CParty: Conditional partition function for density-2 RNA pseudoknots

Luke Trinity
Department of Computer Science, University of Victoria, Canada

Sebastian Will
CNRS/LIX (UMR 7161), Institut Polytechnique de Paris, France

Yann Ponty
CNRS/LIX (UMR 7161), Institut Polytechnique de Paris, France

Ulrike Stege
Department of Computer Science, University of Victoria, Canada

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

Abstract

RNA molecules fold into biologically important functional structures. Efficient dynamic programming RNA (secondary) structure prediction algorithms restrict the the search space to evade NP-hardness of general pseudoknot prediction. While such prediction algorithms can be extended to provide a stochastic view on RNA ensembles, they are either limited to pseudoknot-free structures or extremely complex. To overcome this dilemma, we follow the hierarchical folding hypothesis, i.e. the bio-physically well-motivated assumption that non-crossing structures fold relatively fast prior to the formation of pseudoknot interactions. Thus, we efficiently compute the conditional partition function (CPF) given a non-crossing structure G for a subset of pseudoknotted structures i.e. density-2 structures $G \cup G'$ for non-crossing disjoint G'. Notably, this enables sampling from the hierarchical distribution $P(G'|G)$. As our main contribution, we devise the algorithm *CParty*, which transfers the dynamic programming scheme of HFold (which minimizes free energy of pseudoknots in a realistic model) to a partition function variant by for the first time *de-ambiguating* its decomposition of density-2 structures. Compared to the only other available pseudoknot partition function algorithm, which covers simple pseudoknots, our method covers a much larger structure class; at the same time, it is significantly more efficient—reducing the time as well as the space complexity by a quadratic factor. Summarizing, we provide a highly efficient, cubic time, algorithm for the stochastic analysis of pseudoknotted RNAs, which enables novel applications. For example, we discuss how the CPF for a pseudoknotted therapeutic target in SARS-CoV-2 provides insights into RNA structure formation kinetic paths.

2012 ACM Subject Classification Applied computing → Computational biology; Theory of computation → Dynamic programming

Keywords and phrases RNA structure prediction, free energy, partition function, pseudoknots

Digital Object Identifier 10.4230/LIPIcs.WABI.2023.23

Funding *Hosna Jabbari*: [NSERC DG, Microsoft AI for Health]

1 Introduction

RNA molecules play a vital role in cellular processes; many possess functional structures. Due to the costly and time-consuming nature of experimental methods to detect RNA structure, computational tools to predict the formation of RNA structure conserve valuable resources.

[1] Corresponding author

© Luke Trinity, Sebastian Will, Yann Ponty, Ulrike Stege and Hosna Jabbari;
licensed under Creative Commons License CC-BY 4.0
42nd Conference on Very Important Topics (CVIT 2016).
Editors: John Q. Open and Joan R. Access; Article No. 23; pp. 23:1–23:21
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Our focus is on the prediction of RNA structure via a 2D model, known as secondary structure, whose accurate prediction can also inform state-of-the-art 3D prediction tasks.

RNA secondary structure prediction involves finding a set of base pairs for a given input RNA sequence, such that each base appears in at most one pair [12]. When an upstream RNA segment loops back and forms base pairs with a downstream segment, disrupting their linear alignment, we refer to these as crossing base pairs. Two widely used thermodynamic approaches for predicting non-crossing RNA secondary structure are the ones by Zuker [31] and McCaskill [15]. The Zuker algorithm uses dynamic programming (DP) to recursively compute the free energy of all possible RNA substructures and find the most energetically stable, minimum free energy (MFE) structure [31]. To further improve their prediction accuracy, MFE-based prediction methods can be extended to integrate experimental data (e.g. SHAPE profiles) as soft-constraints [9].

Adopting a stochastic view of RNA, the McCaskill algorithm computes the *partition function* that describes the secondary structure of an arbitrary length RNA. In order to efficiently calculate the partition function, the algorithm must carefully consider the equilibrium ensemble of structures: all possible base pairs and associated energetic contributions from each type of *loop* (i.e., hairpin, internal/bulge, stacked pair, and multiloop). The partition function is useful for predicting the relative probability of any structure as well as base pair probabilities [15, 14]. The likelihood of any particular RNA structure occurring can be determined based on the energy of the structure itself relative to the total energy in the system.

The hierarchical folding hypothesis postulates that RNA first folds into a simple non-crossing structure, then additional bases pair to lower the MFE of the structure [23]. Hierarchical folding can also be thought of as a special kind of kinetics, where the initial non-crossing base pairs form more quickly, followed by crossing or non-canonical base pairs. Assuming fast convergence to the equilibrium over non-crossing structures, and an overly large energy barrier, the probability to observe a given crossing structure can be expressed as the product of the classic Boltzmann probability $P(G)$ of the non-crossing subset G, and the conditional probability $P(G'|G)$. If a Boltzmann-distributed non-crossing structure is irrevocable, then the distribution to consider is not the usual Boltzmann distribution:

$$\mathbb{P}(G, G') = \mathbb{P}(G \cup G') = \frac{e^{-\beta(E_{2D}(G)+E_{PK}(G'))}}{\sum_{H \cup H'} e^{-\beta(E_{2D}(H)+E_{PK}(H'))}}$$

but instead the *hierarchical distribution:*

$$\mathbb{P}(G, G') = \mathbb{P}(G) \times \mathbb{P}(G' \mid G) = \frac{e^{-\beta E_{2D}(G)}}{\sum_{H} e^{-\beta E_{2D}(H)}} \times \frac{e^{-\beta E_{PK}(G')}}{\sum_{H \cup H'} e^{-\beta E_{PK}(H')}}$$

$E_{2D}(G)$ and $E_{PK}(G')$ are the energies associated with the non-crossing and crossing structures of the RNA, respectively. β is the inverse temperature factor given the Boltzmann constant, and the usual summation is over all possible structures with and without crossing base pairs, respectively. Here, the hierarchical distribution probability coincides with the emission probability of structures generated by a process that first samples non-crossing structures in the Boltzmann distribution $\mathbb{P}(G)$ using stochastic backtrack [5], and then samples from the conditional distribution $\mathbb{P}(G' \mid G)$. However, sampling from the conditional distribution, requires computing the *conditional partition function* (CPF), followed by an adaptation into a conditional version of the stochastic backtrack.

There is a one-to-one correspondence between the search spaces considered by MFE algorithms and McCaskill partition function methods [15] provided that the underlying DP

scheme is unambiguous and complete [16]. This relationship only holds if each substructure is unambiguously computed a single time during the DP MFE calculations [15], i.e. redundant recurrence relation schemes are incompatible with the partition function approach.

Examples of crossing (pseudoknotted) MFE predictions methods that can be extended to the partition function approach include the most general algorithm PKnots [19] ($O(n^6)$ time), NUPACK [28] ($O(n^5)$ time), and CCJ [4] ($O(n^5)$ time), the latter allows for more complex structures like four stems interleaved. Note that $O(n^5)$ time complexity for hypothetical partition function version of CCJ [4] follows. Above-mentioned algorithms that can predict complex RNA structures are slow, motivating novel partition function methods that handle a pseudoknotted RNA structure class efficiently. Calculating a partition function for the general pseudoknotted case is NP-hard [10, 1], therefore pseudoknotted partition function methods require assumptions that limit the complexity of predicted pseudoknots [6]. Heuristic methods or methods that rely on approximations cannot be extended to compute partition function. One example of a heuristic method is HotKnots V2.0 [2], which iteratively selects lowest energy stems to add to the predicted structure, another is IPKnots [21], a maximum-expected-accuracy method. Both methods handle a wide class of pseudoknotted structures efficiently.

Here, we develop an unambiguous non-redundant algorithm based on the hierarchical-folding algorithm HFold [8], which for a given non-crossing structure G, identifies a non-crossing structure G' that is disjoint from G. Specifically, HFold finds the *density-2* [8] structure $G \cup G'$ that is the MFE structure given G, (see [8] for a formal definition, for intuition cf. Fig. 2). The density-2 structure class allows for arbitrary depth and length of nested crossing base pairs including H-type pseudoknots and kissing hairpins. This class of structures includes some structures not handled by CCJ [4], and is larger and more comprehensive than the structure class described by the partition function algorithm from Dirks and Pierce [6]. HFold has a time complexity of $O(n^3)$, significantly improving over PKnots [19], CCJ, and NUPACK.

The HFold recurrence relation scheme [8] is redundant in decomposing density-2 structures for a given non-crossing G, making it incompatible with a partition function approach. Our novel algorithm, *CParty*, unambiguosly decomposes each possible density-2 structure, ensuring a one-to-one mapping to calculating the partition function. CParty computes the CPF with respect to the input structure, which can be used to obtain conditional base pair probabilities and unveil RNA kinetic structure formation paths.

Although a 'full' partition function is ultimately desirable, our focus on a CPF algorithm for density-2 pseudoknotted structures allows us to achieve cubic run time complexity. A CPF is particularly useful in case partial information is available for a given sequence. For example, different structure reactivity datasets may only align for specific subparts of a functional structure e.g. initial non-crossing 'stems', while the rest of the structure is unknown [24]. Furthermore, as structures have been solved at increasingly higher resolution [20], significant differences in features (e.g. variable stem length) remain. Experimental methods provide an important view of RNA structure, but precision of results is an ongoing struggle. Ultimately, the conformational flexibility of RNA makes obtaining conditional base pair propensities valuable in enabling additional analyses.

Contributions. We introduce the CPF algorithm CParty and demonstrate that our algorithm completely and unambiguously decomposes the density-2 structure class. CParty can identify high probability base pairs and unveil changes in base pair probability relative to the most likely initial folding of an RNA molecule. Computing the CPF for density-2 structures with CParty is also useful when the non-crossing structure G is unknown for a specific RNA,

or when there are multiple candidate choices for G. The CPF obtained by CParty informs the choice of a best candidate for non-crossing G, explaining kinetic paths in a novel way. Integrating experimental data from various sources (e.g. SHAPE reactivity, crystallography, cryo-EM) within this CPF approach will enable further statistical analysis for the entire landscape of density-2 secondary structures. We discuss an immediate application for the algorithm: targeting the SARS-CoV-2 frameshifting pseudoknot structure. The procedure defined here to compute the CPF can be expanded to an additional functional structure path analysis for therapeutic development, where conditional base pairing probability has future applications for coronaviruses.

Outline. In Section 2, we define the CPF problem and relevant variants that have runtime benefits. Then, we provide a brief notation review. In Section 3, we proceed with a high level description of CParty and define the constrained partition subfunctions, proving correctness of the DP scheme. Finally, in Section 4, we discuss how CParty can advance functional RNA structure formation analysis in specific applications (e.g. SARS-CoV-2 frameshifting pseudoknot).

2 Preliminaries and problem statement

We define energy $E(G)$ of a RNA secondary structure G according to the realistic pseudoknot energy model of [6] with parameters from [2]. For an RNA sequence S, these energies induce a *Boltzmann distribution* of its structures that characterizes its equilibrium ensemble. In this distribution, the frequency of each structure G of S is proportional to its Boltzmann weight $B(E(G))$. This weight is defined as $B(x) = e^{x/KT}$ at a given temperature T; K is the universal gas constant.

The sum over the Boltzmann weights of the structures of an RNA S is known as its *partition function* Z. More concretely, partition functions are defined for RNAs and specific structure classes Ω, e.g. all density-2 structures, as $Z = \sum_{s \in \Omega} B(E(s))$. Finally, the *ensemble probability* of a structure is $\Pr[G] = \frac{B(E(G))}{Z}$.

Problem Statement. Given an RNA sequence S and a non-crossing structure G, the *Conditional Partition Function problem (CPF problem)* is to compute the partition function Z for the set of all RNA secondary structures $G \cup G'$ of S, where G' is a non-crossing RNA structure that is disjoint from G and $G \cup G'$ is density-2 (the definition of *density-2* follows below).

An interesting variant of the CFP problem forces all base pairs of G' to cross those of G. From experience with HFold, where this variant was introduced as "PKonly", we predict that an algorithm can be derived relatively easily from our work.

Notation. We briefly review the required definitions from HFold [8] to define and predict density-2 structures. Density-2 structures are a subset of the *bi-secondary structure* class (cf. Fig. 3), which is the union of two disjoint pseudoknot-free secondary structures [25], e.g. $G \cup G'$. Here, we refer to bases in the given sequence by their indices, e.g., if the input is 'AUCG', base 0 is A, base 1 is U, and so on. If i is paired with j, then $i.j$ is a *base pair*. A *structure*, R, is a set of base pairs for an RNA sequence over alphabet {A, U, C, G}. The index of the base paired with base at index i in structure R is $bp_R(i)$.

To optimize computation and properly allow nested and non-nested base pairs, we define regions that are self-contained and can or cannot be bifurcated (split). Of key interest in the algorithm design is pseudoloops, which we formally define beginning with crossing base pairs and then carefully grouping these pseudoknotted base pairs into bands and associated

regions. Density of a structure is calculated with respect to all pseudoloops in the structure. Abbreviated nomenclature included below.[2]

- **weakly closed region**: $\forall k \in [i.j]$ either $bp_R(k) \in [i, j]$ or $bp_R(k) = 0$ (unpaired)
- **closed region**: $[i, j]$ is weakly closed, and $\forall l \in [i, j-1]$, neither $[i, l]$ nor $[l+1, j]$ is weakly closed [17]
- **cover(s/ed)**: Base pair $i.j$ *covers* k if $i < k < j$ and $\nexists i'.j'$ such that $i < i' < k < j' < j$
- **cross(es)**: $\exists i, i', j, j'$, such that $i.j$, $i'.j'$, and $i < i' < j < j'$, then $i.j$ *crosses* $i'.j'$
- **pseudoknotted base pair**: $i.j.$ is *pseudoknotted* if $\exists i'.j'$ such that $i.j$ crosses $i'.j'$
- **directly banded in**: Pseudoknotted $i.j$ is *directly banded in* $i'.j'$, denoted $i.j \preceq i'j'$, if $i' \leq i < j \leq j'$ and $[i'+1, i-1]$ and $[j+1, j'-1]$ are weakly closed regions
- **band**: Consider a maximal chain of $\preceq$, if $i.j$ is the band outer closing pair, and $i'.j'$ is the inner closing pair, then $[i, i'] \cup [j'.j]$ is the *band* region
- **band, base pairs, or unpaired bases associated with closed region $[i, j]$**: are not in any closed region or band which are subregions of $[i, j]$ (cf. Fig. 1)
- **pseudoloop**: Unpaired bases and base pairs associated with pseudoknotted closed region $[i, j]$, together with the closing base pairs of the bands associated with $[i, j]$
- **density**: Let L be a pseudoloop and $i.bp_R(i)$ and $bp_R(j).j$ be the closing base pairs of L. Band $[i_1, i'_1] \cup [j'_1, j_1]$ crosses k if $i'_1 < k < j'_1$. Let $\#B(L, k)$ be the number of bands associated with L that cross k. Then the *density* of L is: $max_{i \leq k \leq j}(\#B(L, k))$. It follows that the *density* of a structure, R, is maximum density of L over all *pseudoloops* L of R. R is *density-k* if the density of R is at most k.

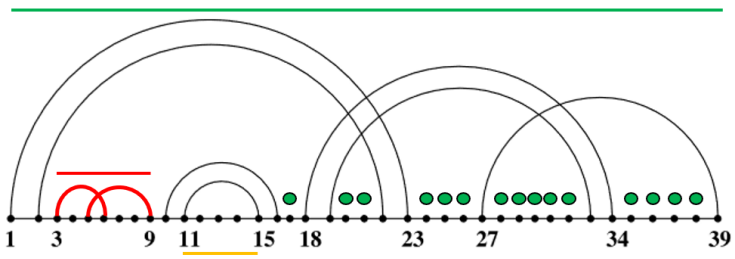

Figure 1 The three *bands* $[1, 2] \cup [22, 23]$, $[18, 19] \cup [33, 34]$ and $[27, 27] \cup [39, 39]$ are associated with closed region $[1, 39]$, whereas $[3, 3] \cup [6, 6]$ and $[5, 5] \cup [9, 9]$ are associated with closed region $[3, 9]$. Unpaired bases associated with closed region $[1, 39]$ are represented with a green dot (c.f. •). Closed regions $[3, 9]$ and $[10, 16]$ are associated with *pseudoloop* $[1, 39]$ but closed region $[11, 15]$ is *not*. Figure modified from [8].

3 Efficient computation of the conditional partition function

HFold review. In the HFold hierarchical folding algorithm, $W(i, j)$ denotes the energy of the MFE pseudoknot-free secondary structure for a given subsequence $s_i s_{i+1} \ldots s_j$ and a given pseudoknot-free structure $G_{i,j}$ taken over all choices of pseudoknot-free G'_{ij} which is disjoint from G_{ij}, and such that $G_{i,j} \cup G'_{i,j}$ is of density-2. W is computed via structure class $V(i, j)$ with $i.j$, and structure class $WMB(i, j)$, where i and j form a pseudoloop. $WMB(i, j)$ is further classified to distinguish pseudoloops with rightmost band in G' as

[2] See [8] for full definitions.

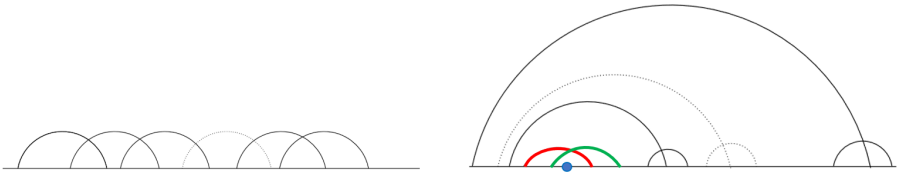

Figure 2 Examples of density-2 structures with arbitrary number and depth of bands [8]. Blue dot covered by red and green bands indicates how association with closed region maintains the density-2 property. Figure modified from [8].

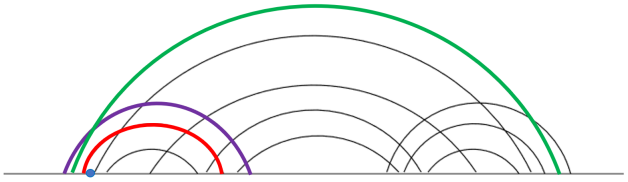

Figure 3 Example of *bi-secondary structure* that is not a density-2 structure [8]. Blue dot covered by red, green, and purple bands indicate density-2 property not satisfied. Figure modified from [8].

structure class $WMB'(i,j)$. Each successive band in a pseudoloop is computed via structure classes BE for bands in G, and VP for bands in G'. Following the approach of HFold, our CParty algorithm utilizes additional structure classes to ensure unambiguous decomposition of the density-2 structure class (cf. Table 1).

We now introduce the CParty DP scheme in detail, conditions for terms in summation are shown in blue. CParty utilizes an energy model requiring complex recurrences and additional penalties (cf. Appendix Table 2) compared to other more simple models [6] e.g. base pair maximization. Due to space constraints, energy parameter table (cf. Table 2) and structure classes with straightforward transition from HFold [8] (WI, WI', V, BE) are described in the Appendix.

3.1 General density-2 structures

$Z_W(i,j)$ denotes the partition function over all structures $G_{i,j} \cup G'_{i,j}$ for the subsequence $s_i \dots s_j$ taken over all choices of $G'_{i,j}$ (which is pseudoknot-free, disjoint from $G_{i,j}$, and such that $G_{i,j} \cup G'_{i,j}$ is density-2) and i and j are not covered by a base pair in G, denoted $\overline{isCovered(G,i)}$ and $\overline{isCovered(G,j)}$. Note that the structure is weakly closed and can possibly be partitioned into two independent substructures. There is no penalty for having unpaired bases at either end of the structure. Base cases are as follows: if $i \geq j$, $Z_W(i,j) = 1$, since the substructure is empty and the only possibility for structure $R_{i,j}$ is the empty structure. In addition, $Z_W(i,j) = 0$, if $isCovered(G,i)$ or $isCovered(G,j)$, to enforce the exterior structure condition.

$$Z_W(i,j) = \sum \begin{cases} (1) \quad \sum\limits_{\substack{i \leq r < j \\ \overline{isCovered(r)}}} Z_W(i,r-1) \cdot Z_V(r,j) \\ (2) \quad Z_W(i,j-1) \\ (3) \quad \sum\limits_{\substack{i \leq r < j \\ \overline{isCovered(r)}}} Z_W(i,r-1) \cdot Z_P(r,j) \cdot B(P_s) \end{cases} \tag{i}$$

Table 1 CParty Structure Classes. First column denotes each structure class abbreviation with description in second column; for example, W: general density-2 structure class. In the third column we delineate additional structure classes that need to be computed; for example, in the first row W is computed via structure class V such that $i.j$, structure class P such that i and j form a pseudoloop, and also via recursive call again to W via bifurcation.

Structure Class	Description	Computed via
W	General density-2 structures	W, V, P
V	$i.j$	VM
P	Pseudoknotted density-2 structures	PG', BE, WI
PG'	Pseudoknotted density-2 structures rightmost band in G'	PG', PG'^w, VP, BE, WI
PG'^w	Pseudoknotted density-2 structures rightmost band in G', weakly closed region between bands	PG', WI
WI	Weakly closed region inside pseudoloop	V, P, WI
WI'	Non-empty weakly closed region inside band	V, P, WI'
VP	$i.j$ in G' crosses base pair in G	VP, VP^R, VP^L, WI, WI'
VP^R	$i.bp(i)$ in G' crosses base pair in G bp(i) $\neq j$	VP, WI'
VP^L	$bp(j).j$ in G' crosses base pair in G bp(j) $\neq i$	VP
VM	$i.j$ close a multiloop	WM, WM^1, WM^P
WM	$i.j$ on a multiloop	WM, V, P, WM^P
WM^1	$i.j$ on a multiloop terminal helix pseudoknot-free	WM^1, V
WM^P	$i.j$ on a multiloop terminal helix pseudoknotted	WM^P, P
BE	$[i, i'] \cup [bp(i'), bp(i)]$ band region	BE, WI'

Case (1) bifurcates such that $\exists r$, $i \leq r < j$, $bp_R(i) \leq r - 1$ (i.e., i is either unpaired or paired with another base inside region $[i, r-1]$), and $bp_R(j) = r$ (i.e., j is paired with r).

Case (2) enforces j as unpaired: $bp_R(j) = 0$ and $bp_R(i) \leq r$ (i.e., i is either unpaired or paired with another base inside region $[i, j-1]$).

Case (3) bifurcates such that $bp_R(i) \leq r - 1$ (i.e., i is either unpaired or paired with another base inside region $[i, r-1]$), and $\exists q, x$ such that $r < q < x < j$, $bp_R(r) = x$ and $bp_R(j) = q$ (i.e., $r.x$ crosses $j.q$). In this case, we add a penalty, P_s, for introducing an exterior pseudoknot.

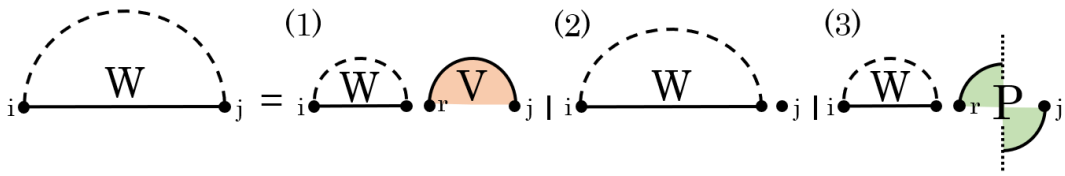

Figure 4 $Z_W(i, j)$ structure class. Case (1) illustrating $r.j$ close a loop. Case (2) enforces j unpaired. Finally for Case (3) $r.j$ form a pseudoloop. Dotted vertical lines indicate overlapping chain of bands can continue and that the chain can initialize or terminate via either G (above horizontal line) or G' (below horizontal line).

Z_W fully decomposes the density-2 structure class (cf. Fig. 4), and is unambiguous to ensure compatibility with the partition function approach. The proof of correctness works by structural induction, showing the correctness of each single recurrence. Proving the correctness of $Z_W(i, j)$ serves as an example.

▶ **Lemma 1.** *The recurrence of $Z_W(i,j)$ is complete, correct, and unambiguous.*

$Z_W(i,j)$ is the partition function over the class of $W(i,j)$ structures for all i and j that are not covered by a base pair in G. The structure class $W(i,j)$ is the set of structures $G_{i,j} \cup G'_{i,j}$ for the subsequence $s_i \dots s_j$ taken over all choices of $G'_{i,j}$ (which is pseudoknot-free, disjoint from $G_{i,j}$ and such that $G_{i,j} \cup G'_{i,j}$ is density-2). The non-covered condition restricts for which i and j the class is defined at all. Here we define everything for one fixed G.

By definition of density-2, for structure class $W(i,j)$, j is either unpaired, or pairs with $i \leq k < j$ such that $k.j$ closes a loop or j is the rightmost end of a chain of crossing base pairs. Since these cases are disjoint, the recurrence is unambiguous; since every structure falls into one of these cases, it is complete. Finally, it is correct, since partition functions can be correctly inferred from smaller subproblems (which are correct by induction hypothesis): If j is unpaired, $Z_W(i,j) = Z_W(i,j-1)$ by $Z_W(2)$. If j is paired with k and is not pseudoknotted i.e. $\nexists h,l$ such that $h.l$ crosses $bp(j).j$, $Z_W(i,j) = Z_W(i,k-1) \cdot Z_V(k,j)$ by $Z_W(1)$. If j is the rightmost end of a chain of crossing base pairs i.e. $\exists x,z$ such that $x.z$ crosses $bp(j).j$, $Z_W(i,j) = Z_W(i,k-1) \cdot Z_P(k,j) \cdot B(P_s)$ by $Z_W(3)$. ■

3.2 Pseudoknotted density-2 structures

When $G_{i,j} \cup G'_{i,j}$ includes pseudoknotted base pairs and cannot be partitioned into two independent substructures for two regions $[i,r]$ and $[r+1,j]$ for any r, such a structure must contain a chain of two or more successively-overlapping bands. This chain of overlapping bands must alternate between $G_{i,j}$ and $G'_{i,j}$, possibly with nested substructures interspersed throughout.

$Z_P(i,j)$ is the partition function over all pseudoknotted density-2 structures between i and j where the rightmost band of the pseudoloop may be in G or G' and the structure is closed (cannot be further partitioned).

Z_{BE} is the partition function over all structures of bands spanned by base pairs of $G_{i,j}$, with an associated band initiation penalty, P_b. Z_{VP} is the partition function over structures enclosed by pseudoknotted base pairs of $G'_{i,j}$ (excluding subregions that are within a band of $G_{i,j}$). Z_{WI} is the partition function over all structures in weakly closed subregions of a pseudoloop.

For $Z_P(i,j)$ base case, if $i \geq j$, then $Z_P(i,j) = 0$, since the substructure is empty. Otherwise, there are two cases (cf. Fig. 5), Case (1): j is paired in G, or Case (2): j is not paired in G. If j is paired in G, then in the MFE structure, some base l with $bp_G(j) < l < j$ must be paired in G', causing $bp_G(j).j$ to be pseudoknotted. We consider all possible choices of l. Once l is fixed, the inner base pair of the band whose outer base pair is $bp_G(j).j$ is also determined (cf. [8]). If j forms a pseudoloop with i and is not paired in G, we move to $Z_{PG'}$, partition function over pseudoknotted density-2 structures with rightmost unconsumed band in G'. Here we note these two cases are unambiguous, with j either paired in G or G'.

$$Z_P(i,j) = \sum \begin{cases} (1) \sum\limits_{\substack{bp_G(j)<l<j \\ bp_G(l)=0}} Z_{BE}(b_{(i,l)}, b'_{(i,l)}) \cdot Z_{PG'}(i,l) \cdot Z_{WI}(l+1, bp_G(b'_{(i,l)})-1) \cdot B(P_b) \\ (2) \quad Z_{PG'}(i,j) \end{cases} \tag{ii}$$

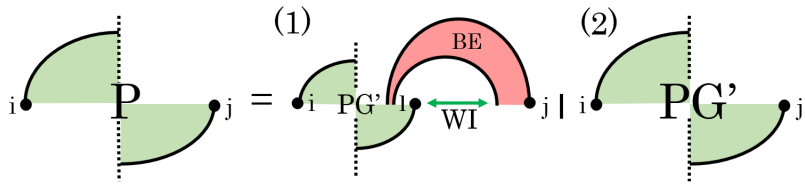

Figure 5 Cases of Z_P. (1) j is paired in G and there must be some base, l, between $bp(j).j$ that is paired in G'. (2) j is not paired in G, then move directly to $Z_{PG'}$.

3.2.1 Pseudoknotted structures with rightmost unconsumed band in G'

When i and j form a pseudoloop where the rightmost unconsumed band of the pseudoloop is not in G, i.e. $bp_G(j) = 0$, it must be part of the structure of G'. To avoid allowing multiple adjacent weakly closed subregions in the pseudoloop, we introduce $Z_{PG'^w}(i, j)$, the partition function over all pseudoknotted density-2 structures between i and j where the unconsumed rightmost band of the pseudoloop is part of the structure of G' and the structure is weakly closed.

Therefore, Cases $(1-2)$ of $Z_{PG'}$ (cf. Fig. 6) are used to account for the energy of the region spanned by the rightmost two bands using Z_{BE} and $Z_{VP}(l, j)$; and recursively continuing again to $Z_{PG'}$, or to $Z_{PG'^w}$ if there is a weakly closed region between the bands. For band border details see [8]. $Z_{PG'}$ Cases $(3-4)$ are end cases, where only one or two bands, respectively, need to be accounted for. For the base case when $i \geq j$, $Z_{PG'} = Z_{PG'^w} = 0$.

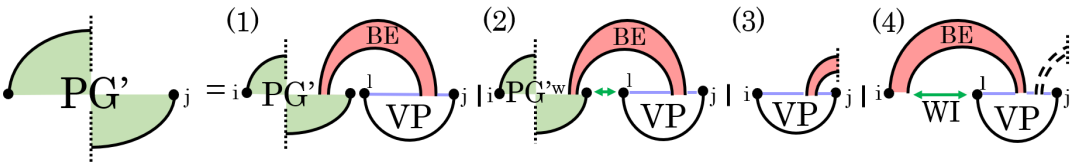

Figure 6 Cases of $Z_{PG'}$. (1) handles two elements of the chain and continues. (2) is similar except there is a weakly closed region between the bands, this will be handled by $Z_{PG'^w}$ structure class to preserve the cubic time complexity. For the end cases we have (3) initial band of chain in G'; and (4) initial band in G.

$$Z_{PG'}(i,j) = \sum \begin{cases} (1) \text{ if } bp_G(j) = 0 \\ \displaystyle\sum_{\substack{i<l<b_{(i,l)} \\ isCovered(G_{i,j},l)}} Z_{BE}(b_{(i,l)}, b'_{(i,l)}) \cdot Z_{PG'}(i, l-1) \cdot Z_{VP}(l, j) \cdot B(2P_b) \\ (2) \text{ if } bp_G(j) = 0, \text{ and } bp_G(l-1) < (l-1) \\ \displaystyle\sum_{\substack{i<l<b_{(i,l)} \\ isCovered(G_{i,j},l)}} Z_{BE}(b_{(i,l)}, b'_{(i,l)}) \cdot Z_{PG'^w}(i, l-1) \cdot Z_{VP}(l, j) \cdot B(2P_b) \\ (3)\ Z_{VP}(i,j) \cdot B(P_b) \\ (4) \text{ if } 0 = bp_G(j) < bp_G(i) \\ \displaystyle\sum_{i<l<bp_G(i)} Z_{BE}(b_{(i,l)}, b'_{(i,l)}) \cdot Z_{WI}(b'_{(i,l)} + 1, l-1) \cdot Z_{VP}(l, j) \cdot B(2P_b) \end{cases} \tag{iii}$$

$$Z_{PG'^w}(i,j) = \sum_{\substack{i<l<j \\ cover(l)=cover(j)}} Z_{PG'}(i,l) \cdot Z_{WI}(l+1,j) \tag{iv}$$

3.3 $i.j$ in G' crosses base pair in G

$Z_{VP}(i,j)$ is the partition function over all structures $R_{i,j}$ in which $i.j \in G'$ and crosses a base pair in G. The energy of $R_{i,j}$ is the energy of all loops within $R_{i,j}$ that are not inside a band whose base pairs are in G and which crosses $i.j$. If $i \geq j$, i or j is paired in G', or $i.j$ does not cross any base pair of G, then $Z_{VP}(i,j) = 0$, otherwise $Z_{VP}(i,j)$ is computed as follows (cf. Fig. 7):

$Z_{VP}(i,j)$ Cases $(1-3)$ handle structures where there are no other base pairs in $R_{i,j}$ that cross the band(s) $i.j$ crosses.[3] These cases are unambiguous: either i is covered, j is covered, or both. In Case (4), $(i+1).(j-1)$ forms a stacked pair (e_{stP} penalty). In Case (5), $i.j$ and $r.r'$ close an internal loop (e_{intP} penalty). In Case (6), $r.(j-1)$ crosses a base pair in G and $[i+1, r-1]$ is a weakly closed non-empty region (multiloop spanning band initiation and base pair penalties, a' and b', respectively). In Case (7), $(i+1).r$ crosses a base pair in G and $[r+1, j-1]$ is a weakly closed non-empty region. In Case (8), $[i+1, r-1]$ is a weakly closed non-empty region, $r.bp(r)$ crosses a base pair in G, and $[bp(r)+1, j-1]$ is either empty or non-empty and weakly closed. Therefore, we introduce Z_{VP^R} (cf. Fig. 8), the partition function over all structures such that $i.bp(i) \in G'$ crosses base pair in G, and $bp(i) \neq j$ (distinct from Case (6)). Finally, in Case (9), $[r+1, j-1]$ is a weakly closed non-empty region, $r.bp(r)$ crosses a base pair in G, and $[i+1, bp(r)-1]$ is empty. Therefore, we introduce Z_{VP^L}, the partition function over all structures such that $bp(j).j \in G'$ crosses base pair in G, $bp(j) \neq i$ (distinct from Case (7)), and $[i+1, bp(j)-1]$ is empty. Note that c' is penalty for unpaired base in a multiloop spanning a band.

$$Z_{VP}(i,j) = \sum \begin{cases} (4) \text{ if } (bp_G(i+1) = 0, \text{ and } bp_G(j-1) = 0) \\ \quad B(e_{stP}(i, i+1, j-1, j)) \cdot Z_{VP}(i+1, j-1) \\ (5) \text{ if } cover(G,i) = cover(G,r) \text{ and } cover(G,j) = cover(G,r') \\ \quad \displaystyle\sum_{\substack{i<r<\min(B'_{(i,j)}, b_{(i,j)}) \\ \max(b'_{(i,j)}, B_{(i,j)})<r'<j}} (B(e_{intP}(i,r,r',j)) \cdot Z_{VP}(r,r')) \\ (6) \displaystyle\sum_{\substack{i<r<\min(B'_{(i,j)}, j) \\ bp_G(r)=0}} Z_{WI'}(i+1, r-1) \cdot Z_{VP}(r, j-1) \cdot B(a' + 2b') \\ (7) \displaystyle\sum_{\substack{max(i, b'_{(i,j)})<r<j \\ bp_G(r)=0}} Z_{VP}(i+1, r) \cdot Z_{WI'}(r+1, j-1) \cdot B(a' + 2b') \\ (8) \displaystyle\sum_{\substack{i<r<\min(B'_{(i,j)}, j) \\ bp_G(r)=0}} Z_{WI'}(i+1, r-1) \cdot Z_{VP^R}(r, j-1) \cdot B(a' + 2b') \\ (9) \displaystyle\sum_{\substack{max(i, b'_{(i,j)})<r<j \\ bp_G(r)=0}} Z_{VP^L}(i+1, r) \cdot Z_{WI'}(r+1, j-1) \cdot B(a' + 2b') \end{cases} \tag{v}$$

[3] Recurrences moved to Appendix for space, for band border lemmas see [8].

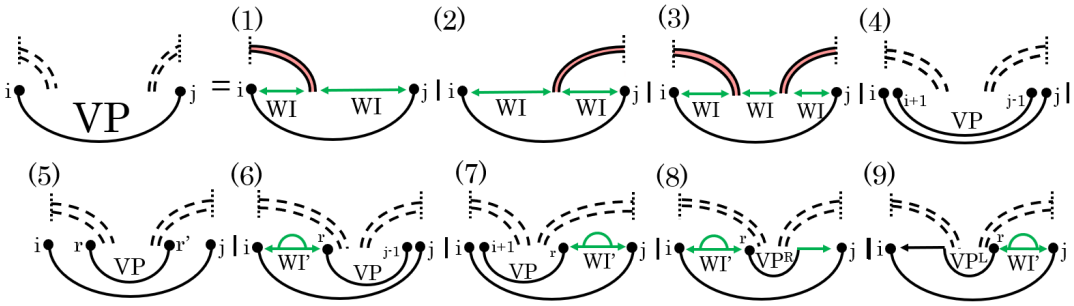

Figure 7 Cases of VP. $(1-3)$ either two or three WI subregions (green) between i and j, band regions excluded. $(4-5)$, stacked pair and internal loop, respectively. $(6-9)$, $i.j$ closes a multiloop spanning a band.

$i.bp(i)$ **in** G' **crosses base pair in** G**,** $bp(i) \neq j$

$$Z_{VP^R}(i,j) = \sum \begin{cases} (1) \displaystyle\sum_{max(i,b'_{(i,j)})<r<j} Z_{VP}(i,r) \cdot Z_{WI'}(r+1,j) \\ (2) \text{ if empty}(G,[(r+1),j]) \\ \displaystyle\sum_{max(i,b'_{(i,j)})<r<j} Z_{VP}(i,r) \cdot B(c'(j-r)) \end{cases} \tag{vi}$$

$bp(j).j$ **in** G' **crosses base pair in** G**,** $bp(j) \neq i$

$$Z_{VP^L}(i,j) = \begin{array}{c} \text{if empty}(G,[i,(r-1)]) \\ \displaystyle\sum_{i<r<min(B'_{(i,j)},j} B(c'(r-i)) \cdot Z_{VP}(r,j) \end{array} \tag{vii}$$

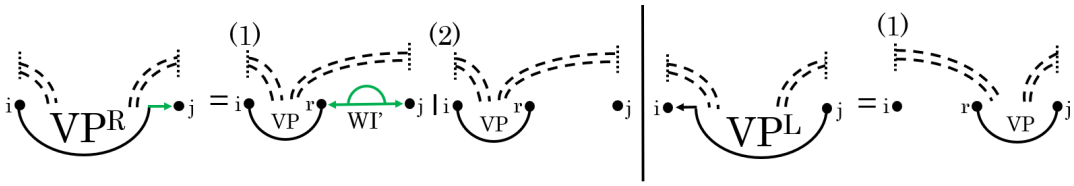

Figure 8 Cases of VP^R and VP^L case. VP^R (1) weakly closed non-empty region $[r+1,j]$, (2) empty region $[r+1,j]$. VP^L (1) empty region $[i,r-1]$.

3.4 $i.j$ close a multiloop

$Z_{VM}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed and $i.j$ closes a multiloop (cf. Fig. 9). Otherwise, $Z_{VM}(i,j) = 0$

In Case (1) we have at least two branches, with the terminal pseudoknot-free helix (multiloop initiation and base pair penalties are denoted a and b, respectively). Z_{WM} is the partition function over all structures such that i and j are on a multiloop; and Z_{WM^1}, the partition function over all structures such that $i.bp(i)$ is the terminal helix on a multiloop and is pseudoknot-free. In Case (2) we have at least three branches, with the terminal pseudoknotted helix. Z_{WM^P} is the partition function over all structures such that $i.bp(i)$

is the terminal helix on a multiloop and forms a pseudoloop. Finally, in Case (3) we have exactly two branches, which form a pseudoloop (c is unpaired base in multiloop penalty).

$$Z_{VM}(i,j) = \sum \begin{cases} (1) \sum\limits_{(i+1)<h\leq(j-1)} Z_{WM}(i+1,h-1) \cdot Z_{WM^1}(h,j-1) \cdot B(a+b) \\ (2) \sum\limits_{(i+1)<h\leq(j-1)} Z_{WM}(i+1,h-1) \cdot Z_{WM^P}(h,j-1) \cdot B(a+b) \\ (3) \sum\limits_{i<r<(j-1)} B(c(r-i-1)) \cdot Z_{WM^P}(r,j-1) \cdot B(a+b) \end{cases} \tag{viii}$$

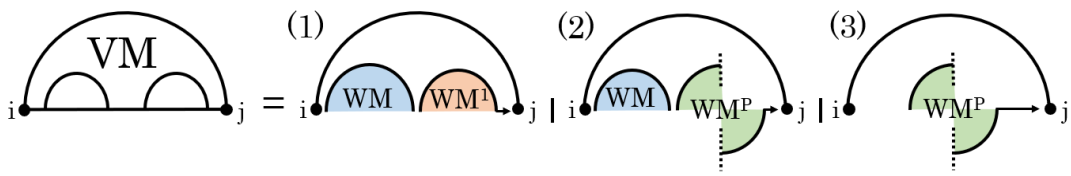

Figure 9 Cases of VM. (1) terminal pseudoknot-free helix, (2) terminal pseudoknotted helix with additional branches. (3) terminal pseudoknotted helix, no additional branches.

i and j on a multiloop $Z_{WM}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed, not empty, and i and j are on a multiloop (cf. Fig. 10). For base case where $i \geq j$, $Z_{WM}(i,j) = 0$. In Case (1), $r.j$ is pseudoknot-free and is the initial helix, in Case (2), r and j form a pseudoloop and the initial helix is pseudoknotted (P_{sm} penalty for introducing pseudoknot inside a multiloop). In Case (3), $(r+1).j$ is pseudoknot-free and is not the initial helix. In Case (4), $(r+1)$ and j form a pseudoloop that does not contain the initial helix. In Case (5), j is unpaired.

$$Z_{WM}(i,j) = \sum \begin{cases} (1) \sum\limits_{i\leq r<j} B(c(r-i)) \cdot Z_V(r,j) \cdot B(b) \\ (2) \sum\limits_{i\leq r<j} B(c(r-i)) \cdot Z_P(r,j) \cdot B(P_{sm}+b) \\ (3) \sum\limits_{i<r<(j-1)} Z_{WM}(i,r) \cdot Z_V(r+1,j) \cdot B(b) \\ (4) \sum\limits_{i<r<(j-1)} Z_{WM}(i,r) \cdot Z_P(r+1,j) \cdot B(P_{sm}+b) \\ (5)\ Z_{WM}(i,j-1) \cdot B(c) \end{cases} \tag{ix}$$

i and j on a multiloop, terminal helix pseudoknot-free $Z_{WM^1}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed, not empty, and i and j are on a multiloop (terminal helix, cf. Fig. 11). With Case (1), $i.j$ form a pseudoknot-free loop, and Case (2), j is unpaired. For base case where $i \geq j$, $Z_{WM^1}(i,j) = 0$.

$$Z_{WM^1}(i,j) = \sum \begin{cases} (1) \quad Z_V(i,j) \cdot B(b) \\ (2) \quad Z_{WM^1}(i,j-1) \cdot B(c) \end{cases} \tag{x}$$

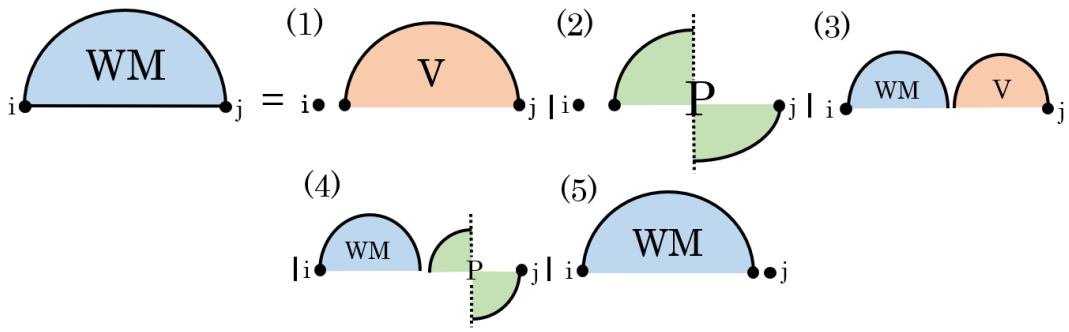

Figure 10 Cases of WM. (1) initial helix pseudoknot-free, (2) initial helix pseudoknotted. (3) intermediate helix pseudoknot-free, additional branches. (4) intermediate helix pseduoknotted, additional branches. (5), j is unpaired.

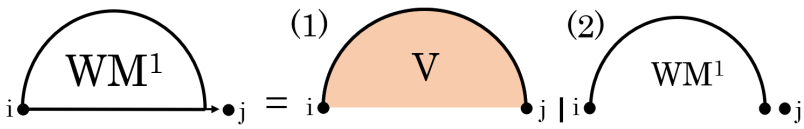

Figure 11 Cases of WM^1. (1) terminal helix pseudoknot-free, (2) j is unpaired.

i and j on a multiloop, terminal helix pseudoknotted $Z_{WM^P}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed, not empty, and i and j are on a multiloop (pseudoknotted terminal branch, cf. Fig. 12). With Case (1), $i.j$ form a pseudoloop, and Case (2), j is unpaired. For base case where $i \geq j$, $Z_{WM^P}(i,j) = 0$.

$$Z_{WM^P}(i,j) = \sum \begin{cases} (1) & Z_P(i,j) \cdot B(P_{sm} + b) \\ (2) & Z_{WM^P}(i,j-1) \cdot B(c) \end{cases} \tag{xi}$$

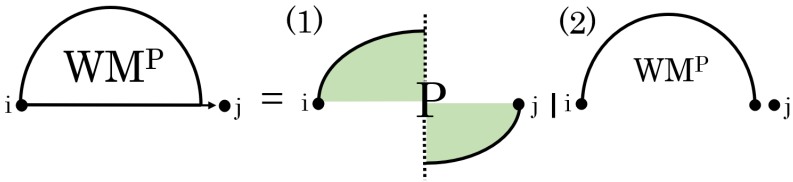

Figure 12 Cases of WM^P. (1) terminal helix pseudoknotted, (2) j is unpaired.

4 Discussion

Efficiently computing the partition function for RNA secondary structure is the problem solved 20 years ago by the McCaskill algorithm and subsequent variants. McCaskill described a pseudoknot-free $O(N^4)$ DP algorithm to compute the total sum of energy in the system (partition function), further reduced to $O(N^3)$ using a simplified energy model [15]. Recently, a linear time approximation algorithm was proposed for pseudoknot-free partition function [29]. However, the best known time complexity for more general pseudoknotted MFE secondary structure prediction is $O(N^5)$ [1]. Existing pseudoknotted partition function methods require

certain assumptions that limit the complexity of predicted pseudoknots [6], and methods with restricted partition function implementations for the pseudoknotted case are too slow for long RNA sequences (e.g. NUPACK [28]). To circumvent these issues, different approaches have been adopted to enable more efficient DP for restricted landscape of pseudoknotted structures.

In light of the hierarchical folding hypothesis (non-crossing base pairs form first followed by additional possibly pseudoknotted base pairs), we introduce the biologically motivated CParty algorithm. CParty achieves our goal of an efficient CPF algorithm for the density-2 structure class via an $O(N^3)$ DP scheme. This allows us to sample from the hierarchical distribution $\mathbb{P}(G) \times \mathbb{P}(G' \mid G)$. To improve runtime for computing the partition function, we followed the approach of the HFold MFE algorithm, which restricts the search space from the general pseudoknotted case to a subset of bi-secondary structures i.e. density-2 structures. CParty achieves a cubic runtime, which is a significant improvement over other partition function methods that handle complex pseudoknots [28, 19, 6, 4].

By incorporating structure constraints into sampling efforts, a CPF can provide more accurate predictions of RNA secondary structures in specific biological contexts. With CParty we can obtain probabilistic structural information to assist identification of novel target sites and contribute to therapeutics. CParty is valuable in comprehensively investigating the landscape of possible structural paths which is complicated and expensive experimentally. Next, we describe how a CPF for complex RNA structures will allow probabilistic investigation of specific base pairs or base-pairing patterns to advance understanding of RNA structure formation.

4.1 CParty for analyzing pseudoknot motifs in SARS-CoV-2

There has been widespread research focusing on structures for the SARS-CoV-2 frameshift site, including multiple viral genome reactivity probing experiments [27, 11, 30, 7]. Based on the available SHAPE datasets and thermodynamic-based predictions [2], multiple candidates exist for initial non-crossing structure at the frameshifting site of SARS-CoV-2 [24]. It was shown previously [24, 18] that suboptimal structures (i.e. structures close to the MFE structure in the energy landscape) may play important roles in function of the viral sequence. Finding folding paths from original base pairs to the MFE as well as suboptimal structures can be essential for designing effective therapeutics. While multiple implementations of partition function calculations exist that can estimate base pair probabilities in pseudoknot-free structures, there are few partition function algorithms that can handle pseudoknotted structure and none that can provide conditional partition function to shed light on folding paths. Employing CParty, we can provide a stochastic view of suboptimal folding for the pseudoknotted structure of SARS-CoV-2 frameshift element, and help in identification of novel target sites for therapy. CParty can also help in evaluating existing predicted structural motifs [22, 3], length-dependent conformations [26], and structural alignment between multiple coronaviruses [24]. Identifying such structural formation pathways for SARS-CoV-2 frameshifting pseudoknot site can motivate future pan-coronaviral therapeutic RNA structure research.

CParty can be similarly applied in other applications to better harness available structure information to isolate key kinetic paths from initial non-crossing structure to pseudoknotted structure, with implications for subsequent tertiary structure modeling.

5 Conclusion

Here, we presented CParty, a novel biologically motivated algorithm following the hierarchical folding hypothesis that efficiently computes the conditional partition function for density-2 structures. Our algorithm CParty efficiently removed the ambiguity of the dynamic programming algorithm of HFold, enabling a conditional partition function approach that samples from the hierarchical distribution $G \cup G'$ for given non-crossing G. CParty is significantly more efficient in time and space compared to the NUPACK partition function algorithm for simple pseudoknots, while also handling a larger structure class including kissing hairpins and arbitrary depth of nested substructures of the same complexity. Computing the conditional partition function can be a smart alternative to an unconditional pseudoknot partition function, because it is very fast and can still be informative in many scenarios. In future work, we will implement CParty for stochastic analysis of pseudoknotted RNAs, enabling novel applications to evaluate hierarchical kinetics. We plan to systematically apply CParty to analyze pseudoknotted structure motifs as sketched in Subsection 4.1.

References

1. Tatsuya Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1-3):45–62, 2000.
2. 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.
3. Pramod R Bhatt, Alain Scaiola, Gary Loughran, Marc Leibundgut, Annika Kratzel, Romane Meurs, René Dreos, Kate M O'Connor, Angus McMillan, Jeffrey W Bode, et al. Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome. *Science*, 372(6548):1306–1313, 2021.
4. Ho-Lin Chen, Anne Condon, and Hosna Jabbari. An o (n 5) algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. *Journal of Computational Biology*, 16(6):803–815, 2009.
5. Ye Ding and Charles E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic acids research*, 31:7280–7301, December 2003. doi:10.1093/nar/gkg938.
6. 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.
7. Nicholas C Huston, Han Wan, Rafael de Cesaris Araujo Tavares, Craig Wilen, and Anna Marie Pyle. Comprehensive in-vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *BioRxiv*, 2020.
8. Hosna Jabbari, Anne Condon, Ana Pop, Cristina Pop, and Yinglei Zhao. HFold: RNA pseudoknotted secondary structure prediction using hierarchical folding. In *International Workshop on Algorithms in Bioinformatics*, pages 323–334. Springer, 2007.
9. Ronny Lorenz, Ivo L. Hofacker, and Peter F. Stadler. RNA folding with hard and soft constraints. *Algorithms Mol. Biol.*, 11(1):1–13, December 2016. doi:10.1186/s13015-016-0070-z.
10. 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, 2000.
11. Ilaria Manfredonia, Chandran Nithin, Almudena Ponce-Salvatierra, Pritha Ghosh, Tomasz K Wirecki, Tycho Marinus, Natacha S Ogando, Eric J Snijder, Martijn J van Hemert, Janusz M Bujnicki, et al. Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. *bioRxiv*, 2020.
12. David H Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna*, 10(8):1178–1190, 2004.

13 David H Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology*, 288(5):911–940, 1999.

14 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.

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

16 Yann Ponty and Cédric Saule. A combinatorial framework for designing (pseudoknotted) RNA algorithms. In M.-F. Sagot T. Przytycka, editor, *Algorithms in Bioinformatics*, number 6833 in LNBI, pages 250–269. Springer, January 2011. URL: `http://dx.doi.org/10.1007/978-3-642-23038-7_22`, `doi:10.1007/978-3-642-23038-7_22`.

17 Baharak Rastegari and Anne Condon. Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. *Journal of computational biology*, 14(1):16–32, 2007.

18 Dustin B Ritchie, Daniel AN Foster, and Michael T Woodside. Programmed- 1 frameshifting efficiency correlates with rna pseudoknot conformational plasticity, not resistance to mechanical unfolding. *PNAS*, 109(40):16167–16172, 2012.

19 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.

20 Christina Roman, Anna Lewicka, Deepak Koirala, Nan-Sheng Li, and Joseph A Piccirilli. The SARS-CoV-2 programmed- 1 ribosomal frameshifting element crystal structure solved to 2.09 å using chaperone-assisted RNA crystallography. *ACS Chemical Biology*, 2021.

21 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):i85–i93, 2011.

22 Tamar Schlick, Qiyao Zhu, Abhishek Dey, Swati Jain, Shuting Yan, and Alain Laederach. To knot and not: Multiple conformations of the SARS-CoV-2 frameshifting RNA element. *bioRxiv*, 2021.

23 Ignacio Tinoco Jr and Carlos Bustamante. How RNA folds. *J. Mol. Biol.*, 293(2):271–281, 1999.

24 Luke Trinity, Ian Wark, Lance Lansing, Hosna Jabbari, and Ulrike Stege. Shapify: Paths to SARS-CoV-2 frameshifting pseudoknot. *PLOS Computational Biology*, 19(2):e1010922, 2023.

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

26 Shuting Yan, Qiyao Zhu, Swati Jain, and Tamar Schlick. Length-dependent motions of SARS-CoV-2 frameshifting RNA pseudoknot and alternative conformations suggest avenues for frameshifting suppression. *Research Square*, 2022.

27 Siwy Ling Yang, Louis DeFalco, Danielle E Anderson, Yu Zhang, Jong Ghut Ashley Aw, Su Ying Lim, Xin Ni Lim, Kiat Yee Tan, Tong Zhang, Tanu Chawla, et al. Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions. *Nat. Commun.*, 12(1):1–15, 2021.

28 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.

29 He Zhang, Liang Zhang, David H Mathews, and Liang Huang. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics*, 36(Supplement_1):i258–i267, 2020.

30 Kaiming Zhang, Ivan N Zheludev, Rachel J Hagey, Marie Teng-Pei Wu, Raphael Haslecker, Yixuan J Hou, Rachael Kretsch, Grigore D Pintilie, Ramya Rangan, Wipapat Kladwang, et al. Cryo-electron microscopy and exploratory antisense targeting of the 28-kda frameshift stimulation element from the SARS-CoV-2 RNA genome. *bioRxiv*, 2020.

31 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.

A Appendix

Table 2 ENERGY PARAMETERS

Name	*Description*
P_s	Exterior pseudoloop initiation penalty
P_{sm}	Penalty for introducing pseudoknot inside a multiloop
P_{sp}	Penalty for introduce pseudoknot inside a pseudoloop
P_b	Band penalty
P_{up}	Penalty for unpaired base in a pseudoloop
P_{ps}	Penalty for closed subregion inside a pseudoloop
$e_H(i,j)$	Energy of a hairpin loop closed by $i.j$
$e_S(i,i+1,j-1,j)$	Energy of a stacked pair closed by $i.j$
$e_{stP}(i,i+1,j-1,j)$	Energy of a stacked pair that spans a band
$e_{int}(i,r,r',j)$	Energy of a pseudoknot-free internal loop
$e_{intP}(i,r,r',j)$	Energy of internal loop that spans a band
a	Multiloop initiation penalty
b	Multiloop base pair penalty
c	Penalty for unpaired base in a multiloop
a'	Penalty for introducing a multiloop that spans a band
b'	Base pair penalty for a multiloop that spans a band
c'	Penalty for unpaired base in a multiloop that spans a band

A.1 $i.j$ in G' crosses base pair in G

$Z_{VP}(i,j)$ is the partition function over all structures $R_{i,j}$ in which $i.j \in G'$ and crosses a base pair in G. The energy of $R_{i,j}$ is the energy of all loops within $R_{i,j}$ that are not inside a band whose base pairs are in G and which crosses $i.j$. If $i \geq j$, i or j is paired in G', or $i.j$ does not cross any base pair of G, then $Z_{VP}(i,j) = 0$, otherwise $Z_{VP}(i,j)$ is computed as follows:

$$Z_{VP}(i,j) = \sum \begin{cases} (1) \text{ if } isCovered(G,i), \text{ and } \overline{isCovered(G,j)} \\ \quad Z_{WI}(i+1, B'_{(i,j)} - 1) \cdot Z_{WI}(B_{(i,j)} + 1, j-1) \\ (2) \text{ if } \overline{isCovered(G,i)}, \text{ and } isCovered(G,j) \\ \quad Z_{WI}(i+1, b_{(i,j)} - 1) \cdot Z_{WI}(b'_{(i,j)} + 1, j-1) \\ (3) \text{ if } isCovered(G,i), \text{ and } isCovered(G,j) \\ \quad Z_{WI}(i+1, (B'_{(i,j)} - 1)) \cdot Z_{WI}(B_{(i,j)} + 1, b_{(i,j)} - 1) \cdot Z_{WI}(b'_{(i,j)} + 1, j-1) \\ (4) \text{ if } (bp_G(i+1) = 0, \text{ and } bp_G(j-1) = 0) \\ \quad B(e_{stP}(i, i+1, j-1, j)) \cdot Z_{VP}(i+1, j-1) \\ (5) \text{ if } cover(G,i) = cover(G,r) \text{ and } cover(G,j) = cover(G,r') \\ \quad \sum\limits_{\substack{i<r<\min(B'_{(i,j)}, b_{(i,j)}) \\ \max(b'_{(i,j)}, B_{(i,j)})<r'<j}} (B(e_{intP}(i,r,r',j)) \cdot Z_{VP}(r,r')) \\ (6) \quad \sum\limits_{\substack{i<r<\min(B'_{(i,j)}, j) \\ bp_G(r)=0}} Z_{WI'}(i+1, r-1) \cdot Z_{VP}(r, j-1) \cdot B(a' + 2b') \\ (7) \quad \sum\limits_{\substack{max(i, b'_{(i,j)})<r<j \\ bp_G(r)=0}} Z_{VP}(i+1, r) \cdot Z_{WI'}(r+1, j-1) \cdot B(a' + 2b') \\ (8) \quad \sum\limits_{\substack{i<r<\min(B'_{(i,j)}, j) \\ bp_G(r)=0}} Z_{WI'}(i+1, r-1) \cdot Z_{VP^R}(r, j-1) \cdot B(a' + 2b') \\ (9) \quad \sum\limits_{\substack{max(i, b'_{(i,j)})<r<j \\ bp_G(r)=0}} Z_{VP^L}(i+1, r) \cdot Z_{WI'}(r+1, j-1) \cdot B(a' + 2b') \end{cases} \tag{xii}$$

A.2 Weakly closed subregions inside pseudoloop

$Z_{WI}(i,j)$ is the partition function over all structures $R_{i,j}$ given that $[i,j]$ is weakly closed ($cover(i) = cover(j) \neq 0$), and $R_{i,j}$ is inside a pseudoloop. If $i = j$ and $bp_G(i) = 0$, $[i,j]$ is empty and $Z_{WI}(i,j) = P_{up}$ i.e. penalty for unpaired base in a pseudoloop. $Z_{WI}(i,j) = 0$, if $i > j$. Otherwise, $Z_{WI}(i,j) = 0$ ($cover(i) \neq cover(j)$, subregion not weakly closed). Z_{WI} is similar to Z_W with additional penalties for base pair, unpaired bases, and pseudoknot initiation inside a pseudoloop.

$$Z_{WI}(i,j) = \sum \begin{cases} (1) \text{ if } r.j \in G, \text{or } (bp_G(r) = 0 \text{ and } bp_G(j) = 0) \\ \quad \sum\limits_{i \leq r < j} Z_{WI}(i, r-1) \cdot Z_V(r,j) \cdot B(P_{ps}) \\ (2)\ Z_{WI}(i, j-1) \cdot B(P_{up}) \\ (3) \sum\limits_{i \leq r < j} Z_{WI}(i, r-1) \cdot Z_P(r,j) \cdot B(P_{sp} + P_{ps}) \end{cases} \tag{xiii}$$

A.3 Non-empty weakly closed subregion inside band

$Z_{WI'}(i,j)$ is the partition function over all nonempty structures $R_{i,j}$, if $[i,j]$ is weakly closed with respect to G, given that $R_{i,j}$ is inside a band. Otherwise, $Z_{WI'}(i,j) = 0$. $Z_{WI'}$ is similar to Z_{WI} with additional penalties for base pair or unpaired base in multiloop spanning a band, and pseudoknot initiation inside a multiloop.

$$Z_{WI'}(i,j) = \sum \begin{cases} (1) \text{ if } r.j \in G, \text{or } (bp_G(r) = 0 \text{ and } bp_G(j) = 0) \\ \quad \sum_{i \le r < j} Z_{WI'}(i, r-1) \cdot Z_V(r,j) \cdot B(b') \\ (2) \text{ if } bp_G(j) = 0 \\ \quad Z_{WI'}(i, j-1) \cdot B(c') \\ (3) \sum_{i \le r < j} Z_{WI'}(i, r-1) \cdot Z_P(r,j) \cdot B(P_{sm} + b') \end{cases} \tag{xiv}$$

A.4 i.j close a loop

$Z_V(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed or empty and $i.j$. Otherwise $Z_V(i,j) = 0$. This subfunction and $Z_{VBI}(i,j)$ to follow are unchanged from Mathews et al. [13] pseudoknot-free algorithm i.e. penalties forhairpin loop, stacked base pairs, internal/bulge loop, or multiloop.

A.4.1 $i.j$ close an internal/bulge loop

$Z_{VBI}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed or empty and $i.j$ closes a bulge or internal loop. Otherwise $Z_{VBI}(i,j) = 0$ [13].

A.5 $[i,i'] \cup [bp(i'), bp(i)]$ band region

$Z_{BE}(i,i')$ is the partition function over the band $[i',i] \cup [bp_G(i), bp_G(i')]$, if $i \le i' < bp_G(i') \le bp_G(i)$; otherwise $Z_{BE}(i,i') = 0$. In Case (1), $bp(i+1).bp(i) - 1$ form a stacked pair in G. In Case (2), $i.bp(i)$ and $l.bp(l)$ in G close an internal loop. In Case (3), $[i+1, l-1]$ and $[bp(l)+1, bp(i)-1]$ are both weakly closed non-empty region. In Case (4), $[i+1, l-1]$ is a weakly closed non-empty region and $[bp(l)+1, bp(i)-1]$ is empty. In Case (5), $[bp(l)+1, bp(i)-1]$ is a weakly closed non-empty region and $[i+1, l-1]$ is empty (cf. Fig. 13). For base case, $Z_{BE}(i,i) = 0$ if $i < bp_G(i)$.

$$
Z_{BE}(i,j) = \sum \begin{cases}
(1) \text{ if } bp_G(i+1) = bp_G(i) - 1 \\
\quad B(e_{stP}(i, bp_G(i))) \cdot Z_{BE}(i+1, i') \\
(2) \text{ if } bp_G(l) > 0,\ \text{empty}(G, [i+1, l-1]), \\
\qquad \text{and empty}(G, [bp_G(l)+1, bp_G(i)-1]), \\
\qquad i < l \leq i',\ \text{and } (bp_G(i') \leq bp_G(l) < bp_G(i)) \\
\quad B(e_{intP}(i, l, bp_G(l), bp_G(i))) \cdot Z_{BE}(l, i') \\
(3) \text{ if } bp_G(l) > 0,\ \text{weakly closed}(G, [i+1, l-1]), \\
\qquad \text{and weakly closed}(G, [bp_G(l)+1, bp_G(i)-1]), \\
\qquad i < l \leq i',\ \text{and } bp_G(i') \leq bp_G(l) < bp_G(i) \\
\quad Z_{WI'}(i+1, l-1) \cdot Z_{BE}(l, i') \cdot Z_{WI'}(bp_G(l)+1, bp_G(i)-1) \cdot B(a' + 3b') \\
(4) \text{ if } bp_G(l) > 0,\ \text{weakly closed}(G, [i+1, l-1]), \\
\qquad \text{and empty}(G, [bp_G(l)+1, bp_G(i)-1]), \\
\qquad i < l \leq i',\ \text{and } bp_G(i') \leq bp_G(l) < bp_G(i) \\
\quad Z_{WI'}(i+1, l-1) \cdot Z_{BE}(l, i') \cdot B(a' + 2b' + c'(bp_G(i) - bp_G(l) + 1)) \\
(5) \text{ if } bp_G(l) > 0,\ \text{weakly closed}(G, [bp_G(l)+1, bp_G(i)-1]), \\
\qquad \text{and empty}(G, [i+1, l-1]), \\
\qquad i < l \leq i',\ \text{and } bp_G(i') \leq bp_G(l) < bp_G(i) \\
\quad Z_{BE}(l, i') \cdot Z_{WI'}(bp_G(l)+1, bp_G(i)-1) \cdot B(a' + 2b' + c'(l - i - 1))
\end{cases} \tag{xv}
$$

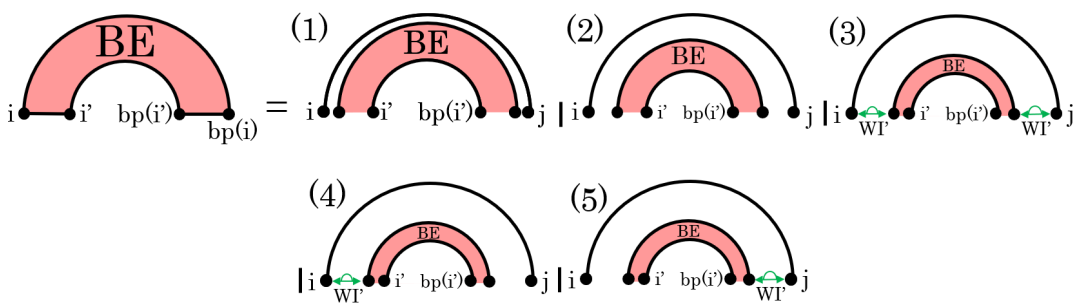

Figure 13 Cases of BE. (1) stacked pair in G; (2) internal loop; (3) initial and terminal weakly closed non-empty regions. (4) initial weakly closed non-empty region and terminal loop. (5) initial loop and terminal weakly closed non-empty region.