PKProbDesign:
RNA inverse folding including pseudoknots by optimizing thermodynamic folding probability

Takumi Otagaki, Junichi Iwakiri, Goro Terai, Kiyoshi Asai, Kengo Sato

Sato lab, Institute of Science Tokyo

PhD student

New Results

Follow this preprint

PKProbDesign: RNA inverse folding including pseudoknots by optimizing thermodynamic folding probability

Takumi Otagaki, Junichi Iwakiri, goro terai, Kiyoshi Asai, Kengo Sato

doi: https://doi.org/10.64898/2026.07.09.736945

Preprint July 10th.

Institute of SCIENCE TOKYO

This is my second time Benasque!

I really enjoyed my previous Benasque.








Professor Asai cooked lunch for us every day!
I was spoiled - just like last time in Benasque.

RNA inverse folding

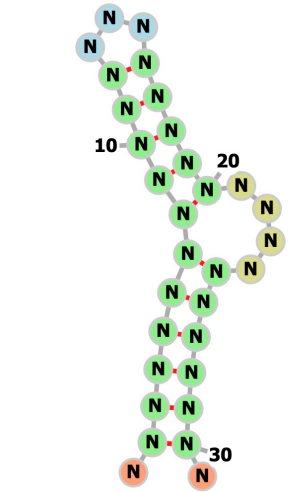
- Inverse Folding problem is
  to find the sequence q^* which folds the target structure σ^t.

Sequence q^*

??????????????

Folds into

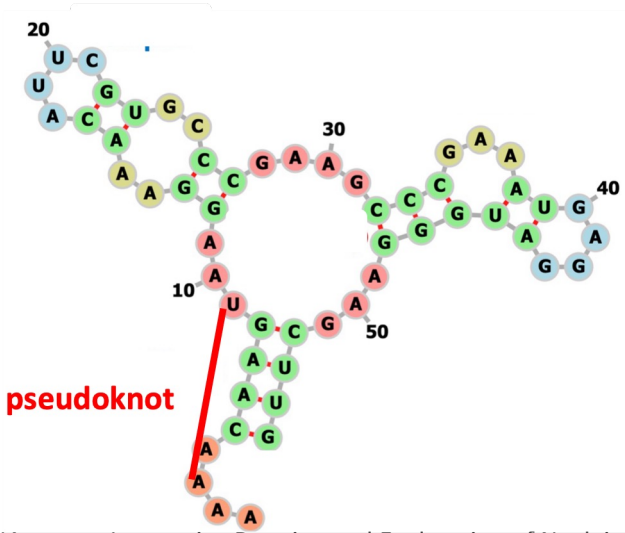
Target Structure σ^t

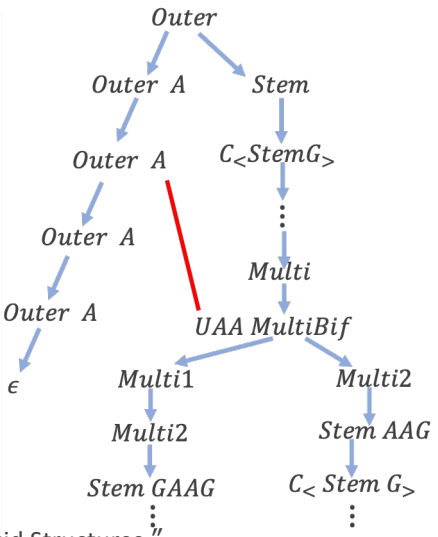


.((((((((((...))))))....))))))

RNA inverse folding with pseudoknots

- **Pseudoknots are important motifs** in functional and viral RNAs and often contribute to tertiary folding.
- They **violate the nested structure** assumed by standard secondary-structure models, making them **computationally challenging**.





Pseudoknots break the tree.

Johnson, Philip Z., and Anne E. Simon. 2023. "RNAcanvas: Interactive Drawing and Exploration of Nucleic Acid Structures." *Nucleic Acids Research* 51 (W1): W501–8.

Minimum Free Energy (MFE) based inverse folding

Find **the sequence s whose MFE structure is σ^t**.

$$s^* \in \{s \mid \text{MFE\_structure}(s) = \sigma^t\}$$

$$\text{MFE\_stucrure(s)} = \text{argmin}_\sigma \, \text{Energy}(\sigma, s)$$

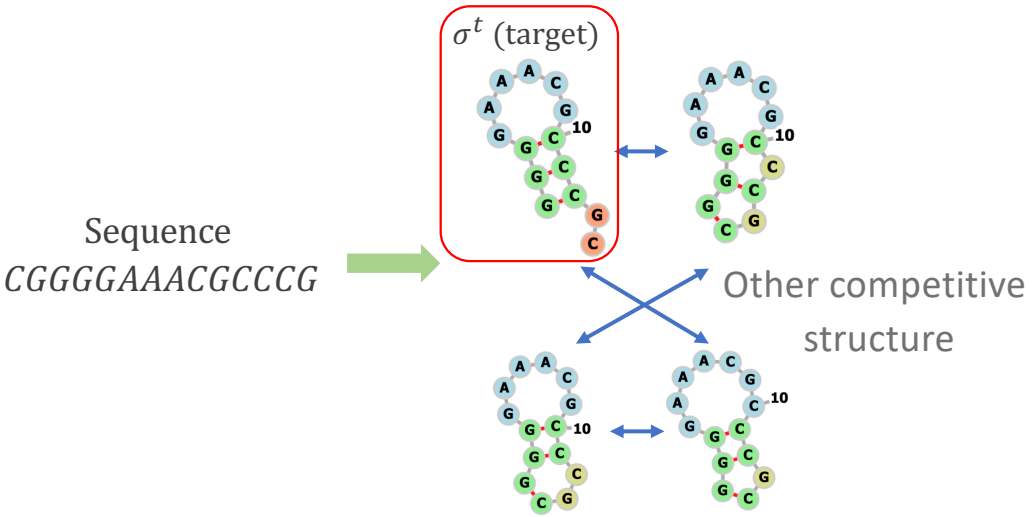
Why MFE agreement is not enough

RNA folds into an ensemble of competing structures,

not just a single MFE structure.

How strongly is the target supported in the ensemble?

→ **Folding Probability: $p(\sigma^t \mid s)$**

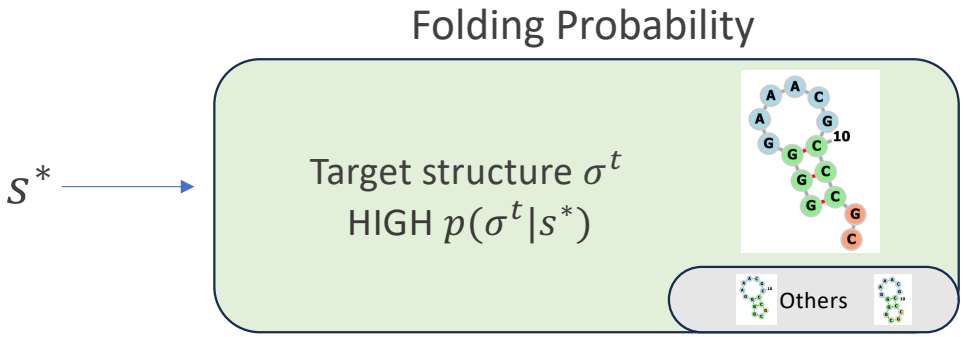


Probability-based inverse folding

- Find the **sequence that maximizes the target folding probability.**

$$s^* = \text{argmax}_s \, p(\sigma^t \mid s)$$

- The objective naturally extends to multiple target structures.
  - e.g., maximize $p(\sigma_1|s) + p(\sigma_2|s)$



The gap: probability-based design for pseudoknots

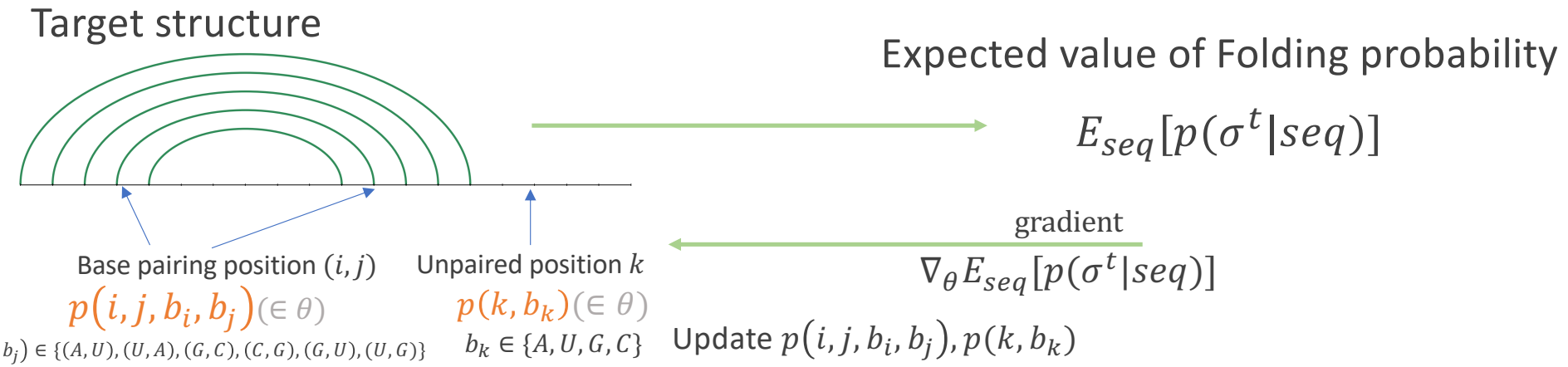
- Probability-based design requires a folding-probability evaluator.
- For pseudoknotted structures, integrating such an evaluator into a practical inverse-folding framework remained challenging.

→We propose PKProbDesign;

A pseudoknot-aware probability-based framework for RNA inverse folding.

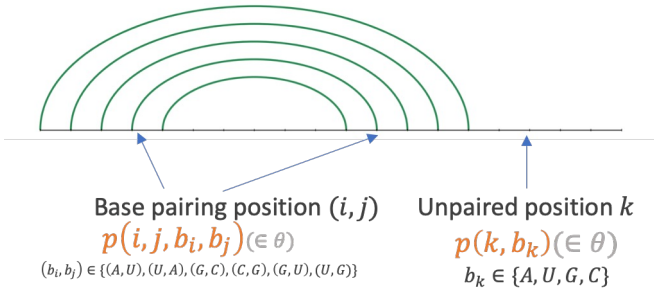
Previous work
SamplingDesign: probability-based inverse folding

- Sear 21the sequence which maximizes **the target folding probability** based on the gradient descent.
  - **Folding probability of the Pseudoknot-free structure is possible!**
- The gradient of folding probability is approximated by Monte Carlo Sampling.



Tang, Wei Yu, Ning Dai, Tianshuo Zhou, David H. Mathews, and Liang Huang. 2026. "SamplingDesign: RNA Design via Continuous Optimization with Coupled Variables and Monte-Carlo Sampling." *Nature Communications* 17 (1): 2950.

SamplingDesign formulation



$$J(\theta) = E_{S \sim p(\cdot|\,\theta)}\,[f(S)]$$

The gradient $\nabla_\theta J(\theta)$is; $\mathbb{E}_{S\sim p_\theta}\left[f(S)\nabla_\theta \log p_\theta(S)\right].$

We can approximate the gradient by sampling,

$$\nabla_\theta J(\theta) \approx \frac{1}{K}\sum_{k=1}^{K} f(S_k)\nabla_\theta \log p_\theta(S_k).$$

PKProbDesign replaces the standard pseudoknot-free evaluator with a pseudoknot-aware thermodynamic folding probability.

$f(S)$ can be black box score for sequence S.

Only sequence scores are required.

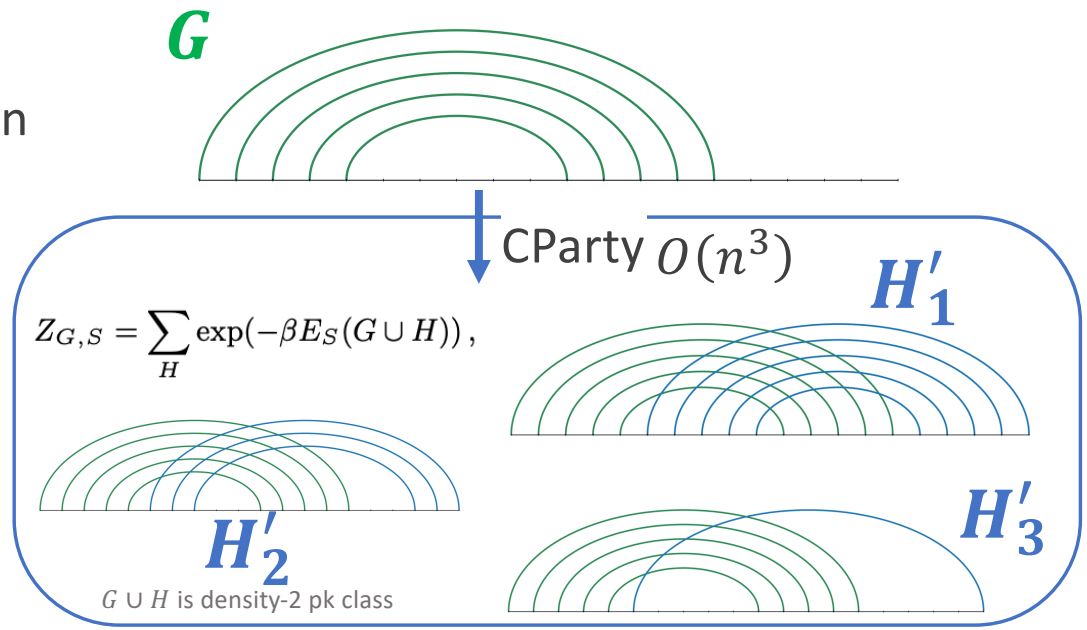
The evaluator f does not need to be differentiable.

CParty enables conditional probabilities for density-2 pseudoknots

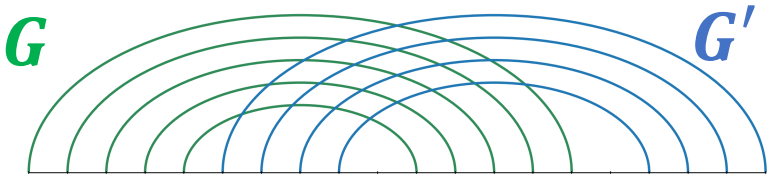
Given a pseudoknot-free scaffold G,
CParty computes a conditional partition function over extensions H s.t. $G \cup H \in$ density2 class.

It provides the conditional thermodynamic evaluator required by PKProbDesign.

Pseudoknot energy model (DP09)



$$P(G' \mid G, S) = \frac{\exp(-\beta E_S(G \cup G'))}{Z_{G,S}}.$$



Gray, Mateo, Luke Trinity, Ulrike Stege, Yann Ponty, Sebastian Will, and Hosna Jabbari. 2024.
"CParty: Hierarchically Constrained Partition Function of RNA Pseudoknots." *Bioinformatics (Oxford, England)* 41 (1): btae748.

Pseudo-joint folding probability for density-2 pseudoknotted structure

A density-2 pseudoknot struct R can be decomposed as.

$\rightarrow R = G \cup G', G \cap G' = \emptyset$, G, G' are pseudoknot-free.

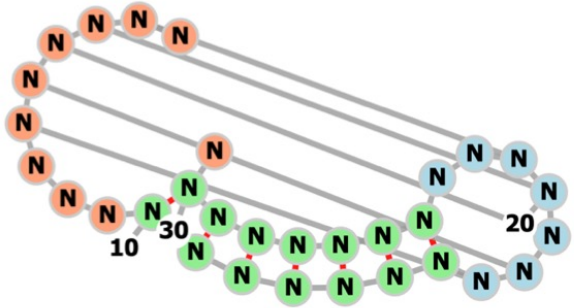
Under the hierarchical folding assumption,
we define the pseudo-joint folding probability as:

$$\tilde{p}(R \mid S) = p(G \mid \mathbf{S})\, \mathbf{p}(\mathbf{G}' \mid \mathbf{G}, \mathbf{S})$$

$p(G|S) \leftarrow$ LinearPartition$O(nb^2)$ (b: beam size)
$p(G'|G,S) \leftarrow$ CParty $O(n^3)$.

Exact $p(R|S)$ is $\sum_{(G,G') \in \mathcal{D}(R)} P(G \mid S)\, P(G' \mid G, S)$.

Target Structure R (density2)



R = .(((((...[[[[[[]))))).]]]]]].

Decomposition

G = .(((((.....))))).....
G' = ...[[[[[[[.....]]]]]]].

Greedy scaffold decomposition and its limitation

We decompose R in a greedy way.

Algorithm 1 Greedy scaffold-extension decomposition

Input: target base-pair set R

Output: scaffold G, extension component G'

1. **Sort** the pairs in R by increasing span $|j - i|$.
2. **Initialize** $G \leftarrow \emptyset$.
3. **For each** pair p in sorted order, add p to G if p does not cross any pair already in G.
4. **Set** $G' \leftarrow R \setminus G$.

Limitation

- Scaffold decomposition is heuristic (greedy).
- No guarantee that G' is pseudoknot-free.
- This issue did not occur in our benchmark dataset.

Dataset

Dataset: Pseudobase++ [Taufer et al., 2009]

- 357 pseudoknotted structures (≤100 nt)
- Removed 55 non-density-2 structures
- Removed 2 structures incompatible with the energy model.
- **300 benchmark targets**

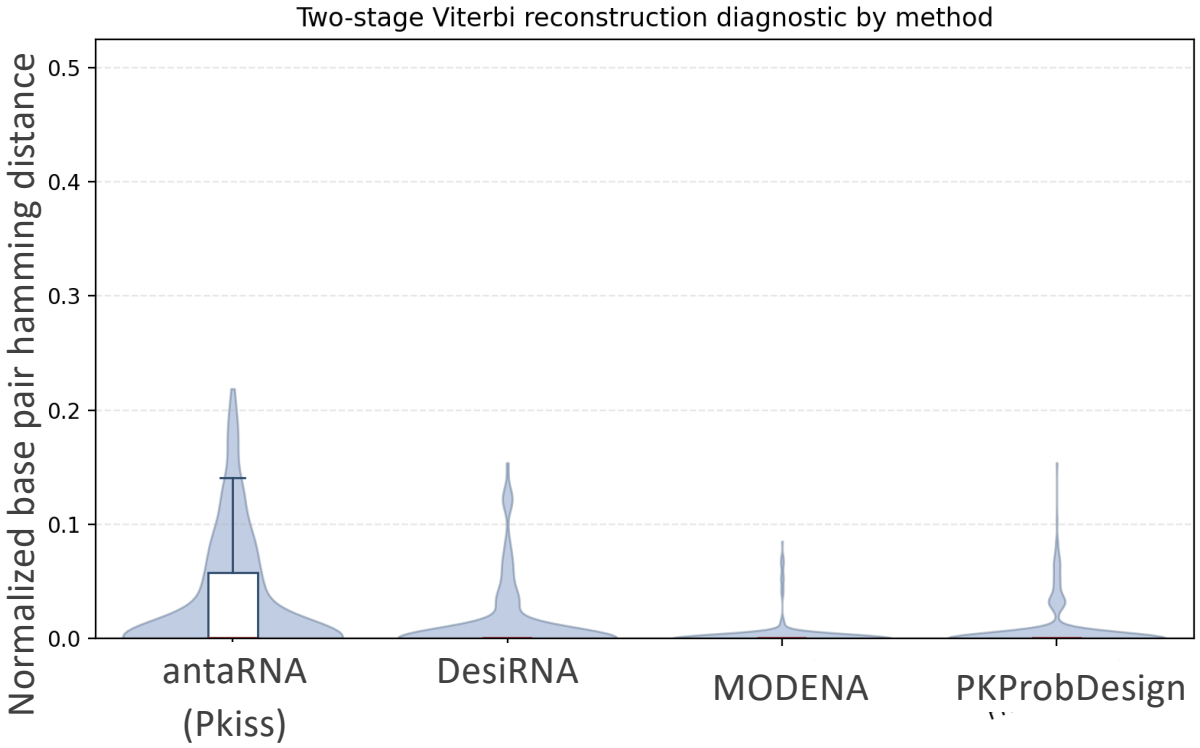
Compared methods:

- DesiRNA [Wirecki et al., 2025]
- MODENA [Taneda, 2011]
- antaRNA [Kleinkauf et al., 2015]

Structure-prediction based evaluation

1. 20 designed seqs for each methods and targets.
2. For each designed sequences;
   2.1. predict MFE G_1 excluding pk.
   2.2. predict MFE $G_2 \cup G_1$ on G_1.
3. Plot the min base_pair_distance$(G_1 \cup G_2, \sigma_t)$
   $\mathrm{Min}_{i=1\ldots20}\ d_{BP}^{norm}(G_1^i \cup G_2^i, \sigma^t)$

$$d_{\mathrm{BP}}^{\mathrm{norm}}(P_{\mathrm{pred}}, P_{\mathrm{tgt}}) = \frac{1}{|S|} \sum_{1 \le i < j \le |S|} \mathbf{1}[\, P_{\mathrm{pred}}[i,j] \neq P_{\mathrm{tgt}}[i,j]]$$

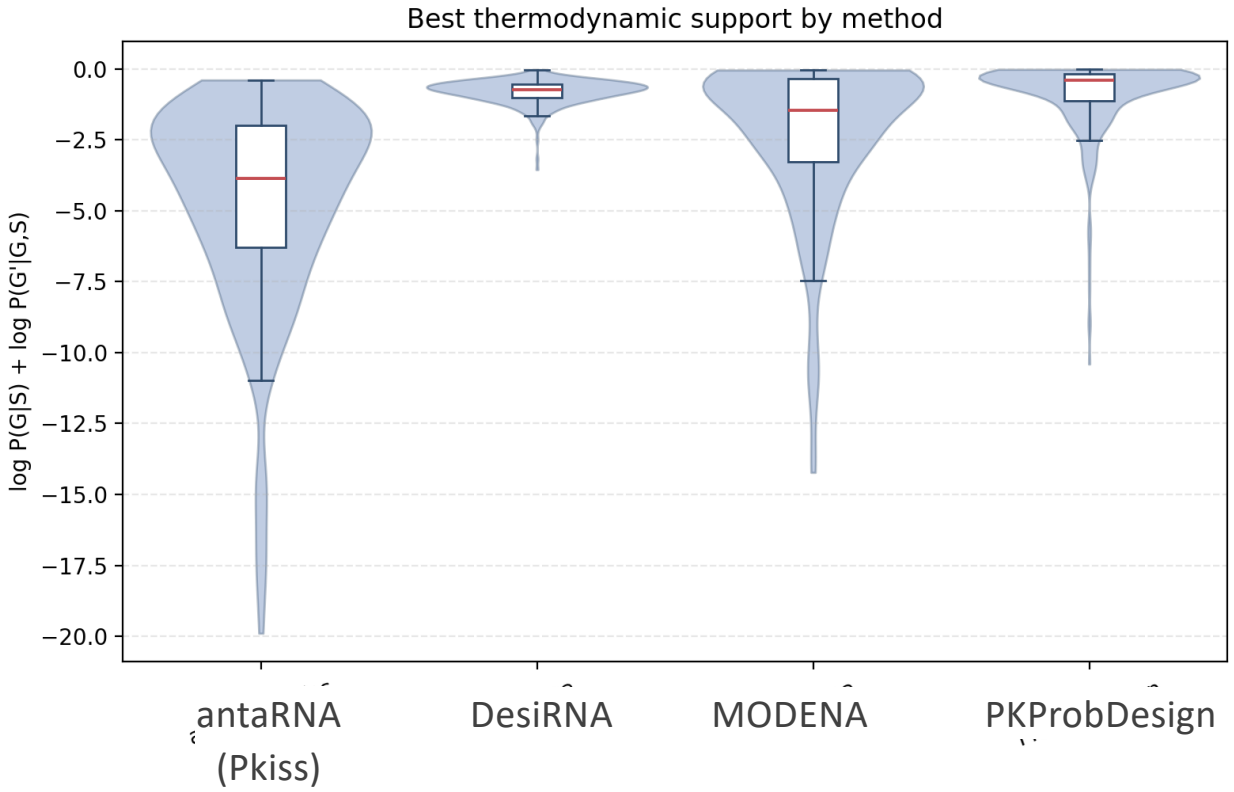


Pseudo-joint folding probability comparison

Higher better

1. 20 designed seqs for each methods, targets.
2. Plot the maximum $\log p(G|S)p(G'|G,S)$

PKProbDesign achieves high pseudo-joint folding probability comparable to DesiRNA



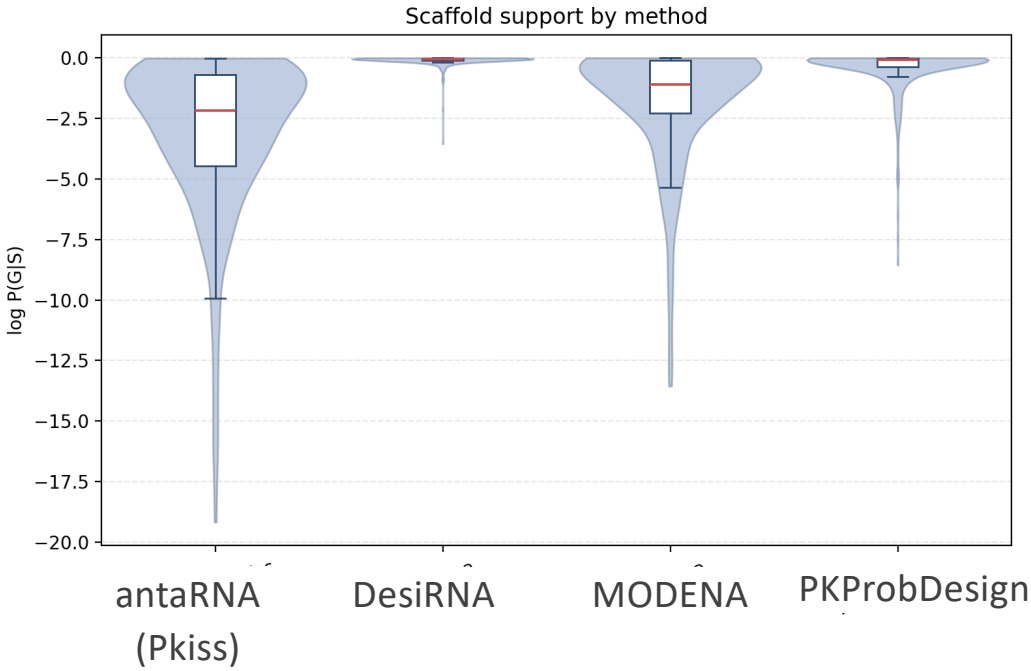
$p_{value} = 0.01$
(pkprobdesign v.s. desirna)

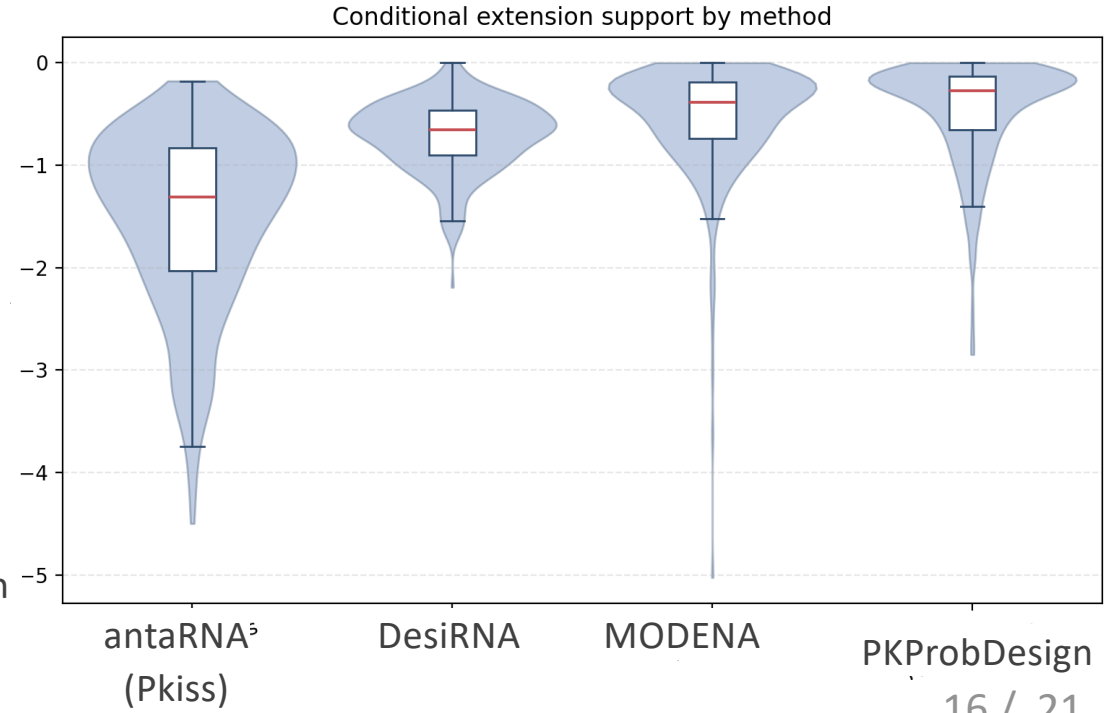
The same target decomposition was used for all methods.

Results

DesiRNA favors scaffold support; PKProbDesign favors conditional extension support

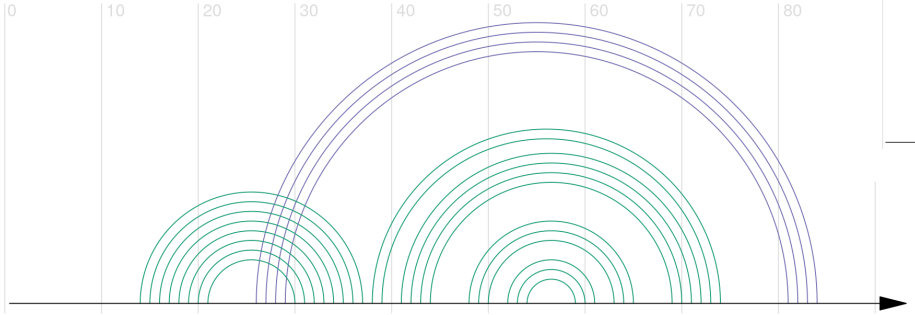
1. 20 designed seqs for each methods, targets.
2. For each target and method, select the sequence with the highest pseudo-joint probability.
3. Decompose its score into scaffold prob $\log p(G|S)$ (left) and extension prob $\log p(G'|G,S)$ (right).



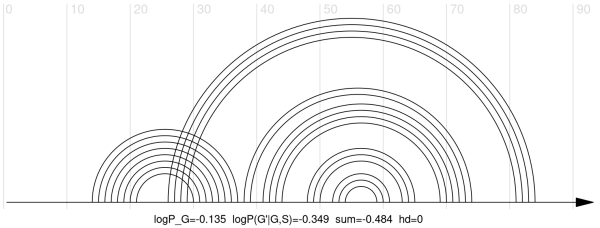


Case Study: PKB00362, viral 5'UTR
A target showing clear differences among methods.

Target σ^t



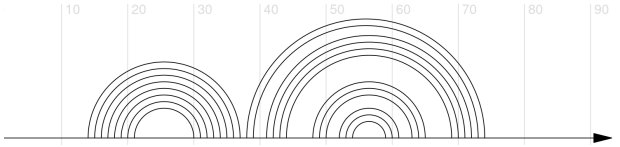
PKProbDesign



$\log p(G|S)p(G'|G,S) = -0.48$

HD=0

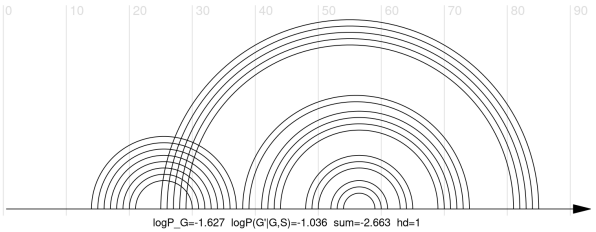
DesiRNA



$\log p(G|S)p(G'|G,S) = -0.82$

HD=4

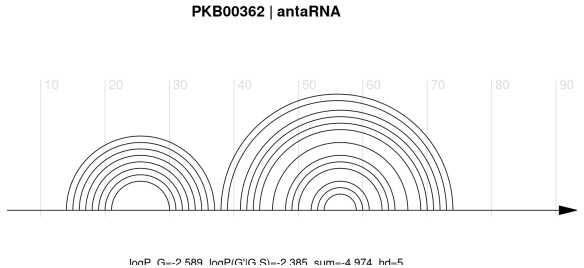
MODENA



$\log p(G|S)p(G'|G,S) = -2.66$

HD=1

antaRNA



$\log p(G|S)p(G'|G,S) = -5.00$

HD=5

1. 20 designed seqs for each methods, targets.
2. For each designed sequence;
   2.1. compute pkfree mfe G.
   2.2. compute mfe $G_2 \cup G_1$ on G_1

Interpretation of method-specific performance

- PKProbDesign showed strong conditional-extensionprobability, consistent with directly optimizing a pseudoknot-aware thermodynamic objective.
- MODENA performed well in MFE-based evaluation, consistent with its predictor-based optimization objective.
- DesiRNA excelled in scaffold probability, reflecting its design strategy (maximizing folding prob)
- antaRNA may have been affected by the fixed GC-content soft-constraint:GC ratio fixed at 0.5, which we could not disable.

Current limitations

- The objective is a decomposition-dependent pseudo-joint probability, not the exact probability of σ.
- The greedy decomposition is heuristic and may fail to produce a pseudoknot-free G'.
  - Future work:
    Use graph-coloring-based decomposition and maximizing the num of bps in G
- The current implementation is restricted to density-2 pseudoknots.
- The model relies on the hierarchical folding assumption.

Toward longer and more diverse pseudoknot-aware design

Algorithmic extensions

- Beam-pruned CParty for longer RNA targets
- Marginalization over, or optimization of, multiple decompositions to obtain the exact joint probability

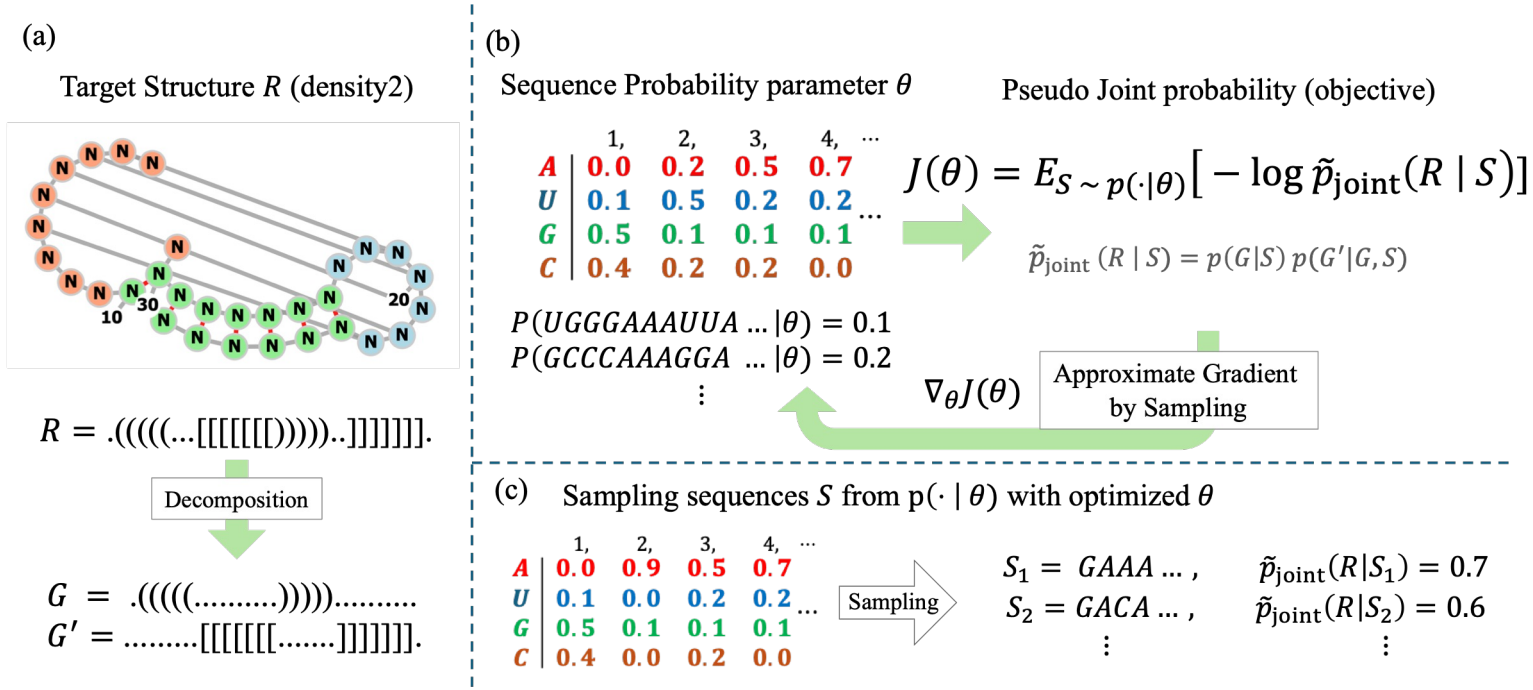
Biological applications

- IRES design
- Riboswitch design

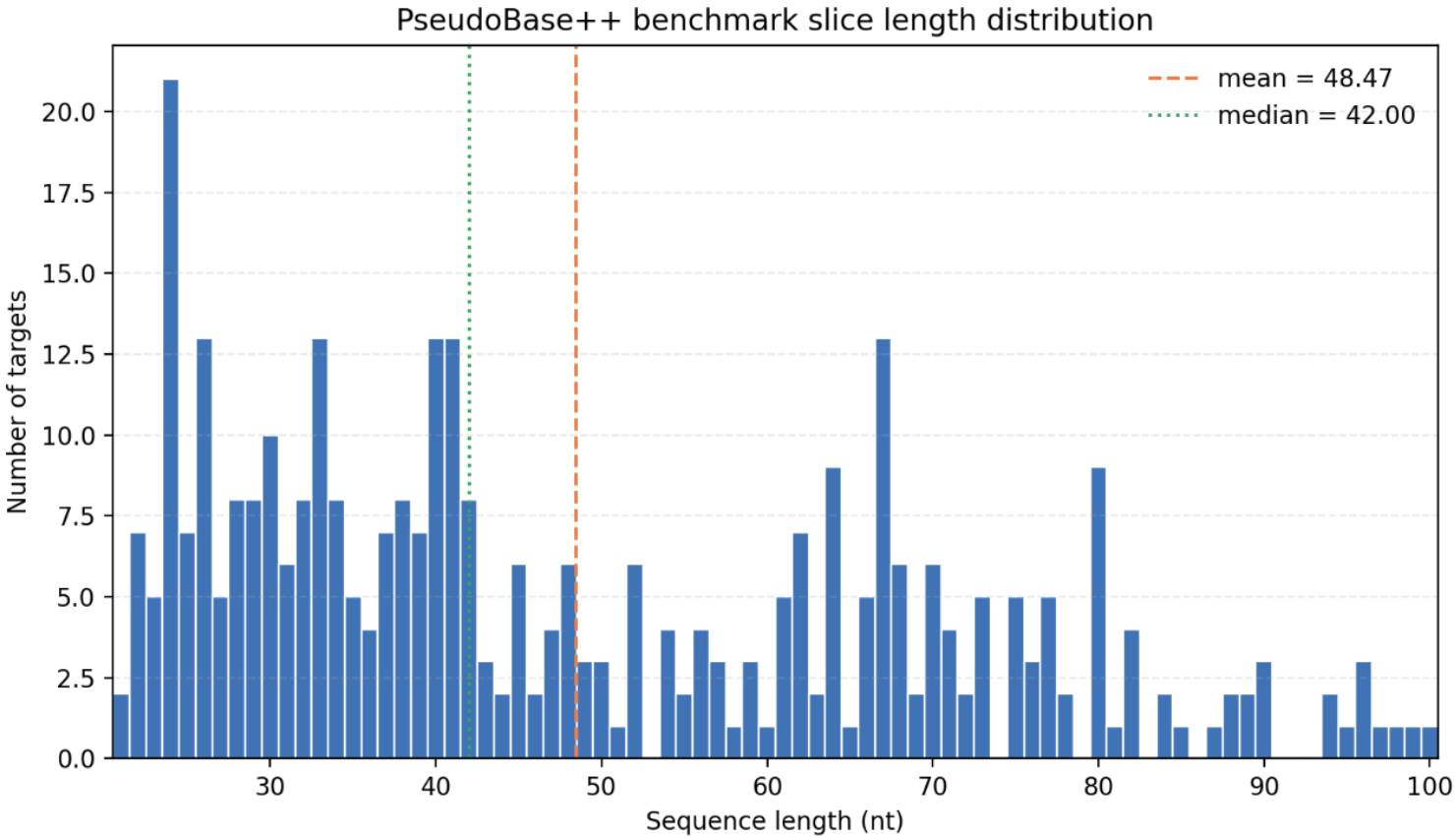
Summary

Thank you!

- Probability-based inverse folding for density-2 pseudoknots
- SamplingDesign optimization with LinearPartition and CParty evaluators
- Improved thermodynamic folding support on 300 PseudoBase++ targets

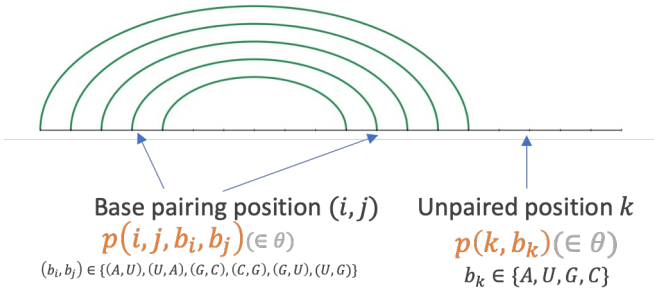


The sequence length distribution



N=300

SamplingDesign formulation



$$J(\theta) = E_{S \sim p(\cdot|\,\theta)}\left[f(S)\right]$$

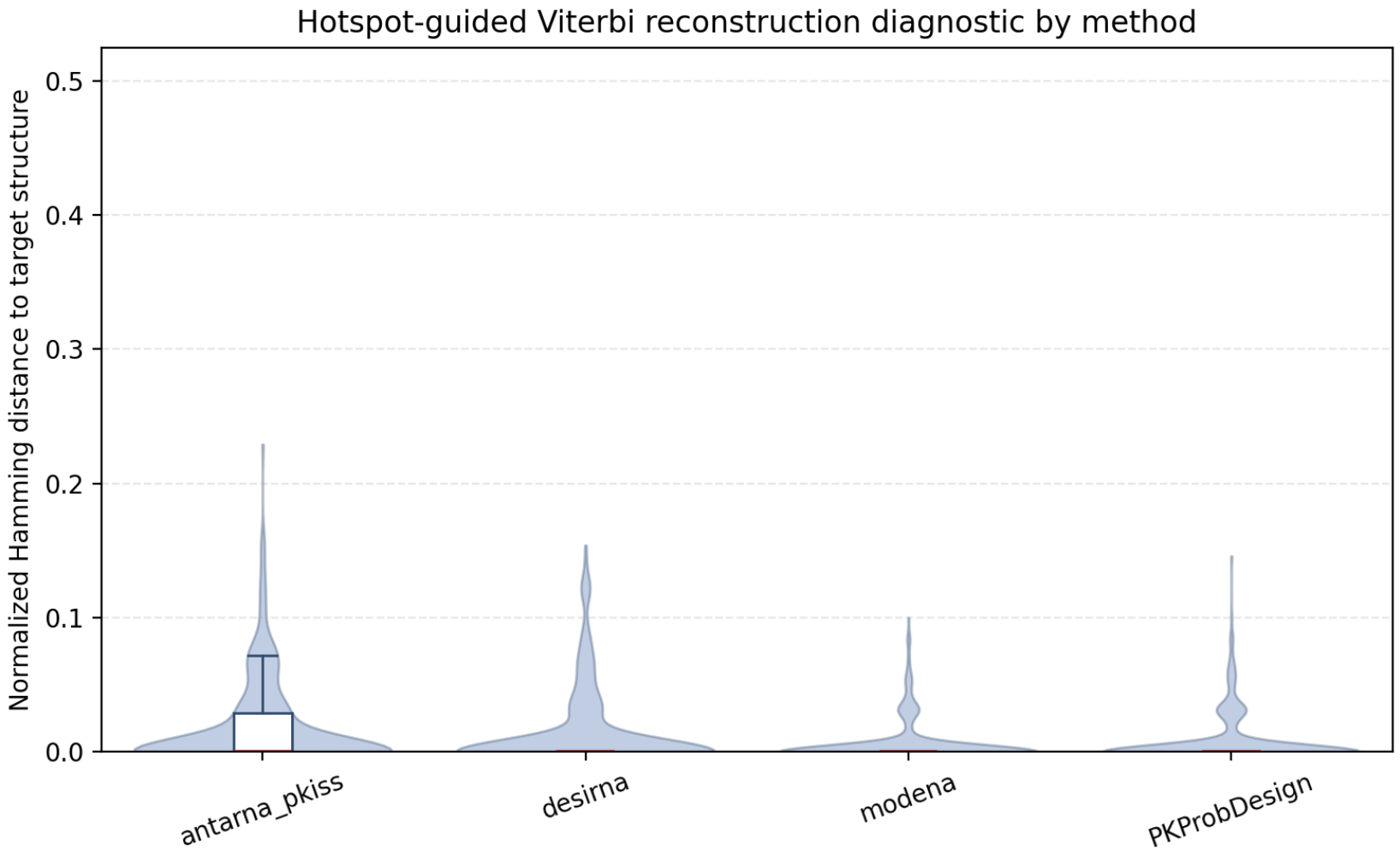
The gradient $\nabla_\theta J(\theta)$is $\nabla_\theta J(\theta) = \nabla_\theta \sum_S p_\theta(S) f(S)$

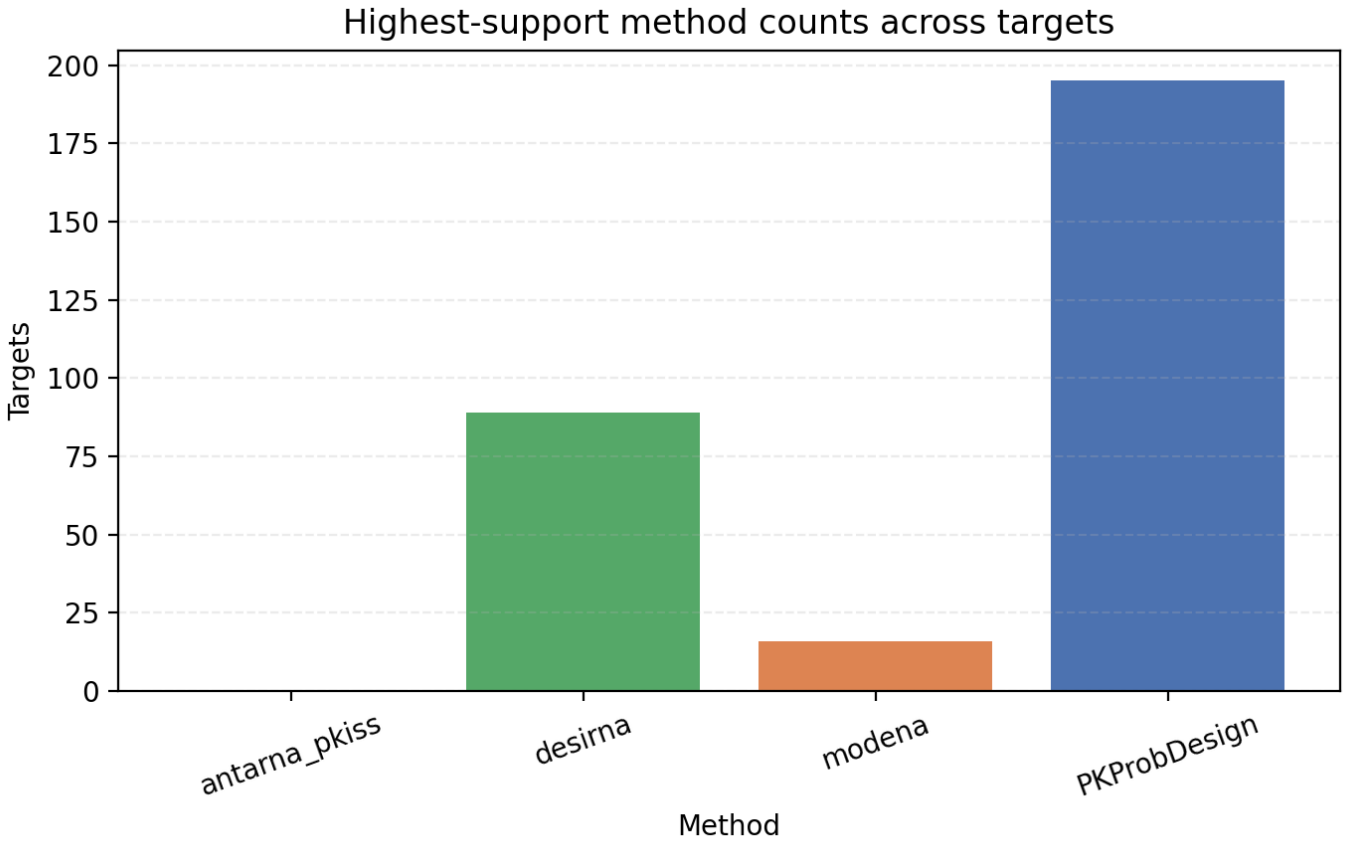
$$= \sum_S f(S)\, p_\theta(S)\, \nabla_\theta \log p_\theta(S)$$

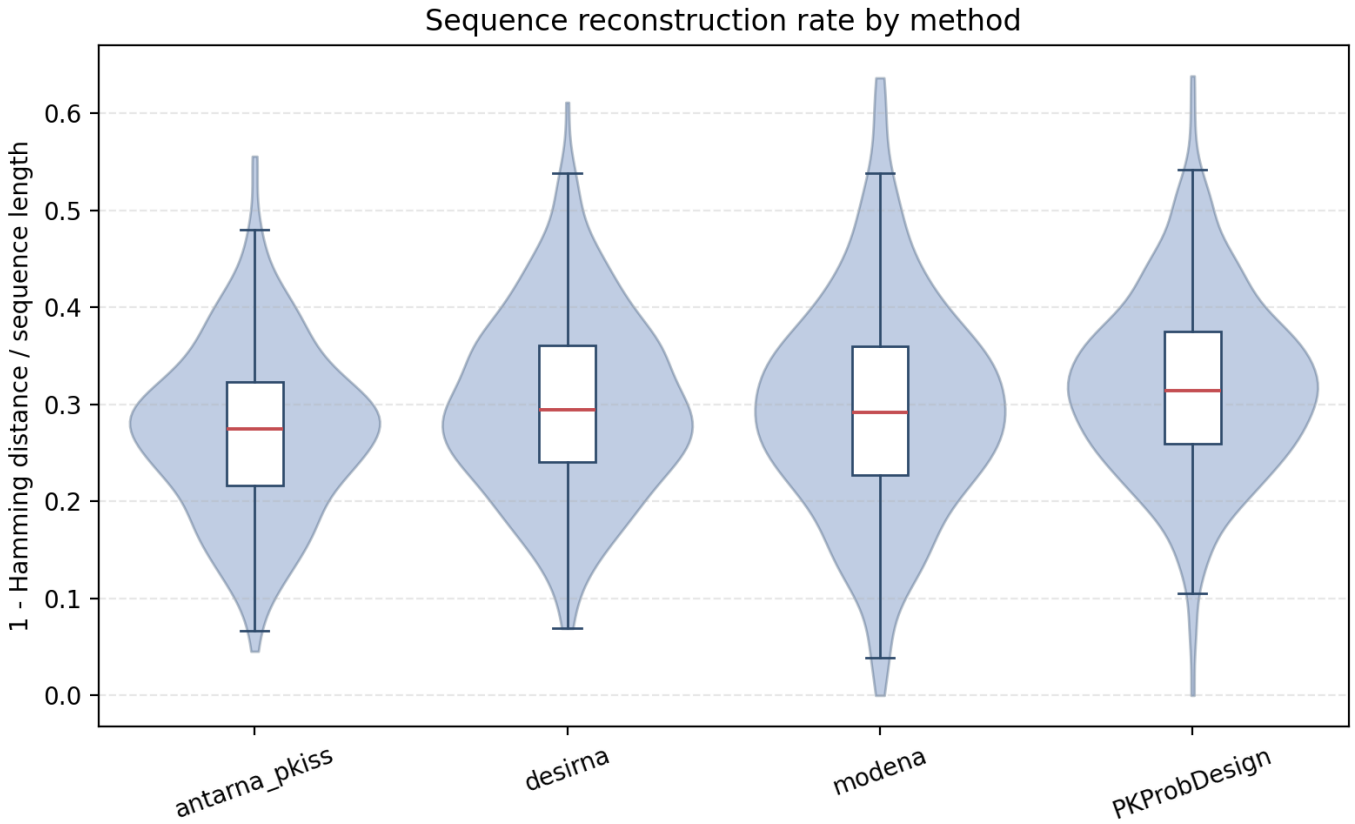
$$= \mathbb{E}_{S \sim p_\theta}\left[f(S) \nabla_\theta \log p_\theta(S)\right].$$

We can approximate the gradient by sampling,

$$\nabla_\theta J(\theta) \approx \frac{1}{K} \sum_{k=1}^{K} f(S_k) \nabla_\theta \log p_\theta(S_k).$$


Hotspot-guided Viterbi reconstruction diagnostic by method
Normalized Hamming distance to target structure
0.0
0.1
0.2
0.3
0.4
0.5
antarna_pkiss
desirna
modena
PKProbDesign


Highest-support method counts across targets
Targets
200
175
150
125
100
75
50
25
0
antarna_pkiss
desirna
modena
PKProbDesign
Method


Sequence reconstruction rate by method
1 - Hamming distance / sequence length
0.0
0.1
0.2
0.3
0.4
0.5
0.6
antarna_pkiss
desirna
modena
PKProbDesign