MASTERARBEIT | MASTER'S THESIS

Titel | Title

Computational Design of Synthetic Flaviviral xrRNAs

verfasst von | submitted by

Katrin Gutenbrunner BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Acknowledgements

First, I would like to thank Dr. Michael Wolfinger, for giving me the opportunity to write my master's thesis and be part of this project. Without his support, this thesis and the overall project would not have been possible.

I would also like to extend my heartfelt thanks to Leonhard Sidl, who guided me throughout this thesis. His invaluable feedback and advice played a significant role in shaping this work. In addition, I am especially grateful to him for contributing analyses and insight beyond the scope of my expertise, which were crucial to this project.

I would also like to thank the the entire TBI team, and especially Prof. Hofacker, for welcoming me into their institute and making me feel like a part of the group.

Lastly, I want to thank my family for their ongoing support throughout my academic journey. Their encouragement and love have been the foundation of my success, and I could not have reached this point without them.

Abstract

This thesis presents the development of a computational framework for designing synthetic RNAs that mimic the structural and functional properties of natural flaviviral exoribonuclease-resistant RNAs (xrRNAs). The design focuses on two subgroups of *orthoflaviviruses*: mosquito-borne flaviviruses (MBFV) and tick-borne flaviviruses (TBFV), with an emphasis on achieving 5'-3' exoribonuclease 1 (Xrn1) resistance. The computational model was built using the Infrared framework and combined sequence constraints with structural scaffolds to replicate critical elements such as pseudoknots and the unique ring-like tertiary structure required for resistance. Through iterative refinement, the model successfully generated synthetic sequences that demonstrated strong similarity to natural xrRNAs in both secondary and tertiary structure analyses. In particular, the MBFV-based model produced sequences that exhibited resistance to Xrn1, validated by experimental testing, while the TBFV-based model showed only partial success. This study demonstrates the possibility of computationally designing RNA sequences with specific resistance properties. Future work aims to integrate synthetic xrRNAs into functional riboswitches, paving the way for innovative applications in synthetic biology.

Kurzfassung

Diese Arbeit präsentiert die Entwicklung eines computergestützten Modells für das Design synthetischer RNAs, die die strukturellen und funktionellen Eigenschaften natürlicher flaviviraler exoribonuklease-resistenter RNAs (xrRNAs) imitieren. Der Fokus des Designs liegt auf zwei Untergruppen des *Orthoflavivirus*: Mosquito-borne Flavivirus (MBFV) und Tick-borne Flavivirus (TBFV), wobei der Schwerpunkt auf der Erreichung der 5'-3'-Exoribonuklease 1 (XRN1)-Resistenz liegt. Das Modell wurde unter Verwendung des Frameworks Infrared erstellt und kombiniert Sequenzbeschränkungen mit strukturellen Vorgaben, um kritische Elemente wie Pseudoknoten und die einzigartige ringartige Tertiärstruktur zu reproduzieren, die für die Resistenz erforderlich sind. Durch iteratives Verfeinern konnte das Modell erfolgreich synthetische Sequenzen generieren, die sowohl in der Sekundär- als auch in der Tertiärstrukturanalyse eine starke Ähnlichkeit zu natürlichen xrRNAs aufwiesen. Insbesondere erzeugte das MBFV-basierte Modell Sequenzen, welche eine durch experimentelle Tests validierte Resistenz gegen Xrn1 zeigten, während das TBFV-basierte Modell nur teilweise erfolgreich war. Diese Studie demonstriert die Möglichkeit, RNA-Sequenzen mit spezifischen Resistenzeigenschaften rechnerisch zu entwerfen. Zukünftige Arbeiten könnten die Integration von synthetischen xrRNAs in funktionale Riboswitches anstreben und so den Weg für innovative Anwendungen in der synthetischen Biologie ebnen.

Contents

1. Introduction

1.1. Motivation and Research Problem

Ribonucleic acid (RNA) plays a critical role in cellular processes, with its structural versatility enabling diverse biological functions. Among these, exoribonuclease-resistant RNA (xrRNA) represent a unique class of RNA structures capable of resisting degradation by certain exoribonucleases such as 5'-3' exoribonuclease 1 (Xrn1). These structures are often associated with the virus genus *Orthoflaviviruses*, as many members of this group contain one or more xrRNA elements within their genomes. The study of flaviviral xrRNAs provides significant insight into their structural and functional mechanisms.

Understanding the mechanisms underlying xrRNA structures opens the possibility of creating artificial RNA elements that exhibit resistance to exoribonuclease while being able to incorporate additional structural or sequence-specific features. The design and manipulation of synthetic xrRNAs could enable numerous applications, ranging from the use as novel tools in synthetic biology and RNA biochemistry to the study of essential features of xrRNA functionality.

Despite their intriguing properties and potential applications, xrRNAs remain understudied, leaving many questions unanswered. Developing synthetic xrRNAs would not only help validate assumptions about their structural characteristics but also uncover new insights into the features critical for their unique resistance property. These investigations could distinguish essential structural elements from those that are not, paving the way for more targeted and effective applications.

As part of an international project focused on the design of synthetic RNase resistant structures and their integration in novel riboswitch constructs to regulate mRNA stability, this master's thesis builds upon previous work conducted in this field. Within the scope of the project, various xrRNA sequences from *Orthoflavivirus* were analysed, providing valuable information on their key properties and structural characteristics. This thesis extends this research by attempting to design synthetic xrRNA sequences that replicate the critical properties of the biological sequences studied previously. In collaboration with the Mörl Lab in Leipzig, these synthetic sequences can also be tested *in vitro* to evaluate their resistance to exoribonucleases.

1.2. Goals and Objectives

The primary goal of this thesis is to develop and optimize a computational model to design xrRNA sequences that replicate the structural and functional characteristics of natural flaviviral xrRNAs. The computational framework Infrared [1], designed to address various bioinformatics challenges, including RNA design, is employed for this purpose.

The specific objectives are as follows:

- Develop a model to design xrRNA sequences that demonstrate exoribonuclease resistance, mimicking the natural behavior of flaviviral xrRNAs.
- Validate the resistance to exoribonuclease of the designed sequences.
- Evaluate the importance of specific structural features by designing xrRNAs with targeted modifications and testing their resistance properties.
- Benchmark the Infrared-based model against an alternative approach, comparing the designed sequences from both frameworks to each other and to natural data.

The ultimate goal is to establish a reliable design model capable of producing sequences that exhibit Xrn1 resistance while maintaining sufficient variability for subsequent manipulation.

2. Background

2.1. RNA

RNA is a versatile and essential part of all known life forms, playing a critical role in various biological processes. RNA is a linear polymer, where the backbone consists of ribose sugar rings linked by phosphate groups. Each of these sugars is defined by being connected to one of the four bases: adenine (A), cytosine (C), guanine (G), or uracil (U). The order of these bases, the sequence, describes the RNA. Unlike DNA, RNA occurs solely single-stranded, making the RNA more flexible and allowing it to bend back to itself to form hydrogen bonds. The sequence of the RNA molecule determines these intramolecular interactions and leads to the formation of different conformations, which can be divided into secondary (2D) and tertiary (3D) structures [2, 3].

2.1.1. Secondary Structure

The secondary structure of an RNA molecule is defined by the interaction of the RNA bases that form Watson-Crick canonical base pairs [4]. Additionally, the non-canonical pair between G and U, called a wobble base pair, contributes to the secondary structure [2]. These base pairs create two-dimensional conformations through hydrogen bonds, which are assembled into specific motifs such as helices or stems, loops, and bulges.

A helix or stem within an RNA is formed when two parts of the strand create base pairs between (nearly) continuous sets of complementary nucleotides. This structure is primarily stabilized by base stacking, with G-C pairs providing higher stability because of their three hydrogen bonds, whereas other pairs only form two hydrogen bonds. A bulge occurs when one or more nucleotides on one side of the helix remain unpaired, causing a kink in the helix. Hairpin loops consist of a single-stranded loop of nucleotides closed off by a stem, resembling a hairpin in shape. An internal loop forms when two or more nucleotides are mismatched within a helix, remaining unpaired and creating symmetric or asymmetric loops. Multi loops, or multi-helix junctions, are formed when three or more helices converge. The most common multi loops are three-way and four-way junctions [5, 6]. The graphical representation of these elements of the secondary structure can be seen in Figure 2.1.

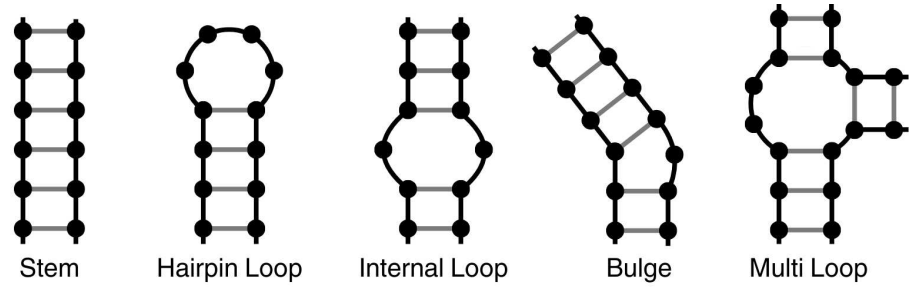


Figure 2.1.: **Different types of RNA secondary structures.** The figure illustrates common RNA secondary structure motifs formed by base-pairing and unpaired regions.

2.1.2. Tertiary Structure

The tertiary structure of an RNA molecule is defined as the arrangement of single RNA building blocks in a three-dimensional space. The arrangement is caused by the use of canonical and non-canonical base pair interaction. It is assumed that the RNA molecule folds into a stable secondary structure first and afterwards the RNA adopts its tertiary conformation. The ability to form tertiary structures is believed to be caused by the flexibility of its unpaired parts [4].

Non-canonical base pair interactions include base triples and pseudoknots. A pseudoknot minimally consists of two helical elements, where the loop region of one helix is paired with a complementary region outside of the loop. Pseudoknots can differ in various characteristics, such as the length and composition of the loop and helix, or the type of interaction. Therefore, pseudoknots represent a very diverse structural group, creating complex structures in RNA. They can play a crucial role in the RNA's function [7].

A base triple is defined by three RNA nucleobases that interact with each other edge-to-edge by hydrogen bonding, creating a more complex structure than a standard base pair. They have been found in structured RNA molecules, stabilizing many tertiary RNA interactions. [8]

2.1.3. RNA Structure Prediction

The structure of RNA is crucial for understanding both its function and its origin. Therefore, predicting the most favorable structure into which an RNA molecule folds is essential. The most thermodynamically stable secondary structure that the RNA can adopt is known as its Minimal Free Energy (MFE) structure. This structure minimizes the Gibbs free energy by accounting for contributions from base pairing, stacking interactions, and loop formation. While the MFE structure is only one of the many

conformations an RNA can adopt, it has the highest probability within the structural ensemble, which includes all possible RNA folding configurations. Computational prediction of the secondary structure typically involves calculating the MFE structure for a given RNA sequence. Various approaches have been developed to address this prediction problem, leading to the creation of several computational tools [5]. One widely used tool is ViennaRNA [9], which has been cited more than 5000 times and has demonstrated its effectiveness in numerous applications over the past several decades. For example, it has been used to analyse DNA secondary structure formation by Szlachta et al. [10]. A comparative study further confirmed RNAfold, part of the ViennaRNA package, as one of the most effective tools for single-stranded nucleic acid secondary structure prediction [11].

Calculating potential RNA structures includes the energetic contributions of different structural components, such as base pairing, stacking interactions, and others. Several energy parameters have been developed and are commonly used for RNA structure prediction. Turner and colleagues developed one of the foundational energy parameters in the 1980s and 1990s [12, 13, 14]. These were further refined by Mathews et al. in 2004, incorporating recent experimental results on terminal mismatches, as well as hairpin, bulge, internal, and multibranch loops [15]. Today, Turner parameters are among the most widely used energy parameters in RNA structure prediction.

However, other energy parameters also exist. One notable example are the Andronescu energy parameters [16], which build on the Turner model by refining its parameters and addressing some of its limitations, particularly in predicting complex RNA structures. The Andronescu parameters combine thermodynamic measurements with machine learning techniques to improve the accuracy of energy parameters. This integration achieves a closer alignment between the predictions and the experimental RNA folding data. As shown in Skibinski [17], the Andronescu parameters appear to be better suited to predict the structure of xrRNA sequences.

2.1.4. Structure Representation

The structure of an RNA molecule can be represented as a dot-bracket notation. In this notation, matching parentheses indicate base pairs, while dots represent unpaired bases. This representation is simple and minimalistic, requiring very little space, but only provides information about the location of the base pairs without further details.

Alternatively, the structure can be described using a dot plot, where each dot indicates a base pair. The intensity of the dot can represent the likelihood of the corresponding base pair. This method offers insights into the structure and also potential alternative structures that the RNA can adopt, but it requires more space than the dot-bracket

notation.

Another way to represent RNA structures is through graph-based representations, where each node corresponds to a base, and each edge represents an interaction between two bases. This method allows for visualization of both the sequence and the structure of the RNA. Additional information, such as the type of base pairs or complex interactions such as base triples, can also be included, providing a more detailed view of the RNA structure [5].

Figure 2.2 illustrates an example of the aforementioned representations, depicting the identical RNA structure using three distinct visual formats.

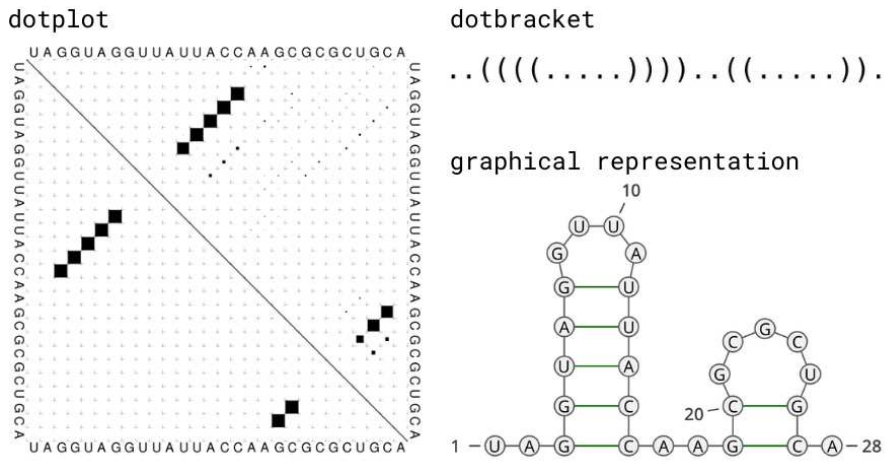


Figure 2.2.: **Different types of RNA representations.** The figure displays the same RNA sequence and secondary structure using three distinct representations, each presenting varying levels of structural and positional information.

2.1.5. RNA Design

RNA design is the process of creating RNA sequences that fold into specific secondary and potentially tertiary structures. This approach reverses the problem of RNA folding in which, given a sequence, the corresponding structure is predicted. Instead, starting with a desired secondary structure, the aim is to predict a sequence that adopts this structure, ideally as its MFE configuration. Therefore, RNA design is also called the inverse folding problem [18]. If there is a sequence that folds into the given secondary structure, it is called a design. Since the secondary structures of RNA sequences can play very functional roles, solving the inverse folding problem has applications in various fields such as pharmaceutics, biochemistry and synthetic biology.

The RNA design process can be divided into two principal paradigms: positive design and negative design. Positive design focuses on maximizing the sequence's affinity for the target structure. This involves increasing the stability and consequently lowering the sequence's free energy. In contrast, negative design aims to optimize the sequence by minimizing the formation of unwanted outcomes, such as unintended interactions or alternative structures. A robust design must integrate both paradigms: it should exhibit high affinity, ensuring strong thermodynamic stability within the target structure, and high specificity, meaning that the sequence is highly likely to fold exclusively into the target structure while energetically disfavoring alternative folds [19, 20].

Different approaches exist to solve the inverse folding problem. One promising method involves constraining the RNA sequence by specifying certain features, such as specific base pairs. These specified base pairs guide the RNA in folding into the desired structure. By progressively constraining the sequence, one can create a design with the desired structure as its MFE configuration [21].

2.1.6. Covariance Model

A Covariance Model (CM) is a statistical model that analyses RNA sequences and structures. It employs stochastic context-free grammars [22] and extends the concept of Hidden Markov Models [23] to RNA sequences. The primary goal of CMs is to identify covariances within an RNA group by analyzing a Multiple Sequence Alignment (MSA) of homologous RNAs. These covariances refer to compensatory mutations between paired bases in the RNA secondary structure, where changes in one nucleotide are often coupled with changes in its pairing partner to maintain the base pairing. Such simultaneous mutations reflect evolutionary pressures to preserve base pairing and overall structural conservation. These covariances are used to construct and train the CM. Once trained, the model can be applied to identify new RNA sequences that match its characteristics. This process enables the discovery of novel RNA sequences that fit the patterns observed in the MSA [24].

2.2. RNA Metrics

When RNA molecules are compared to each other, it can be helpful to obtain a numerical measurement to quantify the similarity or difference between the molecules. This task is inherently challenging due to the complexity involved in comparing three-dimensional objects, which can differ in shape and also in size. As a result, various metrics have been developed, each focusing on different aspects of the RNA structure.

Some methods rely on aligning the molecules as accurately as possible before com-

paring their shapes, which can become even more challenging when the objects are of different sizes. These alignment-based approaches, such as the root mean square deviation, ensure that structural features are optimally matched, enabling a more direct comparison. Other methods do not require alignment and instead compare RNA structures based on intrinsic properties.

Given that each metric targets different RNA characteristics, combining multiple metrics provides a more comprehensive understanding of molecular similarity. Metrics may evaluate aspects such as base-pairing interactions, the overall 3D conformation, or local structural motifs. Using a variety of metrics allows for a robust determination of similarity.

2.2.1. RMSD

The Root Mean Square Deviation (RMSD) [25] of atomic positions is a commonly used metric for structural comparison. Using the Kabsch Algorithm [26], the selected atoms of two (or more) molecules are aligned. Then, the Euclidean distance of the aligned atoms is calculated. The formula of RMSD is defined as:

$$RMSD(X,Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)} \tag{2.1}$$

where X is the reference and Y the predicted structure, and X_i or Y_i is the position of the i-th atom.

The resulting value, represented in Ångstrom (Å), ranges from 0 (all atoms are perfectly aligned) up to infinite, since the metric has no upper bound. Although RMSD can provide a fast intuition of conformational similarity, it does not provide any information about the discrepancy of local structures.

2.2.2. GDT_TS

The Global Distance Test Total Score (GDT_TS) [27] determines the similarity between two molecule structures by evaluating the extent to which the predicted structure aligns with the reference structure. The GDT measures how many alpha carbon atoms line up by superimposing the two structures. Using a specified distance cutoff, the percentage of superimposed atoms that fall into this cutoff is determined. The GDT_TS is the average of the percentages at four different cutoffs.

$$\text{GDT\_TS} = \frac{P_1 + P_2 + P_4 + P_8}{4} \tag{2.2}$$

where P_d represents the percentage of residues within the distance cutoff d in Å [28]. A higher GDT_TS score indicates a closer match to the reference structure, with a score of 1 indicating perfect alignment. Random predictions typically score around 0.2, while a score of 0.5 suggests a roughly correct topology of the molecule.

2.2.3. TM-Score

The Template Modeling Score (TM-score) [29] is a normalized metric designed to consider the residual alignment and also the distances between the two RNA molecules. Normalizing the distance makes it less sensitive to the size of the molecules, which has been shown to be a problem in the GDT and RMSD metrics. The TM score can be calculated as follows:

$$\text{TM-score}_{\text{RNA}} = \frac{1}{L}\sum_{i=1}^{L_{ali}} \frac{1}{1+\frac{d_i}{d_0}^2} \tag{2.3}$$

Where L is the length of the target RNA, L_{ali} the number of aligned nucleotides, d_i the distance between the i-th pair of aligned residues, and d_0 a normalization factor, which prevents the score from being dependent on the RNA length. This scaling factor is defined as:

$$d_0 = 0.6\sqrt{L-0.5} - 2.5 \tag{2.4}$$

The parameters for calculating d_0 were obtained from a vast PDB dataset of random pairs. The TM-score ranges from 0 to 1, where 1 indicates a perfect match between the two structures, and scores closer to 0 indicate little to no similarity.

2.2.4. CAD Score

The Contact Area Difference Score (CAD-score) [30] quantifies the similarity of intramolecular contacts between two molecules. Let the contact between residues i and j in the target structure be (i,j), with the corresponding contact area denoted as $T(i,j)$, and let $M(i,j)$ represent the contact areas in the model structure. The set of all contact areas in the target structure T is denoted as G. By calculating the absolute difference between the contact areas of the target T and the candidate structure M, the CAD gets determined:

$$CAD_{(i,j)} = |T_{(i,j)} - M_{(i,j)}| \tag{2.5}$$

To prevent any overestimation or underestimation of contact areas, a bounded version of the CAD is used:

$$CAD^{bonded}_{(i,j)} = \min(CAD_{(i,j)}, T_{(i,j)}) \tag{2.6}$$

The final CAD score is formulated as follows:

$$CAD\text{-}score = 1 - \frac{\sum_{(i,j)\in G} CAD^{bonded}_{(i,j)}}{\sum_{(i,j)\in G} T_{(i,j)}} \tag{2.7}$$

Ranging between 0 and 1, a higher CAD score implies a good prediction, with a CAD score of 1 indicating that the prediction and the target are identical. In contrast, a lower CAD score indicates a significant discrepancy between the reference structure and the model structure.

2.3. Orthoflavivirus

Orthoflavivirus (FV), formerly known as *Flaviviruses*, are single-stranded RNA viruses belonging to the family *Flaviviridae* [31]. FVs can infect birds and vertebrates, including humans [32]. Due to their high morbidity and mortality rates, they are significant human pathogens responsible for various epidemics in recent history [33]. Flaviviral infections can have various results, ranging from asymptomatic cases to severe diseases such as hemorrhagic fever or meningoencephalitis. The prominent members of FVs include *Zika Virus* (ZIKV), *Dengue virus*, *Yellow fever virus*, and *West nile virus* [34]. Due to the 2016 outbreak, the World Health Organization declared that Zika-related microcephaly is a public health emergency of international concern [35]. Since then, the ZIKV has spread rapidly, causing outbreaks worldwide [34]. ZIKV infections have been linked to severe neurological diseases such as Guillain-Barré Syndrome and congenital Zika syndrome, and infections can lead to severe complications during pregnancy [36]. Several factors contribute to the potential for FV epidemics, including the unique characteristics of their vectors, urbanization that creates ideal breeding habitats, and climate change. In addition, extensive global travel facilitates the spread of viruses and vectors. These conditions lead to the expansion of arthropods into new areas, bringing along the virus [34].

FVs are transmitted primarily through various vectors, with mosquitoes being the most important. Mosquitoes acquire the virus from a host during blood feeding and spread it to other hosts, often humans. In addition to mosquitoes, other species can also act as transmitters. Certain FV members can be transmitted among humans through blood transfusion or oral transmission, but the primary vectors remain invertebrates [37]. *Orthoflaviviruses* can be grouped by their transmission vector, such as mosquito-borne (mosquito borne flavivirus (MBFV)), tick-borne (tick borne flavivirus (TBFV)), insect-specific (insect-specific flavivirus (ISFV)) and those with unknown vectors (no known vector flavivirus (NKV)) [38].

2.3.1. Genome Organization

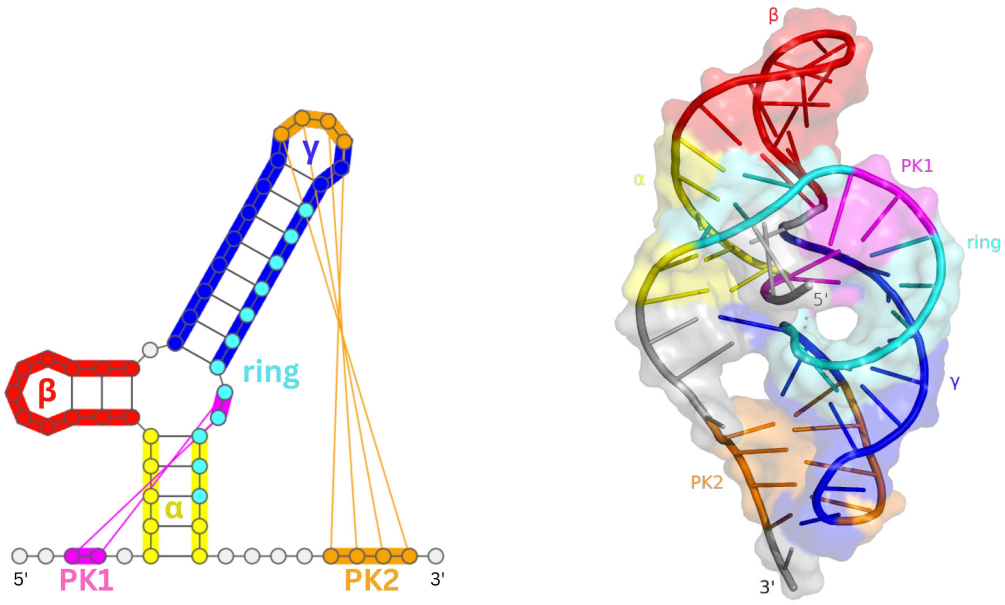
FVs are enveloped, positive-sense (+) single-stranded RNA viruses with a genome size of 10-11 kb. The RNA genome comprises a single open reading frame flanked by 5' and 3' untranslated regions. The open reading frame is translated into three structural proteins and seven non-structural proteins [34].

During viral replication, the host's Xrn1 degrades the vRNA. Xrn1 belongs to the family of 5' $\rightarrow$ 3' exoribonucleases and is highly conserved across species. Most eukaryotic mRNAs are degraded by Xrn1 [39]. This degradation by the Xrn1 proceeds until the enzyme encounters a specific region in the 3' UTR of the virus, where it stalls and appears to dissociate, leading to the production of shorter viral RNAs known as subgenomic flaviviral RNAs (sfRNAs) [40]. These sfRNAs are directly linked to diseases and play a crucial role in pathogenicity in fetal mice and in the occurrence of cytopathic effects in cultured cells through several proposed mechanisms [41]. sfRNAs can also bind to host proteins, potentially indicating their role in intervening in the infection process, including inhibiting the interferon response [42]. The RNA sequences responsible for the stalling of Xrn1 and the generation of sfRNAs are referred to as xrRNAs. Various studies have shown that this region of resistance to Xrn1 is present in several members of the FV genus, and alignments of xrRNA sequences have revealed high conservation [17, 41]. Furthermore, a single FV member can have multiple xrRNAs, producing varying sfRNAs [43].

2.4. xrRNA

The conserved structure of xrRNA has been classified into two distinct classes. Class 1 xrRNAs mainly include members of MBFVs, some NKVs, and potentially one ISV member, while Class 2 xrRNAs are predominantly found in TBFVs and some NKVs. This classification is based on their proposed secondary structures, the halt point of Xrn1 relative to these secondary structures, and the conservation patterns of the sequence [44]. Further refinement of this classification can be achieved by comparing tertiary structures, particularly for Class 1, as no three-dimensional structures are available for Class 2 [42].

Despite their differences, both classes share several similarities. Each class has three stems: α, β, and γ. The stems β and γ form a hairpin loop and, together with the stem α, these stems create a three-way junction. An illustration of these features is shown in Figure 2.3(a). The size of individual parts, such as the stem α or β, varies between the two classes.



(a) Visualization of the Class 1 xrRNA secondary structure.

(b) Visualization of the Class 1 xrRNA tertiary structure.

Figure 2.3.: **Structural example of Class 1 xrRNA member.** Panel (a) shows the secondary structure of xrRNA of ZIKV, a Class 1 xrRNA member of the MBFV group. Panel (b) presents the tertiary structure of xrRNA of ZIKV. Key structural elements are highlighted. The color coding indicates distinct structural features: the ring-like topology is shown in cyan, while pseudoknots PK1 and PK2 are highlighted in magenta and orange, respectively. The three conserved structural elements, α, β, and γ, are shown in yellow, red, and blue, respectively.

Both classes also display tertiary interactions that form two pseudoknots. These pseudoknots differ between the two classes. In class 1, the first pseudoknot (PK1) forms between the 5' end of the structure and a region within the multi loop, whereas in class 2 PK1 ranges from the multi loop to the 3' end. The second pseudoknot (PK2) spans from the hairpin loop of the γ stem to the unpaired 3' end of the sequence in both classes. The difference in PK2 between the two classes is defined by the number of unpaired nucleotides between the stem α and PK2 [45, 42]. It is noteworthy that, despite the presence of pseudoknots, the structure of xrRNA remains knot-free. This implies that during folding, no topological knot is formed within the RNA strand. Consequently, when enough tension is applied to both ends of the RNA strands, the structure unwinds completely, leaving a linear strand.

In addition to pseudoknots, significant tertiary interactions have been identified that contribute to the structural stability of the xrRNA. Among these, conserved base triples within the FV groups are particularly notable due to their potential role in xrRNA. Specifically, in MBFV xrRNA species, a highly conserved U·A-U base triple has been

identified. This base triple is involved in the mechanism to stall Xrn1, thereby preserving the functional RNA structure [44].

Moreover, a vital characteristic of the FV xrRNA structure has been observed, which could be essential for its resistance. The xrRNA forms a unique ring-like structure as a result of various tertiary interactions through which the 5' end of the RNA passes. This conformation is illustrated in Figure 2.3(b), where the ring-like structure is highlighted in cyan. This conformation is formed by structural elements from the γ and α stems, as well as one half of PK1. This structure is believed to be the main factor in resistance to xrRNA, as it mechanically blocks the Xrn1, preventing it from moving any further. This blockage causes Xrn1 to stall and eventually dissociate from the RNA, protecting the remaining sequence [44].
xrRNA or xrRNA-like structures are found in nearly all known FVs and can appear multiple times within a single virus. The xrRNAs located downstream of the primary xrRNA (xrRNA1) appear to function as a backup mechanism, ensuring resistance if the first xrRNA does not provide sufficient protection. Although some FV members have only one additional xrRNA (xrRNA2), others contain up to five xrRNA structures [46].

Sequential conservation has been observed alongside structural conservation. Based on the findings of Skibinski [17], highly conserved regions were identified within the FV groups. These conserved regions likely indicate essential characteristics of the sequence, as these motifs have been preserved over time. Consequently, these motifs are also significant for RNA design, as they can represent key properties of xrRNA.

2.4.1. MBFV xrRNA1

As previously described, MBFV xrRNAs belong to class 1. Comparative analyses of various MBFV sequences have revealed notable features of the secondary structure of MBFV xrRNA1. An exemplary MBFV structure is depicted in Figure 2.4, with highly conserved regions highlighted in red. The structure includes a fully conserved stem α, followed by a three-way multi loop that connects the stems α, β, and γ. This multi loop region is also highly conserved [17]. A notable conserved nucleotide is located between the β and γ loops, which may play a critical role in structural stability [45]. The MBFV xrRNA1 structure contains two pseudoknots. The first pseudoknot, which extends from the 5' end of the structure to the multi loop, is highly conserved. In contrast, the second pseudoknot exhibits less conservation and spans from the γ hairpin loop to several nucleotides at the 3' end of the structure [17].

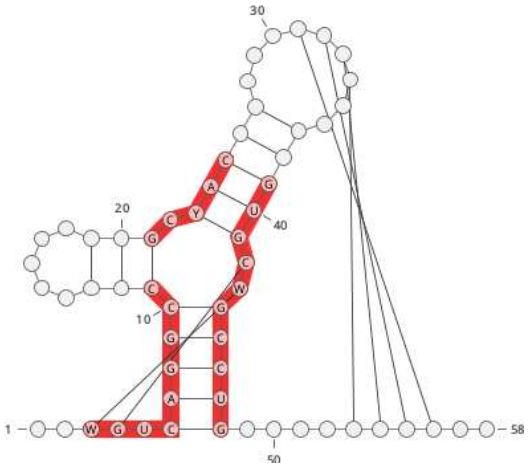


Figure 2.4.: **Secondary Structure of Members from Class 1 xrRNAs.** Depicted is a representative secondary structure of xrRNA1, including pseudoknots and highly conserved regions highlighted in red. These conserved regions exhibit at least 90% sequence identity, based on the sequence alignment presented in Figure 4.3 of [17].

2.4.2. TBFV xrRNA1

The xrRNAs of TBFV are classified as Class 2. Comparative analyses of TBFV sequences have identified conserved and likely essential features of the secondary structure of TBFV xrRNA1. An example of this structure is shown in Figure 2.5, where conserved regions are marked in red. A key feature of TBFV xrRNA1 is a highly conserved motif that forms a three-way junction. Pseudoknot 1 spans from this conserved motif to the unpaired 3' end of the structure. Furthermore, the stem γ is highly conserved and consists mainly of GC pairs, contributing to its stability. The second pseudoknot spans from the γ hairpin loop to the unpaired 3' end of the structure. Both PK1 and PK2 include highly conserved sequence motifs [17].

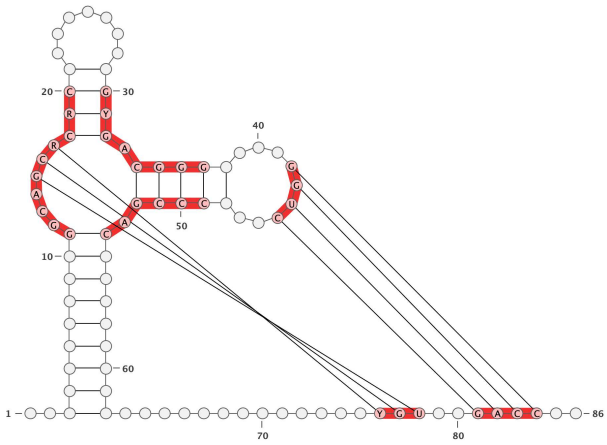


Figure 2.5.: **Secondary Structure of Members from Class 2 xrRNAs.** Depicted is a representative secondary structure of xrRNA1, including pseudoknots highly conserved regions highlighted in red. These conserved regions exhibit at least 90% sequence identity, based on the sequence alignment presented in Figure 4.16 of [17].

2.5. Software Tools

2.5.1. ViennaRNA

ViennaRNA [9] is a comprehensive software package designed to analyse and predict RNA secondary structures. It is widely used in computational biology for RNA research because of its robust algorithms and versatility in handling various RNA-related tasks. The primary function of ViennaRNA is to predict the secondary structures of the RNA based on thermodynamic models. RNA secondary structure prediction involves calculating the MFE structure for a given RNA sequence. The software can also compute the thermodynamic ensemble of possible structures, offering insight into the range of structures an RNA sequence can form. Furthermore, the package offers the computation of the partition function, which quantifies the probabilities of different structures in the ensemble and can be used to calculate the Base Pairing Probability (BPP) of a certain sequence. In addition to these core functionalities, ViennaRNA provides an extensive range of tools and functions, making it a versatile solution for addressing a variety of RNA-related problems. Its wide applicability has made ViennaRNA a cornerstone tool in computational RNA biology.

2.5.2. Infrared

Infrared [1] is a versatile framework designed to address a broad range of bioinformatics problems and can also be used for non-biological applications, enhancing its utility and power. Infrared uses tree decomposition techniques to efficiently solve declaratively

modeled problems. The framework provides an easy-to-use Python interface for problem modeling. Users can describe their problems using simple Boolean constraints, allowing them to focus on biological questions without the need to understand the underlying algorithmic implementations. This abstraction simplifies defining bioinformatics problems, streamlining the problem-solving process. Infrared also offers extensive documentation, with comprehensive tutorials and examples, making it highly user-friendly and easy to adopt. In general, Infrared simplifies the development of bioinformatics applications, enabling rapid and efficient problem solving. Its ability to handle complex data structures and its user-friendly Python interface make it a powerful tool for researchers.

2.5.3. SimRNA

SimRNA [47] is a powerful computational tool designed to predict and model the three-dimensional structures of RNA molecules. It uses a coarse-grained model in which the representation of the RNA is simplified to five atoms per ribonucleotide residue. SimRNA uses knowledge-based potentials, derived from statistical analyses of known RNA structures, along with Monte Carlo simulations, to explore the conformational space of an RNA sequence. This approach helps identify the most likely 3D configurations of the RNA. During Monte Carlo simulations, the RNA sequence undergoes a series of random movements and structural changes to achieve energetically favorable foldings. Through numerous iterations, the simulation converges on the most stable and probable 3D structures.

SimRNA does not produce a single outcome but generates an ensemble of RNA structures. This ensemble provides further insight into the nature of the RNA molecule, representing the range of structures the RNA sequence can fold into.

A useful feature of simRNA is its ability to incorporate constraints into the simulation. These constraints, which can be derived from computational data or other predictions, help refine the RNA structure by restricting the conformational space explored during the MC simulations.

2.5.4. QRNAs

QRNAs [48] is a computational tool designed to refine nucleic acid structures, focusing primarily on RNA molecules. Predictions of 3D RNA structures produced *in silico* often suffer from inaccuracies due to the simplifications used in these methods, such as coarse-grained simulations. The simulation can seem broken or be biologically impossible. Therefore, the initial structure of the prediction needs refinement to achieve a physically reasonable result. By incorporating detailed energetic and structural information, QRNAs enhance the precision of RNA models, making them more reliable for subsequent biological

analyses and applications.

2.5.5. PyMOL

PyMOL [49] is a powerful and widely used molecular visualization tool that allows the user to view, analyse, and manipulate structures of biological molecules such as proteins, small molecules, and nucleic acids. A key feature of PyMOL includes the ability to align RNA molecules or specific RNA regions, which is crucial for comparative studies of different molecules. Moreover, PyMOL can calculate the RMSD, providing a fast quantitative measure of the structural similarity between two molecules. These functionalities make PyMOL an essential tool for gaining insights into the structural relationships and dynamics of biological macromolecules. Furthermore, PyMOL can visualize molecular dynamics simulations, allowing one to observe molecules' conformational changes and dynamic behavior over time. In addition, hydrogen bonds can be identified and displayed, which are essential for understanding interactions within a molecular structure and its stability. In general, PyMOL is a vital tool for analyzing the structural relationships, dynamics, and interactions of biological macromolecules.

2.5.6. Infernal

Infernal ('INFERence of RNA ALignment') [50] is a bioinformatic tool used to analyse RNA sequences and find similarities. Unlike other tools that focus only on RNA sequence conservation, Infernal also considers the conservation of RNA secondary structure, which provides important insights into their relation. Infernal creates a CM that captures the secondary structure and sequence consensus of a group of aligned RNAs. This model can then be used to search databases for RNAs with similar properties. Infernal also includes additional tools for enhancing or using CM in various ways. For example, the program `cmemit` can generate new RNA sequences based on a provided CM, producing sequences that reflect the key characteristics of the original RNA group used to build the model.

2.5.7. LocARNA

LocARNA [51, 52, 53] is a software package designed for fast and high-quality alignment of RNA sequences. The package consists of several tools for the structural analysis of RNAs. One of these tools, called MLocarna, can be used to generate MSA from several unaligned RNAs and produce a consensus structure based on the MSA. MLocarna's strength lies in incorporating RNA secondary structure additionally to the sequence into the alignment and allowing users to extend the alignment by specifying structures or anchoring constraints.

3. Materials and Methods

3.1. Workflow

An overview of the xrRNA design workflow is illustrated in Figure 3.1. The RNA design model was iteratively developed, and the newly designed sequence offered insights that refined the model. The process began by creating and integrating a sequence and structure scaffold based on biological data, as described in Sections 3.4.1–3.4.3. This step enabled the model to generate sequences that reflect the characteristics of the biological data. Next, the generated sequence was optimized using a Monte Carlo approach, described in Section 3.4.4. After optimization, the sequence should satisfy all the criteria set for the secondary structure of the xrRNA.

Since, at this stage, no information about the tertiary structure of the design is available, further investigation was needed to confirm the similarity of the design to the biological xrRNA structures. Using SimRNA simulation and QRNAs refinement, a 3D model of the designed sequence was obtained. Various metrics were then used to assess the differences between the RNA design and the target structure. Analyzing the optimized sequence and comparing it with a reference structure revealed potential problems of the model and led to further refinement. If available, comparisons were made with experimentally obtained data; if not, the design was compared to an optimized artificially generated reference structure. Visualization of the RNA structure using PyMOL helped determine the presence of the ring-like structure and other important structural features.

If the designed RNA did not fulfill the xrRNA condition, such as the absence of a ring-like structure or a missing pseudoknot, the design process was revisited for further optimization. Insights gained from the identification of design issues were used to adapt the Infrared model, further constraining or refining the RNA layout. This iterative process continued until a satisfactory RNA sequence was obtained displaying the desired tertiary structure.

Once the 3D simulation of the RNA sequence appeared satisfactory, the stability of the design was evaluated through Molecular Dynamics (MD) simulations. These simulations, provided by Leonhard Sidl, were analysed and compared to simulations of reference structures.

The analysis of MD simulations provided further insight into discrepancies between the designed sequence and the reference sequence. These observations were then incorporated into the Infrared model, further optimizing the xrRNA design. If the results of the MD simulation were comparable to those of biological sequences, the designed sequence was ready for *in vitro* experiments.

Before conducting experiments, a leader sequence had to be designed. This leader sequence is necessary to allow efficient loading of Xrn1, which degrades the RNA in the 5' to 3' direction. This leader had to meet specific constraints to prevent complications during the experiments. Once a satisfactory leader was designed and added to the sequence, the sequence was sent to the partner laboratory for experimental validation of its resistance to Xrn1, as described in Section 3.7.

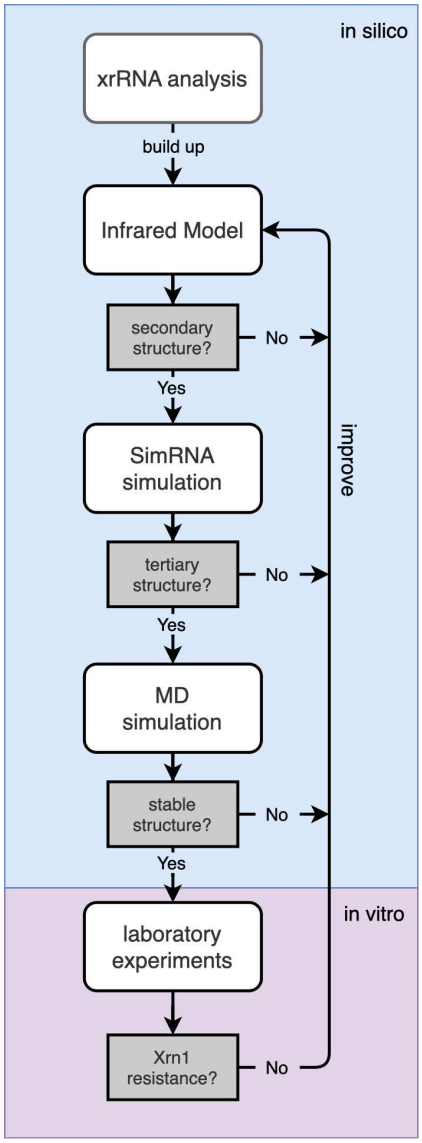


Figure 3.1.: **xrRNA Design Workflow.** A flowchart representing the iterative process of creating the xrRNA design model.

3.2. Materials

The data used in this study was collected and provided by Denis Skibniski, who, as part of his master's thesis *'Computational studies of exoribonuclease-resistant rna structures in flaviviruses'* [17], computationally analysed xrRNA sequences within the FV family. In his study, he identified FV sequences along with their corresponding xrRNA structures. For each FV subgroup, a single sequence was selected to generate an MSA, from which

a consensus structure was predicted. This procedure was performed for all ecological FV groups. For each FV group, the sequences, MSA, and structure predictions were made available for further use. These files enabled the study of the structural properties of xrRNA. Furthermore, the consensus sequence and structure for each FV group, as provided by Skibniski, were incorporated into the RNA design process.

Most of the currently available information on xrRNA is related to the subgroups MBFV and TBFV. Consequently, this study focuses on these two groups. Ideally, biological data, such as tertiary structures of certain xrRNAs, would be available to aid in the quantification and evaluation of designed sequences. However, only four crystal structures of flaviviral xrRNA have been resolved to date: *Murray Valley Encephalitis virus* (MVEV), *Tamana Bat Virus* (TABV), and two structures of ZIKV, with the corresponding PDB IDs 4PQV, 7K16, 5TPY, and 7U4A, respectively. The structure of ZIKV has been determined from both a wild-type sequence (5TPY) and an xrRNA1 mutant (7U4A). In the xrRNA1 mutant (7U4A), pseudoknot 2 was weakened by replacing the second and fourth G-C base pairs with wobble G-U pairs, achieved through changing the C residues at positions 55 and 57 to U [54]. The crystal structure of TABV was obtained from a wild-type RNA sequence [42], while the MVEV structure includes a single mutation at the start of the sequence, as indicated in the PDB entry [41].

MVEV and ZIKV are classified within the MBFV family, while TABV does not appear to align with these clusters and is proposed to be assigned to its own distinct cluster [55]. The crystal structure of ZIKV was mainly used as reference for comparison with the MBFV-based designs.

Unfortunately, no equivalent data was available for TBFV or other *Orthoflavivirus* families. Therefore, an *in-silico* generated and refined structure of *Deer Tick Virus* (DTV), which belongs to the TBFV family, was used as a reference for the TBFV-based designs.

3.3. xrRNA Analysis for RNA-Design

After obtaining the corresponding *Orthoflavivirus* subgroup data, the data set was cleaned and filtered to include only sequences associated with xrRNA1. To create an average xrRNA structure, non-conforming xrRNA structures to the conserved xrRNA1 were removed.

Next, the lengths of the characteristic parts of xrRNA1 for the MBFV family were analysed, specifically examining the following regions: stem α , stem β, hairpin loop β, stem γ, hairpin loop γ , multi loop, length of PK1 and PK2, nucleotides between the α

stem and PK2 (spacer), and the positioning of the PK2 within the γ hairpin loop. The corresponding data can be found in Section A.1, Table A.1 and a representation of these structural regions can be seen in Figure 3.2.

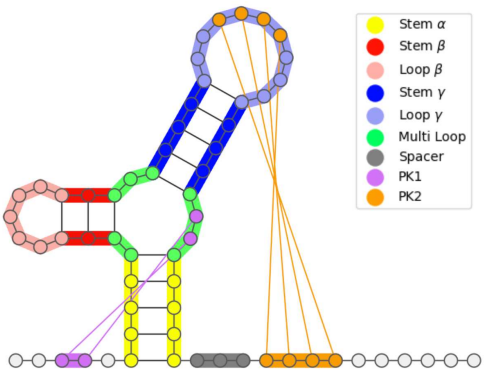


Figure 3.2.: **Segmentation of MBFV xrRNA structures into distinct regions.** Each panel displays the xrRNA structure divided into individual regions, with each segment highlighted in a different color to provide a clear overview of the specific structural parts. The coloring scheme is consistent with that used in the visualization of ZIKV in Figure 2.3.

Similarly, for the TBFV family, the lengths of the corresponding regions of xrRNA1 were analysed, including stem α and its internal loops or bulges, stem β and its internal loops or bulges, hairpin loop β, stem γ and its internal loops or bulges, hairpin loop γ, multi loop, nucleotides between stem α and PK1 (spacer), nucleotides between PK1 and PK2, positioning of the PK2 within the hairpin loop γ, length of PK1 and PK2. The corresponding data can be found in Section A.1, Table A.2 and a representation of these structural regions can be seen in Figure 3.4.

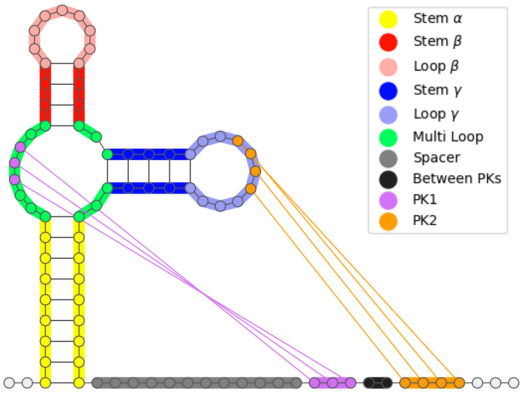


Figure 3.3.: Structural parts of TBFV.

Figure 3.4.: **Segmentation of TBFV xrRNA structures into distinct regions.** Each panel displays the xrRNA structure divided into individual regions, with each segment highlighted in a different color to provide a clear overview of the specific structural parts. The coloring scheme is consistent with that used in the visualization of ZIKV in Figure 2.3.

The minimal and maximal lengths for each of these characteristics were determined to construct the xrRNA design scaffold. Additionally, Spearman rank correlation coefficients were calculated between individual features to identify any correlations between the distinct parts of the xrRNA. The analysis was performed using the lengths of the individual structural parts of the xrRNAs, as presented in Table A.1. The Spearman rank correlation test [56] is a nonparametric measure used to assess the strength and direction of a monotonic relationship between two variables. When calculating the Spearman rank correlation coefficient (ρ), each value in the data set is replaced by its rank. Tied values receive the average rank. The differences between the ranks of corresponding values (d_i) are computed, and the coefficient is calculated using the formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.1}$$

where n is the number of observations. The coefficient ranges from -1 to 1, with 1 indicating a perfect positive monotonic relationship, -1 indicating a perfect negative monotonic relationship, and 0 indicating that there is no monotonic relationship. These correlations, if present, can be incorporated into the xrRNA design to enhance its precision and reliability.

3.4. xrRNA-Design

The implementation aims to generate sequences with the same properties as natural xrRNA structures. The initial step involved collecting all available information on the existing xrRNA structures and their corresponding sequences, which was described in Section 3.3. Since xrRNA structures vary between groups, each RNA design model focuses solely on a single group. Separate scripts were created for the two groups MBFV and TBFV to design RNA sequences that exhibit the specific xrRNA characteristics of the corresponding group.

The general framework of the xrRNA design is based on the structures and sequences of the *Orthoflavivirus* species. The model scaffold should contain the necessary essential features of the conserved region, but still have enough variability to create sequences significantly distinct from the available *Orthoflavivirus* sequence data.

3.4.1. Infrared Modelling

Infrared offers several advantages for defining detailed and customized problems as models. It uses variables, domains, constraints, functions, and features to construct a network model. The model is set up by first specifying the variables and their corresponding domains. A domain is the set of possible values that a variable can take within the model. In the context of this design problem, the possible values for each variable are the four RNA nucleotides `A, C, G, U` plus a gap `-`. To create the model, we define `N` variables, where `N` represents the length of the sequence. Each variable can take one of five values (`A, C, G, U, -`), forming the sample space for the model. Next, we can further refine this sample space by introducing constraints. Constraints are logical expressions that restrict variables to take certain values. These constraints can represent various structural or functional requirements of the RNA sequence, such as specific base pairs or specific sequential motifs. Infrared further allows for the integration of functions that compute values based on the model's variables. For example, a function could output `0` if the given variable is a gap (`-`) and `1` otherwise, enabling the computation of the sequence length. These functions form the foundation for defining features. A feature represents a measurable property or characteristic of a sequence and can group together one or more functions. Each feature can be assigned a specific weight, allowing weighted sampling. By adjusting the function's weight, we can influence the sampling process, thereby manipulating the sequence design to meet desired properties. Using the example from before, a feature could evaluate the length of the sequence by grouping functions that count non-gap bases across all variables. The assignment of a specific weight to this feature allows the model to control the sequence length, thereby adjusting the design process to meet the desired specifications.

3.4.2. Sequence and Structural Constraints

Sequence constraints consist of the IUPAC code and an additional token `X`, which can be replaced by any of the four RNA nucleotides or a gap. The framework for the sequence then consisted of hard constraints and variable parts. The hard constraints are derived from conserved sequences in xrRNA structures within each specific family, based on the work of Skibinski [17] and also mentioned in Sections 2.4.1 and 2.4.2. These hard constraints enforce specific motifs in certain regions of the sequence. For example, the stem α in the MBFV family is highly conserved; therefore, it is hard constrained into the sequence scaffold. Other parts of the sequence are less conserved and show high variability. Therefore, these regions are filled with the `N` and `X` tokens and can be replaced by any of the four nucleotides.

The overall xrRNA structure consists of both secondary and tertiary interactions. As shown in Figure 3.2 and Figure 3.4, the characteristic structure of xrRNA can be divided into the following regions: stem α, stem β, hairpin β, stem γ, hairpin γ, PK1 and PK2. The region between the stem α and the 3' end can be further divided into three sections: nucleotides up to PK1 or PK2, nucleotides between PK1 and PK2, and nucleotides after PK2. The part within the hairpin γ can also be divided into three parts: PK2, unpaired nucleotides towards the 5' end and towards the 3' end of PK2. Together, these regions make up the complete structure of the xrRNA.

Since the lengths of these structural elements vary between species, the model must account for this variability. To accomplish this, the structural scaffold must exhibit flexibility in length. Based on data from FV, each structural element is constrained by the minimum and maximum lengths observed. This flexibility is modeled using the predefined `X` and `N` tokens. For each structural segment, the token `N` is used to reach the minimum length for a particular family, followed by the addition of `X` tokens to reach the maximum observed length for each segment. This approach allows for variable lengths between the observed minimum and maximum values for each region. By applying this method to each defined structural segment, we create a flexible framework in which each structural element obeys the observed length range for a given FV subgroup. However, since the β and γ structures are subdivided into stems and hairpin loops, the total length of these structures may occasionally exceed the observed range. Thus, additional constraints are applied to ensure that the total length of the structures β and γ falls within the correct range. These constraints are implemented directly within the Infrared model.

The overall sequence length is controlled by a function integrated into the Infrared model. It can be fixed to a specific value, constrained within a defined range, or left unspecified. When no explicit length is provided, a targeted sampling approach is used to

determine the relevant parameters for the function. In this process, the design length is derived by sampling from the model with a set target length and a specified tolerance. This sampling generates a weight, which is then used in the length function, resulting in sequences that maintain a length distribution similar to those in which a specific length and tolerance were initially set.

Furthermore, structural constraints, comprising secondary structure and pseudoknots, are added to the Infrared model. This structure is derived from the conserved structure of the corresponding FV family. These structural constraints ensure that the nucleotide positions intended to form a base pair can do so, allowing canonical Watson-Crick pairs and the non-Watson-Crick pair G-U.

3.4.3. Further Constraints

Since gaps can be inserted at any position marked by `X` in the sequence, it is possible to generate the same sequence with gaps at different positions. For example, sequences such as `-GAC-` and `G-A-C` are distinct in terms of gap placement, but yield the same sequence `GAC`, once the gaps are removed. To avoid redundant representations of identical sequences, specific constraints on gap positioning have been implemented in the Infrared model. Within a loop region, gaps are permitted only at the 3' end and must be continuous, with no nucleotides interrupting the sequence of gaps. For example, `NNN-` is allowed, but configurations like `-NNN` or `N-N-` are not. In stem regions, gaps are allowed only on the 'inner' side, where a gap is acceptable if the complementary nucleotide in the base pair is also a gap. As a result, each base pair in the stem consists of either two nucleotides or two gaps. Consider the secondary structure `((.))` as an example, assuming this is purely illustrative. With the specified gap constraints, the model could produce sequences such as `NNNNN` or `N-N-N`. However, sequences with gaps on the outer side of the helix or with gaps on only one strand, such as `-NNN-` or `N-NN-`, are not permitted.

These constraints ensure that the Infrared model generates unique sequences. Moreover, these constraints are also crucial for ensuring the correctness of uniform Boltzmann sampling. By enforcing these constraints, the model prevents depicting the same sequence in multiple ways, which would otherwise bias the sampling process. Thus, the gap-placement constraints ensure accurate and unbiased sampling.

The predicted energy of PK2 was included in the MBFV-based model to ensure stable base pairing. A function was added to the Infrared model to calculate the base pairing energy of PK2. By assigning a negative weight to this function, the model aims to minimize the energy, thus generating sequences that form a stable pseudoknot 2.

If significant correlations between different regions of the xrRNA were identified, as described in 3.3, they were incorporated into the model. The interquartile range of length ratios between corresponding RNA regions was calculated for these correlations. When designing a new sequence, the ratio between these lengths was checked and the design was accepted if the ratio fell within the interquartile range. This method allowed for further refinement of the RNA design based on available data.

3.4.4. Optimization

Using Infrared and the method described earlier, we defined a sample space for possible RNA sequences. Infrared can generate samples from this space until it meets a specific criterion or iteratively optimize a given sample within the defined sample space until it satisfies the requirements. Given the complexity of our design problem, simply drawing samples from the space was unlikely to yield satisfactory results. Therefore, an optimization approach was chosen. Specifically, a Monte Carlo approach was employed, a method widely used in computational biology to solve complex optimization problems.

Monte Carlo simulations use random sampling to explore possible solutions to problems, such as in simulations or optimizations. In this context, the sequence is iteratively modified, and an objective function evaluates the 'fitness' or quality of each new sequence. Based on this fitness score, the modified sequence is either accepted as the current solution or rejected. The general structure of this MC approach is adapted from the Infrared tutorial [57]. In this MC method, two parameters must be defined: the step size and the temperature. The size of the steps sets the number of sequence modifications made. The temperature controls the likelihood of accepting a sample with a lower fitness score, enabling exploration of diverse possible solutions rather than getting stuck in local optima.

The MC algorithm can start from a predefined sequence. If no starting sequence is provided, a random sequence is drawn from Infrared's sample space, set as the current sequence, and saved as the initial 'best sample'. For each MC step, a part of the current sequence is randomly selected and altered to generate a new sequence. This new sequence is evaluated by the objective function to determine its fitness score. If the fitness of the new sequence is better than that of the current sequence, it is accepted as the new current sequence. If the new sequence also has a better fitness than the 'best sample' so far, it is saved as the 'best sample'.

If the fitness of the new sequence is lower than that of the current sequence, it may still be accepted with a probability defined by the function $e^{\frac{\text{newval}-\text{curval}}{\text{temp}}}$, where *newval* is the fitness of the new sequence, *curval* is the fitness of the current sequence, and *temp* is the temperature parameter. This acceptance probability allows for occasional acceptance of

lower-fitness solutions, which helps avoid getting trapped in suboptimal solutions.

If a new sequence is accepted, the current sequence is updated to this new sequence. If it is rejected, the process restarts from the current sequence to create a new variation. This iterative process continues for a specified number of steps, usually between 50,000 and 250,000 iterations, until a satisfactory solution is achieved.

The objective function used in MC optimization has been continuously refined to improve the RNA design. Ward et al. [58] provide an overview of fitness functions for RNA design, recommending the use of Equilibrium Probability (EP) and Ensemble Defect (ED) as objective functions in the optimization process.

The primary goal of the optimization is to generate a sequence that reliably folds into the desired target structure. Thus, maximizing the equilibrium probability is an effective approach to achieving this goal. The EP represents the probability of a sequence folding into the specified structure, with values ranging from 0 to 1, where 1 indicates the highest likelihood of correct folding. Calculating EP involves using the thermodynamic equilibrium partition function for RNAs within a thermodynamic model [59]. The formula of the equilibrium probability $P(s)$ is given as [60]:

$$P(s) = \frac{e^{\frac{-E(s)}{RT}}}{Z} \quad \text{with} \quad Z = \sum_{s} e^{\frac{-E(s)}{RT}}$$

Here, $P(s)$ denotes the probability of observing a specific RNA secondary structure s in the thermodynamic ensemble. $E(s)$ is the free energy of structure s, R is the universal gas constant, T is the absolute temperature in Kelvin.

The optimization also utilizes ensemble defect, a measure first described by Dirks et al. [19]. The term 'ensemble' refers to the set of all possible secondary structures an RNA sequence can form, along with their associated probabilities. ED quantifies the average number of nucleotides that fail to pair as intended in the target structure, taking into account the probabilities of all possible structures in the ensemble. The formula of the ensemble defect $ED(s)$ is given as [60]:

$$ED(s) = 1 - \frac{1}{n} \sum_{ij,(i,j)\in s} p_{ij} - \frac{1}{n} \sum_{i} (1 - s_i) q_i$$

Here, s denotes the target secondary structure, and n is the length of the sequence. The term p_{ij} represents the probability of the base pair (i, j) in the ensemble, and q_i is the probability that nucleotide i is unpaired. The indicator variable s_i is defined such that

$s_i = 1$ if nucleotide i is paired in the target structure s, and $s_i = 0$ otherwise.

The optimization process occurs in two stages: minimizing the ED and maximizing the EP. In the first stage, ED is minimized over the initial half of the total optimization steps. The sequence resulting from ED minimization serves as the starting point for the second stage. In the second stage, the EP is maximized over the remaining steps. By the end of these two optimization stages, the final sequence should ideally exhibit a low ED and a high EP.

The following functions of the ViennaRNA package [9] were used to calculate the EP and ED:

- `fold_compound()`: This class initializes a fold compound object for a given RNA sequence, providing the necessary data for subsequent calculations.
- `pf()`: This function computes the partition function for a given RNA sequence, enabling the calculation of the probabilities of all possible secondary structures within the structure ensemble.
- `pr_structure()`: This function calculates the equilibrium probability from a given secondary structure, providing insights into the likelihood of specific secondary structure.
- `ensemble_defect()`: This function computes the ensemble defect for a given RNA sequence and target structure, estimating the average number of incorrectly paired nucleotides in the structure ensemble compared to the target structure.

3.5. Tertiary Structure Analysis

After creating a design using the model described above, the next step is to determine its tertiary structure. To achieve this, SimRNA is used to generate a three-dimensional simulation of the design. As explained in 2.5.3, SimRNA produces a 3D model of an RNA molecule by processing the input sequence, while incorporating structure constraints to guide the simulation. The input sequence is derived from the designed sequence by removing all the gaps. The structural constraints include the secondary structure and the two pseudoknots. These constraints are also based on the designed sequence, ensuring consistency with the input sequence. Specifically, the gaps removed from the sequence are also removed from the target structure, resulting in an input structure that aligns with the input sequence used for the SimRNA simulation.

The simulation is then configured with the following parameters:

- `NUMBER_OF_ITERATIONS`: 20000000
  Specifies the number of iterations for the annealing simulation.
- `TRA_WRITE_IN_EVERY_N_ITERATIONS`: 20000
  Sets the number of iterations that must pass before appending the conformation to the trajectory file.
- `INIT_TEMP`: 1.65
  Defines the starting temperature of the simulation.
- `FINAL_TEMP`: 0.60
  Defines the ending temperature of the simulation.
- `BONDS_WEIGHT`: 1.0
  Controls the strength of the bonds holding the structure together.
- `ANGLES_WEIGHT`: 0.60
  Regulates the response of planar angles to bending forces.
- `ETA_THETA_WEIGHT`: 0.40
  Influences helical orientation of the RNA strand.

The configurations for `BONDS_WEIGHT`, `ANGLES_WEIGHT`, and `ETA_THETA_WEIGHT` are set to their default values. Furthermore, simulations are conducted using the Replica Exchange Monte Carlo method, activated with the flag `-E`. The number of replicas is set to 16, which results in the following command:

Listing 3.1: SimRNA simulation

```
SimRNA -s inputSequence -S inputStructure -c configFile -E 16
-o outputFile 2> errorFile
```

This command creates 16 parallel trajectories based on the specified configuration and input sequence with the corresponding structural constraints. For each input, the simulation is run 25 times, producing multiple trajectories. After all simulations are completed, the trajectories are clustered using a clustering program provided by SimRNA. This program group similar structures together.

Listing 3.2: SimRNA clustering

```
clustering all.trafl 0.01 5
```

Here, `all.trafl` is a concatenated file of all generated trajectory files. The argument `0.01` specifies that only 1% of the the lowest energy frames from the input `all.trafl` will be used for clustering, while `5` is the RMSD threshold for the clustering.

After clustering, each cluster's trajectory files are converted to PDB format using the

`SimRNA_trafl2pdbs` command. This step results in several clusters, each containing the trajectory and PDB files that belong to that cluster.

Once the clustered structures from the SimRNA simulation have been obtained, they are further analysed. Clusters are ordered by the number of structures they contain, with the first cluster holding the largest number of conformations. The relative frequency of each cluster is calculated to gain deeper insight into the distribution and stability of structural variations.

Using PyMOL to visualize the 3D simulation, an intuition can be obtained about the potential of the design. The absence or presence of the ring-like structure can be easily determined without comparison to a reference structure. Afterward, the designed sequence is aligned with a reference structure, either to experimentally obtained data in the case of MBFV designs or to an optimized artificially generated reference structure in the case of TBFV designs. The differences between the designed sequence and the reference show essential insight into the current model problems. Depending on the relative frequency of the clusters, additional simulations of the top clusters were visualized to further analyse the results. The different metrics described the Chapter 2.2 gave a numerical measurement of the difference between prediction and target structure. The TM-Score and RMSD were calculated using the SWISS-MODEL Structure Assessment tool [61], while the CAD-score was determined using the web tool provided by the Vilnius University Institute of Biotechnology [62]. Using the metrics together with the visualization and the characteristic of the design contributed to an analysis that led to further restrictions and improvements incorporated into the design model.

3.6. Stability Analysis

Various MD simulations were provided to analyse structural stability. These simulations aim to mimic the Xrn1 degradation process by applying a pulling force to the 5' end of the xrRNA structure. Given the knot-free nature of the structure, as mentioned earlier, pulling on the 5' end is expected to lead to its unwinding. The force required to initiate this unwinding is therefore a key factor to examine. The simulations were conducted using a range of constant forces (CFMD) or a gradually increasing force (GFMD). Comparing these simulations with reference simulations provided more profound insights into the stability of the design and helped assess whether it is Xrn1 resistant.

CFMD simulations were performed with pulling forces of 200, 300, 350, 400, and 500 pN. For each force, four simulations were generated. In each simulation, the initial position of the 5' end was determined and monitored throughout the process. At each time step,

the distance between this initial point and the current position of the 5' end was calculated. The distances from all four simulations for each force level were then compiled into histograms. These histograms provide insight into the RNA behavior at each pulling force.

A peak in a histogram indicates a distance that frequently occurred during the simulation, suggesting a conformational state that the RNA adopted for an extended period. Thus, such peaks can serve as indicators of structural configurations. The area of each peak provides additional information, with taller peaks indicating longer-lasting configurations and shorter peaks pointing to short, transient states. The peaks appearing within a distance range of 0 to 10 Å indicate that the RNA remained in its initial configuration. The presence and height of other peaks at greater distances suggest whether the RNA refolded into alternate configurations or unfolded. Therefore, by examining these peaks, the structural stability of the RNA can be estimated, as well as its resistance to unfolding.

Furthermore, the distances between specific base pairs provided additional information on structural stability. Special attention was paid to crucial base pairs, such as those involved in PK1 and PK2. These distances indicate how long the base pair remains intact and when the bond starts to break. Combining these analyses with visualizations of the simulations helps identify potential problems in the design. It provides insights into possible solutions, which could then be included again in the design model.

Since the histograms are less intuitive to interpret, making it challenging to analyse structural dynamics, also GFMD simulations were employed to gain deeper insights into the structural behavior. During these simulations, the position of the 5' end was tracked continuously. The distance between the initial position of the 5' end and its position at each time step was computed and used for subsequent analysis. Distances below 10 Å were interpreted as the molecule retaining its initial conformation, indicating that there were no significant structural rearrangements. Conversely, abrupt increases in this distance were indicative of refolding or unfolding events within the RNA structure. Monitoring the 5' end provides an intuitive and efficient way to assess the stability of the structure.

3.7. xrRNA Resistance Testing

If the designed sequence satisfies the requirements of the secondary and tertiary structure, and the MD simulation yielded satisfactory results, it can be sent to the lab for *in vitro* tests to determine its resistance to Xrn1. A leader sequence must be added to the designed RNA sequence for these experiments. This leader must meet specific criteria to be considered suitable. One requirement is that the leader must start with a specific sequence. Additionally, the leader should not consist of a simple PolyA strand. Most

importantly, the leader sequence must not interact with the xrRNA sequence. After creating a leader, which meets the set requirements, it can be sent to the partner lab, where the resistance of the Xrn1 sequence is tested. For this, in-line probing [63] and Xrn1 resistance assays [41] were used.

3.8. Design Comparison

The primary goal of this study is to demonstrate how RNA can be designed to possess specific structural and functional characteristics. Therefore, it is essential to highlight how the choice of tools improved the design process. For this purpose, the Infrared software tool was chosen because of its user-friendly interface and the ability to generate precise, custom solutions for complex RNA design challenges. The RNA designs generated with Infrared were compared to those generated using a CM-based approach, implemented through the Infernal software. Infernal was selected for this comparison because of its capability to produce sequences based on a CM. By comparing these models with biological data, we can assess how well each model performs in replicating real-world RNA structures and characteristics and its potential flexibility for further RNA design applications, particularly in future xrRNA design.

3.8.1. Covariance Model Design

The CM was created to generate sequences based on the structural and sequential characteristics of a specific RNA group. To construct this CM, the sequences and their corresponding secondary structures of an xrRNA group were aligned to create an MSA. The MSA was generated using `mlocarna` from the LocARNA package [51, 52, 53], which aligns the sequences by integrating both sequential and structural information, which is crucial for building a precise CM. Consistent with previous steps, the same sequences used for the Infrared model's xrRNA scaffold were employed. Using this MSA, the CM was constructed with `cmbuild`. The CM can then generate sequences using the `cmemit` function, with the `-a` flag added to store the sequences in an aligned format for subsequent analysis.

3.8.2. Sequence Diversity

Sequence diversity is a critical factor in comparing RNA design approaches. To evaluate this, the designed sequences generated by the Infrared and CM-based methods were aligned to create MSAs. For the Infrared-generated sequences, two alignments were prepared: one directly based on structural information derived from the target structure, and another using `mlocarna`, which incorporates both structural and sequential information to create the alignment. In contrast, the RNA sequences designed using the CM approach did not

require additional alignment adjustments, as they were already in MSA format due to the appropriate flag being applied during the design process. From the CM MSA and the Infrared MSA generated using `mlocarna`, sequence logos were created to visualize nucleotide diversity at each position. In these plots, the height of each nucleotide at a given position represents its frequency, providing a clear representation of the variability of the designed sequences. Additionally, using the CM MSA and the target-structure-based MSA of Infrared designs the Shannon entropy was calculated for each position in the alignments using the method described by Schmitt [64]:

$$H(x) = -\sum_{i} p_i \log p_i \tag{3.2}$$

where i represents all possible symbols in the alignment (`A, C, G, U`), and p_i is the probability of symbol i occurring at position x.

A high entropy value at a given position indicates high variability, which means that different nucleotides frequently appear across the alignment. The maximum entropy, 2, occurs when all symbols are equally likely. Conversely, a low entropy value suggests conservation at a position, with one nucleotide consistently appearing. The minimum entropy, 0, represents positions where only a single nucleotide appears in all sequences.

3.8.3. Structure Diversity

The structural diversity of the predicted RNA designs was analysed to evaluate the performance of the two design approaches. First, the lengths of the structural parts were compared, as described in Section 3.3. This analysis aims to determine whether the models successfully capture the length distributions of different structural components.

In addition, the base pairing probabilities for each design were calculated using ViennaRNA. These probabilities were visualized as heat maps, illustrating the likelihood of each potential base pairing. By comparing these probabilities, we gain insight into the structural stability and diversity of RNA designs from both approaches, thereby providing a comprehensive evaluation of the performance of each method. Additionally, this analysis highlights alternative structures that frequently emerge in the model designs, offering further insights into their structural variability and potential conformational landscapes.

Given the importance of pseudoknots in the formation of the xrRNA structure, they were specifically analysed. To focus on pseudoknot formation, regions that did not contribute to the base pairing of pseudoknots were constrained to remain unpaired using ViennaRNA, allowing only potential pseudoknot regions to form base pairs. The MFE structure of the constrained sequences was then calculated to assess whether a design could form

base pairs between these regions. The occurrence of base pairs and the number of base pairs formed were recorded. Special consideration was given to pseudoknot 1 in MBFV, which consists of only two base pairs. Since ViennaRNA can fail to predict such short base-pairing interactions, each design was manually evaluated to determine whether the nucleotides at the PK1 position could form Watson-Crick pairs or wobble base pairs.

Finally, the distribution of design lengths was analysed and compared to the length distribution of the original dataset, adding another metric to evaluate the similarity of the designs to the biological data.

4. Results

This chapter presents the results of the analyses and simulations conducted during this master's thesis. The workflow outlined in Section 3.1 serves as the basis for the following discussions. First, structural and sequence analyses of the xrRNAs from the MBFV and TBFV subgroups are presented. These analyses provide the foundational information necessary for the RNA design model. The resulting data are then used to develop the Infrared RNA design model.

Next, the construction of the RNA design model is described, detailing the specific constraints applied for MBFV and TBFV xrRNAs. In addition, the methodology for generating new RNA sequences using the design model is explained.

Finally, the generated sequences are analysed to evaluate their mimicry of the structural and functional properties of the MBFV and TBFV xrRNAs. This includes an assessment of tertiary structure and structural stability using 3D and MD simulation. After the model's capability to produce sequences exhibiting xrRNA-like properties was validated, the proposed design approach was compared to an alternative RNA design method. This comparison aims to evaluate the different design models and their similarities to natural sequences.

4.1. xrRNA Analysis

An analysis of xrRNAs was performed for the MBFV and TBFV groups. The goal was to identify the structural features shared by each group and use this information to develop group-specific RNA design models.

4.1.1. Structural Characteristics of MBFV xrRNAs

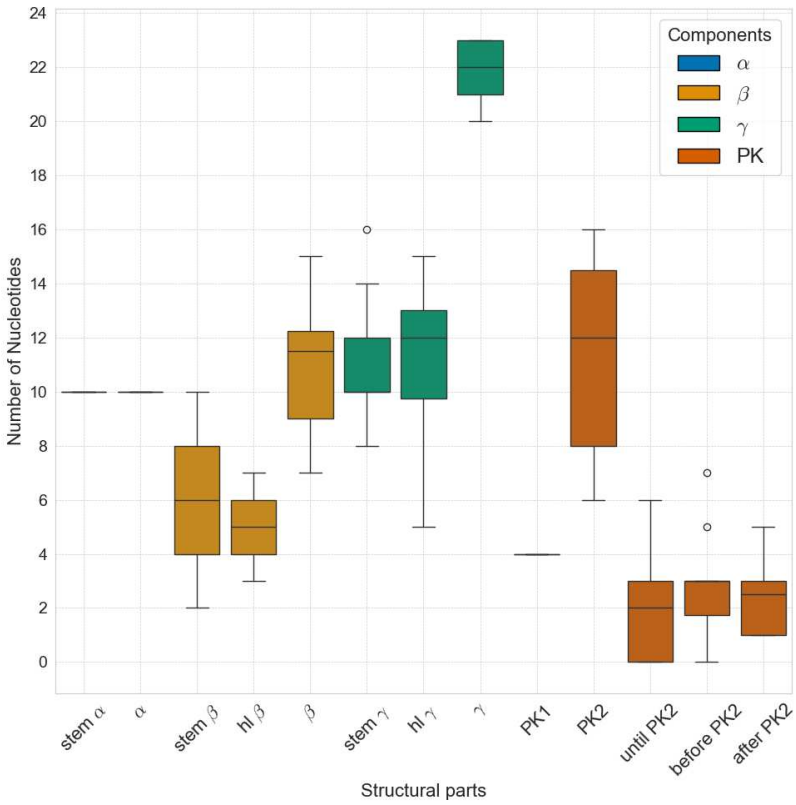
To obtain a representative structure for the MBFV group, outliers were excluded from the analysis. For example, the structure of the *Yellow fever virus* was removed due to its distinct consensus structure, which includes a significantly larger β hairpin loop compared to other members of MBFV. The lengths of each structural component were determined for the remaining sequences. In addition, the total lengths of the structural components α, β, and γ were calculated by summing all corresponding subelements,

including stems, hairpin loops, and bulges. A box plot summarizing these lengths can be seen in Figure 4.1(b).

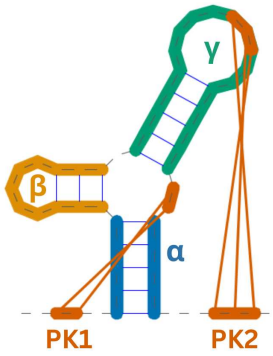
The stem α is fully conserved at a length of 10 nucleotides in all sequences. In contrast, the β and γ regions exhibit more variation. The length of the β stem ranges from 2 to 10 nucleotides, with the hairpin loop spanning 3 to 7 nucleotides, resulting in a median total length of the β region of 11.5 nucleotides. In comparison, the γ region is larger, ranging from 20 to 23, with a median length of 22 nucleotides.

Additional details on pseudoknots and unpaired nucleotides are presented in Figure 4.1(a). Pseudoknot 1 is fully conserved, consistently forming two base pairs, while pseudoknot 2 varies in size, ranging from 3 to 8 base pairs. On average, three nucleotides remain unpaired within the γ loop preceding PK2, with another three unpaired nucleotides typically following PK2. These observations provide valuable information on the positioning of PK2 within the γ hairpin loop.

The total length of MBFV xrRNA structures ranges from 50 to 62 nucleotides, with an average of 57 nucleotides.



(a) Distribution of structural region lengths in MBFV xrRNAs.



(b) Schematic scaffold of MBFV-based xrRNA structures.

Figure 4.1.: **Structural length evaluation of MBFV-xrRNAs.** (a) Boxplot showing the distribution of lengths for different structural regions in MBFV xrRNAs. (b) A schematic scaffold illustrating conserved structural features in the MBFV subgroup.

The Spearman correlation coefficients between the lengths of different structural regions in MBFV are shown in Figure 4.2. Correlations involving the structure α and PK1 were

excluded, as there is no variation in their length. There is a strong positive correlation between the length of the stem β and the total length of the region β, indicating that the length of the stem significantly contributes to the overall size of β. A negative correlation is observed between the lengths of the stem and the hairpin loop of the structure β, suggesting a compensatory relationship that maintains the length of β. Similar trends are observed in the γ region, where a strong negative correlation can be seen between the γ stem and its hairpin loop. Positive correlations are also observed between PK2 and the lengths of the stem and the total structure of β.

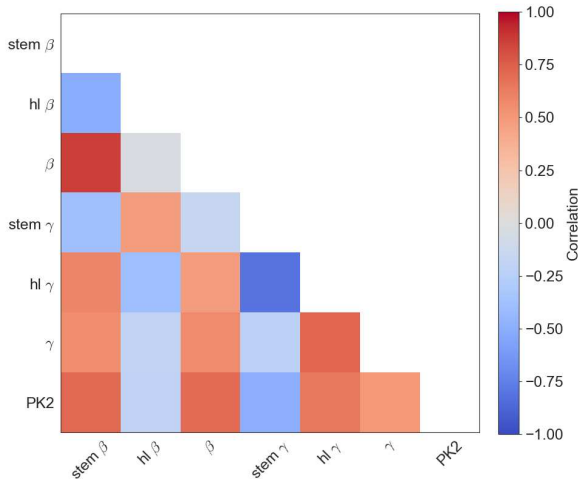


Figure 4.2.: **Correlation analysis of MBFV xrRNAs.** Spearman correlation coefficients between the lengths of different structural regions in MBFV xrRNAs.

4.1.2. Structural Characteristics of TBFV xrRNAs

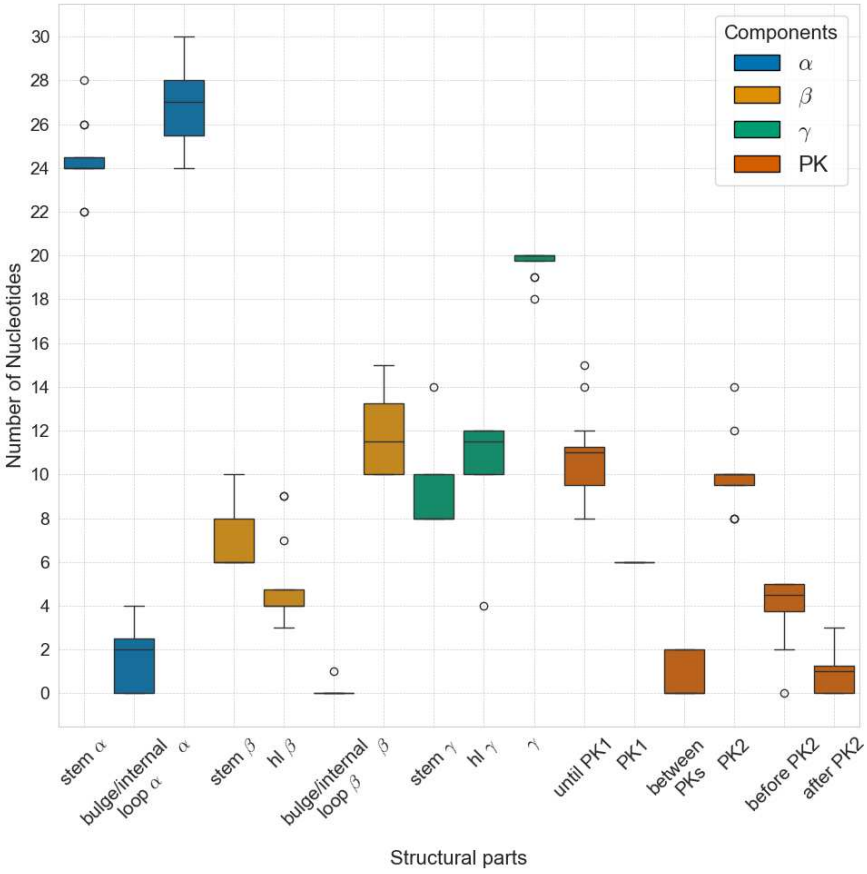
Certain members of the TBFV group were excluded from the analysis due to their distinct structural characteristics. For example, *Tyuleniy virus* and *Saumarez reef virus* lack a β structure, while *Xinyang flavivirus* and *Meaban virus* exhibit unusual secondary structures with shorter stems α, β and γ. Furthermore, PK1 data was unavailable for *Meaban virus* and *Xinyang flavivirus*.

The remaining sequences were analysed and Figure 4.3(a) summarizes the lengths of the structural regions with a schematic representation of the conserved secondary structure of the TBFV members in Figure 4.3(b). The stem α ranges from 22 to 28 and, including bulges or internal loops, the structure spans 24 to 30 nucleotides. In contrast, the β region

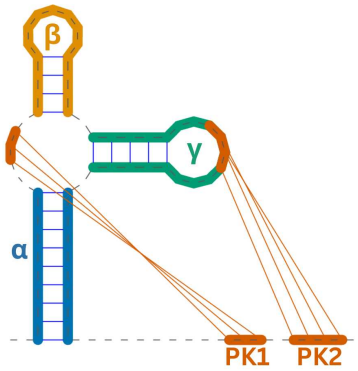
has a more compact length of 10 to 15 nucleotides. The γ region is nearly conserved, with a length of 18 to 20 nucleotides, consisting on average of 8 nucleotides in the stem and 11 in the hairpin loop.

The pseudoknot regions exhibit greater variability. The unpaired region between the stem α and PK1 ranges between 8 and 15 nucleotides. Between the two pseudoknots, there can stay between 0 and 2 nucleotides unpaired. PK2 consists of 4 to 7 base pairs. The number of unpaired nucleotides between PK2 in the loop γ ranges from 0 to 5 before PK2 and from 0 to 3 after PK2.

The total length of TBFV xrRNA structures ranges from 84 to 90 nucleotides, with an average length of 86 nucleotides.



(a) Distribution of structural region lengths in TBFV xrRNAs.



(b) Schematic scaffold of TBFV-based xrRNA structures.

Figure 4.3.: **Structural length evaluation of TBFV-xrRNAs.** (a) Boxplot showing the distribution of lengths for different structural regions in TBFV xrRNAs. (b) A schematic scaffold illustrating conserved structural features in the TBFV subgroup.

Figure 4.4 illustrates the correlations between the lengths of the structural regions in the TBFV xrRNAs. A negative correlation is observed between the lengths of α and β, while α is positively correlated with γ. This suggests a trade-off between the sizes of α and β and a concurrent increase in α and γ length. Similarly, within the γ region, a longer stem is associated with a shorter hairpin loop.

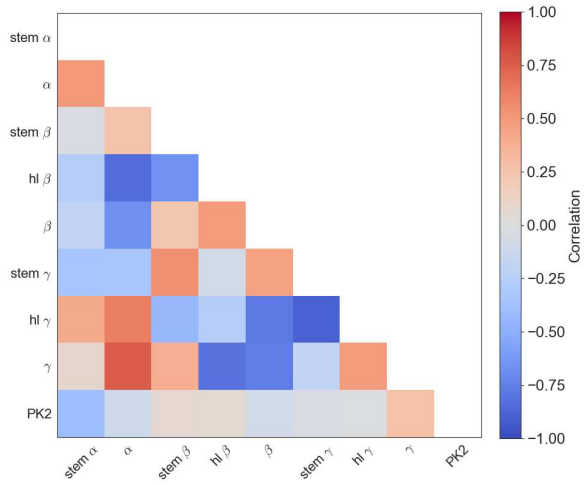


Figure 4.4.: **Correlation analysis of TBFV xrRNAs.** Spearman correlation coefficients between the lengths of different structural regions in TBFV xrRNAs.

4.2. RNA Design Model

4.2.1. Design Constraints for the MBFV-Based Model

The structural constraints of the MBFV model based on the analysis described before is defined as follows: The α structure is restricted to a stem length of exactly 5 base pairs. The β structure has a total length of 7 to 15 nucleotides, its stem consisting of 1 to 5 base pairs and its hairpin loop of 3 to 7 nucleotides. The structure γ is constraint to be within a length of 20 to 23 nucleotides, with a stem of 4 to 8 base pairs and a hairpin loop spanning 5 to 15 nucleotides.

Pseudoknot 1 consists of two base pairs. It spans from upstream of the alpha stem, leaving one unpaired nucleotide between the alpha stem and PK1, to two nucleotides within the multi loop located between the gamma structure and the alpha stem. Pseudoknot 2 consists of 3 to 8 base pairs. The region between the stem α and the downstream base

pairs of pseudoknot 2 can include 0 to 6 unpaired nucleotides. Within the γ hairpin loop, the positions of the upstream base pairs of pseudoknot 2 are also constrained. Specifically, the model permits 0 to 7 unpaired nucleotides before pseudoknot 2 and 1 to 5 unpaired nucleotides after it. These constraints ensure that pseudoknot 2 is appropriately positioned within the γ hairpin loop, while allowing for some structural flexibility.

From the MBFV correlation analysis, only the significant correlation between the lengths of the γ hairpin loop and its stem was incorporated into the model. This was achieved by constraining the relationship between the two structures to fall within the interquartile range, as determined from the biological data. The other structural constraints built into the model should account for the relationship between the total length of β and its stem, as well as the total length of γ and its hairpin loop. In contrast, other observed correlations were excluded because they fell below a defined threshold and were likely incidental. Such correlations were believed to not provide additional predictive value to the model. The structural and sequence constraints combined ensure the model's robustness and functionality in achieving the desired xrRNA structure.

Based on the work of Skibinski [17], the conserved sequence of MBFV xrRNA was incorporated into the design process. By integrating structural and sequence constraints, the fundamental parameters of the model were established, as depicted in Figure 4.5. The RNA sequence in the figure is represented in IUPAC code, which defines the sequence constraints that serve as the scaffold for the model. These constraints enforce specific nucleotide assignments in conserved regions, such as the stem α, while also imposing length restrictions on structural components by using the tokens `N` and `X`. Additional length restrictions for the β and γ regions were applied to achieve the desired structural specificity. To ensure a sufficiently long 5' end, two extra nucleotides were added at the 5' end. The base pairing shown in the figure represents the structural constraints that describe the target secondary structure. This representation ensures that the designed sequences can form the required base pairs, including those within the stems and the two pseudoknots.

Most of the designs were created without specifying a target length, leading to the calculation of weights for the length function. In this case, the target mean was first set to 56 nucleotides with a tolerance of 7, resulting in sequences ranging from 49 to 62 nucleotides (excluding the two additional nucleotides at the 5' end). This range was intended to reflect the true sequence length distribution, which spans 50 to 62 nucleotides, with a mean of 57. However, the model tended to generate longer sequences. As a result, the set length for the weight calculation was adjusted to 54 nucleotides with a tolerance of 7, yielding sequences ranging from 48 to 62 nucleotides, with a median length around

55-56 nucleotides.

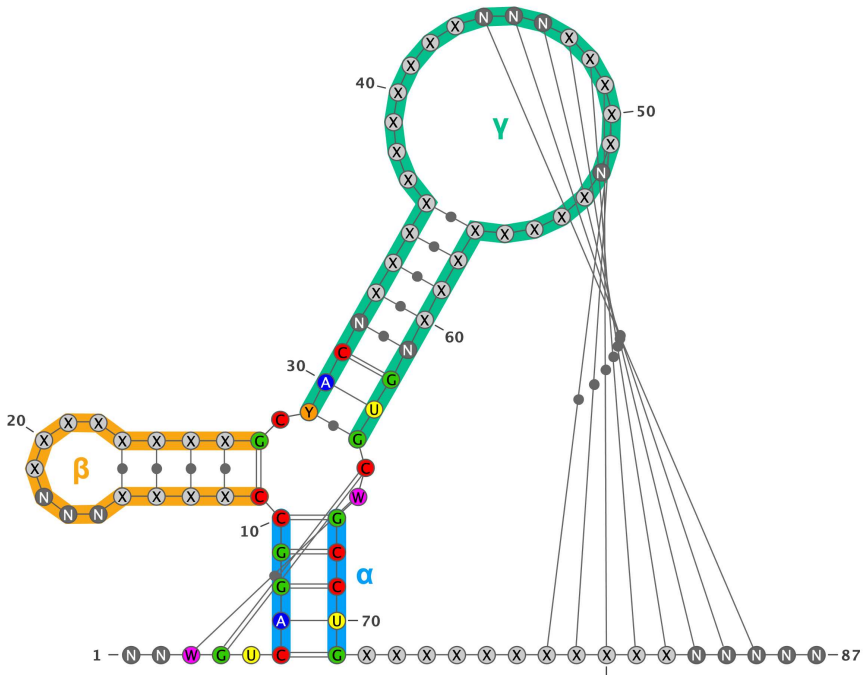


Figure 4.5.: **Design scaffold of the MBFV-based model.** The constraints guiding the RNA design process for the Infrared model based on the subgroup MBFV. Conserved sequence motifs and IUPAC codes define the sequence constraints, while structural constraints ensure accurate base-pair formation to achieve the xrRNA target structure.

4.2.2. Design Constraints for the TBFV-Based Model

The structures in the TBFV model are constrained as follows: The stem α consists of 11 to 14 base pairs, with up to four nucleotides allowed to remain unpaired within the stem. The total length of the α structure ranges from 24 to 30 nucleotides. The β structure has a stem comprising 3 to 5 base pairs and a hairpin loop containing 3 to 9 nucleotides. The total length of the β structure ranges from 10 to 15 nucleotides. The structure γ has a total length of 18 to 20 nucleotides, with the stem γ consisting of 4 to 7 base pairs. Between the stem and pseudoknot 2, there are 4 to 12 unpaired nucleotides. One segment of pseudoknot 2 is located within the γ hairpin loop, consisting of 4 to 7 base pairs. Depending on its position, there may be 0 to 3 unpaired nucleotides preceding PK2 and 0 to 2 unpaired nucleotides following it, allowing PK2 to be flexibly positioned within the γ hairpin loop. Following the stem α, there are 8 to 15 unpaired nucleotides before pseudoknot 1 starts, which is composed of three base pairs. Between PK1 and PK2, there can be up to two unpaired nucleotides.

The observed correlations in the lengths of the TBFV xrRNA structures led to the incorporation of the following correlation into the model:

- Correlation between α and β.
- Correlation between α and γ.
- Correlation between β and γ.

In addition, the relationship between the length of the stem and the hairpin of γ was included in the model. This was achieved by constraining the relationship to fall within the interquartile range, calculated from biological data.

The conserved nucleotides of the TBFV xrRNA sequence, derived from Skibinksi [17], were incorporated into the Infrared design. The combination of structural and sequence constraints resulted in the design scaffold shown in Figure 4.6. The sequence depicted includes conserved motifs along with tokens `N` and `X` for variable regions, representing the sequence constraints in IUPAC code. The base pairs in the figure illustrate the structural constraints, ensuring compatibility with the target structure. Sequential constraints were used to enforce the inclusion of specific motifs and impose length restrictions on structural components. Additional length constraints within the Infrared model were applied to the β and γ regions to ensure structural specificity. The total sequence length was modeled as normally distributed around the mean length of the TBFV sequences, approximately 86 nucleotides. Together, structural and sequence constraints ensure that the generated sequences meet both functional and structural requirements, including the proper formation of stems and pseudoknots.

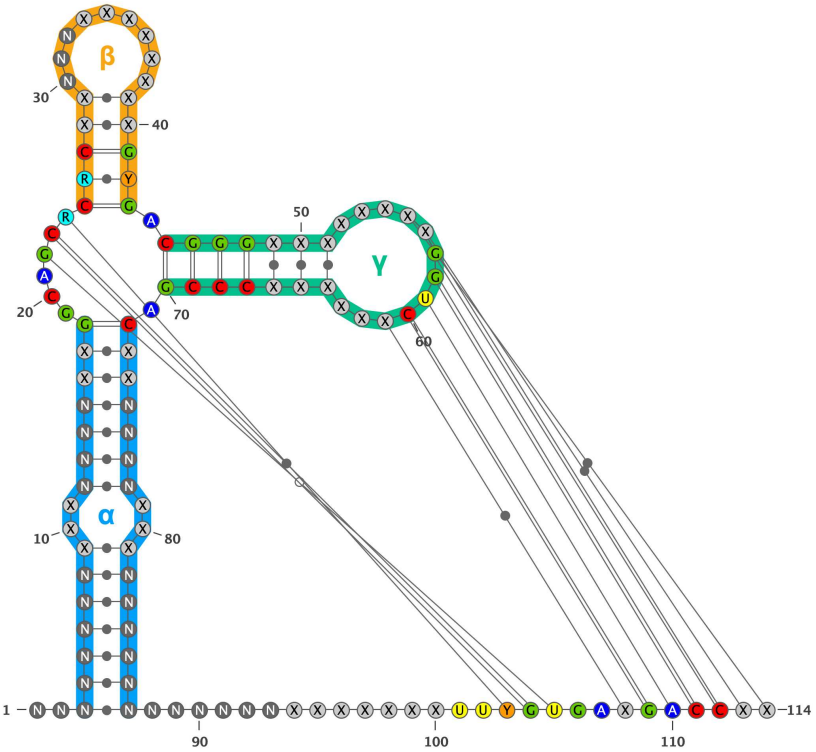


Figure 4.6.: **Design scaffold of the TBFV-based model.** The constraints guiding the RNA design process for the Infrared model based on the subgroup TBFV. Conserved sequence motifs and IUPAC codes define the sequence constraints, while structural constraints ensure accurate base-pair formation to achieve the xrRNA target structure.

4.2.3. Objective Function for Sequence Optimization

The objective function was iteratively refined to optimize the generated sequences while maintaining simplicity. Initially, the objective function included only the equilibrium probability. However, as additional constraints were incorporated, further improvements in EP became increasingly challenging during optimization. To address this, the final strategy employed a two-stage approach: first, minimizing the ensemble defect, followed by maximizing the equilibrium probability.

Maximizing EP is particularly challenging for longer sequences, as small sequence changes often do not immediately improve EP, leaving it at zero. If no improvement is observed, the sequence remains unchanged, leading to the risk of becoming trapped. In contrast, the ensemble defect responds more sensitively to sequence modifications, providing immediate feedback on improvements or deteriorations. This sensitivity allows for rapid and consistent adjustments to the sequence. Therefore, the adopted strategy leverages the strengths of both approaches. In the first stage, ED is minimized to achieve rapid initial

progress. In the second stage, the process aims to maximize the EP, refining the sequence to meet the desired structural requirements more precisely.

4.3. Tertiary Structure Evaluation

Sequences were generated and optimized using the Infrared model. Sequences exhibiting promising properties, such as high equilibrium probability and low ensemble defect, were selected for further analysis through 3D simulations. These simulations were primarily evaluated using structural visualization in PyMOL. Due to the distinct and well-characterized configuration of xrRNA structures, the evaluation of simulation results was relatively straightforward. It was possible to visually determine whether a simulation successfully achieved the desired target conformation or deviated significantly from it. For designs that did not fold into the target conformation, a more detailed analysis was conducted. This analysis primarily involved identifying incorrectly paired nucleotides as well as base pairs that were expected to form but did not. These evaluations provided valuable information on potential challenges in the design process, highlighting specific features or constraints that may require adjustment to improve the model.

4.3.1. Predicted Tertiary Structures of MBFV-Based Designs

The initial 3D-simulated designs generated using the MBFV-based model demonstrated promising results. Visualization of the SimRNA simulations revealed the characteristic ring-like structure, an indicator of the target conformation. Alignment of the simulated structures with the available crystal structure showed a high degree of similarity, suggesting that the model is capable of producing sequences that reliably fold into the desired target structure. This alignment is illustrated in Figure 4.7, where the reference structure is depicted in blue, and the 3D simulation of the design is shown in red. To further evaluate the flexibility and reliability of the model, variations in specific structural characteristics in the design model, such as the length of the stem β, were introduced and tested. Notably, designs with a significantly shortened β stem, consisting of only a single base pair, were still able to fold into the desired tertiary structure. These outcomes highlight the flexibility of the design process and the ability of the model to produce sequences that fold as intended.

After multiple sequences were successfully generated that folded into the reference structure, further experiments were conducted to explore the importance of specific structural features for the xrRNA fold. Given that variations in β stem length did not significantly affect the tertiary structure, and considering that some xrRNA structures of FV do not have the β stem altogether, the first experiment involved completely removing the β stem. Two designs were generated and simulated using SimRNA. Visualization in PyMOL revealed that both simulations retained the ring-like structure and were highly similar

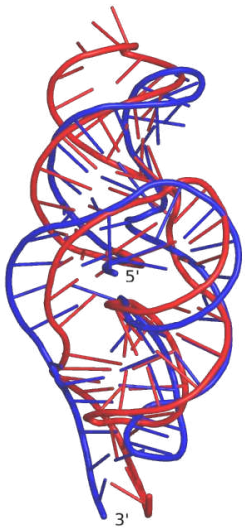


Figure 4.7.: **3D Alignment of Wild-Type and Synthetic Design.** The figure shows the alignment of the wild-type reference structure (ZIKV) in blue with the 3D simulation of the artificial design in red, illustrating their structural similarity.

to each other. Alignment with the ZIKV crystal structure again demonstrated strong similarity, suggesting that the stem β may not be as critical for proper folding as other regions. One of these two designs was selected for further MD analysis to investigate the role of the stem β in greater detail, as discussed in Section 4.4.

Another variation tested the role of base triples by using them as sequence constraints while removing all other constraints. Two sequences were created and simulated. Of these, one design formed a well-defined ring-like structure, while the other adopted a conformation that showed a high similarity to the reference structure but without a clearly defined ring. The design that successfully formed a ring was further analysed to gain insight into the role of the base triples, as discussed in Section 4.4.

Further findings suggested the importance of a conserved cytosine residue located within the multi loop between the β and γ stems. To test this, a single nucleotide variation was introduced, which replaced cytosine with another nucleotide. Two experiments were conducted: one with all remaining sequential constraints removed and another with those constraints preserved. The resulting designs retained a well-defined ring-like structure upon visualization and exhibited high similarity to reference structures, ZIKV. These designs were selected for the subsequent MD analysis, as discussed in Section 4.4.

Furthermore, a set of 10 designs was generated using the standard RNA design model to evaluate the model's ability to produce functional xrRNA sequences. These 10 designs

were 3D-simulated and analysed. Each design was compared to the ZIKV reference structure using metrics such as RMSD, GDT, TM score, and CAD score, with the results summarized in Table 4.1.

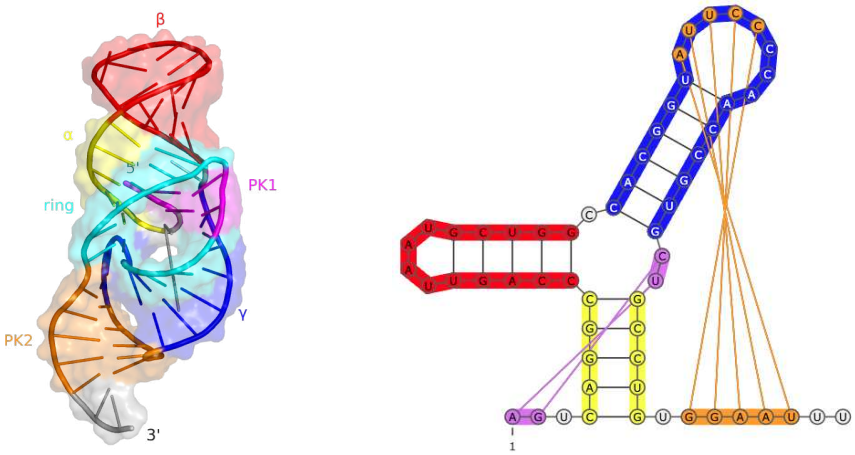
Table 4.1.: MBFV designs metrics table

design	RMSD	GDT	TM	CAD
1	12.13	0.36	0.30	0.48
2	8.03	0.32	0.34	0.47
3	6.95	0.25	0.33	0.49
4	7.76	0.38	0.28	0.48
5	7.34	0.37	0.32	0.46
6	8.64	0.28	0.27	0.45
7	4.55	0.30	0.34	0.48
8	6.46	0.24	0.31	0.49
9	8.93	0.30	0.24	0.39
10	8.22	0.25	0.32	0.52

RMSD analysis showed scores ranging from approximately 4.5 to 12 Å, indicating varying levels of performance between designs. Design 1 exhibited the highest RMSD, while Design 7 had the lowest, suggesting a closer structural alignment to the reference structure. In contrast, the GDT and TM scores provided a different perspective. GDT scores highlighted Design 1, Design 4, and Design 5 as the top-performing designs. Notably, Design 7 and Design 2 achieved the highest TM-scores, followed by Design 3, although the TM-scores across all designs remained within a similar range. CAD scores showed low variability, with Design 10 attaining the highest value. In general, the results from CAD, TM, and GDT indicate nearly equal structural similarity between the designs and the reference structure. Although the CAD scores show intermediate alignments, the low TM and GDT scores suggest a restricted degree of similarity between the designs and the reference.

Despite this, visualization revealed that most designs folded into a similar structure, with the ring-like feature observed in nearly all simulations, except for two designs. These two designs (Design 4 and Design 6) exhibit structures resembling the reference; however, the ring-like feature fails to assemble correctly, instead forming a convex-like opening. This leads to issues where the 5' end of the RNA fails to pass through the ring-like structure. Interestingly, these two designs do not show poor scores in the metrics compared to the other designs, highlighting the importance of visualization in the evaluation of designs, as metrics alone may not capture structural anomalies. An example of a poor design

(Design 4) is shown in Figure 4.8, where panel (a) demonstrates that the 5' end does not pass through the ring-like structure (highlighted in cyan). Instead, the structure formed in this design resembles an ellipse rather than a proper ring. In contrast, Figure 4.9(a) illustrates a well-structured ring-like conformation, with the 5' end passing through the ring. Four designs following this structural pattern, with a well-defined ring-like structure, were selected for further MD analysis, as discussed in Section 4.4.



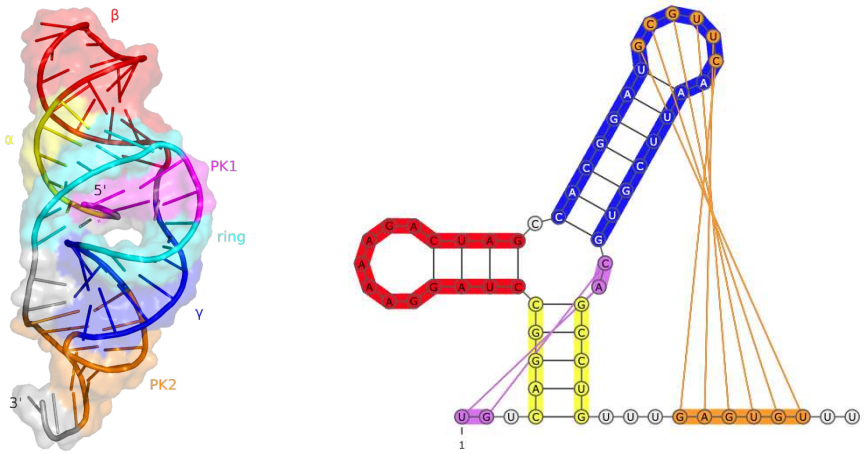
(a) 3D simulation.

(b) Predicted structure.

```
AGUCAGGCCCAGUUAAUGCUGGCCACGGUAUUCCCCAACCGUGCUGCCUGUGGAAUUU
<<.(((((((((((....)))))).((((((([[[[[...))))))>>)))))..]]]]]..
1.......10........20........30........40........50......58
```

(c) RNA sequence and secondary structure

Figure 4.8.: **Visualization of MBFV Design 4.** (a) The 3D simulation of the RNA structure. (b) The predicted structure of the RNA. (c) The RNA sequence alongside its corresponding dot-bracket notation for the structure.



(a) 3D simulation.

(b) Predicted structure.

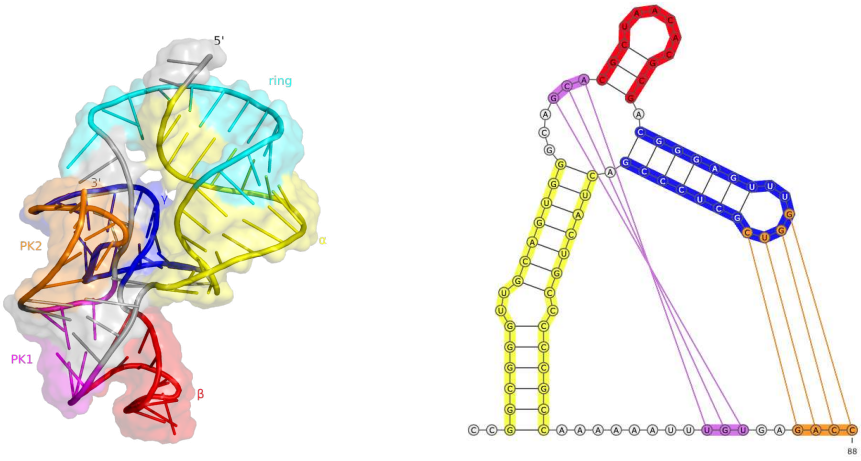
(c) RNA sequence and secondary structure

Figure 4.9.: **Visualization of MBFV Design 2.** (a) The 3D simulation of the RNA structure. (b) The predicted structure of the RNA. (c) The RNA sequence alongside its corresponding dot-bracket notation for the structure.

4.3.2. Predicted Tertiary Structures of TBFV-Based Designs

The design process for TBFV presented more challenges compared to MBFV. The initial designs exhibited substantial deviations from the reference structure, with the ring-like configuration often missing or poorly defined. Despite adjustments to the system and insight into the design process achieved, significant differences between designs and the reference structure persisted.

To better understand the design process and identify its limitations, 10 designs were generated and subsequently simulated in 3D using SimRNA. Although most of the designs formed a ring-shaped structure, many appeared to be loosely organized. This is also evident in Figure 4.10, which shows a design with a ring-like structure; however, the remaining elements appear to be poorly organized. Among these, two designs (Design 8 and Design 9) seem the most promising, as they showed the highest similarity to the reference structure.



(a) 3D simulation.

(b) Predicted structure.

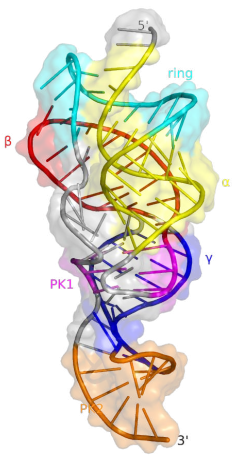
(c) RNA sequence and secondary structure

Figure 4.10.: **Visualization of TBFV Design 1.** (a) The 3D simulation of the RNA structure. (b) The predicted structure of the RNA. (c) The RNA sequence alongside its corresponding dot-bracket notation for the structure.

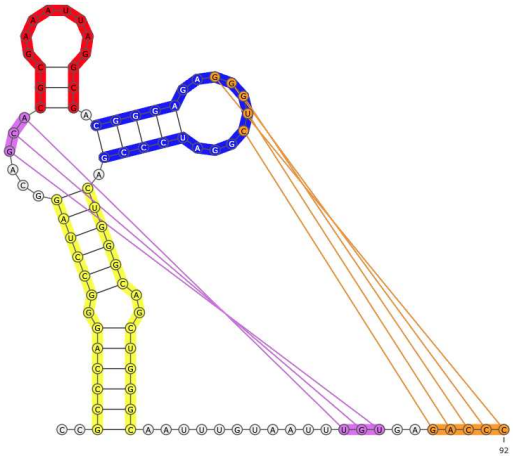
The results of various structural metrics, including RMSD, GDT, and TM scores, are summarized in Table 4.2. Design 9 emerged as the best performing structure, with the lowest RMSD and the highest scores for GDT, TM, and CAD. A visualization of this sequence can be seen in Figure 4.11 Consequently, Design 9 was selected for further MD analysis. In addition, Design 8 was manually adjusted to correct for a few mispaired nucleotides within the ring-like structure and was subsequently used for further MD analysis.

Table 4.2.: TBFV designs metrics table

design	RMSD	GDT	TM	CAD
1	15.76	0.20	0.24	0.50
2	18.24	0.13	0.26	0.51
3	15.27	0.14	0.23	0.46
4	22.24	0.14	0.22	0.49
5	18.05	0.15	0.22	0.52
6	19.40	0.12	0.20	0.52
7	16.40	0.17	0.25	0.46
8	16.20	0.16	0.25	0.50
9	11.68	0.27	0.37	0.54
10	14.98	0.14	0.25	0.47



(a) 3D simulation.



(b) Predicted structure.

(c) RNA sequence and secondary structure

Figure 4.11.: **Visualization of TBFV Design 9.** (a) The 3D simulation of the RNA structure. (b) The predicted structure of the RNA. (c) The RNA sequence alongside its corresponding dot-bracket notation for the structure.

4.4. Structural Stability Evaluation

To evaluate the stability of the designed sequences and their potential resistance to Xrn1, MD simulations were conducted. The initial 3D model generated by SimRNA was used as input for these simulations, following the methodology outlined in Section 3.6. Two

distinct approaches were used: pulling with a range of constant forces (CFMD) and pulling with a gradually increasing force (GFMD). The MD simulation results were visualized using PyMOL, enabling detailed tracking of the unfolding process and identification of structural deviations over time.

As an additional analysis, histograms from the CFMD simulations were generated. These histograms depict the distance of the 5' end throughout the pulling process, providing insight into structural stability. A structure is assumed to remain stable if the 5' end maintains a distance under 10 Å relative to its original position. These histograms illustrate the frequency with which the structure retains similarity to the original configuration. Frames with larger distances indicate potential unfolding events or structural rearrangements. Furthermore, monitoring the distances between specific base pairs provides information about the persistence of base pairing during simulations. By combining these histograms with base-pair distance data and visual analysis, valuable insights into the RNA stability were obtained.

GFMD simulations revealed the force threshold at which the structure begins to unfold. By monitoring the position of the 5' end throughout the simulation, significant increases in distance were identified as markers of refolding or unfolding events. This approach enabled the determination of the maximum force at which the design retained its initial configuration. Comparisons with a reference structure provided insight into the likelihood that the design exhibits Xrn1 resistance. These analyses offer a comprehensive understanding of the stability of the designed RNA sequences under varying force conditions.

4.4.1. MBFV reference structure

Figure 4.12(a) presents the distance histograms from the CFMD simulations of the reference ZIKV structure, based on the crystal structure with PDB ID `5TPY`. To improve visibility across all pulling forces, the histograms are divided into two separate plots. These results offer insights into the behavior of an xrRNA-resistant structure in MD simulations.

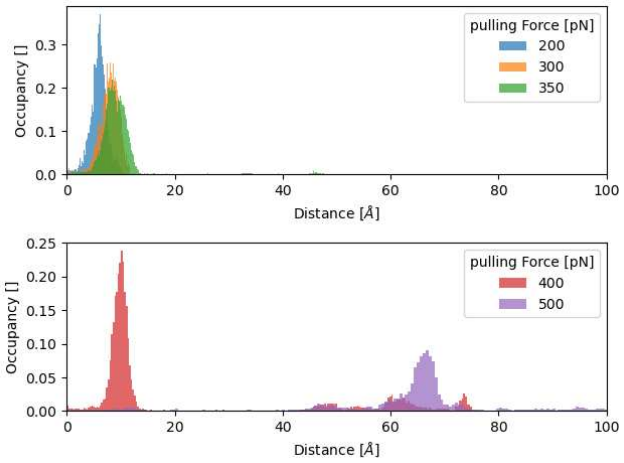
The top plot in Figure 4.12(a) presents histograms for pulling forces of 200, 300, and 350 pN. All three histograms show a peak at or below 10 Å, suggesting that the structure remains predominantly folded throughout the simulations. As the pulling force increases, the peak shifts slightly to the right. However, the histograms indicate that the structure largely retains its initial configuration during the MD simulation for pulling forces between 200 and 350 pN.

The bottom plot shows histograms for pulling forces of 400 and 500 pN. For the 400 pN

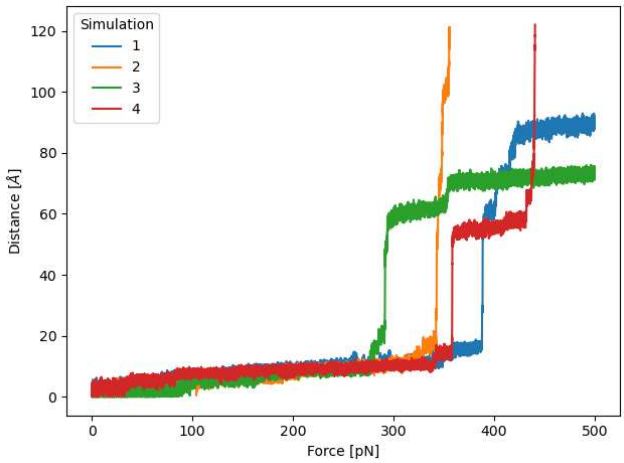
simulation, the primary peak remains near 10 Å, indicating that the structure is mostly stable. Secondary local maxima are observed around 60 and 80 Å, suggesting that at certain time points the structure changes into alternative configurations. Despite these transitions, the majority of frames show the 5' end remaining close to its original position. In contrast, the histogram for the 500 pN pulling force reveals a significant shift to the right, with the peak centered around 65–70 Å. Most frames in this simulation exhibit distances within the range of 60–75 Å, indicating that the structure is unfolded or refolded in nearly all frames. These results suggest that the structure remains intact up to a pulling force of 400 pN, beyond which it no longer retains its initial configuration.

The results of the GFMD simulations show a similar trend. Figure 4.12(b) illustrates the distance between the initial position of the 5' end and its current position during simulations using a gradient force. In all simulations, the 5' end remains close to its original position up to a pulling force of 300 pN, indicating that the RNA sequence retains its initial configuration up to this point. Around 300 pN one simulations shows an increase in distance. Beyond 350 pN, two other simulations exhibit an increase in distance, while one simulation shows an increase just before 400 pN. These increases in distance correspond to refolding or unfolding events. Three simulations plateau around 70 Å, indicating an alternative stable configuration adopted by the RNA sequence.

In summary, the GFMD results demonstrate that the RNA structure remains stable under pulling forces up to 300 pN, similiar to the distance histogram results from the CFMD simulations. At a pulling force of 350 pN, the structure begins to transition, and forces beyond this threshold result in significant structural changes, suggesting re- or unfolding events.
The results of the GFMD simulations show a similar trend. Figure 4.12(b) illustrates the distance between the initial position of the 5' end and its current position during simulations using a gradient force. Across all simulations, the 5' end remains close to its original position up to a pulling force of 300 pN, indicating that the RNA sequence retains its initial configuration within this range.



(a) Histogram from CFMD simulations.



(b) Distance from GFMD simulations.

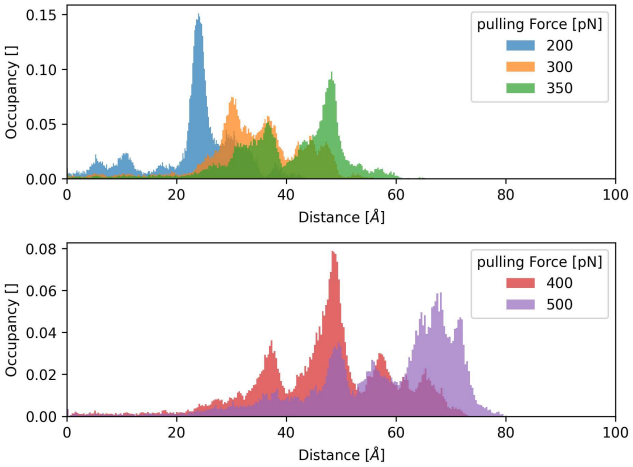
Figure 4.12.: **MD Analysis of MBFV reference structure.** (a) displays two histograms of distances calculated from CFMD simulations under different pulling forces of ZIKV. (b) shows the distance between the 5' end initial position and its current location during GFMD simulations of ZIKV.

4.4.2. MBFV design lacking a β structure

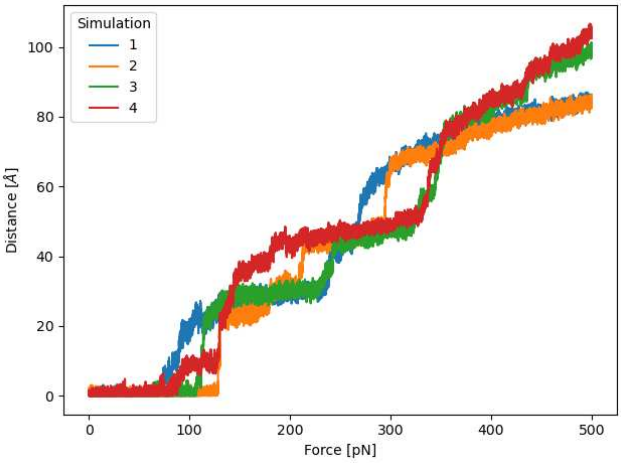
First, we examined sequences modified to meet specific constraints, testing the significance of certain sequence features. Figure 4.13 presents the results of the MD simulations for a sequence designed without the stem β. The tertiary structure of this design closely resembles the ZIKV reference structure, showing a well-defined ring-like structure when simulated using SimRNA. However, analysis of the MD simulations reveals that the structure is relatively unstable. The histograms of the CFMD simulations are shown in

Figure 4.13(a). Under a pulling force of 200 pN, a peak is observed around 25 Å, while for 300 pN, the peak shifts to approximately 30 Å. Pulling with a higher force shifts the peak further to the right, indicating greater structural displacement with greater force. These results suggest that the structure begins to unfold under a force as low as 200 pN. Furthermore, the results of the GFMD simulation in Figure 4.13(b) show a rapid increase in distance. The 5' end can hardly contain its original positional, and even when pulling with a force of 75 pN an increase can be seen, indicating an unstable structure. Compared with the reference structure, this design, which lacks the stem β, exhibits significantly reduced structural stability.

Although the stem β appears to play a critical role in maintaining the stability of the structure, there are biological structures that naturally lack this feature. It is possible that such structures compensate for the reduced stability through other mechanisms, which remain to be explored. Further experiments would be necessary to uncover how these structures adapt to ensure stability in the absence of the stem β.



(a) Histogram from CFMD simulations.



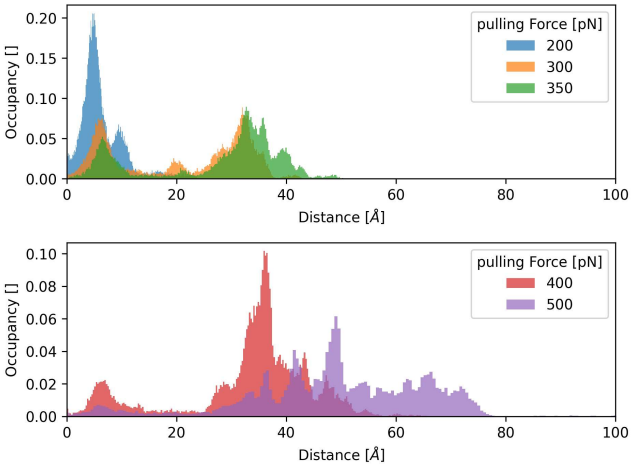
(b) Distance from GFMD simulations.

Figure 4.13.: **MD simulation analysis of a MBFV design lacking β stem.** The two figures illustrate the distance measurements obtained from MD simulations of an MBFV design lacking a β stem. Figure (a) presents two histograms of distances from CFMD simulations conducted under different pulling forces. Figure (b) shows the distance between the 5' end initial position and its current location during GFMD simulations.

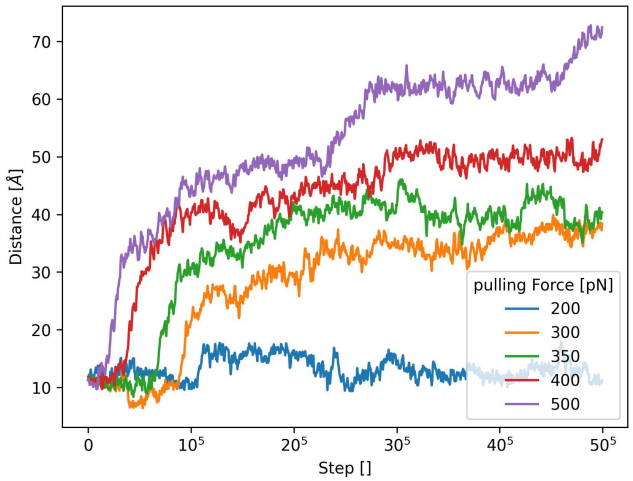
4.4.3. Importance of base triples in MBFV designs

Figure 4.14 shows the histograms of the simulations in which only the base triples were conserved within the sequence. No additional sequence constraints were applied beyond the conservation of base triples. At the lowest pulling force, the structure remains largely intact. However, as the force increases to 300 pN, a large fraction of the frames begins to exhibit RNA unfolding. This fraction grows further with increasing forces, and by 500 pN,

nearly all frames indicate a fully unfolded RNA structure. Although some resistance to unfolding is observed, it is quickly overcome with small increases in pulling force. A similar trend is seen in the distance between a base pair of PK1. At low forces, the base pair remains stable, but as the force increases, the nucleotide distance rapidly exceeds 10 Å, indicating bond rupture. These results demonstrate that conserving only base triples does not provide sufficient sequential and structural information to achieve stability comparable to that obtained with xrRNA sequences.



(a) Distance histogram during CFMD simulations.



(b) PK1 distances during CFMD simulations.

Figure 4.14.: **MD simulation analysis of a MBFV design with only base triple constraints.** The two figures illustrate the distance measurements obtained from MD simulations of an MBFV design, where only the base triple was sequentially constrained. Figure (a) presents two histograms of distances from MD simulations conducted under different pulling forces. Figure (b) shows the mean distance between a PK1 base pair during CFMD simulations conducted under different pulling forces.

4.4.4. Importance of conserved cystosine in MBFV designs

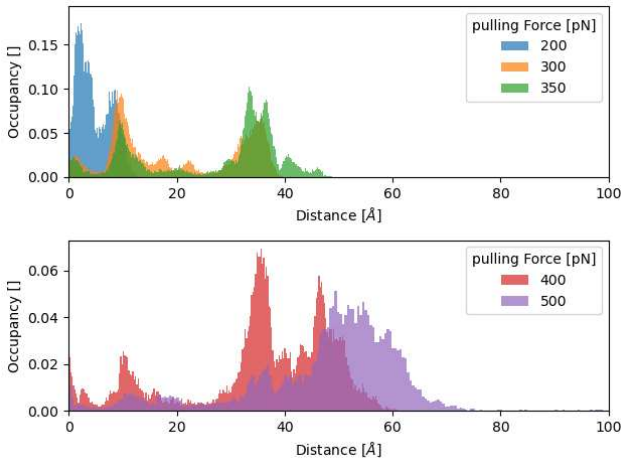
Figure 4.15 presents the CFMD simulations performed with sequences lacking the conserved cytosine within the multiloop. Panel (a) shows simulations for the sequence in which no additional sequential constraints were applied, which means that IUPAC information was not included except for the absence of the conserved cytosine. In contrast,

panel (b) shows simulations of the sequence without the conserved cytosine but with the remaining sequential constraints preserved.

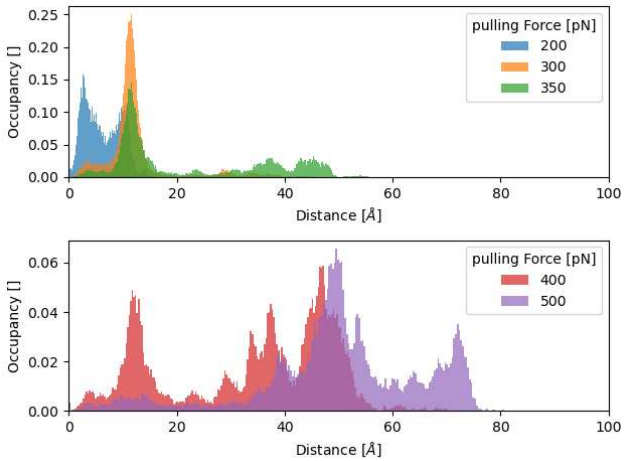
In panel (a), the sequence demonstrates low stability. Even with a pulling force of 300 pN, a second peak begins to emerge, indicating structural disruption. At higher forces, such as 400 pN, the structure can no longer maintain its initial configuration and rapidly unfolds. Panel (b), however, reveals a different trend. When pulled with forces up to 350 pN, the maxima remain between 0 and 15 Å, reflecting a structure that closely resembles its initial configuration. Beyond 400 pN, a shift to the right is observed in the histogram, suggesting a partial refolding of the structure.

When we compare these results, it becomes evident that removing sequential constraints leads to a more unstable molecule. This observation aligns with the findings for sequences in which only the base triple information was conserved, as shown in Figure 4.14. Both sequences – those lacking conserved cytosine and those with only base triple constraints – exhibit low stability and rapid unfolding. This instability could arise from the lack of sufficient sequential constraints, which allows the remaining sequence to be randomized.

In contrast, adding back the sequential constraints, as seen in panel (b), partially restores the molecule's stability. The resulting histograms indicate a certain degree of structural stability; however, it remains inferior to that of the reference structure, which exhibits a higher level of stability. This suggests that the conserved C within the sequence plays a critical role in contributing to the stability of the molecule. However, the stability observed in panel (b) is notably greater than that in panel (a). Sequences lacking comprehensive sequence constraints demonstrate pronounced structural instability, emphasizing that structural constraints alone are insufficient. Therefore, both structural and sequential constraints are essential to achieve stability and behavior comparable to the reference structure.



(a) Distance histogram of the MBFV design with additional sequence-constraints.



(b) Distance histogram of the MBFV design without additional sequence-constraints.

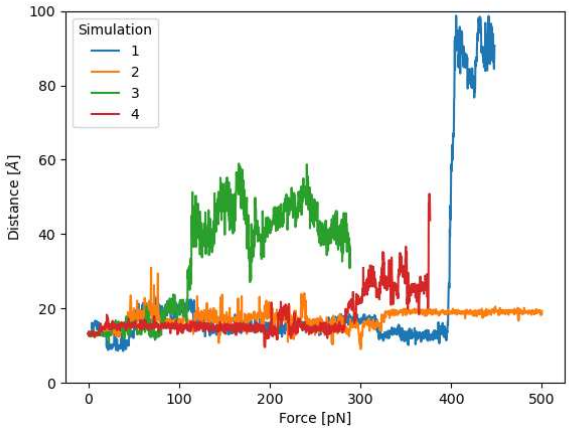
Figure 4.15.: **CFMD simulation analysis of MBFV designs without conserved cytosine.** Panel (a) shows the MD simulations for a sequence where no sequential constraints were applied, except for the absence of the conserved cytosine. Panel (b) displays the MD simulations for a sequence lacking the conserved cytosine but retaining all other sequential constraints.

4.4.5. Stable Pseudoknot 2 in MBFV designs

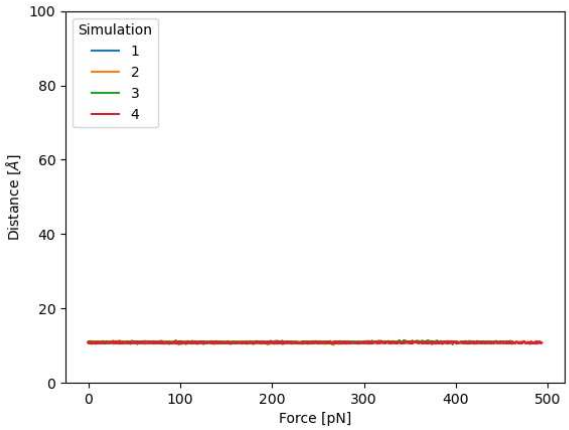
A key observation from these MD analyses is the importance of pseudoknot 2. Figure 4.16 shows the distances between the base pairs of PK2 during MD simulations in which the pulling force was gradually increased. The two plots show two different designs, in (a), an earlier design is shown in which certain constraints were not yet applied. We can see that the distance between base pairs increased significantly in three out of four

simulations. This trend is also seen in another MD simulation of a similar design. In contrast, Figure 4.16(b) shows simulations of a different design where the distance remains constant around 10 Å. These observations show that the PK2 bond breaks in some simulations but remains stable in others.

Analysis of the base-pairing energy of PK2 reveals distinct differences between stable and unstable configurations. In simulations where PK2 dissociates, high energy values (+4.3, +4.6) are observed, indicating instability. In contrast, simulations where the distance of PK2 remains low show negative energy values, suggesting a stable base pairing. These findings indicate that pseudoknot 2 must exhibit negative energy to ensure structural stability. To address this, the energy of pseudoknot 2 was incorporated as a constraint into the MBFV-based Infrared model. Incorporating this insight into the model led to the generation of designs in which pseudoknot 2 remains stable under simulation conditions.



(a) Distance between base pair in PK2 with positive energy.



(b) Distance between base pair in PK2 with negative energy.

Figure 4.16.: **PK2 Distance During MD Simulations.** The plots depict the distances between a base pair of PK2 during GFMD simulations.

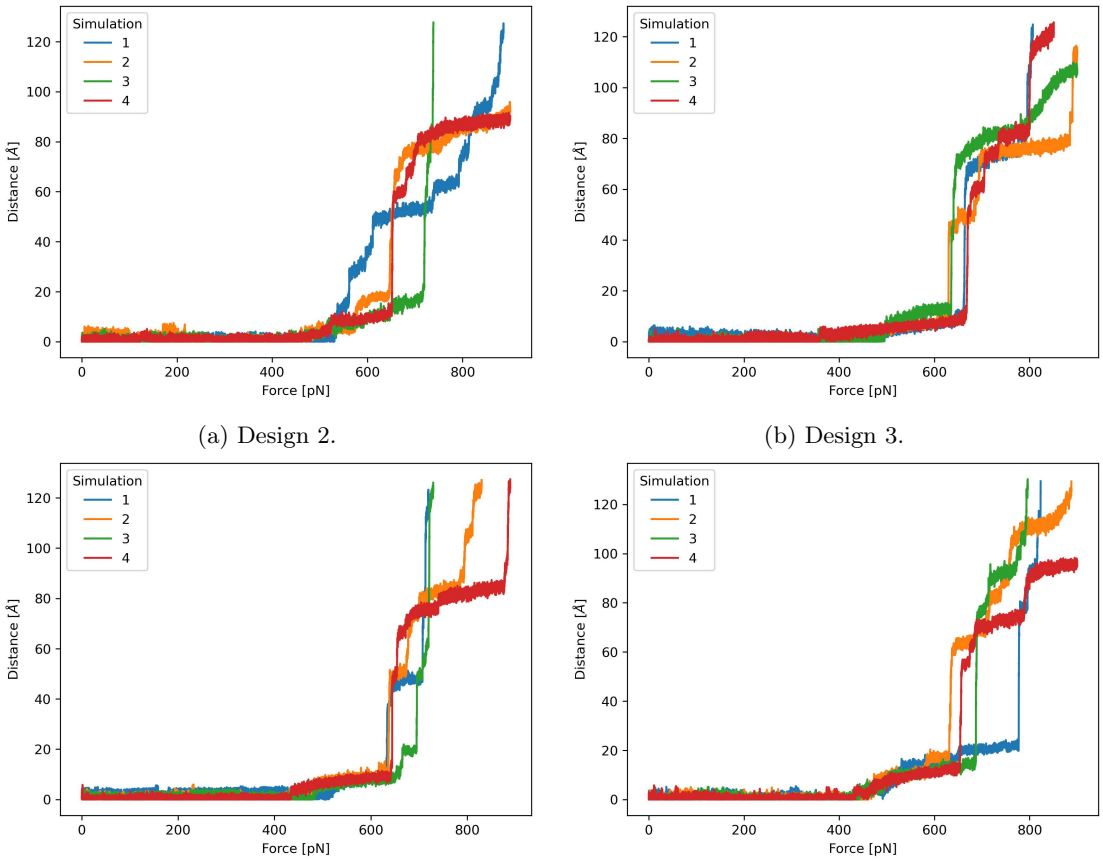
4.4.6. Consistency of the MBFV-Based Design Model

In addition to analyzing specific features of the xrRNA structure, a total of 10 sequences were designed to evaluate the reliability of the design process. After filtering the designs based on their SimRNA 3D simulation results, four designs (Design 2, 3, 7, and 10) were selected for further analysis using MD simulations.

The tracking of the 5' end distance during the GFMD simulations is shown in Figure 4.17. In all four plots, the distance remained low up to a force of 500 pN. In some simulations, such as Simulation 1 of Design 10, the distance remained low even at forces

of 700 pN or nearly 800 pN. However, between 500 and 800 pN, all designs exhibited a significant increase in distance, representing the refolding or unfolding of the RNA sequence.

Comparison of these results with the reference structure, shown in Figure 4.12(b), reveals that the selected designs exhibit structural stability equal to or greater than the reference, as determined by the GFMD simulations. Given that the reference structure is known to exhibit resistance to Xrn1, these results strongly suggest that the designed sequences may also demonstrate resistance to Xrn1.



(a) Design 2.

(b) Design 3.

(c) Design 7.

(d) Design 10.

Figure 4.17.: **GFMD Analysis of MBFV designs.** The plots show the distance between the initial and current positions of the 5' end during MD simulations of four selected MBFV designs.

4.4.7. TBFV Reference structure

No experimental structure was available to serve as a reference for the TBFV sequences. Therefore, as in the tertiary structure analysis, a reference structure was derived from a TBFV sequence of DTV. This structure was computationally modified and optimized to align with structural assumptions and subsequently used for GFMD simulations. The results of the simulations of this reference structure are shown in Figure 4.18. The data indicate that the structure exhibits limited stability. Early in the simulation, the distance between the current position of the 5' end and its initial position increases consistently. Upon applying forces in the range of 250–300 pN, a pronounced increase in distance is observed, indicating a structural transition. Some simulations temporarily stabilize at a distance around 80 Å, suggesting the presence of an alternative configuration. However, certain simulations move beyond this state with further increases in the pulling force. In particular, Simulation 2 demonstrates a rapid increase in distance after exceeding a pulling force of 300 pN without achieving a stable intermediate configuration, unlike the other simulations.

When comparing the stability of the DTV structure with the GFMD analysis of the ZIKV structure (Figure 4.12(b)), it is evident that the ZIKV structure exhibits greater stability, maintaining its conformation over more simulation steps than the DTV structure. This disparity could be attributed to the absence of an experimentally validated structure for DTV.

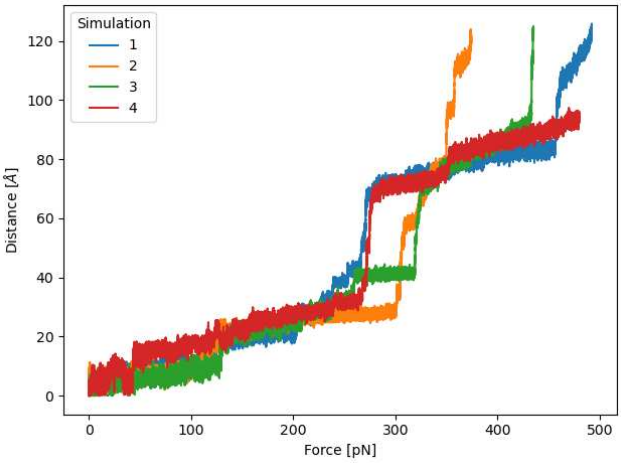


Figure 4.18.: **GFMD Analysis of TBFV reference structure.** The plots show the distance between the initial and current positions of the 5' end during MD simulations of DTV

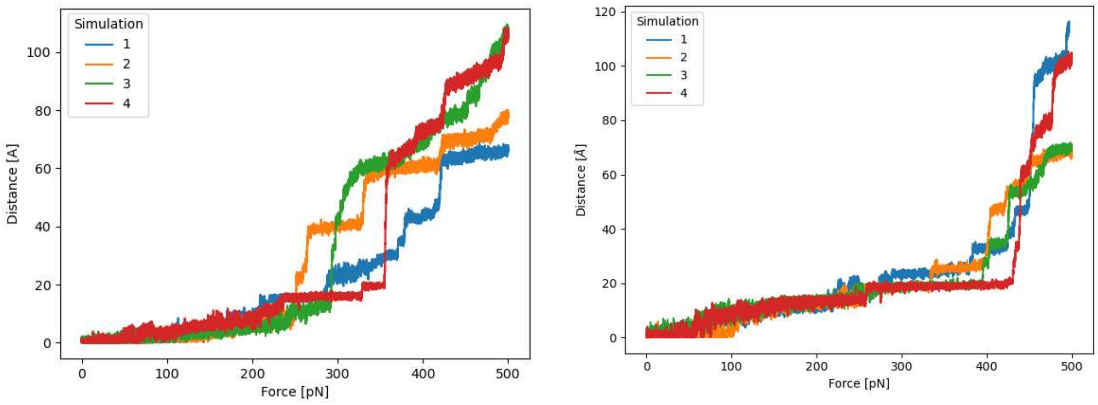
4.4.8. Consistency of the TBFV-Based Design Model

As was also done for the MBFV-based design model, ten sequences were generated and simulated for the TBFV-based model to evaluate the reliability of the design process. Of the 10 generated sequences, only two showed potential resistance to xrRNA and were selected for MD simulations, as shown in Figure 4.19. One of these designs exhibited low structural stability, while the other demonstrated limited pulling stability. Despite manual attempts to improve the sequences, the modifications did not lead to significant improvements in stability.

In Figure 4.19(b), Design 9 shows greater stability compared to Design 8, as seen in panel (a). However, both designs exhibit a steady increase in distance with increasing force, even with low pulling forces, indicating structural instability. In contrast, the MBFV-based designs, shown in Figure 4.17, display a different behavior, where increases in pulling force do not consistently correlate with increases in distance at low pulling forces.

In panel (a), a steep increase in distance is observed even with a pulling force of only 250 pN. The remaining simulations show unfolding occurring at forces around 300 and 350 pN. In the other design (panel (b)), a sharp increase in distance is observed between 400 and 450 pN across the four simulations. Although this might suggest some degree of stability, the distance steadily increases to 20-30 Å up to a pulling force of 400 pN, which already indicates early signs of instability.

These results demonstrate that, although the TBFV-based designs exhibit some resistance to pulling, their stability is noticeably weaker compared to the MBFV-based designs.



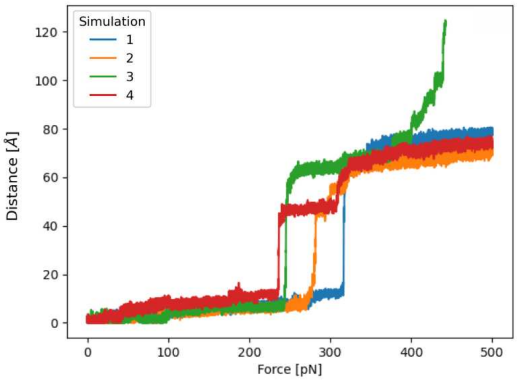
(a) Design 8.

(b) Design 9.

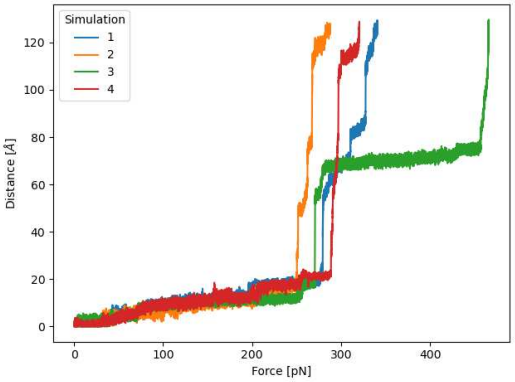
Figure 4.19.: **GFMD Analysis of TBFV designs.** The plots show the distance between the initial and current positions of the 5' end during MD simulations for two selected TBFV designs.

4.5. Experimental Results

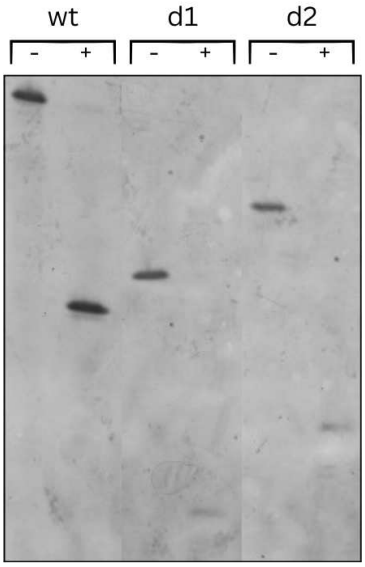
One designed sequence from the TBFV-based model and two sequences generated from the MBFV-based model were selected for laboratory testing. For the TBFV design, Design 9 was chosen, with GFMD results shown in Figure 4.19(b). However, no experimental results for the TBFV design were available at the time of completing this master's thesis. The two MBFV-based designs exhibited stable structures comparable to the reference structure in the GFMD analysis, with one example shown in Figure 4.20(a). Both designs demonstrated partial resistance to Xrn1, although this resistance was significantly lower than that of the wild-type sequence, as shown in Figure 4.20(c). Manual optimization at the nucleotide level was performed on one of the two sequences, which improved its reliability. After these modifications, the sequence achieved resistance values consistently in the range of 250-300 pN across all four simulations, as illustrated in Figure 4.20(b). These modifications ultimately improved the sequence's performance, achieving resistance levels comparable to those of the wild type, as seen in Figure 4.20(d). In addition, in-line probing further validated that the structure matched the expected conformation, providing greater confidence in the design and its functionality.



(a) GFMD Analysis of MBFV Design before modification.



(b) GFMD Analysis of MBFV Design after modification.



(c) MBFV Designs showing partial Xrn1-resistance.

(d) Modified MBFV Design showing improved Xrn1-resistance.

Figure 4.20.: **GFMD and in vitro Testing Results of MBFV Designs.** Panel (a) shows GFMD analysis of the MBFV design before manual optimization, demonstrating partial stability. Panel (b) presents GFMD analysis after optimization, showing improved stability. Panel (c) displays the results from the Xrn1 resistance assay, comparing the two MBFV designs (d1 and d2) with the wild-type (wt) sequence. In panel (d) we see the results after one of the design got further manually modified to improve its resistance (d2-alt).

4.6. Design Comparison

The dataset described in Section 4.1 was utilized to generate a CM. From this, CM sequences were created and used for this analysis. Since the Infrared model based on the MBFV group demonstrated highly promising results, with experimental validation that confirmed the resistance to Xrn1, the comparison focuses exclusively on designs derived from the MBFV model. This analysis aims to evaluate and compare the different design approaches, emphasizing differences in sequence composition and structural properties.

4.6.1. Sequence Diversity

Figure 4.21 shows the sequence logo counts for three datasets: biological data from MBFV members, the `mlocarna` alignment of the designs produced using the Infrared model and the designs generated by the CM model. The MBFV dataset consists of 20 biological sequences, the Infrared dataset includes 1,000 designed sequences, and the CM model dataset contains 100 designed sequences. In addition, the proposed structural motif is displayed. To improve alignment between the datasets, the logo counts were manually adjusted. Specifically, columns with predominantly gaps ($> 95\%$) were removed from the Infrared and CM alignments. In addition, gaps (highlighted in gray) were introduced to create a more comparable plot. The black underlined regions correspond to sequence positions that were completely or partially fixed by constraints in the Infrared design model.

A strong similarity is observed between the CM designs and the biological MBFV data, which can be attributed to the methodology used for the CM generation. Completely conserved positions in the MBFV alignment are also reflected in both the CM and the Infrared models. Notably, a highly conserved pattern, highlighted in yellow, is visible in the MBFV and CM alignments, but not in the Infrared model. This region corresponds to part of PK2. Although this specific sequence motif was not included in the Infrared constraints, the structural constraints in the Infrared model ensure that these positions are capable of base pairing, effectively yielding a similar result as if the motif were explicitly included.

A similar observation can be made for the cyan-highlighted positions, which are again conserved in the MBFV and CM alignments but not in the Infrared alignment. These positions correspond to an extension of the γ helix. Again, the structural constraints in the Infrared model ensure that these nucleotides can form base pairs, replicating similar functional requirements of the conserved positions.

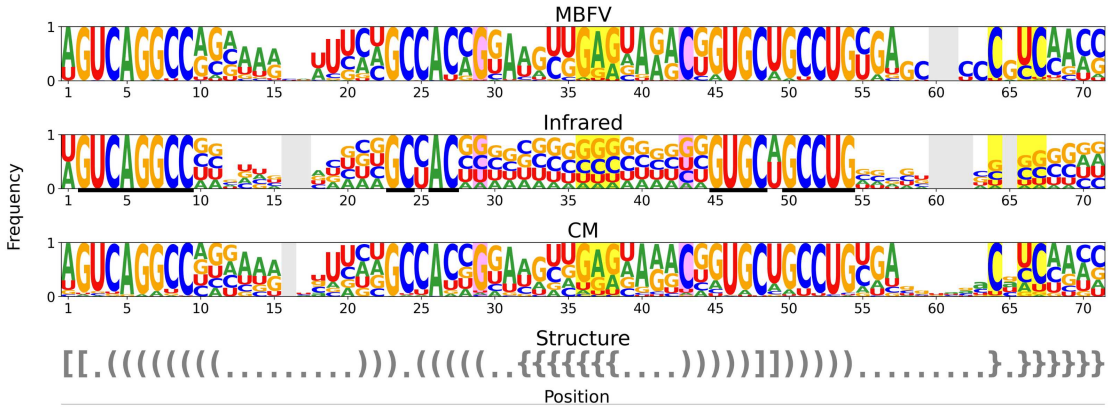
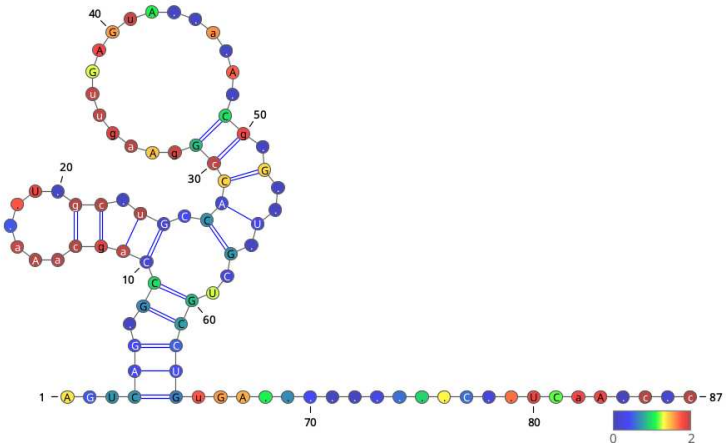
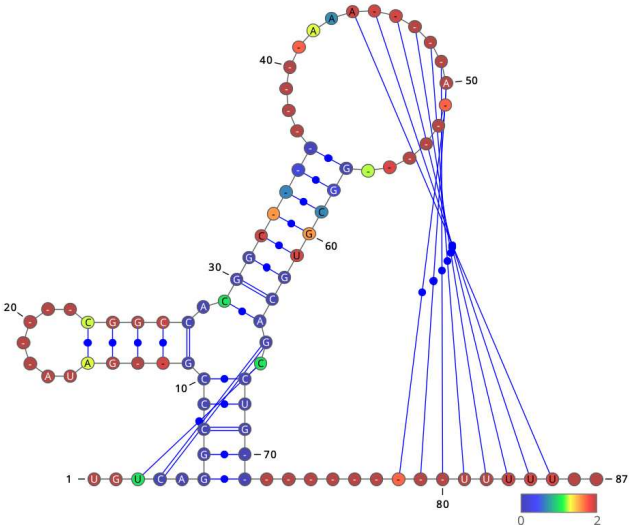


Figure 4.21.: **Sequence Analysis of MBFV models.** Sequence logo counts for MBFV sequences and designs created using the Infrared and CM models based on MBFV data. Gray regions represent manually introduced gaps for alignment purposes. Cyan or yellow highlighted regions indicate interesting positions in the comparison. Black underlined regions correspond to positions that were fixed by sequence constraints in the Infrared RNA design model.

Upon examining the two graphs showing the consensus structure and the Shannon entropy of the different designs (Figure 4.22), we observe a high degree of similarity in both the structural elements and the conserved regions. The CM captures the conservation within the stem α and the multi loop, both of which show low entropy in the CM designs while being entirely fixed in the Infrared model. However, regions such as the structure β display a high entropy in both design models, indicating a variability in the sequence. CM introduces unpaired nucleotides into various stems, a feature that does not align with the biological data. However, these inserted bulges exhibit high entropy associated with gap characters, suggesting that they are rarely incorporated into the sequence. Overall, both models demonstrate comparable consensus structures and Shannon entropy patterns, with only minor differences between them.



(a) Consensus structure and Shannon entropy of CM designs.



(b) Consensus structure and Shannon entropy of Infrared designs.

Figure 4.22.: **Conservation analysis of MBFV models.** Shannon entropy and consensus structure comparisons of CM and Infrared designs based on MBFV data.

4.6.2. Structure Diversity

Figure 4.23 presents a comparison of the structural length distributions for different structural parts of the xrRNA sequences, based on the same analysis as in Section 4.1. For the CM-generated designs, pseudoknot information was derived from an analysis using ViennaRNA. Only sequences with predicted pseudoknot interactions were included

in the comparison of the CM pseudoknot lengths. Specifically, 78% of the CM sequences were included in the PK1 analysis, while 60% of the total CM designs were included in the PK2 analysis.

The Infrared and CM models effectively mimic the MBFV data for many structural parts. The Infrared model generally produces lengths within the observed biological ranges, as these ranges are explicitly built into the model. However, the CM model occasionally generates outliers typically shorter than the observed lengths. For example, in the α region, CM designs show greater variability than Infrared designs or the biological data.

Infrared tends to generate shorter structural parts, particularly for the β region. For example, the stem of β has a mean length of 6 nucleotides in the MBFV data, whereas Infrared designs have a mean of 2 nucleotides. In contrast, CM designs for the stem β exhibit a mean length of 8 nucleotides with no observed variance, indicating that the model does not capture this structural diversity. Both models also struggle to replicate the variability in the stem γ. For both Infrared and CM designs, the length of the stem γ is nearly invariant at 10 nucleotides, whereas the biological data range from 8 to 15.

When examining the pseudoknot, we observe that CM also generates sequences capable of forming only one base pair for pseudoknot 2. The formation of base pairs for pseudoknot 2 in both Infrared and CM designs shows high similarity to the biological data. However, the CM design appears to partially miss the accurate positioning of PK2 within the γ hairpin loop and after the stem α, as it shows longer unpaired regions compared to biological data.

Although both models successfully reproduce many characteristics of the MBFV data, they exhibit limitations in capturing the full variability of specific structural parts, particularly in regions such as stem β and stem γ. Regarding reliability, the CM model shows a higher frequency of generating sequences outside the observed biological ranges, whereas the Infrared model enforces stricter adherence to these constraints.

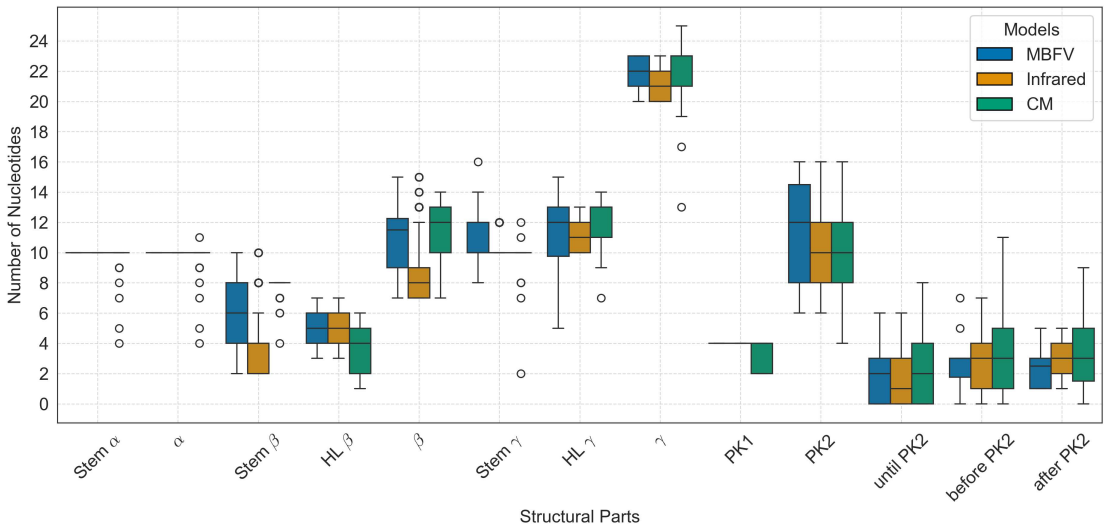


Figure 4.23.: **Structural part analysis of MBFV-models.** Box plot comparing the structural lengths of different xrRNA parts from MBFV biological data, Infrared model designs, and CM model designs.

The base pairing probabilities of the three datasets, comprising natural MBFV sequences, Infrared model designs, and CM designs, were calculated and visualized as heat maps (Figure 4.24). Panel (a) shows the base pairing probabilities of the natural xrRNA sequences from the MBFV group. Panel (b) depicts the probabilities for sequences designed using the Infrared model, with the lower triangle of the matrix highlighting the expected base pairing. Panel (c) presents the probabilities for CM designs, where the lower triangle again represents the expected or desired structure. In all panels, diagonal lines indicate helices or continuous base pairing.

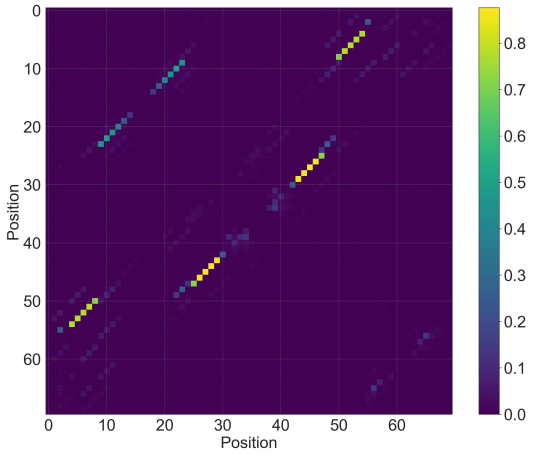
The BPPs of MBFV sequences show three distinct diagonal lines, representing the stems α, β and γ. The α stem (starting around position 5/55) and the γ stem (starting around position 25/48) exhibit high base pairing probabilities, both reaching approximately 0.8. The stem β shows lower probabilities, around 0.6. Pseudoknots cannot be clearly identified.

In panel (b), corresponding to Infrared model designs, the three stems are visible alongside an additional diagonal line, which represents pseudoknot 2 (extending from approximately Position 50/80 to 40/90). These helices show lower base pairing probabilities compared to the MBFV sequences, with a maximum probability of around 0.4. Additionally, some base pairing probabilities between Positions 60 and 70 suggest unwanted interactions. Pseudoknot 1 remains undetectable in the heat map, consistent with the MBFV sequences.

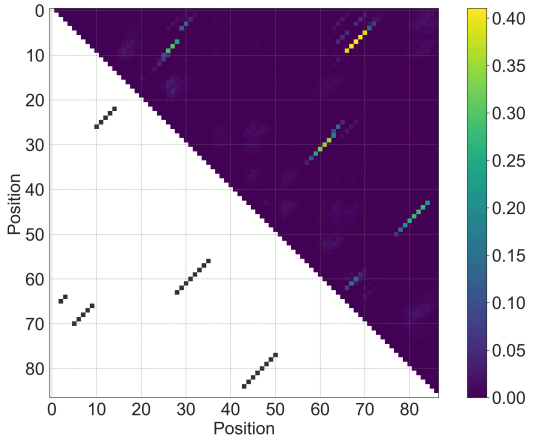
The BPP heat map for CM designs exhibits several diagonal lines, although these are

much less well defined compared to the MBFV and Infrared datasets. The strongest line, corresponding to the stem α, reaches a maximum base pairing probability of 0.35, while the other stems exhibit probabilities of only 0.10 to 0.15. Despite the low probabilities, faint patterns that resemble the stems of α and γ can still be discerned, although these patterns are unclear and poorly defined.

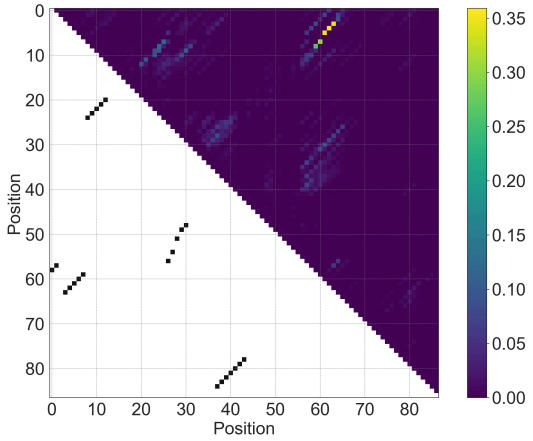
In summary, all three data sets exhibit the motif of three stems, but the clarity and intensity of these features vary significantly. The natural MBFV sequences display the highest base pairing probabilities, followed by the Infrared designs with moderate probabilities, and finally the CM designs, which exhibit low probabilities and less defined structures. Notably, pseudoknot 2 is visible only in the Infrared designs, whereas pseudoknot 1 is absent in all three datasets. This absence may be attributed to the small size of pseudoknot 1, which consists of only two base pairs, making it unlikely that ViennaRNA would accurately capture these interactions.



(a) MBFV sequences.



(b) Infared designs.



(c) CM designs.

Figure 4.24.: **Base pairing probabilities of MBFV-models.** The plots show heat maps of the BPPs. (a) uses biological data of MBFV members. (b) uses sequences created from the Infrared design model. (c) uses sequences created from the CM model.

As explained in Section 3.8.2, the ability to create pseudoknots was also analysed, the results can be seen in Table 4.3. Examination of the BPPs of designed sequences revealed that the Infrared model reliably builds pseudoknot 2, whereas the CM model showed no clear evidence of pseudoknot formation. However,also the biological data did not show evidence of any pseudoknots. Notably, no signs of pseudoknot 1 were observed in the BPPs of either model.

Each design was subsequently evaluated to determine whether the nucleotides within the pseudoknot 1 regions could theoretically form two base pairs. For the Infrared designs, all sequences exhibited the ability to form the two base pairs required for pseudoknot 1, which is unsurprising given that this feature was explicitly included in the model. In contrast, only 58% of the sequences designed using the CM model were able to form both base pairs that comprise pseudoknot 1. The ability to form pseudoknot 2 was further analysed using ViennaRNA. Among Infrared designs, 89.7% were able to form base pairs between the two pseudoknot 2 regions, with an average of approximately 5 base pairs per pseudoknot. By comparison, CM designs formed pseudoknot 2 in 60% of the cases, with an average of 5 base pairs per pseudoknot 2 as well.

In general, 89.7% of the Infrared designs were able to form both pseudoknot 1 and pseudoknot 2, while only 34% of the CM designs achieved this. These results demonstrate that the Infrared model reliably produces the pseudoknots required for the characteristic xrRNA structure. Although the CM model, by capturing conserved sequential information, occasionally generates sequences capable of forming one or even both pseudoknots, the Infrared model ensures consistent pseudoknot formation across nearly all designs.

Table 4.3.: Fraction of sequences that can form pseudoknots in Infrared and CM designs based on ViennaRNA analysis

model	PK1	PK2	both PKs
Infrared	1	0.897	0.897
CM	0.58	0.6	0.34

In Figure 4.25, the sequence length distributions of three datasets are shown: biological MBFV data, designs generated using the Infrared model and designs created using the CM model. Both models are able to approximate the sequence length range of biological data; however, the CM model occasionally generates sequences exceeding the observed biological lengths, some reaching up to 69 nucleotides.

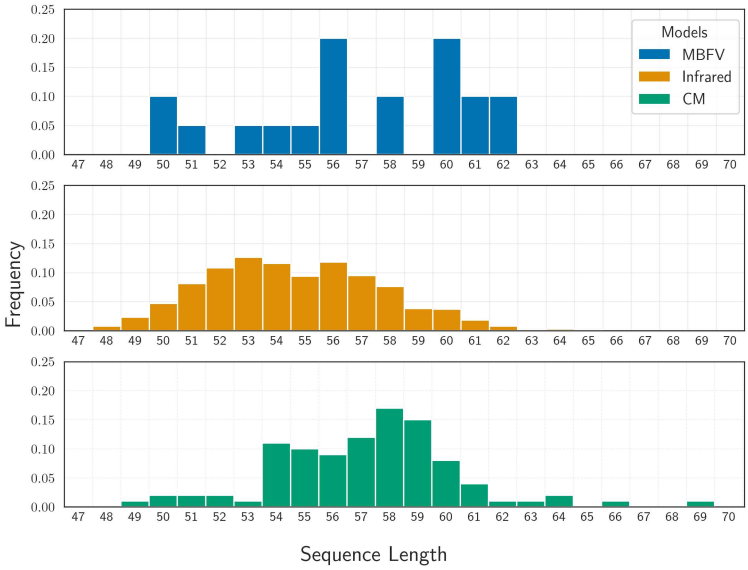


Figure 4.25.: **Length Distribution of MBFV Sequences.** Histograms showing sequence length distributions for MBFV biological data, Infrared model designs, and CM model designs.

In summary, the comparison between the Infrared and CM models highlights significant differences in sequence composition, structural stability, and the ability to replicate the features of natural MBFV xrRNA sequences. Both models effectively capture many aspects of the biological data; however, a key advantage of the Infrared model is its reliable formation of pseudoknots. In particular, the Infrared model achieves superior accuracy in forming both pseudoknots, with a success rate of nearly 90%. In contrast, the CM model successfully forms both pseudoknots in only about one-third of the cases, highlighting its limitations in this use case. A key drawback of the CM model is that it can only include information about sequence and secondary structure, lacking the ability to incorporate tertiary structural features such as pseudoknots. This constraint represents a significant disadvantage when trying to replicate the complex structural characteristics of natural MBFV xrRNA sequences.

Sequence diversity analysis revealed that both models replicate conserved elements, such as the stem α and multi loop, while allowing variability in other regions, such as stems β and γ. The base pairing probability analysis further demonstrated the Infrared model's reliability in producing sequences that form the desired base pairs, a feature not consistently observed in the CM designs.

Another key advantage of the Infrared model lies in its ability to optimize a designed sequence efficiently. The Infrared approach builds a design scaffold that allows for iterative

modification of sequences while ensuring the preservation of critical features, such as base-pairing compatibility and conservation of important motifs. In contrast, the CM model generates individual sequences based solely on the provided covariance model. Adjusting or refining these sequences requires the manual incorporation of additional constraints, a process that can be laborious and time-consuming, particularly when dealing with complex RNA structures. This lack of flexibility limits the CM model's utility for applications requiring iterative refinement or sequence manipulation.

In conclusion, the Infrared model offers a more robust and adaptable framework for RNA design, effectively balancing structural accuracy, pseudoknot formation, and sequence optimization. Given that pseudoknot formation is a crucial aspect of the unique xrRNA structure, and considering that Infrared is better equipped to incorporate this important feature into the design process, it emerges as the more suitable choice for designing xrRNA structures. Moreover, its ability to preserve critical features while enabling targeted modifications represents a key advantage over the CM model. In addition, Infrared enables further manipulation of the sequence or structure to meet specific requirements, enhancing its utility in RNA design.

5. Discussion

The design and analysis of exoribonuclease-resistant RNA sequences represent a significant advancement in the understanding of these unique structures and their applications in synthetic biology. This study aimed to develop a computational framework for designing synthetic xrRNA sequences that replicate the properties of natural flaviviral xrRNAs, with a focus on structural stability and resistance to degradation by Xrn1. Here, research goals, key findings, including limitations, and potential directions for future research are discussed.

5.1. Research Goal

The goal of this thesis was to create a robust design model based on flaviviral sequences that exhibit resistance to exoribonuclease. By analyzing biological data derived from two subgroups, MBFV and TBFV, critical structural and sequential features were identified. Based on these findings, a scaffold for the two design models was developed. This scaffold was designed to mimic the biological data while ensuring variability in the design process. Using this information and the Infrared framework, two xrRNA design models were created, one for each subgroup.

Within these models, sequences were constrained to meet various structural and sequential requirements. Using a Monte Carlo optimization process, the models optimized sequences to adopt the desired target structures. A two-stage optimization approach ensured a reliable process capable of generating sequences that accurately fold into the target secondary structure. Further tertiary structure and stability analyses were conducted to ensure exoribonuclease resistance. Sequences demonstrating the same structural features as their biological counterparts were subjected to experimental validation. For each subgroup, one of these designs was selected and sent to the partner lab in Leipzig for experimental testing of its resistance to Xrn1.

However, the TBFV model exhibited some unresolved uncertainties by the end of this thesis and did not reliably generate exoribonuclease-resistant sequences. Nevertheless, sequences were created that, on the basis of tertiary and stability analyses, exhibited structural similarities to biological data.

In contrast, a reliable computational model was developed for MBFV, which is capable of producing synthetic xrRNA sequences with structural features similar to biological sequences of the MBFV subgroup. Experimental testing validated that a sequence generated using the MBFV model exhibited resistance to Xrn1. Although the resistance level was initially lower than that of natural sequences, manual refinements elevated the resistance to match the level of a biological sequence derived from ZIKV. Notably, this work represents the first successful creation of an artificially designed, functional structural non-coding RNA.

5.2. Key Findings

5.2.1. Computational Modeling and RNA Design

The Infrared-based computational model developed in this study demonstrated its ability to design RNA sequences with structural characteristics that mimic natural xrRNAs. By incorporating flexible structural scaffolds and constraints derived from conserved sequences and secondary structures, the model successfully replicated critical elements such as pseudoknots and the unique ring-like tertiary structure essential for Xrn1 resistance. These findings underscore the potential of computational design to generate functional RNA molecules with customizable properties.

5.2.2. Resistance Validation

Different validation methods were used to further ensure the resistance of the designed sequences. Methods like 3D simulation, as well as MD, gain insight into the similarities between the designed sequences and a natural reference structure. Furthermore, experimental validation confirmed the effectiveness of the computational designs. These findings highlight the interaction between computational modeling and experimental validation in advancing RNA design.

5.2.3. Limitations

This study identified several limitations that affected the reliability of the TBFV-based model in generating Xrn1-resistant sequences. Although some of the sequences designed exhibited the desired structural features and partial stability, the model was unable to consistently produce sequences that reliably folded into the intended tertiary structure. The evaluation of these structures in three-dimensional space was further constrained by the limited availability of tertiary structure data for TBFVs, making comprehensive comparisons challenging.

Furthermore, a limitation arises when using 3D structural simulations for validation.

Although some sequences appeared to adopt the correct target structure in static 3D simulations, they exhibited poor structural stability when tested in MD simulations. Furthermore, the metrics used to evaluate 3D structures are not entirely reliable. Although some metrics indicated a low similarity between the reference and the designs, the MD simulations revealed that some designs exhibited structural stability similar to the reference structure. This discrepancy underscores the limitations of relying solely on 3D simulations and static metrics to validate the resistance or structural similarity of xrRNA, as such methods may not fully capture the dynamic nature of RNAs.

5.2.4. Comparison of RNA Design Models

The selection of an appropriate RNA design approach was critical for generating a model that is both reliable and variable. To validate the chosen approach, a comparison was conducted with an alternative RNA design method, specifically the CM. This comparison highlighted the fundamental differences in how each model addresses the RNA design problem and provided information on their respective strengths and limitations. Although CM effectively identifies conserved secondary structures and important sequence motifs, it struggles to integrate certain complex but essential structural constraints. For example, the analysis of pseudoknot formation revealed that CM could not consistently generate sequences capable of forming pseudoknots. In contrast, the Infrared model allows for the direct incorporation of structural constraints into the design process. As a result, sequences generated using the Infrared model exhibited a much higher likelihood of successfully folding into structures containing pseudoknots. The analysis further demonstrated that pseudoknots could theoretically form in nearly all sequences produced by the Infrared model. This is not surprising, given that the Infrared model allows for the direct inclusion of complex structures like pseudoknots as constraints, a feature that is not supported by the CM model. Moreover, the Infrared model offers significant advantages in sequence optimization. By providing a flexible design scaffold, the Infrared model allows for targeted adjustments to the sequence while ensuring that critical base pairs or conserved motifs remain intact. This flexibility enables fine-tuning of the sequence to enhance its structural and functional properties. In contrast, sequences generated using CM are less suitable for optimization, as modifications to improve certain features often risk disrupting conserved motifs or critical structural elements.

5.3. Future Directions

This master's thesis was conducted as part of a larger project focusing on the design of synthetic RNase resistant structures and their integration in novel riboswitch constructs to regulate mRNA stability. The primary objective of this work was to design synthetic xrRNA sequences that mimic the RNase-resistant structure, specifically those found

in *orthoflavivirus*. This goal was successfully achieved, with the MBFV-based model developed demonstrating a reliable design process capable of generating sequences that closely replicate the structural and functional features of natural xrRNAs.

The next step in this project involves the design of xrRNA inducible riboswitches *in silico*. This will require further manipulation of the designed sequences to meet specific structural and functional requirements. The current model has shown sufficient flexibility to allow targeted modifications, such as altering the stem β, while retaining secondary and tertiary configurations. The stem β exhibits high variability and is therefore a prime candidate for customization. Furthermore, as mentioned earlier, certain FV members lack a whole β structure. Focusing on this region could enable the creation of artificial xrRNA sequences that either lack the β stem or incorporate highly customized versions. Although completely removing the β stem without destabilizing the structure remained unsuccessful, its relative variability makes it the most feasible region for targeted manipulation.

An important future direction also involves expanding the dataset used to construct the design scaffold, as only a subset of available xrRNA data was utilized in this thesis. Incorporating a broader range of sequences could enhance the flexibility of the model by identifying additional conserved regions that must remain fixed, while also revealing areas with potential for more variability. This expanded dataset could also highlight new sequential and structural properties, that contribute to increased stability or to a higher flexibility. By gathering, processing, and analyzing more xrRNA sequences and their structural features, the design scaffold could be refined to allow for more versatile applications. This enhanced model could then be more readily adapted for tasks such as creating riboswitches.

This thesis also provides a framework for developing an Infrared design model based on the analysis of conserved sequences and structural motifs. This approach can be extended to other *orthoflavivirus* subgroups like NKV and ISFV or even to unrelated RNA families with intriguing structures of interest. Artificially designed RNA sequences can be tailored to address specific challenges and applied in a variety of novel contexts.

The field of RNA design is rapidly emerging, offering significant potential for applications in pharmaceutical research, biochemistry, synthetic biology, and nanotechnology. Despite the challenges, such as complex intramolecular interactions, ensuring thermodynamic stability, and maintaining functional integrity, overcoming these obstacles enables the design of highly versatile and specific RNA molecules. Advances in the design of dynamic and responsive RNAs will further expand the scope of applications, enabling the creation of molecules that respond to specific environmental cues or stimuli.

6. Conclusion

This thesis developed a computational framework for designing synthetic xrRNAs that mimic the structural and functional properties of natural flaviviral xrRNAs, with a specific focus on achieving structural stability and resistance to Xrn1. The computational model was built on the analysis of conserved sequences and structural motifs found in natural xrRNAs. Using the Infrared framework for model development, as well as tertiary structure simulations and stability testing, the design process underwent iterative refinement to generate synthetic RNA sequences that accurately mimic the structural features of their natural counterparts. By incorporating flexible structural scaffolds and specific constraints, the model successfully generated sequences that replicate critical elements, such as pseudoknots and the ring-like tertiary structure, while allowing for targeted modifications and adaptability to meet specific design requirements.

The designed sequences demonstrated structural similarity to biological xrRNAs in both secondary and tertiary structure simulations. Experimental validation further confirmed the capability of the framework to generate synthetic xrRNAs, as a designed sequence exhibited Xrn1 resistance, underscoring the practical applicability of the approach.

This work demonstrates that, with the appropriate computational models and design strategies, it is possible to recreate the critical structural features of naturally occurring RNA. The successful development of exoribonuclease-resistant sequences represents not only an important step toward future applications in RNA design but also a groundbreaking achievement in RNA biology, being the first instance of a functional structural non-coding RNA synthesized de novo. Future work could extend this design approach to other RNA families and explore applications such as riboswitch constructs for mRNA stability regulation.

List of Tables

List of Figures

Listings

Bibliography

[1] Hua Ting Yao, Bertrand Marchand, Sarah J. Berkemer, Yann Ponty, and Sebastian Will. Infrared: a declarative tree decomposition-powered framework for bioinformatics. *Algorithms for Molecular Biology*, 19, 12 2024.

[2] Paul G. Higgs. RNA secondary structure: Physical and computational aspects, 2000.

[3] Bruce Alberts. *Molecular Biology of the Cell*. Garland Science Taylor & Francis, 5 edition, 2002.

[4] Xin Li and Peixuan Guo. *RNA Structure and Folding*, pages 3–4. CRC Press, 8 2022.

[5] Dagmar Klostermeier and Christian Hammann. *RNA structure and folding : biophysical techniques and prediction methods*. Walter de Gruyter, 2013.

[6] Sohini Bhattacharya, Eshita Mutt, and Abhijit Mitra. RNA structural bioinformatics, 2013.

[7] David W. Staple and Samuel E. Butcher. Pseudoknots: RNA structures with diverse functions, 6 2005.

[8] Amal S. Abu Almakarem, Anton I. Petrov, Jesse Stombaugh, Craig L. Zirbel, and Neocles B. Leontis. Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Research*, 40:1407–1423, 2 2012.

[9] Ronny Lorenz, Stephan H. Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA package 2.0, 2011.

[10] Karol Szlachta, Ryan Thys, Naomi Atkin, Levi Pierce, Stefan Bekiranov, and Yuh-Hwa Wang. Alternative dna secondary structure formation affects RNA polymerase ii promoter-proximal pausing in human. *Genome Biology*, 19, 07 2018.

[11] Thomas Binet, Séverine Padiolleau-Lefvre, Stéphane Octave, Bérangère Avalle, and Irene Maffucci. Comparative study of single-stranded oligonucleotides secondary structure prediction tools. *BMC Bioinformatics*, 24:422, 11 2023.

[12] Susan M. Freier, Ryszard Kierzek, John A. Jaeger, Naoki Sugimoto, Marvin H. Caruthers, Thomas Neilson, and Douglas H. Turner. Improved free-energy parameters for predictions of RNA duplex stability, 1986.

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

[14] Tianbing Xia, John SantaLucia, Mark E. Burkard, Ryszard Kierzek, Susan J. Schroeder, Xiaoqi Jiao, Christopher Cox, and Douglas H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with watson - crick base pairs. *Biochemistry*, 37:14719–14735, 10 1998.

[15] 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, 2004.

[16] Mirela Andronescu, Anne Condon, Holger H. Hoos, David H. Mathews, and Kevin P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. In *Bioinformatics*, volume 23. Oxford University Press, 7 2007.

[17] Denis Skibinski. Computational studies of exoribonuclease-resistant RNA structures in flaviviruses, 2024.

[18] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures, 1994.

[19] Robert M. Dirks, Milo Lin, Erik Winfree, and Niles A. Pierce. Paradigms for computational nucleic acid design. *Nucleic Acids Research*, 32:1392–1403, 2 2004.

[20] Yann Ponty, Stefan Hammer, Hua-Ting Yao, and Sebastian Will. Advanced design of structural RNAs using RNAredprint. In Ernesto Picardi, editor, *RNA Bioinformatics*, volume 2284 of *Methods in Molecular Biology*, pages 1–15. Springer, 2021.

[21] Édouard Bonnet, Paweł Rzazewski, and Florian Sikora. Designing RNA secondary structures is hard. In *Journal of Computational Biology*, volume 27, pages 302–316. Mary Ann Liebert Inc., 3 2020.

[22] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 04 1998.

[23] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[24] Sean R. Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088, 1994.

[25] Marc Parisien, José Almeida Cruz, Éric Westhof, and François Major. New metrics for comparing and assessing discrepancies between RNA 3d structures and models. *RNA*, 15:1875–1885, 10 2009.

[26] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta crystallographica*, 32:922–923, 7 1976.

[27] Clément Bernard, Guillaume Postic, Sahar Ghannay, and Fariza Tahi. RNAdvisor: a comprehensive benchmarking tool for the measure and prediction of RNA structural model quality-supplementary materials supplementary note 1: State-of-the-art metrics a. general metrics.

[28] Adam Zemla. Lga: A method for finding 3d similarities in protein structures. *Nucleic Acids Research*, 31:3370–3374, 7 2003.

[29] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Genetics*, 57:702–710, 12 2004.

[30] Kliment Olechnovič and Česlovas Venclovas. The use of interatomic contact areas to quantify discrepancies between RNA 3d models and reference structures. *Nucleic Acids Research*, 42:5407–5415, 2014.

[31] Thomas S. Postler, Martin Beer, Bradley J Blitvich, Jens Bukh, Xavier de Lamballerie, Jan Felix Drexler, Allison Imrie, Amit Kapoor, Galina G. Karganova, Philippe Lemey, Volker Lohmann, Peter Simmonds, Donald B. Smith, Jack T. Stapleton, and Jens H. Kuhn. Renaming of the genus flavivirus to orthoflavivirus and extension of binomial species names within the family flaviviridae. *Archives of Virology*, 168, 08 2023.

[32] John S. Mackenzie, Duane J. Gubler, and Lyle R. Petersen. Emerging flaviviruses: the spread and resurgence of japanese encephalitis, west nile and dengue viruses. *Nature Medicine*, 10:S98–S109, 11 2004.

[33] Ernest A. Gould and Tom Solomon. Pathogenic flaviviruses. *The Lancet*, 371:500–509, 02 2008.

[34] Hui Ying Chong, Chiuan Yee Leow, Abu Bakar Abdul Majeed, and Chiuan Herng Leow. Flavivirus infectio–a review of immunopathogenesis, immunological response, and immunodiagnosis, 12 2019.

[35] World Health Organization. Zika virus. `https://www.who.int/news-room/fact-sheets/detail/zika-virus`, 12 2022. Accessed: 2024-12-10.

[36] David Baud, Duane J. Gubler, Bruno Schaub, Marion C. Lanteri, and Didier Musso. An update on zika virus infection. *The Lancet*, 390:2099–2109, 11 2017.

[37] Erica Diani, Anna Lagni, Virginia Lotti, Emil Tonon, Riccardo Cecchetto, and Davide Gibellini. Vector-transmitted flaviviruses: An antiviral molecules overview, 10 2023.

[38] Andrii Slonchak and Alexander A. Khromykh. Subgenomic flaviviral RNAs: What do we know after the first decade of research, 11 2018.

[39] Vinay K. Nagarajan, Christopher I. Jones, Sarah F. Newbury, and Pamela J. Green. Xrn 5' $\rightarrow$ 3' exoribonucleases: Structure, mechanisms and functions. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829:590–603, 06 2013.

[40] Stephanie L. Moon, John R. Anderson, Yutaro Kumagai, Carol J. Wilusz, Shizuo Akira, Alexander A. Khromykh, and Jeffrey Wilusz. A noncoding RNA produced by arthropod-borne flaviviruses inhibits the cellular exoribonuclease xrn1 and alters host mRNA stability. *RNA*, 18:2029–2040, 11 2012.

[41] Erich G. Chapman, David A. Costantino, Jennifer L. Rabe, Stephanie L. Moon, Jeffrey Wilusz, Jay C. Nix, and Jeffrey S. Kieft. The structural basis of pathogenic subgenomic flavivirus RNA (sfRNA) production, 2014.

[42] Rachel A. Jones, Anna-Lena Steckelberg, Quentin Vicens, Matthew J. Szucs, Benjamin M. Akiyama, and Jeffrey S. Kieft. Different tertiary interactions create the same important 3d features in a distinct flavivirus xrRNA. *RNA*, 2021.

[43] Jeffrey S. Kieft, Jennifer L. Rabe, and Erich G. Chapman. New hypotheses derived from the structure of a flaviviral xrn1-resistant RNA: Conservation, folding, and host adaptation. *RNA Biology*, 12:1169–1177, 1 2015.

[44] Andrea MacFadden, Zoe O'donoghue, Patricia A.G.C. Silva, Erich G. Chapman, René C. Olsthoorn, Mark G. Sterken, Gorben P. Pijlman, Peter J. Bredenbeek, and Jeffrey S. Kieft. Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs. *Nature Communications*, 9, 12 2018.

[45] Matthew J. Szucs, Parker J. Nichols, Rachel A. Jones, Quentin Vicens, and Jeffrey S Kieft. A new subclass of exoribonuclease-resistant RNA found in multiple genera of flaviviridae. *mBio*, 11, 2020.

[46] Roman Ochsenreiter and Michael T. Wolfinger. Strukturierte RNAs in viren. *BioSpektrum*, 29:156–158, 3 2023.

[47] Michal J. Boniecki, Grzegorz Lach, Wayne K. Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M. Rother, and Janusz M. Bujnicki. SimRNA: A coarse-grained method for RNA folding simulations and 3d structure prediction. *Nucleic Acids Research*, 44, 12 2015.

[48] Juliusz Stasiewicz, Sunandan Mukherjee, Chandran Nithin, and Janusz M. Bujnicki. QRNAs: Software tool for refinement of nucleic acid structures. *BMC Structural Biology*, 19, 3 2019.

[49] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. „ November 2015.

[50] Eric P. Nawrocki and Sean R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29:2933–2935, 11 2013.

[51] Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, 3:e65, 4 2007.

[52] Sebastian Will, Tejal Joshi, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. LocaRNA-p: Accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18:900–914, 5 2012.

[53] Martin Raden, Syed M. Ali, Omer S. Alkhnbashi, Anke Busch, Fabrizio Costa, Jason A. Davis, Florian Eggenhofer, Rick Gelhausen, Jens Georg, Steffen Heyne, Michael Hiller, Kousik Kundu, Robert Kleinkauf, Steffen C. Lott, Mostafa M. Mohamed, Alexander Mattheis, Milad Miladi, Andreas S. Richter, Sebastian Will, Joachim Wolff, Patrick R. Wright, and Rolf Backofen. Freiburg RNA tools: a central online resource for RNA-focused research and teaching. *Nucleic Acids Research*, 46:W25–W29, 7 2018.

[54] Rhese D. Thompson, Derek L. Carbaugh, Joshua R. Nielsen, Ciara M. Witt, Edgar M. Faison, Rita M. Meganck, Atul Rangadurai, Bo Zhao, Jeffrey P. Bonin, Nathan I. Nicely, William F. Marzluff, Aaron T. Frank, Helen M. Lazear, and Qi Zhang. Lifetime of ground conformational state determines the activity of structured RNA. *Nature Chemical Biology*, 2 2025.

[55] X. de Lamballerie, S. Crochu, F. Billoir, J. Neyts, P. de Micco, E. C. Holmes, and E. A. Gould. Genome sequence analysis of tamana bat virus and its relationship with the genus flavivirus. *Journal of General Virology*, 83:2443–2454, 10 2002.

[56] Jerrold H. Zar. Significance testing of the spearman rank correlation coefficient, 1972.

[57] Hua-Ting Yao, Yann Ponty, and Sebastian Will. *Developing Complex RNA Design Applications in the Infrared Framework*, chapter RNA Folding - Methods and Protocols, pages 285–313. Springer US, 2024.

[58] Max Ward, Eliot Courtney, and Elena Rivas. Fitness functions for RNA structure design. *Nucleic Acids Research*, 51:E40, 4 2023.

[59] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure, 1990.

[60] Ivo L. Hofacker and Ronny Lorenz. Predicting various thermodynamic properties — viennarna package. https://www.tbi.univie.ac.at/RNA/ViennaRNA/doc/html/partfunc/thermodynamics.html#c.vrna_pr_structure, 2024. Accessed: 2025-04-02.

[61] SWISS-MODEL. Structure assessment tool, 2025. Accessed: December 19, 2024. Available at: https://swissmodel.expasy.org/assess/.

[62] Vilnius University Institute of Biotechnology. Cad-score: Contact area difference-based function for evaluation of protein structural models, 2025. Accessed: December 19, 2024. Available at: https://bioinformatics.lt/cad-score/.

[63] Elizabeth E. Regulski and Ronald R. Breaker. *In-Line Probing Analysis of Riboswitches*, pages 53–67. Humana Press, Totowa, NJ, 2008.

[64] Armin O. Schmitt and Hanspeter Herzel. Estimating the entropy of dna sequences. *Journal of Theoretical Biology*, 188:369–377, 10 1997.

Acronyms

BPP Base Pairing Probability. 15

CAD-score Contact Area Difference Score. 9

CM Covariance Model. 7

DTV *Deer Tick Virus*. 21

ED Ensemble Defect. 28

EP Equilibrium Probability. 28

FV *Orthoflavivirus*. 10

GDT_TS Global Distance Test Total Score. 8

ISFV insect-specific flavivirus. 10

MBFV mosquito borne flavivirus. 10

MD Molecular Dynamics. 18

MFE Minimal Free Energy. 4

MSA Multiple Sequence Alignment. 7

MVEV *Murray Valley Encephalitis virus*. 21

NKV no known vector flavivirus. 10

RMSD Root Mean Square Deviation. 8

RNA Ribonucleic acid. 1

TABV *Tamana Bat Virus*. 21

TBFV tick borne flavivirus. 10

TM-score Template Modeling Score. 9

Xrn1 5'-3' exoribonuclease 1. 1

xrRNA exoribonuclease-resistant RNA. 1

ZIKV *Zika Virus*. 10

A. Supplementary Resources

A.1. Supplementary Data

This section provides supplementary data from MBFV and TBFV used in the context of this Master's thesis.

Table A.1.: **Length of structural features of MBFV xrRNAs.** The table shows the lengths of individual structural features of xrRNA1 from members of the MBFV group. Values represent the number of nucleotides. "Spacer" refers to the number of unpaired nucleotides located between the α stem and the 3′ end of PK2. Asterisks (*) indicate unpaired nucleotides located within the γ loop, upstream of PK2 (towards the 5′ end) and downstream of PK2 (towards the 3′ end).

Virus	**Stem α**	**Stem β**	**Loop β**	**Stem γ**	**Loop γ**	**Multi Loop**	**Spacer**	**PK1**	**PK2**	**unpaired 5′***	**unpaired 3′***	**Total length**
AROAV	10	4	4	10	12	9	2	4	12	3	3	54
BAGV	10	10	5	10	12	9	0	4	16	1	3	62
BANV	10	4	5	14	9	9	3	4	10	3	1	56
DENV1	10	2	7	12	9	9	6	4	8	5	0	56
DENV2	10	2	7	12	9	9	6	4	8	5	0	56
DENV3	10	2	7	12	9	9	6	4	6	5	1	55
DENV4	10	4	4	10	10	9	2	4	8	3	3	50
JEV	10	8	4	10	13	9	0	4	14	1	3	60
KOKV	10	4	4	10	10	9	0	4	12	1	3	51
KOUV	10	6	6	10	13	9	3	4	12	4	3	60
KUNV	10	6	7	10	13	9	0	4	16	1	3	61
MVEV	10	10	5	10	12	9	0	4	16	1	2	62
NMV	10	2	5	10	10	9	2	4	10	2	3	50
SLEV	10	6	6	10	13	9	2	4	14	3	3	60
TMUV	10	10	3	10	12	9	0	4	12	1	5	58
UGSV	10	6	3	8	15	9	0	4	8	3	7	53
USUV	10	8	4	10	13	9	0	4	16	1	3	60
WESSV	10	8	4	12	11	9	2	4	10	3	3	58
WNV	10	10	3	10	13	9	0	4	16	1	3	61
ZIKV	10	6	5	16	5	9	4	4	8	1	0	56

Table A.2.: **Length of structural features of TBFV xrRNAs.** The table shows the lengths of individual structural features of xrRNA1 from members of the TBFV group. Values represent the number of nucleotides. "Spacer" denotes the number of unpaired nucleotides between the α stem and the 3′ end of PK1, or PK2 in the case of MPFV and XiFV. "Between PKs" indicates the number of unpaired nucleotides between PK1 and PK2. Asterisks (*) indicate unpaired nucleotides located within the γ loop, upstream of PK2 (towards the 5′ end) and downstream of PK2 (towards the 3′ end). Double asterisks (**) indicate the number of nucleotides being in a bulge and/or internal loop within the corresponding stem.

Virus	**Stem α**	**B./Int. α****	**Stem β**	**Loop β**	**B./Int. β****	**Stem γ**	**Loop γ**	**Multi Loop**	**Spacer**	**PK1**	**between PKs**	**PK2**	**unpaired 5′***	**unpaired 3′***	**Total length**
ALKV	22	4	8	4	0	10	10	14	8	6	0	12	3	0	84
DTV	24	0	6	9	0	8	11	14	12	6	2	14	2	0	90
GGV	24	0	6	7	1	14	4	14	15	6	2	8	0	0	88
KFDV	24	2	8	4	0	10	10	14	8	6	2	10	4	2	84
KSIV	28	2	8	3	0	8	12	14	8	6	2	8	5	3	86
LGTV	24	4	10	3	0	10	10	14	11	6	0	10	4	0	89
LIV	24	3	6	4	0	8	12	14	11	6	0	10	5	1	85
MPFV	16	1	4	7	0	14	5	11	9	0	0	10	0	0	53
NEGV	26	1	6	4	0	8	12	14	11	6	0	10	5	1	85
OHFV	24	4	6	4	0	8	12	14	11	6	0	10	5	1	86
POWV	24	0	6	9	0	8	11	14	14	6	2	8	4	3	89
SGEV	26	3	6	4	0	8	12	14	10	6	0	10	5	1	86
SREV	20	6	0	0	0	12	8	16	6	10	2	14	0	1	78
TBEV	22	5	6	4	0	8	12	14	11	6	0	10	5	1	85
TYUV	28	0	0	0	0	12	8	16	5	10	2	8	3	1	76
XiFV	12	0	6	6	0	8	9	10	12	0	0	10	2	1	62

A.2. Code

A.2.1. Infrared Model Example

The Python script in Listing A.1 demonstrates an example of model creation using the Infrared framework. The code is based on the MBFV design approach but has been simplified for demonstration purposes. While the essential scaffold constraints for sequences and structures are included, more detailed constraints, such as the control of feature length relations, have been omitted.

```
import infrared as ir
import infrared.rna as rna

# Define the input sequence
consensus_sequence =
'NNWGUCAGGCCXXXXNNNXXXXXXXXGCYACNXXXXXXXXXXXXNNNXXXXXNXXXXXXXXNGUGC
    WGCCUGXXXXXXXXXXXNNNNN'

# Target secondary structure and pseudoknot structures
target_structures = [
'.....((((((((((.......)))))).(((((((.....................)))))))..)))))
    ..............',
'..((.............................................................)).....
    ..............',
'.........................................(((((((.................
    ......)))))))..'
]

# Variable regions in stems and loops (index ranges)
variable_stems = [[(11, 14), (22, 25)], [(32, 35), (56, 59)], [(46, 50),
    (77, 81)]]
variable_loops = [(18, 21), (36, 42), (52, 55), (71, 76)]

# Define the Infrared model
model = ir.Model()
sequence_length = len(consensus_sequence)
# Set up new variable X with 5 possible values (A, C, G, U, Gap)
model.add_variables(sequence_length, 5, name='X')

# Mapping from IUPAC symbols to nucleotides + gap
extended_iupac = {
    'A': 'A',
    'C': 'C',
    'G': 'G',
    'U': 'U',
    'R': 'AG',
    'Y': 'CU',
    'W': 'AU',
```

```
    'N': 'ACGU',  # Any base
    'X': 'ACGU-', # Any base or gap
    '-': '-',     # Gap
    '.': '-',     # Gap
}

# Define base-pair compatibility table including gaps
_bpcomp_tab_gap = [
    (0, 3),  # A-U
    (1, 2),  # C-G
    (2, 1),  # G-C
    (2, 3),  # G-U
    (3, 0),  # U-A
    (3, 2),  # U-G
    (4, 4)   # Gap-Gap
]

# Base pair compatibility constraint
ir.def_constraint_class(
    'BPComp',
    lambda i, j: [i, j],
    lambda x,y: 1 if (x,y) in _bpcomp_tab_gap else 0,
)

# Base pair incompatibility constraint
ir.def_constraint_class(
    'BPNotComp',
    lambda i, j: [i, j],
    lambda x, y: 0 if (x, y) in _bpcomp_tab_gap else 1,
)

# Define functions to check if position is a gap or not
ir.def_function_class(
    'NotGap',
    lambda i: [i],
    lambda x: 1 if x != 4 else 0,
)

# Define constraints for allowed gap placement in loops and stems
ir.def_constraint_class(
    'GapsRight',
    lambda i: [i, i + 1],
    lambda left, right: 0 if left == 4 and right != 4 else 1,
    # NNN-- True
    # --NNN False
)

ir.def_constraint_class(
```

```
    'GapsLeft',
    lambda i: [i, i + 1],
    lambda left, right: 0 if left != 4 and right == 4 else 1,
    # --NNN True
    # NNN-- False
)

# Function to convert IUPAC symbol to numerical values
def iupac_values(symbol):
    return [rna.nucleotide_to_value(nt) for nt in extended_iupac[symbol]]

# Add sequence constraints (IUPAC)
for pos, symbol in enumerate(consensus_sequence):
    model.add_constraints(ir.ValueIn(pos, iupac_values(symbol)))

# Add base pair constraints from target structures
for structure in target_structures:
    base_pairs = rna.parse(structure)
    model.add_constraints(BPComp(i, j) for (i, j) in base_pairs)

# Gap placement rules to avoid redundant sequences
for start, end in variable_loops:
    model.add_constraints(GapsRight(i) for i in range(start, end))

for stem_a, stem_b in variable_stems:
    x, y = stem_a
    model.add_constraints(GapsRight(i) for i in range(x, y))
    x,y = stem_b
    model.add_constraints(GapsLeft(i) for i in range(x, y))

# Define objective features (simplified for illustration)
# Example: Feature to minimize total sequence length (excluding gaps)
model.add_functions([NotGap(i) for i in range(sequence_length)], '
    totLength')

# Set weights for the features (illustrative values)
model.set_feature_weight(-1.0, 'totLength')  # Minimize total length

# Model is set up and can be used for sampling or optimization

# Example usage of Model
# Initialize a sampler for the model
sampler = ir.Sampler(model)

# Generate 10 sequence samples
samples = [sampler.sample() for _ in range(10)]

# Print the sampled sequences (without gaps)
```

```
for sample in samples:
    sequence = rna.values_to_seq(sample.values()[:sequence_length])
    print(sequence.replace('-', ''))  # Remove gaps for clean output
```

Listing A.1: **Example model:** RNA design with Infrared.

A.2.2. Code Availability

The code and scripts developed as part of this thesis are available at:

- GitHub Repository: `https://github.com/GutKat/MA.git`