Algorithms for prediction of RNA secondary structure: coronavirus pseudoknots via Shapify & CParty

by

Luke Trinity
B.A., University of Vermont, 2018
M.Sc., University of Vermont, 2019

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Computer Science

University of Victoria

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Algorithms for prediction of RNA secondary structure:
coronavirus pseudoknots via Shapify & CParty

by

Luke Trinity
B.A., University of Vermont, 2018
M.Sc., University of Vermont, 2019

Supervisory Committee

Dr. H. Jabbari, Supervisor
(Department of Computing Sciences)

Dr. U. Stege, Co-Supervisor
(Department of Computer Science)

Dr. I. Numanagić, Departmental Member
(Department of Computer Science)

Dr. G. Owens, Outside Member
(Department of Biology)

Supervisory Committee

Dr. H. Jabbari, Supervisor
(Department of Computing Sciences)

Dr. U. Stege, Co-Supervisor
(Department of Computer Science)

Dr. I. Numanagić, Departmental Member
(Department of Computer Science)

Dr. G. Owens, Outside Member
(Department of Biology)

ABSTRACT

RNA molecules play a vital role in cellular processes, and many possess functional structures. Due to the complex nature of experimental methods to detect RNA structure, computational tools to predict RNA structure formation are invaluable for building comprehensive knowledge. We seek to predict RNA structure algorithmically, with a focus on the following concepts from the literature: (1) Minimum Free Energy (MFE) methods, (2) the hierarchical folding hypothesis, and (3) partition function ensemble approaches. The MFE framework is an RNA folding hypothesis stating that each RNA molecule folds into the structure with the minimum free energy. In conjunction with MFE, we employ the biologically motivated hierarchical folding hypothesis, stating that an RNA molecule will first fold once (initial fold), before a subsequent folding may occur that lowers the structure's free energy. The accuracy of MFE and hierarchical folding methods can be improved by effective incorporation of known RNA structure information such as experimental reactivity data. We introduce *Shapify*, an algorithm incorporating experimental data within hierarchical RNA folding prediction. Shapify receives SHAPE data as input to guide RNA structure prediction, allowing the unification of multiple experimental results to determine structure-function patterns. The time complexity of Shapify is $O(N^3)$ time, where N is the RNA sequence length, enabling faster prediction compared with other methods that also handle a complex RNA structure class.

We then consider the partition function model, based on the MFE approach, where we compute the sum of free energies for each possible RNA structure in the ensemble at equilibrium. The likelihood of any particular RNA structure occurring can then be determined based on the energy of the structure itself relative to the total energy in the system. Currently, partition function methods are restricted to predicting a limited set of RNA structures, because existing algorithms that allow complex RNA structures are too slow, at best $O(N^5)$ time complexity. We introduce *CParty*, an $O(N^3)$ time complexity partition function algorithm that includes complex RNA structures in the ensemble. The development of CParty's recursive decomposition schemes was non-trivial to integrate within the algorithmic implementation. By providing an input structure to algorithm CParty, we compute a 'conditional' partition function, enabling probabilistic calculation that advances understanding of RNA structure formation patterns.

In this dissertation, we (1) incorporate partial RNA structure information into

hierarchical secondary structure prediction via Shapify to understand important secondary structure motifs affecting viral function, (2) design and implement CParty, a conditional partition function algorithm to handle complex RNA structures, and (3) apply these and other related algorithms to provide RNA structural information for COVID-19 therapeutic targets. Here, we pinpoint key secondary structure folding motifs in our quest to predict functional RNA structures. Our hierarchical folding algorithms push the frontier of prediction accuracy for functional RNA secondary structures, contributing to coronavirus treatments.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

To each and every person who helped me to complete this dissertation, thank you. I want to extend my sincere gratitude to my COBRA lab mates, it has been amazing discussing bioinformatics with you each week. To my supervisor, Hosna Jabbari, thank you for bringing me to Canada (BC specifically) and pushing and motivating me to write the best thesis humanly possible. This has been an exciting dive into a new subject area for me, I look forward to continuing to learn from you and contribute to inspiring research. To my supervisor, Ulrike Stege, thank you so much for your contributions to each and every aspect of this work. Your feedback has improved the quality immeasurably. To my committee members, I truly appreciate your time reviewing and improving our research from different perspectives. To my family, especially my parents, thank you for always being there for me. I would be nowhere without you! Finally, to my friends and my girlfriend, thank you for continuing to love me and monitor my sanity over the years even as it becomes for difficult.

DEDICATION

Sometimes it's a little better to travel than to arrive.
Robert M. Pirsig
Zen and the Art of Motorcycle Maintenance: An Inquiry Into Values

Chapter 1

Introduction

We introduce relevant background information, selected topics in RNA structure prediction, goals and contributions for this dissertation, and provide an outline for the remaining chapters.

1.1 Background

Technological advances continue to improve research into the inner workings of the human body. The discovery of the cell in the 17th century began a long journey that is now culminating in an ongoing realization of the importance of DNA and RNA in enabling cellular function. However, there remain a multitude of unknowns precluding a comprehensive understanding for DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). RNA, the single stranded counterpart to the double helix of DNA, is most commonly thought of as a mechanism to transfer genetic information from DNA into proteins [16]. Research continues to describe the functions of RNAs that do not encode proteins, e.g. non-coding, ribosomal, and transfer RNAs, which also perform essential functions at different cellular levels, e.g. gene translation, expression, regulation, and coding for proteins [17, 18, 19, 20, 21]. Knowledge of functional RNA families continues to increase. The RFAM database [22] includes RNA structure information for 4108 RNA families as of the latest release in November, 2022.

RNA is composed of four bases: adenine (A), uracil (U), cytosine (C), and guanine (G). A critical aspect of understanding RNA is identifying its structure, which contributes to various functional properties. Characterizing RNA structure can aid in the design of RNA molecules for applications in biotechnology [23]. Due to the time

and financial costs of experiments determining RNA structure [24], computational approaches to predict the structure of RNA will be invaluable for the foreseeable future.

In brief, the function of RNA is primarily dependent on its structure, which is determined based on the sequence of nucleotide bases (e.g., 'AUCG...') and how the bases interact with each other (base pairing) or other molecules. The wide array of functions that RNA can perform motivates continued efforts to model it. Due to chemical and energetic constraints, RNA is rarely found as one single strand. Following the harmonious order of our universe, also known as the second law of thermodynamics, RNA demands stability. To achieve stability, RNA folds back on itself, forming structures that lower its free energy. The three-dimensional conformations that result from RNA folding regulate molecular mechanisms in the cell [25, 26]. Despite advances, many facets of RNA structure are not yet fully understood for a variety of reasons, e.g., due to a limited number of RNA-template structures to compare against [27].

Given the relative youth of focused RNA research, it is not surprising that comprehensive information about RNA structure is unavailable at this time. A complete understanding is a lofty goal since over 75% of the human genome is transcribed into RNA (over four million bases) [28, 29]. Multiple confounding factors limit our understanding of RNA structure, including its interplay with other molecules in the cell, as well as the ability for a single RNA molecule to fold into different structures.

Computational RNA structure prediction algorithms determine the set of base pairs formed by an RNA molecule folding (cf. Section 2). Under the minimum free energy (MFE) paradigm, the RNA structure with the lowest free energy is classified as the functional structure because there is evidence that RNA molecules usually fold to their MFE structure [30]. This approach is derived from thermodynamic principles: when a system is at equilibrium, it is considered to be globally optimal. Therefore, to compute the free energy of an RNA molecule, each sequence of unpaired bases within the structure (i.e., a *loop*) is assigned a free energy value based on the underlying energy model. The energy of a structure is calculated as the sum of the loop energies. In cases where multiple sequences have a common ancestor, i.e., when they are homologous, comparative algorithmic methods can be used to predict RNA secondary structures provided the sequence alignment is well-justified and other confounding variables are accounted for [31].

Here we focus on thermodynamic MFE methods, although non-MFE energy-based approaches like the heuristics described in [5] or maximum expected accuracy methods

such as in [32] do exist and can be highly useful in the right context. In reality, RNA molecules do not always reach global optimum; *kinetic traps* can occur when RNA folds into a locally optimal structure requiring a high energy barrier to escape [33]. Furthermore, the ability of RNA to fold into a multitude of possible structures, i.e. conformational plasticity, has been linked to its very function [34]. Based on these factors, RNA structure-function research should account for 'suboptimal' structures, especially those that are energetically favourable despite not being globally optimal.

Methods to identify RNA structure in living cells (*in vivo*) or cells outside an organism (*in vitro*) include probing RNA molecules for reactivity [35, 36], cryo-electron microscopy (cryo-EM) [12], and crystallography [37] experiments. Structural reactivity (SHAPE-MaP, i.e., selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling) [38] probing can provide single base resolution structural data for entire RNA transcriptomes, while cryo-EM and crystallography are more localized and have solved specific RNA structures accurately to less than 10 angstroms (less than 0.000001 millimeters). Despite these advances, new RNA folding challenges continue to emerge. An ensemble of possible RNA structures [39] along with inherently difficult challenges in crystallization and 3-D modeling [40] (time consuming, complex, costly) makes computational methods (*in silico*) critically necessary to simulate and predict RNA structure formation.

Tertiary structure prediction, which determines important long-range interactions defined by atomic coordinates in 3-D physics-based modeling [41, 42, 43, 44, 45, 46], is much more challenging than secondary structure prediction because it allows each base to appear more than once in predicted structures (e.g. *three* bases pairing together, $\{(A, B), (B, C)\}$). In my dissertation I focus on secondary structure prediction, where each base of an RNA sequence can appear in at most one predicted base pair of a structure. Note that secondary structure prediction efforts can also contribute to subsequent tertiary structure prediction [30], when experiments or simulations take into account the most stable secondary structures shaping the complete landscape of possible RNA structure formation.

1.2 RNA Secondary Structure Prediction

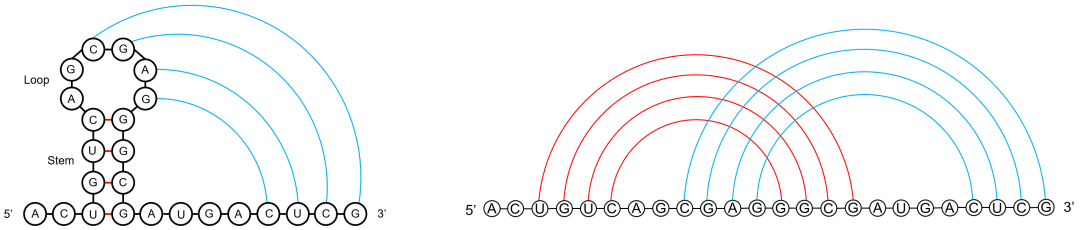
The computational problem of RNA secondary structure prediction can be defined as follows: given an input RNA sequence, determine a set of paired bases from the sequence such that each base appears in at most one pair and the free energy of the

structure is minimized [7]. RNA folding refers to the process by which RNA acquires its structure through interacting nucleotide bases (also referred to as *bases*). While realistically predicting the full RNA tertiary structure (or 3D structure) is ultimately desirable, prediction of the secondary structure (i.e. , the determination of the desired set of all base pairs) appears easier and sheds light on the tertiary structure. For a recent review of methods for RNA tertiary structure prediction, we refer to Li et al. [47]. The majority of existing computational methods for the prediction of RNA structure, therefore, focus on the prediction of secondary structure. We represent an RNA molecule by its sequence S of length n. We refer to bases by their position in S indexed from 1 to n from 5' (left) to 3' (right) end. Recall an RNA sequence is made up of four bases: adenine (A), cytosine (C), guanine (G), and uracil (U). When an RNA structure forms, complementary bases pair together and form hydrogen bonds. 'A' pairs with 'U' and 'G' pairs with either 'C' or 'U'–referred to as *canonical base pairs*. A *base pair* is then defined as the pairing of two bases i and j where $1 \leq i < j \leq n$, and represented as $i.j$. Consecutive base pairs are referred to as a *stem*. Recall that within a secondary structure approach each base can pair with at most one other base (i.e., no base triplets are allowed).

We say base pairs $i.j$ and $i'.j'$ are *nested* if $1 \leq i < i' < j' < j \leq n$, and *disjoint* if $1 \leq i < j < i' < j' \leq n$. An RNA structure with only nested and disjoint base pairs is referred to as a *pseudoknot-free* structure.

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*. We call structures with crossing base pairs, *pseudoknotted*. RNA folding can be described sequentially, with the *initial* stems forming first (Fig 1.1, red lines) followed by additional stems (Fig 1.1, blue lines). Formally, an RNA structure is considered *pseudoknotted* when at least two of its base pairs, $i.j$ and $i'.j'$ cross: $1 \leq i < i' < j < j' \leq n$, in which case both $i.j$ and $i'.j'$ are considered pseudoknotted base pairs. We note that, while pseudoknotted base pairs are sometimes considered as part of the tertiary structure, we consider them as part of the secondary structure.

The hierarchical folding hypothesis posits that RNA first folds into a simple non-crossing (pseudoknot-free) structure, then additional (possibly crossing) bases pair to lower the free energy of the structure [30]. RNA structure analysis finds that the initial structure may be modified locally to accommodate formation of more stable base pairs, specifically, non-crossing structure can be followed by additional crossing base pairs [48, 49]. In particular, hierarchical folding was experimentally identified in



(a) (b)

Figure 1.1: Pseudoknot Diagram. (a) *Hairpin* loop with stem base pairs shown in red. Potential *H-type* pseudoknot pairing shown in blue between bases in the loop and complementary downstream bases. (b) Representation of 1.1a as an arc diagram for display of overlapping structures.

formation of pseudoknots [50] including functionally important pseudoknots (frameshift stimulating structures) [51]. 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 that the most stable RNA structure is the one with the lowest free energy, algorithms like HFold [52] or Iterative HFold [3] (cf. Figure 1.2) find the MFE structure for a given sequence S, and a pseudoknot-free input structure G, where each RNA loop is assigned an energy value. The free energy of an RNA structure is calculated as the sum of the energies of its loops. Free energy of some loops were experimentally determined, and others were extrapolated based on the experimental values [53, 39, 4].

Both the HFold [3] and Iterative HFold [52] algorithms are MFE methods that allow pseudoknots. While HFold adheres strictly to the hierarchical folding hypothesis, Iterative HFold allows for minimal modification in the input structure to accommodate formation of base pairs that can lower the energy of the structure. Iterative HFold utilizes four biologically sound methods and outputs the MFE structure among these four predictions as the final structure. These four distinct methods explore different local structural modifications and allow for identification of *suboptimal* structures– structures with free energy only slightly higher than that of the MFE structure. We utilize all of these four structures (as opposed to just the one with the MFE) to obtain a glimpse of the suboptimal structure landscape.

Algorithms for prediction of RNA secondary structure found in the literature can be divided into two categories: algorithms that predict structures that can include pseudoknots, and algorithms that only predict structures that are pseudoknot-free (no

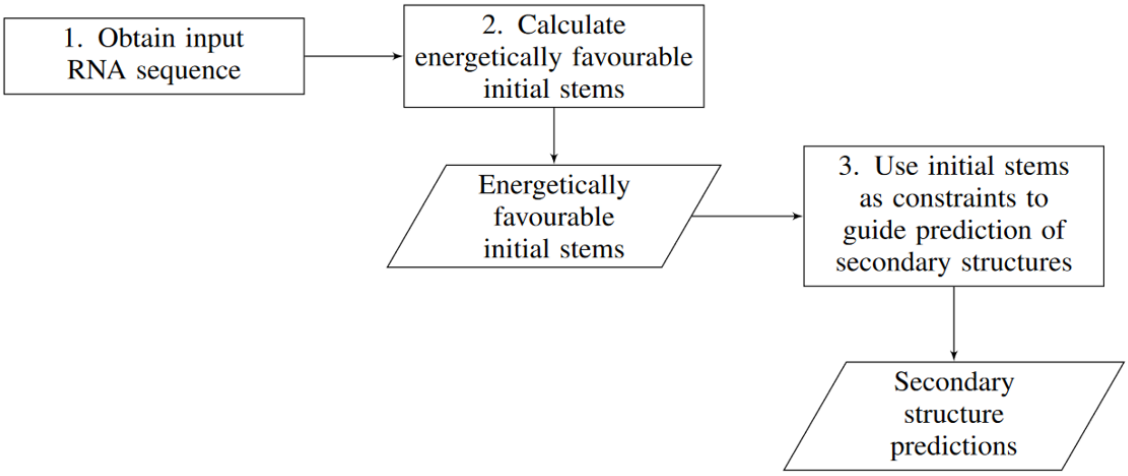


Figure 1.2: Hierarchical Folding Pipeline. Rectangles dictate actions, parallelograms denote input/output.

crossing base pairs). Progress towards computational prediction of pseudoknot-free secondary structures has been remarkable [54]. However, there are many functionally important pseudoknotted structures, especially those in viral RNA [55] which we focus on in this dissertation. Although pseudoknots are ubiquitous in RNA [56, 57, 58], prediction of the MFE pseudoknotted RNA structure, given an input RNA sequence, is computationally intractable (i.e., NP-hard) and cannot be approximated [59, 60, 61]. Therefore, many algorithmic approaches focus on pseudoknot-free structures. Algorithms for pseudoknotted structure prediction, like Shapify and CParty, consider specific classes of pseudoknots that are a subset of the general pseudoknot structure class. Specifically, current polynomial-time MFE-based pseudoknotted structure prediction algorithms [59, 62, 63, 64] find the MFE structure for a given input sequence, from a restricted class of structures defined by allowable patterns of crossing base pairs.

Two widely used thermodynamic approaches for predicting non-crossing (i.e. pseudoknot-free) RNA secondary structure are the ones by Zuker [65] and McCaskill [66]. The Zuker algorithm uses dynamic programming (DP) to recursively compute the free energy of all possible RNA substructures and find the MFE structure [65]. To further improve their prediction accuracy, MFE-based prediction methods can be extended to integrate experimental (e.g., SHAPE) data as soft-constraints [67].

Adopting a stochastic view of RNA, the McCaskill algorithm computes the *partition function* that describes the pseudoknot-free 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, cf. Section 2.6). The partition function is useful for predicting the relative probability of any structure as well as base pair probabilities [66, 68]. 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.

1.3 Objectives

In this dissertation we introduce, develop, implement and apply novel thermodynamic-based algorithmic approaches to efficiently predict complex functional RNA structures. RNA secondary structure is intrinsically linked to tertiary long range interactions. Overall, the accuracy of secondary structure prediction can be improved by incorporating experimental data [39]. Yet, there is no available secondary structure prediction algorithm to include such data as a soft-constraint within a hierarchical folding approach. We seek to better unite experimental datasets with the secondary structure prediction algorithms themselves. Motivated by this limitation, we introduce the Shapify algorithm (cf. Chapter 3), which uses reactivity data as input to guide pseudoknotted secondary structure prediction. Shapify is based on the efficient hierarchical folding algorithm design of Iterative HFold [52], which handles a restricted class of pseudoknotted structures that is quite broad, allowing arbitrary depth and length of pseudoknotted base pair 'chains'. Effectively tuning the Shapify method to incorporate fixed or guide input constraints will contribute to understanding of how RNA folds into a stable pseudoknotted structure.

Within the field of RNA secondary structure prediction, we seek to more comprehensively investigate the landscape of possible structures than existing methods currently do. It is well known that a partition function for pseudoknot-free structures can provide the base pairing probability for each position in a structure relative to the input constraint. There are algorithms that compute a partition function in $O(N^3)$ time for pseudoknot-free secondary structure prediction, but nothing better than $O(N^5)$ time for a pseudoknot-allowed partition function method [59]. Here we introduce a novel partition function algorithm, CParty, which computes the partition function in $O(N^3)$ time for a restricted class of pseudoknots conditioned on a given input structure.

With our novel algorithms, Shapify and CParty, the goal is to improve secondary structure pseudoknot prediction capabilities compared with the best available methods, ShapeKnots [1] and RNAFold [69]. We demonstrate how our publicly available software implementations can be used to predict most stable secondary structures and investigate RNA folding. By applying our algorithms to predict secondary structures for the SARS-CoV-2 ribosomal frameshifting pseudoknot, we contribute to potential viral therapeutics for one of the most impactful diseases of our time, COVID-19. Beyond this specific case study, our methodology can be used as a roadmap for future work in design, implementation, and application of algorithms to predict complex RNA secondary structure folding dynamics.

1.4 Contributions

The contributions of this thesis are as follows:

1. We proposed the Shapify algorithm, which uses the relaxed hierarchical hypothesis with $O(N^3)$ time and $O(N^2)$ space complexity. Shapify takes as input a pseudoknot-free structure, and a SHAPE reactivity experimental dataset, to compute a possibly pseudoknotted output structure whose energy is at least as low as that of any (density-2) pseudoknotted structure containing the input structure (cf. Chapter 3).

2. We completed the Shapify algorithm implementation by employing cross validation to determine optimal parameters. To accomplish this we aggregated a novel database of RNAs with known structure and associated SHAPE data (cf. Chapter 3).

3. We applied Shapify to predict energetically favourable secondary structures for coronavirus frameshift stimulating pseudoknots (cf. Chapter 3) towards viral therapeutics.

4. We further applied Shapify to extended SARS-CoV-2 RNA sequence lengths surrounding the frameshift site to provide information about how the moving of the ribosome changes secondary structure motifs which may regulate the frameshift event (cf. Chapter 4).

5. We utilized sequence covariation methods (RNAz [70] and KnotAli [71]) to obtain consensus structure predictions for coronaviruses frameshift pseudoknot site. We discuss implications of these predictions for structure-function investigations (cf. Chapters 3 & 4).

6. We proposed the CParty algorithm, which computes the conditional partition function in $O(N^3)$ time for a restricted but general class of pseudoknotted structures (density-2). CParty takes as input a pseudoknot-free structure to compute the conditional partition function, representing two orders of magnitude time complexity improvement over previous pseudoknotted partition function methods (cf. Chapter 5).

7. We implemented and applied the CParty algorithm to identify ensemble energies for the SARS-CoV-2 frameshift pseudoknot. Here we compare our results with the RNAFold [69] partition function, and discuss future applications of this novel algorithm towards coronaviruses structure-function prediction (cf. Chapter 5).

The research in this dissertation is beneficial to the extended RNA and biomolecular community. By providing the efficient Shapify and CParty algorithms, it is now possible to obtain more energetically favourable predictions for pseudoknotted RNAs and associated ensembles. Here we have improved the accuracy of structure prediction, which is critical in identifying RNA function. Shapify incorporates experimental reactivity data more effectively than previous methods, and CParty is the first $O(N^3)$ time algorithm to compute the conditional partition function for pseudoknotted RNA.

1.5 Outline

Chapter 1 begins with background, including higher-level context before more detailed review of RNA structure prediction, goals, and contributions of this dissertation.

Chapter 2 describes important background knowledge related to predicting RNA secondary structure: definitions, tiers of RNA structure, how RNA secondary structure is visualized, dynamic programming and foundational algorithmic approaches, loop decomposition and energy models, suboptimal structures, and stochastic sampling.

Chapter 3 This is the first results chapter, including introduction of the Shapify hierarchical folding algorithm and application of Shapify to identify paths to the SARS-CoV-2 frameshifting pseudoknot at length 68 nt. Structural similarity between coronaviruses is identified and discussed.

Chapter 4 In the second results chapter we present extended evolutionary analysis within a relaxed hierarchical folding framework. We apply Shapify to extended length SARS-CoV-2 frameshift element sequences, 'tying the knot' to unveil coronavirus frameshift pseudoknot motifs.

Chapter 5 The third and final results chapter presents CParty, the first conditional partition function algorithm for density-2 RNA pseudoknots. We validate and apply CParty to describe a critical secondary structure transition of the SARS-CoV-2 frameshift element.

Chapter 6 Conclusions, and future work for RNA secondary structure hierarchical folding algorithms and their applications.

Chapter 2

RNA Secondary Structure

In this chapter we provide background knowledge which will be beneficial for understanding the key RNA secondary structure results in this dissertation. The following topics are covered: definitions, tiers of RNA structure, visualization of RNA secondary structure, earliest algorithms for prediction of RNA secondary structure, energy models, loop decomposition, suboptimal structures, and stochastic sampling.

2.1 Definitions

- **RNA molecule**: sequence S of length n
- **RNA folding**: the process by which RNA acquires its structure through interacting nucleotide bases
- **Base pair**: the pairing of two bases i and j where $1 \leq i < j \leq n$, represented as $i.j$
- **RNA secondary structure prediction**: for an input RNA sequence, determine a set of paired bases such that each base appears in at most one pair
- **RNA structure**, R: a set of base pairs for an RNA sequence over alphabet {A, U, C, G}
- $bp_R(i)$: the index of the base paired with base at index i in structure R
- **Region** $[i, j]$: a sequence of indices between i, the left border, and j, the right border, inclusive

- **Weakly closed region** $[i, j]$: $\forall k \in [i, j]$ either $bp_R(k) \in [i, j]$ or $bp_R(k) = 0$ (unpaired)
- **Closed region** $[i, j]$: $[i, j]$ is weakly closed, and $\forall l \in [i, j-1]$, neither $[i, l]$ nor $[l+1, j]$ is weakly closed [72]
- **Base pair** $i.j$ **covers** k: if $i < k < j$ and $\nexists i'.j'$ such that $i < i' < k < j' < j$
- **Stem**: consecutive base pairs
- **Loop**: sequence of unpaired bases closed by a base pair
- **Nested base pairs** $i.j$ **and** $i'.j'$: if $\exists i, i', j, j'$, such that $i.j$, $i'.j'$, and $1 \leq i < i' < j' < j \leq n$
- **Disjoint base pairs** $i.j$ **and** $i'.j'$: if $\exists i, i', j, j'$, such that $i.j$, $i'.j'$, and $1 \leq i < j < i' < j' \leq n$
- **Disjoint structures**: two structures are disjoint if and only if there is no base that is present in both structures (recall each structure is a set of base pairs). For example, if G and G' are disjoint and $i.j \in G$, then neither i nor j can be paired in G'. Similarly, if G and G' are disjoint and $h.l$ in G', then neither h nor l can be paired in G.
- **Pseudoknot-free structure**: an RNA structure with only nested and disjoint base pairs
- $i.j$ **crosses** $i'.j'$: if $\exists i, i', j, j'$, such that $i.j$, $i'.j'$, and $i < i' < j < j'$
- **Pseudoknotted structure**: an RNA structure where $i.j$ and $i'.j'$ cross
- **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 region**: consider a maximal chain of $\preceq$, if $i.j$ is the outermost pair, and $i'.j'$ is the innermost pair, then $[i, i'] \cup [j'.j]$ is the band region
- **Associated with closed region**: Bands, 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 2.1)

- **Pseudoloop**: the unpaired bases and base pairs associated with pseudoknotted closed region $[i, j]$, together with the closing base pairs of the bands associated with $[i, j]$
- $\#B(L, k)$: 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 $i_1 \leq k \leq j_1$. Then $\#B(L, k)$ is the number of bands associated with L that cross k.
- **Pseudoloop density**: Density of a pseudoloop L is: $max_{i \leq k \leq j}(\#B(L, k))$
- **RNA structure density**: density of a structure, R, is the maximum density of L over all *pseudoloops* L of R. R is *density-k* if the density of R is at most k.

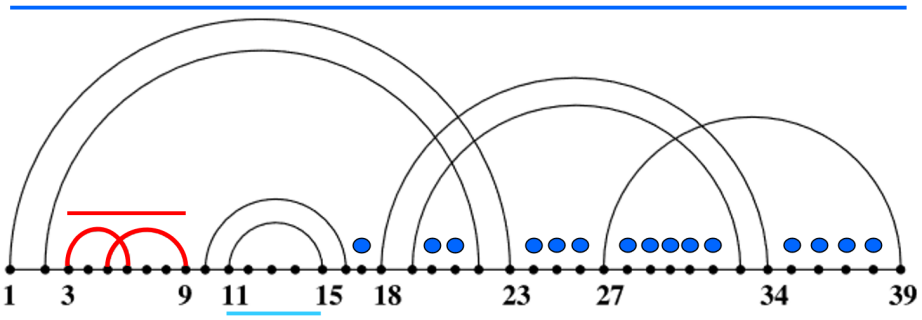


Figure 2.1: The three *bands* $[1, 2] \cup [22, 23]$, $[18, 19] \cup [33, 34]$ and $[27, 27] \cup [39, 39]$ as well as unpaired bases not in closed subregions (c.f. blue dots •) are associated with closed pseudoknotted region $[1, 39]$, i.e., density-2 pseudoloop $[1, 39]$. Closed regions $[3, 9]$ and $[10, 16]$ are associated with pseudoloop $[1, 39]$ but closed region $[11, 15]$ is *not*. Bands $[3, 3] \cup [6, 6]$ and $[5, 5] \cup [9, 9]$ are associated with closed pseudoknotted region $[3, 9]$, i.e., density-2 pseudoloop $[3, 9]$. Note that bases $12 - 14$ are covered by 11.15 but not 10.16. Figure modified from [3].

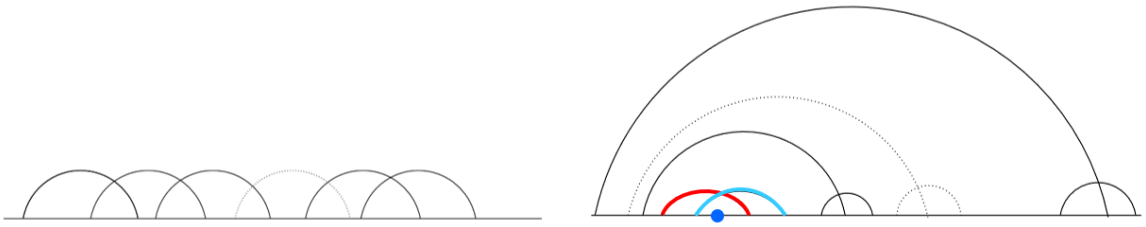


Figure 2.2: Arc diagram representation of two density-2 structures, each structure contains an arbitrary number and depth of bands [3]. Blue dot covered by red and light blue bands indicates how association with closed region maintains the density-2 property. Figure modified from [3].

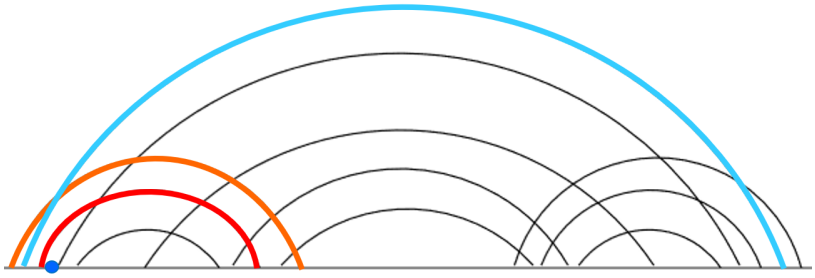


Figure 2.3: Example of a *bi-secondary structure* that is not a density-2 structure [3]. Blue dot covered by red, orange, and light blue bands indicates density-2 property not satisfied. Due to the crossing pattern, the bands cannot be efficiently separated into nested pseudoloops. As a result, the blue dot is not associated with a closed subregion, this increases the density of the structure above 2. Figure modified from [3].

- ***G***: a pseudoknot-free structure, (e.g. the input to our hierarchical folding algorithm Shapify)
- ***G'***: a pseudoknot-free structure such that G and G' are disjoint and $G \cup G'$ is density-2
- **Minimum free energy framework**: RNA folding hypothesis stating each molecule folds into the structure with minimum free energy
- **Suboptimal structure**: RNA structure that is not the minimum free energy structure
- **Hierarchical folding hypothesis**: RNA first folds into a non-crossing structure, than additional bases pair to lower the free energy of the structure, possibly forming pseudoknots
- **Standard energy model for pseudoknot-free structure** [62, 53]: $E(s) = \sum_{L \in s} F_L$ where $E(s)$ is the total free energy of structure s, F_L is the free energy of a loop that is associated with each loop L in structure s
- **Structure class**: a category of RNA secondary structures based on common features like topological complexity
- **Partition function**: $Z = \sum_{s \in S} e^{E(s)/(KT)}$ where S is the set of all possible secondary structures, $E(s)$ is the free energy of structure s, K is the universal gas constant, at a given temperature T

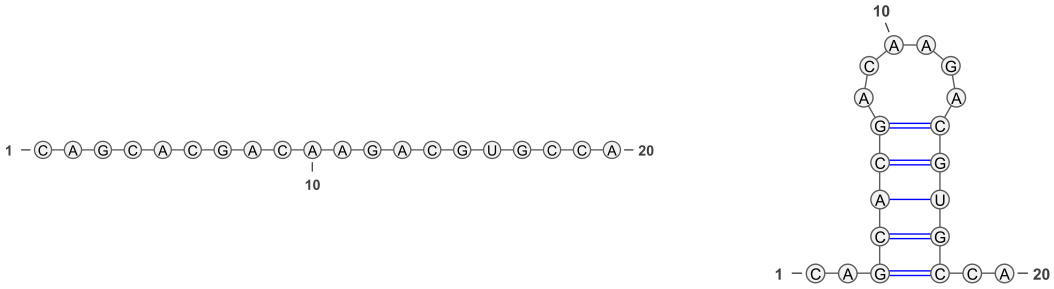
- **Conditional partition function**: $Z_W(i,j)$ denotes the conditional partition function given input G over all structures $G_{i,j} \cup G'_{i,j}$ for the subsequence $s_i \dots s_j$ taken over all choices of $G'_{i,j}$ (which is pseudoknot-free, disjoint from $G_{i,j}$, and such that $G_{i,j} \cup G'_{i,j}$ is density-2) and i and j are not covered by a base pair in G
- **Unambiguous decomposition scheme**: a method for decomposing a structure into its components where each individual substructure is considered exactly once
- **in vivo**: refers to experiments within a living organism
- **in vitro**: refers to experiments conducted outside the living organism, typically in a laboratory environment such as a culture dish
- **in silico**: process modeling via computational methods
- **SHAPE data**: Selective 2′ Hydroxyl Acylation analyzed by Primer Extension (SHAPE) is a chemical modification technique used to investigate the secondary structure of RNA molecules. Floating-point numbers, typically in the 0 to 1 range, reflect the degree of reactivity for each base in the sequence. A higher value suggests an unpaired base, while a lower value indicates a structure or paired base. Note that the data must be normalized before being introduced as pseudo-energies in the thermodynamic energy model. The normalization process typically involves scaling and transforming the raw SHAPE reactivity values to help correct for variations in experimental conditions, e.g., reagent concentrations.
- **Slippery sequence**: RNA sequence where the ribosome is prone to slipping into a different reading frame (e.g., in coronaviruses typically a heptanucleotide sequence 'UUUAAAC')
- **Native frameshift structure**: expected structure that contributes to the programmed ribosomal frameshift (e.g., in SARS-CoV-2 the native structure is the experimentally identified three-stemmed pseudoknot)
- **Conformational plasticity**: ability of an RNA molecule to form different configurations or structures
- **Non-native structures**: RNA conformations that differ from the native-type

- **Frameshift efficiency**: the frameshift rate, determining the ratio of viral proteins
- **RNA secondary structure motif**: recurring, conserved pattern of interactions within the secondary structure of RNA molecules
- **Dual graph**: RNA dual graphs are mathematical representations used to describe the secondary structure of RNA molecules by representing stems as vertices and loops as edges

2.2 Tiers of RNA Structure

As mentioned, RNA is normally found in a single strand, and consists of four nucleotides: A, G, C, and U (uracil is substituted for thymine in DNA). Here we define the hierarchy of structure 'levels' in RNA:

- **RNA Primary Structure**: The linear sequence of nucleotides (cf. Fig 2.4a).
- **RNA Secondary Structure**: At this level (*which is our focus for the dissertation*) we take into account hydrogen bonds between the nucleotides in the sequence. More precisely, the secondary structure for an RNA sequence is a set of base 'pairs'. Base pairs that are considered include A with U, and also C with G; the latter has a stronger connection [73] because C and G pair with three hydrogen bonds as opposed to two between A and U [74]. Other types of base pairing include 'wobble' base pairs (G with U, cf. [75]) and non-canonical base pairs [76], the latter are not considered in this work.
- **RNA Tertiary Structure**: Here, the intramolecular bonds are considered. For RNA, tertiary structure includes triple stranded bonding, i.e., 'triplets', where three bases form bonds together. To aid in understanding we provide a comparison via DNA (cf. [77]), which has a well-known tertiary structure of a double helix, its secondary structure would be a simple 'ladder'.
- **RNA Quaternary Structure**: Finally, at the highest level one must consider intermolecular bonds, i.e., bonds between different molecules (i.e., RNA-RNA, RNA-DNA, or RNA-protein).



(a) (b)

Figure 2.4: Tiers of RNA structure. (a) Primary structure: a sequence of nucleotides. (b) Secondary structure: a hairpin loop with stem base pairs shown in blue. Visualization created using VARNA [6].

2.3 Visualization of RNA Secondary Structure

We use different strategies to represent RNA secondary structures:

- **Radial**: Base pairs form a 'ladder' branching off of circular stretches of unpaired bases (cf. Fig 2.4b).
- **Linear (arc-diagram)**: RNA sequence of bases is shown from left to right (5′ to 3′) in a single horizontal line. Base pairs are represented as arcs between bases (cf. Fig 2.5).
- **Dot-bracket**: Computer-readable RNA structure is represented as a sequence of characters corresponding with the actual sequence of RNA bases. Any base that is unpaired is denoted '.', paired bases are represented by parenthesis '()', and typically, pseudoknotted base pairs are represented by square brackets '[]'.

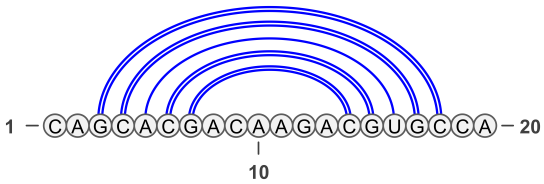


Figure 2.5: Arc diagram representation of secondary structure. This structure is analogous to the radial representation in Fig 2.4b. Visualization created using VARNA [6].

2.4 Earliest Algorithms for Prediction of RNA Secondary Structure

In order to design any algorithm to predict RNA secondary structure, we must define a scoring function. Here we utilize free energy, thus each particular structure has an associated free energy, and lower free energy corresponds with a more stable structure. Theoretically, the MFE structure is the most stable structure. The algorithm by Smith and Waterman introduced in 1978 [78, 79] used recursive decomposition with a simple energy model (no differentiation between different types of base pairing) to determine the total number of possible secondary structures. It was subsequently shown that the number of secondary structures rises exponentially with the length of the sequence, making such an exhaustive recursive method inefficient (exponential complexity). Thus, *Dynamic Programming* (DP) is used to efficiently decompose each secondary structure into substructures, by removing redundant computations from the recursive scheme. The underlying assumption of DP is 'Bellman's Principle of Optimality' [80], which states that the optimal solution of a problem must be based on subproblems that each have optimal solutions. DP defines a relationship where the optimal solution to a subproblem is equal to the combination of subproblems. The Nussinov algorithm [81] combines recursive decomposition with DP to find the maximum number of possible base pairs, where each base pair has the same weight, 1.

In the recurrence relation for the Nussinov algorithm, $\mathrm{N}[i,j]$ represents the maximum number of base pairs in the subsequence i, j. For the base case, where $1 \leq i \leq n$, $N[i,i] = 0$ and $N[i,i-1] = 0$. The recursion for $\mathrm{N}[i,j]$ is then computed where $1 \leq i < j \leq n$:

$$\mathrm{N}[i,j] = \max \begin{cases} \mathrm{N}[i,j-1] & (\text{if } j \text{ unpaired}) \\ \max_{i \leq k < j} \mathrm{N}[i,k-1] + \mathrm{N}[k+1,j-1] + 1 & (\text{if } k \text{ and } j \text{ can pair}) \end{cases}$$

The recurrence demonstrates the optimal substructure property, building the global solution from optimal solutions to subproblems.

2.5 Energy Models

For more complex methods to predict RNA secondary structure, an energy model is needed to compute the free energy of the structure. Free energy is measured in

kcal/mol, and the most energetically favourable secondary structure is the one with MFE. In order to compute the total free energy of each structure and find the MFE structure, we sum up the contributions of each substructure, each with an associated free energy. The energy model determines the free energy value assigned to each substructure assuming the free energy of a secondary structure is additive in terms of its loops, e.g., $E(s) = \sum_{L \in s} F_L$ where $E(s)$ is the total free energy of structure s, F_L is the free energy of a loop that is associated with each loop L in structure s [62, 66].

Simple energy models may assign a value based on the energy of different types of base pairs, but more complex models will distinguish between many different specific types of loops. In general, secondary structures are decomposed into loops, and empirical results such as melting experiments [82] are used to determine energy values for each loop. Where experimental data is unavailable, energy values are extrapolated from experimental results. One of the key principles for energy model compatibility with DP, is that the free energy of each substructure is dependent only on itself and neighboring substructures. Hence the name, 'nearest neighbor' energy model, which is the style implemented in this dissertation and includes the Turner parameters or Turner energy rules which experimentally determined terms for stacking base pair energies, structure initiation energies, and various types of mismatch penalties [83]. The Zuker algorithm [84, 85, 65] was the first to use a nearest neighbor energy model to find the MFE secondary structure for an RNA sequence. This builds on the work of Nussinov which simply maximized the number of base pairs.

In the recurrence relation for the Zuker algorithm, $\mathrm{W}[i,j]$ represents the MFE of the subsequence i,j and $\mathrm{V}[i,j]$ represents the MFE of any substructure where i is paired with j. Let m be the minimum loop size, usually 3. For the base case, where $j - i \leq m$, $W[i,j] = 0$, $V[i,j] = \infty$. The recursion for $\mathrm{W}[i,j]$ is then computed where $i < j - m$, and for $\mathrm{V}[i,j]$ where $1 \leq i \leq j \leq n$:

$$\mathrm{W}[i,j] = \min \begin{cases} \mathrm{W}[i,j-1] & (\text{if } j \text{ unpaired}) \\ \min_{i \leq k < j-m}(\mathrm{W}[i,k-1] + \mathrm{V}[k,j]) & (\text{if } k \text{ and } j \text{ can pair}) \end{cases}$$

2.6 Loop Decomposition

DP recurrences can be stratified into cases, where similar elements are grouped and each handled in the same manner. For decomposition of an RNA structure into its component substructures we define elements as different types of loops. A k:loop

consists of u unpaired bases and $k-1$ base pairs [66]. Note that we exclude the closing base pair. The simplest type of loop enclosed by a base pair $i.j$ is a *hairpin* ($k=1$), this is a strand of unpaired bases with no additional base pairs, i.e., *branches*, between i and j. When $i.j$ closes two branches such that all bases between the closing base pairs are unpaired, this is an *internal* loop ($k=2$, $u>0$). A *bulge* loop is a special type of internal loop where there are no unpaired bases in one side of the loop. The highly stable *stacked pair* ($k=2$, $u=0$) can also be thought of as a type of internal loop, one with no unpaired bases in either branch, i.e., $i.j$ and $(i+1).(j-1)$. When a base pair $i.j$ contains at least three *branches*, this is referred to as a multiloop ($k>2$). Due to the complexity in multiloop decomposition, a linear approximation is used to compute the free energy of the loop: $F_L = a + b(k-1) + cu$ where a is the penalty for initiating a multiloop, b is the penalty for each additional branch, and c is the penalty for an unpaired nucleotide in the multiloop [86, 87, 66].

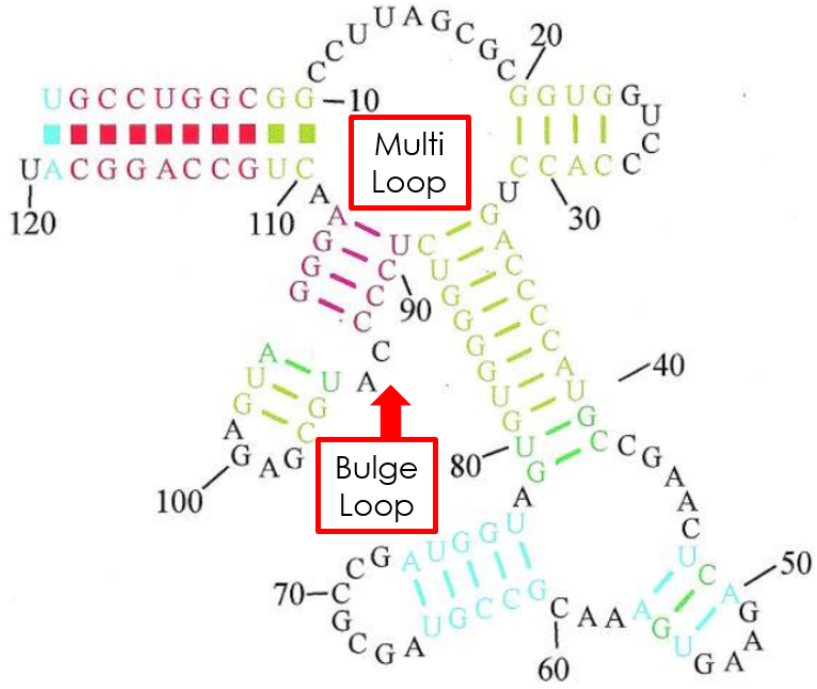


Figure 2.6: Radial representation of an RNA structure. Indicated multiloop (top red box) has four branches, the bulge loop (pointer from bottom red box) has unpaired bases on only one side of the helix as opposed to both sides which would be an internal loop. Figure modified from [7].

2.7 Suboptimal Structures

In defining an energy model and designing DP algorithms to find the MFE structure for an RNA sequence, we assume that the structure which has MFE is the most stable. However, even the best available theoretical energy model has an accuracy barrier [4], therefore the computed MFE structure is not always the structure reached in reality. In some biological applications, e.g. riboswitches [88], the same RNA sequence may fold into multiple different structures. Other molecules in the cell such as ribosomes [89] or co-transcriptional folding [90] may occlude part of the RNA sequence, limiting the ability of certain bases to pair. Based on these important empirical results, there is an interest and value in predicting structures with energy slightly higher than the MFE structure, i.e., *suboptimal structures* (cf. Fig 2.7). The best candidate choices for suboptimal structures are those that are close in free energy to the MFE structure, as they are most energetically favourable and probable to be reached. The goal of studying RNA kinetics is to understand how one secondary structure folds into another, a critical aspect of this is describing pathways through the folding landscape that include suboptimal structures.

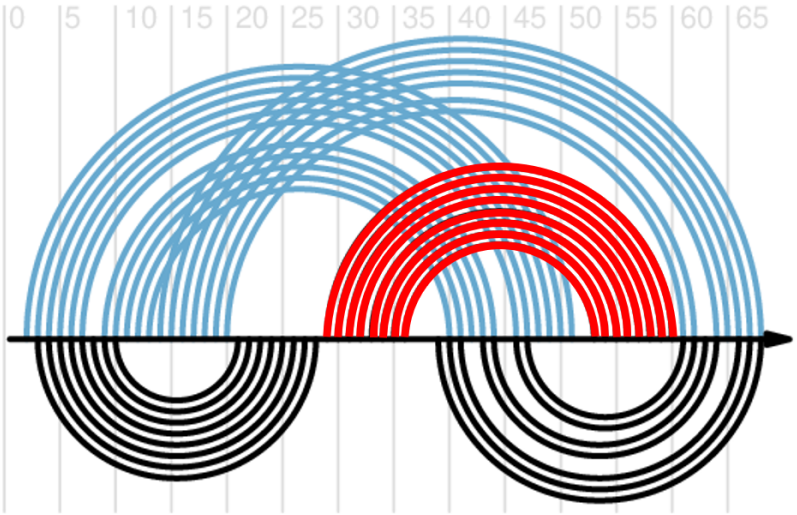


Figure 2.7: Suboptimal structures predicted for the SARS-CoV-2 RNA virus -1 programmed ribosomal frameshift (PRF) stimulating pseudoknot via Shapify. The second most favourable initial stem (in red) was used as constraint. Blue arcs combined with red arcs correspond with the MFE structure, black arcs combined with red arcs represents a suboptimal structure with only slightly higher free energy. Visualization generated using R-chie web server [8].

2.8 Stochastic Sampling

In selecting the best suboptimal structures, one can use a random, i.e., *stochastic*, sampling strategy, provided a distribution of RNA secondary structures is defined. Valuable insights into the sampling strategies for RNA secondary structures were contributed by Ding and Lawrence [91], when they demonstrated how we can efficiently sample any structure, given exactly its weight in the ensemble. We adopt the Boltzmann [92] distribution from the field of thermodynamics, which is a probability distribution for a system which has different possible states. Building on this idea is the statistical mechanics concept of a *partition function*. Note that this is a function because it is dependent on the temperature in the system, for RNA typically this is human body temperature 37° C or 98.6° F. Assuming the temperature, the partition function is then a scalar which corresponds with all the possible states in a system at thermodynamic equilibrium. The McCaskill [66] algorithm is the first to compute the partition function for an ensemble of RNA secondary structures (cf. Chapter 5). In the general case, S is the set of all possible secondary structures. Therefore, the partition function Z is defined as a sum of the energies of each structure in the ensemble, with respect to the universal gas constant, K, and the temperature, T: $Z = \sum_{s \in S} e^{-[E(s)/KT]}$. As introduced by McCaskill [66], this sum may be replaced by a sum over all possible loops closed by (i, j) where F_L is the free energy of an associated loop, including the summed contributions of their subloops: $Z^b_{ij} = \sum_L e^{-[F_L/KT]} \prod_{(h,l) \in L} Z^b_{hl}$. For the base case, $Z^b_{ii} = 0$ where $i < j$ or $h \leq l$. The complete partition function Z_{ij} must also include all structures where $bp(i) \neq j$. Thus, for $h \leq l$ we have $Z_{ii} = 1.0$ and $Z_{i+1,i} = 1.0$; and where $\exists h, l$ such that $i \leq h < l \leq j$: $Z_{ij} = 1.0 + \sum_{h,l} Z_{i,h-1} Z^b_{h,l}$ [66].

In order to utilize DP to compute the partition function, certain properties must be maintained within the algorithm. First, the algorithm must be complete, meaning each possible secondary structure must be decomposed to fully describe the ensemble. With regard to completeness, it is important to define the space or class of secondary structures which are included, e.g., pseudoknot-free secondary structures. Second, the decomposition scheme must be unambiguous, i.e., each secondary structure in the class should be reached by the algorithm exactly once. Certain algorithms like the one introduced by Zuker violate the unambiguity principle in the DP scheme when the same structure can be reached more than once by the decomposition. With the goal of the Zuker algorithm to find the MFE structure, unambiguous decomposition is not an issue. However, including the same ensemble structure twice when computing the

partition function will introduce bias and must be avoided in the algorithm design.

We now dive into the implementation of Shapify, the relaxed MFE hierarchical folding algorithm, with applications to the SARS-CoV-2 frameshift RNA sequence.

Chapter 3

Shapify: Paths to SARS-CoV-2 Frameshifting Pseudoknot

Multiple coronaviruses including MERS-CoV causing Middle East Respiratory Syndrome, SARS-CoV causing SARS, and SARS-CoV-2 causing COVID-19, use a mechanism known as -1 programmed ribosomal frameshifting (-1 PRF) to replicate. SARS-CoV-2 possesses a unique RNA pseudoknotted structure that stimulates -1 PRF. Targeting -1 PRF in SARS-CoV-2 to impair viral replication can improve patients' prognoses. Crucial to developing these therapies is understanding the structure of the SARS-CoV-2 -1 PRF pseudoknot. Our goal is to expand knowledge of -1 PRF structural conformations. Following a structural alignment approach, we identify similarities in -1 PRF pseudoknots of SARS-CoV-2, SARS-CoV, and MERS-CoV. We provide in-depth analysis of the SARS-CoV-2 and MERS-CoV -1 PRF pseudoknots, including reference and noteworthy mutated sequences. To better understand the impact of mutations, we provide insight on -1 PRF pseudoknot sequence mutations and their effect on resulting structures. We introduce *Shapify*, a novel algorithm that given an RNA sequence incorporates structural reactivity (SHAPE) data and partial structure information to output an RNA secondary structure prediction within a biologically sound hierarchical folding approach. Shapify enhances our understanding of SARS-CoV-2 -1 PRF pseudoknot conformations by providing energetically favourable predictions that are relevant to structure-function and may correlate with -1 PRF efficiency. Applied to the SARS-CoV-2 -1 PRF pseudoknot, Shapify unveils previously unknown paths from initial stems to pseudoknotted structures. By contextualizing our work with available experimental data, our structure predictions motivate future

RNA structure-function research and can aid 3-D modeling of pseudoknots.

3.1 Shapify Chapter Summary

Identifying inter-viral structural similarity in frameshifting pseudoknots is valuable for treatment development as existing viruses mutate or novel diseases emerge. We followed a structural alignment approach to identify the consensus structure for SARS-CoV and MERS-CoV -1 PRF pseudoknots, with SARS-CoV-2 as reference. We developed Shapify for improved prediction of possibly pseudoknotted RNA secondary structures guided by SHAPE reactivity data. Then, we shed light on the structure formation of the SARS-CoV-2 and MERS-CoV frameshifting pseudoknots by analyzing possible structural conformations. Previous structural predictions obtained via different methods all have significant differences. Our results demonstrate innate resiliency via converging paths of the SARS-CoV-2 virus in achieving the native -1 PRF stimulating pseudoknot. Fully accounting for pan-coronaviral structural conformations, which include transient and suboptimal structures, is vital for comprehending viral function.

3.2 Introducing Shapify

Sequence analyses of the SARS-CoV-2 genome classify it as a member of the Betacoronavirus subfamily, which includes SARS-CoV and MERS-CoV [93, 94]. All coronaviruses use a particular replication strategy called ribosomal frameshifting, which is a promising target for therapeutic drug development [95]. They utilize the combination of a *slippery sequence* (where the ribosome is prone to slipping forwards or backwards into a different reading frame) and an *RNA pseudoknot* to cause the ribosome to shift from one reading frame into the other [96, 97]. The expected structures of -1 PRF stimulating pseudoknots, referred to as *native* or *wild-type* pseudoknots, have been identified and studied for multiple viruses [34, 98]. In addition, it was found that some non-native pseudoknot conformations, those that differ from the native structure, play a role in regulating frameshifting [34]. Specifically, the *conformational plasticity* of pseudoknots, i.e., their ability to form non-native structures, was established to be correlated with -1 PRF efficiency [34].

In the case of SARS-CoV-2, two different long open reading frames comprise two-thirds of the viral genome: ORF1a and ORF1b (cf. Fig 3.1) [99]. ORF1b is out

of frame with respect to ORF1a, meaning the ribosome will not translate both frames without shifting from one frame to the other.

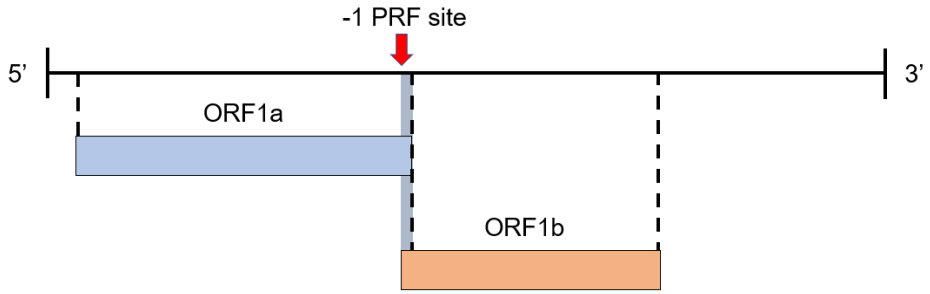


Figure 3.1: SARS-CoV-2 Viral Genome. −1 PRF site marked by the red arrow. Normally, ribosome translates the complete ORF1a. Sometimes at the −1 PRF site, ribosome shifts from ORF1a to ORF1b, resulting in synthesis of fusion ORF1a/1b polypeptide.

The native structure of the frameshift stimulating pseudoknot in SARS-CoV-2 possesses a unique three-stemmed structure, an H-type pseudoknot (cf. Fig 3.2). Kelly et al. [55] observed the rate of frameshifting in SARS-CoV-2 to be approximately 15%-30%, indicating there is further variation in structure beyond what the authors could capture in their experimental procedure. Indeed, unfolding experiments to investigate structural dynamics of the SARS-CoV-2 frameshifting pseudoknot revealed multiple distinct conformations [100].

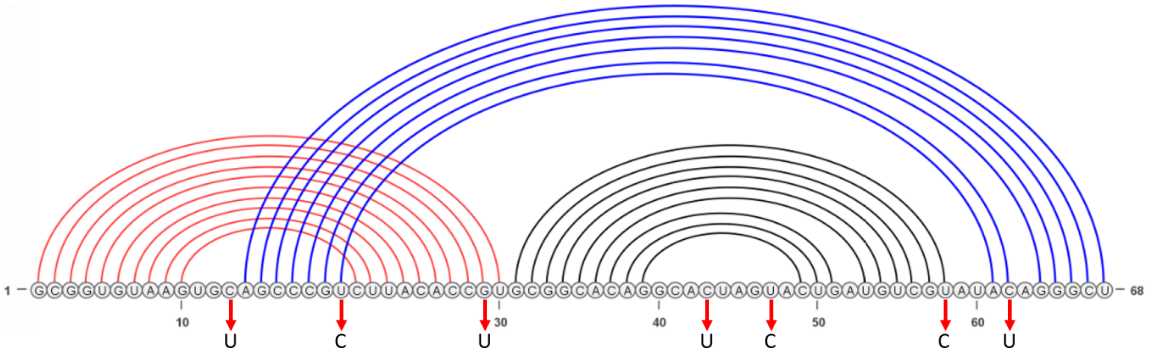


Figure 3.2: SARS-CoV-2 −1 PRF pseudoknot Native Structure. Stem 1 in red, Stem 2 in black, and pseudoknotted Stem 3 in blue. Visualization generated using VARNA [6]. Mutations of interest shown with red arrows.

The structural model proposed for the SARS-CoV-2 −1 PRF pseudoknot continues to evolve. Huston et al. [10] found complex folding dynamics with multiple conformational clusters. In [9, 15, 12], additional structures for the pseudoknot have

been proposed. Furthermore, Yan et al. [101] suggest generalized structure motifs that may be affected by unfolding dynamics during translocation. However, there has not yet been a focus on identifying the ensemble of energetically favourable structures in proximity to the MFE structure of the pseudoknot. Given that non-native folding paths correlate with −1 PRF efficiency and are relevant to structure function, further research on identifying energetically favourable SARS-CoV-2 −1 PRF pseudoknot conformations is valuable in structural prediction.

Most of the existing software packages to determine 3-D conformations of RNA utilize *secondary structure* (set of base pairs) as an input constraint [41, 42, 43, 44, 45, 46]. Indeed, Omar et al. [102] used the known native secondary structure as a starting point in their 3-D physics-based modeling of the SARS-CoV-2 −1 PRF pseudoknot and subsequently identified three unique stable conformations. The non-native secondary structures of the SARS-CoV-2 −1 PRF pseudoknot, however, have yet to be identified and incorporated into structural prediction efforts. Here, we provide alternate starting points and input constraints, which can improve both the accuracy and interpretability of 3-D physics-based modeling of the pseudoknot.

Mutations have been reported in all components of the SARS-CoV-2 −1 PRF pseudoknot sequences [103, 104] available on GenBank [105] and GISAID [106]. Although there are far less available sequenced genomes for MERS-CoV, mutations similar to the ones found in SARS-CoV-2 were observed in the MERS-CoV −1 PRF pseudoknot sequence in multiple samples [2]. Yet, the possible effects of such mutations on RNA structure formation and resulting secondary structures have not been fully characterized. Understanding the impact of mutations is vital to long-term success of treatments that rely on RNA structure. Interestingly, the most prevalent mutation to the SARS-CoV-2 −1 PRF pseudoknot sequence increased its similarity to the MERS-CoV −1 PRF pseudoknot [103]. Further validation of this similarity between the SARS-CoV-2 and MERS-CoV −1 PRF pseudoknots is highly relevant for future coronavirus structure-function prediction efforts.

Here, we computationally explore how the single mutation observed in the MERS-CoV −1 PRF pseudoknot sequence changes structure prediction. Kelly et al. [55] evaluated effects of 14 mutations aiming to experimentally disrupt −1 PRF function of SARS-CoV-2. We seek to expand their work by assessing single nucleotide (nt) mutations observed in the general population that were not considered in their work. A single nucleotide change from cytosine to uracil at position 62 (C62U) was identified in a significant proportion of sequences [103]. Based on information in

the COVID-19 CG database [107], which contains over 10.2 million sequences via GISAID [106] as of September 2022, C62U continues to be the most prevalent mutation in the −1 PRF pseudoknot sequence (0.33% of global sequences). Only two other mutations in the −1 PRF pseudoknot sequence were detected in over 0.1% of global samples: A8G in 0.25%, and C43U in 0.16% of all sequences. Notably, no mutations have been detected in the −1 PRF pseudoknot sequences of any of the following lineages: B.1.1.7 (a.k.a. 20I/501Y.V1 Variant of Concern), B.1.351 (a.k.a. 501.V2 variant, 20C/501Y.V2), B.1.617.2 (a.k.a. Variant of Concern Delta G/478K.V1), and B.1.1.52A+BA.* lineage (a.k.a. Variant of Concern Omicron GRA) including sublineages (i.e., B.A.2, B.A.3, B.A.4 and B.A.5 [108, 109]) see Table S in S1 File for accession IDs.

Even a single nucleotide change in an RNA sequence can affect the resulting structure, with functional implications. There is evidence that as little as one or two mutations or deletions can completely disrupt the native structure formation of the SARS-CoV-2 frameshifting pseudoknot [26, 110]. Neupane et al. [104] identified that a uracil to cytosine mutation (U20C) for the SARS-CoV-2 −1 PRF pseudoknot resulted in more than a three-fold reduction in −1 PRF efficiency [104]. We further explore the mutations presented in [104], as these mutations occur in regions important for the formation of the −1 PRF pseudoknot. Three of these mutations are located near the junction of stems (U20C, G29U, U58C), one is adjacent to an adenine bulge identified as critical in frameshifting (C62U) [111], and two are in Loop 2 where mutations have been shown to reduce frameshifting efficiency in SARS-CoV (C43U, U47C) [112]. Incorporating known mutations in pseudoknot structure prediction will advance the understanding of potential conformations, which could alter frameshifting efficiency with implications for viral infectivity and pathology.

It was noted that reducing frameshifting efficiency of SARS-CoV attenuated viral propagation to a significant degree [113, 114, 111]. Distinct small molecules have been identified, which can bind to the SARS-CoV −1 PRF pseudoknot and disrupt its function [115, 116, 117]. In 2014, Ritchie et al. observed that the introduction of the small molecular ligand 2-{[4-(2-methylthiazol-4-ylmethyl)-[1,4]diazepane-1-carbonyl]amino} benzoic acid ethyl ester (MTDB) effectively decreased alternate folding by binding to the SARS-CoV −1 PRF pseudoknot [116]. The result is attributed to the possible hydrogen bonds formation of MTDB with the nucleotides in Loop 3, inhibiting a non-native folding path resulting in reduction of the −1 PRF rate to near 0%. Additional experiments find certain small molecules to be

effective at inhibiting frameshifting for multiple coronaviruses–in both human and bat–showcasing the potential for pan-coronaviral therapeutic treatment design [118]. *In vitro* experiments for SARS-CoV-2 determined that MTDB binding reduced −1 PRF by 60% [55]. This result was replicated and expanded upon by experiments that found that various mutations in the SARS-CoV-2 −1 PRF pseudoknot sequence did not have a significant effect on anti-frameshifting activity of MTDB [104]. More recently, the compound merafloxacin was identified as significantly decreasing the efficiency of the −1 frameshift in SARS-CoV-2 [117]. This result was further validated by *in vivo* experiments [26], confirming the viability of targeting the frameshift element as a high-efficacy treatment. Another compound, 2-(5-acetylthiophen-2yl)furo[2,3-b]quinoline (KCB261770), was found to reduce frameshift efficiency in SARS-CoV, SARS-CoV-2, and MERS-CoV [119]. This points to possible structural similarity between the three coronaviruses, and makes KCB261770 a promising potential therapeutic for a range of coronaviruses. Increasing comprehensive structural knowledge of −1 PRF pseudoknots in coronaviruses can therefore be valuable in enhancing small molecule therapeutics.

SHAPE-MaP [38] combines chemical probing with an energy model to identify structural motifs on entire viral RNA genomes [1]. High-throughput structure probing methods including SHAPE-MaP and DMS-MaPseq [35] (dimethyl sulfate (DMS) mutational profiling with sequencing) measure many individual molecules and average the results to generate single-base resolution reactivity data. Recent studies performed *in vivo* SHAPE-MaP analyses of SARS-CoV-2 [9, 10, 11]. Despite improvements, high-throughput methods targeting the entire $\sim 30,000$nt genome are prone to noise and may suffer inaccuracies when predicting the structure for a specific region less than 100nt long. To focus on a smaller genomic region it can be beneficial to analyze predictions guided by *in vitro* SHAPE data on specific fragments of RNA, for example the SARS-CoV-2 frameshifting pseudoknot sequence [12].

In this work we present *Shapify*, an algorithm for prediction of RNA pseudoknotted structure based on the hierarchical folding hypothesis while incorporating reactivity data, to provide new insight to the SARS-CoV-2 −1 PRF pseudoknot structure. We utilized four structural reactivity probing experimental datasets [9, 10, 11, 12] as constraints to Shapify and compared predicted structures with those obtained using ShapeKnots [1], an existing heuristic algorithm for predicting RNA pseudoknotted structure and incorporating reactivity data. We demonstrate that Shapify improves the identification of probable structure formation paths for the SARS-CoV-2 −1 PRF pseudoknot.

In an attempt to expand knowledge of the SARS-CoV-2 −1 PRF pseudoknot structural conformation, in this work we first evaluate and report structural similarities between −1 PRF pseudoknots of SARS-CoV, SARS-CoV-2, and MERS-CoV. To further explore the hypothesized similarity between the SARS-CoV-2 and MERS-CoV −1 PRF pseudoknot structures [103], we enumerate the specific loci of similarity as well as the consensus structure. Following the hierarchical folding hypothesis, we predict structures for SARS-CoV-2 and MERS-CoV −1 PRF pseudoknots. To assess effects of mutations on the frameshifting structure, we provide predictions for seven SARS-CoV-2 −1 PRF mutated sequences and one MERS-CoV −1 PRF mutated sequence. These mutations were selected because they were observed in the population [107, 2] as well as experimentally validated for their effect on −1 PRF frameshifting [104].

Our results contribute to RNA structural prediction by providing a sampling of the landscape of notable–and previously unidentified–non-native secondary structures, which may play a role in regulating frameshifting [34, 120]. Whereas previous work recognized the importance of non-native pseudoknotted structures, the structural paths delineated here disclose non-native structure formation from initial stable stems. We discuss the relationship between each set of predicted initial stems and pseudoknotted structures. This approach is unique in providing information about possible paths to the final pseudoknotted structure. By analyzing predictions based on structural reactivity data, we provide novel information regarding application of such methods in this specific context (see Discussion). Finally, we contextualize our predictions with available experimental results including crystallography and cryo-electron microscopy [12, 37, 101]. Our pseudoknot structural predictions represent alternate starting points, which can improve the accuracy of existing 3-D physics-based modeling [41, 42, 43, 44, 45, 46].

3.3 Shapify Materials and Methods

We first provide a present our sequence data sources and availability. Next we introduce methods for structural similarity detection and secondary structure prediction. That is, we introduce our *Shapify* algorithm and its closest competitor, ShapeKnots. Lastly, we provide information on the source and availability of the SHAPE reactivity data used in this work and introduce our procedure for SHAPE data analysis.

Name	ID	Position	Length	Reference
SARS-CoV	NC_004718.3	13405-13472	68	[55]
SARS-CoV-2	NC_045512.2	13475-13542	68	[55]
MERS-CoV	NC_019843.3	13440-13510	71	[103]

Table 3.1: List of viral sequences with their NCBI accession ID, position of frameshifting structure, length, and their reference.

3.3.1 Sequence Data

We obtained the reference genomes for SARS-CoV, SARS-CoV-2, and MERS-CoV from the National Center for Biotechnology Information (NCBI) [2]. The position of frameshifting structure for each viral sequence was obtained from [55], and [103] as presented in Table 3.1.

The datasets generated and/or analysed during the current study are available in the frameshifting repository: `github.com/HosnaJabbari/frameshifting`.

3.3.2 Structural Similarity Detection

We aligned whole genomes of SARS-CoV-2, SARS-CoV, and MERS-CoV using Clustal Omega [121] version 1.2.4, with SARS-CoV-2 as the reference. To detect *de novo* structural similarities in non-coding regions, we used RNAz v2.0 [122]. RNAz uses evolutionary conservation of functional secondary structures as well as thermodynamic stability of the secondary structure in detecting structural signal. We ran RNAz with our multiple sequence alignment of three sequences and false positive rate of about 1% (setting parameter $P > 0.9$; default value is 0.5).

3.3.3 Hierarchical Folding Prediction Pipeline

Following the HotSpots package of HotKnots V2.0 [4], we identified up to 20 most stable unique stems for each sequence. The stems were ranked based on their free energy and referred to by their IDs. These stems (referred to as *initial stems*) were used as constraints in Iterative HFold, Shapify, and ShapeKnots to explore the suboptimal structural landscape. The secondary structures produced based on each initial stem constraint were ranked by their free energy. Each predicted secondary structure was linked to the initial stem(s) that produced it. We report initial stems and the resulting pseudoknotted structures for eight SARS-CoV-2 -1 PRF stimulating pseudoknot sequences: reference, and seven mutated sequences. Additionally, we

report initial stems and the resulting pseudoknotted structures for two MERS-CoV −1 PRF stimulating pseudoknot sequences: reference, and one mutated sequence.

3.3.4 Shapify Design and Validation

In order to incorporate SHAPE reactivity data to guide the hierarchical folding approach, we adapted Iterative HFold [52] to develop Shapify. Shapify takes as input an RNA sequence, a SHAPE dataset, and a pseudoknot-free secondary structure (cf. Figure 3.3) and outputs the predicted secondary structure. Both SHAPE data and the pseudoknot-free input structure guide Shapify's prediction with known RNA structural information.

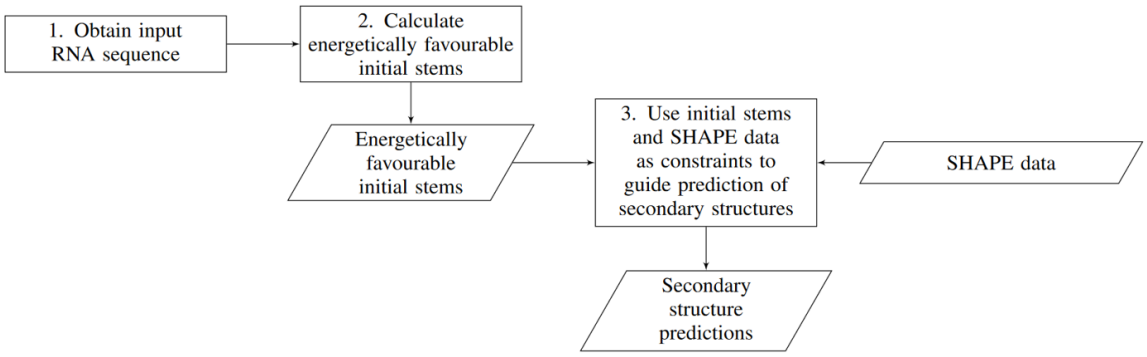


Figure 3.3: Shapify Hierarchical Folding Pipeline. Rectangles dictate actions, parallelograms denote input/output.

We used the pseudo energy terms created by Deigan et al. [123] from SHAPE reactivity data, as a means of integrating such data into our prediction algorithm. This pseudo energy term at index i of an RNA sequence includes a penalty for base pairing that increases with experimentally-derived SHAPE reactivity, m, (also referred to as *slope*) and an intercept that encourages base pairing for nucleotides with low SHAPE reactivity, b:

$$m[\log(SHAPE(i)+1)]+b, \tag{3.1}$$

and is applied to stem energies only. Here, $SHAPE(i)$ refers to the SHAPE reactivity score at position i of the sequence.

The slope and intercept parameters must be determined empirically [1] for each prediction method, thus we aggregated a database of 30 RNAs with known structure and available SHAPE datasets. Data from the original ShapeKnots cross-validation [1]

included five RNAs with lengths > 300 nt, five riboswitch RNAs, four RNAs with structures that are not well predicted by thermodynamic parameters, and three RNAs whose structures are likely modulated by protein interaction. We supplemented this with six RNAs with known structure and available SHAPE data via the RNA Mapping Database [124]: three riboregulators involved with translation [125], SARS-CoV-2 3' and 5' UTR regions [126], and a ribonuclease domain of *Bacillus subtilis* [127].

We implemented a leave-one-out cross validation to determine the optimal values for slope m and intercept b, searching over a grid of 29 possible slope values and 21 possible intercept values (cf. Fig 3.4). For each combination we determined the geometric mean of the sensitivity and positive predictive value (ppv) for each of the RNA. We then averaged all the values excluding one RNA respectively, repeating the procedure for each possible RNA to be excluded and taking an average of averages to calculate the final result. Note that base pairs are considered to be predicted correctly even when one of the two indices is different by up to one nucleotide; this helps account for uncertainty and dynamism in RNA structure [128]. The longest RNA sequence of length 530 nt took 25 minutes for Shapify to deliver a prediction utilizing a 2.1GHz processor with 4GB memory. We determined the optimal values for Shapify parameters as slope $m = 1.4$ and intercept $b = -0.5$ (cf. Fig 3.4).

3.3.5 Shapify Software Availability

Shapify is available at `https://github.com/ltrinity/Shapify` (DOI: 10.5281/zenodo.6100185).

3.3.6 ShapeKnots

ShapeKnots [1] is a heuristic method similar to that of Ren et al. [5] in that it iteratively adds stems while incorporating SHAPE reactivity data for prediction of pseudoknotted structures. ShapeKnots utilizes the pseudo energy terms created by Deigan et al. [123]. ShapeKnots parameters were adjusted from default as follows: maximum number of internally generated structures = 200 (default 100), maximum percent difference in folding free energy change for internally generated suboptimal structures = 100% (default = 20%), maximum number of structures provided to user = 100 (default 20), maximum percent difference in folding free energy change for generating suboptimal structures = 50 (default 20). We compared performance of ShapeKnots and Shapify to provide a baseline.

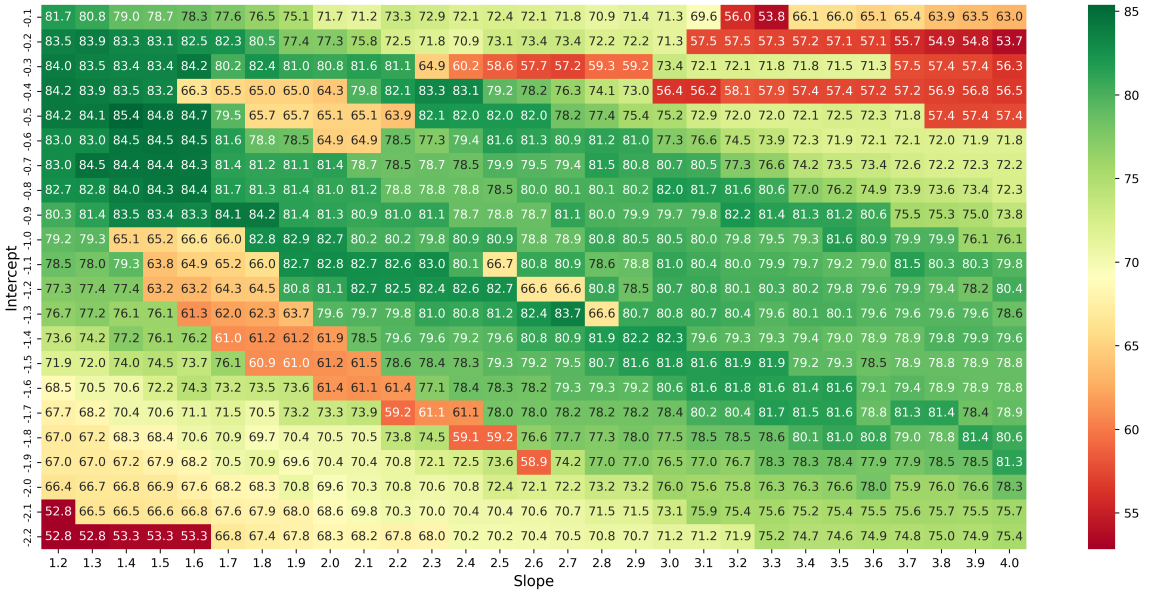


Figure 3.4: Cross Validation Results For each possible combination of the intercept and slope, the value within the grid represents the geometric mean of the sensitivity and positive predictive value over 30 values of results from 29 RNA (each of the 30 RNA left out exactly once). Final result is the average of averages using a leave-one out scheme to avoid bias toward any one RNA. Optimal parameters identified by white box: intercept -0.5, slope 1.4. Color indicates performance relative to optimal with darker green as the best and red as the worst performance.

3.3.7 SARS-CoV-2 SHAPE Data

We obtained four available SARS-CoV-2 SHAPE datasets as presented in Table 3.2.

Huston et al., Manfredonia et al., and Yang et al. are *in vivo* genome-wide probing, while Zhang et al. is *in vitro* probing specific to the SARS-CoV-2 frameshifting pseudoknot sequence. For each dataset, the values corresponding with the 68nt SARS-CoV-2 -1 PRF sequence were used as constraints in RNA secondary structure prediction. For both Shapify and ShapeKnots, in addition to the SHAPE reactivity

Name	Type	Genome-wide	Reference
Huston et al.	*in vivo*	Yes	[10]
Manfredonia et al.	*in vivo*	Yes	[9]
Yang et al.	*in vivo*	Yes	[11]
Zhang et al.	*in vitro*	No	[12]

Table 3.2: List of SARS-CoV-2 SHAPE reactivity datasets used in this work with their referred name, type and reference.

information, a set of initial stems were provided as input to the algorithm.

3.3.8 Bootstrapping

To further explore the degree to which the information in each SHAPE dataset was used in the structure prediction of Shapify and ShapeKnots, we implemented a bootstrap method as follows–for each original dataset, respectively. For each of the 68 indices of the SARS-CoV-2 -1 PRF sequence, the SHAPE reactivity value was selected randomly from the SHAPE dataset, with replacement. This newly generated bootstrap SHAPE information was then used as constraint for secondary structure prediction. This process was repeated $10,000$ times for each of the four SHAPE datasets, for each of our prediction methods (i.e. Shapify and ShapeKnots). Having predicted $10,000$ secondary structures for the bootstrap datasets, the secondary structures were converted into binary information–with 0 representing unpaired and 1 representing paired–and averaged for each position of the SARS-CoV-2 -1 PRF sequence.

3.4 Shapify Results

We first describe the identified structural similarity among SARS-CoV, SARS-CoV-2, and MERS-CoV. Next, we present the predicted secondary structures of the SARS-CoV-2 and MERS-CoV -1 PRF pseudoknot using Iterative HFold and following the hierarchical folding hypothesis. Then we explore the effect of the single nucleotide mutations, as stated in Section 3.3, for SARS-CoV-2 and MERS-CoV. Finally, we visualize SHAPE-based predictions of the SARS-CoV-2 -1 PRF pseudoknot using Shapify, comparing them to ShapeKnots predictions as well as their dependence on the SHAPE data.

3.4.1 -1 PRF Structural Similarity

One predicted loci of the structural similarity output from RNAz encompassed the region containing the -1 PRF pseudoknot for SARS-CoV-2. Structural similarity between SARS-CoV-2, SARS-CoV, and MERS-CoV was detected at locations 13439 to 13555, 13369 to 13485, and 13404 to 13523 in their respective aligned genomes (cf. Table 3.3). Structures are represented in dot-bracket format. Open parentheses show the base on the 5' side of the sequence and the closed parentheses represent the

base on the 3' side of the sequence that are binding together. Each period "." (dot) identifies an unpaired base in the structure.

The alignment produced using RNAz identifies a highly conserved region for the -1 PRF structure of SARS-CoV-2 and MERS-CoV. This is interesting because there is not enough sequence similarity between the SARS-CoV-2 and MERS-CoV -1 PRF sequences to be detected using the Basic Local Alignment Search Tool [129]. For example, Lu et al. found SARS-CoV-2 genome to be $\sim 50\%$ homologous to MERS-CoV, as compared to $\sim 79\%$ homologous to SARS-CoV [130].

We note that while RNAz's prediction identifies conserved similarity in structures of SARS-CoV-2 and MERS-CoV, it only predicts pseudoknot-free structures (cf. Table 3.3), necessitating our further study of pseudoknots with MFE methods.

3.4.2 SARS-CoV-2 -1 PRF Pseudoknot (68 nt)

To predict the structure of the SARS-CoV-2 -1 PRF pseudoknot, following the hierarchical folding hypothesis we first generated a set of initial stems based on the reference sequence (cf. Section 3.3). Table 3.4 presents most stable initial stems for the SARS-CoV-2 -1 PRF stimulating pseudoknot ranked based on their free energies.

Using these 18 distinct initial stems as input constraints, we predicted 16 unique secondary structures for the SARS-CoV-2 -1 PRF pseudoknot reference sequence using Iterative HFold [52]. The top 11 most stable structures (based on their free energy) are presented in Table 3.5, see Table C in S1 File for the complete list. We use parentheses "( )" and square brackets "[ ]" to represent crossing base pairs that identify pseudoknotted structures. The first column in Table 3.5 lists the initial stem ID(s) corresponding to the constraint that resulted in the predicted secondary structure. Note that the same structure can result from different initial stems because Iterative HFold allows for minor modifications to the input constraint.

The secondary structure with the lowest free energy as predicted by Iterative HFold is 100% consistent with the native SARS-CoV-2 -1 PRF pseudoknot structure (cf. Table 3.5, row 1). While some structures were native-adjacent–structurally close to the native structure–we also predicted native non-adjacent (markedly different) structures. In the third row of Table 3.5, for example, we have a structure with free energy of -18.26 kcal/mol that has the pseudoknot forming further towards the 3' end of the RNA compared to the native structure. Native non-adjacent pseudoknotted structures can be seen (cf. Table 3.5, rows 3, 8, 9 and 11) where native Stem 3 (cf. blue

Table 3.3: −1 PRF Structural similarity for SARS-CoV, SARS-CoV-2, and MERS-CoV as predicted by RNAz. Gaps in the alignment are represented as hyphen. Asterisks (*) in bottom row (PK location) correspond to the known location of SARS-CoV-2 and MERS-CoV −1 PRF native structures.

Virus	−1 PRF Partial Predicted Secondary Structure
SARS-CoV	((((((..((((((((((.((((....)))))....
SARS-CoV-2	.(((((((...(((...((((....)))))...)))
MERS-CoV	.....((((((((((..((.............(..(((((((((((.....)))))))))))..)..(((((.....)))))..))..))))))))))((((.....
Consensus	AGUCAGCUGAUGCAUCAACGUUUUUAAACGGGUUUGCGGUGUAAGUGCA- - -GCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUCUACAGGGCUUUUGAC
	.(((((....(((.....((((....))))).....(((((((((.((....)).....))))))))))))))...)))))....(((..((((((.............))))
PK location	----------**********------

Table 3.4: The most stable initial stems in SARS-CoV-2 −1 PRF stimulating pseudoknot reference sequence, ranked based on their free energies. These stems were used as structural constraint for predicting the SARS-CoV-2 −1 PRF stimulating pseudoknot secondary structure following the hierarchical folding hypothesis. First column provides stem ID (i.e. rank) and the third column lists free energy of the stem. Input sequence is provided in the bottom row.

ID	Initial Stem	Free Energy (kcal/mol)
1	((((((((((.....)))))))))).....	-10.79
2	.....(((((((((.....))))))))).....	-4.67
3	.....(((((.....))))).....	-3.77
4	.....((((.....)))).....	-3.47
5	.....(((((.....))))).....	-2.74
6	.....((((.....)))).....	-2.54
7	.....((((.....)))).....	-2.53
8	.....(((.....))).....	-2.42
9	.....((((.....)))).....	-2.35
10	((((.....)))).....	-2.26
11	.....(((.....))).....	-2.10
12	.....((((.....)))).....	-2.07
13	.....(((((.....))))).....	-1.36
14	.....((((.....)))).....	-1.32
15	..(((((.....))))).....	-0.66
16	.....(((((.....))))).	-0.51
17	...((((.....)))).....	-0.38
18	(((((((.....))))))).....	-0.24
Sequence	GCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUAUACAGGGCU	

Table 3.5: Predicted secondary structures for the SARS-CoV-2 −1 PRF stimulating pseudoknot based on the reference sequence. These structures are predicted by Iterative HFold given the initial stems in Table 3.4 as structural constraints. Certain suboptimal structures (e.g. 6_s and 2_s, cf. Fig 2.7) are reported, because in these cases the structure has only slightly higher free energy than the MFE structure predicted by Iterative HFold. Native structure is marked with an asterisk (*) in row 1. As shown in rows 1, 2 and 5 of the table, multiple initial stems can result in a single prediction for the −1 PRF stimulating pseudoknot. Input sequence is provided in the bottom row.

Initial Stem ID	Predicted Secondary Structure	Free Energy (kcal/mol)
1, 3, 9, 18	((((((((((...[[[[[[[)))))))))).(((((((((.....))))))).)))))..]].]]]]]*	-18.86
8, 11	((((((((((.([[[[[[)))))))))).(((((((((.....))))))).))))).....]]]]]	-18.80
4	((((((((((.....))))))))))..(((((((((.[[.[[.[[[)))).))))...]]]]].]].	-18.26
14	.(((((((((...[[[[[[[)))))))))).(((((((((.....))))))).))))))).]].]]]]]	-17.82
5, 13, 15	..(((((((((...[[[[[[[)))))))))).((((((((((.....))))))).)))))))]].]]]]]	-17.63
7	(((((((((....[[[[[[[.)))))))))).(((((((((.....))))))).))))))..]].]]]]]	-17.37
12	(((((((((.....[[[[[[[..)))))))))).(((((((((.....))))))).))))))..]].]]]]]	-16.62
6	((((((((((.....))))))))))).(((((.....[[[.[[[.)))))).....]]].]]]	-16.24
6_s	((((((((((.....))))))))))).....(((..[[[.[[[.)))).....]]].]]]	-16.08
2	.((((((.(((((((((((.....[[[[[[[[...)))))).))))))).]]]]]]]]]].]]]]]	-15.26
2_s	..(((((((.....)))))))).(((((((..[[[.[[.[[.....)))))))))]]]].]]]	-14.84
Sequence	GCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUAUACAGGGCU	

stem in Fig 3.2) does not form and expected unpaired bases in the second stem loop are paired.

3.4.3 MERS-CoV −1 PRF Pseudoknot

Table 3.6 presents initial stems for the MERS-CoV −1 PRF stimulating pseudoknot, as predicted for the reference sequence. For the complete set of stems see Table D in S1 File. Using the 18 initial stems as input constraints, 15 unique secondary structures were predicted for the MERS-CoV −1 PRF pseudoknot reference sequence using Iterative HFold. These predicted pseudoknotted structures are presented in Table 3.7 sorted by their free energy, see Table E in S1 File for the complete list. Note that multiple structures were reached from more than one starting point (cf. Table 3.7, rows 1 & 2). There is no well established native structure for the MERS-CoV −1 PRF pseudoknot. The secondary structure with the lowest free energy as predicted by Iterative HFold is 93% consistent with the structure presented by Fourmy and Yoshizawa [103]. Comparing the MFE structure predicted for MERS-CoV and SARS-CoV-2 −1 PRF pseudoknots (cf. Tables 3.5 and 3.7, row 1), we see the familiar three stems and H-type pseudoknotted structure. In the MERS-CoV −1 PRF pseudoknot predictions, the second and third lowest free energy structures are pseudoknot-free. Furthermore, only six of the 15 predicted secondary structures were pseudoknotted. By contrast, each of the 16 structures predicted for the SARS-CoV-2 −1 PRF pseudoknot (cf. Table 3.5) contained pseudoknotted base pairs. The energy of predicted secondary structures for the MERS-CoV −1 PRF pseudoknot are generally higher compared to structures predicted for SARS-CoV-2, indicating they may be less stable. In addition, there is a bigger energy gap between the MFE predicted structure for MERS-CoV and the second most energetically favourable structure (0.94 kcal/mol); as compared with SARS-CoV-2 where there are 3 structures within 0.94 kcal/mol of the MFE structure.

3.4.4 Effect of Mutations on the SARS-CoV-2 −1 PRF Pseudoknot

We repeated the hierarchical folding method for each mutated sequence, see red arrows in Fig 3.2 for mutation locations in the native structure. We note that Neupane et al. observed a significant decrease of frameshifting efficiency in SARS-CoV-2 −1 PRF for only U20C mutation [104] while Ishimaru et al. observed a decrease in frameshifting

Table 3.6: Top four energetically favourable stems predicted for MERS-CoV −1 PRF stimulating pseudoknot reference sequence. These stems were used as structural constraint to Iterative HFold program to predict the secondary structure of the −1 PRF stimulating pseudoknot. First column provides stem IDs, the second column presents the predicted stems in dot-bracket format and the third column lists free energy of the given stem in kcal/mol. Input sequence provided in bottom row. See Table D in S1 File for complete table.

ID	Stem	Free Energy (kcal/mol)
1	((((((((((.....)))))))))).....	-10.39
2	.....((((.....)))).....	-4.69
3	.((((.....)))).....	-4.25
4	.....(((((.....)))))	-3.79
Sequence	GGGGUUCUAUUGUAAAUGCCCGAAUAGAACCCUGUUCAAGUGGUUUGUCCACUGAUGUCGUCUUUAGGGCA	

Table 3.7: Top four lowest free energy secondary structures predicted for MERS-CoV −1 PRF stimulating pseudoknot based on the reference sequence. These structures are predicted by Iterative HFold given the input stems in Table 3.6 as structural constraints. Input sequence is provided in the bottom row. See Table E in S1 File for the complete table.

Initial Stem ID	Secondary Structure Prediction	Free Energy (kcal/mol)
1, 45	((((((((((.....[[[[[.))))))))))..(((.((((.....)))))))...........]]]]]	-17.36
2, 3, 8	((((((((((.....)))))))))).....	-16.40
12	((((((((((.....)))))))))).....	-16.17
13	((((((((((.....)))))))))).....	-16.14
Sequence	GGGGUUCUAUUGUAAAUGCCCGAAUAGAACCCUGUUCAAGUGGUUUGUCCACUGAUGUCGUCUUUAGGGCA	

efficiency of SARS-CoV with C43U and U47C [112].

The predicted initial stems did change in some cases between the reference sequence and the mutated sequences. Certain stems that were predicted for both the reference sequence and respective mutated sequences had differences in free energy. Some stems were stable for the reference sequence but destabilized by mutated sequences. In addition, specific mutated sequences led to initial stems that were not identified for the reference sequence. We present novel initial stems for mutated sequences ranked by their free energies and referred to by their IDs based on where they would have been ranked by free energy relative to the initial stems of the reference sequence (cf. Table 3.8). Initial stems for mutated sequences are given an ID with a letter to distinguish them from the reference sequence stems (e.g., 5a). If multiple stems from mutated sequences have the same free energy ranking with respect to the initial stems of the reference sequence, they are given IDs with sequential letters (e.g., 5a and 5b).

The initial stems predicted for the C13U mutated sequence destabilized initial stems 5 and 7 (as their free energy increased). One novel stem was detected for this mutation, referred to as 13a (cf. Table 3.8).

The initial stems predicted for the U20C mutated sequence destabilized initial stem 1, but stabilized stems 9, 10, 12, and 16 (their free energies decreased). Two novel stems were detected for the U20C mutated sequence referred to as 5a and 5b.

Repeating the method for the sequence with the guanine/uracil mutation at position 29 (G29U) did not change the predicted energy of any of initial stems, but did not detect four stems (initial stems 1, 2, 4 and 14). Five novel stems were detected based on G29U mutation, referred to as 2a, 3a, 11a, 13b, and 15a.

For the sequence with the cytosine/uracil mutation at position 43 (C43U), initial stems 4 and 5 were destabilized. Initial stems 6 and 15 were not detected for C43U, but a novel stem was detected, referred to as 4a.

Initial stems predicted for the sequence with the uracil/cytosine mutation at position 47 (U47C) stabilized initial stem 13. Two novel stems were detected for the U47C mutation, referred to as 4b and 8a.

For the sequence with the uracil/cytosine mutation at position 58 (U58C), initial stems 2 and 18 were stabilized. One novel stem was detected for the U58C mutation, referred to as 18a.

Finally, for the sequence with the cytosine/uracil mutation at position 62 (C62U), initial stem 16 was not detected. There was a novel stem identified for this mutation, referred to as 13c.

Table 3.8: Predicted initial stems for mutated SARS-CoV-2 −1 PRF stimulating pseudoknot sequences. These stems were used as structural constraint for predicting the secondary structure of mutated SARS-CoV-2 −1 PRF sequences. First column identifies the mutation and its location in the 68 nt −1 PRF sequence. For example C13U identifies a mutation from C to U at index 13. Second column provides the stem ID based on the stem's free energy and the ranking of the initial stem of the reference sequence. For example, in row 1 the stem has a free energy of −1.87 kcal/mol and is denoted by stem ID 13*a*. Relative to the initial stems predicted for the reference sequence (cf. Table 3.4), this stem has the thirteenth lowest free energy. Third column represents the predicted initial stem for the mutated sequence, and the fourth column provides free energy of the given stem (kcal/mol). Input sequence is provided in the bottom row for each mutation section with mutations highlighted in yellow.

Mutation	ID	Stem	Free Energy (kcal/mol)
C13U	13a	.....((((((.....)))))).....	-1.87
		GCGGUGUAAGUG**U**AGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUAUACAGGGCU	
U20C	5a	.....(((.....))).....	-2.99
	5b	..((((.....)))).....	-2.77
		GCGGUGUAAGUGCAGCCCG**C**CUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUAUACAGGGCU	
G29U	2a	..(((((((.....))))))).....	-8.39
	3a	.....(((((((.....))))))).....	-3.86
	11a	.....(((((.....))))))..	-2.13
	13b	....((((((.....))))))....	-2.04
	15a	.....(((((.....))))).....	-0.92
		GCGGUGUAAGUGCAGCCCGUCUUACACC**U**UGCGGCACAGGCACUAGUACUGAUGUCGUAUACAGGGCU	
C43U	4a	.....(((((((....))))))).....	-3.74
		GCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCA**U**UAGUACUGAUGUCGUAUACAGGGCU	
U47C	4b	.....((((.....)))).....	-3.54
	8a	.((((((.....)))))).....	-2.46
		GCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAG**C**ACUGAUGUCGUAUACAGGGCU	
U58C	18a	.....((((.....)))).....	-0.27
		GCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCG**C**AUACAGGGCU	
C62U	13c	.....((((((.....))))))...	-1.96
		GCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUAUA**U**AGGGCU	

We used the 18 initial stems from the reference sequence in addition to the novel initial stems for respective mutated sequences (cf. Table 3.8) as input constraints to predict secondary structures using Iterative HFold. For each of the seven SARS-CoV-2 −1 PRF pseudoknot mutated sequences, $14 - 19$ secondary structures were predicted (cf. Tables G-M in S1 File). The native structure, the structure with the MFE in Table 3.5, was not predicted for U20C, G29U, or C62U mutated sequences based on any of the initial mutated stems. For example, initial stems 1 and 3 that would result in the native structure based on the reference sequence, did not result in the native structure based on the sequence with the U20C mutation (cf. Table H in S1 File). Instead, they joined initial stems 8 and 11 resulting in a non-native structure.

In general two to five structure clusters (identical prediction from multiple different initial stems) were identified for each mutated sequence.

3.4.5 Effect of Mutation on the MERS-CoV −1 PRF Pseudoknot

We obtained 237 MERS-CoV genomes from the NCBI Virus Variation database [2]. There was a mutation observed at position 13479 from C to U in three sequences (KR011263, MG011354, and KR011266). Following the procedure explained in Section 3.3, we obtained initial stems for the mutated sequence: initial stem 12 was destabilized, and there was a novel initial stem detected (cf. Table D in S1 File). We used the 18 initial stems as input constraints to predict secondary structures using Iterative HFold leading to 13 unique secondary structure predictions (cf. Table N in S1 File). The MFE structure predicted for reference and mutated sequence remained the same. The mutated base was only predicted as paired in one out of the 13 structures, although pre-mutation it was paired in 12 out of 15 of the structures predicted for the reference sequence. Notably, initial stems 2, 3, and 8 led to a different structure, accommodating the mutation with a larger loop.

3.4.6 SARS-CoV-2 −1 PRF Pseudoknot with SHAPE (68 nt)

In this section we present Shapify results using the 18 initial stems (cf. Table 3.4) and four SHAPE datasets (cf. Section 3.3) as constraint. We found significant overlap among the four sets of predictions, giving a total of 43 unique secondary structure predictions (cf. Table O in S1 File).

In Fig 3.5 and Fig 3.6 we visualize structural paths from the initial stems to the predicted secondary structures from Shapify. Beginning from the left, each initial stem is labeled with its ID (cf. Table 3.4). There are four predictions obtained for each initial stem, one for each of the SHAPE datasets used. Note that predictions among the four results could be the same. Here we include any additional suboptimal structures within 2 kcal/mol of the MFE prediction. Darker color represents higher agreement among SHAPE datasets, meaning the same prediction was obtained with different SHAPE data.

Fig 3.5 and Fig 3.6 present structural paths from the initial stems to the native and non-native structures predicted by our Shapify algorithm. All non-native structures

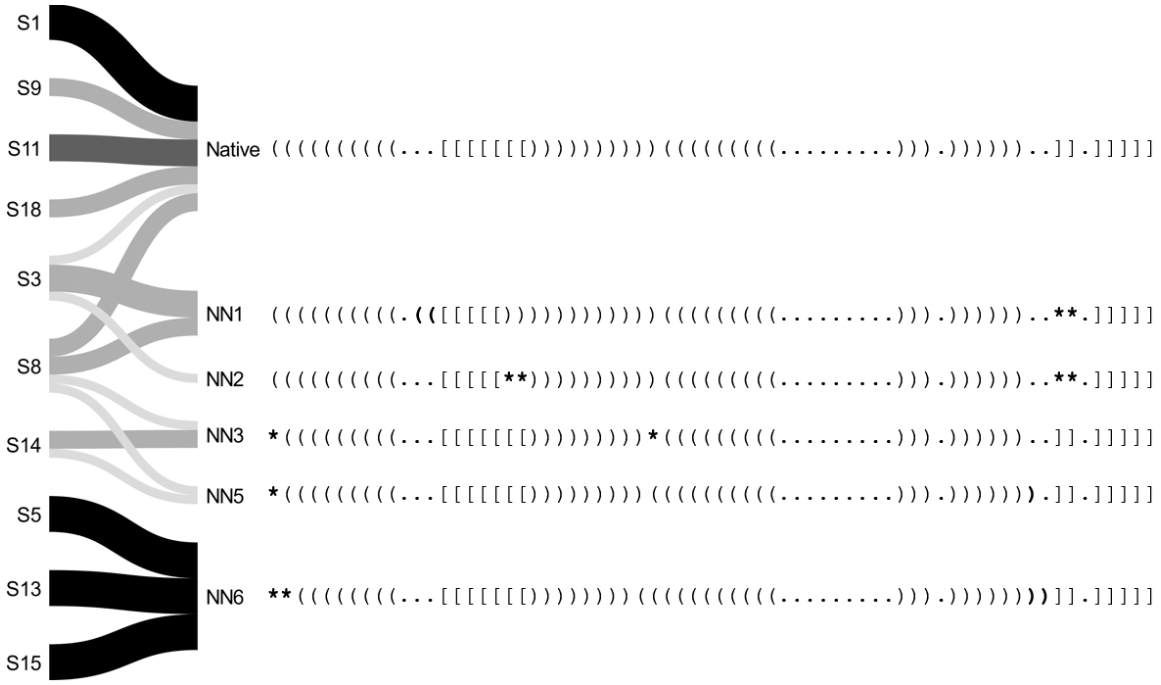


Figure 3.5: Shapify Predicted Native-adjacent Structural Paths. Presentation of structural paths from each initial stem leading to native or non-native (but native-adjacent) secondary structures (e.g., NN1 is the lowest free energy non-native structure; cf. Table O in S1 File). Initial stems labeled on the left (e.g., S1 for initial stem 1; cf. Table 3.4). If the structure predicted for a specific initial stem was the same for all four SHAPE datasets it is presented with a black colored path. In other cases, where the predicted structure was the same for three, two, or only one of the SHAPE datasets, the path is colored dark grey, grey, or light grey, respectively. Differences from the native structure are marked in bold, with parentheses/brackets representing changes in paired bases, and asterisks representing predicted unpaired bases that were paired in the native structure.

are numbered according to their free energy (e.g., NN1 is the lowest free energy non-native structure, NN2 is the second lowest free energy non-native). The initial stem IDs are annotated for each secondary structure and SHAPE dataset along with all structures' free energies in Table O in S1 File. For additional visualization with initial stems highlighted, cf. Fig C-D in S1 File. The native structure and five native-adjacent structures (non-native secondary structures that have only minor differences from the native structure) are included in Fig 3.5. Separately, we visualize nine native non-adjacent structures that are significantly different from the native structure (cf. Fig 3.6). Differences between the native structure in the first row and other structures below are indicated in bold; brackets show differences in paired bases,

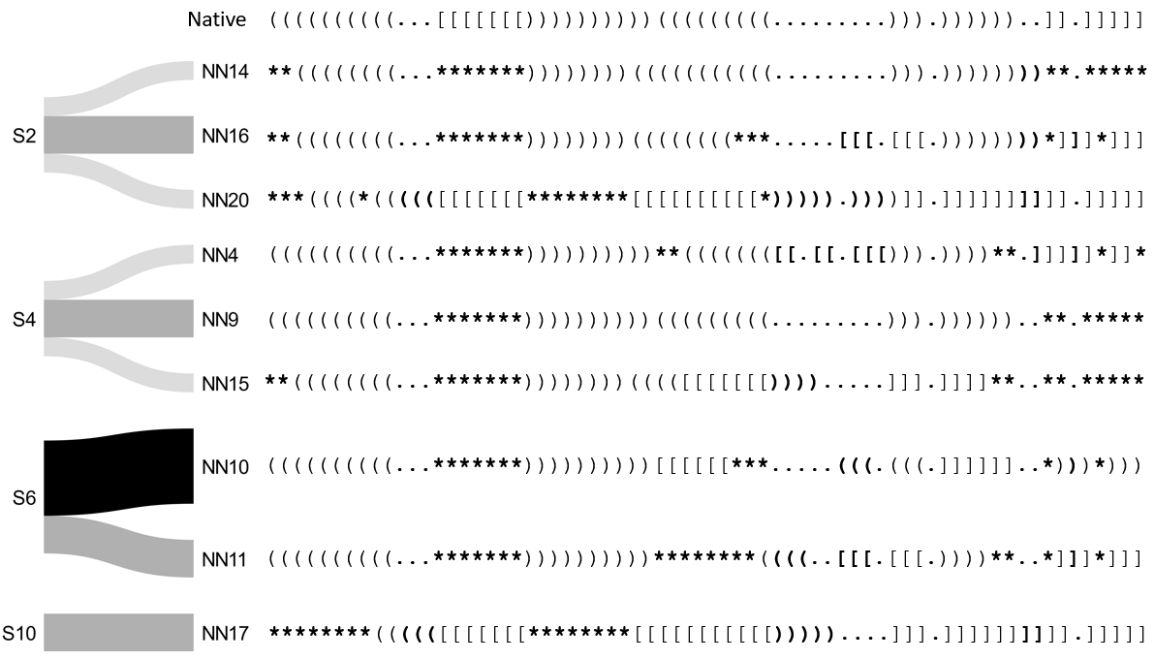


Figure 3.6: Shapify Predicted Native Non-adjacent Structural Paths. Presentation of structural paths from the initial stems leading to native non-adjacent secondary structures. Initial stems are labeled on the left (cf. Table 3.4). If, for a given initial stem, the predicted structure is the same for all four SHAPE datasets, it is presented with a black colored path. In other cases, where the predicted structure is the same for three, two, or only one of the SHAPE datasets, the path is colored dark grey, grey, or light grey, respectively. Differences from native structure are marked in bold, with parentheses/brackets representing changes in paired bases, and asterisks representing predicted unpaired bases that are paired in the native structure.

and asterisks identify bases that are paired in the native structure but predicted unpaired in the non-native structure.

The native structure was reached by six different initial stems (cf. Fig 3.5). This indicates that there are structural paths to the native structure from an array of initial stems. Native-adjacent structures, also display significant diversity in the paths with which they can be reached. Specifically, non-native structures 1, 3, 5, and 6 (i.e. the first, third, fifth, and sixth lowest free energy non-native structures) which are classified as native-adjacent were each reached by two different initial stems. For native non-adjacent structures, however, there were not multiple paths to the same structure from different initial stems. Furthermore, there was less agreement between predictions from different SHAPE datasets in structural paths to the native non-adjacent structures (presented with lighter path colors).

3.4.7 ShapeKnots Predictions

We obtained predictions for the SARS-CoV-2 -1 PRF pseudoknot using ShapeKnots and each of the SHAPE datasets. In addition, since ShapeKnots is capable of receiving input structures, we further execute ShapeKnots with each SHAPE dataset and initial stems (cf. Table 3.4), respectively, as constraints. The lowest free energy structures predicted by ShapeKnots were obtained when no initial stem was used as input to the program (i.e. only RNA sequence and SHAPE data was used as input, cf. Table 3.9). As shown in Fig 3.7, structures predicted by ShapeKnots, have generally higher free energy compared to the structures predicted by Shapify.

Initial stems 1 and 2 as input constraints led to predictions with slightly higher free energy (compared to the ones predicted by Shapify), and each had multiple suboptimal structures. Interestingly, none of the structures predicted by ShapeKnots when given initial stem 1 as input constraint were pseudoknotted (cf. Table 3.9, bottom four rows). This is in contrast to Shapify's prediction that identifies initial stem 1 as a structural path to the native pseudoknotted structure with high agreement among all SHAPE datasets (cf. Fig 3.5). We note that in the absence of SHAPE reactivity data, Iterative HFold also identified the initial stem 1, as one of the initial stems that reach the native pseudoknotted structure (cf. Table 3.5, row 1).

ShapeKnots identified two paths from initial stem 2 leading to energetically favourable native-adjacent structures that contained multiple distinct sets of pseudoknotted base pairs (cf. Table 3.9, rows 5 and 6). These new pseudoknotted base pairs create a kissing-hairpin structure as opposed to the native H-type pseudoknotted structure. The initial stem 2 also has a path to reach an H-type pseudoknotted structure that resembles the native structure with a shift of Stem 1 to the 3' end.

3.4.8 SARS-CoV-2 -1 PRF Pseudoknot SHAPE Data Analysis

Fig 3.8A presents comparison of SHAPE reactivity values for each position in the 68nt -1 PRF SARS-CoV-2 sequence reported by Manfredonia et al. [9], Huston et al. [10], Yang et al. [11], and Zhang et al. [12]. A base is considered reactive (unpaired) if its reactivity score is above 0.3.

There are five regions that are identified as reactive by at least two SHAPE datasets: bases $6-13$, $19-32$, $41-48$, $51-52$, $59-63$, $67-68$. The loop region from positions $60-63$ was well captured by the Zhang et al. and Yang et al. datasets,

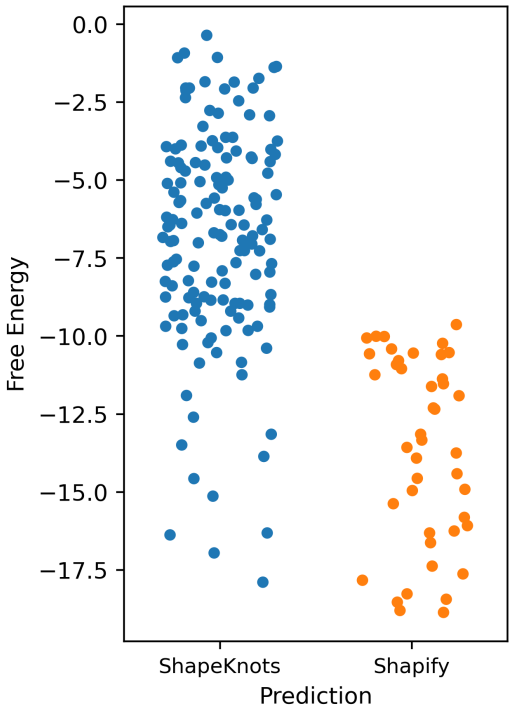


Figure 3.7: Comparison of Free Energy of ShapeKnots and Shapify Structural Predictions. Each dot represents a unique secondary structure prediction for SARS-CoV-2 -1 PRF frameshifting pseudoknot. The y-axis represents free energy in kcal/mol with lower free energy corresponding with more energetically favourable structures. ShapeKnots (blue) failed to predict the most energetically favourable structures when given SHAPE data and initial pseudoknot free stems as constraint as compared to Shapify (orange). All structures are listed in Tables O-P in S1 File.

with Huston et al. and Manfredonia et al. also finding reactivity in this region. There were limited positions that all SHAPE datasets agreed reported no reactivity above the threshold: positions $1-3$, $15-16$, 18, $33-34$, $36-37$, $49-50$, and 54. For significant reactivity, all datasets agreed only on positions 23 and 29.

Fig 3.8B presents ShapeKnots' predictions using bootstrapped SHAPE values repeated $10,000$ times and averaged for each index (where 0 identifies an unpaired base and 1 a paired base). A mean values of 1.0 at index i conveys the nucleotide at position i was predicted as paired for all bootstrapped datasets, and mean value of 0.0 at position i indicates that nucleotide i was always predicted as unpaired for all bootstrapped datasets.

Fig 3.8C similarly presents Shapify's predictions using bootstrapped SHAPE values.

There were significant differences with respect to the predictions via ShapeKnots

Table 3.9: Most energetically favourable ShapeKnots [1] secondary structure predictions for the SARS-CoV-2 −1 PRF stimulating pseudoknot based on the reference sequence. SHAPE dataset source for each prediction indicated in the first column. Second column provides stem IDs, NA if none used. Certain suboptimal structures (e.g., 2_s) are reported because of only slightly higher free energy than the MFE structure predicted by ShapeKnots.

SHAPE Data	Initial Stem ID	Secondary Structure Prediction	Free Energy (kcal/mol)
[10]	NA	.((((((((((.(([[[[[[))))))))))))((((((((((.....)))).)))))))....]]]]]	-18.11
	2	..((((((((.(([[[[[[))))))))))))((((((((((.....)))).))))))))...]]]]]	-17.89
[12]	NA	.((((((((((...[[[[[..))))))))))((((((((((.....)))).)))))))....]]]]]	-17.85
[9]	NA	.((((((((((...[[[[[[[[))))))))))((((((((((....)).))).)))))))].].]]]]]	-17.25
[12]	2_s	[[(((((((((.]][[[[[..))))))))))((((((((((.....)))).))))))))...]]]]]	-16.95
[9]	2_s	[[(((((((((.]][[[[[[[[))))))))))((((((((((....)).))).))))))))].].]]]]]	-16.38
[12]	1	((((((((((.....))))))))))((((((((((.....)))).)))))).....	-16.32
[10]	1_s	((((((((((.((.....))))))))))))((((((((((.....)))).)))))).....	-15.13
[11]	1_s		
[9]	1_s	((((((((((.((.....))))))))))))).((((((((((....)).))).)))))....	-14.57
Sequence		GCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUAUACAGGGCU	

and Shapify following bootstrapping. Shapify predicted the first and second bases in the sequence to be paired far more frequently than ShapeKnots did. This is interesting because, despite general consensus on all SHAPE datasets for no reactivity for these positions (i.e. paired), three energetically favourable native-adjacent structures (NN3, NN5, and NN6, cf. Fig 3.5) were predicted to have these base(s) unpaired.

Furthermore, positions 12, 13, 40, 41, 46, 47, and 59 tend to be predicted as paired via ShapeKnots; however, with Shapify these bases are more likely predicted as unpaired. Multiple SHAPE datasets reported reactivity at these positions, indicating Shapify predictions better match the SHAPE data.

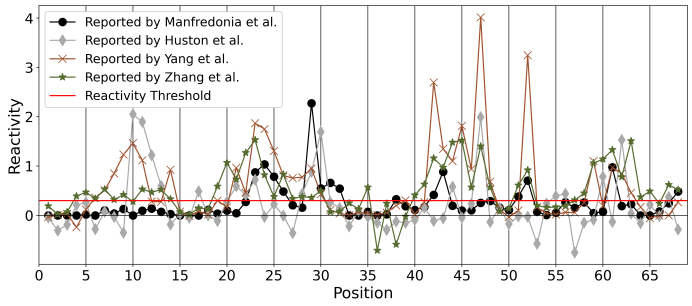
ShapeKnots tends to predict loci 61 and 62 as unpaired (consistent with SHAPE datasets), while Shapify predictions tend towards a paired prediction.

In general, Shapify follows the SHAPE datasets more closely and appears more resilient to extreme values in the SHAPE data, seen by the relatively lower variance especially when data by Yang et al. [11] was used.

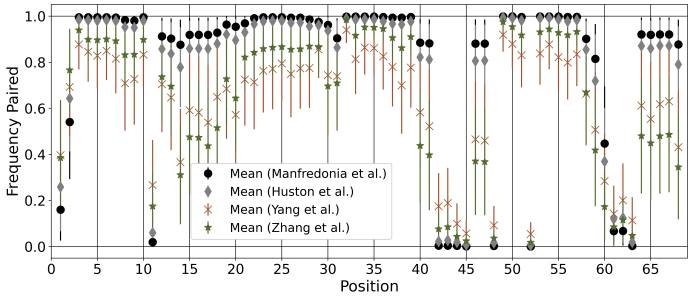
3.5 Shapify Discussion

Here we discuss implications of our SARS-CoV-2 −1 PRF structure prediction results to advance potential treatment development.

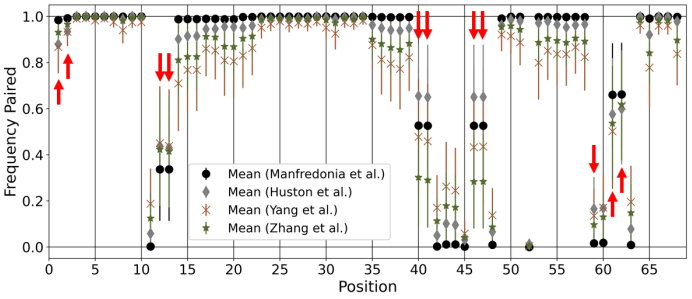
Conserved similarity. First we aligned viral genomes and used RNAz to



(a)



(b)



(c)

Figure 3.8: SHAPE Dataset Analysis. A: Comparison of −1 PRF sequence SHAPE reactivity dataset reported by Manfredonia et al. [9], Huston et al. [10], Yang et al. [11], and Zhang et al. [12]. Reactivity at or below 0.3 is considered to be low or non-reactive indicating the base is paired. B: ShapeKnots' predictions using bootstrapped SHAPE values were obtained 10,000 times and averaged for each index. The mean and variance of the 10,000 predictions for each respective SHAPE dataset are shown here. The y-axis indicates the frequency each nucleotide is predicted as paired. Mean value of 1.0 at a specific position conveys that the nucleotide was predicted as paired for all bootstrapped values, and mean value of 0.0 indicates that the nucleotide was predicted as unpaired for all bootstrapped data. C: Bootstrap procedure was repeated with Shapify used for prediction. Red arrows mark differences from ShapeKnots predictions.

identify a region of conserved structural similarity among SARS-CoV, SARS-CoV-2, and MERS-CoV. In addition to the expected similarity between SARS-CoV and

SARS-CoV-2, we detected structural similarity with MERS-CoV in its −1 PRF pseudoknot region despite low sequence similarity between the two. This result was hypothesized by Fourmy and Yoshizawa [103], and SHAPE data supports the existence of a pseudoknot at the same position in all three coronaviruses [131]. Zhang et al. [12] identified a similar 3D RNA structure using de novo computer modeling in a range of betacoronaviruses including MERS-CoV, SARS-CoV, and SARS-CoV-2. This aligns well with the structural similarity we present in this work. We note that while RNAz detected similarity in the −1 PRF of SARS-CoV-2 and MERS-CoV, its predicted structure is pseudoknot-free whereas the expected structure for this region is indeed a pseudoknotted structure. We therefore, continued our quest to better determine the structure in both cases.

Structural prediction. By following the hierarchical folding hypothesis we explored the structural landscape of the SARS-CoV-2 and MERS-CoV −1 PRF pseudoknots, expanding on previous structure prediction efforts [103, 12]. Recent work using optical tweezers to identify folding pathways of the SARS-CoV-2 frameshifting pseudoknot strongly supports hierarchical folding, as Stem 1 was identified to fold first followed by Stem 3 [100]. Our predictions for energetically favourable initial stems aligns with this result, with initial stem 1 possessing the lowest free energy by a significant margin (ID 1, cf. Table 3.4).

In both SARS-CoV-2 and MERS-CoV, the second most energetically favourable initial stem (ID 2, cf. Tables 3.4 and 3.6) results in a structure prediction that is markedly different from the native structure (cf. Tables 3.5 and 3.7). Initial stem 2's formation is, however, supported by our structural alignment. ShapeKnots predictions identified additional paths from initial stem 2 to native-adjacent structures that intriguingly contained two different sets of pseudoknotted base pairs (kissing-hairpin).

The MFE structure predicted by Iterative HFold for SARS-CoV-2 −1 PRF matched the native structure. The predicted MFE structure for MERS-CoV possesses a recognizable three stemmed structure resembling the native type for the SARS-CoV-2 frameshifting pseudoknot. In addition, the two initial stems (ID 1 and 4, Table 3.6) that led to this prediction for MERS-CoV, strongly resemble initial stems 1 and 3 (Table 3.4) that led to native and native-adjacent structures in SARS-CoV-2. The lowest free energy initial stem (ID 1) for SARS-CoV-2 and MERS-CoV, respectively, is highly conserved in the structural alignment (cf. Table 3.3). These indicators point to possible kinetic paths for frameshifting pseudoknots that may be consistent across coronaviruses. Indeed, similar structures have been observed to initiate frameshifting

in human coronavirus HCV 229E [132]. As a future direction we seek to further expand our structural analysis to additional coronaviruses and upstream targets including attenuator hairpin and palindromic sequences [133].

For SARS-CoV-2 five of the predicted pseudoknotted structures including the native structure, resulted from multiple initial stem input constraints. These structures can be reached from different starting points, which may indicate their increased conformational plasticity. SARS-CoV-2 has greater path resiliency in achieving the native -1 PRF pseudoknot structure and forming pseudoknots in general, when compared with MERS-CoV -1 PRF pseudoknot predictions that were often pseudoknot-free. Pseudoknot predictions for SARS-CoV-2 had lower MFE than those for MERS-CoV. Furthermore, there is a smaller energy gap between the MFE structure of SARS-CoV-2 and the next lowest free energy prediction; where MERS-CoV had a larger gap between the MFE prediction and the next lowest free energy prediction. This may be related to the overall trend that SARS-CoV and SARS-CoV-2 exhibited greater stable RNA structuredness across their genomes compared to MERS-CoV [131]. We hypothesize this may be a factor in the continued spread among various species of COVID-19. Further experiments are needed to confirm the exact structure of non-native conformations for the SARS-CoV-2 -1 PRF pseudoknot and how these specific structures correlate with frameshifting efficiency and viral pathology.

The native pseudoknotted structure for SARS-CoV-2 -1 PRF is formed by crossing Stems 1 and 3 (an H-Type pseudoknot). However, we observed a structure that connects Stems 2 and 3 through a "shifted/smaller" Stem 3 (that crosses Stem 2). This pattern is observed for both SARS-CoV-2 and MERS-CoV. This change of conformation may be functionally important.

Mutations effect. To explore the effect of point mutations on structure of SARS-CoV-2 -1 PRF, we used an array of most observed mutations in the population as well as a set of previously studied mutations [103]. The predicted secondary structures based on the SARS-CoV-2 -1 PRF mutated sequences are demonstrably different from those for the reference sequence; certain mutations stabilized while others destabilized the initial stems. In particular, U20C, and G29U destabilized initial stem 1, and did not reach the native pseudoknotted structure. Reduction in frameshifting efficiency was experimentally observed by Neupane et al. [104] as a result of U20C mutation. Mutations in Loop 2 (C43U and U47C) were previously found to reduce frameshifting efficiency in the original SARS-CoV [112], although the mechanism is unknown. Interestingly, while there were changes in the predicted

secondary structures for the mutated sequences, the function of the SARS-CoV-2 pseudoknot is expected to be conserved for G29U, C43U, U47C, U58C, and C62U [104]. Given that the structure of the pseudoknot with sequence mutations is different but still functional, we hypothesize that the relationship between pseudoknotted structure and function may be more dynamic and flexible than previously expected. In the case of the U20C mutation, we identified a structure that is significantly different from the native expectation and has low free energy (cf. Table H in S1 File, row 3). This native non-adjacent structure is ranked third in terms of free energy, which is not the case for any other mutation. In addition, U20C mutation was the only SARS-CoV-2 sequence that led to a suboptimal structure predicted for initial stem 1. This could be the cause of competition for paths from initial stem 1, which may be functionally crucial. Further experiments are needed to validate this result and confirm the structural cause for the decrease in -1 PRF efficiency based on the U20C mutation.

Although there are limited MERS-CoV genomic sequences available, one particular mutation (C13479U) was observed in multiple samples. The MFE predicted structure was unchanged due to the mutation, in which the third stem was shifted in multiple predictions to accommodate the mutated base in a larger loop instead. Notably, the conformational switch (shifted Stem 3) as described above is preserved despite the mutation.

Incorporating SHAPE. To incorporate chemical modification data into our predictions, we introduced Shapify, an algorithm that takes as input an RNA sequence, SHAPE data, and partial structure information and outputs a possibly pseudoknotted RNA secondary structures based on the hierarchical folding hypothesis and guided by SHAPE reactivity score.

Through Shapify predictions we identified paths from initial stems to secondary structures that showcase remarkable redundancy in the mechanism for achieving the native SARS-CoV-2 -1 frameshifting pseudoknot, as well as native-adjacent structures. Shapify's predictions can help capture both the conformational flexibility, diverse structural landscape, and convergence towards native structure of the SARS-CoV-2 -1 PRF pseudoknot.

We observed more consistency in predictions for native and native-adjacent structures: multiple paths from different initial stems to these structures, and agreement when various SHAPE datasets were each used as constraint. Given that the frameshift is vital for viral replication, we expect structural resilience in achieving the frameshift inducing structure. Such structural resiliency can be seen by convergence to the native

and native-adjacent structures from different initial stems. Note the four black paths in Fig 3.5 that indicate the same predicted pseudoknot for each of the four SHAPE datasets as constraint. In further support of the proposed native structure resiliency, is the lack of path redundancy for non-native adjacent predictions (cf. Fig 3.6). Interestingly, initial stem 2, which is energetically favourable, has multiple paths to different native non-adjacent structures. Further study is required to determine the effect of this mechanism in regulating frameshift efficiency.

Using crystallography for the SARS-CoV-2 -1 PRF pseudoknot accurate to 2.09 angstroms, Roman et al. [37] unveiled marked differences compared to the previously identified crystallography-based structures for the frameshifting structure, including a shorter Stem 2. Here using Shapify we identified multiple new conformations for the SARS-CoV-2 -1 PRF, including a shorter Stem 2, different location for formation of Stem 2 or pseudoknotted base pairs that are not reflected in the native structure.

Recently, using a method combining graph theory, secondary structure prediction, SHAPE structural probing,and thermodynamic ensemble modeling Schlick et al. [15, 101] identified three structural motifs for the SARS-CoV-2 -1 PRF. One of these motifs (denoted as '3_6', coarse grained three stem structure invariant to stem length), corresponds with the first, second, and fourth lowest free energy non-native structures identified by Shapify. Following cryo-EM experiments Bhatt et al. [26] also identified a structure that corresponds with the '3_6' motif.

Our results indicate that NN1, NN2 and NN4 share paths to form following the pairing of initial stems 3, 8, or 14. In addition, both initial stems 8 and 14 had paths to two of the '3_6' motif structures, depending on which SHAPE dataset was used as constraint. In the case of initial stem 8 leading to the formation of the fifth lowest free energy non-native structure, the path is suboptimal, as the initial stem 8 can also lead to the formation of the native structure, the first, or the third lowest free energy non-native structure.

In addition it should be noted that kinetics may also play a role in initial stem formation. Specifically, refolding of RNA after the ribosome passes over and unfolds the pseudoknot may favor pairing towards the 5' end which exits the ribosome first.

Combining our predicted structures in the absence of SHAPE data (using Iterative HFold) for mutated SARS-CoV-2 -1 PRF sequences, and Shapify's predictions with four SHAPE datasets on the reference sequence, we observed that any modification to initial stem 1 (including destabilization or missing initial stem 1 due to mutation) resulted in no native structure prediction, possibly because initial stem 1 is a key

path to the native structure (note in Fig 3.5 structural path from initial stem 1 to the native structure is supported by all four SHAPE datasets). High degree of agreement among all SHAPE datasets is observed also for initial stems 5, 13, 15 (to NN6) and 16 (to NN19). This may justify the decrease in the -1 PRF efficiency previously observed for U20C in SARS-CoV-2 (possibly due to destabilization of initial stem 1), and C43U (possibly due to destabilization of initial stems 5 and 15) and U47C (possibly due to stabilization of initial stem 13) experimentally observed in the original SARS-CoV virus. This can further point to possible decrease/change in efficiency of -1 PRF in C13U (due to destabilization of initial stem 5 and introduction of a novel stem 13a), G29U (due to destabilization of initial stem 1 and introduction of a novel stem 13b), and C62U (missing the initial stem 16 and introduction of a novel stem 13c). Our predicted structures in the absence of SHAPE reactivity data included no native structure for U20C, G29U and C62U. We hypothesize initial stems 1, 5, 13, 15 and 16 to be important transient structures in the structural path of the -1 PRF of SARS-CoV-2 genome. Further experiments are needed to assess this hypothesis.

Comparison with ShapeKnots. Comparing performance of Shapify with that of ShapeKnots, a heuristic method that utilizes SHAPE reactivity data, we demonstrated that Shapify identified lower free energy structures than the ones identified by ShapeKnots. ShapeKnots predictions were not found to effectively utilize the initial stems (i.e., ShapeKnots performance was better without the initial stems as constraint). In addition, Shapify predictions better align with the SHAPE datasets for the -1 PRF structure of SARS-CoV-2. Unlike ShapeKnots, Shapify minimizes the free energy of the possibly pseudoknotted output structure relative to the given input structure and the SHAPE reactivity data. Therefore Shapify's method of adding pseudoknotted stems is better motivated energetically than that of ShapeKnots. In addition, Shapify is flexible, allowing minor modification to the input structure to allow formation of energetically favourable base pairs.

Bootstrap. We observed discrepancies in multiple positions in the secondary structure predictions of ShapeKnots when compared to the SHAPE reactivity data used. For instance, SHAPE datasets indicated high reactivity that was not evident in the actual secondary structure prediction. We therefore performed bootstrapping to assess how much secondary structure prediction for ShapeKnots (and Shapify) are influenced by the accuracy of the given SHAPE reactivity data. For ShapeKnots we observed that approximately half of the nucleotides were not impacted at all by change in the value of SHAPE reactivity induced by the bootstrapping method when

Manfredonia et al. and Huston et al. datasets were used. In case of the Yang et al. dataset we observed more variability in the structures predicted after bootstrapping; this can also be attributed to more extreme values in this dataset. Sensitivity to SHAPE data was less severe in structures predicted by Shapify as evaluated by our bootstrapping method. While ShapeKnots seems to be more sensitive to extreme values in the SHAPE data used, both methods seem to be more influenced by their thermodynamic parameters than by SHAPE data.

We note that *in vivo* SHAPE data is collected by probing the entire $30,000+$ nt SARS-CoV-2 viral genome. If a structure spans across the slippery sequence, it can be difficult to understand how unfolding triggered by ribosome translocation affects formation of the frameshifting pseudoknot and its ability to initiate the frameshift. Here we account for possible inaccuracies in the *in vivo* SHAPE reactivity data by including the *in vitro* SARS-CoV-2 SHAPE dataset from Zhang et al. [12] which focused on a specific fragment of the viral genome, the -1 PRF region, and is corroborated by cryo-EM imaging.

The four SHAPE datasets presented here have some similar trends but are unique at various positions. This could be attributed to the different protocols or approaches for data collection. It may also indicate that the RNA was in different structure conformations for each individual measurement during the high-throughput experiment. Structural probing methods measure multiple molecules and then average the results to obtain final reactivity data, which may lead to noise in the signal when various molecules are each in different structure conformations. Disagreement between SHAPE datasets raises questions to the extent of reliance on using SHAPE data. The SHAPE datasets used in this work, for example, reported significant reactivity for positions 23 and 29 (indicating unpaired region), while both indices are believed to be paired in Stem 1 [100]. Here we acknowledge that tightness of the pseudoknotted loops caused by dimerization with other loops of the structure have been previously cited as possibly affecting correctness of SHAPE data [102]. However, in the SHAPE datasets we analyzed we could not find a non-reactive region (in two or more SHAPE datasets) that covers a loop region of the SARS-CoV-2 -1 PRF native structure.

Previous efforts on RNA 3D modeling of the SARS-CoV-2 -1 PRF pseudoknot acknowledged the relevance of non-native structures and conformational plasticity [34], but this concept has not been fully integrated into the latest prediction efforts [102]. Efforts in identifying unique stable 3D conformations of the SARS-CoV-2 -1 PRF pseudoknot can have a great impact in treatment development [102]. We believe such

endeavors can benefit from a more comprehensive overview of the SARS-CoV-2 −1 PRF pseudoknot secondary structure landscape, one that includes non-native structures. Given the complicated nature of modeling tertiary interactions, the initialization of such simulations should account for multiple initial secondary structures of the pseudoknot. Our structural predictions can be utilized in 3D physics-based modeling of pseudoknots as alternative starting points or to provide context to SARS-CoV-2 structure prediction efforts, where small secondary structure differences can have a profound impact on resulting tertiary interactions and three-dimensional conformations of the pseudoknot [41, 42, 43, 44, 45, 46].

3.6 Shapify Conclusions

We set out to expand knowledge of the SARS-CoV-2 −1 PRF structural conformation to aid with the current efforts for identification of potential treatment options. We detected −1 PRF pseudoknot conserved structural similarity among SARS-CoV, SARS-CoV-2, and MERS-CoV; and discussed varying degrees of RNA structuredness and similar −1 PRF folding paths between the viruses. We identified energetically favourable initial stems for SARS-CoV-2 and MERS-CoV −1 PRF pseudoknots, and used these initial stems as constraint for secondary structure prediction via the Iterative HFold algorithm. To further predict possibly pseudoknotted structures, we introduced Shapify to utilize both SHAPE data and partial structure information. To assess Shapify's predicted structures, we compared them with structures predicted by ShapeKnots; we found Shapify's predicted structures more stable (lower in free energy) than those predicted by ShapeKnots. In order to understand the effect of SHAPE data on predictions, we performed a bootstrap procedure, which demonstrated Shapify is less sensitive to SHAPE data compared to ShapeKnots. Shapify pseudoknot predictions reveal SARS-CoV-2 possesses consistent path resiliency in achieving the −1 PRF pseudoknot native structure, with multiple pathways from initial stems to native and native-adjacent structures as compared with paths to native non-adjacent structures.

We determined the effects of most observed point mutations to the SARS-CoV-2 and MERS-CoV −1 PRF pseudoknot sequences, finding that certain mutations may stabilize or destabilize pseudoknot structure. Our results indicate SARS-CoV-2 initial stem 1 is critical in formation of the native −1 PRF SARS-CoV-2 pseudoknot, with initial stem 2 having paths to native non-adjacent structures. Understanding structural

similarities between coronaviruses such as initial stem/pseudoknot alignment and path convergence can shed light on their mechanisms of function and help with finding effective treatments for existing and emerging contagions. Similarities in structure between coronaviruses may contain vital information that can validate proposed explanations for frameshifting. The individual and consensus structures presented here can further inform structure prediction, providing previously unavailable insights to explain paths of structure formation. In the next chapter we extend this structural analysis to longer frameshifting pseudoknot sequences, in a wider array of coronaviruses, with additional sequence covariation information.

Chapter 4

Tying the Knot of the Coronavirus Frameshift Pseudoknot

Using mutual information from 181 coronavirus sequences, in conjunction with the algorithm KnotAli, we compute secondary structure predictions for the frameshift site of different coronaviruses. We then utilize the Shapify algorithm to obtain most stable SARS-CoV-2 secondary structure predictions guided by frameshift sequence-specific and genome-wide experimental data. We build on our previous secondary structure investigation of the singular SARS-CoV-2 68 nt frameshift element sequence, by using Shapify to obtain predictions for 132 extended SARS-CoV-2 sequences. We include covariation information in our results via a companion hierarchical folding algorithm, KnotAli, which uses a multiple sequence alignment to obtain secondary structure predictions but does not use SHAPE data. Previous investigations have not applied hierarchical folding to extended length SARS-CoV-2 frameshift sequences. By doing so, we simulate the effects of ribosome interaction with the frameshift site, providing insight to biological function. We contribute in-depth discussion to contextualize secondary structure dual-graph motifs for SARS-CoV-2, highlighting the energetic stability of the previously identified 3_8 motif alongside the known dominant 3_3 and 3_6 (native-type) -1 PRF structures. Integrating experimental data within minimum free energy (MFE) hierarchical folding algorithms provides novel structure predictions to distill the relationship between RNA structure and function. In particular, fully categorizing most stable secondary structure predictions via hierarchical folding supports our identification of motif transitions and critical site targets for future therapeutic research.

4.1 Chapter Summary

Finding evolutionary connections between coronaviruses frameshift element RNA structures is a worthwhile goal in contributing to treatment development for afflicted human and animal populations. Predicting the most energetically favourable RNA secondary structures, and how they may form via the hierarchical folding hypothesis, is an efficient use of computational resources to shed light on RNA structure-function.

We used the KnotAli algorithm to obtain mutual information from 181 coronaviruses frameshift RNA sequences. Guided by this evolutionary information, we computed secondary structure predictions to allow comparison of marked similarities and subtle differences between SARS-CoV-2 and other coronaviruses frameshift element RNA structures. In addition, we applied the Shapify algorithm to predict secondary structures for extended SARS-CoV-2 frameshift element sequences informed by SHAPE reactivity data. Here we critically expand the known landscape of most stable -1 PRF secondary structure conformations, isolating the location of key secondary structure motif transitions that can improve site targeting of viral therapeutics. Our application of hierarchical folding algorithms contributes novel predictions of functional RNA structures, enhancing discussion of how secondary structures unfold or refold to regulate frameshifting in coronaviruses.

4.2 Extended Background

Given the high degree of complexity in predicting how the frameshift pseudoknot structure may be wedged or somehow possibly obstruct the entrance of the ribosome mRNA channel to initiate the frameshift event [26], we strive to fully understand the initial and subsequent folding of RNA molecules within a secondary structure model. To better inform and interpret tertiary structure experiments and simulations, our computational analysis explores the landscape of possibly pseudoknotted structures to elucidate key SARS-CoV-2 folding conformations. Experiments show that frameshift efficiency is significantly higher for an extended frameshift sequence (2924 nucleotides) than minimal frameshift sequence (92 nucleotides) [133]. Therefore, unfolding of longer range RNA structures must affect how the RNA refolds into a pseudoknot in proximity with the ribosome [12, 134]. Here we quest to further characterize the ensemble of most stable RNA secondary structures for extended frameshift sequences in SARS-CoV-2 and related coronaviruses.

Predicting and understanding frameshift inducing RNA structures in SARS-CoV-2 and related viruses is a critical target for therapeutic development [95, 134]. One example is intracellular small molecule therapy [135, 136, 137, 104, 117, 26, 138], a strategy to limit viral fitness by disrupting the twisted RNA structure that contacts the ribosome at multiple locations [26]. Experiments demonstrate that specific compounds can inhibit frameshift efficiency in SARS-CoV-2 [117, 26] and other coronaviruses, affecting both humans and bats [118]. For example, KCB261770 was found to reduce frameshift efficiency in SARS-CoV-1, SARS-CoV-2, and MERS-CoV [119]. Indeed, novel viral genome targeting found the SARS-CoV-2 frameshift pseudoknot site to be the most effective location to limit viral reproduction [139].

Vigilance is necessary as coronaviruses continue to evolve in animal reservoirs. SARS-CoV-2 viral evolution analysis finds it most phylogenetically related to bat SARS-like coronaviruses [94]. An intelligent strategy to prepare for future SARS-CoV-2 variants or novel coronaviruses must leverage evolutionary structure information as well as the *hierarchical folding hypothesis*, in order to understand the role of initial and subsequent RNA folding within frameshift element mechanics, e.g., our earlier work in predicting secondary structures for the 68 nucleotide (nt) SARS-CoV-2 sequence [14]. Searching for commonality of structure features between coronaviruses contributes to broad spectrum pseudoknot therapeutic targeting, evidenced by molecular dynamic simulations of previously unknown 3-D structures for bat-coronavirus frameshift mechanics [140]. Previous analysis of length-dependent structures in different coronaviruses identified how sequences evolved to support a range of frameshift element structure motifs [134] (cf. Section 4.3.1).

Functional RNA structures that regulate the frameshift event have been studied for multiple viruses [34, 116, 98]. In particular, RNA sequences in the frameshift region have been observed to possess *conformational plasticity* [34]. For SARS-CoV-2, folding into multiple structures is a functional viral phenomenon evidenced by optical tweezers experiments [100, 141] as well as SHAPE-MaP (selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling) [38, 1] chemical probing of the SARS-CoV-2 viral genome *in vitro* [12] and *in vivo* [10, 9, 11]. Multiple unique stable conformations for the SARS-CoV-2 frameshift pseudoknot have been predicted and observed via various methods including crystallography, cryo-EM, and 3D-physics simulations [9, 15, 12, 102, 37, 142]. Thermal unfolding of RNA found major and minor paths from the folded to unfolded state, concluding that stability of transient (i.e., intermediate) states dictates folding paths [143].

Intense predictions efforts continue to build the structural model proposed for the exact SARS-CoV-2 -1 ribosomal frameshift mechanics. The RNA sequence itself is the primary factor in determining the structure of an RNA, ergo mutations to the frameshift element sequence can disrupt structure-function. Analyses of mutated SARS-CoV-2 RNA sequences and resulting structures find even single nucleotide mutations can substantially reduce frameshift efficiency [26, 110, 104]. Empirical evidence supports the specificity needed for the SARS-CoV-2 frameshift sequence, i.e., that mutations in this structural region are remarkably rare. The most prevalent mutations in the frameshift region (C13536U and C13378U) were each observed in only 0.12% of over $700,000$ sequences most recently recorded via GISAID at the time of this writing [106, 107]. Furthermore, there is some evidence of evolutionary convergence, with the most common mutation C13536U increasing sequence similarity between SARS-CoV-2 and MERS-CoV [103].

To contribute to comprehensive RNA structural knowledge we apply thermodynamic-based minimum free energy (MFE) algorithms to predict *crossing* RNA structures that may regulate ribosome pausing mechanics in betacoronaviruses. Towards resilient SARS-CoV-2 therapeutic treatments, we substantially expand on our previous hierarchical folding investigation of the SARS-CoV-2 68 nt frameshift element sequence [14], by predicting secondary structures for extended length coronavirus frameshift sequences and incorporating sequence covariation information.

4.2.1 Chapter Contributions

First, we sought to detect and utilize coronaviruses sequence covariation by applying *KnotAli* [71], a free energy minimization algorithm that merges conserved evolutionary information within a relaxed hierarchical folding approach. We utilized KnotAli to predict possibly pseudoknotted secondary structures for SARS-CoV-2 and related coronaviruses. We present and discuss our results for inter- and intra-coronavirus RNA structuredness. Predictions via KnotAli showcase evolved conformational flexibility based on detected covariation, and also how mutations change predicted structures for SARS-CoV-2 and related viruses, especially bat coronaviruses. We provide additional context for known covariation and associated base pairing [15] by predicting novel base pairs that possess strong covariation within the multiple sequence alignment.

Second, to explore SARS-CoV-2 frameshift pseudoknot motifs in longer sequences, we applied our hierarchical folding algorithm, *Shapify*, to predict possibly pseudo-

knotted secondary structures while incorporating reactivity data. Following the approach of Schlick et al. [15], which extended to a 222 nt coronavirus frameshift sequence element window, we obtained SARS-CoV-2 frameshift element structure predictions for sequences increasing in length from 90 nt to 222 nt using Shapify, in combination with genome wide *in vivo* and sequence specific *in vitro* SHAPE data. Our predictions allow comparison of energetic stability between known secondary structure motifs [15]. Predictions via Shapify unveil a diverse array of energetically favourable potential pseudoknots for the frameshift element that were previously unknown. Our SHAPE-informed structure prediction analysis includes detailed classification of critical pseudoknot motifs. We report complex pseudoknot predictions both upstream ($5'$) and downstream ($3'$) with respect to the *native* pseudoknot structure (cf. Section 4.3.1), including the traditional *attenuator* hairpin [55]. By extending SHAPE-informed hierarchical-folding structure prediction, our analysis more precisely describes the landscape of energetically favourable RNA structures, facilitating site-specific therapeutic targeting strategies.

4.3 Tying the Knot: Methods

First, we introduce secondary structure motifs [14]. Second, we introduce the hierarchical folding method KnotAli [71] for detection of covariation through alignment and secondary structure prediction informed by covariation. KnotAli is equipped to allow for minor modification to improve prediction accuracy.

4.3.1 SARS-CoV-2 Frameshift Secondary Structure Motifs

SARS-CoV-2 secondary structure prediction efforts find multiple length-dependent structural *motifs* related to the frameshift event [15, 101, 134]. We use *dual graph* nomenclature introduced in [13, 144, 145, 15] to refer to the RNA secondary structure or substructure predicted at or directly $3'$ of the coronavirus frameshift element slippery sequence. A dual graph specifies connectivity and topological aspects of secondary structure by representing each stem by a vertex (cf. Fig 4.1). Dual graph edges between vertices represent junction of stems, or loops. Specifically, each unpaired RNA strand is represented by an edge.

Dual graph representations allow for variable length of stems and loops, which makes RNA secondary structure pattern or motif identification easier. Note that

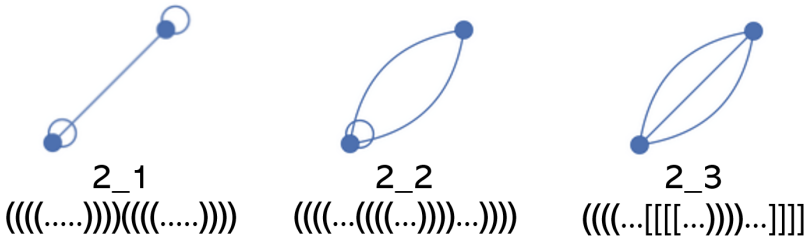


Figure 4.1: RNA dual graph motifs and nomenclature with two vertices. Vertices represent stems. Edges represent junction of stems, or bulge/internal loops with more than one residue on each strand. Self-edges represent hairpin loops. Dual graphs are referred to by two numbers, listed below each respective graph. The first number indicates the number of vertices, the second number specifies the topology, e.g., 2_1 is the dual graph secondary structure motif with two vertices, specifically the first possible topology. For additional details refer to the RNA-As-Graphs database [13]. Dot bracket example structures for each respective motif shown below number labels. Open parentheses/brackets show the base on the 5′ side of the sequence, closed parentheses/brackets represent the base on the 3′ side of the sequence that are binding together. Each period "." represents an unpaired base.

the number of possible topologies rises exponentially with the number of stems, e.g., with three stems, eight possible topologies, with six stems, 508 possible topologies. Therefore, we use dot bracket notation (cf. Fig 4.1) or arc diagrams (cf. Fig 4.2) to represent secondary structure predictions with more than three stems, such as those predicted for the SARS-CoV-2 sequence of length 222 which may have ten or more stems. In arc diagrams, the RNA sequence of bases is shown from left to right (5′ to 3′) in a single horizontal line. Base pairs are represented as arcs between bases.

The *native* frameshift element structure for SARS-CoV-2 [55], also referred to as 3_6 motif [15], is a three-stemmed pseudoknot forming directly downstream of the slippery RNA sequence (cf. Fig 4.2 top arc diagram, pseudoknotted base pairs in red). Structure-function research also finds a frameshift efficiency downregulation mechanism via a simple RNA loop preceding the slippery sequence, i.e., the *attenuator* hairpin, via interaction with the ribosome during translocation or elongation [141, 55] (cf. Fig 4.2, sequence highlighted in pink). Previously, we predicted 3_6 motif structures and structure similarity with SARS-CoV-2 for SARS-CoV-1 and MERS-CoV [14].

Within a hierarchical folding framework for the SARS-CoV-2 frameshift element sequence, the stem occurring directly downstream of the slippery sequence, referred to as stem 1, was identified as the most energetically favourable initial stem (cf. Fig 4.2, dark blue arcs) within a 68 nt window [14]. Likely due to its high relative stability

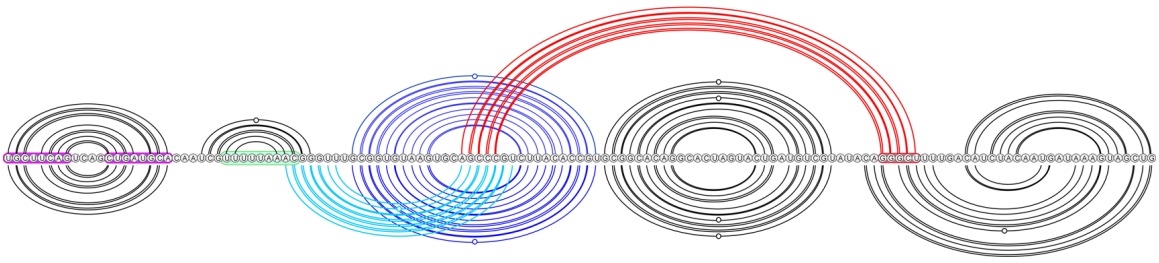


Figure 4.2: Dominant SARS-CoV-2 pseudoknot motif predictions via Shapify. SARS-CoV-2 frameshift element sequence shown as a horizontal line from 5′ (left) to 3′ (right). Arcs represent predicted base pairs. Top arc diagram includes 3_6 motif component (cf. Table 4.2 for dot-bracket format) of the fifth most stable structure predicted via Shapify (cf. Section 4.3.4) for 144 nt sequence, free energy −29.45 kcal/mol, initial stem 5 base pairs in red (free energy −4.22 kcal/mol). Downstream pseudoknot target sequence highlighted in red. Bottom arc diagram includes 3_3 motif component of the MFE structure predicted via Shapify for 144 nt sequence, free energy −30.93 kcal/mol, initial stem 2 base pairs in light blue (free energy −6.1 kcal/mol). Initial stem 1 base pairs in dark blue (free energy −11.48 kcal/mol).

within the energy landscape, stem 1 refolds quickly when unfolded [100]. Stem 1 was also found to be highly conserved among coronaviruses [15], and present in both major structure motifs proposed by Schlick et al. [15] (3_6 for 77 nt window, and 3_3 for 144 nt window). For extended sequence lengths, native stem 1 may not form, instead, upstream base pairing has been identified [133].

Conversely, the second most energetically favourable initial stem for SARS-CoV-2 frameshift element sequence, i.e., stem 2, was found to pair differently depending on window size around the frameshift site considered for prediction. Schlick et al. [15] used secondary structure prediction in combination with SHAPE structural probing and thermodynamic ensemble modeling to identify stem 2 either (A): paired into a simple pseudoknot crossing upstream of stem 1 in the 3_3 motif (95.6% of ensemble at 144 nt, cf. Fig 4.2, light blue arcs), or (B): paired downstream forming the native H-type pseudoknot in the 3_6 motif (98% of ensemble at 77 nt).

4.3.2 Covariation-informed Hierarchical Folding

KnotAli [71] combines the strengths of MFE prediction and alignment-based methods through relaxed hierarchical folding. KnotAli uses a multiple RNA sequence alignment as input to predict possibly pseudoknotted secondary structures for each sequence in the alignment. KnotAli first identifies a set of pseudoknot-free base pairs, based

on mutual information, to guide subsequent free energy minimization. Note that relaxed hierarchical folding indicates predictions can be reached via multiple different paths allowing suboptimal structures, and the initial structure may be modified to accommodate base pairs that lower the free energy of the structure. Output from KnotAli includes base pairs that show strong covariation among the multiple sequence alignment, as well as possibly pseudoknotted predictions for each individual sequence.

4.3.3 Coronavirus Data

We obtained the coronavirus alignment of Schlick et al. [15], where out of 3760 SARS-CoV-2 coronavirus sequences [106], and 2855 other coronavirus sequences [146], 1248 sequences were found to be non-redundant [147]. These 1248 sequences were structurally aligned to the 222 nt SARS-CoV-2 frameshift element SHAPE consensus structure [15] using the Infernal covariance model [148] giving a final result of 182 non-duplicate homologous sites including seven SARS-CoV-2 sequences. Here, we converted the alignment of 182 sequences to FASTA format for input into KnotAli.

For input to Shapify we utilized the reference genome for SARS-CoV-2 from the National Center for Biotechnology Information, $NC\_045512.2$ [2]. We obtained three available SARS-CoV-2 SHAPE datasets to guide Shapify predictions, two *in vivo*: Huston et al. [10], and Yang et al. [11], one *in vitro*: Manfredonia et al. [9].

4.3.4 Shapify Window Procedure

Secondary structure predictions were obtained via Shapify with initial stems and SHAPE data for each SARS-CoV-2 sequence varying in length from 90 nt to 222 nt, with a step size of 1. Therefore, our results are based on removing successive nucleotides from the $5'$ side of the RNA sequence for a total of 132 sequences analyzed (cf. Table 4.1 for window position and length). To compare free energy of structures for different sequence lengths, we divided the energy of each structure by its respective sequence length for an unbiased inter-window comparison, referred to as free energy per nt (for visualization see Fig 4.6):

$$\text{Free energy per nt} = \frac{\text{Structure free energy}}{\text{Sequence length}}$$

Table 4.1: Shortest and longest window sizes used for SARS-CoV-2 structure predictions via Shapify.

Position	Length
$13485 - 13575$	90
$13354 - 13575$	222

4.4 KnotAli and Extended Shapify Results

We first visualize base pairs with strong covariation among the multiple sequence alignment identified by KnotAli. Next, we present secondary structure predictions for coronaviruses frameshift element sequences via KnotAli. Then, we display our Shapify structure investigation applying hierarchical folding to predict SARS-CoV-2 secondary structures for frameshift element sequences up to length 222. All structure predictions and associated data can be found in S2 File.

4.4.1 KnotAli Secondary Structure Predictions

KnotAli finds strong covariation in the multiple sequence alignment for base pairs that align with the known attenuator hairpin [55], and also the native 3_6 motif (cf. Fig 4.3).

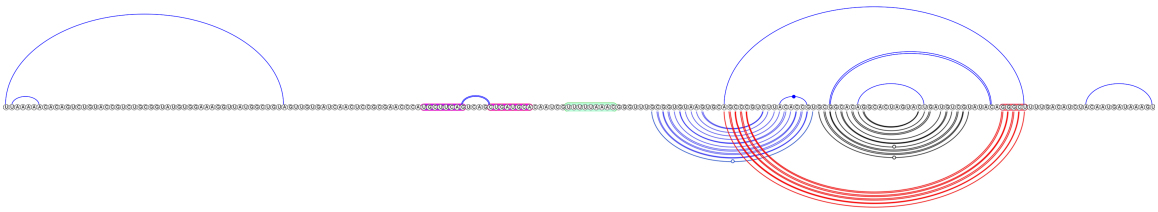


Figure 4.3: Coronavirus frameshift element covariation. Base pairs in the top arc diagram have strong covariation among the multiple sequence alignment identified by KnotAli. Bottom arc diagram displays the SARS-CoV-2 native 3_6 psuedoknot, downstream target sequence in red. SARS-CoV-2 attenuator hairpin sequence highlighted in pink, slippery sequence in green.

Among the seven SARS-CoV-2 sequences in the alignment, structure predictions for five sequences included the 3_3 motif, while two sequences (EPI_ISL_465643, EPI_ISL_426088) resulted in a prediction that included the native 3_6 motif instead (cf. Fig 4.4). A structure containing the 3_6 motif was also predicted for BtRf-BetaCoV (cf. Fig 4.5). BtRf-BetaCoV has 77% overall identity with SARS-CoV-2, and 91%

frameshift sequence identity (FSID) with SARS-CoV-2 as identified by Schlick et al. [15].

The majority of sarbecovirus secondary structure predictions include the 3_3 motif: Pangolin-CoV (98% FSID), SARS-CoV-1 (93% FSID), SARS-like WIV1-CoV (93% FSID, cf. Fig 4.5), BtRs-BetaCoV (92% FSID), Bat-Cov-Cp (91% FSID), and Bat-CoV-Rp (91% FSID).

Sarbecovirus predictions demonstrate significant structuredness. Pseudoknots were predicted upstream of the native frameshift pseudoknot site for SARS-CoV-2, SARS-CoV-1, SARS-like WIV1-CoV, BtRs-BetaCoV, BtRf-BetaCoV, and Bat-CoV-Cp. Pseudoknots were predicted downstream of the native frameshift pseudoknot site for Pangolin-CoV, SARS-like WIV1-CoV, BtRf-BetaCoV, and Bat-HpBetaCoV.

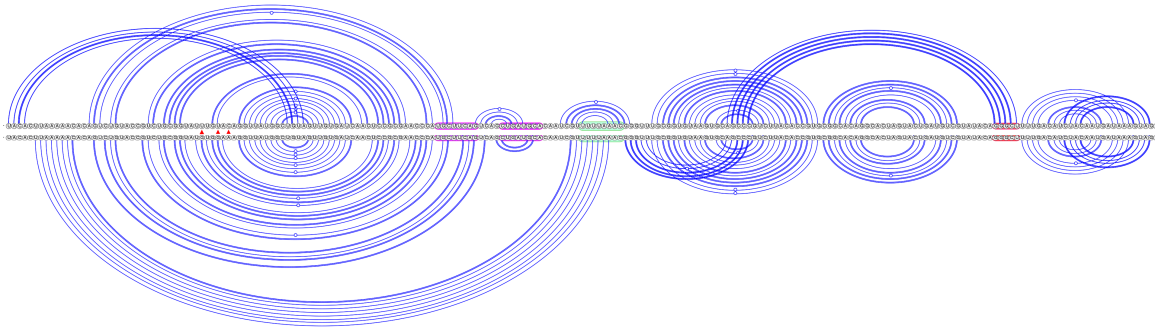


Figure 4.4: SARS-CoV-2 secondary structure predictions via KnotAli. Top arc diagram: free energy -36.47 kcal/mol, EPI_ISL_426088, includes 3_6 motif. Bottom arc diagram: free energy -40.65 kcal/mol, EPI_ISL_426905, includes 3_3 motif. Mutations indicated with a red $\triangle$ symbol between sequences. Attenuator hairpin sequence highlighted in pink, slippery sequence in green, downstream native pseudoknot target sequence in red.

4.4.2 Shapify Extended Length Secondary Structure Predictions

Following the procedure outline in Section 4.3.4, we obtained a total of $10,916$ secondary structure predictions via Shapify (cf. Fig 4.6) for the SARS-CoV-2 frameshift element window of varying length. We observe that the most stable secondary substructures (including only up to three stems at or directly $3'$ of the slippery sequence) can be classified into four pseudoknot dual-graph motifs: 2_3, 3_3, 3_6 (native-type), and 3_8 (cf. Fig 4.2, Fig 4.6, and Table 4.2).

Table 4.2: Secondary structure pseudoknot motifs in dot bracket notation. Note that motif classification allows minimal modification to structures, e.g., in the size of loops. Sequence ID: $NC\_045512.2$ [2], indices $13467 - 13565$. Open parentheses/brackets show the base on the $5'$ side of the sequence, closed parentheses/brackets represent the base on the $3'$ side of the sequence that are binding together. Each period "." represents an unpaired base.

Motif	Secondary Structure
2_3	.....(((((((.[[.[[[.[[[[[[)))))))).....]]]]]..]]]]]
3_3	[[[[[[[.((((((((((...]]]]]]]))))))))))(((((((((.....))).)))))).....
3_6	........((((((((((.(([[[[[)))))))))))(((((((((((....)).))).)))))).....]]]]].....
3_8	..............((((((((((.....[[[[[[[[[[)))))....]]].]]]]]]]]].]]]]].....
Sequence	ACGGGUUUGCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUAUACAGGGCUUUUGACAUCUACAAUGAUAAAGUAGCUGGU

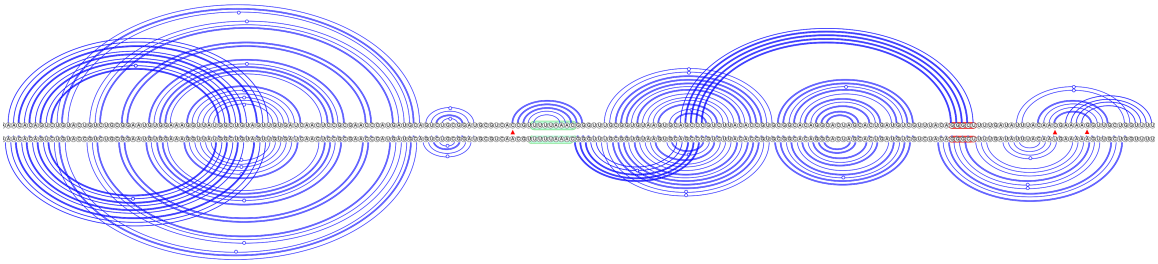


Figure 4.5: Bat coronaviruses secondary structure predictions via KnotAli. Top arc diagram: BtRf-BetaCov, free energy −43.24 kcal/mol, KJ473811, includes 3_6 motif. Bottom arc diagram: SARS-like WIV1-CoV, free energy −39.72 kcal/mol, KU444582, includes 3_3 motif. Mutations indicated with a red △ symbol between sequences. Slippery sequence in green, downstream native pseudoknot target sequence in red.

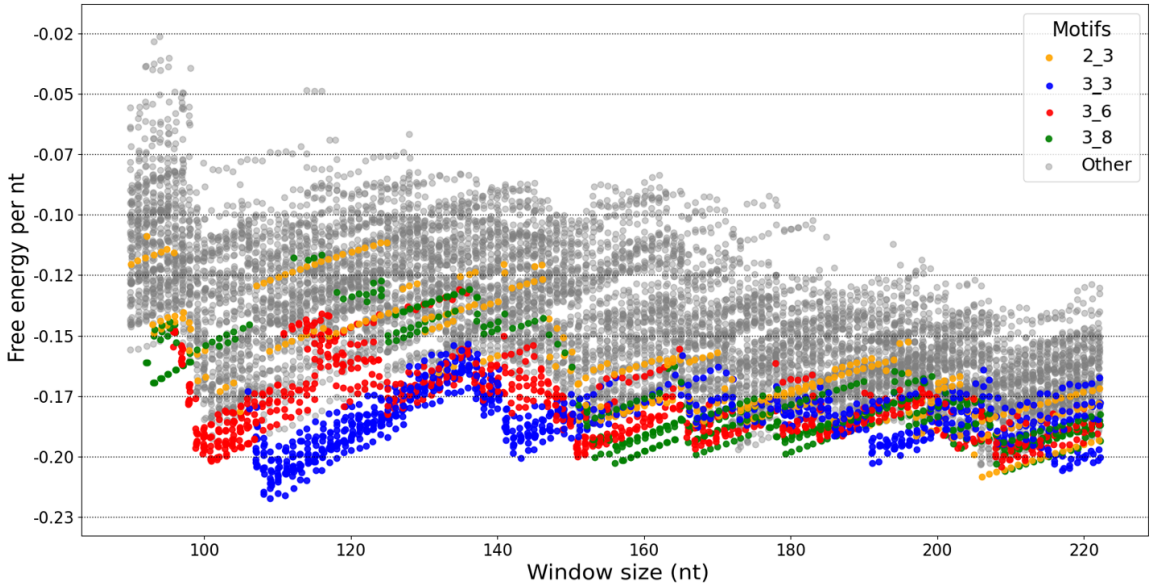


Figure 4.6: SARS-CoV-2 secondary structure motifs free energy per nt. Each point represents a Shapify predicted secondary structure for the SARS-CoV-2 frameshift sequence (cf. Section 4.3.4, Table 4.1). X-axis represents window size, y-axis represents free energy per nt. Dots are colored based on the four listed dual-graph motifs (legend in top-right) detected at or directly 3′ of the slippery sequence (cf. Table 4.2), or grey for none.

Our predictions confirm that initial stem 1 folds into the 3_3 motif, as part of the complete MFE structure, for three specific window size intervals: $107-151$, $191-199$, and $216-222$ (cf. Fig 4.6). There was path convergence, meaning multiple initial stems resulted in different predictions that each contain the 3_3 motif (cf. Fig 4.7). For example, with the 144 nt sequence as input, the MFE structure and the next four

most stable structures all include the 3_3 motif. At the critical window sizes $98 - 106$, which include the region directly preceding the slippery sequence, the 3_3 motif is sufficiently destabilized leading to the MFE structure including the 3_6 motif instead. With window sizes $152 - 153$, the MFE structure again includes the 3_6 motif as a key component. At the shortest length, as the 3_6 motif is destabilized, a 3_8 motif is predicted within the MFE structure for window sizes $93 - 97$.

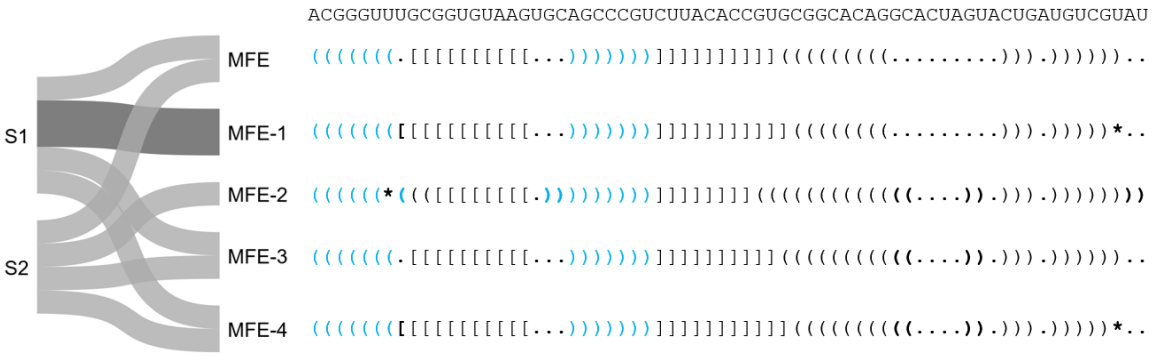


Figure 4.7: Convergence to the most stable structures that contain the 3_3 motif. 144 nt window Shapify predictions with initial stems 1 and 2 as constraint result in the MFE and most stable structures that contain the 3_3 motif. Initial stems labeled on the left (e.g., S1 for initial stem 1). Darker grey path indicates the structure predicted with a specific initial stem was the same for two SHAPE datasets. Light grey path indicates the structure predicted with an initial stem was specific to one SHAPE dataset. Structures on the right are labeled by energy proximity to the MFE structure, e.g. MFE-1 is the lowest free energy structure after the MFE structure. Differences from the MFE structure are marked in bold, with parenthesis representing changes in paired bases, and asterisks representing predicted unpaired bases that were paired in the MFE structure. 3_3 motif pseudoknotted base pairs shown in light blue.

Predictions via Shapify unveil changes in energetic favourability of pseudoknotted motifs for extended length sequences. We find the 3_8 motif predicted within the MFE structure for window sizes of $154 - 172$ and $179 - 185$. In addition, a 2_3 motif is predicted within the MFE structure for window sizes $205 - 208$. The 2_3 motif is energetically close to the 3_8 motif at window sizes $209 - 215$. Beyond local pseudoknotted motifs, we detect additional stable pseudoknotted regions suggesting possible upstream and downstream structure-function of the frameshift element (cf. Fig 4.8,4.9 and 4.10).

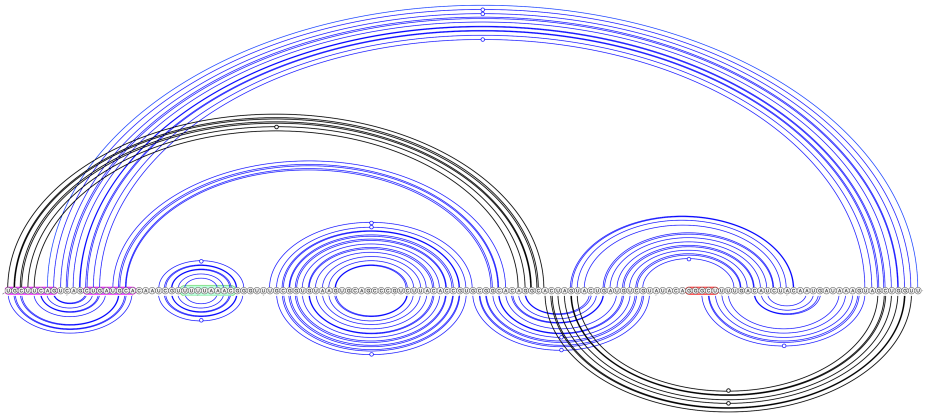


Figure 4.8: Structural regions involving pseudoknots in SARS-CoV-2, 144 nt window via Shapify. Top arc diagram: MFE-20, free energy −24.29 kcal/mol, initial stem 18 in black (free energy −0.35 kcal/mol). Bottom arc diagram: MFE-12, free energy −25.76 kcal/mol, initial stem 11 in black (free energy −1.44 kcal/mol). Attenuator hairpin sequence in pink, slippery sequence in green, downstream native pseudoknot pairing region in red.

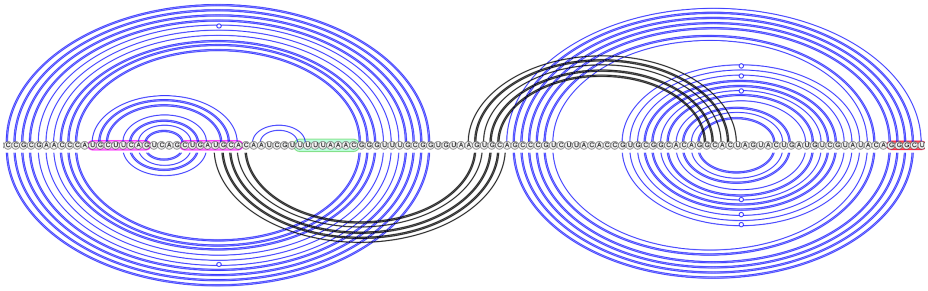


Figure 4.9: SARS-CoV-2 pseudoknot predictions overlap, 222 nt window via Shapify. Top arc diagram: MFE-5, free energy −45.04 kcal/mol, initial stem 15 in black (free energy −2.74 kcal/mol). Note that the MFE-5 pseudoknot was also detected within the 144 nt window (cf. MFE-19) and the 68 nt window [14]. Bottom arc diagram: MFE-29, free energy −39.34 kcal/mol, initial stem 12 in black (free energy −3.41 kcal/mol). Attenuator hairpin sequence in pink, slippery sequence in green, downstream native pseudoknot pairing region in red.

4.5 Chapter Discussion

Fully understanding functional RNA structures remains an elusive but worthwhile goal, and is a necessary step towards effective therapeutics for viral infections. Despite many attempts to distill how coronavirus RNA folds to efficiently regulate a frameshift event, much is still unknown about the mechanism. We employed two hierarchical folding free energy minimization algorithms, KnotAli, and Shapify, for prediction of

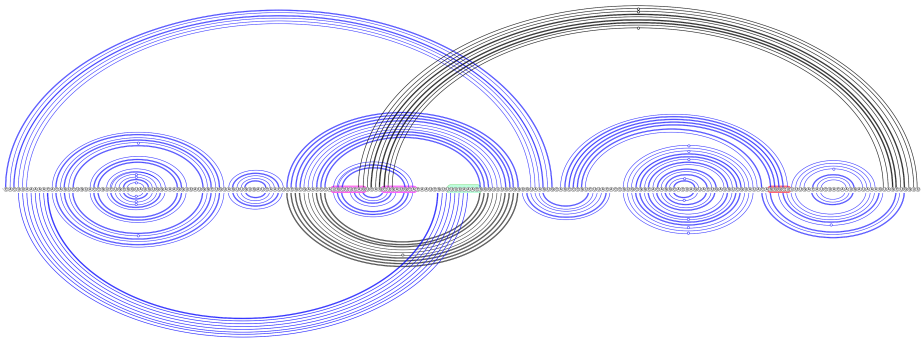


Figure 4.10: SARS-CoV-2 long range pseudoknot predictions, 222 nt window via Shapify. Top arc diagram: MFE-58, free energy −34.16 kcal/mol, initial stem 16 in black (free energy −1.98 kcal/mol). Bottom arc diagram: MFE-10, free energy −42.74 kcal/mol, initial stem 2 in black (free energy −10.08 kcal/mol). Attenuator hairpin sequence in pink, slippery sequence in green, downstream native pseudoknot pairing region in red.

possibly pseudoknotted structures of SARS-CoV-2. Next, we discuss and contextualize our predictions of the coronavirus frameshift element to advance known structure information towards site-specific viral therapeutics, namely: (1) secondary structure predictions for SARS-CoV-2 and bat coronaviruses frameshift elements via KnotAli, and (2) insights from SHAPE-informed hierarchical folding predictions for SARS-CoV-2 via Shapify.

4.5.1 Coronaviruses Frameshift Element Covariation

With regard to the base pairs detected by KnotAli that show strong covariation among the multiple sequence alignment, we first note the innermost base pair of the traditional attenuator hairpin stem-loop (GC base pair at 13442.13447). Covariation of this innermost base pair, along with the attenuator hairpin being functionally retained in SARS-CoV-1 and SARS-CoV-2 despite sequence differences [55], suggests the loop may be structurally conserved. Compensatory mutations related to the GC base pair at 13442.13447 include a potential AU base pair at 13442.13447 in transmissible gastroenteritis virus (TGEV, 59% FSID), porcine respiratory coronavirus (PRCV, 60% FSID), and turkey coronavirus (Turkey-CoV, 60% FSID), also a potential UA base pair at 13442.13447 in Night heron coronavirus (Night-Heron-CoV, 58% FSID). In addition, the outermost base pair of the pseudoknotted stem in the native 3_6 motif structure is found to have strong covariation, supporting previous results of Schlick et al. [15].

A two-branched multiloop directly downstream of the slippery sequence is identified as having strong covariation by KnotAli. The larger branch forms a bulge loop via a GC base pair at 13507.13536. Covariation of the GC base pair at 13507.13536 is supported by possible GU base pair at 13507.13536 in Pangolin-CoV, and also possible AU base pair at 13507.13536 in TGEV, Night-Heron-CoV, PRCV, Feline-CoV (61% FSID), and Canine-CoV (58% FSID). Further support is seen via possible UG base pair at 13507.13536 in infectious bronchitis virus (IBV, 60% FSID) and Turkey-CoV. Covariation of this bulge loop inner base pair, AU at 13512.13524, is supported by possible UA base pair at 13512.13524 in TGEV, PRCV, Feline-Cov, Canine-CoV and BtSk-Alpha-CoV (61% FSID), and also possible GU base pair at 13512.13524 in IBV and Turkey-CoV. In addition, the AU base pair at 13512.13524 is predicted via Shapify in multiple different SARS-CoV-2 pseudoknots including the 3_3, 3_6, and 3_8 motifs (cf. Table 4.2) and longer range predictions (cf. Fig 4.9 and 4.10).

With respect to the native frameshift pseudoknot (cf. Fig 4.3), there are two base pairs with strong covariation predicted upstream, and one base pair with strong covariation predicted downstream. Previously, Schlick et al. [15] identified an upstream AU base pair at 13366.13411 in SARS-CoV-2, Pangolin-CoV, Bat-CoV-Rp, BtRs-BetaCoV, SARS-like WIV-1-CoV, SARS-CoV, BtRf-BetaCoV, and Bat-CoV-Cp. Their covariance analysis highlighted compensatory mutations via a CG base pair at 13366.13411 in TGEV and PRCV, a UG base pair at 13366.13411 in Feline-CoV, and a UA base pair at 13366.13411 in Rousettus-Bat-Cov (64% FSID). KnotAli finds strong covariation in the multiple sequence alignment to support a UA base pair at 13360.13366, with A13366 pairing upstream with U13360 as opposed to the previously identified AU base pair at 13366.13411. This UA base pair at 13360.13366 is supported by compensatory mutations via possible GU base pair at 13360.13366 in TGEV, PRCV, Feline-CoV and Canine-Cov. Finally, the additional upstream base pair predicted via KnotAli, UA at 13359.13410, is supported by possible GC base pair at 13359.13410 in IBV, Rousettus Bat-CoV, and Turkey-CoV. KnotAli also finds strong covariation in a downstream AU base pair at 13553.13565, supported by possible GC base pair at 13553.13565 in IBV and Turkey-CoV.

These novel predicted base pairings, supported by coronavirus multiple sequence alignment covariation, provide additional context to previously identified covariation in coronaviruses [15]. As a constraint guiding secondary structure prediction via KnotAli, we note the covariation-informed initial structure is preserved in certain individual predictions for coronaviruses, but in other cases more stable structures can be reached

by disrupting these base pair(s). We observe that minimal upstream mutations to SARS-CoV-2 sequences lead to either the 3_3 motif or 3_6 motif predicted via KnotAli (cf. Fig 4.4). In addition, minor changes between bat coronaviruses like BtRf-BetaCov and SARS-like WIV1-Cov not only led to different motif predictions, but also affect predicted downstream pseudoknot pairings (cf. Fig 4.5). We conclude that minimal upstream and downstream sequence variation can significantly change conformations of the frameshift element. In supporting resilient frameshift therapeutics, it could be valuable to assess the viability of treatments intended for SARS-CoV-2 across various bat coronaviruses sharing similar sequences and structures. Such investigations could provide more detailed insights into how the frameshift regulation mechanism may be disrupted.

4.5.2 Extended SARS-CoV-2 Predictions via Shapify

Our secondary structure predictions for the SARS-CoV-2 frameshift sequence via Shapify offer insight into frameshift dynamics. By considering a SARS-CoV-2 sequence of varying length for prediction, we simulated how the interaction of the ribosome with the frameshift element [97] affects RNA structural motifs of SARS-CoV-2. In doing so we have characterized the landscape of possibly pseudoknotted structures, finding the 3_8 motif to be the most energetically favourable at specific sequence lengths, while also clarifying interplay between the dominant 3_3 and 3_6 native-type -1 PRF structures.

Irrespective of length, structures containing the 3_3 motif reached the lowest free energy per nt (nearly -0.23, cf. Fig 4.6). Our SHAPE-informed hierarchical folding also demonstrates innate resiliency, i.e., redundant folding paths from different initial stems all leading to the 3_3 motif structure. This extends previous results finding similar path convergence to the 3_6 motif structure within a smaller 68 nt window [14].

Secondary structure predictions demonstrate that as shorter sequence length destabilizes the 3_3 motif, the 3_6 native-type motif seamlessly emerges to dominate the ensemble, suggesting a transition between the two motifs. We observe the shift from 3_3 to 3_6 as occurring at SARS-CoV-2 sequence index 13468, when the second of three successive guanine nucleotides that would otherwise form the 3_3 motif stem is removed from the sequence under consideration. This location for a potential secondary structure transition was also identified within a partition function framework for the SARS-CoV-2 frameshift element [149]. Beyond these two major motifs, additional

pseudoknot patterns are needed to fully describe the MFE structures of the ensemble, especially the 2_3 and 3_8 motifs.

The 3_8 motif has been previously detected as part of the SARS-CoV-2 frameshift element by *in vivo* probing and subsequent folding analysis within a 126 nt window [10]. It was also predicted computationally via hierarchical folding informed by SHAPE data as soft-constraint within a 68 nt window [14]. Observing that the 3_8 motif is the most energetically favourable structure for shorter length sequences, we hypothesize it may act as a transient structure facilitating refolding of either the 3_3 or 3_6 motif. In further support of a link between these three motifs is that they share an adenine bulge (A13524), which has long been identified as critical in frameshift regulation [111].

Further structure analysis is needed to understand function of the 2_3 motif, which is known to be prevalent in viral frameshift elements, especially plant viruses, but also simian retrovirus type-1, and mouse mammary tumor virus [145]. We hypothesize this downstream pseudoknot folding may have some effect on translation termination of the ribosome, which has been linked to frameshift regulation [89].

Our novel predictions suggest function of the attenuator hairpin via previously unknown pseudoknotted base pairing, paving the way for future tertiary modeling of SARS-CoV-2 frameshift RNA structure-function. Attenuator hairpin alternative pairings which include the slippery sequence and initial stem 1 have also been proposed [133]. We predict for the first time attenuator hairpin bases folding into long range pseudoknotted interactions (cf. Fig 4.8, top arc diagram, and Fig 4.10, top arc diagram partially supported by IPKnot structure prediction [32, 15]), and also a potential H-type pseudoknot structure with the stem-loop conserved as part of a pseudoknotted multiloop (Fig 4.9, bottom arc diagram). Notably, we visualized multiple structures that possess significant overlap or regions of structural similarity. While certain stems overlap, other pseudoknotted base pairs form which indicate potential conformational switching between the two structures.

The latest SARS-CoV-2 tertiary structure experimental results required molecular simulation of additional folding to reach higher agreement with previous experimentally derived molecular structures [142]. Understanding how initial most stable stems give way to pseudoknotted motifs via unfolding and refolding may unlock the key to discerning kinetic trajectories, which are currently not well defined. With regard to the most stable initial stems obtained via HotKnots [5], in general we find initial stem 1, i.e, stem 1 (free energy -11.48 kcal/mol, cf. Fig 4.2, dark blue base pairs) is the lowest free energy stem. This confirms previous results finding initial stem 1 to

be most stable within a 68 nt window [14]. Other initial stems can change in their relative energetic stability, e.g., initial stem 2 (free energy −6.1 kcal/mol, cf. Fig 4.2, light blue base pairs) is the second lowest free energy stem for the 144 nt window, but this is supplanted in the 222 nt window by an additional highly stable upstream stem. We hypothesize a possible frameshift downregulation function for this additional highly stable stem, because in our analysis no folding path was detected to the most stable secondary structures. Overall, further study is needed to fully understand the mechanics of the folding from initial stems into secondary structures, especially in pseudoknots.

In the folding process of the SARS-CoV-2 frameshift pseudoknot, the native pseudoknotted stem likely folds last [100]. Hence, we suggest exploring site-specific therapeutic targeting of the downstream native pseudoknot pairing. Since this pseudoknot critically refolds to initiate the frameshift, it should be an accessible location to disrupt the frameshift pseudoknot. In comparing the site target potential of an overlapping $5'$ site that did not include the pseudoknot target, oligonucleotide targeting found the downstream native pseudoknot pairing to be a more effective target for reducing frameshift efficiency [150]. The downstream native pseudoknot site may be even more effective than previously realized, because it includes key structure pairings of the 3_8 and 2_3 motifs, which are each predicted within the MFE structure at specific sequence lengths. For more comprehensive and broadly applicable therapeutics, upstream sites, like the attenuator hairpin, should also be considered in subsequent work. Further exploration of the intricate relationship between RNA structure and function is needed in the field of coronavirus therapeutics, towards more positive outcomes for human and animal health.

To investigate the ensemble of frameshift pseudoknot structures in SARS-CoV-2, next we design and implement the first conditional partition function algorithm for pseudoknotted RNA secondary structures: CParty.

Chapter 5

CParty: Conditional Partition Function for Density-2 RNA Pseudoknots

RNA molecules fold into biologically important functional structures. Efficient dynamic programming RNA (secondary) structure prediction algorithms restrict 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 provide the theoretical framework and implementation for our algorithm, *CParty*, that follows the hierarchical folding hypothesis, i.e., the biophysically 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 structure G'. Notably, this can enable sampling from the hierarchical distribution $P(G'|G)$. With CParty we develop for the first time an *unambiguous* scheme based on *HFold*, i.e., the minimum free energy hierarchical folding algorithm based on a realistic pseudoknot energy model. Thus, we develop the first partition function variant for density-2 structures. Compared to the only other available pseudoknot partition function algorithm, which covers simple pseudoknots (and follows a different strategy, mapped from a pure minimum free energy structure prediction), 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 an efficient, cubic time, algorithm for the stochastic analysis of pseudoknotted RNAs, which enables novel applications. We discuss one such application, i.e., how the CPF for a pseudoknotted therapeutic target in SARS-CoV-2 can provide insight into RNA structure formation.

5.1 Introducing CParty

Recall the two widely used thermodynamic approaches for predicting non-crossing (pseudoknot-free) RNA secondary structure by Zuker [65] and McCaskill [66]. The Zuker algorithm uses DP to recursively compute the free energy of all possible RNA substructures and find the most energetically stable, the MFE structure [65].

Adopting a stochastic view of RNA, the McCaskill algorithm computes the *partition function* that describes the pseudoknot-free 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*. The partition function is useful for predicting the relative probability of any structure as well as base pair probabilities [68, 66]. 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.

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:*

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

$E_{2D}(G)$ 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 [91], 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. In this work, we develop and implement the algorithmic framework for calculating the CPF: *CParty*.

There is a one-to-one correspondence between the search spaces considered by MFE algorithms and McCaskill partition function methods [66], provided the underlying DP scheme is unambiguous and complete [151]. This relationship only holds if each substructure is unambiguously computed a single time during the DP MFE calculations [66], i.e., redundant recurrence relation schemes are incompatible with the partition function approach.

Examples of pseudoknotted MFE predictions methods that can be extended to the partition function approach include the most general algorithm PKnots [64], with $O(N^6)$ time complexity, also NUPACK [152] and CCJ [153] both with $O(N^5)$ time complexity, the latter allows for more complex structures like four interleaved *stems* (consecutive adjacent base pairs). 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 [59, 60], therefore pseudoknotted partition function methods require assumptions that limit the complexity of predicted pseudoknots [62]. Heuristic methods or methods that rely on approximations, such as HotKnots V2.0 [4] and IPKnots [32], cannot be extended to compute partition function, although both methods handle a broad class of pseudoknotted structures efficiently.

Here, we develop an unambiguous recurrence relation scheme based on the hierarchical-folding algorithm *HFold* [3], 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* structure $G \cup G'$ that is the MFE structure given G (see [3] for formal definitions, for intuition cf. Fig 2.2-Fig 2.3). 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, and is larger and more comprehensive than the structure class described by the partition function algorithm from Dirks and Pierce [62]. HFold has a time complexity of $O(N^3)$, significantly improving over PKnots, CCJ, and NUPACK.

The HFold recurrence relation scheme is redundant in decomposing density-2 structures for a given non-crossing structure G, making it incompatible with a partition function approach. Our novel algorithm, *CParty*, unambiguously 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. We note that choosing input structure G is of particular importance, since it is impossible to reach the MFE structure when the fixed structure G is not part of the MFE structure. Previous work [52, 14] identified promising techniques to reach the MFE structure by selecting energetically favourable initial stems, i.e., most stable non-crossing structures, or choosing stems compatible with chemical modification data (such as SHAPE reactivity).

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 cases where partial information is available for a given sequence. Different structure reactivity datasets may align for specific subparts of a functional structure, e.g., initial stems, while the rest of the structure is unknown [14]. As functional RNA structures are solved at higher resolution [37], significant differences remain in predictions (e.g., length dependent conformational flexibility [101]). Conditional base pair probabilities obtained via CParty implementation can enable analyses exploring key hierarchical folding paths.

CParty Software Availability CParty is available at: `https://github.com/ltrinity/CParty`.

Contributions. We introduce the CPF algorithm CParty and demonstrate that our algorithm completely and unambiguously decomposes the density-2 structure class. CParty can be utilized to 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 structure 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 can enable further statistical analysis for the ensemble of density-2 secondary structures. We present an immediate application for the algorithm: analysis of the SARS-CoV-2 frameshift pseudoknot structure [55]. CParty provides the framework for additional functional structure path analyses, e.g., computing conditional base pairing probabilities towards therapeutic development in coronaviruses.

Outline. In Section 5.2, we define the CPF problem and relevant variants. Then, we provide a brief notation review. In Section 5.3, we proceed with a high level description of CParty and define the constrained partition subfunctions, proving correctness of the DP scheme. In Section 5.4 we provide details specific to implementation, validation, and testing of CParty. In Section 5.5, we discuss and contextualize how CParty can inform RNA structure formation analysis, providing an application to the SARS-CoV-2 frameshift pseudoknot sequence.

5.2 Preliminaries and Problem Statement

We define energy $E(G)$ of an RNA secondary structure G according to the realistic pseudoknot energy model of [62] with parameters from [4] (cf. Table 5.1). That is, $F_{pseudoknot} = F_{pseudoloop} + \sum F_{band}$, where $F_{pseudoknot}$ is the free energy of a pseudoknot, $F_{pseudoloop}$ is the free energy of a pseudoloop, F_{band} is the free energy of a band, and the summation is over all bands in the pseudoknot [154]. Note there are different penalties if the pseudoknot is inside a multiloop or another pseudoknot. 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* (definition below). Equivalently: $\Pr[G'|G] = \frac{B(E(G' \cup G|G))}{Z}$, where Z is computed over all choices of G' for a given fixed structure G.

An interesting variant of the CPF problem forces all base pairs of G' to cross those of G. We expect this 'pseudoknot-only' variant of CParty can be derived relatively easily from our work.

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 are 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.

5.3 Efficient Computation of the Conditional Partition Function

HFold review. Recall that density-2 structures are a subset of the *bi-secondary structure* class, which is the union of two disjoint pseudoknot-free secondary structures [155], e.g., $G \cup G'$. In the HFold [3] hierarchical folding algorithm, $W(i, j)$ denotes the energy of the MFE density-2 secondary structure for a given subsequence $s_i s_{i+1} \dots s_j$ and a given pseudoknot-free structure $G_{i,j}$ taken over all choices of pseudoknot-free $G'_{i,j}$ which is disjoint from $G_{i,j}$, 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 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 5.2). We include intuition for runtime analysis of CParty structure classes, for formal proof refer to HFold [3].

We now introduce the CParty DP scheme in detail, conditions for terms in summa-

tion are shown in blue. CParty utilizes the DP09 energy model of HotKnots V2.0 [4], with parameters matching the pseudoknotted energy model of Dirks and Pierce [62] that were re-estimated using available pseudoknotted structures (cf. Table 5.1).

5.3.1 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 as $\overline{isCovered(G,i)}$ and $\overline{isCovered(G,j)}$. Note that $Z_W(i,j)$ is weakly closed and can possibly be partitioned into two independent substructures. 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.

$Z_V(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed or empty and $i.j$. $Z_P(i,j)$ is the partition function over all pseudoknotted density-2 structures between i and j where the structure is closed (cf. Section 5.3.2).

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

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.

Since Z_V handles pseudoknot-free loops, its time and space complexity are $O(N^3)$ and $O(N^2)$, respectively, where n is the length of the input sequence. Time and space complexity of Z_W, therefore, depend on the time and space complexity of Z_P, which

Table 5.1: ENERGY PARAMETERS. All parameters were derived at 37 degrees celsius and 1 M salt (NaCl) concentration or extrapolated from experimental values cf. [3, 4, 5].

Name	*Description*	*Value (kcal/mol)*
P_s	Exterior pseudoloop initiation penalty	-1.38
P_{sm}	Penalty for introducing pseudoknot inside a multiloop	10.07
P_{sp}	Penalty for introduce pseudoknot inside a pseudoloop	15.00
P_b	Band penalty	2.46
P_{up}	Penalty for unpaired base in a pseudoloop	0.06
P_{ps}	Penalty for closed subregion inside a pseudoloop	0.96
$e_H(i,j)$	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	$0.89 \times e_S(i,j)$
$e_{int}(i,r,r',j)$	Energy of a pseudoknot-free internal loop	
$e_{intP}(i,r,r',j)$	Energy of an internal loop that spans a band	$0.74 \times e_{int}(i,d,e,j)$
a	Multiloop initiation penalty	3.39
b	Multiloop base pair penalty	0.03
c	Penalty for unpaired base in a multiloop	0.02
a'	Penalty for introducing a multiloop that spans a band	3.41
b'	Base pair penalty for a multiloop that spans a band	0.56
c'	Penalty for unpaired base in a multiloop that spans a band	0.12

Table 5.2: CParty Structure Classes. The first column denotes each structure class abbreviation with description in second column. In the third column we delineate additional structure classes that need to be computed, e.g., W is computed via V where $i.j$, via P where i and j form a pseudoloop, and also by recursive call again to W.

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 rightmost band in G'	PG', PG'^{w}, VP, BE, WI
PG'^{w}	Pseudoknotted 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 stem pseudoknot-free	WM^{1}, V
WM^{P}	$i.j$ on a multiloop, terminal stem pseudoknotted	WM^{P}, P
BE	$[i, i'] \cup [bp(i'), bp(i)]$ band region	BE, WI'

handles pseudoloops.

Z_W fully decomposes the density-2 structure class (cf. Fig 5.1), 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 of a generalized induction step justifying that all cases are disjoint and unambiguous.

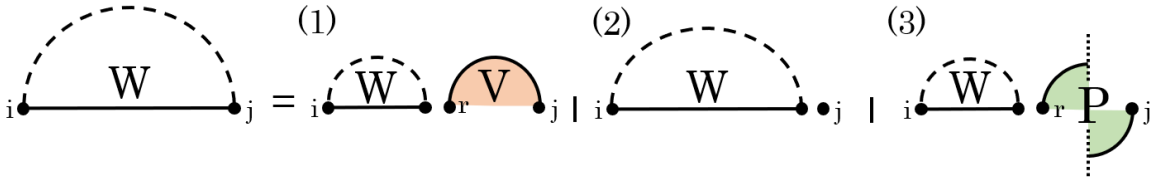


Figure 5.1: $Z_W(i,j)$ structure class. Case (1) illustrating $r.j$ close a loop. Case (2) enforces j unpaired. Finally for Case (3) $r.j$ form a pseudoloop. Dashed arcs indicate possible structure, each solid arc represents a base pair. The dotted vertical line indicates an overlapping chain of bands can continue and that the chain can begin or end via either G (above horizontal line) or G' (below horizontal line). Filled in circles show regions covered by specific structure classes, orange for Z_V, and green for Z_P.

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

$Z_W(i,j)$ is the partition function over 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).

By definition of density-2, j either pairs with k, $i \leq k < j$ such that $k.j$ closes a pseudoknot-free loop (1), j is unpaired (2), or j is the rightmost end of a chain of crossing base pairs (3), cf. Eq. i. These cases are disjoint, i.e., if j is unpaired it cannot also be paired; additionally, if j is paired and closes a pseudoknot-free loop, it cannot also be paired in the rightmost band of a pseudoloop. Therefore, all three cases are disjoint, and the recurrence is unambiguous. Since every density-2 structure falls into one of these three cases, the $Z_W(i,j)$ recurrence 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 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 unpaired, $Z_W(i,j) = Z_W(i,j-1)$ by $Z_W(2)$. 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 $Z_W(3)$. ■

5.3.2 Pseudoknotted Density-2 Structures

$Z_P(i,j)$ is the partition function over $[i,j]$ when $[i,j]$ is a pseudoknotted closed region (cannot be further partitioned), containing a chain of two or more successively overlapping bands that must alternate between $G_{i,j}$ and $G'_{i,j}$. The rightmost band of the pseudoloop may be in G or G', possibly with nested substructures interspersed throughout.

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 5.2), 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 (e.g., $b_{(i,l)}$ or $b'_{(i,l)}$ cf. [3]). If j forms a pseudoloop with i and is not paired in G, we move to $Z_{PG'}$ Case (2), partition function over pseudoknotted density-2 structures with rightmost band in G'. Here we note these two cases are disjoint and Z_P is unambiguous, with j either paired in G or G'.

Since in the region $[i,j]$ for Z_P Case (1) we search over all possible choices of l, time complexity is $O(N^3)$. Therefore, time complexity of Z_P will be maximum of $O(N^3)$ and time complexity of $Z_{PG'}$.

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

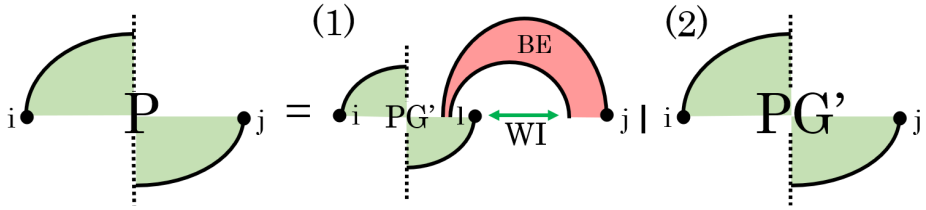


Figure 5.2: Cases of Z_P. (1) j is paired in G and there must be some base, l, between $bp(j).j$ that is paired in G'. (2) j is not paired in G, then move directly to $Z_{PG'}$. Filled in circles show regions covered by specific structure classes, red for Z_{BE}, and green for Z_P and $Z_{PG'}$.

5.3.3 Pseudoknotted Structures with Rightmost Band in G'

When i and j form a pseudoloop where the rightmost band of the pseudoloop is not in G, i.e., $bp_G(j) = 0$, it must be part of the G' structure.

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 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 5.3) 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 [3]. $Z_{PG'}$ Cases $(3-4)$ are end cases, where only one or two bands, respectively, need to be accounted for. For the base case where $i \geq j$, $Z_{PG'} = Z_{PG'^w} = 0$.

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

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

Time complexity for both $Z_{PG'}$ and $Z_{PG'^W}$ is $O(N^3)$, as in both cases we search over all values of l for a given region $[i,j]$.

5.3.4 $i.j$ 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 5.4):

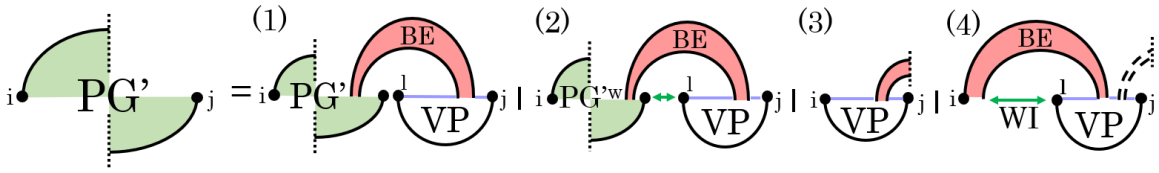


Figure 5.3: Cases of $Z_{PG'}$. (1) handles two rightmost elements of the chain and continues. (2) is similar to case (1) except there is a weakly closed region between the bands, this will be handled by $Z_{PG'^w}$ structure class to preserve the cubic time complexity. For the end cases we have (3) leftmost band of chain in G'; and (4) leftmost band in G. Filled in circles show regions covered by specific structure classes, green for $Z_{PG'^w}$. Colored lines correspond with structure classes that may or may not have any substructures: Z_{WI} in green, and purple for Z_{VP}. Dashed arcs indicate possible structure, each solid arc represents a base pair.

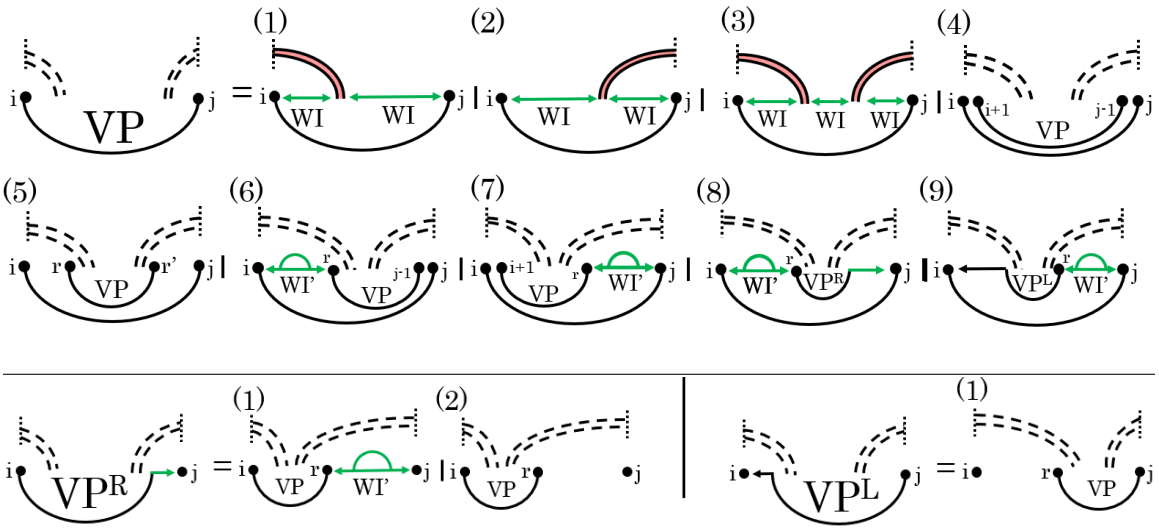


Figure 5.4: Cases of VP, VP^R, and VP^L. Top: VP $(1-3)$ either two or three WI subregions (green) between i and j, band regions excluded. $(4-5)$, stacked pair and internal loop, respectively. $(6-9)$, $i.j$ closes a multiloop spanning a band. Bottom-left: VP^R, i.e., $i.bp(i)$ in G' crosses base pair in G, $bp(i) \neq j$. VP^R (1) weakly closed non-empty region $[r+1, j]$, (2) empty region $[r+1, j]$. Bottom-right: VP^L, i.e., $bp(j).j$ in G' crosses base pair in G, $bp(j) \neq i$. VP^L (1) empty region $[i, r-1]$. Colored lines correspond with structure classes: Z_{WI} in green may or may not have any substructure, but for $Z_{WI'}$ which also has a green arc, there must be some substructure.

$Z_{VP}(i,j)$ Cases $(1-3)$ handle structures where there are no other base pairs in $[i,j]$ that cross the same band(s) $i.j$ crosses.[1] These cases are unambiguous: either i is covered, j is covered, or both. In Case (4), $(i+1).(j-1)$ forms a stacked pair

[1]For band border lemmas see [3].

(with its energy computed through e_{stP}). In Case (5), $i.j$ and $r.r'$ close an internal loop (with its energy computed through e_{intP}). In Cases $(6-9)$, we handle if there is a multiloop spanning a band. In Case (6), $r.(j-1)$ crosses a base pair in G and $[i+1, r-1]$ is a weakly closed non-empty region. 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 5.4), 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.

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

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

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

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

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

Since G is a pseudoknot-free input structure, all base pair information in G, including borders of bands in G, can be computed in linear time. The time complexity of Z_{VP} is then dominated by the search over region $[i.j]$ to find value of r, and hence, is $O(N^3)$. For a high-level understanding: Cases $(6-9)$ of Z_{VP} are of critical importance to unambiguously handle a multiloop that spans a band or band(s). For a complete decomposition that preserves the $O(N^3)$ time complexity, Z_{VP^R} and Z_{VP^L} are introduced asymmetrically such that there is only one possible path to reach each structure. For example, Z_{VP} Case (8) enforces a structure somewhere in the region between i and r, and moving to Z_{VP^R} Case (1) enforces an additional structure in the subregion adjacent to j. To compare with Z_VP Case (9), similarly we enforce a structure somewhere in the region between r and j, but moving to Z_{VP^L} there is no possible case to introduce an additional structure adjacent to i. Thus, we avoid any ambiguity in Z_{VP} decomposition.

5.3.5 $i.j$ Closes 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 5.5). Otherwise, $Z_{VM}(i,j) = 0$

In Case (1) we have at least two branches, with a terminal pseudoknot-free stem. 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 a terminal stem on a multiloop and is pseudoknot-free. In Case (2) we have at least three branches, with a terminal pseudoknotted stem. Z_{WM^P} is the partition function over all structures such that $i.bp(i)$ is a terminal stem on a multiloop and forms a pseudoloop. Finally, in Case (3) we have exactly two branches, which form a pseudoloop.

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

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 5.5). For base case where $i \geq j$, $Z_{WM}(i,j) = 0$. In Case (1), we have a pseudoknot-free initial stem, in Case (2), the initial stem is a pseudoloop. In Case (3), we have an intermediate pseudoknot-free branch that is not the initial stem. In Case (4), we have an intermediate pseudoloop that is not the the initial stem. In Case (5), j is unpaired.

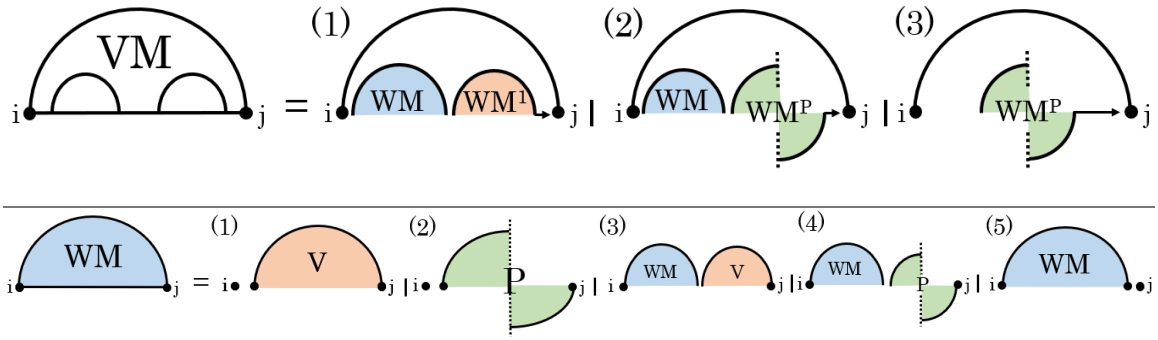


Figure 5.5: Cases of VM and WM. Top: VM (1) terminal pseudoknot-free stem, (2) terminal pseudoknotted stem, with additional branches. (3) terminal pseudoknotted stem, no additional branches. Bottom: WM (1) initial stem pseudoknot-free, (2) initial stem pseudoknotted. (3) intermediate stem pseudoknot-free, additional branches. (4) intermediate stem pseudoknotted, with additional branches. (5), j is unpaired. Filled in circles show regions covered by specific structure classes, blue for Z_{WM}.

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

i and j on a multiloop, terminal stem 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 stem, cf. Fig 5.6). With Case (1), $i.j$ form a pseudoknot-free loop, and Case (2), j is unpaired. For base case where $i \geq j$, $Z_{WM^1}(i,j) = 0$.

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

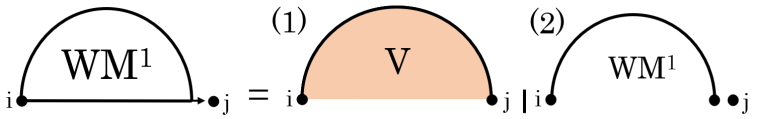


Figure 5.6: Cases of WM^1. (1) terminal stem pseudoknot-free, (2) j is unpaired.

i and j on a multiloop, terminal stem 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 5.7). With Case (1), $i.j$ form a pseudoloop, and Case (2), j is unpaired. For base case where $i \geq j$, $Z_{WM^P}(i,j) = 0$.

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

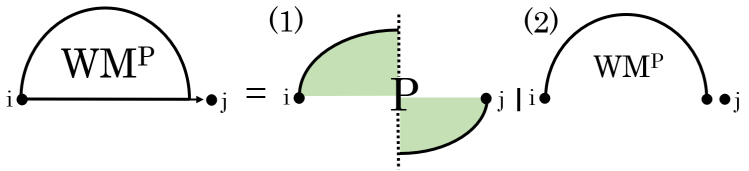


Figure 5.7: Cases of WM^P. (1) terminal stem pseudoknotted, (2) j is unpaired.

5.3.6 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.

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

5.3.7 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.

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

5.3.8 $i.j$ 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. [53] pseudoknot-free algorithm, i.e., penalties for hairpin loop, stacked base pairs, internal/bulge loop, or multiloop.

$i.j$ 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 $Z_{VBI}(i,j) = 0$ [53].

5.3.9 $[i,i'] \cup [bp(i'), bp(i)]$ Band Region

$Z_{BE}(i,i')$ is the partition function over the band $[i',i] \cup [bp_G(i), bp_G(i')]$, if $i \leq i' < bp_G(i') \leq bp_G(i)$; 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 5.8). For base case, $Z_{BE}(i,i) = 0$ if $i < bp_G(i)$.

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

5.4 CParty Implementation

With CParty we compute the first CPF for density-2 structures, therefore, no other algorithm is available for a direct comparison to explicitly verify relative performance. To validate CParty, we first computed the partition function Z_{pkfree} for different test sequences giving no input structure as constraint. Note that, without an input G, the CParty partition function result will be pseudoknot-free. We then compared

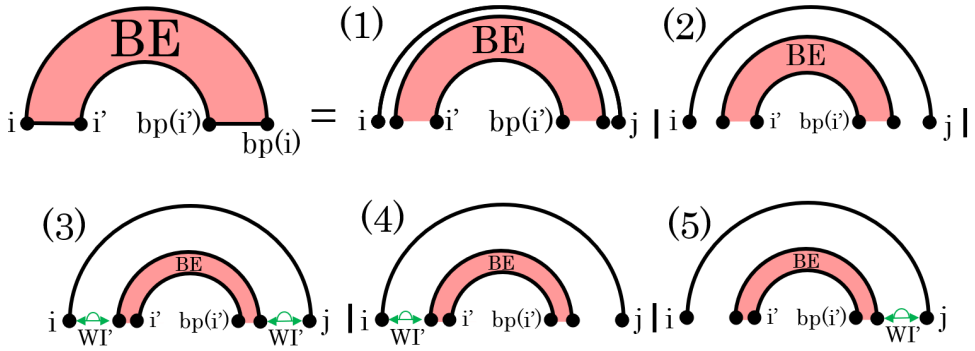


Figure 5.8: Cases of BE. (1) stacked pair in G; (2) internal loop; (3) initial and terminal weakly closed non-empty regions. (4) initial weakly closed non-empty region and terminal loop. (5) initial loop and terminal weakly closed non-empty region.

CParty Z_{pkfree} to the pseudoknot-free partition function result of RNAfold [69], Z'. For accurate comparison between Z_{pkfree} and Z', we used updated parameters [4, 5] for RNAfold, with a dangle-free energy model (i.e., d_0). We found general agreement between pseudoknot-free CParty Z_{pkfree} (with empty input structure G) and RNAFold Z'. For further assessment of the pseudoknotted implementation of CParty, we compute the Boltzmann weight of the MFE structure obtained via HFold with a dangle-free energy model and an input structure G, and divide by CParty Z_{CPF} for the same sequence and structure pair. The resulting fraction is the proportion of the conditional ensemble attributed to the MFE structure, referred to as MFE^Z. We find MFE^Z is in the expected range, and explicitly confirm the completeness and correctness of Z_{CPF} for shorter sequences.

5.5 CParty Results—Discussion

Efficiently computing the partition function for RNA pseudoknot-free secondary structure is the problem solved in 1990 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 [66]. Recently, a linear time approximation algorithm was proposed for pseudoknot-free partition function [54]. Computing partition function for pseudoknotted RNA secondary structure, is however, NP-hard. Existing pseudoknotted partition function methods require certain assumptions that limit the complexity of predicted pseudoknots [62]. These methods are too slow for long RNA sequences

(e.g., NUPACK [152]). New approaches need to be adopted to enable more efficient partition function computation for RNA pseudoknots.

We presented CParty, a novel biologically motivated algorithm following the hierarchical folding hypothesis that efficiently computes the conditional partition function (CPF) for density-2 RNA secondary structures via an $O(N^3)$ DP scheme. CParty receives an RNA sequence and a pseudoknot-free structure, G, as input and calculates the CPF for $G \cup G'$, when G' is a pseudoknot-free structure disjoint from G, and $G \cup G'$ is a density-2 structure. Our algorithm removes the ambiguity of the DP algorithm HFold, enabling an efficient CPF approach that can sample from the conditional distribution, via stochastic analysis of the hierarchical distribution $\mathbb{P}(G) \times \mathbb{P}(G' \mid G)$.

Following the hierarchical folding hypothesis, by restricting the search space using the input structure G, we achieve a runtime complexity of $O(N^3)$ and space complexity of $O(N^2)$ matching the time and space complexity of pseudoknot-free partition function computation. Therefore, CParty represents a significant theoretical improvement in time and space complexity compared to the existing partition function algorithm for simple pseudoknots NUPACK [62, 152], while also handling a broader class of structures including kissing hairpins and arbitrary depth of nested substructures. Computing the CPF can be informative in many scenarios and CParty will be faster for longer sequences compared to unconditional pseudoknot partition function methods like NUPACK.

We implemented CParty for stochastic analysis of pseudoknotted RNAs, enabling novel evaluation of hierarchical kinetics. An important factor in using our algorithm, is to correctly identify the input structure G. As previously identified [52, 14], utilizing the most stable stems (lowest free energy or equivalently the highest probability stems) seems promising in identifying the MFE structure as well as low energy suboptimal structures – structures that are energetically close to the MFE structure. Repeatedly sampling the hierarchical distribution with multiple choices of fixed structure G will be highly informative in identification of possible folding paths to the final structure. By incorporating structure constraints into sampling efforts, a CPF provides 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.

Next, we describe how CParty enables probabilistic investigation to advance understanding of RNA structure formation. By applying CParty to the

SARS-CoV-2 frameshift sequence we demonstrate initial folding of RNA significantly effects pseudoknot pairing propensity of the conditional ensemble. Uncovering hierarchical folding transitions between secondary structure motifs motivates future pan-coronaviral therapeutic RNA structure-function research. CParty can be similarly used in other applications to better harness available structure information, isolating key kinetic paths from initial non-crossing structure to pseudoknotted structure, with implications for subsequent tertiary structure modeling.

5.5.1 CParty for Analyzing Pseudoknot Motifs in SARS-CoV-2

There has been widespread research predicting RNA structures for the SARS-CoV-2 frameshift sequence, including multiple viral genome reactivity probing experiments [10, 9, 11, 156]. Employing CParty with different fixed input structures, G, we provide a stochastic view of suboptimal folding for the SARS-CoV-2 frameshift stimulating RNA structure ensemble. The frameshift stimulating pseudoknot in SARS-CoV-2 [55] possesses a pseudoknot structure which is density-2, making it an excellent use case for CParty.

Based on available SHAPE datasets and thermodynamic-based predictions for the frameshift sequence of SARS-CoV-2 [110, 14], we identify the two most energetically favourable initial stems for the SARS-CoV-2 77 nucleotide frameshift pseudoknot sequence (cf. Table 5.3). With each of these initial stems as structure constraint, we compute Z_{CPF} via CParty. In addition, we calculate MFE^Z (cf. Section 5.4) for the sequence with each initial stem, respectively. With initial stem 2 as constraint, via CParty the conditional ensemble is more stable (larger Z_{CPF}) and diverse (smaller MFE^Z) than the conditional ensemble constrained by initial stem 1. For comparison, we computed the pseudoknot-free conditional partition function via RNAfold [69], referred to as Z', finding CParty Z_{CPF} is significantly larger than Z' due to the energy contribution of density-2 pseudoloops in the ensemble. Investigating the conditional ensemble constrained by initial stem 1, the ratio of RNAfold Z' to CParty Z_{CPF} is an order of magnitude smaller than the ratio of the conditional ensemble energies constrained by initial stem 2. We conclude that initial stem 2 folding leads to the most stable structures and significantly increased pseudoknot pairing propensity of the SARS-CoV-2 frameshift pseudoknot sequence.

We investigate SARS CoV-2 frameshift conditional structure ensembles further

Table 5.3: SARS-CoV-2 frameshift pseudoknot conditional ensemble analysis with most stable initial stems. First column: ID based on free energy rank. Second column: partial structure input for columns $3-6$, "()" indicate paired bases, "." show unpaired bases. Third column: free energy of structure in column two. Fourth column: MFE^Z, cf. Section 5.4. Bottom row: partial input sequence, accession: NC_045512.2, full sequence index $13466-13542$ [2].

ID	*Partial Stem Constraint*	*Kcal/mol*	MFE^Z	Z_{CPF}	Z'	$\frac{Z'}{Z_{CPF}}$
1	.....(((((((((((.....))))))))))))	-11.1	69%	$1.1*10^{14}$	$1.3*10^{12}$	$1.2*10^{-2}$
2	.(((((((.....))))))).....	-6.1	53%	$3.2*10^{14}$	$4.7*10^{11}$	$1.5*10^{-3}$
	AACGGGUUUGCGGUGUAAGUGCAGCCCGUCUUACACCGUG					

by computing Z_{CPF} for decreasing sequence lengths with initial stems 1 and 2 as constraint, in order to simulate the effects of the translocating ribosome [97, 96] (cf. Fig 5.9). We observed the hairpin formed by initial stems 1 and 2 folding (referred to as 3_3 structure motif cf. [15, 134], Fig 5.9-b top arc diagram) dominates the conditional ensemble until a transition occurs (cf. Fig 5.9, red rectangles) to the native-type structure [55] (3_6 motif [15], Fig 5.9-b bottom arc diagram), which dominates the shorter length ensemble. Based on the stable transition of ensemble energies from 3_3 to 3_6 motif we hypothesize that destabilization of initial stem 2 *facilitates* subsequent refolding of the native-type pseudoknot.

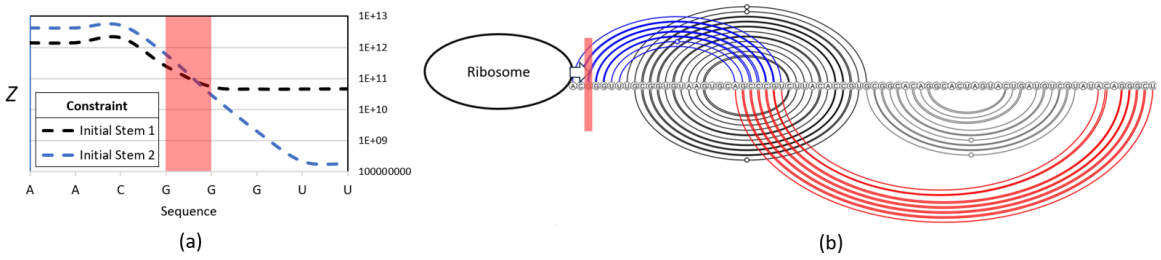


Figure 5.9: SARS-CoV-2 secondary structure motif transition. (a): Z_{CPF} for decreasing SARS-CoV-2 sequence length, Z_{CPF} is divided by each sequence length for normalization. (b): Top arc diagram: 3_3 motif [15], initial stem 1 in black, initial stem 2 in blue. Bottom arc diagram: 3_6 motif. (a & b): Red rectangles highlight the location of a transition from the 3_3 motif to the 3_6 motif.

Having presented our three main results chapters, we now move to the final chapter of this dissertation where we discuss conclusions and future related work.

Chapter 6

Conclusions

In this dissertation we design, implement, and apply algorithms based on thermodynamics for prediction of pseudoknotted RNA secondary structures. We leverage the hierarchical folding hypothesis with (1) Shapify, an MFE algorithm to predict pseudoknotted secondary structures guided by experimental data, and (2) CParty, the first *conditional* partition function algorithm handling a pseudoknotted structure class.

Predicting the most energetically favourable RNA secondary structures and how they may form via the hierarchical folding hypothesis is a promising avenue to understand RNA structure-function. We validate our novel algorithm Shapify by aggregating an RNA structure/SHAPE dataset, demonstrating this hierarchical folding method outperforms ShapeKnots, the best existing algorithm that uses reactivity data to guide secondary structure prediction. In using the most energetically favourable initial stems and reactivity datasets to guide prediction, we find that Shapify is robust to extreme data values. With analysis under the hierarchical folding hypothesis via Shapify, we identify structure patterns such as different coronaviruses that share the same secondary structure motifs. Our work more completely categorized how different types of secondary structures are stable in SARS-CoV-2 at specific sequence lengths, and exactly pinpoint where structure motif transitions occur. We also investigated the SARS-CoV-2 frameshift pseudoknot structure with mutated sequences, finding some mutations can have a destabilizing effect. In targeting functional RNA structures that may regulate frameshifting, more focused analysis is needed to leverage other coronaviruses that possess high structure similarity with SARS-CoV-2 towards treatment design.

We designed and proved the theoretical correctness of our novel CParty algorithm

DP scheme. Then we carefully completed the non-trivial implementation and validation of the conditional partition function algorithm. Our $O(N^3)$ partition function algorithm for pseudoknotted RNA contributes to a more complete understanding of the RNA secondary structure ensemble for an RNA sequence. We demonstrated the functionality of CParty with our analysis of the SARS-CoV-2 frameshift pseudoknot site, unveiling a major secondary structure motif transition location in the conditional ensemble.

CParty enabled us to sample the hierarchical distribution of pseudoknotted RNA secondary structures. With our algorithm CParty we can corroborate the secondary structure predictions from Shapify. For example, independently with each method we identified the same location in SARS-CoV-2 where one major structure motif gives way to another. CParty has profound potential for an in-depth study of RNA molecules by accounting for every possible structure in the density-2 pseudoknotted ensemble. In addition, because CParty computes the *conditional* partition function, we can shed light on how structures fold via kinetic paths from initial stem to possibly pseudoknotted structure. There are multiple fruitful avenues to build additional components on the CParty framework, the first is a pseudoknot-only variant. Such a modification would exclude any pseudoknot-free structures in the ensemble calculation. Another valuable but less trivial project is to compute the base pairing probability of each base in the sequence. This will require implementing a backtrack method to determine how frequently a base is paired in an RNA sequence across all structure classes computed via CParty.

Determining the evolutionary connection between coronavirus frameshift regulating RNA structures is a worthwhile goal in contributing to treatment development for afflicted human and animal populations. Having emphasized SARS-CoV-2 in this work, to enhance the context and discussion of key results we also included sequence covariation information and secondary structure predictions for related coronaviruses. Other betacoronaviruses of note include SARS-CoV-1, the virus responsible for the SARS outbreak in 2003, MERS-CoV, the virus responsible for the MERS outbreak in 2012, and other bat coronaviruses like BtRf-BetaCov or SARS-like WIV1-Cov whose descendants could be responsible for the next pandemic. MERS-CoV structure similarities with SARS-CoV-2 are of particular interest for future investigation with hierarchical folding algorithms because these two viruses belong to different clades (evolutionary branches) [94], and have limited overall sequence identity [130]. Our conserved structural similarity analysis demonstrated SARS-CoV-2 is generally more

stable in the frameshift region than MERS-CoV, and possessed striking similarities to bat coronaviruses.

Overall, the field of RNA secondary structure has a massive room for growth due to the complexities of RNA molecules folding inside of us or other animals. It will take continued meticulous work to fully understand mechanisms like viral frameshift regulation, but the result could be new-age treatments critically limiting the spread of diseases like COVID-19 or human immunodeficiency virus. Any and all computational approaches or other experimental strategies towards a better understanding of RNA kinetic paths will be valuable to improve our description of the real-world model.

Bibliography

[1] Christine E Hajdin, Stanislav Bellaousov, Wayne Huggins, Christopher W Leonard, David H Mathews, and Kevin M Weeks. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *PNAS*, 110(14):5498–5503, 2013.

[2] J Rodney Brister, Danso Ako-Adjei, Yiming Bao, and Olga Blinkova. NCBI viral genomes resource. *Nucleic Acids Res.*, 43(D1):D571–D577, 2015.

[3] 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.

[4] 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.

[5] Jihong Ren, Baharak Rastegari, Anne Condon, and Holger H Hoos. Hotknots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11(10):1494–1504, 2005.

[6] Kévin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinf.*, 25(15):1974, 2009.

[7] 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.

[8] Volodymyr Tsybulskyi, Mohamed Mounir, and Irmtraud M Meyer. R-chie: A web server and R package for visualizing cis and trans RNA–RNA, RNA–DNA and DNA–DNA interactions. *Nucleic Acids Res.*, 48(18):e105–e105, 2020.

[9] 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 SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res.*, 48(22):12436–12452, 2020.

[10] Nicholas C Huston, Han Wan, Madison S Strine, Rafael de Cesaris Araujo Tavares, Craig B Wilen, and Anna Marie Pyle. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol. Cell*, 81(3):584–598, 2021.

[11] 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.

[12] Kaiming Zhang, Ivan N Zheludev, Rachel J Hagey, Raphael Haslecker, Yixuan J Hou, Rachael Kretsch, Grigore D Pintilie, Ramya Rangan, Wipapat Kladwang, Shanshan Li, et al. Cryo-EM and antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome. *Nat. Struct. Mol. Biol.*, 28(9):747–754, 2021.

[13] Daniela Fera, Namhee Kim, Nahum Shiffeldrim, Julie Zorn, Uri Laserson, Hin Hark Gan, and Tamar Schlick. RAG: RNA-As-Graphs web resource. *BMC Bioinf.*, 5(1):1–9, 2004.

[14] Luke Trinity, Ian Wark, Lance Lansing, Hosna Jabbari, and Ulrike Stege. Shapify: Paths to SARS-CoV-2 frameshifting pseudoknot. *PLoS Comput. Biol.*, 19(2):e1010922, 2023.

[15] Tamar Schlick, Qiyao Zhu, Abhishek Dey, Swati Jain, Shuting Yan, and Alain Laederach. To knot or not to knot: Multiple conformations of the SARS-CoV-2 frameshifting RNA element. *J. Am. Chem. Soc.*, 143(30):11404–11422, 2021.

[16] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[17] Pea Carninci, T Kasukawa, S Katayama, J Gough, MC Frith, Norihiro Maeda, Rieko Oyama, T Ravasi, B Lenhard, C Wells, et al. The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, 2005.

[18] Svetlana Deryusheva and Joseph G Gall. Novel small cajal-body-specific RNAs identified in drosophila: probing guide RNA function. *RNA*, 19(12):1802–1814, 2013.

[19] Benjamin J Hale, Cai-Xia Yang, and Jason W Ross. Small RNA regulation of reproductive function. *Mol. Reprod. Dev.*, 81(2):148–159, 2014.

[20] Christine E Holt and Erin M Schuman. The central dogma decentralized: new perspectives on RNA function and local translation in neurons. *Neuron*, 80(3):648–657, 2013.

[21] John S Mattick and Igor V Makunin. Non-coding RNA. *Hum. Mol. Genet.*, 15(suppl_1):R17–R29, 2006.

[22] Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, 49(D1):D192–D200, 2021.

[23] Paul WK Rothemund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006.

[24] Brice Felden. RNA structure: experimental analysis. *Curr. Opin. Microbiol.*, 10(3):286–291, 2007.

[25] Tim R Mercer, Marcel E Dinger, and John S Mattick. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, 10(3):155–159, 2009.

[26] 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.

[27] Fabrizio Pucci, Mehari B Zerihun, Emanuel K Peter, and Alexander Schug. Evaluating DCA-based method performances for RNA contact prediction by a well-curated data set. *RNA*, 26(7):794–802, 2020.

[28] ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799, 2007.

[29] ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements. *PLoS Biol.*, 9(4):e1001046, 2011.

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

[31] William Gao, Ann Yang, and Elena Rivas. Thirteen dubious ways to detect conserved structural RNAs. *IUBMB life*, 75(6):471–492, 2023.

[32] 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. *Bioinf.*, 27(13):i85–i93, 2011.

[33] Ivan Dotu, William A Lorenz, Pascal Van Hentenryck, and Peter Clote. Computing folding pathways between RNA secondary structures. *Nucleic Acids Res.*, 38(5):1711–1722, 2010.

[34] 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.

[35] Meghan Zubradt, Paromita Gupta, Sitara Persad, Alan M Lambowitz, Jonathan S Weissman, and Silvi Rouskin. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Med.*, 14(1):75–82, 2017.

[36] Edoardo Morandi, Ilaria Manfredonia, Lisa M Simon, Francesca Anselmi, Martijn J van Hemert, Salvatore Oliviero, and Danny Incarnato. Genome-scale deconvolution of RNA structure ensembles. *Nat. Methods*, pages 1–4, 2021.

[37] 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 Chem. Biol.*, 2021.

[38] Nathan A Siegfried, Steven Busan, Greggory M Rice, Julie AE Nelson, and Kevin M Weeks. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Med.*, 11(9):959–965, 2014.

[39] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *PNAS*, 101(19):7287–7292, 2004.

[40] Swati Jain, David C Richardson, and Jane S Richardson. Computational methods for RNA structure validation and improvement. *Methods Enzymol.*, 558:181–212, 2015.

[41] Jun Wang, Jian Wang, Yanzhao Huang, and Yi Xiao. 3dRNA v2. 0: An updated web server for RNA 3D structure prediction. *Int. J. Mol. Sci.*, 20(17):4116, 2019.

[42] Tomasz Zok, Maciej Antczak, Michal Zurkowski, Mariusz Popenda, Jacek Blazewicz, Ryszard W Adamiak, and Marta Szachniuk. RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res.*, 46(W1):W30–W35, 2018.

[43] Shuxiang Li, Wilma K Olson, and Xiang-Jun Lu. Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.*, 47(W1):W26–W34, 2019.

[44] Marcin Biesiada, Katarzyna J Purzycka, Marta Szachniuk, Jacek Blazewicz, and Ryszard W Adamiak. Automated RNA 3D structure prediction with RNAComposer. In *RNA Structure Determination*, pages 199–215. Springer, 2016.

[45] Marc Parisien and Francois Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–55, 2008.

[46] Andrey Krokhotin, Kevin Houlihan, and Nikolay V Dokholyan. ifoldrna v2: folding rna with constraints. *Bioinf.*, 31(17):2891–2893, 2015.

[47] Bing Li, Yang Cao, Eric Westhof, and Zhichao Miao. Advances in RNA 3d structure modeling using experimental data. *Front. Genet.*, 11:574485, 2020.

[48] Kevin A Wilkinson, Edward J Merino, and Kevin M Weeks. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNAAsp transcripts. *J. Am. Chem. Soc.*, 127(13):4659–4667, 2005.

[49] Feng Ding, Shantanu Sharma, Poornima Chalasani, Vadim V Demidov, Natalia E Broude, and Nikolay V Dokholyan. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, 14(6):1164–1173, 2008.

[50] Samuel S Cho, David L Pincus, and D Thirumalai. Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. *PNAS*, 106(41):17349–17354, 2009.

[51] Gang Chen, Kung-Yao Chang, Ming-Yuan Chou, Carlos Bustamante, and Ignacio Tinoco. Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of -1 ribosomal frameshifting. *PNAS*, 106(31):12706–12711, 2009.

[52] Hosna Jabbari and Anne Condon. A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures. *BMC Bioinf.*, 15(1):147, 2014.

[53] David H Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288(5):911–940, 1999.

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

[55] Jamie A Kelly, Alexandra N Olson, Krishna Neupane, Sneha Munshi, Josue San Emeterio, Lois Pollack, Michael T Woodside, and Jonathan D Dinman. Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *J. Biol. Chem.*, pages jbc–AC120, 2020.

[56] Ruey-Yi Chang, Ta-Wen Hsu, Yen-Lin Chen, Shu-Fan Liu, Yi-Jer Tsai, Yun-Tong Lin, Yi-Shiuan Chen, and Yi-Hsin Fan. Japanese encephalitis virus non-coding RNA inhibits activation of interferon by blocking nuclear translocation of interferon regulatory factor 3. *Vet. Microbiol.*, 166(1-2):11–21, 2013.

[57] Yizhu Lin, Brigitte F Schmidt, Marcel P Bruchez, and C Joel McManus. Structural analyses of NEAT1 lncRNAs suggest long-range RNA interactions that may contribute to paraspeckle architecture. *Nucleic Acids Res.*, 46(7):3742–3752, 2018.

[58] Irina V Novikova, Scott P Hennelly, Chang-Shung Tung, and Karissa Y Sanbonmatsu. Rise of the RNA machines: exploring the structure of long non-coding RNAs. *J. Mol. Biol.*, 425(19):3731–3746, 2013.

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

[60] 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.

[61] 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.

[62] Robert M Dirks and Niles A Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24(13):1664–1677, 2003.

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

[64] Elena Rivas and Sean R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285(5):2053–2068, 1999.

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

[66] 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.

[67] 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.

[68] David H Mathews and Douglas H Turner. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, 16(3):270–278, 2006.

[69] 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 Mol. Biol*, 6:1–14, 2011.

[70] Stefan Washietl. *RNAz 2.1 Manual*. Department for Theoretical Chemistry, University Vienna.

[71] Mateo Gray, Sean Chester, and Hosna Jabbari. KnotAli: informed energy minimization through the use of evolutionary information. *BMC bioinformatics*, 23(1):159, 2022.

[72] Baharak Rastegari and Anne Condon. Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. *J. Comput. Biol.*, 14(1):16–32, 2007.

[73] E Stofer, C Chipot, and R Lavery. Free energy calculations of Watson-Crick base pairing in aqueous solution. *Journal of the American Chemical Society*, 121(41):9503–9508, 1999.

[74] James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[75] Gabriele Varani and William H McClain. The G·U wobble base pair. *EMBO reports*, 1(1):18–23, 2000.

[76] Jhuma Das, Shayantani Mukherjee, Abhijit Mitra, and Dhananjay Bhattacharyya. Non-canonical base pairs and higher order structures in nucleic acids: crystal structure database analysis. *Journal of Biomolecular Structure and Dynamics*, 24(2):149–161, 2006.

[77] Juraj Michalik. *Non-redundant sampling in RNA Bioinformatics*. PhD thesis, Université Paris Saclay (COmUE), 2019.

[78] Michael S Waterman. Secondary structure of single-stranded nucleic acids. *Adv. math. suppl. studies*, 1:167–212, 1978.

[79] Michael S Waterman and Temple F Smith. RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, 42(3-4):257–266, 1978.

[80] Richard Bellman and RAND Corp Santa Monica CA. Dynamic programming. 1956.

[81] Ruth Nussinov, George Pieczenik, Jerrold R Griggs, and Daniel J Kleitman. Algorithms for loop matchings. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.

[82] Susan J Schroeder and Douglas H Turner. Optical melting measurements of nucleic acid thermodynamics. In *Methods Enzymol.*, volume 468, pages 371–387. Elsevier, 2009.

[83] Douglas H Turner, Naoki Sugimoto, and Susan M Freier. RNA structure prediction. *Annu. Rev. Biophys. and Biochem.*, 17(1):167–192, 1988.

[84] Rune B Lyngsø, Michael Zuker, and Christian NS Pedersen. An improved algorithm for RNA secondary structure prediction. *BRICS Report Series*, 6(15), 1999.

[85] Michael Zuker and David Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.

[86] Valerie Hower and Christine E Heitsch. Parametric analysis of RNA branching configurations. *Bull. Math. Biol.*, 73:754–776, 2011.

[87] Daniel Jost and Ralf Everaers. Prediction of RNA multiloop and pseudoknot conformations from a lattice-based, coarse-grain tertiary structure model. *J. Chem. Phys.*, 132(9), 2010.

[88] Jacob T Polaski, Samantha M Webster, James E Johnson, and Robert T Batey. Cobalamin riboswitches exhibit a broad range of ability to discriminate between methylcobalamin and adenosylcobalamin. *J. Biol. Chem.*, 292(28):11650–11658, 2017.

[89] Frederick Rehfeld, Jennifer L Eitson, Maikke B Ohlson, Tsung-Cheng Chang, John W Schoggins, and Joshua T Mendell. CRISPR screening reveals a dependency on ribosome recycling for efficient SARS-CoV-2 programmed ribosomal frameshifting and viral replication. *Cell Rep.*, 42(2), 2023.

[90] Sha Gong, Yanli Wang, Zhen Wang, and Wenbing Zhang. Co-transcriptional folding and regulation mechanisms of riboswitches. *Molecules*, 22(7):1169, 2017.

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

[92] Kim Sharp and Franz Matschinsky. Translation of ludwig boltzmann's paper "on the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium" sitzungberichte der kaiserlichen akademie der wissenschaften. mathematisch-naturwissen classe. abt. ii, lxxvi 1877, pp 373-435 (wien. ber. 1877, 76: 373-435). reprinted in wiss. abhandlungen, vol. ii, reprint 42, p. 164-223, barth, leipzig, 1909. *Entropy*, 17(4):1971–2009, 2015.

[93] Yu Chen, Qianyun Liu, and Deyin Guo. Emerging coronaviruses: genome structure, replication, and pathogenesis. *J. Med. Virol.*, 92(4):418–423, 2020.

[94] Nicola Petrosillo, Giulio Viceconte, Onder Ergonul, Giuseppe Ippolito, and Eskild Petersen. COVID-19, SARS and MERS: are they closely related? *Clin. Microbiol. Infect.*, 2020.

[95] Jamie A Kelly, Michael T Woodside, and Jonathan D Dinman. Programmed-1 ribosomal frameshifting in coronaviruses: a therapeutic target. *Virology*, 554:75–82, 2021.

[96] Jonathan D Dinman. Mechanisms and implications of programmed translational frameshifting. *Wiley Interdiscip. Rev. RNA*, 3(5):661–673, 2012.

[97] John F Atkins, Gary Loughran, Pramod R Bhatt, Andrew E Firth, and Pavel V Baranov. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.*, 44(15):7007–7078, 2016.

[98] Bo Wu, Haibo Zhang, Ruirui Sun, Sijia Peng, Barry S Cooperman, Yale E Goldman, and Chunlai Chen. Translocation kinetics and structural dynamics of ribosomes are modulated by the conformational plasticity of downstream pseudoknots. *Nucleic Acids Res.*, 46(18):9736–9748, 2018.

[99] Rozhgar A Khailany, Muhamad Safdar, and Mehmet Ozaslan. Genomic characterization of a novel SARS-CoV-2. *Gene Rep.*, page 100682, 2020.

[100] Krishna Neupane, Meng Zhao, Aaron Lyons, Sneha Munshi, Sandaru M Ileperuma, Dustin B Ritchie, Noel Q Hoffer, Abhishek Narayan, and Michael T Woodside. Structural dynamics of single SARS-CoV-2 pseudoknot molecules reveal topologically distinct conformers. *Nat. Commun.*, 12(1):1–9, 2021.

[101] 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. *Nat. Commun.*, 13(1):4284, 2022.

[102] Sara Ibrahim Omar, Meng Zhao, Rohith Vedhthaanth Sekar, Sahar Arbabi Moghadam, Jack A Tuszynski, and Michael T Woodside. Modeling the structure of the frameshift-stimulatory pseudoknot in SARS-CoV-2 reveals multiple possible conformers. *PLoS Comput. Biol.*, 17(1):e1008603, 2021.

[103] Dominique Fourmy and Satoko Yoshizawa. A cytosine-to-uracil change within the programmed -1 ribosomal frameshift signal of SARS-CoV-2 results in structural similarities with the MERS-CoV signal. *bioRxiv*, 2020.

[104] Krishna Neupane, Sneha Munshi, Meng Zhao, Dustin B Ritchie, Sandaru M Ileperuma, and Michael T Woodside. Anti-frameshifting ligand active against SARS coronavirus-2 is resistant to natural mutations of the frameshift-stimulatory pseudoknot. *J. Mol. Biol.*, 432(21):5843–5847, 2020.

[105] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic Acids Res.*, 41(D1):D36–D42, 2012.

[106] Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

[107] Albert Tian Chen, Kevin Altschuler, Shing Hei Zhan, Yujia Alina Chan, and Benjamin E Deverman. COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. *Elife*, 10:e63409, 2021.

[108] Perumal Arumugam Desingu, K Nagarajan, and Kuldeep Dhama. Emergence of omicron third lineage BA.3 and its importance. *J. Med. Virol.*, 2022.

[109] Houriiyah Tegally, Monika Moir, Josie Everatt, Marta Giovanetti, Cathrine Scheepers, Eduan Wilkinson, Kathleen Subramoney, Zinhle Makatini, Sikhulile Moyo, Daniel G Amoako, et al. Emergence of SARS-CoV-2 omicron lineages BA. 4 and BA. 5 in south africa. *Nat. Med.*, pages 1–1, 2022.

[110] Tamar Schlick, Qiyao Zhu, Swati Jain, and Shuting Yan. Structure-altering mutations of the SARS-CoV-2 frameshifting RNA element. *Biophys. J.*, 120(6):1040–1053, 2021.

[111] Ewan P Plant, Gabriela C Pérez-Alvarado, Jonathan L Jacobs, Bani Mukhopadhyay, Mirko Hennig, and Jonathan D Dinman. A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.*, 3(6):e172, 2005.

[112] Daniella Ishimaru, Ewan P Plant, Amy C Sims, Boyd L Yount Jr, Braden M Roth, Nadukkudy V Eldho, Gabriela C Perez-Alvarado, David W Armbruster, Ralph S Baric, Jonathan D Dinman, et al. RNA dimerization plays a role in ribosomal frameshifting of the SARS coronavirus. *Nucleic Acids Res.*, 41(4):2594–2608, 2013.

[113] Ewan P Plant, Rasa Rakauskaitė, Deborah R Taylor, and Jonathan D Dinman. Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. *J. Virol.*, 84(9):4330–4340, 2010.

[114] Ewan P Plant, Amy C Sims, Ralph S Baric, Jonathan D Dinman, and Deborah R Taylor. Altering SARS coronavirus frameshift efficiency affects genomic and subgenomic RNA production. *Viruses*, 5(1):279–294, 2013.

[115] So-Jung Park, Yang-Gyun Kim, and Hyun-Ju Park. Identification of RNA pseudoknot-binding ligand that inhibits the -1 ribosomal frameshifting of SARS-coronavirus by structure-based virtual screening. *J. Am. Chem. Soc.*, 133(26):10094–10100, 2011.

[116] Dustin B Ritchie, Jingchyuan Soong, William KA Sikkema, and Michael T Woodside. Anti-frameshifting ligand reduces the conformational plasticity of the SARS virus pseudoknot. *J. Am. Chem. Soc.*, 136(6):2196–2199, 2014.

[117] Yu Sun, Laura Abriola, Rachel O Niederer, Savannah F Pedersen, Mia M Alfajaro, Valter Silva Monteiro, Craig B Wilen, Ya-Chi Ho, Wendy V Gilbert, Yulia V Surovtseva, et al. Restriction of SARS-CoV-2 replication by targeting programmed -1 ribosomal frameshifting. *PNAS*, 118(26), 2021.

[118] Sneha Munshi, Krishna Neupane, Sandaru M Ileperuma, Matthew TJ Halma, Jamie A Kelly, Clarissa F Halpern, Jonathan D Dinman, Sarah Loerch, and Michael T Woodside. Identifying inhibitors of- 1 programmed ribosomal frameshifting in a broad spectrum of coronaviruses. *Viruses*, 14(2):177, 2022.

[119] Dae-Gyun Ahn, Gun Young Yoon, Sunhee Lee, Keun Bon Ku, Chonsaeng Kim, Kyun-Do Kim, Young-Chan Kwon, Geon-Woo Kim, Bum-Tae Kim, and Seong-Jun Kim. A novel frameshifting inhibitor having antiviral activity against zoonotic coronaviruses. *Viruses*, 13(8):1639, 2021.

[120] Luke Trinity, Lance Lansing, Hosna Jabbari, and Ulrike Stege. SARS-CoV-2 ribosomal frameshifting pseudoknot: Detection of inter-viral structural similarity. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 451–460, 2021.

[121] Fabian Sievers, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D. Thompson, and Desmond G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7, 2011.

[122] Andreas R Gruber, Sven Findeiß, Stefan Washietl, Ivo L Hofacker, and Peter F Stadler. RNAz 2.0: improved noncoding RNA detection. In *Biocomputing 2010*, pages 69–79. World Scientific, 2010.

[123] Katherine E Deigan, Tian W Li, David H Mathews, and Kevin M Weeks. Accurate SHAPE-directed RNA structure determination. *PNAS*, 106(1):97–102, 2009.

[124] Pablo Cordero, Julius B Lucks, and Rhiju Das. An RNA mapping database for curating RNA structure mapping experiments. *Bioinf.*, 28(22):3006–3008, 2012.

[125] Kyle E Watters, Timothy R Abbott, and Julius B Lucks. Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucleic Acids Res.*, 44(2):e12–e12, 2016.

[126] Ramya Rangan, Andrew M Watkins, Jose Chacon, Rachael Kretsch, Wipapat Kladwang, Ivan N Zheludev, Jill Townley, Mats Rynge, Gregory Thain, and Rhiju Das. De novo 3D models of SARS-CoV-2 RNA elements from consensus experimental secondary structures. *Nucleic Acids Res.*, 49(6):3092–3108, 2021.

[127] Julius B Lucks, Stefanie A Mortimer, Cole Trapnell, Shujun Luo, Sharon Aviran, Gary P Schroth, Lior Pachter, Jennifer A Doudna, and Adam P Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *PNAS*, 108(27):11063–11068, 2011.

[128] Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinform.*, 11(1):1–9, 2010.

[129] Jian Ye, Scott McGinnis, and Thomas L Madden. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, 34:W6–W9, 2006.

[130] Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, 395(10224):565–574, 2020.

[131] Wes Sanders, Ethan J Fritch, Emily A Madden, Rachel L Graham, Heather A Vincent, Mark T Heise, Ralph S Baric, and Nathaniel J Moorman. Comparative analysis of coronavirus genomic RNA structure reveals conservation in SARS-like coronaviruses. *BioRxiv*, 2020.

[132] J Herald and SG Siddell. An 'elaborated' pseudoknot is required for high frequency frameshifting during translation of HCV 229E polymerase mRNA. *Nucleic Acids Res.*, 21(25):5838–5842, 1993.

[133] Tammy CT Lan, Matty F Allan, Lauren E Malsick, Jia Z Woo, Chi Zhu, Fengrui Zhang, Stuti Khandwala, Sherry SY Nyeo, Yu Sun, Junjie U Guo, et al. Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. *Nat. Commun.*, 13(1):1–14, 2022.

[134] Shuting Yan, Qiyao Zhu, Jenna Hohl, Alex Dong, and Tamar Schlick. Evolution of coronavirus frameshifting elements: Competing stem networks explain conservation and variability. *PNAS*, 120(20):e2221324120, 2023.

[135] Mo Yang, Feyisola P Olatunji, Curran Rhodes, Sumirtha Balaratnam, Kara Dunne-Dombrink, Srinath Seshadri, Xiao Liang, Christopher P Jones, Stuart FJ Le Grice, Adrian R Ferre-D'Amare, et al. Discovery of small molecules targeting the frameshifting element RNA in SARS-CoV-2 viral genome. *ACS Med. Chem. Lett.*, 2023.

[136] Carmine Varricchio, Gregory Mathez, Trestan Pillonel, Claire Bertelli, Laurent Kaiser, Caroline Tapparel, Andrea Brancale, and Valeria Cagno. Geneticin shows selective antiviral activity against SARS-CoV-2 by interfering with programmed-1 ribosomal frameshifting. *Antiviral Research*, 208:105452, 2022.

[137] Gregory Mathez and Valeria Cagno. Small molecules targeting viral RNA. *Int. J. Mol. Sci.*, 24(17):13500, 2023.

[138] Jens Kurreck, Eliza Wyszko, Clemens Gillen, and Volker A Erdmann. Design of antisense oligonucleotides stabilized by locked nucleic acids. *Nucleic Acids Res.*, 30(9):1911–1918, 2002.

[139] Daseuli Yu, Hee-Jeong Han, Jeonghye Yu, Jihye Kim, Gun-Hee Lee, Ju-Hee Yang, Byeong-Min Song, Dongseob Tark, Byeong-Sun Choi, Sang-Min Kang, et al. Pseudoknot-targeting Cas13b combats SARS-CoV-2 infection by suppressing viral replication. *Mol. Ther.*, 31(6):1675–1687, 2023.

[140] Rohith Vedhthaanth Sekar, Patricia J Oliva, and Michael T Woodside. Modelling the structures of frameshift-stimulatory pseudoknots from representative bat coronaviruses. *PLoS Comput. Biol.*, 19(5):e1011124, 2023.

[141] Lukas Pekarek, Matthias M Zimmer, Anne-Sophie Gribling-Burrer, Stefan Buck, Redmond Smyth, and Neva Caliskan. Cis-mediated interactions of the SARS-

CoV-2 frameshift RNA alter its conformations and affect function. *Nucleic Acids Res.*, 51(2):728–743, 2023.

[142] Weiwei He, Josue San Emeterio, Michael T Woodside, Serdal Kirmizialtin, and Lois Pollack. Atomistic structure of the SARS-CoV-2 pseudoknot in solution from SAXS-driven molecular dynamics. *Nucleic Acids Res.*, page gkad809, 2023.

[143] Xunxun Wang, Ya-Lan Tan, Shixiong Yu, Ya-Zhou Shi, and Zhi-Jie Tan. Predicting 3D structures and stabilities for complex RNA pseudoknots in ion solutions. *Biophys. J.*, 122(8):1503–1516, 2023.

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

[145] Qiyao Zhu, Louis Petingi, and Tamar Schlick. RNA-As-Graphs motif atlas—dual graph library of RNA modules and viral frameshifting-element applications. *Int. J. Mol. Sci.*, 23(16):9249, 2022.

[146] Brett E Pickett, Eva L Sadat, Yun Zhang, Jyothi M Noronha, R Burke Squires, Victoria Hunt, Mengya Liu, Sanjeev Kumar, Sam Zaremba, Zhiping Gu, et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic acids res.*, 40(D1):D593–D598, 2012.

[147] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[148] Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[149] Luke Trinity, Sebastian Will, Yann Ponty, Ulrike Stege, and Hosna Jabbari. CParty: Conditional partition function for density-2 RNA pseudoknots. *bioRxiv*, pages 2023–05, 2023.

[150] Ekaterina Knizhnik, Stepan Chumakov, Julia Svetlova, Iulia Pavlova, Yuri Khodarovich, Vladimir Brylev, Vjacheslav Severov, Rugiya Alieva, Liubov Kozlovskaya, Dmitry Andreev, et al. Unwinding the SARS-CoV-2 ribosomal frameshifting pseudoknot with LNA and G-Clamp-Modified Phosphorothioate Oligonucleotides inhibits viral replication. *Biomol.*, 13(11):1660, 2023.

[151] 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, Saarbrücken, Germany, January 2011. Springer.

[152] 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. *J. Comput. Chem.*, 32(1):170–173, 2011.

[153] 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. *J. Comput. Biol.*, 16(6):803–815, 2009.

[154] Mirela Andronescu, Anne Condon, Holger H Hoos, David H Mathews, and Kevin P Murphy. Computational approaches for RNA energy parameter estimation. *RNA*, 16(12):2304–2318, 2010.

[155] 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.

[156] Kaiming Zhang, Ivan N Zheludev, Rachel J Hagey, Raphael Haslecker, Yixuan J Hou, Rachael Kretsch, Grigore D Pintilie, Ramya Rangan, Wipapat Kladwang, Shanshan Li, et al. Cryo-EM and antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome. *Nat. Struct. Mol. Biol.*, 28(9):747–754, 2021.

[157] Luke Trinity, Ulrike Stege, and Hosna Jabbari. Tying the knot: Unraveling the intricacies of the coronavirus frameshift pseudoknot. *bioRxiv*, pages 2023–12, 2023.

Appendix A

RNA Structures and Associated Data

S1 File. Supplementary Materials Shapify.
Secondary structure predictions via Shapify for SARS-CoV-1, SARS-CoV-2, and MERS-CoV including RNAz-determined consensus structure for the frameshift pseudoknot sequence. `https://doi.org/10.1371/journal.pcbi.1010922.s001`

S2 File. Supplementary Materials Tying The Knot.
Secondary structure predictions for extended length SARS-CoV-2 via Shapify, and homologous coronaviruses via KnotAli. `https://github.com/ltrinity/TyingTheKnot`

Appendix B

Publication Authorship

The candidate contributed to all major ideas and writing of the published and unpublished manuscripts that are the basis of this thesis. The candidate was the lead author in all published and unpublished manuscripts. The candidate collaborated with the candidate's supervisors and other colleagues and mentors in all aspects of research including conceptualization, project administration, methodology, data curation, algorithm design, software development, validation, evaluation, formal analysis, visualization, writing the original draft, reviews, and editing. The introductory and conclusion chapters, Chapters 1–2, and 6, were written by the candidate, but used selected content from publications that he co-authored [120, 14, 157, 149]. Chapter 3 is a collaboration between the candidate, Ian Wark, Lance Lansing, and the candidate's supervisors, Dr. Hosna Jabbari, and Dr. Ulrike Stege. A version of Chapter 3 appeared in the 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI) [120], and a subsequent version was published in PLoS Computational Biology [14]. Chapter 4 is a collaboration between the candidate, and the candidate's supervisors, Dr. Hosna Jabbari, and Dr. Ulrike Stege. A preliminary version of Chapter 4 is publicly available on bioRxiv [157]. Chapter 5 is a collaboration between the candidate, Dr. Sebastian Will, Dr. Yann Ponty, and the candidate's supervisors, Dr. Hosna Jabbari, and Dr. Ulrike Stege. A preliminary version of Chapter 5 is publicly available on bioRxiv [149].