(\omega b)^4}{180}+\mathcal{O}(\omega b)^6\right]+i\omega R}. \tag{S6}$$

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

$$e^{-n\left[\frac{(\omega b)^4}{180}\right]} = 1 - \frac{n(\omega b)^4}{180} + \mathcal{O}(\omega b)^8. \tag{S7}$$

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. (S6) becomes a Gaussian integral, which can be solved analytically to yield

$$P_n(\vec{R}) = \left(\frac{3}{2\pi s b}\right)^{3/2} e^{\left(-\frac{3R^2}{2sb}\right)} \left[1 - \frac{3\left(5s^2b^2 - 10sbR^2 + 3R^4\right)}{20s^3b}\right]. \tag{S8}$$

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. (S8). Keeping only first-order correction terms, and defining $\vec{R}_{21} = \vec{R}_2 - \vec{R}_1$,

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

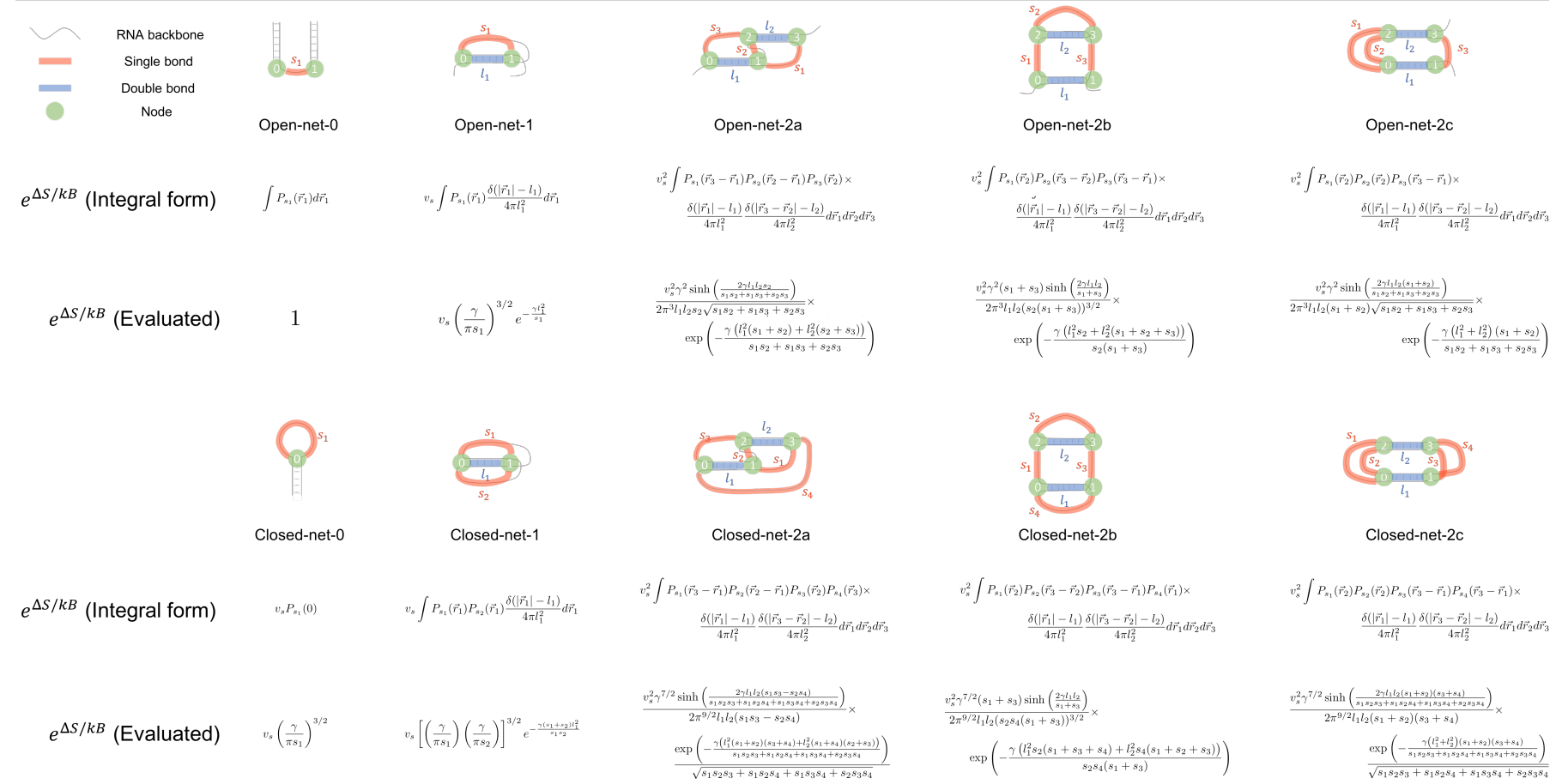


FIG. S2: **Graphs of simple RNA structures.** The 10 graphs with at most two regions of double-stranded RNA and their corresponding RNA backbones are displayed alongside integral and evaluated expressions for the entropy of each graph. Note that stems shown as parallel could be antiparallel if the system considered is comprised of more than one strand. See Fig. 2 of Ref. [52] for comparison.

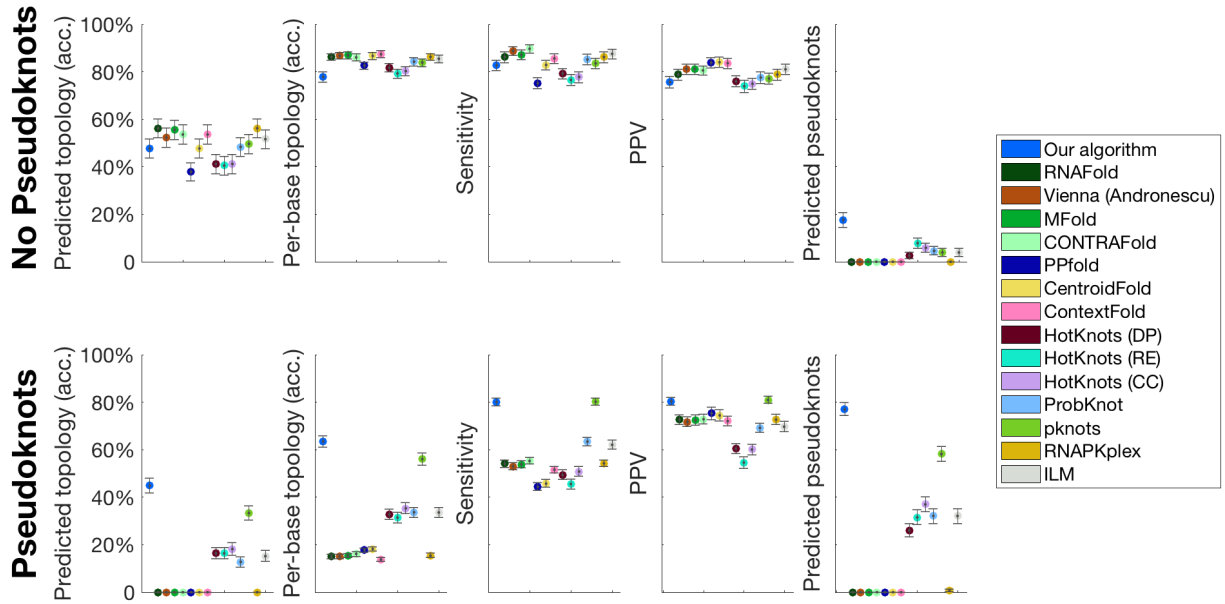


FIG. S3: **Results only including sequences whose structure our algorithm could have predicted.** We consider only the 153 non-pseudoknotted and 165 pseudoknotted sequences whose structures do not include base pairs or topologies disallowed by our algorithm. In this case, we predict the correct topology 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.

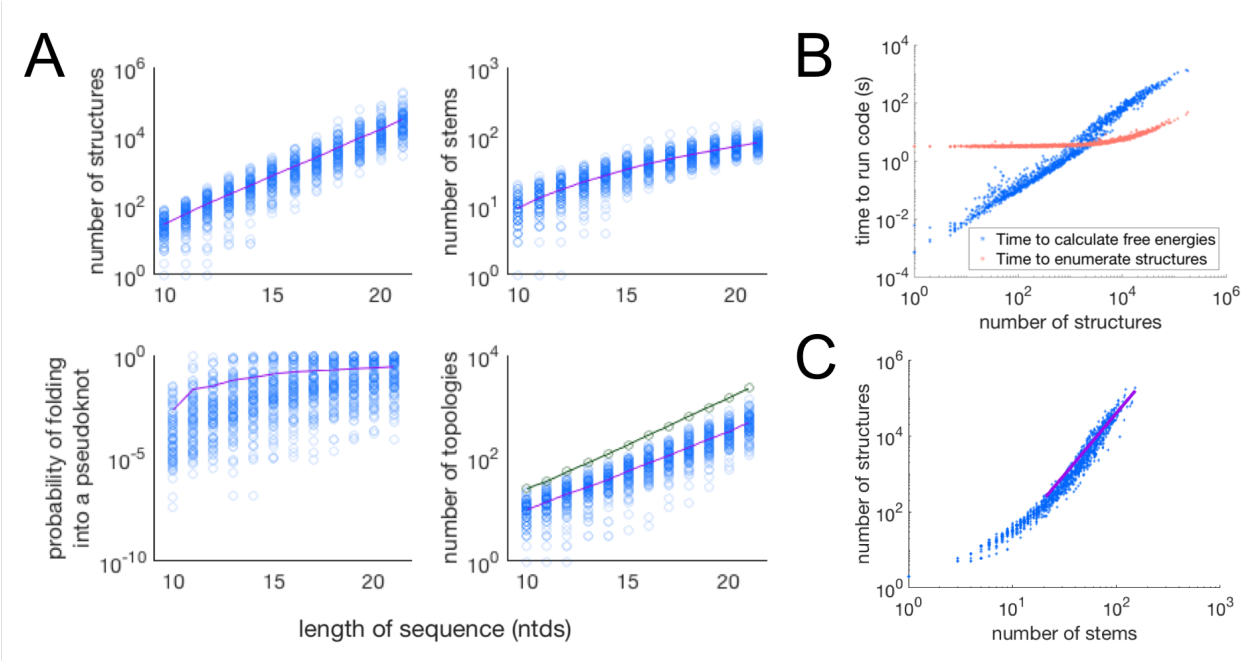


FIG. S4: **Scaling of the algorithm properties with length of sequence.** We input 100 random sequences for each length between 10 and 21 nucleotides into the algorithm. **(A)** Various properties of the results are plotted as a function of the length of the sequence. Blue circles are datapoints for 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 NP-complete 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. 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. **(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 $b = 3.24 \pm 0.11$.