PRISM: Partition-Function Decomposition into Structural Classes for Hierarchically Constrained RNA Pseudoknot Ensembles

Mateo Gray
Department of Biomedical Engineering, University of Alberta, Edmonton, Canada

Sebastian Will
Department of Computer Science, Institut Polytechnique de Paris, Palaiseau, France

Hosna Jabbari[1]
Department of Biomedical Engineering, University of Alberta, Edmonton, Canada

Abstract

While structure ensemble analysis became a valuable routinely applied tool for pseudoknot-free RNA, the extension to pseudoknots remains challenging due to the computational hardness of the general problem. The existing efficient algorithms for the computation of partition function with pseudoknots were still computationally expensive and were restricted to simple pseudoknots. This changed only with CParty, which computes pseudoknotted partition functions with the efficiency of pseudoknot-free folding. At its core, CParty follows the hierarchical folding hypothesis, such that ensemble structures can form pseudoknots only with a given input constraint structure. For an RNA sequence S and pseudoknot-free structure G, CParty limits the ensemble to "density-2" structures $G \cup G'$ for a second, disjoint pseudoknot-free structure G'.

We present PRISM that extends CParty from pure partition function calculation to full-fledged posterior probability analysis. By stochastic traceback through CParty's dynamic programming matrices, it samples structures from the conditional Boltzmann ensemble. From estimated base pair probabilities, it generates ensemble representations, predicts centroid and maximum expected accuracy structures and calculates properties. In addition to position-specific summaries, PRISM maps sampled structures to RNA shapes, producing a posterior distribution over topological abstractions. This shape-level summary captures ensemble diversity even when a conserved pseudoknotted motif appears with shifted base-pair positions across samples.

We validate PRISM in the pseudoknot-free limit, where it reproduces RNAFold quantities for minimum free energy, ensemble free energy, centroid expected distance, and maximum expected accuracy. We further show that stochastic traceback recovers Boltzmann structure probabilities and that sampling error decreases at the expected Monte Carlo rate while runtime grows linearly with the number of samples. Our case study demonstrate that RNA-shape summaries can reveal dominant pseudoknotted topologies that centroid decoding may miss. PRISM thus converts the CParty partition function into a practical framework for posterior decoding and topology-aware analysis of hierarchically constrained pseudoknotted RNA ensembles.

2012 ACM Subject Classification Applied computing → Computational biology

Keywords and phrases RNA, MFE, Secondary Structure Prediction, Pseudoknot, Partition Function, Centroid, MEA, RNA shape, Stochastic traceback

Digital Object Identifier 10.4230/LIPIcs.WABI.2026.30

Supplementary Material *Software (PRISM's implementation and detailed Results)*: `https://github.com/TheCOBRALab/PRISM` [10]
archived at `swh:1:dir:79803e70b92d8457ba08d16fe67b09382501716c`
Dataset: `https://github.com/mateog4712/PRISM-RawData` [11]
archived at `swh:1:dir:4646c334d25f499a6485f62fe913bb90e2cdd115`

[1] Corresponding author

1 Introduction

RNA molecules perform diverse biological functions whose mechanisms often depend on their structural conformations [5, 17, 21, 30, 32]. Structured RNA regions can regulate translation, control pre-mRNA splicing, mediate catalytic activity, and organize long-range transcript architectures. In these settings, RNA structure is not only a consequence of sequence composition, but part of the molecular mechanism by which an RNA exposes or hides functional elements, recognizes binding partners, stabilizes alternative conformations, or catalyzes chemical reactions. Computational models of RNA structure are therefore central to the interpretation of RNA function, the prioritization of experimental hypotheses, and the identification of structural elements relevant to disease or therapeutic design. Because experimental structure determination remains labor-intensive, condition-dependent, and difficult to scale to transcriptome-wide settings, computational prediction and ensemble analysis remain indispensable complements to experimental approaches.

RNA structure ensembles and pseudoknot classes. Let $S = s_1 \cdots s_n$ be an *RNA sequence* of n *nucleotides* (also called, *bases*). An *RNA secondary structure* R is a set of *base pairs* (i, j), $1 \le i < j \le n$, such that each nucleotide participates in at most one pair. Two base pairs (i, j) and (k, ℓ) *cross* if $i < k < j < \ell$ or $k < i < \ell < j$. A structure is considered *pseudoknot-free* if it contains no crossing base pair; otherwise, it is *pseudoknotted*. For a secondary structure class $\mathcal{C}(S)$, secondary structure prediction algorithms assign a *free energy* $E_S(R)$ to each $R \in \mathcal{C}(S)$. For a structure R of S, define its *Boltzmann weight* $w_S(R) = \exp(-E_S(R)/(\mathcal{R}T))$, depending on the absolute temperature T via the gas constant $\mathcal{R}$. Then, the *partition function* $Z_{\mathcal{C}}(S) = \sum_{R \in \mathcal{C}(S)} w_S(R)$ defines the *Boltzmann distribution* in the *Boltzmann ensemble* $\mathcal{C}(S)$ by

$$\mathbf{P}_{\mathcal{C}}[R \mid S] = \frac{w_S(R)}{Z_{\mathcal{C}}(S)} \qquad \text{for } R \in \mathcal{C}(S).$$

Note that the structures with *minimum free energy (mfe)* $\min_{R \in \mathcal{C}} E_S(R)$ have maximum probability in this distribution. The *posterior probability* of the base pair (i, j) is defined as

$$\mathbf{P}_{\mathcal{C}}\big[(i,j) \in R\big] = \frac{1}{Z_{\mathcal{C}}(S)} \sum_{\substack{R \in \mathcal{C}(S) \\ (i,j) \in R}} w_S(R).$$

Such base pair probabilities characterize the base pairing propensities in the thermodynamic equilibrium beyond reporting single predicted structures and are oftentimes used for visualization in dot plots.

Ensemble analysis. For the class of pseudoknot-free structures, established dynamic-programming algorithms provide the toolkit for computing partition functions, base-pair probabilities, stochastic traceback samples, centroid structures, and maximum expected accuracy (MEA) structure [20, 7, 6, 8, 18]. These posterior summaries are often more informative than a single MFE structure as they distinguish well-supported base pairs from uncertain or competing alternatives.

Pseudoknots substantially complicate this ensemble analysis problem. Predicting the minimum free energy secondary structure of pseudoknotted RNA is NP-hard in general [19, 1] and even inapproximable [25]. Existing pseudoknotted MFE prediction algorithms, therefore restrict the class of structures they handle or simplify the energy model [16, 13, 14].

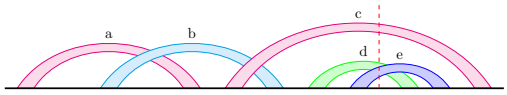

Figure 1 Bands of a density-2 structure. A pseudoknot can be decomposed into bands, each of which is internally non-crossing but crosses the remaining structure. All base pairs within a band share the same crossing pattern with respect to the rest of the structure. Bands that mutually cross one another are grouped into components, here {a,b,c} and {d,e}. A set of bands is density-k, if every position is covered by at most k bands of the same component. An example of this is shown at the dashed red line [15, 16].

Hierarchical folding: HFold and CParty. Particularly fast pseudoknot folding algorithms were proposed based on the biologically motivated hierarchical folding hypothesis: an RNA molecule first folds into a pseudoknot-free structure; additional base pairs are subsequently added to further minimize the structure's energy, allowing the resulting structure to contain crossing base pairs [28]. The tool HFold for hierarchical pseudoknot prediction, finds the MFE structure given a pseudoknot-free input structure G [13, 14]. HFold's class of structures is called *density-2* structures (Fig. 1), a subclass of general bisecondary structures [33]. This large structure class includes kissing hairpins and chains of pseudoknots; moreover, it recursively allows density-2 substructures.

CParty extended HFold's hierarchical model from MFE prediction to partition-function computation [9]. Given a sequence S and a fixed pseudoknot-free input structure G, CParty considers all structures of the form $G \cup G'$, where G' is pseudoknot-free, $G \cap G' = \emptyset$, and $G \cup G'$ is a valid density-2 structure. Equivalently, CParty defines the conditional ensemble

$$\Omega[S, G] = \{G \cup G' : G' \text{ is pseudoknot-free}, \ G \cap G' = \emptyset, \ G \cup G' \text{ is density-2}\},$$

and computes $Z(S \mid G) = \sum_{R \in \Omega[S,G]} w_S(R)$. The induced conditional Boltzmann distribution is

$$\mathbf{P}_{\Omega[G]}[R \mid S, G] = \frac{w_S(R)}{Z(S \mid G)} \qquad \text{for } R \in \Omega[S, G].$$

Thus, CParty provides the normalizing constant and constrained ensemble free energy for $\Omega[S, G]$. However, the scalar value $Z(S \mid G)$ does not by itself identify which base pairs are well supported, which nucleotides are likely unpaired, whether the posterior mass is concentrated on pseudoknotted or pseudoknot-free extensions, or whether distinct base-pair-level structures represent the same higher-level topology.

Pseudoknot ensemble analysis: PRISM. In this work, we introduce PRISM, a posterior-decoding and ensemble-analysis framework built on CParty. The name PRISM reflects the central operation of the method: whereas CParty computes one constrained partition function, PRISM separates the same Boltzmann ensemble into interpretable structural classes and estimates the posterior mass of those classes. Here, a structural class is any subset $\mathcal{A} \subseteq \Omega[S, G]$ defined by a structural property of interest. For example, $\mathcal{A}$ may be the set of structures containing a particular base pair, the set of structures in which a nucleotide is unpaired, the set of structures containing at least one pseudoknot, or the set of structures with a particular RNA shape. The Boltzmann weight assigned to such a class is $Z_{\mathcal{A}}(S \mid G) = \sum_{R \in \mathcal{A}} w_S(R)$, and its posterior probability under the conditional ensemble is

$$\mathbf{P}_{\Omega[G]}[R \in \mathcal{A} \mid S, G] = \frac{Z_{\mathcal{A}}(S \mid G)}{Z(S \mid G)}.$$

If a collection of classes $\mathcal{A}_1, \ldots, \mathcal{A}_m$ is disjoint and covers $\Omega[S, G]$, then the CParty partition function decomposes as

$$Z(S \mid G) = \sum_{\ell=1}^{m} Z_{\mathcal{A}_\ell}(S \mid G).$$

Thus, the same partition function can be viewed not only as a scalar normalizing constant, but also as the total Boltzmann weight distributed across structural classes.

PRISM does not recompute these restricted partition functions exactly. Instead, it draws stochastic traceback samples from $\mathbf{P}_G(\cdot \mid S, G)$ and estimates class posterior probabilities by sample frequencies. For a structural class $\mathcal{A}$, the Monte Carlo estimate is

$$\widehat{\mathbf{P}}[\mathcal{A} \mid S, G] = \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}\left\{R^{(t)} \in \mathcal{A}\right\}.$$

The classes considered in this paper include base-pair events, unpaired-nucleotide events, pseudoknot-conditioned subensembles, decoded structures, and RNA-shape abstractions. PRISM estimates base-pair and unpaired probabilities, constructs a marginal structural-state string, computes centroid and pseudoknot-conditioned centroid structures, computes the MEA structure, and summarizes the ensemble at the level of RNA shapes [26]. The shape-level view is especially useful when a pseudoknotted motif is conserved topologically but its exact base-pair endpoints shift across sampled structures.

Contributions. We introduce PRISM, an algorithmic posterior analysis layer for the hierarchically constrained density-2 ensemble computed by CParty. We first formalize posterior analysis for this ensemble by making explicit the conditional Boltzmann distribution over admissible extensions $G \cup G'$. We then introduce a posterior class-weight view of the CParty partition function: rather than using the partition function only as a scalar normalizing constant, PRISM estimates how its Boltzmann weight is distributed across interpretable structural classes, including base-pair events, unpaired-nucleotide events, pseudoknot-conditioned subensembles, decoded structures, and RNA-shape abstractions. From these probabilities, PRISM computes posterior summaries for the constrained pseudoknot ensemble, including a marginal structural-state string, centroid structures, pseudoknot-conditioned centroid structures, maximum expected accuracy structures, and dot-plot visualizations. Finally, we introduce shape-level summaries of sampled structures, which provide a topology-aware view of ensemble diversity that is robust to small shifts in helix position. We validate the implementation in the pseudoknot-free limit against RNAFold, examine sampling convergence and runtime overhead, and show that RNA-shape summaries can reveal dominant pseudoknotted ensemble features that base-pair-level centroid decoding may miss.

2 Methods

The input to PRISM is an RNA sequence $S = s_1 \cdots s_n$ and a fixed pseudoknot-free input structure G. PRISM utilizes the conditional partition function $Z(S \mid G)$ over the admissible density-2 ensemble $\Omega[S, G]$, and samples structures $R^{(1)}, R^{(2)}, \ldots, R^{(N)} \in \Omega[S, G]$ from the conditional Boltzmann distribution $\mathbf{P}_{\Omega[G]}(R \mid S, G) = \frac{w_S(R)}{Z(S|G)}$. Unless otherwise stated, we use $N = 1000$ stochastic traceback samples to estimate ensemble statistical properties.

Let $\mathcal{B}_G(S)$ denote the set of base pairs that can occur in at least one structure in $\Omega[S, G]$. The posterior probability of a base pair $(i, j) \in \mathcal{B}_G(S)$ in the conditional ensemble is defined in analogy to unconditional base pair probabilities as

$$p_{ij} = \mathbf{P}_{\Omega[G]}[(i, j) \mid S, G] = \sum_{R \in \mathcal{C}(S), (i,j) \in R} \mathbf{P}_{\Omega[G]}[R \mid S, G].$$

Although the probabilities p_{ij} conditionally depend on G, we abstain from more verbose notation like $p_{i,j|G}$, since G will be always clear from the context. Note that in practice, PRISM always estimates base pair probabilities from sampling by stochastic traceback and operates on estimates $\widehat{p}_{ij}$. In the following, we use the hat notation only where this distinction is directly relevant.

2.1 Stochastic traceback

For pseudoknot-free RNA structures, exact base-pair probabilities can be computed by dynamic-programming recurrences derived from the partition function [20, 18]. An alternative is representative sampling: structures are drawn from the Boltzmann ensemble, and base-pair frequencies in the sample are used as Monte Carlo estimates of posterior base-pair probabilities [7]. PRISM extends this representative-sampling strategy to the hierarchically constrained ensemble computed by CParty.

Stochastic traceback follows the same decomposition structure as a standard dynamic-programming traceback. At each traceback step, however, the algorithm samples one of the possible recurrence cases according to its conditional Boltzmann weight. Suppose a dynamic-programming state A has recurrence $Z_A = \sum_{c \in \mathcal{D}(A)} W_c$, where $\mathcal{D}(A)$ is the set of decomposition cases for A, and W_c is the Boltzmann contribution of case c, including the weights of any subproblems introduced by that case. During traceback from state A, PRISM chooses case c with probability

$$\mathbf{P}[c \mid A] = \frac{W_c}{Z_A}.$$

The algorithm then recursively traces back through the subproblems specified by the selected case. At the top level, this procedure samples a complete structure $R \in \Omega[S, G]$ with probability $w_S(R)/Z(S \mid G)$. Thus, the stochastic traceback samples are drawn from the same conditional Boltzmann distribution whose partition function is computed by CParty.

The correctness of this sampling rule follows from the usual recursive decomposition argument. Each dynamic-programming state is a Boltzmann-weighted sum over mutually exclusive decomposition cases. Choosing each case with probability equal to its contribution divided by the state partition function makes the probability of any complete traceback path equal to the product of the conditional probabilities along that path. The intermediate state partition functions cancel telescopically, leaving exactly the Boltzmann weight of the completed structure divided by $Z(S \mid G)$.

Operationally, PRISM draws a uniform random number and compares it with the cumulative Boltzmann weights of the recurrence cases. To reduce the cost of repeated cumulative-weight scans, we use the Boustrophedon sampling strategy [22]. This improves the efficiency of sampling without changing the target Boltzmann distribution.

Given N sampled structures, the posterior probability of base pair (i, j) is estimated as

$$\widehat{p}_{ij} = \frac{f_{ij}}{N} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}\left\{(i, j) \in R^{(t)}\right\},$$

where f_{ij} is the number of sampled structures containing (i, j). This directly yields estimates of the unpaired probabilities p_i^u, since

$$p_i^u = 1 - \sum_{j<i} p_{ji} - \sum_{j>i} p_{ij}.$$

2.2 Marginal structural-state string

PRISM constructs a marginal structural-state string from the estimated posterior probabilities. This string is a position-wise annotation of the ensemble: for each nucleotide, PRISM reports the most probable local structural state of that nucleotide. It is not a globally decoded RNA structure, and its brackets are not required to define a valid set of base pairs.

We use the alphabet { . (,), [,] } corresponding respectively to the five local states: unpaired (u), non-crossing left endpoint (NL) non-crossing right endpoint (NR), crossing left endpoint (CL), and crossing right endpoint (CR). The classification of a base pair as crossing or non-crossing is understood relative to the fixed input structure G. Let

$$\chi_G(i,j) = \begin{cases} 1, & \text{if } (i,j) \text{ crosses at least one base pair in } G, \\ 0, & \text{otherwise.} \end{cases}$$

All base pairs in G are therefore classified as non-crossing, since G itself is pseudoknot-free. For each position i, PRISM computes

$$\widehat{p}_i^{\,\mathrm{NL}} = \sum_{\substack{j>i \\ \chi_G(i,j)=0}} \widehat{p}_{ij}, \qquad \widehat{p}_i^{\,\mathrm{NR}} = \sum_{\substack{j<i \\ \chi_G(j,i)=0}} \widehat{p}_{ji},$$

and

$$\widehat{p}_i^{\,\mathrm{CL}} = \sum_{\substack{j>i \\ \chi_G(i,j)=1}} \widehat{p}_{ij}, \qquad \widehat{p}_i^{\,\mathrm{CR}} = \sum_{\substack{j<i \\ \chi_G(j,i)=1}} \widehat{p}_{ji}.$$

The symbol at position i is chosen by taking the maximum among

$$\widehat{p}_i^{\,u}, \quad \widehat{p}_i^{\,\mathrm{NL}}, \quad \widehat{p}_i^{\,\mathrm{NR}}, \quad \widehat{p}_i^{\,\mathrm{CL}}, \quad \widehat{p}_i^{\,\mathrm{CR}}.$$

Thus, the marginal structural-state string summarizes the most likely state of each nucleotide under the conditional ensemble. Unlike a sampled structure, centroid structure, or MEA structure, it is not obtained by optimizing a global structural objective; it reports local marginal posterior states.

2.3 Ensemble diversity

PRISM computes ensemble diversity from the estimated base-pair probabilities. In analogy to the common definition for non-crossing ensembles, we define the ensemble diversity $\langle d \rangle$ of the conditional ensemble as expected distance between two structures

$$\langle d \rangle(S \mid G) = \sum_{R \in \Omega[G], R' \in \Omega[G]} \mathbf{P}[R \mid S, G] \cdot \mathbf{P}[R' \mid S, G] \cdot d_{\mathrm{BP}}(R, R')$$

Defining the base pair distance d_{BP} as symmetric difference

$$d_{\mathrm{BP}}(R, R') = |R \setminus R'| + |R' \setminus R|,$$

let us compute ensemble diversity from base pair probabilities

$$\langle d \rangle(S \mid G) = \sum_{(i,j) \in \mathcal{B}_G(S)} 2 \cdot p_{ij} \left(1 - p_{ij}\right).$$

This statistic is small when most admissible base pairs have posterior probability near 0 or 1, and it increases when posterior mass is spread across competing alternatives. Each term is maximized at $p_{ij} = 1/2$, so the statistic emphasizes uncertain or mutually competing contacts.

2.4 Centroid structures

The centroid structure is a posterior decoder that minimizes expected base-pair distance to the ensemble [6]. Using posterior base-pair probabilities, the expected distance from $Y \in \Omega[S, G]$ to the conditional ensemble is

$$\widehat{d}_G(Y) = \sum_{R \in \Omega[S,G]} \mathbf{P}_{\mathcal{C}}[R \mid S, G] \cdot d_{\mathrm{BP}}(R, Y) = \sum_{(i,j) \in Y} (1 - p_{ij}) + \sum_{(i,j) \in \mathcal{B}_G(S) \setminus Y} p_{ij}.$$

Thus, we obtain the centroid structure Y^* that minimizes this expected distance by selecting all base pairs (i, j) with $p_{ij} > 0.5$ and excluding all others.

Note that the analogous construction of the centroid in the pseudoknot-free case, necessarily yields a pseudoknot-free centroid structure. In the case of PRISM, where $p_{i,j}$ are probabilities in the conditional ensemble of G, one observes the following properties: Since G is fixed, base pairs in G have probability 1. For the remaining base pairs of the centroid Y^* in $Y^* \setminus G$, we can argue as in the non-crossing case, that they cannot cross each other. Therefore, the centroid structure is bi-secondary. Note that we forgo the stricter claim for density-2 in this contribution to avoid technical digressions.

Pseudoknot-conditioned centroid. PRISM also computes a pseudoknot-conditioned centroid. This decoder uses the same centroid objective, but estimates probabilities only from sampled structures that contain at least one crossing pair. Let

$$\mathcal{T}_{\mathrm{PK}} = \left\{ t \in \{1, \ldots, N\} : R^{(t)} \text{ contains at least one crossing pair} \right\},$$

and let $N_{\mathrm{PK}} = |\mathcal{T}_{\mathrm{PK}}|$. If $N_{\mathrm{PK}} = 0$, no pseudoknot-conditioned centroid is reported. Otherwise, the pseudoknot-conditioned base-pair probability is estimated as

$$\widehat{p}_{ij}^{\mathrm{PK}} = \frac{1}{N_{\mathrm{PK}}} \sum_{t \in \mathcal{T}_{\mathrm{PK}}} \mathbf{1}\left\{ (i, j) \in R^{(t)} \right\}.$$

The pseudoknot-conditioned centroid is computed by applying the same expected-distance decoder to $\widehat{p}_{ij}^{\mathrm{PK}}$ instead of $\widehat{p}_{ij}$. This conditioning removes sampled structures without pseudoknots from the probability estimates, so pseudoknotted alternatives are not diluted by pseudoknot-free structures.

This decoder still operates at the level of exact base-pair identities. If a pseudoknotted helix is present in many sampled structures but shifts by one or two nucleotides across samples, its probability can be split among nearby contacts such as (i, j), $(i, j-1)$, and $(i+1, j)$. In that case, no individual crossing base pair may exceed the 50% centroid threshold, even though the pseudoknotted motif is common in the ensemble. This limitation motivates the shape-level summaries described below.

2.5 Maximum expected accuracy structure

The maximum expected accuracy (MEA) decoder selects a structure that maximizes the expected number of correctly predicted paired and unpaired positions [8]. For a candidate structure Y, let $U(Y)$ be the set of nucleotides unpaired in Y. Given a base-pair weight parameter $\gamma > 0$, PRISM defines the expected accuracy of Y as

$$EA_\gamma(Y) = \sum_{(i,j) \in Y} 2\gamma p_{ij} + \sum_{i \in U(Y)} p_i^u.$$

The MEA structure is $Y_{\mathrm{MEA}} = \arg\max_Y EA_\gamma(Y)$, where the maximization is over structures compatible with the hierarchical decomposition.

PRISM uses the $G \cup G'$ form of hierarchical folding to reduce this calculation to the standard pseudoknot-free MEA calculation. Since G is fixed, its base pairs are present in every admissible candidate structure. The variable part of the prediction is the pseudoknot-free extension G'. PRISM therefore optimizes the MEA objective over the pseudoknot-free extension and then adds the fixed input structure G. Base pairs in G contribute a constant term to the objective, and nucleotides paired in G are not eligible to be unpaired in the candidate structure. The pseudoknot-free MEA recurrence is standard [8, 18], so we do not reproduce it here.

2.6 RNA shape abstraction

RNA shapes provide an abstraction of secondary structure that groups structures according to higher-level topology rather than exact base-pair positions [26, 23]. This is useful for posterior analysis because small shifts in helix endpoints can produce many distinct base-pair-level structures while preserving the same overall structural organization.

For pseudoknot-free structures, Steffen et al. [26] defined several abstraction levels. At low abstraction levels, some unpaired regions and loop details are retained. At abstraction level 5, unpaired nucleotides are removed and stacks and internal loops are collapsed so that only the nesting pattern of helices remains. For example,

```
((.((..((...)))).((...))..)).
```
(Level 1) `(_(_(_))(_)_)_`

(Level 5) `(()())`

PRISM uses an analogous level-5 abstraction for pseudoknotted structures. Given a sampled structure R, we represent non-crossing base pairs using `()` and base pairs that cross the input structure G using `[]`. We then delete unpaired nucleotides and collapse each maximal stack of base pairs into a single base pair, preserving the order and crossing pattern of the remaining helices. Both crossing and non-crossing helices are retained. We denote the resulting shape by $\sigma_5(R)$.

This construction is related to, but distinct from, the *shadow* abstraction of Reidys et al. [23]. Their shadow construction removes non-crossing arcs, removes isolated vertices, and collapses the remaining stacks. In our notation, this corresponds to a further abstraction, denoted $\sigma_6(R)$, in which non-crossing helices are removed after the level-5 representation. For example,

```
((..((...))..[[..)).]]
```
(Level 5) `(()[)]`

(Level 6) `([)]`

Thus, σ_5 retains both the pseudoknotted and non-pseudoknotted components of the topology, whereas σ_6 focuses only on the crossing core. We use σ_5 as the default shape abstraction for ensemble summaries because it preserves more information about the complete structural organization while still removing local positional variation.

Given sampled structures $R^{(1)}, \ldots, R^{(N)}$, PRISM estimates the posterior probability of each shape a by

$$\widehat{q}(a) = \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}\Big\{\sigma_5(R^{(t)}) = a\Big\}.$$

The resulting distribution over shapes characterizes the topology-level diversity of the conditional ensemble. Structures that differ by small shifts in base-pair endpoints can map to the same shape, so shape frequencies can reveal dominant pseudoknotted motifs even when no individual shifted base pair has sufficiently high probability to appear in the centroid.

3 Datasets and input constraints

For the empirical time and space analysis of PRISM, we obtained 2824 sequences from the RNASTRAND V2.0 database [2]. The median sequence length was 242 nucleotides, and the longest sequence contained 1500 nucleotides. For each sequence, we identified the 20 most stable stem-loop structures by scoring candidate stem loops using hairpin and stacking base-pair energies across the sequence. These stem-loop structures were used as fixed input constraints G for PRISM.

For the shape-based case study, we used the ZMP–ZTP riboswitch family from Rfam, identified by RF01750 [12]. We partitioned the Rfam consensus structure for this family into its pseudoknot-free and pseudoknotted components. We denote the pseudoknot-free component by G_{big} and the pseudoknotted component by G_{small}. Each of these two consensus-derived structures was fitted to each ungapped sequence in the family, and the fitted structures were used as alternative input constraints for PRISM.

4 Empirical Results

We evaluate PRISM in three stages. First, we validate correctness in the pseudoknot-free limit $G = \emptyset$, where $\Omega[S, G]$ is the standard secondary-structure ensemble and PRISM should agree with RNAFold under the same energy model. This validation includes both deterministic posterior summaries and stochastic traceback: we compare MFE energy, ensemble free energy, centroid expected distance, and MEA score, and we verify that the observed frequency of the MFE structure in sampled structures matches its Boltzmann probability. Second, we measure empirical runtime and memory usage for the full PRISM workflow and quantify how the stochastic traceback sample count affects runtime and posterior-estimation error. Third, we evaluate topology-level summaries by mapping sampled structures to level-5 RNA shapes. These shape summaries demonstrate how PRISM separates the conditional partition function into interpretable structural classes and can reveal pseudoknotted motifs or alternative conformations that are not apparent from base-pair-level centroid decoding alone.

4.1 PRISM agrees with RNAFold in the pseudoknot-free limit

When the input structure is empty, $G = \emptyset$, the conditional ensemble $\Omega[S, G]$ reduces to the standard pseudoknot-free secondary-structure ensemble. Consequently, $Z(S \mid \emptyset)$ should agree with the pseudoknot-free partition function computed by RNAFold under the same energy model and parameter settings. As a first validation of the implementation, we compared the MFE energy, ensemble free energy, centroid distance, and MEA score computed by PRISM with the corresponding quantities computed by RNAFold [18]. All comparisons used DP09 parameters [3] without dangle energies. PRISM reproduced the RNAFold values to numerical precision across all four quantities (Fig. 2).

We next validated the stochastic traceback procedure. For a structure R, the Boltzmann probability predicted by the partition function is $w_S(R)/Z(S \mid G)$. Therefore, for the MFE structure R_{MFE}, the observed frequency of R_{MFE} among stochastic traceback samples should match $\frac{w_S(R_{\text{MFE}})}{Z(S|G)}$. We performed this comparison for sequences with length $n \leq 100$. Longer

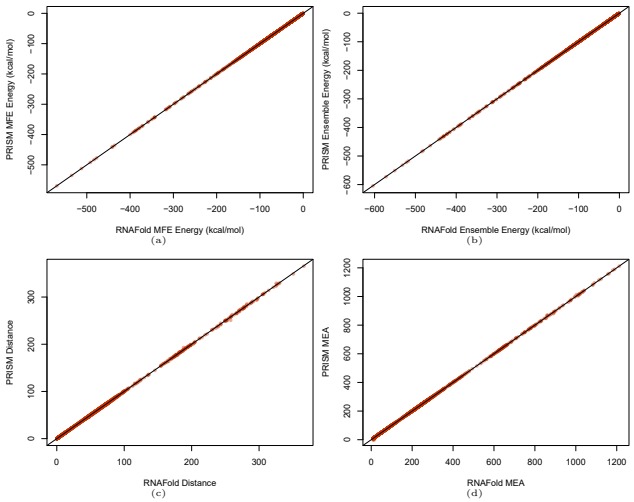

Figure 2 Validation of PRISM against RNAFold in the pseudoknot-free limit $G = \emptyset$. For each sequence, the PRISM output is plotted against the RNAFold output under the same energy parameters. Panels show (a) MFE energy, (b) ensemble free energy, (c) centroid expected distance to the ensemble, and (d) MEA score. Points on the diagonal indicate agreement between the two implementations.

sequences were excluded from this validation because the Boltzmann mass of any single structure, including the MFE structure, is typically very small, making direct frequency comparisons less informative at the sample sizes used. The observed MFE frequencies closely matched their partition-function probabilities (Fig. 3), supporting that stochastic traceback samples from the intended conditional Boltzmann distribution.

4.2 PRISM has empirical time and memory scaling close to RNAFold

We next evaluated the empirical time and memory requirements of the PRISM workflow when calculating pseudoknot partition functions. The measured workflow included computation of the constrained partition function $Z(S \mid G)$, stochastic traceback, centroid decoding, and MEA decoding. We compared PRISM's performance with the pseudoknot-free RNAFold and with NUPACK 3.0. To our knowledge, NUPACK is the only available a tool for computing unconstrained pseudoknot partition functions. It considers the class of simple pseudoknots in $\mathrm{O}(n^5)$ time and $\mathrm{O}(n^4)$ space complexity [34]. Since PRISM requires both a sequence S and an input structure G, we used the stem-loop constraints generated as described in Section 3. The same constrained inputs were provided to RNAFold; NUPACK 3.0 does not support structure constraints and was run on sequence alone.

All experiments were run on an M2 Max Mac Studio with 64 GB of RAM. Runtime was measured as user CPU time, and memory usage was measured as maximum resident set size. Across the benchmark, the maximum runtime observed for PRISM was 23.608 seconds, and the maximum memory usage was 293024 KB. For RNAFold, the corresponding maximum runtime and memory usage were 7.674 seconds and 56256 KB, respectively.

The empirical scaling behavior is shown in Fig. 4. For sequence lengths $n \geq 250$, the fitted runtime exponent for PRISM was close to that of RNAFold and substantially below that of NUPACK 3.0 over the shared range of lengths. The fitted memory exponent showed the same qualitative pattern. Thus, despite extending the pseudoknot-free posterior-analysis toolkit to the hierarchically constrained density-2 ensemble, PRISM remains close to RNAFold in empirical scaling and below NUPACK 3.0 in both runtime and memory usage on this benchmark.

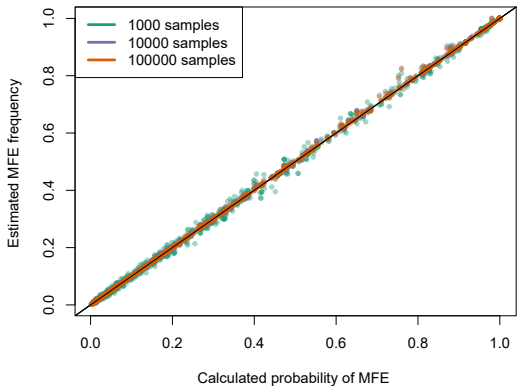

Figure 3 Validation of stochastic traceback using the MFE structure. The x-axis gives the Boltzmann probability of the MFE structure computed from its contribution to the partition function, and the y-axis gives the observed frequency of the MFE structure in stochastic traceback samples. Results are shown for sequences with $n \leq 100$. The diagonal line $y = x$ denotes exact agreement.

4.3 Sampling introduces linear runtime overhead and Monte Carlo convergence

Because PRISM estimates base-pair probabilities from stochastic traceback samples, the number of samples controls a tradeoff between runtime and posterior-estimation accuracy. To quantify this tradeoff, we evaluated three representative sequence lengths: 100, 210, and 320. For each length, we measured runtime as a function of the number of samples N_{samp}. Runtime increased linearly with N_{samp}, consistent with a fixed cost for filling the dynamic-programming tables followed by a per-sample traceback cost (Fig. 5a).

We also measured the accuracy of estimated base-pair probabilities as a function of N_{samp}. For each sample count, we computed the root mean square deviation (RMSD) between the base-pair probabilities estimated from that number of samples and the corresponding estimates obtained from 100000 samples, which were used as a reference:

$$\text{RMSD}(N_{\text{samp}}) = \sqrt{\frac{1}{|\mathcal{B}_G(S)|} \sum_{(i,j)\in\mathcal{B}_G(S)} \left(\widehat{p}_{ij}^{(N_{\text{samp}})} - \widehat{p}_{ij}^{(100000)}\right)^2}.$$

The RMSD decreased consistently with increasing sample count and followed the expected Monte Carlo rate of approximately $N_{\text{samp}}^{-1/2}$ (Fig. 5b) [4]. The observed RMSD was 0.0027 at 1000 samples and 0.001 at 10000 samples, indicating that 1000 samples already provided accurate posterior estimates for the downstream summaries considered here.

4.4 RNA shapes can recover pseudoknotted motifs missed by centroid decoding

The sampled structures produced by PRISM can be summarized either at the level of exact base pairs or at the level of RNA shapes. The centroid structure is a base-pair-level posterior decoder: a base pair contributes positively to the centroid objective only when its posterior probability exceeds 1/2. This makes the centroid useful for identifying well-supported individual contacts, but it can miss structural motifs whose base pairs shift across nearby positions. In particular, if a pseudoknotted helix occurs in most sampled structures but its

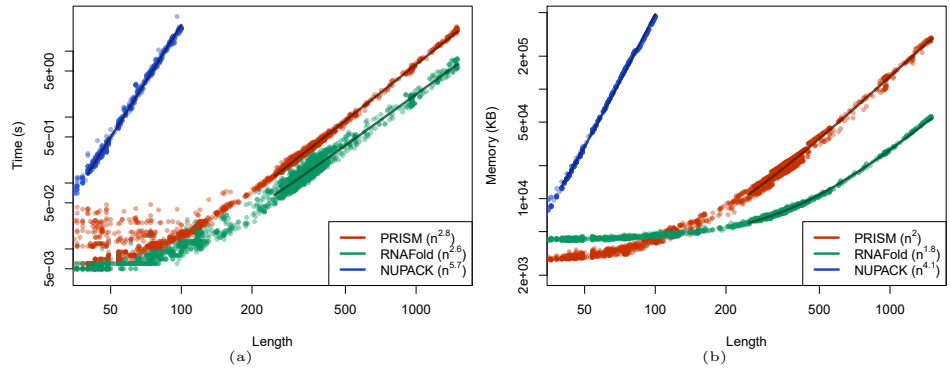

Figure 4 Runtime and memory usage of PRISM, RNAFold, and NUPACK 3.0 on the benchmark dataset with structure constraints. Runtime and memory are shown as functions of input length on log–log axes. Empirical scaling was estimated by fitting power laws for lengths $n \geq 250$, except for NUPACK 3.0, where fitting was performed for $n \geq 40$. Fitted trends are shown as solid lines, and the corresponding exponents x are reported in the legend as empirical complexities of the form $\Theta(n^x)$.

endpoints vary slightly, the posterior mass can be distributed among nearby crossing base pairs. Then no single crossing base pair exceeds the centroid threshold, even though the pseudoknotted motif is common in the ensemble.

RNA-shape summaries address this issue by grouping structures according to topology rather than exact base-pair positions. The following example shows a sequence, its MFE structure, its centroid structure, and the two most frequent level-5 shapes observed among sampled structures:

(1) `GGGUGGUGUGUACCACCCCUGAUGAGUCCGAAAGGACGAA` sequence,
(2) `(((((((.[[[).)))))))...]]].((((....))))...` MFE structure,
(3) `(((((((....)))))))........((((....))))...` centroid structure,
(4) `[][]()` (0.806), `[][]()()` (0.175) top two shapes.

The MFE structure contains crossing base pairs, whereas the centroid structure does not. However, the dominant RNA shape, `[][]()`, has estimated posterior frequency 0.806, showing that the crossing topology is present in the majority of sampled structures. The discrepancy occurs because the pseudoknotted base pairs are shifted across nearby positions in the sample, preventing any single crossing pair from exceeding the centroid threshold. Thus, shape-level posterior summaries can reveal pseudoknotted ensemble features that are hidden from exact base-pair-level centroid decoding.

4.5 RNA-shape distributions reveal alternative conformations in the ZMP–ZTP family

We used the ZMP–ZTP riboswitch family [31], Rfam identifier RF01750, as a test case for evaluating whether RNA-shape distributions can summarize alternative conformations across a structured RNA family. RF01750 is a particularly informative example because the annotated aptamer contains a conserved pseudoknotted fold [24, 29], whereas the full riboswitch is thought to regulate transcription through competition between the ligand-stabilized aptamer and alternative, pseudoknot-disrupting expression-platform structures [27]. Thus, this family provides a natural setting in which to ask whether shape distributions recover both the conserved pseudoknotted architecture and plausible pseudoknot-free alternatives.

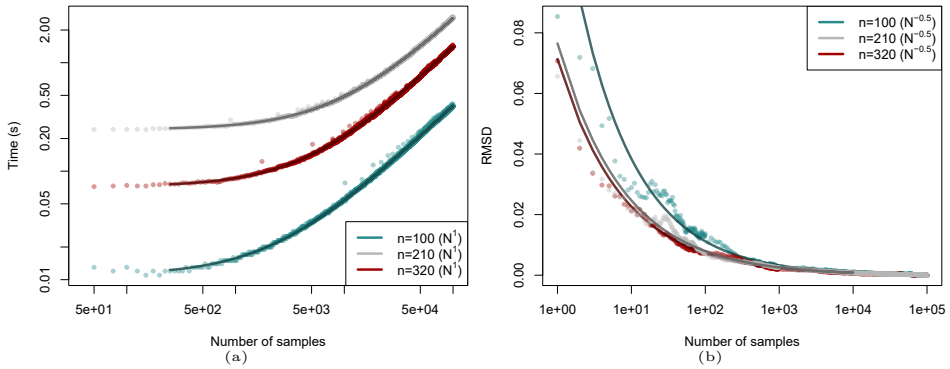

Figure 5 Effect of stochastic traceback sample count on PRISM runtime and posterior probability estimates. (a) Runtime as a function of the number of samples for three sequence lengths. Increasing the number of samples adds linear overhead after the dynamic-programming tables have been computed. (b) RMSD of estimated base-pair probabilities relative to estimates obtained from 100000 samples. The fitted exponents are close to $-1/2$, matching the standard Monte Carlo convergence rate for estimating Bernoulli probabilities.

As described in Section 3, we decomposed the Rfam consensus structure into two partial input structures. The first, G_{big}, is the maximal pseudoknot-free component obtained after removing crossing base pairs. The second, G_{small}, contains the removed crossing component. Each partial structure was fitted to each ungapped RF01750 sequence and then used as an input constraint for PRISM. For every sequence/input pair, we generated stochastic traceback samples, mapped each sampled structure to its level-5 RNA shape, and retained at most the five most frequent shapes whose estimated posterior frequency exceeded 0.1. Shapes below this threshold were omitted.

Figure 6 shows the resulting shape distributions. Each point represents one retained shape for one sequence/input pair. Circles and triangles distinguish samples generated from G_{big} and G_{small}, respectively. Point color gives the rank of the shape within that sequence/input pair, with rank 1 corresponding to the highest estimated posterior frequency. The sequences are ordered by length on the y-axis, and the horizontal reference line marks the transition near length 95. The dendrogram above the x-axis clusters the observed level-5 shapes; shapes without square brackets correspond to pseudoknot-free classes, whereas shapes containing square brackets retain a crossing, pseudoknotted component.

For sequences of length at most 95, the highest-ranked shapes under the two input constraints are concentrated around the consensus pseudoknotted shape, `([()(])`, shown in bold in Figure 6. This indicates that, for shorter RF01750 sequences, both the pseudoknot-free component and the removed crossing component provide enough information for PRISM to recover the same family-level architecture. In this length range, the two partial constraints therefore lead to compatible posterior shape distributions rather than to distinct dominant folds.

For longer sequences, the distributions become more heterogeneous and depend more strongly on the input constraint. Under G_{small}, many sequences shift toward the pseudoknotted shape `([(()()))(])`, a variant of the consensus shape with an additional nested loop, while pseudoknot-free shapes such as `(()())` remain recurrent. Under G_{big}, the sampled structures are spread across several pseudoknot-containing and pseudoknot-free classes, including shapes in which the crossing component appears in a different shape context or is absent from the retained high-probability class. Thus, the longer sequences show greater shape-level plasticity: the pseudoknotted component is not lost uniformly, but its placement and compatibility with surrounding helices vary across the family.

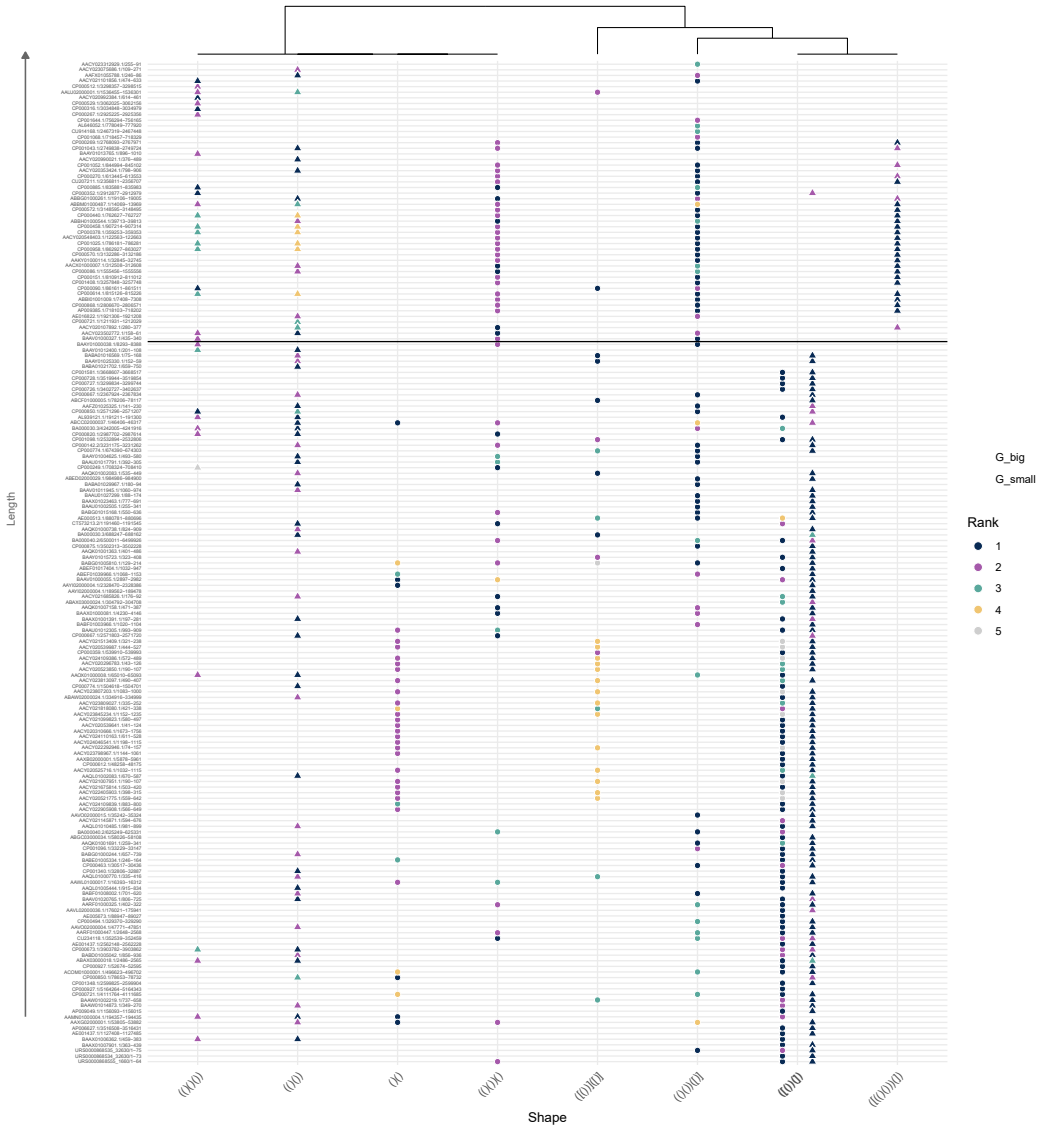

Figure 6 Distribution of level-5 RNA shapes for the ZMP–ZTP riboswitch family RF01750. For each sequence and input constraint, the five most frequent shapes are shown. The x-axis gives the RNA shape, and the y-axis lists sequences ordered by length. A horizontal line marks the boundary at sequence length 95. Point color indicates the estimated posterior frequency of the shape among sampled structures, and point shape distinguishes the two input constraints G_{big} and G_{small}. Points with $p < 0.1$ are excluded.

These results illustrate the value of RNA-shape distributions for family-level analysis. A single decoded structure would obscure the coexistence of consensus-like pseudoknotted structures, pseudoknot-free alternatives, and pseudoknotted variants with additional loops. By contrast, the distributional view reveals how these alternatives depend on both sequence length and the partial structural constraint used as input. We interpret the pseudoknot-free classes as computationally recovered alternatives that are consistent with the known regulatory competition between aptamer and expression-platform structures, not as evidence for a second ligand-stabilized aptamer fold. In the biological ZMP–ZTP riboswitch, ligand binding is expected to stabilize the pseudoknotted aptamer state; the pseudoknot-free conformations are more naturally interpreted as apo-compatible, terminator-compatible, or off-pathway alternatives.

5 Discussion and Future Directions

PRISM substantially extends the practical value of efficient hierarchical partition function computation over our previous algorithm CParty. Where CParty computes only a scalar ensemble quantity, we obtain posterior probabilities of structural features and structural classes in PRISM. Our empirical results demonstrate original ensemble-based analysis of pseudoknotted RNAs, as enabled by PRISM's novel functionality. CParty's algorithm efficiently computes $Z(S \mid G)$; as a side effect of efficient computation by dynamic programming, it additionally computes partition functions of subproblems. In turn, PRISM utilizes this information to estimate posterior masses for interpretable classes, including base-pair events, unpaired-nucleotide events, pseudoknot-conditioned subensembles, decoded structures, and RNA-shape abstractions. This class-decomposition view is the main conceptual distinction between CParty and PRISM: CParty computes the constrained partition function, whereas PRISM separates its Boltzmann mass into summaries that are directly useful for structural interpretation.

The validation experiments support the correctness of this posterior-analysis layer. When $G = \emptyset$, the conditional ensemble reduces to the standard pseudoknot-free secondary-structure ensemble, and PRISM reproduces RNAFold values for MFE energy, ensemble free energy, centroid expected distance, and MEA score under the same energy parameters. The comparison between the Boltzmann probability of the MFE structure and its observed frequency in stochastic traceback samples further supports that the traceback procedure samples from the intended conditional Boltzmann distribution.

The computational experiments show that the posterior-analysis layer remains practical over the tested sequence lengths. The full workflow, including computation of $Z(S \mid G)$, stochastic traceback, centroid decoding, and MEA decoding, remains close to RNAFold in empirical scaling and below NUPACK 3.0 over the shared benchmark range. Sampling introduces a transparent tradeoff: increasing the number of samples increases runtime linearly, while the RMSD of estimated base-pair probabilities decreases at approximately the standard Monte Carlo rate. This behavior makes the sample count an explicit accuracy parameter. In our experiments, 1000 samples gave small RMSD relative to estimates obtained from 100000 samples, suggesting that moderate sample sizes are sufficient for many posterior-decoding tasks.

The results also show that base-pair-level summaries and topology-level summaries answer different questions. Centroid decoding is appropriate when the same exact contacts recur with high probability, but it can miss a pseudoknotted motif whose helix endpoints shift across sampled structures. The pseudoknot-conditioned centroid reduces dilution from pseudoknot-

free samples, but it still operates at the level of exact base-pair identities. RNA-shape summaries provide a complementary view by grouping structures according to topology. In the example from Section 4, the centroid structure did not contain crossing base pairs, yet the dominant level-5 shape was pseudoknotted and appeared with high posterior frequency. This illustrates why posterior mass can be weak at the level of individual shifted base pairs but strong at the level of structural classes. The ZMP–ZTP riboswitch [31] analysis demonstrates how shape distributions can be used to compare conditional ensembles across related sequences and input constraints. By fitting G_{big} and G_{small} to each ungapped sequence and computing the most frequent level-5 shapes, PRISM summarizes how the posterior distribution over topologies changes with sequence length and with the chosen input scaffold. For shorter sequences, both inputs frequently recover the consensus-like shape. For longer sequences, the dominant shapes diverge: one input favors variants with additional loops, while the other shows changes in the placement of the pseudoknotted component or shifts toward pseudoknot-free conformations. These results suggest that shape-level posterior summaries can help identify alternative conformational classes within an RNA family, especially when single decoded structures are too coarse to represent ensemble variation.

Several limitations remain. First, PRISM analyzes a conditional ensemble defined by a fixed pseudoknot-free input structure G. The resulting posterior probabilities are therefore conditional on the chosen scaffold. If G is inaccurate or if the biological system occupies multiple competing scaffolds, then a single $\Omega[S, G]$ may not capture all relevant alternatives. Second, the admissible ensemble remains restricted to hierarchically generated density-2 structures. This class is broad enough to include many important pseudoknotted motifs, but it does not represent arbitrary pseudoknots. Third, base-pair probabilities and shape frequencies are estimated by stochastic traceback rather than computed exactly. The sampling experiments show predictable Monte Carlo convergence, but rare structural features may require larger sample sizes. Finally, the level-5 shape abstraction intentionally removes positional details. This is useful for detecting conserved topology, but it can merge structures that differ in functionally relevant loop lengths or exact contact positions.

These limitations suggest natural extensions. One direction is to develop adaptive sampling procedures that continue stochastic traceback until base-pair probabilities or shape frequencies reach a specified confidence threshold. Another is to compute selected class-specific partition-function masses more directly, rather than estimating all class probabilities by sampling. A third direction is to integrate external evidence, such as comparative information or probing-derived constraints, into the choice of G or into posterior reweighting of sampled structures. Finally, shape-aware decoders could return representative structures from high-probability shape classes, combining the interpretability of a concrete structure with the robustness of topology-level posterior analysis.

References

1 Tatsuya Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1-3):45–62, 2000. doi:10.1016/S0166-218X(00)00186-4.

2 Mirela Andronescu, Vera Bereg, Holger H. Hoos, and Anne Condon. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 9:340, 2008. doi:10.1186/1471-2105-9-340.

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

4 Robert P Christian and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics, 2004.

5 Jose Almeida Cruz and Eric Westhof. The dynamic landscapes of RNA architecture. *Cell*, 136(4):604–609, 2009. doi:10.1016/j.cell.2009.02.003.

6 Ye Ding, Chi Yu Chan, and Charles E. Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11(8):1157–1166, 2005. doi:10.1261/rna.2500605.

7 Ye Ding and Charles E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31:7280–7301, December 2003.

8 Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. In *Proceedings of the Tenth Annual International Conference on Research in Computational Molecular Biology*, pages 90–98, 2006. doi:10.1093/bioinformatics/btl246.

9 Mateo Gray, Luke Trinity, Ulrike Stege, Yann Ponty, Sebastian Will, and Hosna Jabbari. CParty: hierarchically constrained partition function of RNA pseudoknots. *Bioinformatics*, 41(1):btae748, 2025. doi:10.1093/bioinformatics/btae748.

10 Mateo Gray, Sebastian Will, and Hosna Jabbari. PRISM. Software, swhId: swh:1:dir:79803e70b92d8457ba08d16fe67b09382501716c (visited on 2026-08-13). URL: https://github.com/TheCOBRALab/PRISM, doi:10.4230/artifacts.27651.

11 Mateo Gray, Sebastian Will, and Hosna Jabbari. PRISM-RawData. Dataset, swhId: swh:1:dir:4646c334d25f499a6485f62fe913bb90e2cdd115 (visited on 2026-08-13). URL: https://github.com/mateog4712/PRISM-RawData, doi:10.4230/artifacts.27652.

12 Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R Eddy, and Alex Bateman. Rfam: annotating non-coding rnas in complete genomes. *Nucleic Acids Research*, 33(suppl_1):D121–D124, 2005.

13 Hosna Jabbari, Anne Condon, Ana Pop, Cristina Pop, and Yinglei Zhao. Hfold: Rna pseudoknotted secondary structure prediction using hierarchical folding. In Raffaele Giancarlo and Sridhar Hannenhalli, editors, *Algorithms in Bioinformatics*, pages 323–334, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. doi:10.1007/978-3-540-74126-8_30.

14 Hosna Jabbari, Anne Condon, and Shelly Zhao. Novel and Efficient RNA Secondary Structure Prediction Using Hierarchical Folding. *Journal of Computational Biology*, 15(2):139–163, March 2008. doi:10.1089/cmb.2007.0198.

15 Hosna Jabbari, Ian Wark, Carlo Montemagno, and Sebastian Will. Sparsification Enables Predicting Kissing Hairpin Pseudoknot Structures of Long RNAs in Practice. In *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*, volume 88 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 12:1–12:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPIcs.WABI.2017.12.

16 Hosna Jabbari, Ian Wark, Carlo Montemagno, and Sebastian Will. Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics*, 34:3849–3856, November 2018. doi:10.1093/bioinformatics/bty420.

17 Marilyn Kozak. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361:13–37, 2005. doi:10.1016/j.gene.2005.06.037.

18 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*, 6:1–14, 2011.

19 Rune B Lyngsø and Christian NS Pedersen. Pseudoknots in RNA secondary structures. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 201–209, New York, NY, United States, 2000. Association for Computing Machinery.

20 John S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

21 Stephanie A. Mortimer, Mary Anne Kidwell, and Jennifer A. Doudna. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, 15:469–479, 2014. doi:10.1038/nrg3681.

22 Yann Ponty. Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy. *Journal of Mathematical Biology*, 56:107–127, 2008. doi:10.1007/s00285-007-0137-z.

23 Christian M. Reidys, Fenix W. D. Huang, Jørgen E. Andersen, Robert C. Penner, Peter F. Stadler, and Markus E. Nebel. Topology and prediction of RNA pseudoknots. *Bioinformatics*, 27(8):1076–1085, 2011. doi:10.1093/bioinformatics/btr090.

24 Aiming Ren, Kanagalaghatta R Rajashankar, and Dinshaw J Patel. Global rna fold and molecular recognition for a pfl riboswitch bound to zmp, a master regulator of one-carbon metabolism. *Structure*, 23:1375–1381, 2015. doi:10.1016/j.str.2015.05.016.

25 Saad Sheikh, Rolf Backofen, and Yann Ponty. Impact of the energy model on the complexity of rna folding with pseudoknots. In *Annual Symposium on Combinatorial Pattern Matching*, pages 321–333. Springer, 2012. doi:10.1007/978-3-642-31265-6_26.

26 Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006. doi:10.1093/bioinformatics/btk010.

27 Eric J Strobel, Luyi Cheng, Katherine E Berman, Paul D Carlson, and Julius B Lucks. A ligand-gated strand displacement mechanism for ztp riboswitch transcription control. *Nature Chemical Biology*, 15:1067–1076, 2019. doi:10.1038/s41589-019-0382-7.

28 Ignacio Tinoco Jr and Carlos Bustamante. How RNA folds. *Journal of Molecular Biology*, 293(2):271–281, 1999.

29 Jeremiah J Trausch, Joan G Marcano-Velázquez, Michal M Matyjasik, and Robert T Batey. Metal ion-mediated nucleobase recognition by the ztp riboswitch. *Chemistry and Biology*, 22:829–837, 2015. doi:10.1016/j.chembiol.2015.06.007.

30 M. Bryan Warf and J. Andrew Berglund. Role of RNA structure in regulating pre-mRNA splicing. *Trends in Biochemical Sciences*, 35(3):169–178, 2010. doi:10.1016/j.tibs.2009.10.004.

31 Zasha Weinberg, Joy X Wang, Jarrod Bogue, Jingying Yang, Keith Corbino, Ryan H Moy, and Ronald R Breaker. Comparative genomics reveals 104 candidate structured rnas from bacteria, archaea, and their metagenomes. *Genome Biology*, 11, 2010. doi:10.1186/gb-2010-11-3-r31.

32 Timothy J. Wilson and David M. J. Lilley. RNA catalysis—is that it? *RNA*, 21(4):534–537, 2015. doi:10.1261/rna.049874.115.

33 Christina Witwer, Ivo L Hofacker, and Peter F Stadler. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 1(2):66–77, 2004. doi:10.1109/TCBB.2004.22.

34 Joseph N Zadeh, Conrad D Steenberg, Justin S Bois, Brian R Wolfe, Marshall B Pierce, Asif R Khan, Robert M Dirks, and Niles A Pierce. NUPACK: analysis and design of nucleic acid systems. *Journal of Computational Chemistry*, 32(1):170–173, 2011. doi:10.1002/JCC.21596.