THE UNIVERSITY OF CHICAGO

THE STRUCTURAL DETERMINATION OF VIRAL RNA ELEMENTS USING CHAPERONE ASSISTED RNA CRYSTALLOGRAPHY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOCHEMISTRY AND MOLECULAR BIOPHYSICS

BY

CHRISTINA ALEXANDRA ROMAN

CHICAGO, ILLINOIS

DECEMBER 2021

This work is dedicated to my father Miguel Roman, whose tireless support of my dreams sustained me each day and continues to drive my passion for pursuit of discovery truth and excellence.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I'd like to thank my mentor Dr. Joseph Piccirilli for directing my graduate education in biochemistry. Additionally, the faculty of the entire department of biochemistry and molecular biology each made valuable contributions to my development and training. Dr. Phoebe Rice served as my advocate and champion since I first joined the University of Chicago community as a PREP student. My co-author Anna Lewicka was a foundation of support and guidance through out each of the projects enumerated here in, her direct scientific contributions and others are enumerated at the end of each chapter. My peers have also played an instrumental role in my development and the connections we've build in the student organizations we've created and lead together will surely stand the tests of time. The numerous Uchicago administrators who mentored me in Diversity Equity and Inclusion work have opened the door for my next career and their training was as important as my scientific mentorship. My work was also supported by generous funding from the HHMI Gilliam Fellowship and NIH Chemical Biology Training Grant T32GM008720. It is with profound gratitude that I recognize all of the support these individuals and institutions have provided.

Abstract

RNA is an incredibly diverse molecule, simultaneously encoding for proteins and adopting tertiary structures capable of recruiting macromolecular components, binding ligands and performing catalysis. The nature of RNA structure and how it performs these biologically essential functions is a rapidly advancing field in biochemistry. Of particular focus are the structures and mechanisms of viral RNA elements which support and propagate viral replication and proliferation. While RNAs are generally recalcitrant to Crystallization, Chaperone Assisted Crystallography is a new technology that enables rapid crystallization and resolution of pertinent structured viral RNAs. In this work I present the high-resolution structures of two viral RNA elements. Both represent the first structure of either element resolved using crystallography. They provide insights into a new RNA structural motif and the mechanisms these viruses employ to regulate translation of their genes. The Programmed Frameshifting element of SARS-CoV-2 adopts two conformations to induce a frameshifting event during translation of the SARS-CoV-2 gRNA. Here the first high resolution structure of the Programmed Frameshifting Element was solved to 2.09Å and reveals what may be a second linear conformation of this frameshifting element. The Pea Enation Mosaic Virus 2 Translation Enhancer is a cap-intendent translation enhancer found in the 3' UTR of plant viruses. Presented here is the first crystal structure of this class of translation enhancers solved to 2.75Å. This structure reveals a new C-turn motif which creates a solvent exposed binding site for the eukaryotic initiation factor eIF4E. This work demonstrates the potential for chaperone assisted RNA crystallography to improve RNA crystallography and advance the study of viral RNA element structures.

1. Chapter 1: Thesis Introduction:

1.1 A review of RNA structural biology and the role of crystallography in resolving pathogenic RNA structures.

RNA structure is a rapidly growing field with ever increasing potential to contribute to advances in medicine, basic science, and biotechnology. As our ability to detect, predict and model structured RNA elements improves, their prevalence in every aspect of biology becomes increasingly obvious. As a result, RNA elements are emerging as promising new drug targets, especially in the midst of a global pandemic driven by a positive sense RNA virus[51,64]. Computational modeling of RNA structures is at the forefront of many of the recent advances[8, 39, 68]. While secondary and tertiary structure modeling and dynamics simulations are becoming increasingly powerful and accurate, they are limited by the existing reference datasets of experimentally determined RNA structures[18]. To address the growing gaps in our understanding of RNA folding, function, and structure we must employ highly interdisciplinary approaches to characterizing RNA structures to fully represent their dynamic, conditional and interactive nature.

Experimentally determined structures are a keystone of this multifaceted field. These structures provide motif templates that new RNA sequences can be mapped upon[55]. They also serve as a biophysical reference to calculate the likelihood and stability of the variety of interactions ribonucleotides can form in complex folds. The impact of experimentally determined RNA structures is the comprehensive atomic resolution snapshot they provide of one RNA in one conformation. Not only do these structures generate information about the general architectures, but they can be broken down into smaller motifs and structural elements, which can be classified and re-classified as new representatives of these folds as they emerge[16, 39]. Accurate classification of each level of structural organization for motifs generates logical parameters defining the

requirements for specific folds that can then be applied to sequence analysis and tertiary modeling[25]. However, the low throughput nature of crystallography limits the rate of discovery of new elements and motifs. Additionally, RNA elements often function through dynamics, switching from one conformation to another in response to different conditions[40]. This capacity for structural dynamics makes structured RNAs highly adaptable genetic regulatory elements which become particularly useful for viruses and bacteria. Unfortunately, crystallography is ill-equipped to yield data about dynamics and multiple conformations. Therefore, additional biochemistry and biophysical characterization is often combined with crystal structures to generate a comprehensive understanding of the way the structure relates to the mechanism of action. Alternative methods of structure determination can also provide valuable dynamics data such as; cryo-EM, which can separate images of differently folded RNAs in the grid; NMR, which can detect real time changes in interacting regions of RNAs; and Single Angle Xray Scattering (SAXS) which can generate low resolution envelopes that sum a conformationally heterogenous population of RNAs in solution. Each of these methods is informed by, and informs computational structural techniques. Therefore, RNA structural studies demand a highly interdisciplinary approach to structure determination and functional characterization; no one method is sufficient on its own[27,50,68].

Using crystallography to investigate novel structured RNAs is an almost paradoxical endeavor, however, as crystallography is rarely successful on poorly studied RNA targets. Often an RNA element must have its secondary structure determined through chemical probing to inform the design of suitable crystallization constructs. This is where computational modeling of novel RNAs, identified through phylogenetics and deep sequencing, is most helpful in supporting experimental structural determination. Understanding what we know, what information we are

missing and what information a structure may provide are all crucial to making savvy investments of time and effort in RNA crystallographic studies.

RNAs are recalcitrant to crystallization due to their biophysical properties[59]. Unlike proteins, which have 20 unique residues in their polymers, RNA only uses four monomers, making the specific repeating set of surface interactions needed to form a lattice less likely to occur[59]. The 2'OH moiety also contributes to the instability of RNAs as it enables spontaneous hydrolysis, making a long crystallization process less likely to occur. Many RNAs are also highly flexible or adopt multiple native conformations, often these dynamics are a critical component of their native functions[60]. In other words, RNAs are flexible by nature's design. The flexibility and conformational diversity of RNAs interferes with forming a consistent and repeating crystal lattice. Errors in the lattice, for example due to one region being poorly ordered in the crystal, leads to a loss of data as x-rays do not diffract consistently from the same atoms in different positions. Therefore, conformationally heterogenous populations of RNAs or RNAs with large flexible regions are poor crystallizers[40]. Small, homogenous, densely packed RNAs do crystalize readily, continuous helices, dense pseudoknots or other rigid structures have been known to crystalize. However, the majority of structured RNA elements are not so rigidly folded often RNAs contain large single stranded regions of solvent exposed loops that don't adopt a consistent position. Often RNAs which do crystalize are missing density in flexible regions where no repetitive position of atoms exists to consistently bend x-rays into distinct spots on a diffraction pattern. RNA constructs can be cloned to maximize properties which favor crystallization in the same way proteins are mutated into crystallizable constructs[62]. Flexible regions like over hangs or large loops can be mutated to blunts end or stable GNRA tetraloops respectively[62]. However, each modification has the potential to remove information about the native structure.

Chaperone assisted RNA crystallography provides another route to increase the throughput of crystallographic analysis of novel structured RNAs. In this approach engineered RNA-binding antibody fragments (Fabs) serve as crystallization chaperones[23]. These antibody fragment RNA complexes can be generated in one of two ways. For a novel RNA, with minimal biochemical characterization, a phage displayed library of antibody fragments can be used to perform an iterative selection for Fabs that bind the RNA with high affinity[62]. These Fabs are then expressed and purified and made into complexes with their RNA targets, and the complexes are then advanced to crystallization trials.

If the secondary structure of the RNA is known or predicted with reasonable accuracy, we can apply the grafting approach of chaperone assisted RNA crystallography. In this method, a Fab-binding epitope from an existing RNA/Antibody pair is installed into a similarly shaped motif along the new RNA target[23]. Thus far our lab has had success replacing small solvent exposed loops with the BL3-6 binding sequence AAACA which must be closed by a GC pair (Figure 1.1). We replace larger loops and bulges with a twelve-nucleotide loop of the sequence AUAGUAUAUCAA, which binds Fab HAVx (Figure 1.1)[22,23]. Hairpin loops can be replaced with a bent loop, paired region, loop motif which binds Fab-BRG (Figure 1.1)[63].

Both approaches, either a de-novo selection or grafting, are able to generate more stable native conformations of their RNA counter parts in the crystal. Phage display selected Fabs are able to bind the native structure of the RNA and hold it in place[62]. In the grafting approach the Fab likely restrains the stem-loop or bulge that it binds, there by taking a flexible region of the RNA and locking it in place[22,23,63]. Our Chaperone assisted RNA crystallization method has also generated structures of RNAs that were too large for NMR and too small for cryo-EM but unable

to crystalize in free RNA crystallization trials. Examples of such RNAs are discussed later on in the thesis namely the SARS CoV-2 Frameshifting Element and the PEMV2 PTE.

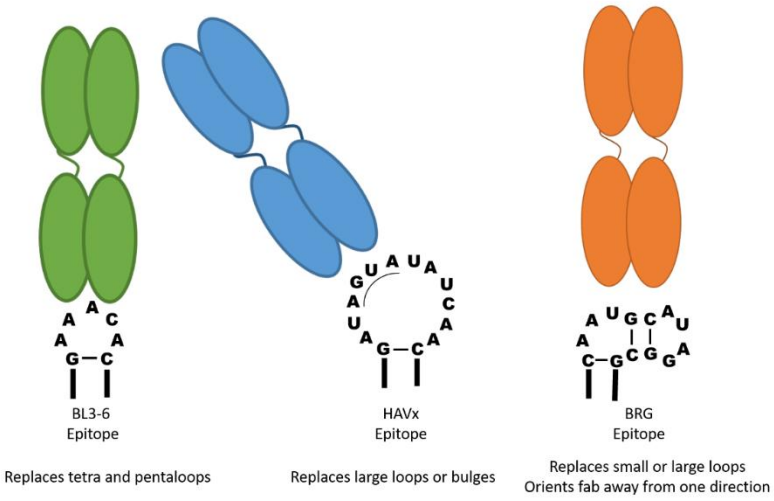


Figure 1.1 Graftable fAB-Binding RNA epitopes and the relative orientation of their respective fABs

1.1.1 Computational RNA structure prediction and modeling approaches and limitations

To understand the importance and intersection of RNA crystallography and pathogenic RNA elements we must first put high resolution crystal structures into context of the current state of RNA structural studies. Computational RNA structure modeling serves to analyze, classify, predict and simulate RNA structures based on their primary sequence[39]. RNA folding can be thought of as a hierarchal process where the secondary structure is formed before the tertiary structure[45]. Exceptions to this would be RNAs that fold co-transcriptionally or fold into their functional states only with the aid of chaperones or remolding proteins.

In the hierarchal approach to structure predictions, RNA sequences are analyzed for the potential to form complementary stretches and single stranded regions. This folding strategy works well for most RNAs with one conformation, but RNAs with multiple conformations, or which get

remodeled post transcriptionally can easily be misfolded by this approach [27,40]. Important paired regions are often conserved across homologs, although the precise sequences that create these paired regions can vary while retaining the same structure[39,65]. Therefore, to detect conservation algorithms look for consistent patterns of co-variation of paired bases[27,65]. Patterns of co-variation across multiple related sequences known to perform the same functions can direct hierarchal folding strategies[65]. This phylogenetic approach to RNA structure prediction remains the gold standard and serves as the foundation for many of the more refined approaches discussed going forward[65].

In most RNA folding software the hierarchal folding assumption is combined with free energy minimization[66]. Thermodynamics adds a second level of evaluation were the potential secondary and tertiary interactions are generated, modified, and iteratively adjusted to find a final conformation with the lowest free energy[67]. Thermodynamics based RNA folding algorithms work particularly well for secondary structure prediction compared to tertiary structure prediction because generally the free energy of WC base pairs is much larger (-0.9 to -3.4 kcal/mol) than that of individual tertiary interactions which tend to only contribute between -0.3 to -1.5 kcal/mol to the structure[68]. The small energetic differences between two tertiary conformations is due to the fact that they differ primarily in helical orientation but not bonding interactions. Both conformations likely have the same number of interactions or only low energy interactions to differentiate them which makes the two conformations appear energetically highly similar[68]. Algorithms may simply choose the marginally more stable conformation as the final structure, and this error can be thought of as getting trapped in an energy well that may not necessarily represent the functional structure[2].The artificial energy well problem limits the accuracy of free energy minimization strategies especially when applied to dynamic RNAs.

To more accurately predict long distance interactions like pseudoknots or kissing loop interactions, folding algorithms increasingly rely on experimental data like structural probing and known intramolecular interactions from existing RNA structure models[68, 69]. Modeling programs like RNA As a Graph (RAG) use a strategy called atomistic modeling where the backbone of single stranded regions connecting helices are broken down into combinations of their rotatable bonds to determine all the possible positions the helices may occupy relative to one another[70]. The radius of gyration of the helices eliminates some of the possible arrangements due to steric clashes and from there the remaining possible arrangements or loops and helices are assayed for the potential for pseudoknots or kissing loops to form between them[68, 70]. The RNA As a Graph (RAG) approach has played an essential role in determining the structures of RNA elements in the SARS-CoV-2 genome [28, 38].

Determining if an RNA sequence likely occupies multiple conformations can help determine the best modeling approach for a given target. To detect the potential for multiple conformations an analysis of the Boltzmann ensemble of structures resulting from the energy minimization procedure can reveal the most probable relevant conformations among a set of low energy folds[68]. Examples of these programs include GT fold and the Most Informative Base Pair program [37, 71]. Maximum entropy models can be even more accurate when dealing with RNAs with multiple conformations because it accounts for probability of each conformation, expressed as a Shannon entropy[52]. Researchers can also model the effects of different environmental conditions on the probabilities of each fold using maximum entropy models. For example, idealized solvent assumptions can generate errors in the RNA's fold, maximum entropy models can calculate the probably of each fold in idealized and non-idealized solvent to check for the effect of those assumptions on the population of low energy models [52]. The probability of any given fold can be

expressed in the form of Shannon entropies which is a statistical probability that can represent any number of biophysical parameters[44]. Not only can Shannon entropies contribute to maximum entropy models, but they can also be applied to biophysical and biochemical data to build a more detailed picture of the dynamics at hand. For instance, calculating the Shannon entropies from force extension curve data for RNAS has proven useful to identify the number of stable conformations a given RNA is most likely to occupy[15].

Existing crystal structures inform each of these methods. As the dataset of experimentally determined structures, which provide reference points for each of these algorithms, expands the more accurate they will become[68]. For example, the thermodynamic stability of a modeled fold is calculated from the known stability of the individual base interactions that comprise it[36,58]. However, working off of a small pool of examples of non-canonical pairs and unusual stacking arrangements can lead to inaccuracy in these calculations[68,39]. Improving the diversity of contributing known RNA motifs improves the accuracy [36,58]. The frequency of occurrence of different types of long-distance interactions observed in experimentally determined RNA structures also defines the likelihood of each of type of interaction in the statistical modeling methods[27,68]. This means that flexible, crystallization resistant motifs are necessarily underrepresented and as a consequence are treated as improbable when its more likely that highly flexible loops and bulges are common in native RNA structures. Both underrepresentation and inconsistent classification of these rarer motifs in databases limit the accuracy and precision of tertiary folding predictions [27, 39, 68]. One way to address this problem is to generate more and more diverse RNA crystal structures. Technologies like chaperone assisted RNA crystallography create an avenue to crystalize RNAs in conformations that might normally be unable to form stable crystal contacts.

In the same way that additional experimentally determined structures can improve the accuracy of modeling algorithms, biochemical probing and experimentally determined folding pathways can also help by imposing further restraints on tertiary folding predictions[50]. SHAPE reactivities, for example, can help identify flexible regions like loops and paired regions, but RNA constructs that populate multiple conformations can confound the inference of structural information from SHAPE reactivates[69]. Structural constraints derived from SHAPE reactivities combined with thermodynamic and probabilistic folding algorithms can help identify and differentiate alternate conformations of the same RNA. These combinatorial approaches to RNA folding can help escape the flaws of each method in isolation filling in gaps with information from multiple sources[40,50]. Techniques such as smFRET and cryo-EM can also make important contributions to charactering the structure of RNAs with multiple conformations, and these techniques are often combined with computational modeling to determine multi-domain RNA structures with complex junctions[40]. Therefore, an interdisciplinary approach to developing a deeper, more comprehensive understanding of RNA folds is required to advance RNA structural prediction, which is a crucial step in developing therapeutics for functional RNAs in pathogens.

1.2 RNAs as potential drug targets

As RNA structural studies advance, RNA elements have become increasingly promising new drug targets[47]. However, the biochemical properties of RNAs present numerous challenges to specific targeting with small molecules. While RNAs can have site specific binding pockets, as demonstrated by riboswitches and aptamers, the properties that a ligand must have to form the stable and specific interactions with a reputative set of four motifs is not yet known[47]. The hydrophilic nature of RNA nucleotides limits the availability of hydrophobic binding pockets in comparison to proteins, instead drugs must be designed to use a combination of stacking

interactions and hydrogen bond networks to create site specific docking[47, 51]. Moreover, functional RNAs are present at much lower levels in the cell than proteins, in part, due to their short half-lives. Structured RNAs often serve as scaffolds for complexes or as single sites for recognition along an mRNA, only a hand full of a given RNA element may be present in each cell[47]. Therefore, ligands which target them would require a very high affinity to compensate for the low levels of these structured RNAs present in the cell[47]. Nevertheless, technological advances have provided solutions to many of these challenges developing highly specific, cell permeable compounds that use the nature of RNA's pockets to their advantage[47, 51].

The discovery and study of riboswitches revealed the capacity for structured RNAs to bind small molecules, often aromatic compounds, ions and nucleotides with high specificity[7, 3]. From there small molecule screens were developed and tested against well characterized viral RNAs [30, 11]. Small molecules have now been discovered to bind to functional viral RNAs in SARS CoV, HIV, and HCV, albeit with limited specificity[17]. As of 2018 more than 30 compounds have been identified that can bind and modulate the function of structured mammalian RNAs[10]. These discoveres present opportunities for deeper investigation of these RNAs structures and development of therapeutics for health conditions ranging from neurological disorders to cancers[10].The chemical properties RNA targeting ligands need to possess are still being evaluated, but evidence suggests that this avenue of therapeutics is viable and wrought with potential[31]. A more comprehensive understanding of RNA structure, dynamics, and function is a crucial component to bring this medical and biochemical technology to fruition.

1.3 An introduction to RNA structural motifs with examples of their roles in viruses

Many viruses use RNA as their genetic material due to its dual role as an mRNA and as a functional macromolecule capable of binding other macromolecules and regulating their

processes[33]. RNA viruses can be broken down into four categories: Double-stranded, negative-strand, ambisense and positive-strand viruses. While all four employ structured RNAs, positive sense RNA viruses provide examples closest to the work presented here. RNA elements of positive sense RNA viruses often serve multiple roles regulating non-canonical translation, regulating genomic replication and aiding in genomic packaging [20, 32]. Sometimes one RNA element serves a combination of these functions through dynamic switching between conformations[34].

Positive sense RNA viral genomes are divided into three regions, a 5' untranslated region (5' UTR), the open reading frames (ORFs), and a 3' untranslated region (3' UTR)[33]. Structured RNA elements can be found in all three genomic regions. The sub-genomic RNAs, which are transcribed from the genomic RNA, also utilize the structured RNA elements encoded there in [56, 20,33]. Structured RNAs in the 5' and 3' UTR often regulate translation and replication [20]. These RNA elements can circularize the RNA, recruit host translation and replication factors, and ablate host immune responses [20]. Structured RNAs found within ORFs can regulate translation by modulating the rate of translocation of the translating ribosome or by recruiting and regulating various macromolecules[5]. Frameshifting elements are some of the most common RNA elements found within viral ORFs [9]. These RNA elements switch the translating ribosome to read an alternate ORF nested in a second reading frame contained within the same genomic space as the original genes[9]. Not only does this save genomic space but it also allows for production of protein isoforms and regulates the stoichiometry of the proteins produced[9].

While the structured RNA elements viruses employ can be massive and complex multi domain structures, the functional capacity and importance of a viral RNA element does not depend on the complexity of the RNA's fold[20]. Even the simplest structural motifs play crucial roles in viral biology. RNA secondary structure can be categorized into single stranded regions and helical

or paired regions made up of canonical Watson Crick base pairs[16]. These components can combine to form structural elements like the dinucleotide platform, base triples, or A-minor interactions or motifs like hairpin loops, kink turns or T loops. Structural motifs are combined to create tertiary and quaternary structures like pseudoknots[16].

The nature of ribonucleotides makes the RNA folding problem notably different than the protein folding problem. First, ribonucleotides have more rotatable backbones and more hydrogen-bonding groups than amino acids[72]. Secondly, bases are connected to flexible pentose rings and can twist about their glycosidic bond[72]. The bases can also form many different kinds of specific interactions using their Watson crick face, Hoogsteen face or sugar edge face as well as form stable pi-pi stacking interactions between their aromatic rings[49]. The flexible sugar phosphate backbone can form hydrogen bond interactions through the ribose 2'OH or the phosphate oxygens[49]. The ribose itself is also highly flexible but mainly occupies two pucker conformations, while the 2' OH group generally favors the 2' endo conformation certain backbone orientations can twist the ribose into the 3' exo conformation as well[54]. The position of the 2'OH group not only determines what it can form hydrogen bonds with, but also alters its reactivity making it a good representative of backbone positions when chemically probing an RNA structure[54]. Together this creates a huge diversity of potential orientations of the backbone and bases in dense structural cores, selective pockets, and solvent exposed surfaces, which can serve as binding motifs. Common RNA structural motifs have been identified and classified through analysis of experimentally determined RNA structures. Understanding the structure and diverse functions of these motifs builds into a more comprehensive understanding of how viruses use RNA elements to perform essential tasks. This enables both the development of better pharmaceutical interventions but also synthetic RNAs for use in chemical and molecular biology research.

1.3.1 Stem-loop structures in viral RNAs

Stem-loop structures are one of the most fundamental and versatile structural motifs in structured RNAs. A stem-loop, also known as a hairpin loop, is simple A form helix capped by a single stranded turn that connects the ascending strand to the descending strand[16]. Stem-loops are a fundamental motif that can build into larger interactions either within an RNA or between other macromolecules thereby performing essential biochemical functions for viral propagation. Despite

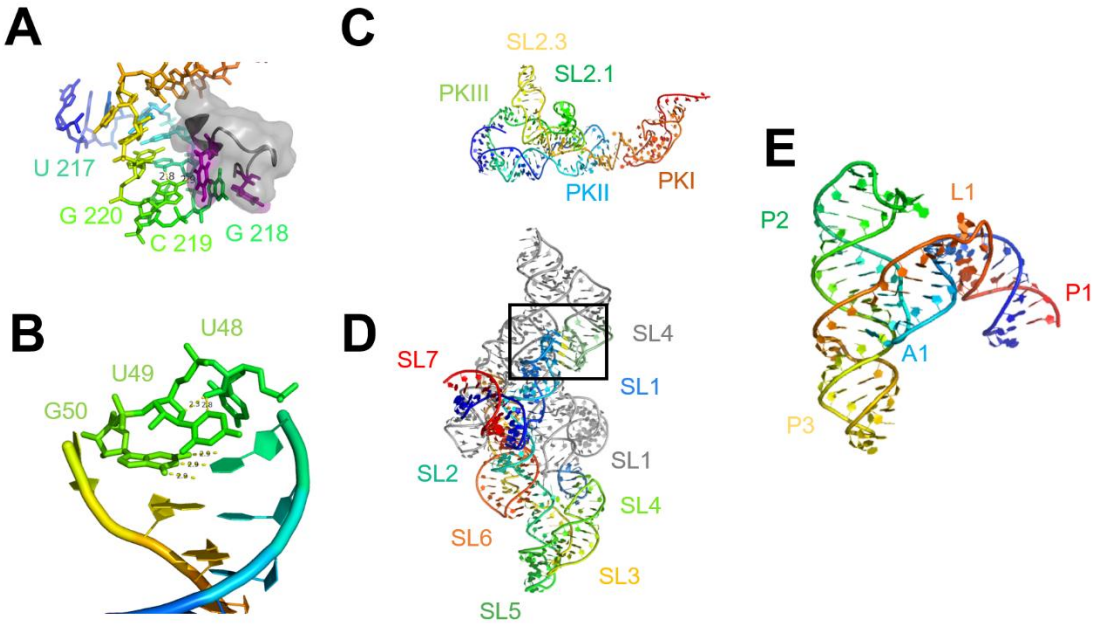


Figure 1.2 Viral RNA elements of common structural motifs A) PDB code 2IHX Ronus Sarcoma Virus UGCG tetraloop of stem-loop C of the Mψ packaging signal bound to the nucleocapsid protein G218 is sandwiched by two tyrosines[48]. B) U-turn contained in the SARS-CoV-1 SL 2 in the 5' UTR tetraloop[21] C) Cricket Paralysis Virus IRES, pseudoknots PKI PKII and PKIII where PK III is nested within PKI and interactions between L1 and S2 position nearby SL2.3 and SL 2.1 to poke out into solvent for protien binding[35]. D) kissling loop interaction between SL1 and SL4 between two copies of the VS ribozyme, positioning near by helices to create the VS ribozyme active site intermolecular

their apparent structural simplicity, stem loops can often perform surprisingly complex functions even when isolated from other structural RNA motifs[12]. Tetraloops are the most common hairpin

loop structure observed thus far. Tetraloops can be divided into several classes such as GNRA, UNCG, GUYG, GANC, (A/U)GNN and UUUM, with N standing for any nucleotide, R standing for G or A, Y standing for U or C and M standing for A or C [12,14]. Tetraloops form more stable compact structures than both potentially smaller or larger loops due to their ability to create π stacking interactions and hydrogen bond networks within the loop [16]. Tetraloop classes such as the UNCG, GNRA and CUUG families are the most stable. These classes form pairs between positions 1 and 4 often stack residues 2 or 3 and further restrain flexibility through 2' OH hydrogen bonding[12,14]. As a result, GNRA and UCG tetraloops are the most common class of tetra loop obsered in RNA crystal structures. Their specific and stable folds also make tetraloops sites of RNA-RNA interactions and protein recognition[48]. Many viruses contain lone tetraloop hairpins throughout their 5' or 3' UTRs[14, 48]. In these cases, tetraloops often recruit essential proteins or form critical regulatory long-distance pairing with other structure elements[48,32].

Despite their small size compared to other structural motifs, these hairpin structures often perform multiple functions in viral replication. For example, in Hepatitis C Virus and other Flaviviridae family viruses a conserved GCAC motif contained in the stem of a stem-loop IV of the HCV IRES serves as the recognition site for human La protein[24]. When this hairpin binds La it induces a conformational change in La to facilitate recruitment of NS5B or NS3 and the interplay between the two mediates the translation to replication switch event[24].

Due to its stability, numerous NMR structures have been solved of the UNCG tetraloop from viral RNA elements, often bound to various protein partners[14]. The UUCG tetraloop is a bi-loop, having a bifurcated hydrogen bond between the first U and the fourth G which adopts a syn conformation about the glycosidic bond thereby stabilizing the tetraloop (Figure 1.2 A)[48]. In the Rous Sarcoma Virus, the nucleocapsid protein binds to a UGCG tetraloop of stem-loop C of the

Mψ packaging signal in the 5' UTR (Figure 1.2 A) [57]. In this interaction the second G is sandwiched by two tyrosines (Figure 1.2 A). When this stem-loop is mutated to other common tetraloop sequences the interaction is lost and the virus loses considerable infectivity [57].

Stem-loops are present in each of the higher order structures discussed going forward. There are also many kinds of stem-loops beyond the tetraloop and while they are generally less stable, they can still be essential elements of the RNA's structure and function. The dynamic nature of loops beyond the tetraloop classes make them more challenging to observe using crystallography and their absence from our structural data bases can contribute to gaps in our understanding of RNA structure.

1.3.2 Single Stranded Regions as structured and functional turns in Viral RNAs

Loops often create turn motifs, the most common of these being the U-turn, which is a sharp bend in the RNA backbone between the first and second residues of a UNR consensus sequence (where N is any nucleotide and R is G or A)[21]. The U residue of the U turn creates two hydrogen bonds with the N and the R of the UNR sequence, stabilizing the sharp turn[21]. GNRA tetraloops can form U turns with the G and A residues paired and hydrogen bonds connecting the G to the N and the R (Figure 1.2 B)[21]. There are also other turns such as Z-turns, which are a variation of a U-turn, and S-turns with two consecutive bends in the phosphate backbone[16]. Internal loops create kink turns, reverse kink turns and T-loops each of these motifs allows a single stranded region of RNA to create a stable structure that can then form protein or RNA-RNA interactions [16]. For example, a stem–loop in coronavirus' 5′ untranslated region adopts a U-turn which plays a functional role in replication[28]. The loop of the SL3 region of the 5' UTR of coronaviruses contains the sequence CUUG, where the U's form the hydrogen bonding pattern of a U turn as detected by NMR (Figure 1.2 B)[28]. Mutations that disrupt the U-turn structure interfere with viral translation

and replication while mutations that maintain the predicted hydrogen bonds do not alter viral propagation[28]. Careful identification and categorization of the diversity of turn structures is an integral part to making accurate RNA structural predictions.

1.3.3 The roles and organization of pseudoknots in viruses

A pseudoknot is a structure formed though base pairing of a single stranded region of an RNA hairpin loop to another distal region of the RNA[5]. By default, all pseudoknots are defined as H-type with additional levels of complexity building on top of this fundamental class. An H type pseudoknot has at least two base paired helices were referred to as stems S1 and S2 connected by two or three single stranded loops[5]. In the majority of H type pseudoknots S1 and S2 are coaxially stacked with loop 1 tucking into the major grove of S2 and loop 3 tucking into the minor groove of S1[5]. Pseudoknots can also form between single stranded bulges or interior or multi branched loops[5]. These kinds of higher order pseudoknots can be classified into six groups based on 200 representative structures[34]. The LL type contains an additional hairpin between S1 and S2, which inserts into the apical loop of the pseudoknot[34]. The HL-in type is similar to the LL type in that is has an insertion into the apical loop, but in this case the additional hairpin forms as a bulge interceding S1[34]. In the HL-out type the additional hairpin forms within the apical loop preceding the long-distance pseudoknot interactions[34]. The HH type contains an additional hairpin interceding the 5' end of stem 1 without creating a separate structural element downstream of the apical loop[34]. If two hairpin loops form interactions between one another this is classified as an HHH type pseudoknot, also known as a kissing loop interaction[34].

While pseudoknots have topological constraints and must be able to form complete stems connected by loops, there is no conserved general sequence for a pseudoknot, making them difficult to detect by phylogenetic sequence analysis[55]. A pseudoknot's function depends entirely

on their tertiary structure which can often be maintained by a variety of conserved pairing patterns. While pseudoknots are difficult to predict, computational folding algorithms are improving the accuracy of pseudoknot prediction[55]. Atomistic molding methods like RNA As a Graph are successful at predicting pseudoknots because it employs topological constraints in addition to energy minimization[70].

In viral RNAs pseudoknots enable RNA elements to fold into a highly diverse set of tertiary structures, which conduct a large array of functions[5]. Pseudoknots are common in IRES elements, often creating highly organized structural cores that position helices to form recognition motifs for the ribosomal subunits[5]. For example, the cricket paralysis virus IRES contains three H type pseudoknots called PKI, PKII and PKIII, making up the three structural domains of the IRES. PK III is nested within PKI and forms interactions with PKII, creating a dense, complex fold (Figure 1.2 C)[41, 35]. Cryo-EM combined with x-ray crystallography revealed that this highly knotted fold ultimately projects two stem-loops that form crucial interactions with rpS5 [41, 35]. These helices are, in part, positioned by loop helix interactions in PK III, where L1 interacts with S2's major groove to position the solvent exposed stem-loops SL2.1 and SL2.3 for rPS5 binding[41,35].

Pseudoknots can also partake in autoregulation mechanisms; because these structures are dynamic and conditional, they can act as switches or as semi-passable barriers during translation[9]. Frameshifting pseudoknots are among the most well studied pseudoknot structures[19,46]. However, the mechanism of -1 frameshifting is unknown despite its ubiquity in viral genomes[15]. These pseudoknots form in the middle of open reading frames and block ribosome translocation along the mRNA[9]. The stall and the tension this blockage creates in the codon recognition sites of the ribosome then causes the tRNAs in the A and P sites to slip backwards by one position, altering the reading frame of the transcript[9]. The -1 frameshifting pseudoknot of the mouse mammary

tumor virus was solved by NMR[43]. This 34 nucleotide-long RNA element forms two GC-rich stems connected by two compact loops[43]. An intercalated adenine between the two helices at their junction interferes with helical stacking and induces a 60-degree bend[43]. This bend is further stabilized by a short two nucleotide loop connecting the junction to the base of stem 2[43]. These structural features enable the pseudoknot to bend at the intercalated hinge point[43]. This dynamic nature of the pseudoknot is a critical feature required to induce frameshifting[19,43].

1.3.4 Kissing loop interactions

A kissing loop is formed when two complementary single stranded loops pair with one another. This can also be considered an HHH-type pseudoknot[1]. Kissing loop interactions can form intra or intermolecularly. Kissing loop interactions often facilitate long distance interactions in structured RNAs and play a major role in stabilizing folds due to the strong energetic contributions of base pairing[32]. Kissing loop interactions can also contribute to forming catalytic sites as in the case of the Varkud Satellite Ribozyme[4,73]. In this case a SLV forms a kissing loop interaction with SLI, which alters the conformation of SLI to form the cleavage site (Figure 1.2 D)[4,73].

The T shaped structure (TSS) element of Pea Enation Mosaic Virus 2 (PEMV2) also forms a long-distance kissing loop interaction with hairpins located within a viral ORF and at the 5' UTR[13]. The TSS element adopts a three stemmed T shaped structure. Within the TSS, the 5' hairpin contains highly flexible loop residues that protrude into solution[13]. These residues base pair with a complementary stem-loops found both in the 5' UTR and within an ORF thereby circularizing the RNA for translation of each downstream region[13]. Although this interaction only consists of four base pairs it is sufficient to circularize the viral genome and recruit the 80S or 60S ribosomal subunit to the translation start site.

1.3.5 Three- and four-way junctions

Junction loops are defined as the single stranded regions connecting three or more continuous paired regions[16]. The orientation of the paired regions is in part determined either by interactions between the junctions and the paired regions' major or minor grooves or through packing between the paired regions themselves[16]. When there are three paired regions a three-way junction is formed. These junctions can be further classified as A-type, B-type, or C-type depending on the length of the junction regions between the paired regions and the orientation of the paired regions relative to each other[26]. The paired regions of a three-way junction are denoted P1, P2 and P3 with P1 and P2 stacked on top of one another[26]. A-type three-way junctions tend to have shorter junctions with P3 oriented in any position but most often perpendicularly to the P1, P2 stack[26]. B-type junctions form a lambda shape with the P3 region associating with the P1 region[26]. C-type junctions form a Y shape with P3 associating with P2[26].

When there are four paired regions in a junction, the resulting four-way junction can be divided into nine families[25]. Four-way junctions with two sets of stacked helices fall into the H, cH and cL, families. If there is only one coaxial stack of helices then they are either cK or π[25]. Junctions with no stacking belong in cW, ψ, X or cX families[25]. The subdivisions are determined by the parallel or antiparallel orientation of the helices relative to the continuous strands[25].

The HAV viral IRES is believed to use its domain V to recruit eIF4E to the IRES[22]. This domain forms a C-type three-way junction with P2 and P3 coaxially stacked and P1 protruding perpendicularly from that axis (Figure 1.2 E)[22]. The junction between P1 and P2, J12, consists of an A-rich bulge which forms a lonepair triloop (LPTL) structure[22]. LPTL motifs contain one base-pair followed by a three-nucleotide long hairpin[22]. LPTL motifs have been observed in other type II picornaviral IRES domain V regions[22]. The LPTL structure forms interactions with the minor

grove of P3 creating two A minor interactions in two base triples[22]. These interactions likely orient the P3 relative to P1[22]. The structural similarity of this three-way junction to other domain V structures from this class of IRESes implies that these motifs use this structure to bind eIF4E, although, currently no structure of an IRES bound to eIF4E yet exists, so the nature of this interaction remains unknown[22].

1.4 Tying it back together: Knowledge gap in RNA structural biology due to a lack of experimentally determined high resolution RNA structures

RNAs are highly multifunctional molecules, their flexible monomeric units enable the polymers they form to adopt a wide variety of structural elements, which combine to form unique motifs. Viruses in particular make excellent use of this potential for diverse three-dimensional arrangements. Over millions of years viruses have evolved RNAs to hijack host cell targets through the use of highly adaptable and varied structured RNA elements. We are faced with a combinatorial problem when predicting a RNA's fold via an algorithm. Our technologies are only just beginning to catch up with evolution, as our understanding of RNA structure has improved enough to allow for initial pharmaceutical interventions against structured RNA targets. A way forward in this frontier is to simultaneously advance each of the fields that build into the interdisciplinary field of RNA structural biology. RNA crystallography is among the more time-consuming disciplines in this modern rush, which leaves the reference data sets crystallography generates to guide molecular modeling programs precariously small. The examples of viral RNA structures presented here are only a tiny subset of the diversity of structures actually at play in nature. RNA functions in large part due to its flexibility and existing structure determining methods are hampered by the structural inconsistency and heterogeneity that are often intrinsic to RNA function. Technologies like chaperone assisted crystallography may provide one way around the problem this inherent

flexibility presents to crystallography. Additionally, by generating raw structures more rapidly, advances in RNA crystallography are poised to improve the computational modeling of structure and dynamics.

Chaperone-assisted RNA crystallography is a tool we are developing and refining into a broadly applicable strategy to improve RNA crystallization. In this work I present two new viral RNA structures solved to 2.09Å and 2.75Å resolution using the grafting approach to chaperone assisted RNA crystallography. While these structures only represent two single RNA's each has revealed important new information about the potential mechanisms these viral RNA's use to regulate critical functions in viral replication. One structure is the first to identify a new conformation of the SARS-CoV-2 programmed ribosomal frameshifting element, which is known to operate though conformational switching. This structure presents a new opportunity to model the mechanism of -1 ribosomal frameshifting in SARS-CoV-2. The other characterizes a new fold and set of motifs that may be commonly used by cap intendent translation enhancers found broadly throughout plant viruses. This work demonstrates the potential chaperone assisted crystallography presents to studying viral RNA structural biology.

1.5 Works Cited in Chapter 1

1) Andersen, A. A., & Collins, R. A. (2001). Intramolecular secondary structure rearrangement by the kissing interaction of the Neurospora VS ribozyme. Proceedings of the National Academy of Sciences, 98(14), 7730–7735. https://doi.org/10.1073/PNAS.141039198

2) Bisaria, N., Greenfeld, M., Limouse, C., Mabuchi, H., & Herschlag, D. (2017). Quantitative tests of a reconstitution model for RNA folding thermodynamics and kinetics. Proc Natl Acad Sci U S A, 114, E7688–E7696. https://doi.org/https://doi.org/10.1073/pnas.1703507114

3) Blount, K. F., & Breaker, R. R. (2006). Riboswitches as antibacterial drug targets. Nature Biotechnology, 24(12), 1558–1564. https://doi.org/10.1038/nbt1268

4) Bonneau, E., Girard, N., Lemieux, S., & Legault, P. (2015). The NMR structure of the II-III-VI three-way junction from the Neurospora VS ribozyme reveals a critical tertiary interaction and provides new insights into the global ribozyme structure. Rna, 21(9), 1621–1632. https://doi.org/10.1261/rna.052076.115

5) Brierley, I., Pennell, S., & Gilbert, R. J. C. (2007). Viral RNA pseudoknots: versatile motifs in gene expression and replication. Nature Reviews Microbiology 2007 5:8, 5(8), 598–610. https://doi.org/10.1038/nrmicro1704

6) D'Ascenzo, L., Leonarski, F., Vicens, Q., & Auffinger, P. (2017). Revisiting GNRA and UNCG folds: U-turns versus Z-turns in RNA hairpin loops. RNA, 23(3), 259–269. https://doi.org/10.1261/RNA.059097.116

7) Deigan, K. E., & FerrÉ-D'AmarÉ, A. R. (2011). Riboswitches: Discovery of Drugs That Target Bacterial Gene-Regulatory RNAs. Accounts of Chemical Research, 44(12), 1329–1338. https://doi.org/10.1021/ar200039b

8) Dersch, P., Khan, M. A., Mühlen, S., & Görke, B. (2017). Roles of regulatory RNAs for antibiotic resistance in bacteria and their potential value as novel drug targets. Frontiers in Microbiology, 8(MAY), 1–12. https://doi.org/10.3389/fmicb.2017.00803

9) Dinman, J. D. (2012). Mechanisms and implications of programmed translational frameshifting. Wiley Interdisciplinary Reviews: RNA. https://doi.org/10.1002/wrna.1126

10) Donlic, A., & Hargrove, A. E. (2018). Targeting RNA in mammalian systems with small molecules. Wiley Interdisciplinary Reviews: RNA, 9(4). https://doi.org/10.1002/wrna.1477

11) Fernandes, J., Jayaraman, B., & Frankel, A. (2012). The HIV-1 Rev response element. RNA Biology, 9(1), 6–11. https://doi.org/10.4161/rna.9.1.18178

12) Fiore, J. L., & Nesbitt, D. J. (2013). An RNA folding motif: GNRA tetraloop–receptor interactions. Quarterly Reviews of Biophysics, 46(3), 223–264. https://doi.org/10.1017/S0033583513000048

13) Gao, F., Gulay, S. P., Kasprzak, W., Dinman, J. D., Shapiro, B. A., & Simon, A. E. (2013). The Kissing-Loop T-Shaped Structure Translational Enhancer of Pea Enation Mosaic Virus Can Bind Simultaneously to Ribosomes and a 5' Proximal Hairpin. Journal of Virology, 87(22), 11987–12002. https://doi.org/10.1128/jvi.02005-13

14) Geary, C., Baudrey, S., & Jaeger, L. (2008). Comprehensive features of natural and in vitro selected GNRA tetraloop-binding receptors. Nucleic Acids Research, 36(4), 1138–1152. https://doi.org/10.1093/NAR/GKM1048

15) Halma, M. T. J., Ritchie, D. B., & Woodside, M. T. (2021). Conformational Shannon Entropy of mRNA Structures from Force Spectroscopy Measurements Predicts the Efficiency of -1 Programmed Ribosomal Frameshift Stimulation. Physical Review Letters, 126(3), 38102. https://doi.org/10.1103/PhysRevLett.126.038102

16) Hendrix, D. K., Brenner, S. E., & Holbrook, S. R. (2005). RNA structural motifs: Building blocks of a modular biomolecule. Quarterly Reviews of Biophysics, 38(3), 221–243. https://doi.org/10.1017/S0033583506004215

17) Hermann, T. (2016). Small molecules targeting viral RNA. Wiley Interdisciplinary Reviews: RNA, 7(6), 726–743. https://doi.org/https://doi.org/10.1002/wrna.1373

18) Hofacker, I. L., & Lorenx, R. (2013). Predicting RNA structure: Advances and Limitations. In RNa folding. Retrieved from https://link-springer-com.proxy.uchicago.edu/protocol/10.1007%2F978-1-62703-667-2_1

19) Houck-Loomis, B., Durney, M. A., Salguero, C., Shankar, N., Nagle, J. M., Goff, S. P., & D'Souza, V. M. (2011). An equilibrium-dependent retroviral mRNA switch regulates translational recoding. Nature, 480(7378), 561–564. https://doi.org/10.1038/nature10657

20) Jaafar, Z. A., & Kieft, J. S. (2019). Viral RNA structure- based strategies to manipulate translation. Nature Reviews Microbiology, 17(february), 110–123. https://doi.org/10.1038/s41579-018-0117-x

21) Jucker, F. M., & Pardi, A. (1996). GNRA tetraloops make a U-turn. RNA, 1, 219–222.

22) Koirala, D., Shao, Y., Koldobskaya, Y., Fuller, J. R., Watkins, A. M., Shelke, S. A., … Piccirilli, J. A. (2019a). A conserved RNA structural motif for organizing topology within picornaviral internal ribosome entry sites. Nature Communications, 10(1), 3629. https://doi.org/10.1038/s41467-019-11585-z

23) Koldobskaya, Y., Duguid, E. M., Shechner, D. M., Suslov, N. B., Ye, J., Sidhu, S. S., … Piccirilli, J. A. (2011). A portable RNA sequence whose recognition by a synthetic antibody facilitates structural determination. Nature Structural and Molecular Biology, 18(1), 100–107. https://doi.org/10.1038/nsmb.1945https://doi.org/10.1038/s41467-019-11585-z

24) Kumar, A., Ray, U., & Das, S. (2013). Human La Protein Interaction with GCAC near the Initiator AUG Enhances Hepatitis C Virus RNA Replication by Promoting Linkage. Journal of Virology, 87(21), 6713–6726. https://doi.org/10.1128/JVI.00525-13

25) Laing, C., & Schlick, T. (2009). Analysis of Four-Way Junctions in RNA Structures. Journal of Molecular Biology, 390(3), 547–559. https://doi.org/10.1016/J.JMB.2009.04.084

26) Lescoute, A., & Westhof, E. (2006). The interaction networks of structured RNAs. Nucleic Acids Research, 34(22), 6587. https://doi.org/10.1093/NAR/GKL963

27) Li, B., Cao, Y., Westhof, E., & Miao, Z. (2020). Advances in RNA 3D Structure Modeling Using Experimental Data, 11(October), 1–19. https://doi.org/10.3389/fgene.2020.574485

28) Liu, P., Li, L., Millership, J. J., Kang, H., Leibowitz, J. L., & Giedroc, D. P. (2007). A U-turn motif-containing stem – loop in the coronavirus 5 9 untranslated region plays a functional role in replication. RNA, 13, 763–780. https://doi.org/10.1261/rna.261807.5

29) Manfredonia, I., Nithin, C., Ponce-Salvatierra, A., Ghosh, P., Wirecki, T. K., Marinus, T., … Incarnato, D. (2020). Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Research, 48(22), 12436–12452. https://doi.org/10.1093/nar/gkaa1053

30) Mei, H.-Y., Cui, M., Heldsinger, A., Lemrow, S. M., Loo, J. A., Sannes-Lowery, K. A., … Czarnik, A. W. (1998). Inhibitors of Protein−RNA Complexation That Target the RNA: Specific Recognition of Human Immunodeficiency Virus Type 1 TAR RNA by Small Organic Molecules. Biochemistry, 37(40), 14204–14212. https://doi.org/10.1021/bi981308u

31) Morgan, B. S., Forte, J. E., Culver, R. N., Zhang, D. Y., & Hargrove, A. E. (2017). Discovery of Key Physicochemical, Structural, and Spatial Properties of RNA-Targeted

Bioactive Ligands. Wiley Interdisciplinary Reviews: RNA, 56(43), 13498–13502. https://doi.org/https://doi.org/10.1002/anie.201707641

32) Nicholson, B. L., & White, K. A. (2014). Functional long-range RNA-RNA interactions in positive-strand RNA viruses. Nature Reviews Microbiology, 12(7), 493–504. https://doi.org/10.1038/nrmicro3288

33) Payne, S. (2017). Introduction to RNA viruses. In Viruses (pp. 97–105). https://doi.org/10.1016/B978-0-12-803109-4.00010-6

34) Peselis, A., & Serganov, A. (2014). Structure and function of pseudoknots involved in gene expression control. Wiley Interdisciplinary Reviews. RNA, 5(6), 803–822. https://doi.org/10.1002/wrna.1247

35) Pfingsten, J. S., Costantino, D. A., & Kieft, J. S. (2006). Structural basis for ribosome recruitment and manipulation by a viral IRES RNA. Science (New York, N.Y.), 314(5804), 1450–1454. https://doi.org/10.1126/SCIENCE.1133281

36) Phan, A., Mailey, K., Saeki, J., Gu, X., & Schroeder, S. J. (2017). Advancing viral RNA structure prediction: measuring the thermodynamics of pyrimidine-rich internal loops. RNA, 23(5), 770–781. https://doi.org/10.1261/rna.059865.116

37) Rogers, E., & Heitsch, C. E. (2014). Profiling small RNA reveals multimodal substructural signals in a bolzmann ensemble. Nucleic Acids Research, 42(e171).

38) Schlick, T., Zhu, Q., Jain, S., & Yan, S. (2020). Structure-Altering Mutations of the SARS-CoV-2 Frame Shifting RNA Element.pdf. Biophysical Journal.

39) Schlick, Tamar, & Pyle, A. M. (2017). Biophysical Perspective Opportunities and Challenges in RNA Structural Modeling and Design. Biophysj, 113(2), 225–234. https://doi.org/10.1016/j.bpj.2016.12.037

40) Schroeder, S. J. (2018). Challenges and approaches to predicting RNA with multiple functional structures. RNA, 24, 1615–1624. https://doi.org/10.1261/rna.067827.118.RNA

41) Schüler, M., Connell, S. R., Lescoute, A., Giesebrecht, J., Dabrowski, M., Schroeer, B., … Spahn, C. M. T. (2006). Structure of the ribosome-bound cricket paralysis virus IRES RNA. Nature Structural & Molecular Biology, 13(12), 1092–1096. https://doi.org/10.1038/nsmb1177

42) Shao, Y., & Zhang, Q. C. (2020). Targeting RNA structures in diseases with small molecules. Essays in Biochemistry, 64(6), 955–966. https://doi.org/10.1042/EBC20200011

43) Shen, L. X., & Tinoco, I. (1995). The Structure of an RNA Pseudoknot that Causes Efficient Frameshifting in Mouse Mammary Tumor Virus. Journal of Molecular Biology, 247(5), 963–978. https://doi.org/10.1006/JMBI.1995.0193

44) Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., & Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nature Methods, 11(9), 959. https://doi.org/10.1038/NMETH.3029

45) Šponer, J., Bussi, G., Krepl, M., Banáš, P., Bottaro, S., Cunha, R. A., … Otyepka, M. (2018). RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. Chemical Reviews, 118(8), 4177–4338. https://doi.org/10.1021/acs.chemrev.7b00427

46) Staple, D. W., & Butcher, S. E. (2005). Pseudoknots : RNA Structures with Diverse Functions, 3(6), 956–959. https://doi.org/10.1371/journal.pbio.0030213

47) Sztuba-Solinska, J., Chavez-Calvillo, G., & Cline, S. E. (2019). Unveiling the druggable RNA targets and small molecule therapeutics. Bioorganic and Medicinal Chemistry. https://doi.org/10.1016/j.bmc.2019.03.057

48) Thapar, R., Denmon, A. P., & Nikonowicz, E. P. (2014). Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins. Wiley Interdisciplinary Reviews: RNA, 5(1), 49–67. https://doi.org/10.1002/wrna.1196

49) Ulyanov, N. B., & James, T. L. (2010). RNA Structural Motifs That Entail Hydrogen Bonds Involving Sugar-Phosphate Backbone Atoms of RNA. New Journal of Chemistry, 34(5), 910–917. https://doi.org/10.1039/b9nj00754g

50) Wang, J., Williams, B., Chirasani, V. R., Krokhotin, A., Das, R., & Dokholyan, N. V. (2019). Limits in accuracy and a strategy of RNA structure prediction using experimental information. Nucleic Acids Research. https://doi.org/10.1093/nar/gkz427

51) Warner, K. D.; Hajdin, C. E.; Weeks, K. M. (2018). Principles for Targeting RNA with Drug-like Small Molecules. Nat. Rev. Drug Discov., 17(8), 547–558.

52) Weinreb, C., Riesselman, A. J., Ingraham, J. B., Gross, T., Sander, C., & Marks, D. S. (2016). 3D RNA and Functional Interactions from Evolutionary Couplings. Cell, 165(4), 963–975. https://doi.org/10.1016/j.cell.2016.03.030

53) Westhof, E., & Lescoute, A. (2006). Topology of three-way junctions in folded RNAs. RNA Society, 12, 83–93. https://doi.org/10.1261/rna.2208106.In

54) White, N. A., Sumita, M., Marquez, V. E., & Hoogstraten, C. G. (2018). Coupling between conformational dynamics and catalytic function at the active site of the lead-dependent ribozyme. RNA (New York, N.Y.), 24(11), 1542–1554. https://doi.org/10.1261/rna.067579.118

55) Xu, X., & Chen, S.-J. (2020). Topological constraints of RNA pseudoknotted and loop-kissing motifs: applications to three-dimensional structure prediction. Nucleic Acids Research, 48(12), 6503. https://doi.org/10.1093/NAR/GKAA463

56) Yang, J., Xia, H., Qian, Q., & Zhou, X. (2015). RNA chaperones encoded by RNA viruses. Virologica Sinica, 30(6), 401–409. https://doi.org/10.1007/s12250-015-3676-2

57) Zhou, J., Bean, R. L., Vogt, V. M., & Summers, M. (2007). Solution Structure of the Rous Sarcoma Virus Nucleocapsid Protein: μΨ RNA Packaging Signal Complex. Journal of Molecular Biology, 365(2), 453–467. https://doi.org/10.1016/j.jmb.2006.10.013

58) Zuber, J., Sun, H., Zhang, X., McFadyen, I., & Mathews, D. H. (2017). A sensitivity analysis of RNA folding nearest neighbor parameters identifies a subset of free energy parameters with the greatest impact on RNA secondary structure prediction. Nucleic Acids Research, 45(10), 6168–6176. https://doi.org/10.1093/nar/gkx170

59) Felden, B. (2007). RNA structure: experimental analysis. Current Opinion in Microbiology, 10(3), 286–291. https://doi.org/10.1016/J.MIB.2007.05.001

60) Murray, L. J. W., Arendall, W. B., Richardson, D. C., & Richardson, J. S. (2003). RNA backbone is rotameric. Proceedings of the National Academy of Sciences, 100(24), 13904–13909. https://doi.org/10.1073/PNAS.1835769100

61) Murray, L. J. W., Arendall, W. B., Richardson, D. C., & Richardson, J. S. (2003). RNA backbone is rotameric. Proceedings of the National Academy of Sciences, 100(24), 13904–13909. https://doi.org/10.1073/PNAS.1835769100

62) Ye, J. D., Tereshko, V., Frederiksen, J. K., Koide, A., Fellouse, F. A., Sidhu, S. S., … Piccirilli, J. A. (2008). Synthetic antibodies for specific recognition and crystallization of structured RNA. Proceedings of the National Academy of Sciences of the United States of America, 105(1), 82–87. https://doi.org/10.1073/pnas.0709082105

63) Shao, Y., Huang, H., Qin, D., Li, N. S., Koide, A., Staley, J. P., … Piccirilli, J. A. (2016). Specific Recognition of a Single-Stranded RNA Sequence by a Synthetic Antibody

Fragment. Journal of Molecular Biology, 428(20), 4100–4114. https://doi.org/10.1016/j.jmb.2016.08.029

64) Simmonds, P., Tuplin, A., & Evans, D. J. (2004). Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus-evolution and host persistence. Rna, 10(9), 1337–1351. https://doi.org/10.1261/rna.7640104

65) Rivas E, Clements J, Eddy SR. 2017. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat Methods 14: 45–48.

66) Miao Z, Adamiak RW, Antczak M, Batey RT, Becka AJ, Biesiada M, Boniecki MJ, Bujnicki JM, Chen SJ, Cheng CY, et al. 2017. RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. RNA 23: 655–672.

67) Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9: 133–148.

68) Schroeder, S. J. (2018). Challenges and approaches to predicting RNA with multiple functional structures. RNA, 24, 1615–1624. https://doi.org/10.1261/rna.067827.118.RNA

69) Wang, J., Williams, B., Chirasani, V. R., Krokhotin, A., Das, R., & Dokholyan, N. V. (2019). Limits in accuracy and a strategy of RNA structure prediction using experimental information. Nucleic Acids Research. https://doi.org/10.1093/nar/gkz427

70) St-Onge K, Thibault P, Hamel S, Major F. Modeling RNA tertiary structure motifs by graph-grammars. Nucleic Acids Res. 2007;35(5):1726-36. doi: 10.1093/nar/gkm069. Epub 2007 Feb 21. PMID: 17317683; PMCID: PMC1865062.

71) Lin L, McKerrow WH, Richards B, Phonsom C, Lawrence CE. 2018. Characterization and visualization of RNA secondary structure Boltzmann ensemble via information theory. BMC Bioinformatics 19: 82.

72) Murray, L. J. W., Arendall, W. B., Richardson, D. C., & Richardson, J. S. (2003). RNA backbone is rotameric. Proceedings of the National Academy of Sciences, 100(24), 13904–13909. https://doi.org/10.1073/PNAS.1835769100

73) Suslov NB, DasGupta S, Huang H, Fuller JR, Lilley DM, Rice PA, Piccirilli JA. Crystal structure of the Varkud satellite ribozyme. Nat Chem Biol. 2015 Nov;11(11):840-6. doi: 10.1038/nchembio.1929. Epub 2015 Sep 28. PMID: 26414446; PMCID: PMC4618023.

2. Chapter 2 The Crystal Structure of the SARS CoV 2 Programmed Frameshifting Element Solved to 2.09Å

2.1 Introduction

The historic and deadly COVID-19 pandemic is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Researchers around the world are searching for treatments for this catastrophic disease, in part by targeting drug design efforts towards the structured RNA elements in the nearly 30 kb RNA genome of the virus. However, a lack of high-resolution, three-dimensional structural information about structured regions of the genome make development of drugs to target them difficult. Computational modeling and structural probing techniques are able to identify structured regions within RNAs and can suggest whether an RNA element might contain a pocket sufficient for ligand binding but, these estimations often lack certainty about the chemical arrangement of binding pockets.[1,2] By contrast, broad screens of RNA binding chemicals do not require high resolution structural information as starting point and can yield lead molecules, but these chemicals are rarely drug-like due to their toxicity, lack of cell permeability or lack of bioavailability.[1] Experimentally derived structures of viral RNA elements can provide another route to drug discovery. Existing drugs can be screened against experimentally determined structural models using structural dynamics simulations to identify potential binders.[3]

One potentially druggable RNA target in the SARS-CoV-2 genome is the programmed -1 ribosomal frameshifting element (PFSE). Thus far it has been shown to bind the ligand 2-{[4-(2-methyl-thiazol-4ylmethyl)-[1,4]diazepane-1-carbonyl]-amino}-benzoic acid ethyl ester (MTDB) and in cell culture this ligand can compromise ribosomal frameshifting and inhibit viral replication by three orders of magnitude.[3-8] Chemical probing and homology modeling consistently predict this programmed frameshifting element to form a stable and well-ordered three stemmed H-type

pseudoknot.[8-14] For structural biologists, its small size also makes it a promising candidate for x-ray crystallography. Thus, we saw an opportunity to support the drug design process and mechanistic investigations by generating a high-resolution structure of the programmed frameshifting element using our lab's method of chaperone assisted RNA crystallography.

Ribosomal frameshifting from Orf1a to Orf1b is a critical step in coronavirus propagation.[15,18] Orf1a and its out-of-frame continuation Orf1b, are the first open reading frames to be translated directly from the SARS-CoV-2 RNA genome upon infection and disassembly (Fig. 1A). They encode the nonstructural proteins (NSPs) 1-16, which are involved in evading the host immune response, replicating the genomic RNA, and producing the sub-genomic mRNAs that encode the structural proteins. NSPs 1-10 are produced from Orf1a as the self-cleaving polypeptide pp1a.[13,16] As the ribosome approaches the stop codon of Orf1a the programmed frameshifting element pseudoknot can cause the ribosome to slip backwards by one position.[16-18] If frameshifting occurs the ribosome will continue translating into Orf1b, producing pp1ab, which comprises NSPs 1-16. Known as the golden mean hypothesis of ribosomal frameshifting, incorrect stoichiometry of early replication products, in this case pp1a and pp1ab, disrupts the replication cycle and reduces virus propagation.[4,13,19]

The basic mechanism of -1 ribosomal frameshifting is known, although there are many levels of regulation at play that are not understood.[17,20-22,46] Generally, a structured region of the RNA causes a translating ribosome to pause over a so-called slippery site with a nucleotide sequence pattern of X <u>XXY</u> <u>YYZ</u> composition.[16,18,23,24] This structured region is most often a pseudoknot, which forms 6 – 8 nts downstream of the slippery site. The pseudoknot structure opposes the translocating ribosome, which creates tension that causes the slippery site codon

interactions with P- and A-site tRNAs to slide backwards by one nucleotide from X XXY YYZ to XXX YYY Z, resulting in a -1 shift in the reading frame.[16,18,23-25]

When incorporated into luciferase reporter mRNA constructs, the minimal sequence of frameshifting pseudoknots from different viruses induce frameshifting at an internally consistent frequency, but this frequency varies widely from virus to virus.[27] For example, the programmed frameshifting element from West Nile virus induces frameshifting around 70 to 80% of the time while the SARS-CoV-1 PFSE induces frameshifting around 15 to 30% of the time.[5,27,28] Force extension curves of the different pseudoknots in an optical trap reveal that those elements populating more conformations induce frameshifting more frequently.[5,27,28] A linear correlation between the rate of frameshifting in vitro and the number of conformations a pseudoknot can adopt can be drawn using the calculated Shannon entropy of each pseudoknot.[29] The SARS-CoV-1 and SARS-CoV-2 PFSEs adopt only two conformations under tension, consistent with other pseudoknots that induce comparable rates of frameshifting.[5,28,29,50] The conformational dynamics of frameshifting pseudoknots are clearly one important component of this highly regulated mechanism. An experimentally determined tertiary structure may help to illuminate how these conformations contribute to the regulation of frameshifting.

In SARS-CoV-2, the PFSE encompasses the roughly 80-nt long sequence (residues 13462-13542 in the HB01 strain gen bank number NC_45512.2) that includes a hepta-nucleotide slippery site, nucleotides 13462 – 13468 in the genomic numbering, and a spacer region followed by the proposed RNA pseudoknot structure in positions 13474 - 13542 in the genome (Fig. 1A).[30] Interestingly, this pseudoknot sequence is nearly perfectly conserved from SARS-CoV-1, the etiological agent of the 2003 coronavirus pandemic, with the exception of a single point mutation at position C 13444 in SARS-CoV-1 (Gen bank number AY278488.2) which corresponds to an A

in position 13533 in SARS-CoV-2 [12,33,59]. The pseudoknot structure of the frameshifting element from SARS-CoV-1 has been well-characterized previously by chemical probing, mutational and NMR studies of minimal constructs.[12,14,21] The secondary structure of the SARS-CoV-1 PFSE derived from these studies (Fig. 1B) has been used to guide molecular dynamics and tertiary structure prediction models.[8-10,32] Briefly, the PFSE's secondary structure comprises Stem 1 (labeled S1 in Fig. 1) at its 5' end, which leads into Loop 1 (L1) followed by Stem 2 (S2) (Fig 1B). Stem 3 (S3) forms below Stem 1 and folds the RNA structure back on itself by forming the twelve-nucleotide long Loop 2. Stem 3 leads into Loop 3 (L3) which spans the gap between the end of Stem 3 and the start of Stem 2. Stem 2 forms the long-distance interactions in the pseudoknot, which defines it as an H-type, and ties the 5' end of the RNA to the 3' end (Fig. 1B).[12,31]

Despite the one nucleotide substitution, the SARS-CoV-2 PFSE likely adopts the same conformation as the SARS-CoV-1 PFSE. Computational modeling predicts similar structures for both sequences and NMR studies show close agreement with the predicted three stem structure in both cases.[8-10,31,33-35] Small-angle x-ray scattering diffraction analysis by Kelly et al. has shown that the SARS-CoV-1 frameshifting pseudoknot and the SARS-CoV-2 frameshifting pseudoknot have nearly identical topology.[33] Functionally, both SARS-CoV-1 and SARS-CoV-2 pseudoknots frameshift to appreciable rates in vitro and in cells.[33] Likewise, in both cases frameshifting is inhibited in the presence of the ligand MTDB, which implies both PFSEs bind the ligand.[7,33] This evidence allows us to apply the established research on the SARS-CoV-1 PFSE to this new SARS-CoV-2 PFSE in our investigation of the structure.

To contribute higher resolution information about the SARS-CoV-2 PFSE structure, we have applied the chaperone assisted RNA crystallography method to the PFSE. We created a modified construct of the PFSE lacking the upstream slippery site and spacer region. This

structured RNA was then bound to an engineered humanized murine antibody fragment to facilitate co-crystallization. Using iridium hexammine, we solved the structure using a combination of single angle anomalous diffraction phasing and molecular replacement. Here we report the structure of the SARS-CoV-2 programmed frameshifting element pseudoknot solved by x-ray diffraction to 2.09 Å.

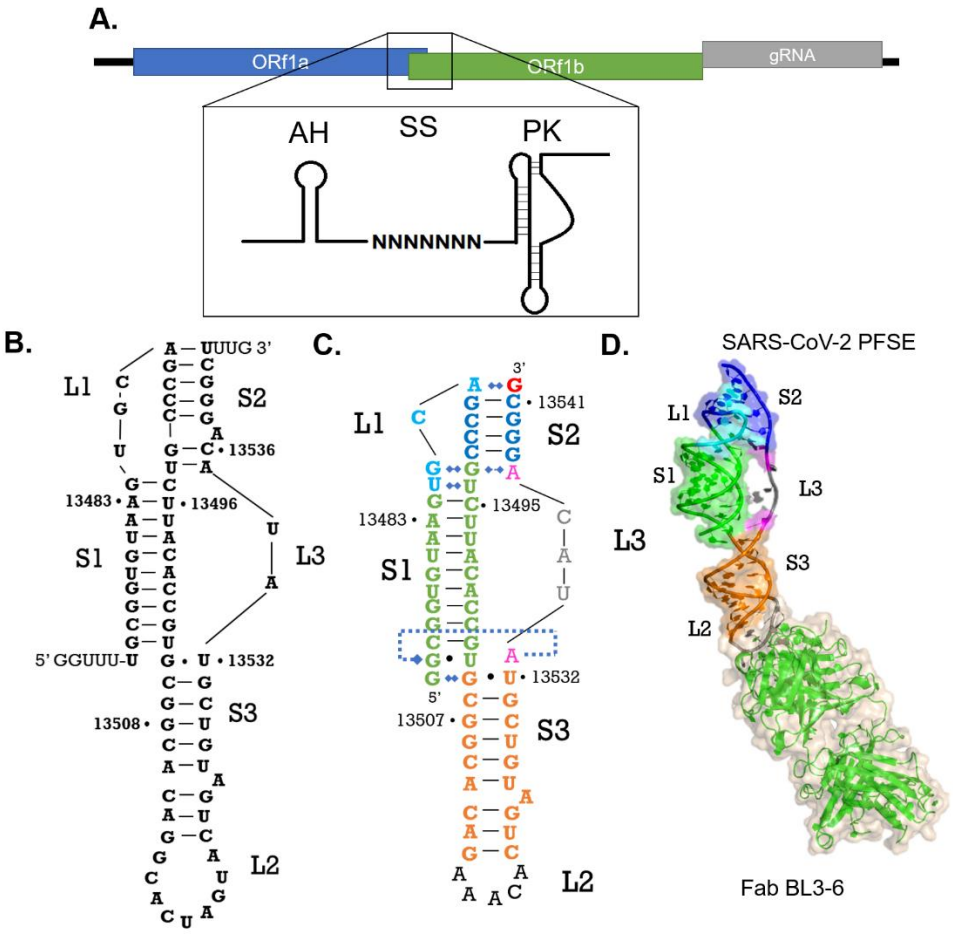


Figure 2.1 Overall structure of SARS-CoV-2 programmed -1 ribosomal frameshifting element pseudoknot. (A) Diagram of frameshifting element relative to SARS-CoV-2 genome. AH indicates the attenuator hairpin, SS indicates the slippery site and PK indicates the pseudoknot structure.[6] (B) Predicted secondary structure of the programmed frameshifting element.[6] (C) Secondary structure derived from the crystal structure; Stem 1 colored green, Loop 1 colored cyan, Stem 2 colored navy, Loop 3 colored magenta with nucleotides lacking density colored grey, Stem 3 colored orange, Loop 2 mutant pentaloop colored black, nucleotide 13542 shown in red was added via nonspecific addition in in vitro transcription reaction.

Fig. 2.1 Continued (D) Crystal structure of the SARS-CoV-2 frameshifting element bound to Fab BL3-6 through its mutated loop structure shown as cartoon and transparent surface; surface excluded for nucleotides with no electron density in Loop 3. Color scheme of crystal structure matches that of the secondary structure in (C).

2.2 Results

2.2.1 Design and characterization of crystallization construct.

The PFSE region comprises the so-called the slippery site (13462 – 13468) followed by a spacer region and finally the core pseudoknot spanning nucleotides 13474 to 13542 in the genomic numbering for strain HB01 (Fig. 1A).[12,15,30] Our crystallization construct includes only the minimal sequence for the pseudoknot (13474 - 13541). We omitted the spacer and the slippery site sequences from the crystallization construct due to the high flexibility expected for single-stranded regions at the ends of RNAs, which can reduce crystallization efficiency. In this construct the starting residue, nucleotide 13474, was mutated from U to G to enhance in vitro transcription efficiency. This residue is predicted to be unpaired and is not considered essential for pseudoknot formation. It was unclear whether the final uridine of the pseudoknot sequence (residue 13542) would be paired.[9] To avoid a dangling nucleotide in the construct, which can compromise the crystallization construct rigidity, this nucleotide was excluded from the template sequence. However, upon sequencing RNA extracted directly from the crystal, we found that the construct that crystallized had an untemplated G added to the 3' end (See below and Supplementary Figure S3). We refer to this minimal version the PFSE as the wild type PFSE construct.

RNA elements often resist forming well-diffracting crystals due to their biophysical properties such as instability, negatively charged backbone, conformational heterogeneity, and limited functional group diversity for mediating specific lattice contacts. Our lab has found that antibody fragment (Fab) RNA complexes crystallize more readily than RNA alone. In previous

work we reported developing a suite of humanized murine antibody fragments and their cognate RNA motifs, which can be grafted into structured RNA targets to create a Fab binding site.[36,37] Among these, Fab BL3-6 binds to hairpins with the loop sequence (AAACA) closed by a GC pair, which we call the BL3-6-binding epitope. The most suitable location to graft the BL3-6-binding epitope into the PFSE was Loop 2, at the base of Stem 3. The SARS-CoV-1 PFSE has been shown to retain wild type levels of frameshifting activity in-vitro when Loop 2 was replaced with a GUUG tetraloop, therefore we expected that the PFSE structure could tolerate a AAACA pentaloop in the Loop 2 position[14]. We refer to this Fab-binding PFSE construct as PFSE BL3-6. In an electrophoretic mobility shift assay (EMSA) the PFSE BL3-6 construct formed a mobility shifted species in the presence of Fab BL3-6, suggesting that the grafted sequence formed the expected loop and had not altered the predicted RNA secondary structure as shown in supplementary Figure 2.4 (Fig B.4).

2.2.2 Crystallization and structure determination.

RNA, in the complex with antibody fragment, was concentrated to 6 mg mL -1 and crystallized in 2 % v/v tacsimate pH 4; 0.1 M sodium acetate trihydrate pH 4.6; 16 % w/v polyethylene glycol 3,350 (which has a final a measured pH of 4.8) and were further optimized with the addition of either 0.01 M sarcosine or 0.01 M betaine hydrochloride. Some crystals were soaked in cryogenic protectant containing iridium hexammine to aid in phasing using anomalous dispersion. These crystals, with and without heavy metal soaking, yielded multiple data sets with the best diffracting to 2.09 Å. To solve the structure, data sets with strong anomalous signal and low resolution were phased in Phenix using a combination of single-wavelength anomalous diffraction (SAD) experimental phasing and molecular replacement using the Fab BL3-6 model.[36,38-40] An initial electron density map was generated, and a partial model of the RNA was then built-in coot.[41] This

partial RNA model and the Fab model were then used to find a molecular replacement solution for a separate high resolution native diffraction data set. This solution gave complete electron density maps. Iterative rounds of building and refinement in this high-resolution data set yielded a complete structure for the PFSE pseudoknot at 2.09Å with an R_{work} of 19.93% and an R_{free} of 23.08% (additional statistics reported in Table 1). The coordinates for this structure have been deposited in the PDB under accession code 7MLX.

Data collection	
Space group	C 2 2 21
Resolution range	67.71 - 2.09 (2.17 - 2.09)
Cell dimensions	
a, b, c (Å)	76.83, 143.24, 133.64
α, β, γ (◦)	90, 90, 90
R-merge	0.083 (0.730)
I/ σ (I)	15.98 (2.50)
$CC_{1/2}$	0.99 (0.80)
CC*	1 (0.94)
CC work/CC free	0.95(0.87)/0.88 (0.78)
Completeness (%)	99.89 (99.93)
Multiplicity	6.8 (6.7)
Refinement	
No. Unique reflections	43945 (4321)
R_{work}/ R_{free} (%)	0.1993 (0.2550)/0.2308 (0.2886)
RMS deviations	
Bond lengths (Å)	0.007
Bond angles (◦)	0.97
Average B-factor all atoms ($Å^2$)	40.78
Ramachandran plot of all protein residues	
favored (%)	97.64
allowed (%)	2.36
Number of residues	
RNA	65
Protein residues	430

Table 2.1 SARS-CoV-2PFSE X-ray crystallography data collection and refinement statistics.

2.2.3 The global structure reveals a three-stemmed pseudoknot matching previous predictions.

The SARS-CoV-2 PFSE forms a three-stemmed H-type pseudoknot structure with 3 loops, consistent with original predictions for the SARS-CoV-1 PFSE by Plant et al. 2005 and confirmed as the general secondary structure for the SARS-CoV-2 PFSE more recently.[12,14,31] In our structure (Fig. 1C and 1D) Stem 1, shown in green in all figures, begins at the 5' end and forms between nucleotides 13474 - 13484 on the 5'-side and nucleotides 13493 – 13504 on the 3'-side (Fig. 1C). Interestingly, the 5' end of Stem 1 threads through the ring created by Loop 3 (Fig. 2.4E and Fig B.2), a feature inferred from mid-resolution cryo-EM models (see discussion). [4,43] The 5' strand of Stem 1 leads into Loop 1, colored cyan in all figures, which encompasses nucleotides 13485 – 13488. Within Loop 1, nucleotides 13485 and 13486 form interactions with nucleotides 13494 and 13493 of Stem 1, respectively. The Loop 1 interactions position G13489 for pairing with the 3' end of the RNA to form Stem 2, shown in navy in all figures. Stem 2 is G-rich and comprises nucleotides 13489 – 13492 pairing with nucleotides 13541 – 13537, respectively. The 5' end of nucleotide 13538 connects to Loop 3, shown in magenta or grey when residues lack electron density. Loop 3 encompasses nucleotides 13533 - 13537 and connects the end of Stem 2 with Stem 3. Stem 3, colored orange in all figures, coaxially stacks below Stem 1. Stem 3 encompasses nucleotides 13505 - 13513 which pair with nucleotides 13523 to 13532 with residue 13526 unpaired. This helically stacked three-stemmed structure approximately matches the predicted secondary structure of the PFSE (Fig. 1B).[12] Nevertheless, we observe several base pairing differences, detectable because of the high-resolution data.

2.2.4 Loop 1 organization facilitates formation of the pseudoknot in Stem 2.

In the helically stacked conformation of the PFSE, Loop 1 forms interactions that position the descending strand (in our illustrations) of Stem 2 to form the long-distance H-type pseudoknot

interaction. Starting from the 5' end of Loop 1, nucleotide U13485, which was predicted to be unpaired (Fig. 2.2A), forms a water-mediated base-pairing interaction with U13494 in Stem 1 involving the O2 keto group and N3 imino group, respectively (Fig. 2.2D). Additionally, the N3 imino NH of U13485 donates a hydrogen bond to the O4 keto group of U13494. While U13485 was predicted to be unpaired, nucleotide U13494 was predicted to pair with A13535 as part of Stem 2. However, in our structure U13494 faces inwards towards Stem 1, and is sandwiched in place by base stacking interactions with G13493 and C13495 (Fig. 2.2F). The base pairing between U13494 and U13485 keeps the Loop 1 strand close to the core of Stem 1, which helps position G13486 to form its interactions.

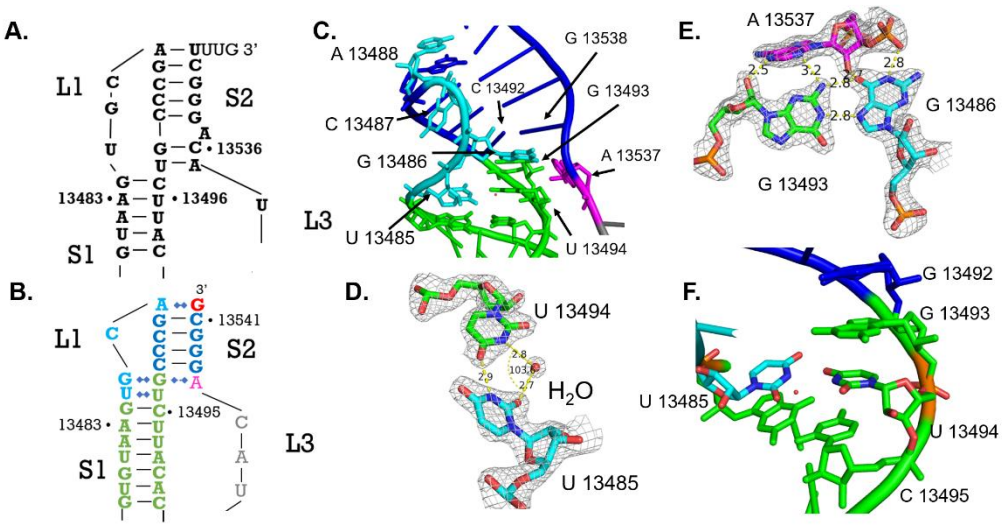


Figure 2.2 Organization of Loop 1 (indicated as L1 in figures). (A) Predicted secondary structure interactions of Loop 1; some secondary structure models predict that A13488 forms a base pair with U13541[9,12] (B) Crystal-structure-derived secondary structure of Loop 1 colored cyan as defined in Fig. 1. (C) Crystal structure model of Loop 1 (cyan). (D) Noncanonical pairing between U13485 (cyan) and U13494 (green) mediated by an ordered water molecule. (E) G13486 - G13493 - A13537 base triple at the Stem 1 (green) - Stem 2 (blue) junction. G13486 (cyan) from L1 interacts with the phosphate oxygen of A13537 (magenta) and forms a Hoogsteen base pair with G13493 (green). (F) Base stacking interactions holding U13495 in the Stem 1 helical stack.

G13486 engages in a base triple interaction with G13493 and A13537. G13486 uses its Hoogsteen face to interact with the Watson-Crick face of G13493 and uses its exocyclic keto group

to accept a hydrogen bond from A13537's 2' hydroxyl group (Fig. 2.2E and Fig. 2.3F). We consider this G13486-G13493-A13537 base triple the transition between Stem 1 and Stem 2 because G13486 is no longer stacking on the Stem 1 helical axis. Above this point C13487 twists away from G13486 leaving C13492 to base pair with G13538 (Fig. 2.2C). C13487's nucleobase plane has rotated relative to the base planes of G13486 and A13488 to run parallel to the helical axis of Stem 1 (Fig. 2.2C). After the turn at C13487, the ribose-phosphate backbone reverses direction again to allow A13488 to engage in a non-canonical pairing interaction with the 3'-terminal G13542, forming the uppermost base pair of the pseudoknot duplex, Stem 2. This positions the nucleotides that follow, G13489 and C13490 - C13492 to base pair with the opposing G rich strand at the 3' end of the RNA, G13538 - 13540 and C13541.

2.2.5 Stem 2 is shorter than predicted.

Stem 2 was originally predicted to consist of six base pairs between nucleotides 13488 - 13494 and nucleotides 13542 - 13535 with A13537 bulged out from the helix (Fig. 2.3A). However, our structure shows that Stem 2 consists of Watson-Crick base pairs between nucleotides 13488 - 13492 and nucleotides 13542 - 13538, (Fig. 2.3B) with A13488 's Watson-Crick face forming a Hoogsteen interaction with G13542 and capping the helix from the top (Fig. 2.3E). Residue A13537, ends the Stem 2 stack and forms an A-minor interaction with the aforementioned G13486 - G13493 Hoogsteen pair at the base of Stem 2. In this base triple A13537 forms two hydrogen bonds: one involves its N6 amine, and the 2'OH of G13493, and the other involves N3 and the exocyclic amine of G13493 (Fig. 2.3F). A hydrogen bond between G13538's nonbridging phosphate oxygen and the imino NH of G13486 further stabilizes the base triple, in addition to the 2' OH and G13486 keto group hydrogen bond mentioned previously (Fig. 2.2 E and Fig. 2.2F). A key difference between the predicted Stem 2 and the crystal structure is that nucleotides C13536

and A13535 do not form the predicted base pairs with G13493 and U13494, respectively. Instead, there is clear density showing that G13493 and U13494 direct their Watson-Crick faces away from the central axis of Stem 2, leaving C13536 and A13535, which lack clear density, without their predicted pairing partners (Fig. 2.3D). We conclude that in this conformation, Stem 2 is shorter than predicted.

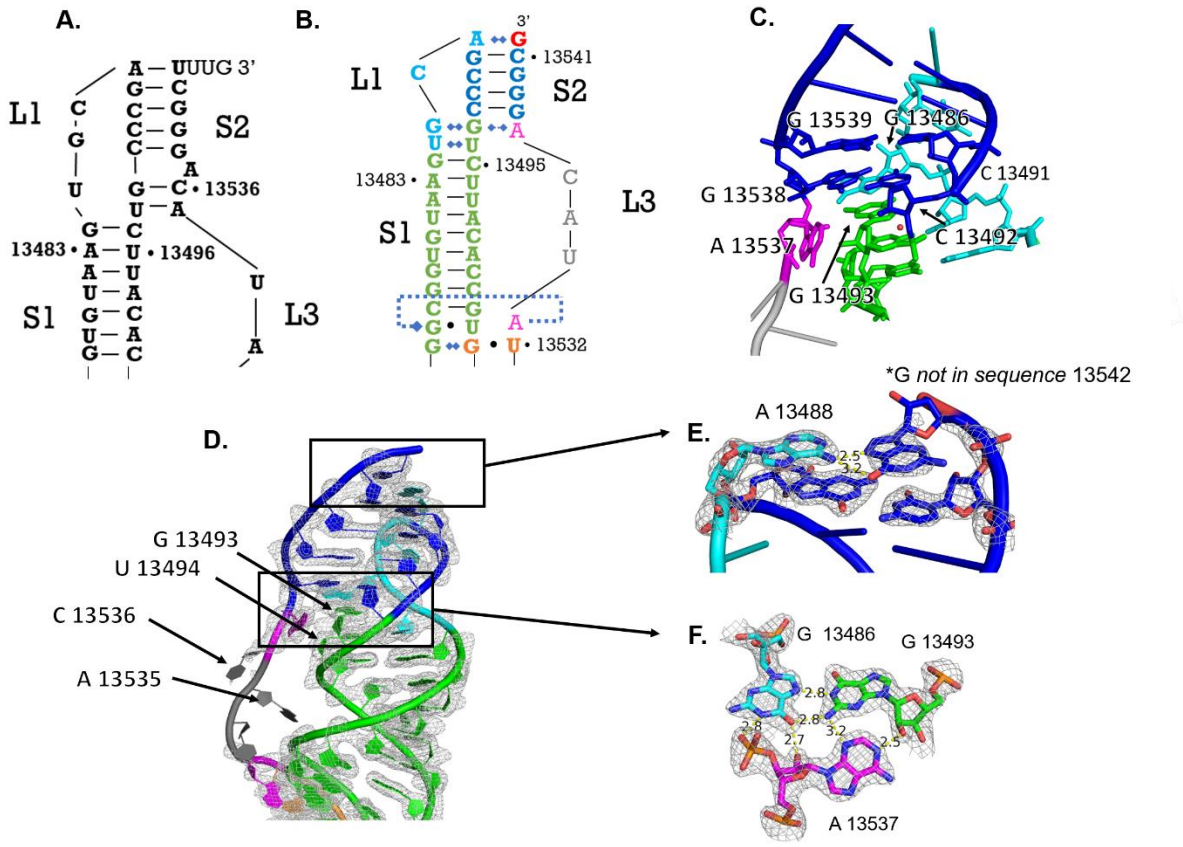


Figure 2.3. Stem 2 (labeled S2 and colored navy) and Loop 3 (labeled L3 colored magenta for residues with density and grey for residues without density) differ from predictions. (A) Predicted secondary structure of Stem 2.[12] (B) Secondary structure derived from the crystal structure. (C) Stem 2/Stem 1 junction showing inflection point at G13493 and C13492. (D) Crystal structure displaying electron density for Stem 2 and Loop 3; electron density map displayed as mesh. (E) A13488 - G13542 Hoogsteen interaction with the electron density map displayed as mesh. (F) A13537, G13493 and G13486 base triple with electron density map displayed as a mesh.

Although the DNA template used for transcription of the PFSE construct was designed to terminate transcription at nucleotide 13541, the electron density map showed density for an

additional 65th nucleotide. As T7 RNA polymerase is known to add untemplated nucleotides to the 3' terminus of transcripts we used a simple method to sequence the 3' ends of in vitro transcribed RNAs extracted from crystals to confirm the presence of an additional nucleotide and reveal its identity.[42] Briefly, the RNA was poly-adenylated to create a primer binding site for a poly(T) reverse primer for reverse transcription into cDNA. The resulting cDNA was amplified and sequenced. This not only revealed the identity of the terminal residues but also quantitatively measured the enrichment for each of the four nucleotides randomly added by T7. RNA extracted from washed crystals shows a G predominating after C13541 (Fig B.3). By contrast the mother liquor contains transcripts with a mix of G and C (Fig B.3). The enrichment for G over C in the crystal may be due to the crystal contact involving stacking of G13542 with a mirrored symmetry mate of the RNA. Whether the native U in position 13542 forms a base pair with A13488 remains unclear.[9]

2.2.6 The C56A point mutation between SARS-CoV-1 to SARS-CoV-2 PFSEs forms a base triple.

Secondary structure predictions and computational models of the PFSE vary in the arrangement of A13533.[9,10,34,35] This residue is often modeled as unpaired in Loop 3 or paired to U13504 in Stem 3. However, in this structure we see clear density for a triple base pair involving G13475, U13504 and A13533 at the S1/S3 junction (Fig. 2.4C). Stem 3 ends at a wobble base pair between G13505 and U13532; above this, A13533 breaks away from this helical stack to form a Watson-Crick sugar-edge interaction with G13475, which forms a wobble pair with U13504 (Fig. 2.4C and 2.4E). This base triple interaction may help stabilize the vertical/stacked/co-axial arrangement of Stem 1/Stem 3.

Compared to the PFSE of SARS-CoV-1 the PFSE of SARS-CoV-2 has a single C to A substitution at residue 13533 in SARS-CoV-2 which corresponds to 13444 in SARS-CoV-1. The

C to A mutation itself seems to have no effect on frameshifting efficiency as both frameshift to appreciable extents, although the in vitro and in vivo efficiencies vary significantly.[33] When position 13533 is modeled as a C in our structure, we see that the guanidinium group is poised to make the same Watson Crick sugar edge hydrogen bonding interactions that the guanidinium group of A makes (Fig B.1). Thus, the SARS-CoV-1 PFSE also appears capable of forming a triple base interaction at the S1/S3 junction, supporting the functional relevance of the base-triple.

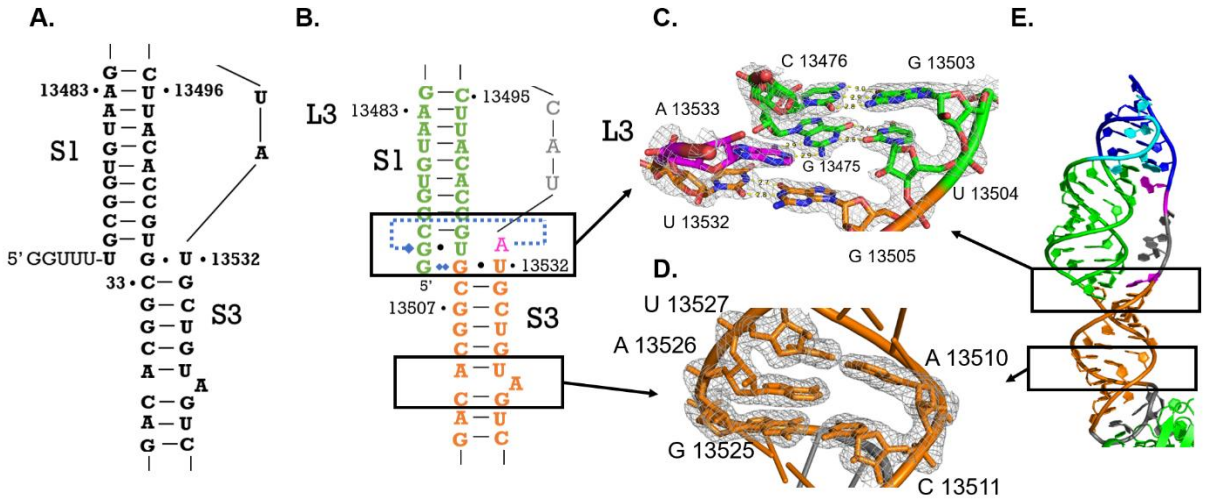


Figure 2.4. Stem 1 and Stem 3 junction. (A) Predicted secondary structure of the Stem 1/Stem 3 junction.[12] (B) Crystal structure derived secondary structure of the Stem 1/Stem 3 junction. (C) Stem 1/Stem 3 base triple formed between A13533 (pink), G13475 (green) and U13504(green). G13505 – U13532 (orange) wobble pair that ends the Stem 3 helical stack. C13476 – G13503 (green) base pair illustrating the start of the Stem 1 helical stack. Electron density for this region displayed as mesh. (D) Unpaired nucleotide A13526 does not bulge from the helix as predicted and resides within the helix via stacking interactions with nucleotides U13527 and G13525; electron density displayed as mesh. (E) Locations of these regions in the PFSE structure.

The S1/S3 junction is also the site of the 5' end threading though the ring created by Stem 1 and Loop 3, consistent with observations from other structural studies of the PFSE.[4,43] Our structure shows that threading is due to the twist of the Stem 1 helix holding the first 5 nucleotides in place as Loop 3 and the 3' end of the RNA fold over top (Fig B.2). Here we also find that

G13475, included at the 5' end for transcriptional efficiency in place of the native U, interacts with the G13505 – U13532 wobble pair of Stem 3 (Fig. 6.2). In this base triple, the N1H imino group and $N2H_2$ exocyclic amine of G13474 donate hydrogen bonds to N7 and the O6 keto group of G13505, respectively (Fig. 6.2). The native U nucleotide in position 13474 would not be able to make the analogous interactions. While the nonnative interaction of G13474 may stabilize the vertical conformation, its absence in the native structure would not preclude formation of the vertical conformation.

2.2.7 Geometric constraints restrict the conformation of nucleotides with missing density.

Loop 3, which connects Stems 2 and 3, consists of nucleotides 13533, 13534, 13535, 13536, 13537 in our structure. Density for residues 13534, 13535, and 13536 is missing from the data sets, suggesting this loop is flexible and not well organized in the crystal (Fig. 2.3D). Nevertheless, the 3' end of A13533 and the 5' phosphate of A13537, which are well-defined by the electron density, are separated by 17.4 Å. The length of one fully extended nucleotide of ssDNA is 6.7 Å.[44] Single-stranded RNA is expected to be more compact than ssDNA due to steric restraints imposed by the 2'OH on the sugar pucker conformations. If nucleotides 13534, 13535, and 13536 in Loop 3 were fully extended, they would span a maximum distance of 20.1 Å, leaving only ~2.7 Å of slack for bending. In a study of a forty-nucleotide strand of poly(U) RNA a total contour length was 196 Å was measured, giving 4.9 Å as the average length of each nt.[44] In a relaxed state we would therefore expect three nucleotides to span 14.3 Å meaning that in spanning 17 Å the nucleotides in Loop 3 are likely extended and less flexible than they would be if their ends were free. We anticipate that the residues in this region likely possess enough flexibility to wiggle back and forth like a short string held at both ends but must maintain a relatively extended backbone conformation to bridge the distance between residues 13533 and 13537.

2.2.8 Stem 3 organization matches predictions.

The base pairing for Stem 3 observed in our structure agrees with the predicted pairing. In previous work A13526 frequently exhibits sensitivity to chemical modification indicative of a single-stranded nucleotide.[14] We find this residue unpaired and stacked within Stem 3 (Fig. 2.4D). Previous mutational studies have shown that deletion of A13526 or insertion of a corresponding U to form a base pair both reduce frameshifting efficiency in the SARS-CoV-1 PFSE, suggesting that this residue's unpaired state contributes to the frameshifting mechanism. Nevertheless, complete deletion of Stem 3 rescues frameshifting.[14] Often computational modeling predicts that the PFSE is bent at the junction between Stem 1 and Stem 3, and cryo-EM structures exhibit this bend.[4,8,9,10,32,35,43] Having A13526 unpaired could contribute to the dynamic character of Stem 3, facilitating sampling of the bent conformation.

2.3 DISCUSSION

2.3.1 Comparing existing structural probing data with our crystal structure.

It is possible that our Fab-hairpin crystallization module and crystal packing forces facilitated formation of the linear conformation we observe in this crystal structure. However, at a minimum, the secondary structure shows close agreement with the base pairing pattern inferred from chemical probing, mutational analysis and NMR[11,34,35,43,45]. Nevertheless, the biological relevance of the linear conformation and associated base triples observed in this high-resolution structure await further investigation either in the context of frameshifting or another stage of the viral lifecycle. We note that alternate orientations of Stem 3 relative to Stem 1 and Stem 2 do not

seem to require major rearrangements to the secondary structure as illustrated in Figure 2.5, suggesting that the PFSE potentially samples the linear conformation in solution.

Analysis of the SARS-CoV-1 PFSE by chemical and enzymatic probing and NMR, supported by mutational studies, indicate that frameshifting depends on the PFSE adopting a three stemmed H-type pseudoknot like structure.[12,14,31] Recent DMS and SHAPE probing performed on minimal constructs of the highly homologous SARS-CoV-2 PFSE implicate an analogous three stemmed structure.[11,34,35,43,45] NMR analysis detected base pairing interactions consistent with the same three stemmed arrangement of the SARS-CoV-2 PFSE in solution as well.[31] SAX data collected on the SARS-CoV-1 and SARS-CoV-2 PFSE in combination with in vitro translation assays in the presence and absence of the frameshifting inhibitory ligand MTDB all support the conclusion that these SARS-CoV's PFSEs adopt the same structure and perform frameshifting through the same mechanism.[5,33]

This first experimentally determined high resolution structure of a SARS-CoV PFSE has revealed interactions that were not identified via chemical probing and NMR assays previously applied to CoV-1 or CoV-2 PFSE. These new interactions provide a structural hypothesis that may explain the C to A variation at residue 13533 in SARS-CoV-2 and 13444 in SARS-CoV-1, which are otherwise absolutely conserved within the PFSE. While the secondary structure closely matches predictions for both SARS-CoV-1 and SARS-CoV-2 PFSEs, the base triple interactions detected at the S1/S2 and S1/S3 junctions in our 2.09 Å resolution data set may contribute to conformational sampling, a process critical for frameshifting in cells or influence interactions with cellular factors.[27,29] The roles that the vertical conformation and the accompanying base triples play in the mechanism of frameshifting can now be investigated with directed tests informed by the structure.

The PFSE region may only sample this three stemmed arrangement occasionally, as the in vivo probing reactivities of the PFSE in its genomic context exhibit differences compared to in vitro chemical probing reactivities of minimal constructs.[11,34,43,47] Lan et al. showed that the in vitro DMS reactivities of the PFSE become more similar to those observed in vivo as the construct is elongated to include more of the genomic sequence.[11] This suggests that the PFSE region likely samples different or additional conformations in its genomic context, potentially forming long distance interactions. However, during translation with translocating ribosomes upstream and downstream, the PFSE region would be unfolded and isolated from the rest of the genomic RNA enabling formation of the frameshifting structure.[4,17,46] Additionally, the finding that the MTDB can inhibit viral replication and alters the ratios of the conformations the PFSE can adopt suggests that structures adopted by the minimal PFSE element have relevance for viral function.[4,5]

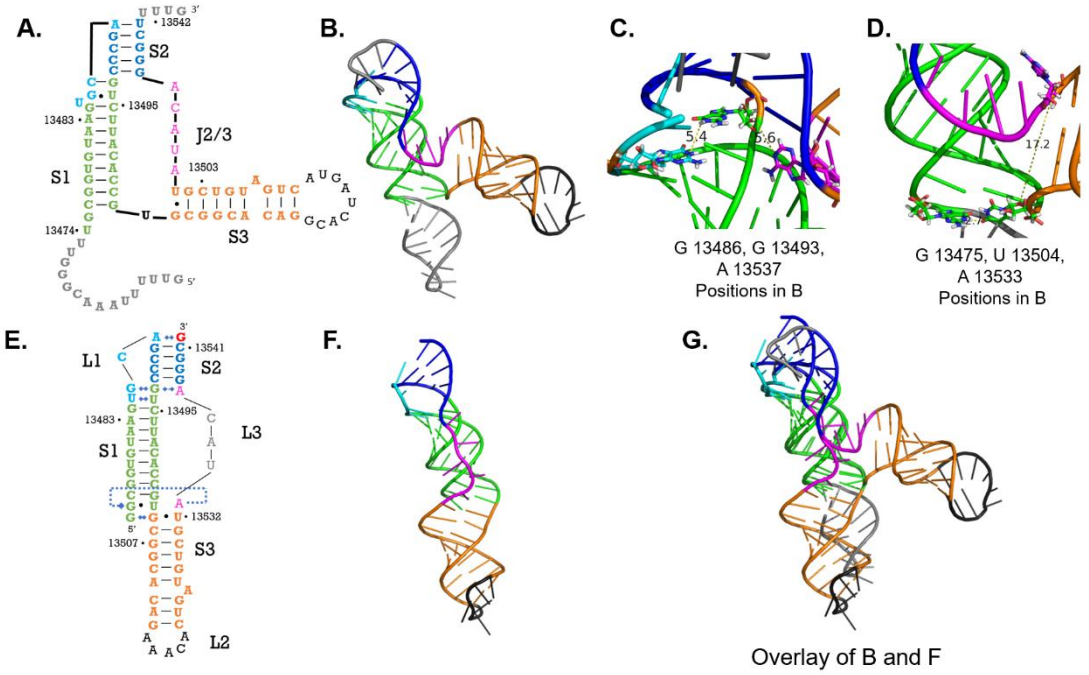


Figure 2.5. SARS-CoV-2 PFSE cryo-EM structure solved by Zhang *et al.* [43] overlaid with crystal structure reported here. (A) Secondary structure model Zhang *et al.*[43] cryo-EM reported renumbered to correspond to our numbering for comparison; Stem 1 colored green; Loop 1 colored

Figure 2.5 continued from previous page: cyan; Stem 2 colored navy; Loop 3 colored magenta; Stem 3 colored orange; Loop 2 colored black; nucleotides present in this cryo-EM structure but not in our structure are colored grey. (B) cryo-EM model (PDB code: 6XRZ) displayed as a cartoon colored as in A. (C) Stem 1/Stem 2 base triple residues, G13486, G13493, A13537 displayed as sticks in the cryo-EM structure; measurements display distance between atoms that interact in the crystal structure. (D) Stem 1/Stem 3 base triple residues G13475, U13504, A13533 displayed as sticks in cryo-EM structure. (E) Secondary structure map derived from the crystal structure colored to match A. (F) Crystal structure displayed as a cartoon colored as in B. (G) Overlay of cryo-EM model (B) and our crystal structure (F).

2.3.2 Comparison of our crystal structure to cryo-EM structures.

Currently two mid-resolution cryo-EM models exist for the PFSE. Both share important features with our model, and the differences may provide hints as to the mechanism of ribosomal frameshifting. Zhang et al. used cryo-EM data at 6.7 Å - 6.9 Å resolution as a constraint for the Ribosolve pipeline to model the structure of the PFSE and slippery site (nt 13,459 - 13,548 in NC.045512.2).[43] The second cryo-EM structure was reported by Bhatt et al.[4] In this case the PFSE pseudoknot is present in the context of the genomic mRNA encoding NSP 10 - 12 bound to a ribosome paused over a mutated slippery site.[4] Here the PFSE region is solved to 5 Å - 7 Å.[4] The coordinates for the ribosome bound structure have not yet been deposited, so making direct comparisons between this structure and our crystal structure is not yet possible.

Both cryo-EM structures find the PFSE in a “bent” arrangement, where Stem 1 and Stem 2 are helically stacked while Stem 3 bends perpendicularly away from Stem 1.[4,43] Figure 2.5 illustrates the similarities and differences between the Zhang et al. structure (PDB code: 6XRZ) and our crystal structure.[43] Our 2.09 Å structure confirms many of the base pairing interactions that remained ambiguous at 6.7 Å. The secondary structures reported by each work identify the same differences from the ‘literature’ secondary structure as we have found. The structures agree that Stem 2 ends at A13537, and the crystal structure reveals additional details about the specific base pairing interactions at the S1/S2 helical junction. Our structure also shows the same 5’ end

threaded topology first identified in these cryo-EM structures. In contrast to the cryo-EM structures, our crystal structure adopts a vertically stacked conformation, which has been consistently predicted by computational modeling but not yet observed experimentally.[9,10,34] This structure could represent the second SARS-CoV-2 PFSE conformation.

Curiously, neither of the base triple pairing interactions we report are present in the cryo-EM models, although this may be due to the limited resolution of these structures. The bent conformation of the PFSE and slippery site observed in the solution cryo-EM model (PDB code: 6XRZ), positions the S1/S2 base triple nucleotides (G13486, G13493 and A13537) too far apart to form the base triple (Fig. 2.5C).[43] Additionally, in the ribosome-bound structure G13486 forms direct interactions with the N-terminal domain of US3 while A13537 remains unpaired in the J2/3 region.[4] Therefore, the S1/S2 base triple we observe would have to be dissolved in the bent conformation and in the ribosome bound state. Interestingly, mutations to G13486 or A13537 in isolation were shown to reduce frameshifting frequency markedly, although the role of A13537 is unknown.[4] The PFSE's intolerance to A13537 mutations combined with these structural observations suggest that G13486 may serve an additional structural role to orient G13493's sugar edge to interact with A13533's WC face. In other -1 frameshifting pseudoknots elimination of known base triple interactions does reduce frameshifting efficiency by destabilizing the pseudoknot.[48] These observations suggest that the S1/S2 base triple may act as a conformational switch or decision point between the bent conformation and the vertical conformation.

The base triple involving G13475, U13504 and A13542 at the S1/S3 junction may further stabilize the vertical conformation as it cannot be fully formed in the bent or ribosome-bound cryo-EM structures. In the ribosome bound structure G13475, of the S1/S3 base triple, is inside the mRNA entry channel but A13533 and U13504 are modeled as paired.[4] In the cryo-EM structure

of the element in isolation (PDB code: 6XRZ) G13475 and U13504 are close enough to base pair while A13533 is more than 10 Å away (Fig. 2.5D).[43] Unfortunately, we lack mutational information for this region of the PFSE to illuminate the contributions of these nucleotides to frameshifting. In nature, however, the residue corresponding to A13533 in SARS-CoV-2 is a C (13444) in SARS-CoV-1. The role of this C to A substitution is not yet understood, but when A13533 is modeled as a C, the hydrogen bonding interactions with G13475 can still be formed (Fig B.1).[33] The role of this S1/S3 junction region now warrants deeper investigation as it could dictate a dynamical relationship between bent and linear conformations and thereby influence frameshifting efficiency.

Our crystal structure, the cryo-EM structure of the free PFSE with the slippery site and the ribosome-bound PFSE-containing gRNA fragment have similar topology and differ only at junctions where alternate pairing interactions or ribosomal interactions can occur. The crystal structure model represents a vertically stacked conformation and the cryo-EM structures represent a bent or wedged conformation. Modeling of the PFSE by Omar et al. and Rangan et al. demonstrated that the arrangement of Stem 1 and Stem 2 relative to Stem 3 can be flexible.[9,10] Indeed, substantial data exist that support the hypothesis that the frameshifting element populates two distinct conformations that govern the efficiency of PFSE frameshifting[27,28,29,50].

We caution, however, that conformational differences detected across the various models of the PFSE may be due to differences in RNA construct design and or ribosomal interactions. For example, our construct lacks the 5' slippery site sequence, which forms a helix that coaxially stacks beneath stem 1 in the cryo-EM model. The slippery site helix could preclude formation of the vertically stacked conformation just as the ribosome does. Additionally, recent investigation of the SARS-CoV-2 RNA genome both within the cytoplasm and the virion infers entirely new

secondary structures for the PFSE.[11,47,58] Rather than frameshifting itself, the vertical conformation observed here could have significance for PFSE structural rearrangements that occur during other stages of the virial lifecycle. Moreover, as there are limited structures available for RNA pseudoknots, the high-resolution structure data presented herein have intrinsic value for structure-based design of small molecule binders, as a starting point for molecular dynamics simulations, and as a framework for testing the relationship of the RNA structure to its function.

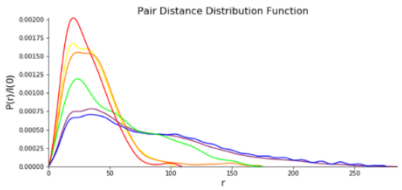
2.3.3 Correspondence of this structure with PFSE dynamics and folding data.

In programmed ribosomal -1 frameshifting pseudoknots, helical stacking is often a required feature to induce frameshifting, and tertiary interactions have been shown to stabilize frameshifting pseudoknots and promote efficiency.[48,49] Frameshifting pseudoknots must also be conformationally dynamic to function, refolding into one among multiple conformations each time the ribosome reads through the mRNA.[28,49] Optical tweezers can be used to both unfold the pseudoknot and mimic the tugging forces the ribosome would apply to a pseudoknot during translation.[17,29] In reporter assays, the rate of frameshifting of any given pseudoknot correlates linearly with its conformational Shannon entropy, a statistical metric for conformational plasticity.[29] Pseudoknots that occupy two conformations often induce frameshifting on the order of 20% efficiency.[27] Nevertheless, frameshifting rates measured in-vitro can be different than those measured in vivo or in infected cells due to additional interactions with proteins and the gRNA.

Unfolding force extension curves show that the SARS-CoV-2 PFSE can adopt two distinct conformations of roughly similar stability, falling in line with expectations given a frameshifting frequency in the 15% - 30% range. [50] The more stable form (N), which unfolds at an average force of 30 pN and the slightly less stable form (N') unfolds at an average force of 16 pN. Possibly the two states correspond in part to the vertical and bent conformations, although we emphasize that

this awaits careful validation. The base triples in the vertical conformation would be expected to provide additional stability and rigidity relative to a conformer lacking them. In line with this possibility, another H-type pseudoknot found in human telomerase RNA (ΔU177) was found to lose ~16 pN of unfolding force upon elimination of two minor grove triple bases, which is a similar difference in unfolding force between N and N'.[48,50] It will be important to address whether the vertical conformation corresponds to one of the states in the unfolding experiments, and our structure will inform design of atomic mutations for these tests.

2.3.4 Ongoing work to validate the presence of the linear conformation in solution



	Guinier Rg	P(r) Rg	Guinier I(0)	P(r) I(0)	Mw (Vc) KD	Mw (actual) KDa
PFSE BL36	24.8 (+/-0.2)	25.55 (+/-0.3)	6.52 (0.02)	6.05 (0.03)	21.3	20
PFSE BL36 A13533U	28.3 (0.11)	32.63(1.0)	9.33 (0.02)	9.65 (0.08)	30.8	20
PFSE BL36 A13537U	27.47 (0.25)	28.29 (0.21)	6.82(0.02)	5.83 (0.03)	29.7	20
PFSE wt	41.04 (0.2)	46.65 (0.48)	12.09 (0.05)	12.68 (0.08)	35 (42 B) (52 Vc)	22
PFSE wt A13533U	60.67 (0.84)	71.51 (1.75)	19.85 (0.26)	21.31 (0.39)	66.3	22
PFSE wt A13537U	55.19 (0.55)	66.33 (2.08)	84.42 (0.69)	91.25	58.3	22

Figure 2.6. SAXS p(R) functions plotted as a distance distribution PFSE BL36 monomer shown in red with A13533U PFSE BL36 mutant shown in orange and A13537U shown in yellow. PFSE WT shown in green, the absence of a compact bell shape plot is indicative of an extended possibly aggregated structure mutants in the wily type background shown in blue and purple amplify this aggregation.

To determine if the PFSE adopts a linear and bent conformation we collected SAXS profiles on wild type PFSE, PFSE BL3-6 and versions of these constructs with A13533U and A13537U mutations. The p(r) function was calculated from the merged images and a low-

resolution envelope was then calculated DAMIF program in the RAW program suite. This work was supported by collaborations with Dr. James Fuller and Dr. Sichun Yang who aided in the SAXS data processing and analysis respectively.

Because of the degeneracy in RNA pairing, the SAXS profiles of the PFSE constructs with mutated base triples are likely to be ambiguous due to the possibility of alternate secondary structures forming. Therefore, the initial experiments reported here will need to be repeated with atomic mutagenesis. Nevertheless, as described below we do have preliminary SAXS data that suggest structural contributions from the base triples.

The PFSE wild type construct has been shown to dimerize with its self, this dimerization seems to play a functional role in packing the gRNA into the viral capsid[21]. Unfortunately, that made the construct adopt the dimer form in solution, detectable as a radius of gyration that is twice the molecular weight of the construct (46kDa vs 22kDa)(Figure 2.6). This occluded information about a bent or linear conformation. The PFSE BL36 construct however maintained its monomer form in solution and its experimental SAXS data fits reasonably well with the linear structure (up to $q = 0.2\ A^{-1}$). To get a better fit of this data to both the bent and linear conformations and to calculate the proportion of each in the envelope Dr. Yang plans to use MC-Fold and MC-Sym to modify the starting models to represent states between these two extremes. In the BL3-6 background A56U and A60U SAXS profiles are nearly identical to each other but are dramatically different from the unmutated PFSE BL3-6 (Figure 2.6). Neither profile matches either the linear nor the bent model of the PFSE. The p(r) function of the mutants has an unusually large Dmax that cannot be fit to zero which is indicative of a more expanded structure (Figure 2.6). It may be that if L3 loses its interactions with S1 and S2 the strand may become loose enough to allow for additional structural flexibility which could account for the extended

conformation detected with SAXS. From this we surmise that the conformational ensemble has indeed been changed and that these residues do play a role in that conformational ensemble.

2.4 Methods

2.4.1 Construct Design.

Our crystallization target was based on nucleotides 13475 – 13541 in the SARS-CoV-2 genome gen bank number NC_045512.2; these are the minimal residues predicted to form base pairs.[30] We chose to exclude nucleotide 13542 in this crystallization construct because its binding to A13488 was unclear, and incorporating an unpaired nucleotide at the 5' end could compromise the structural integrity of the crystallization complex. Typically, the PFSE is defined as this region as well as the 14 nucleotides upstream, which include the slippery site; these were excluded from the crystallization construct because long single-stranded regions of RNA can disrupt folding and crystallization of the RNA. This truncated construct we refer to as the wild type PFSE construct.

To enable binding to antibody fragment, Fab BL3-6, we mutated the nucleotide corresponding to Loop 2 (13514 to 13522) to the sequence AAACA. This crystallization construct is referred to as the BL3-6 PFSE construct. We chose to mutate these residues because mutational studies have shown that Loop 2 can be mutated to a common RNA tetraloop without altering the ratio of frameshifting in the SARS-CoV-1 PFSE, which is believed to be structurally identical to the SARS-CoV-2 PFSE.

V-fold[51] predictions suggest that the BL3-6 PFSE does not contain non-native regions of complementarity that might be prone to disrupt the native secondary structure. A gel shift assay was performed on refolded PFSE RNA constructs and antibody fragment BL3-6 to show that the

grafted nucleotides still form the expected solvent-exposed loop, and that Fab binding was only observed for the mutated construct.

2.4.2 RNA Transcription and Purification.

Single stranded DNA templates and primers for PCR and transcription were ordered from IDT encoding the transcription template for each RNA construct with a T7 promoter. Forward primers were ordered matching the T7 promoter region and reverse primers contained a single 2'O-methyl modification at the 3' end to avoid untemplated additions by T7 polymerase.[52] Transcription template DNA was amplified into double stranded DNA using PCR. RNA was transcribed from the purified PCR product using an in vitro transcription reaction as follows 50 pmol mL^{-1} DNA template was incubated for 3 h at 37 °C in buffer containing 40 mM Tris-HCl, pH 8.0, 2 mM spermidine, 10 mM NaCl, 25 mM $MgCl_2$, 10 mM DTT, 40 U mL^{-1} RNase inhibitor, 5 U mL^{-1} thermostable inorganic pyrophosphatase, 5 mM of each NTP, and 50 μg mL^{-1} T7 RNA polymerase. Reactions were halted by addition of RNase free DNAase1 at 5U mL^{-1} and incubation at 37 °C for 30 minutes. RNA was purified on a 10% denaturing polyacrylamide gel in 0.5x TBE running buffer. The RNA was visualized with UV shadowing, extracted and eluted into 10 mM Tris, pH 8.0, 2 mM EDTA, 300 mM NaCl buffer via overnight incubation at 4 °C. The eluted RNA was then concentrated and exchanged into double distilled H_2O using a 10K Amicon filter and stored at -80 °C until further use.

2.4.3 Fab Purification.

The BL3-6 Fab expression vector (available upon request) was transformed into 55244 chemically competent cells (www.atcc.org) and grown on LB plates supplemented with carbenicillin at 100μg mL^{-1}. Nine colonies from the plates were chosen and inoculated to a starter

culture with 100 μg mL^{-1} carbenicillin, which was grown at 30 ºC for 8 hours. Once the starter culture reached an OD 600 of 8, 15mLs of starter culture was used to inoculate 1L of 2×YT media and grown for 24 h at 30 °C. The cells were then pelleted via centrifugation at RT, and the cell pellet was resuspended in 1L of freshly prepared CRAP-Pi media supplemented with 100 ug mL^{-1} carbenicillin. The cells were set to grow for 24 h at 30 °C, harvested via centrifugation at 4ºC and frozen at -20 ºC. Frozen cell pellets were lysed in PBS buffer supplemented with 0.4 mg mL^{-1} of Lysozyme and 0.01 mg mL^{-1} of DNase I. After 30 minutes PMSF was added to a final concentration of 0.5 mM. After 30 minutes, the cells were centrifuged, 45 min, 12000 rpm, at 4 ºC. Lysate was transferred to new sterile bottles and centrifuged again for 15 minutes, 12000 rpm, at 4 ºC. Supernatant was filtered through 0.45 μm filters into a sterile bottle (Millipore Sigma, www.sigmaaldrich.com), and Fab proteins were purified using the AKTAxpress fast protein liquid chromatography (FPLC) purification system (Amersham, www.gelifesciences.com) as described previously.[37] The lysate in PBS buffer (pH 7.4) was loaded into a protein A column, and the eluted Fab in 1 M acetic acid was buffer exchanged back into the buffer PBS (pH 7.4) using 30kDa cutoff Amicon filter and loaded into a protein G column. The Fab was eluted from protein G column in 0.1 M glycine (pH 2.7) and then buffer-exchanged into 50 mM NaOAc, 50 mM NaCl buffer (pH 5.5) and loaded into a heparin column. Finally, the eluted Fab in 50 mM NaOAc, 2 M NaCl (pH 5.5) was dialyzed back into 1× PBS (pH 7.4), concentrated, and analyzed by 12% SDS-PAGE using Coomassie Blue R-250 staining for visualization. Aliquots of Fab samples were tested for RNase activity using the RNaseAlert kit (Ambion, www.thermofisher.com). The aliquots of Fab samples were flash frozen in liquid nitrogen and stored at −80 °C until further use.

2.4.4 Electrophoretic Mobility Shift Assay (EMSA).

To validate Fab binding to the BL3-6 PFSE RNA construct, purified RNA constructs in double distilled H_20 were heated to 90 ^oC for 1 minute, then cooled on ice for 2 minutes then held at room temperature for 3 minutes. Refolding buffer (50 mM HEPES pH 8, 5mM $MgCl_2$, 100 mM KCl) was added, and the RNA was then incubated at 50 ^oC for 10 minutes to facilitate refolding. RNA was then mixed with either PBS as a negative control or a 1.1 M ratio of RNase-free Fab BL3-6 and incubated at room temperature for 30 minutes to establish equilibrium binding. Fab RNA complexes were separated by gel electrophoresis in a 12% polyacrylamide gel made in 0.5x TBE buffer supplemented with 5mM $MgCl_2$. The gel was stained with ethidium bromide and visualized via UV light and photographed (Fig B.4).

2.4.5 Sequencing Reactions.

To determine the identity of the untemplated nucleotide observed in the electron density map, we sequenced the RNA from three sources: the transcription product, the mother liquor of the crystal drop and from the crystal itself. For the transcribed RNA we followed the standard procedure for poly(A)-tailing (NEB) using 3 μg of RNA in reaction with *E. coli* poly(A) polymerase. For the mother liquor-derived RNA samples: mother liquor was harvested from a 200 nL drop of crystallized complex. To ensure all the mother liquor solution was harvested, 0.5 μL of well solution was added to the drop prior to transferring the samples to an eppendorf tube. For the crystal-derived RNA samples: after the mother liquor was harvested from the drop, the remaining crystals were washed 3 times with well solution. Then the crystals were transferred to an eppendorf tube with 2 μL of RNase free water and crushed via pipetting. Samples with transcribed RNA, mother liquor- and crystal-derived RNA were denatured for 1 minute in 90 °C. Then cooled down on ice for 2 minutes and incubated for 3 minutes at room temperature. Denatured RNA samples

were elongated with *E. coli* poly(A) polymerase 1μL of 5000 U mL^{-1} (NEB), and 2 μL of 10x *E. coli* Poly(A) Polymerase Reaction Buffer with addition of 2 μL of 10 mM ATP and RNase free water up to 20 μL as per manufacturer's instructions. Reaction was incubated at 37 ºC for 30 minutes in water bath. Reaction was halted by addition of EDTA to a final concentration of 10 mM. RNA was ethanol precipitated and checked for poly(A) elongation on a 10% polyacrylamide gel stained with Ethidium Bromide. Polyadenylated RNA was used as the template in a reverse transcription reaction using SuperScript III (Invitrogen) according to manufactures instructions. cDNA was amplified using end specific primers and PCR with 30 cycles of amplification using an annealing temperature of 52 ºC and *Taq* DNA polymerase (NEB). The double stranded DNA products where then submitted for sequencing.

2.4.6 Crystallization.

PFSE BL3-6 RNA was denatured in water by incubation at 90 ºC for 1 minute, ice for 2 minutes and room temperature for 3 minutes. The RNA was then refolded by addition of refolding buffer (50 mM HEPES pH 8, 5mM $MgCl_2$, 100 mM KCl) and incubated at 50 ºC for 10 minutes. Fab was added to the RNA at a 1:1.1 molar ratio of RNA:Fab and incubated at room temperature for 30 minutes before concentrating the complex to 6mg mL^{-1} RNA via centrifugal filtration in a 10 kDa cutoff Amicon Centrifugal Filter Unit. Concentrated complexes were then filtered using 0.2 μm cutoff Millipore centrifugal filter units and used to set high-throughput hanging-drop vapor-diffusion crystallization screens at RT using commercially available screening kits from Hampton Research and Jena Bioscience using the Mosquito liquid handling robot (TTP Labtech). Crystals grew in of 2 % v/v tacsimate pH 4; 0.1 M sodium acetate trihydrate pH 4.6; 16 % w/v polyethylene glycol 3,350, and were further optimized with the addition of either 0.01 M sarcosine or 0.01 M betaine hydrochloride. Crystals appeared and grew to full size within a week at 21 ºC.

Some crystals were looped and transferred to new drops of a solution containing 80% glycerol and 3 mM MTDB in addition to the original crystallization conditions to incorporate the PFSE ligand. Other crystals were grown in the presence of 10 molar equivalents of ligand for each mole of RNA-Fab complex. In neither case was density for the ligand detected in the electron density map. Other crystals were looped and transferred to new drops containing the original crystallization conditions with added 20 % glycerol (v/v) and 3 mM iridium hexammine as cryo-protectant and to incorporate iridium hexammine into the crystal lattice. Iridium hexammine was synthesized in house following the protocol established by Batey et al.[53] A set of crystals were allowed to incubate for 24 hours while another set of crystals were only allowed to incubate for 2 hours before looping and freezing. The crystals incubated with iridium hexammine for only two hours diffracted to higher resolution than those soaked for 24 hours.

2.4.7 Crystal Diffraction Data Collection Processing and Analysis.

The X-ray diffraction data sets were collected at the Advanced Photon Source NE-CAT section beamline 24-ID-C and 24-ID-E. Crystals soaked with iridium hexammine were shot with both the default wavelength of 0.979180 and a wavelength of 1.04040 to illicit anomalous diffraction of the iridium. All the data sets were then integrated and scaled using its on-site RAPD automated programs (https://rapd.nec.aps.anl.gov). Initial SAD-MR phases were obtained from a data set which diffracted to only 3.27 Å using a partial molecular replacement (MR) solution of the Fab (PDB code: 6DB8) in Phenix Autosol[38,40]. A low-resolution election density map was able to be calculated for a portion of the RNA from which a partial model of the FSE was built. This partial RNA model was used in addition to the fab model (PDB code: 6DB8) to find a molecular replacement solution for a higher resolution (2.09Å) data set collected from a native unsoaked crystal. The electron density map using the MR phases of the partial RNA-Fab model were vastly

improved and allowed for unambiguous model building nucleotide by nucleotide. [26,38] The model was iteratively built and refined in coot and phenix refine until the R_{work} and R_{free} could not be further improved. [39-41] Water was automatically added during refinement and later validated visually in COOT[41] according to the electron density map and difference map and potential hydrogen bonding interactions. All structure related figures were made in PyMOL[54] (www.pymol.org), and figures labels were edited in Microsoft PowerPoint.

2.4.8 SAXS Sample prep data collection and analysis

PFSE RNA constructs were denatured in water by incubation at 90 ºC for 1 minute, ice for 2 minutes and room temperature for 3 minutes. The RNA was then refolded by addition of refolding buffer (50 mM HEPES pH 8, 5mM $MgCl_2$, 100 mM KCl) and incubated at 50 ºC for 10 minutes. RNA was concentrated to 4mg/mL using filter centrifugation and 30uL samples were prepared at 4mg/mL, 2mg/mL and 1mg/mL concentrations by diluting with flow through. Flow through for each refolded construct was submitted along side each concentration series as a blank. The buffer blank signal were averaged and subtracted from the data by the ALS beamline hosts. Buffer subtracted images were merged into raw data files from which the p(r) functions were calculated. P(r) profiles were then used to generate a three-dimensional bead model using DAMIF.

2.5 Additional Details

2.5.1 Accession Numbers

Atomic coordinates and structure factors for the reported crystal structures have been deposited with the Protein Data Bank (PDB, https://www.rcsb.org/) under accession number 7MLX.

2.5.2 Conflict of Interest

The authors of this publication report no conflicts of interest.

2.5.3 Acknowledgements

We are grateful to lab members M. Moore, H. Reese and M. Disare for insightful comments and critical review of the manuscript. We thank P. Rice, E. Ozkan, B. Dickinson, and J. Staley for valuable insights and discussion throughout the course of this project. This work was supported by NIH/NIGMS Grant GM102489. C.R. was supported by the HHMI Gilliam Fellowship and NIH Chemical Biology Training Grant T32GM008720. This work also used the Advanced Photon Source on the Northeastern Collaborative Access Team beamline, 24-ID-C and 24-ID-E (GM124165), a Pilatus detector (RR029205), and an Eiger detector (OD021527) at the APS (DE-AC02-06CH11357). We are very grateful for the support from the APS NE-CAT Beamline hosts.

2.5.4 Author Contributions

Contributing Authors: Anna Lewicka, Deepak Koirala, Nan-Sheng Li, and Joseph A. Piccirilli

A. L. and D. K. conceived and designed the initial experiments under direction from J.A.P. A. L. prepared the samples and conducted biochemical and crystallographic experiments and collected x-ray diffraction data. N.-S. L. synthesized the frameshift inhibitory ligand MTBD and the iridium hexammine. A.L. and D. K. performed the initial crystallization trials and data collection together. C. R. solved the crystal structures, analyzed the biochemical and crystallographic data, interpreted the results, and wrote the manuscript with J.A.P.

2.5.5 Rights and Permissions

ACS Chemical Biology 2021 16 (8), 1469-1481 DOI: 10.1021/acschembio.1c00324

2.6 Works Cited in Chapter 2

1) Warner, K. D., Hajdin, C. E., Weeks, K. M., (**2018)** Principles for Targeting RNA with Drug-like Small *Molecules. Nat. Rev. Drug Discov*. 17 (8), 547–558. https://doi.org/10.1038/nrd.2018.93.

2) Shao, Y., Zhang, Q. C., (**2020)** Targeting RNA Structures in Diseases with Small Molecules. *Essays Biochem.*, 64 (6), 955–966. https://doi.org/10.1042/EBC20200011.

3) Park HJ, Park SJ. (**2012**) Virtual Screening for RNA-Interacting Small Molecules. *Biophysical approaches to translational control of gene expression.* 1:235-252. doi:10.1007/978-1-4614-3991-2_12

4) Bhatt, P. R., Scaiola, A., Loughran, G., Leibundgut, M., Kratzel, A., McMillan, A., O' Connor, K. M., Bode, J. W., Thiel, V., Atkins, J. F., Ban, N. (**2020)** Structural Basis of Ribosomal Frameshifting during Translation of the SARS-CoV-2 RNA Genome. *bioRxiv*, 2020.10.26.355099. https://doi.org/10.1101/2020.10.26.355099.

5) Ritchie, D. B., Soong, J., Sikkema, W. K. A., Woodside, M. T., (**2014)** Anti-Frameshifting Ligand Reduces the Conformational Plasticity of the SARS Virus Pseudoknot. *J. Am. Chem. Soc.* 136 (6), 2196–2199. https://doi.org/10.1021/ja410344b.

6) Kelly, J. A., Woodside, M. T., Dinman, J. D., (**2021**) Programmed -1 Ribosomal Frameshifting in Coronaviruses: A Therapeutic Target. *Virology* 554, 75–82. https://doi.org/10.1016/j.virol.2020.12.010.

7) Neupane, K., Munshi, S., Zhao, M., Ritchie, D. B., Ileperuma, S. M., Woodside, M. T., (**2020**) Anti-Frameshifting Ligand Active against SARS Coronavirus-2 Is Resistant to Natural Mutations of the Frameshift-Stimulatory Pseudoknot. *J. Mol. Biol*. 432 (21), 5843–5847. https://doi.org/10.1016/j.jmb.2020.09.006.

8) Park, S.-J., Kim, Y.-G., Park, H.-J.(**2011**) Identification of RNA Pseudoknot-Binding Ligand That Inhibits the -1 Ribosomal Frameshifting of SARS-Coronavirus by Structure-Based Virtual Screening. *J. Am. Chem. Soc.* 133 (26), 10094–10100. https://doi.org/10.1021/ja1098325.

9) Rangan, R., Watkins, A. M., Chacon, J., Kretsch, R., Kladwang, W., Zheludev, I. N., Townley, J., Rynge, M., Thain, G., Das, R., (**2021**) De Novo 3D Models of SARS-CoV-2 RNA Elements from Consensus Experimental Secondary Structures . *Nucleic Acids Res.* 49 (6), 3092–3108. https://doi.org/10.1093/nar/gkab119.

10) Omar, S. I., Zhao, M., Sekar, R. V., Moghadam, S. A., Tuszynski, J. A., Woodside, M. T., (**2021**) Modeling the Structure of the Frameshift-Stimulatory Pseudoknot in SARS-CoV-2 Reveals Multiple Possible Conformers. *PLoS Comput. Biol.* 17 (1), e1008603. https://doi.org/10.1371/journal.pcbi.1008603.

11) Lan, T. C. T., Allan, M. F., Malsick, L. E., Khandwala, S., Nyeo, S. S. Y., Sun, Y., Guo, J. U., Bathe, M., Griffiths, A., Rouskin, S., (**2021**) Insights into the Secondary Structural

Ensembles of the Full SARS-CoV-2 RNA Genome in Infected Cells. *bioRxiv* 2020.06.29.178343. https://doi.org/10.1101/2020.06.29.178343.

12) Plant, E. P., Pérez-Alvarado, G. C., Jacobs, J. L., Mukhopadhyay, B., Hennig, M., Dinman, J. D., (**2005**) A Three-Stemmed MRNA Pseudoknot in the SARS Coronavirus Frameshift Signal. *PLOS Biol.* 3 (6), e172.

13) Plant, E. P., Rakauskaite, R., Taylor, D. R., Dinman, J. D., (**2010**) Achieving a Golden Mean: Mechanisms by Which Coronaviruses Ensure Synthesis of the Correct Stoichiometric Ratios of Viral Proteins. *J. Virol.* 84 (9), 4330–4340. https://doi.org/10.1128/JVI.02480-09.

14) Plant, E. P., Sims, A. C., Baric, R. S., Dinman, J. D., Taylor, D. R., (**2013**) Altering SARS Coronavirus Frameshift Efficiency Affects Genomic and Subgenomic RNA Production. *Viruses* 5 (1), 279–294. https://doi.org/10.3390/v5010279.

15) Masters, P. S., (**2006**) The Molecular Biology of Coronaviruses, *Academic Press.* 66, 193–292. https://doi.org/https://doi.org/10.1016/S0065-3527(06)66005-3.

16) Brierley, I., Ramos, F. J. D., (**2006**) Programmed ribosomal frameshifting in HIV-1 and the SARS–CoV, *Virus Research*. 119, i. 1, p. 29-42, https://doi.org/10.1016/j.virusres.2005.10.008

17) Chang, K.-C., (**2012)** Revealing -1 Programmed Ribosomal Frameshifting Mechanisms by Single-Molecule Techniques and Computational Methods. *Comput. Math. Methods Med.*, 569870. https://doi.org/10.1155/2012/569870.

18) Atkins, J. F., Loughran, G., Bhatt, P. R., Firth, A. E., Baranov, P. V. (**2016**) Ribosomal Frameshifting and Transcriptional Slippage: From Genetic Steganography and Cryptography to Adventitious Use. *Nucleic Acids Res*. 44 (15), 7007–7078. https://doi.org/10.1093/nar/gkw530.

19) Sun, Yu., Abriola, L., Surovtseva, Y. V., Lindenbach, B. D., Guo, J. U., (**2020**) Restriction of SARS-CoV-2 Replication by Targeting Programmed -1 ribosomal Frameshifting in Vitro, *BioRvx*, doi.org/10.1101/2020.10.21.349225

20) Lin, Z., Gilbert, R. J. C., Brierley, I., (**2012**) Spacer-Length Dependence of Programmed -1 or -2 Ribosomal Frameshifting on a U6A Heptamer Supports a Role for Messenger RNA (mRNA) Tension in Frameshifting. *Nucleic Acids Res.* 40 (17), 8674–8689. https://doi.org/10.1093/nar/gks629.

21) Ishimaru, D., Plant, E. P., Sims, A. C., Yount, B. L. J., Roth, B. M., Eldho, N. V, Pérez-Alvarado, G. C., Armbruster, D. W., Baric, R. S., Dinman, J. D., Taylor, D. R., Hennig, M., (**2013**) RNA Dimerization Plays a Role in Ribosomal Frameshifting of the SARS Coronavirus. *Nucleic Acids Res.* 41 (4), 2594–2608. https://doi.org/10.1093/nar/gks1361.

22) Wu, B., Zhang, H., Sun, R., Peng, S., Cooperman, B. S., Goldman, Y. E., Chen, C., (**2018**) Translocation Kinetics and Structural Dynamics of Ribosomes Are Modulated by

the Conformational Plasticity of Downstream Pseudoknots. *Nucleic Acids Res.* 46 (18), 9736–9748. https://doi.org/10.1093/nar/gky636.

23) Caliskan, N., Katunin, V. I., Belardinelli, R., Peske, F., Rodnina, M. V., (**2014**) Programmed –1 Frameshifting by Kinetic Partitioning during Impeded Translocation. *Cell* 157 (7), 1619–1631. https://doi.org/https://doi.org/10.1016/j.cell.2014.04.041.

24) Caliskan, N., Peske, F., Rodnina, M. V., (**2015**) Changed in translation: mRNA recoding by −1 programmed ribosomal frameshifting *Cell Press Trends in Biochemical Sciences* Vol. 40. No. 5 256 – 274. https://doi.org/10.1016/j.tibs.2015.03.006

25) Dinman, J. D., (**2012**) Mechanisms and Implications of Programmed Translational Frameshifting. *Wiley Interdiscip. Rev. RNA* 3 (5), 661–673. https://doi.org/10.1002/wrna.1126.

26) McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., Read, R. J., (**2007**) Phaser Crystallographic Software. *J. Appl. Crystallogr.* 40 (4), 658–674. https://doi.org/10.1107/S0021889807021206.

27) Halma, M. T. J., Ritchie, D. B., Cappellano, T. R., Neupane, K., Woodside, M. T., (**2019**) Complex Dynamics under Tension in a High-Efficiency Frameshift Stimulatory Structure. *Proc. Natl. Acad. Sci.* 116 (39), 19500 LP – 19505. https://doi.org/10.1073/pnas.1905258116.

28) Ritchie, D. B., Foster, D. A. N., Woodside, M. T., (**2012)** Programmed −1 Frameshifting Efficiency Correlates with RNA Pseudoknot Conformational Plasticity, Not Resistance to Mechanical Unfolding. *Proc. Natl. Acad. Sci.* https://doi.org/10.1073/pnas.1204114109.

29) Halma, M. T. J., Ritchie, D. B., Woodside, M. T., (**2021**) Conformational Shannon Entropy of MRNA Structures from Force Spectroscopy Measurements Predicts the Efficiency of -1 Programmed Ribosomal Frameshift Stimulation. *Phys. Rev. Lett.* 126 (3), 38102. https://doi.org/10.1103/PhysRevLett.126.038102.

30) Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G., Jiang, T., (**2020**) Genome Composition and Divergence of the Novel Coronavirus (2019-NCoV) Originating in China. *Cell Host Microbe* 27 (3), 325–328. https://doi.org/10.1016/j.chom.2020.02.001.

31) Wacker, A., Weigand, J. E., Akabayov, S. R., Altincekic, N., Bains, J. K., Banijamali, E., Binas, O., Castillo-Martinez, J., Cetiner, E., Ceylan, B., Chiu, L.-Y., et. el. Zetzsche, H., (**2020**) Secondary Structure Determination of Conserved SARS-CoV-2 RNA Elements by NMR Spectroscopy. *Nucleic Acids Res.* 48 (22), 12415–12435. https://doi.org/10.1093/nar/gkaa1013.

32) (32). Schlick, T., Zhu, Q., Dey, A., Jain, S., Yan, S., Laederach, A., (**2021**) To Knot and Not: Multiple Conformations of the SARS-CoV-2 Frameshifting RNA Element. *bioRxiv* April https://doi.org/10.1101/2021.03.31.437955.

33) Kelly JA, Olson AN, Neupane K, Munshi S, San Emeterio J, Pollack L, Woodside MT, Dinman JD., (**2020**) Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). J Biol Chem. Jul 31,295(31):10741-10748. doi: 10.1074/jbc.AC120.013449. Epub 2020 Jun 22. PMID: 32571880, PMCID: PMC7397099.

34) Manfredonia, I., Nithin, C., Ponce-Salvatierra, A., Ghosh, P., Wirecki, T. K., Marinus, T., Ogando, N. S., Snijder, E. J., van Hemert, M. J., Bujnicki, J. M., Incarnato, D., (**2020**) Genome-Wide Mapping of SARS-CoV-2 RNA Structures Identifies Therapeutically-Relevant Elements. *Nucleic Acids Res.* 48 (22), 12436–12452. https://doi.org/10.1093/nar/gkaa1053.

35) Schlick T, Zhu Q, Jain S, Yan S., (**2021**) Structure-altering mutations of the SARS-CoV-2 frameshifting RNA element. Biophys J. Mar 16,120(6):1040-1053. doi: 10.1016/j.bpj.2020.10.012. Epub 2020 Oct 21. PMID: 33096082, PMCID: PMC7575535

36) Koldobskaya, Y., Duguid, E. M., Shechner, D. M., Suslov, N. B., Ye, J., Sidhu, S. S., Bartel, D. P., Koide, S., Kossiakoff, A. A., Piccirilli, J. A., (**2011**) Portable RNA Sequence Whose Recognition by a Synthetic Antibody Facilitates Structural Determination. *Nat. Struct. Mol. Biol.* 18 (1), 100–106. https://doi.org/10.1038/nsmb.1945.

37) Koirala, D., Lewicka, A., Koldobskaya, Y., Huang, H., Piccirilli, J. A., (**2020**) Synthetic Antibody Binding to a Preorganized RNA Domain of Hepatitis C Virus Internal Ribosome Entry Site Inhibits Translation. *ACS Chem. Biol*. 15 (1), 205–216. https://doi.org/10.1021/acschembio.9b00785.

38) Liebschner, D., Afonine, P. V, Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V, Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J., Adams, P. D. (**2019**) Macromolecular Structure Determination Using X-Rays, Neutrons and Electrons: Recent Developments in Phenix. Acta Crystallogr. Sect. D, *Struct. Biol.* 75 (Pt 10), 861–877. https://doi.org/10.1107/S2059798319011471.

39) Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B. 3rd, Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., Richardson, D. C., (**2018**) MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Sci.* 27 (1), 293–315. https://doi.org/10.1002/pro.3330.

40) Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., Adams, P. D., (**2012**) Towards Automated Crystallographic Structure Refinement with Phenix.Refine. Acta Crystallogr. *Sect. D Biol. Crystallogr.* 68 (4), 352–367. https://doi.org/10.1107/S0907444912001308.

41) Emsley, P., Lohkamp, B., Scott, W. G., Cowtan, K., (**2010**) Features and Development of Coot. *Acta Crystallogr. D. Biol. Crystallogr.* 66 (Pt 4), 486–501. https://doi.org/10.1107/S0907444910007493.

42) Gholamalipour Y, Karunanayake A. M., Martin C. T., (**2018**) 3' End Additions by T7 RNA Polymerase are RNA Self-Templated, Distributive and Diverse in Character-RNA-Seq Analyses. *Nucleic Acids Res.* 46(18):9253-9263. doi:10.1093/nar/gky796

43) Zhang, K., Zheludev, I. N., Hagey, R. J., Wu, M. T.-P., Haslecker, R., Hou, Y. J., Kretsch, R., Pintilie, G. D., Rangan, R., Kladwang, W., Li, S., Pham, E. A., Bernardin-Souibgui, C., Baric, R. S., Sheahan, T. P., D Souza, V., Glenn, J. S., Chiu, W., Das, R., (**2020**) Cryo-Electron Microscopy and Exploratory Antisense Targeting of the 28-KDa Frameshift Stimulation Element from the SARS-CoV-2 RNA Genome. *bioRxiv*. 2020.07.18.209270. https://doi.org/10.1101/2020.07.18.209270.

44) Chen, H., Meisburger, S. P., Pabit, S. A., Sutton, J. L., Webb, W. W., Pollack, L., (**2012**) Ionic Strength-Dependent Persistence Lengths of Single-Stranded RNA and DNA. *Proc. Natl. Acad. Sci.* 109 (3), 799 LP – 804. https://doi.org/10.1073/pnas.1119057109.

45) Iserman, C., Roden, C., Boerneke, M., Sealfon, R., McLaughlin, G., Jungreis, I., Park, C., Boppana, A., Fritch, E., Hou, Y. J., Theesfeld, C., Troyanskaya, O. G., Baric, R. S., Sheahan, T. P., Weeks, K., Gladfelter, A. S., (**2020**) Specific Viral RNA Drives the SARS CoV-2 Nucleocapsid to Phase Separate. *bioRxiv* 2020.06.11.147199. https://doi.org/10.1101/2020.06.11.147199.

46) Harrington, H. R., Zimmer, M. H., Chamness, L. M., Nash, V., Penn, W. D., Miller, T. F., Mukhopadhyay, S., Schlebach, J. P., (**2020**) Co-translational folding stimulates programmed ribosomal frameshifting in the alphavirus structural polyprotein, *Journal of Biological Chemistry*, v. 295, i. 20, p. 6798-6808, https://doi.org/10.1074/jbc.RA120.012706.

47) Huston, N. C., Wan H., Strine, M. S., Tavares, R. C. A., Wilen, C. B., Pyle, A. M., (**2021)** Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms, *Molecular Cell*, 81, 584–598 doi.org/10.1016/j.molcel.2020.12.041

48) Chen, G., Chang, K.-Y., Chou, M.-Y., Bustamante, C., Tinoco, I., (**2009**) Triplex Structures in an RNA Pseudoknot Enhance Mechanical Stability and Increase Efficiency of –1 Ribosomal Frameshifting. *Proc. Natl. Acad. Sci.* 106 (31), 12706 LP – 12711. https://doi.org/10.1073/pnas.0905046106.

49) Giedroc, D. P., Cornish, P. V., (**2009**) Frameshifting RNA Pseudoknots: Structure and Mechanism. Virus Res. 139 (2), 193–208. https://doi.org/10.1016/j.virusres.2008.06.008.

50) Neupane, K. P., Zhao, M., Hoffer, N. Q., Lyons, A., Munshi, S., Ritchie, D., Woodside, M. T., (**2021**) Structural Dynamics of SARS-CoV-2 Frameshift Signal Studied by Single-

Molecule Force Spectroscopy Reveal Topologically Distinct Conformers. *Biophys. J.* 120 (3), 314a-314a. https://doi.org/10.1016/j.bpj.2020.11.1987.

51) Xu, X., Zhao, P., Chen, S.-J., (**2014**) Vfold: A Web Server for RNA Structure and Folding Thermodynamics Prediction. *PLoS One* 9 (9), e107504. https://doi.org/10.1371/journal.pone.0107504.

52) Rio, D. C., (**2013**) Expression and Purification of Active Recombinant T7 RNA Polymerase from E. Coli. *Cold Spring Harb. Protoc.* 2013 (11). https://doi.org/10.1101/pdb.prot078527.

53) Batey, R. T., Kieft, J. S., (**2016**) Soaking Hexammine Cations into RNA Crystals to Obtain Derivatives for Phasing Diffraction Data. *Methods Mol. Biol.* 1320, 219–232. https://doi.org/10.1007/978-1-4939-2763-0_14.

54) The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

55) Nilsen T. W., (2014) 3'-end labeling of RNA with [5'-32P]Cytidine 3',5'-bis(phosphate) and T4 RNA ligase 1. Cold Spring Harb Protoc, (4):444-6. doi: 10.1101/pdb.prot080713. PMID: 24692494.

56) Cold Spring Harb Protoc (2014) Cold Spring Harbor Laboratory Press doi:10.1101/pdb.rec082123

57) Grosjean, H., Keith, G., Droogmans, L., (2004) Detection and quantification of modified nucleotides in RNA using thin-layer chromatography. Methods Mol Biol.,265:357-91. doi: 10.1385/1-59259-775-0:357. PMID: 15103084.

58) Cao, C., Cai, Z., Xiao, X., et al., (2021) The architecture of the SARS-CoV-2 RNA genome inside virion. Nat Commun 12, 3917 https://doi.org/10.1038/s41467-021-22785-x

59) Wu, Q., Zhang, Y., Lü, H., Wang, J., He, X., Liu, Y., Ye, C., Lin, W., Hu, J., Ji, J., Xu, J., Ye, J., Hu, Y., Chen, W., Li, S., Wang, J., Wang, J., Bi, S., Yang, H., (2003) The E protein is a multifunctional membrane protein of SARS-CoV. Genomics Proteomics Bioinformatics. May,1(2):131-44. doi: 10.1016/s1672-0229(03)01017-9. PMID: 15626343, PMCID: PMC5172412.

3. Chapter 3: The Crystal Structure of The Cap-Independent Translation Enhancer PEMV2 PTE solved to 2.75Å

3.1 Introduction

In eukaryotes, processed mRNA is covalently modified with an inverted 5' N7 methyl guanine triphosphate cap and 3' poly A tail, which serve to identify the mRNA to the eukaryotic translation initiation factors[3]. Protein translation begins when eukaryotic initiation factor eIF4E recognizes the 5' cap structure[3]. From there, eIF4E recruits eIF4G forming the eIF4F complex which in turn binds the poly A binding protein[3]. The poly A binding protein then binds the poly A tail and circularizes the mRNA[3]. The circularized mRNA with the eIF4F complex is bound such that the 5' end is now recognizable to the 43S pre-initiation complex, which subsequently assembles the Ribosome at the AUG start site[3]. The 5' M^7GpppN cap is a crucial beacon that distinguishes mRNA from other cellular RNAs and is only added to the mRNA during post-transcriptional processing in the nucleus[3]. This presents a problem for parasitic viruses, whose mRNA is generated in the cytoplasm[4,11]. These viral mRNAs necessarily lack the 5' cap structure needed for cap-dependent translation[4,11]. Instead of encoding capping enzymes in their small genomes, viruses have evolved cap-independent translation mechanisms[4,11]. In some cases, this cap-independent transition mechanism involves covalently modifying viral RNAs with the Viral Protein genome-linked protein (VPg)[15]. In other cases, structured RNA elements in the 5' or 3' UTR bind and hijack the host cell translation initiation machinery [4,11].

Internal Ribosome Entry Sites (IRESs) are perhaps the most common example of cap-independent translation elements found in viral RNAs. IRESs are large multi-domain structured RNAs located within the 5'UTR of viral genomes[13]. IRESs function as cis-acting elements that recruit the initiation factors and/or the ribosomal subunits directly to the start of the open reading

frame[13]. They do this by using structural motifs that often mimic the host's factors native binding targets[13]. In isolation from their viral 5' UTRs, IRES elements act as autonomous elements capable of inducing translation of any downstream open reading frame[13]. The mechanisms and structures of IRESs are well characterized compared to other classes of functional RNAs, in part due to their prevalence in well studied animal and human viruses such as Human Immune Deficiency Virus (HIV) and Hepatitis C Virus (HCV)[13]. IRESs are notably less common in plant viruses, instead these viruses more often deploy an alternative cap independent translation mechanism[13].

Recently, structured RNA translation elements were discovered in the 3' UTRs of plant viruses[19]. These 3' Cap Independent Translation Enhancers (CITEs) are still being identified in plant viruses through deep sequencing and phylogenetic analysis. Although they are expected to be utilized more broadly, so far CITEs have only been identified in viruses belonging to the *Tombusviridae* Family and the *Luteovirus* genus[24]. While it is possible that 3' CITEs could exist in other viruses or even in cellular RNAs so far none have been identified outside of plant viruses[24].

Like IRESs, CITEs are also structured, cis-acting, autonomous RNA elements that induce translation of viral RNA genomes by recruiting various host translation initiation factors or components. However, CITEs are generally smaller than IRESs, usually only spanning 100-200 nucleotides in length and generally only encompassing a single domain[24]. Their complex tertiary structures are capable of both binding host translation factors and circularizing the RNA to bring the ribosome to the 5' end of the RNA to scan for the AUG start site[24]. CITEs often operate independently, with one CITE per 3' UTR. However, two instances of virus harboring multiple CITEs have recently been discovered[24]. In the first case of this phenomena Pea Enation Mosaic Virus 2 (PEMV2), was found to use three cites at once one to bind eIF4E, another to recruit the 80S ribosomal subunit and a third to circularize the RNA[7]. Each of these CITEs functions

independently of one another[7]. In the second case of a virus harboring multiple CITEs, horizontal gene transfer between two coinfecting viruses lead to the acquisition of a second functional, and resistance breaking, CITE. This finding raises the possibility that CITEs are readily exchanged between viruses, enabling rapid viral evolution and immune evasion[4,14].

Viral epidemics and pandemics have always threatened food security and a growing global population only magnifies their devastating effects[11]. Additionally, the proliferation of genetically modified crops drives mono-culture farming, a practice that leaves crops particularly susceptible to disease due to an absence of natural genetic variation[11]. Climate change also increases plant's susceptibility to viral infection by upregulating abiotic stress response gene expression pathways[10]. Despite the increasing threat plant viruses pose, study of their molecular biology lags behind that of human pathogens, which leaves society less prepared to respond to agricultural disasters[11]. As crucial components to many plant viral replication mechanisms, CITEs are an excellent target to study using structural biology with the long-term goal of developing drugs to target them or biotechnologies to evade them and improve agricultural resilience[18,24,25]. Understanding the structural mechanism of translation factor recruitment of each CITE could guide the development of genetically modified crops that use mutant transcription factors resistant to CITE hijacking [18,24,25].

There are seven distinct classes of CITEs; each adopt different well conserved structures and employ different mechanisms of host factor recruitment[24]. The first CITE was identified in Satellite Tabaco Necroses Virus (STNV) and represents the Translation Enhancer Domain (TED) class. TED CITEs form a long stem loop structure with several internal bulges and terminal loops[5]. This class binds eIF4F with its main contacts forming with eIF4E[5]. The Barley Yellow Dwarf Virus Translation Enhancer (BTE) class is found in the genera Luteovirus (Family Luteoviridae),

Necrovirus, Dianthovirus, Alphanecrovirus, betanecroviruses, and Umbravirus (Family Tombusviridae). BTE CITEs adopt a cloverleaf or cruciform structure with two to four helices branching from a central stem[26]. The sequence of this central stem is critical to binding eIF4G to recruit the eIF4F complex[26]. BTE's also bind the 18S rRNA through sequence complementarity[26]. The Y-Shaped Structure (YSS) CITE class is exclusive to tombusviruses[2]. YSS CITEs bind eIF4F or eIF4(Iso)F and form a three stemmed structure with one base stem and two branches of roughly equal length[2]. The T-Shaped Structure (TSS) CITE resembles a tRNA and is found in Carmovirus and Umbravirus viruses of the Tombusviridae family[22]. TSS CITEs recruit the 60S or 80S subunit of the ribosome by mimicking a tRNA and interacting with the P-site in the ribosome[22]. They are also capable of circularizing the gRNA. This is the only CITE with an experimentally determined representative structure. This TSS structure closely matched secondary structural predictions and was solved using solution NMR (PDB accession code 2KRL)[30]. There are two TSS CITES found along site a PTE class CITE in the PEMV2 virus 3' UTR [4,6,7]. The most newly discovered CITE CABYV-Xinjiang-like translation element (CXTE) was discovered in a resistance breaking strain of the Melon Necrotic Spot Virus[14]. This CITE adopts an I-shaped secondary structure with two helices emerging from a central bulge and binds both wild type eIF4E and a CITE resistant eIF4E mutant[12].

The seventh class of CITE is the Panicum Mosaic Virus Translation Element or PTE, first identified in Panicum Mosaic Virus and then identified in PEMV2[1,28]. PTEs have been discovered in Panicoviruses, Aureviruses, Carmoviruses and Umbravirsues[24]. They adopt a T-shaped structure which binds plant, and in one case mammalian, eIF4E[12]. PTEs cannot bind any other eIF4 proteins like eIF4G or eIF4(Iso)E, but PTE affinity for eIF4E is enhanced in the presence of eIF4G (Figure 3.1)[28]. These PTE elements are among the most well biochemically characterized

CITE, making them an ideal target for structural studies. A PTE's ability to induce cap independent translation, or compete with the native cap structure relates directly to its affinity for eIF4E as determined by Wang et al. 2009 [28]. PTE activity can therefore be detected and characterized through a translation inhibition assay, where increasing concentrations of an isolated PTE construct competes for eIF4E against a PTE-regulated luciferase reporter in wheat germ extract[27,28].

Chemical structure probing with SHAPE and homology analysis showed that PTEs adopt a T-shaped secondary structure[27]. A basal stem is made up of three coaxially stacked helices, Helix 1 (H1), Helix 2 (H2) and Helix 3 (H3), separated by small internal bulges (Figure 3.1A)[27]. Between H2 and H3 there is an approximately 10 nucleotide long asymmetrical G rich bulge on the 5' side known as the G-domain[27]. At the top of H3 a three-way junction branches the structure into stem-

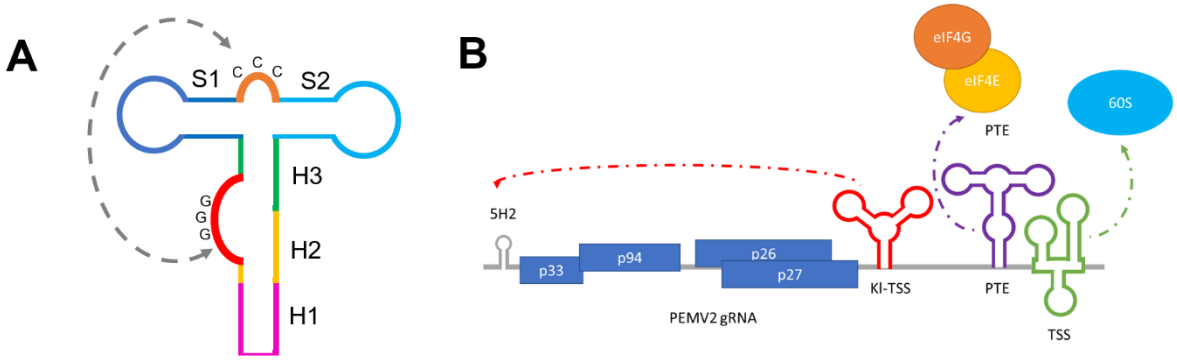


Figure 3.1. A. General secondary structure of PTE class of CITEs using nomenclature established in previous literature, regions colored to match coloring of regions used throughout. B. Genomic organization of Pea Enation Mosaic Virus 2 illustrating the relative positions of the 3' UTR CITEs, their binding partners, and the 5' SH2 hairpin structure used for circularization.

loop1 (SL1) and stem-loop 2 (SL2). SL1's 3' end is connected to SL2's 5' end by a C rich bulge known as the C-domain[27]. The length of SL1 and SL2 vary across PTEs[27]. In almost all PTEs, with the exception of PEMV2, Loop 1 shares sequence complementarity with a hairpin (5H2) in the 5'

UTR[27]. The interaction between L1 and 5H2 circularizes the RNA and helps recruit the preinitiation complex in the same way as eIF4E and poly-A binding protein[19].

Generally, the G-domain bulge contains a poly G tract that exhibits complementarity to the poly C tract in the C-domain; mutations to either of these tracts eliminate or severely reduce PTE induced translation, translation inhibition, and eIF4E binding (Figure 3.1A)[27,28]. The C-domain is largely unreactive to SHAPE structure probing[27]. It was then hypothesized that the G- and C-domains form a pseudoknot in the PTE[27]. Curiously, the C- and G-domains from different PTEs cannot be mutationally transferred into new PTEs and retain their function, suggesting their eIF4E binding activity is dependent on the structural context of their respective PTEs[27]. The G-domain of every PTE examined by SHAPE probing contains a single G which is "hyper-modified" in the presence of magnesium, suggesting it is in a solvent exposed position along the structure[27]. This so called hyper-modified G is also protected from modification in the presence of eIF4E, suggesting that this residue is the site of eIF4E binding. Using this assumption, a tertiary model of the PEMV2 and PMV PTEs were made with a single G of the G-domain flipped out and docked inside eIF4E's cap binding pocket[26].

Pea enation mosaic virus 2 (PEMV2) has three CITEs in its 3' UTR; one PTE and two TSS elements (Figure 3.1 B)[6,7]. The ks-TSS element contributes to genome circularization during translation by binding a small hairpin structure in the 5' UTR, thus alleviating the PEMV2 PTE's role of genome circularization[6,7]. Interestingly, PEMV2 PTE is the only PTE which does not use its L1 to hybridize with the 5' UTR to circularize the RNA[19]. The other TSS element in the PEMV2 3' UTR binds the 60S ribosome subunit. Even in the context of the full length 3' UTR containing all three CITEs, mutations to critical regions of the PTE element reduce translation efficiency, and eliminate PTE eIF4E's 58nM binding affinity [27,28]. Mutations to alter the sequence of either loop

1 or 2 do not alter eIF4E affinity or translation inhibition, however, mutating residues in the G- and C-domains reduce translation efficiency by 95%[28]. In addition, extensive chemical probing identifies a homogeneous monodisperse conformation for this element[27,28].

Using the extensive biochemical characterization as a guide, we are able to generate a crystallization construct of the PEMV2 PTE to determine its structure experimentally. To crystallize RNAs which are usually recalcitrant to crystallization, we apply chaperone assisted RNA crystallography[9]. By mutating a solvent exposed loop to a Fab-binding epitope, we can create a complex of antibody fragment and RNA element which is much more crystallizable than RNA alone[9]. The antibody fragment not only stabilizes the RNA, but also serves as a model for molecular replacement phasing. While considerable evidence exists to support the PTE secondary structure and the C- and G-domain pseudoknot, we do not know how the PTE forms a structure that eIF4E can bind with such high affinity to compete with the native 5' cap structure. This crystal structure of the PTE has now revealed a novel fold that creates an eIF4E binding site with a single solvent exposed G flipped out from the G-domain.

3.2 Results

3.2.1 Construct design and validation

To create a version of the PEMV2 PTE element, that was most likely to crystallize in the native conformation, we began with a construct that represented nucleotides 3820-3907 in the PEMV2 viral genome[28]. This region was identified as the minimal sequence necessary to induce cap independent translation [27]. To create continuous helix at the base of the construct, residue C3822 was mutated to a U to create a pair with the opposite A3906. This construct was also designed to contain an extra G at the 5' end to improve transcriptional efficiency in an in vitro

transcription reaction. Additionally, in our hands, RNA constructs with a 5' overhang G can crystalize more efficiently than constructs which end in a pair because the overhanging G can form a crystal packing interaction with a mirrored copy of the RNA element.

To create a version of the PEMV2 PTE element which can bind our Fab-crystallization chaperone, one of the solvent exposed loops needed to be mutated to the Fab-binding sequence in such a way that it did not disrupt the native fold or function. Previous mutational assays have determined that eIF4E binding activity is independent to loop 1 sequence, therefore loop 1 can tolerate a mutation to our BL3-6 Fab-binding epitope, the AAACA penta-loop closed by a GC pair[28]. Similarly, loop 2 was also shown to tolerate mutations without altering PTE function, however mutating loop 2 to the fab binding epitope resulted in a construct which could not bind Fab efficiently. Therefore, crystallization trials were only attempted using the mutated L1 construct, this construct is hence forth referred to PEMV2 BL3-6.

The PEMV2 BL3-6 construct was confirmed to retain the same fold as the wildtype construct using an EMSA gel shift assay under non-denaturing conditions. Refolded wild type and BL3-6 minimal RNA constructs ran at the same positions in the native gel supplemented with an effective concentration of 10 mM Magnesium Chloride. In the presence of a molar equivalent amount of eIF4E both constructs shifted quantitatively to the bound state. The same shift did not occur for a non-PTE RNA construct, indicating this interaction is not the result of non-specific RNA binding activity of eIF4E. From this one can conclude that the BL3-6 construct folds and functions in a way that closely resembles the wild type PEMV2 PTE such that it retains the native eIF4E binding activity.

3.2.2 Crystallization and model building

PEMV2 BL3-6 was refolded and incubated with a 1.1 molar equivalent of a mutant Fab BL3-6[1] to forma complex of RNA and Fab. The complex was then concentrated to a final concentration of 6 mg/mL RNA. Four sets of 96 well screens (Hampton PEG Ion, Natrix, Index, Jena Bioscence Nuc-Pro) were set using a 1:1 ratio of complex to well solution in the hanging drop format. After two weeks at room temperature crystals grew in 5mM MgCl2, 50mM HEPES pH 7.0 25% PEG 550. These crystals were further optimized with an additive screen and the final crystals grew with 2% Benzamidine Hydrochloride. Crystals were looped and cryo-protected with well solution containing 50% glycerol and frozen in liquid nitrogen. A data set was collected out to 2.1 Å, but the data were highly anisotropic. After anisotropy correction, the data set was usable out to 2.5 Å, but concerns about the low figure of merit for the highest resolution shells justified truncating the data further to 2.75 Å. The structure was solved using molecular replacement of the Fab model in Phenix. Only one copy of the RNA and Fab were discovered in the P1211 space group. Using the initial phases from the molecular replacement solution, the RNA was able to be built into the emerging density after multiple rounds of refinement. Simulated annealing and composite omit maps were applied to the model during refinement and building to combat accumulating model bias as identified by a growing gap between the R_{work} and R_{free} values. The final structure was solved to 2.75 Å with an R_{work} of 25% and an R_{free} of 29%. Additional statistics reported in Table 3.1.

[1] Crystal trials were set with wild type BL3-6 and mutants with surface entropy reducing point mutations as well. The crystal which gave the highest resolution diffraction presented here actually contained a point mutation at residue 170 form lysine to alanine however this mutation is not believed to have contributed to the crystal packing. Despite this, crystals of PEMV2 with wild type Fab BL3-6 typically do not diffract below 3Å

Data collection	
Wavelength	0.9792
Resolution range	44.08 - 2.75 (2.848 - 2.75)
Space group	P 1 21 1
Unit cell	63.4924 65.579 90.1035 90 110.347 90
Total reflections	69577 (6704)
Unique reflections	18179 (1194)
Multiplicity	3.8 (3.7)
Completeness (%)	88.41 (65.75)
Mean I/sigma(I)	14.37 (2.36)
Wilson B-factor	32.90
R-merge	0.1921 (0.5634)
R-meas	0.2195 (0.6542)
CC1/2	0.978 (0.821)
CC*	0.994 (0.95)
Refinement	
Reflections used in refinement	16141 (1194)
Reflections used for R-free	1464 (113)
R-work	0.2574 (0.4900)
R-free	0.2954 (0.5550)
CC(work)	0.806 (0.344)
CC(free)	0.748 (0.240)
Protein residues	440
RMS(bonds)	0.002
RMS(angles)	0.59
Ramachandran favored (%)	96.33
Ramachandran allowed (%)	3.67
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	0.27
Clashscore	8.64
Average B-factor	91.71

Table 3.1. PEMV2 PTE Structure X-ray crystallography data collection and refinement statistics

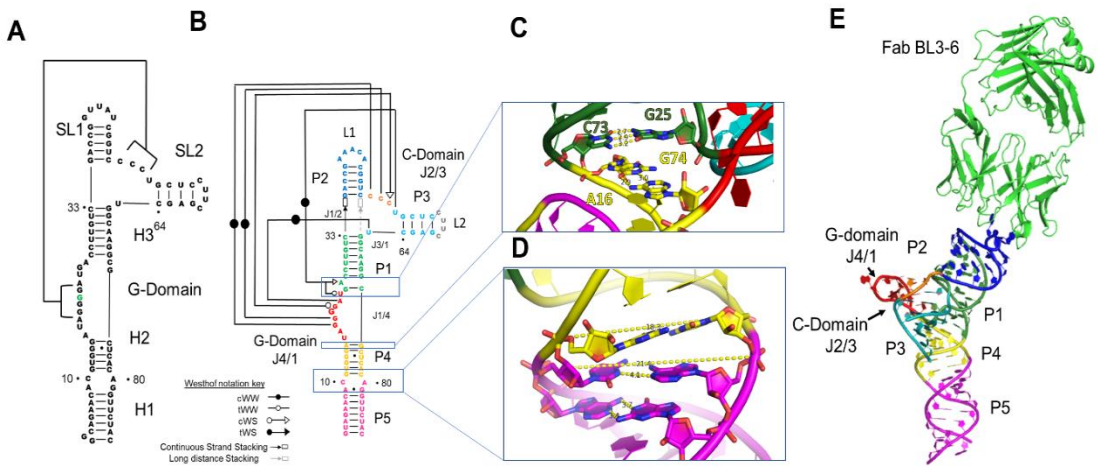


Figure 3.2. A) PEMV2 PTE secondary structure determined by Wang et al 2009 with helix and stem nomenclature established therein. B) PEMV2 PTE secondary structure determined by the crystal structure, interactions denoted using the Westhof notation[31] and illustrations Pair and Junction nomenclature revision illustrated. C) Zoom in on P4 P1 helical stack region illustrating the relative rotation of P1(green) to P4(yellow) D) Zoom in on P5 P4 helical stack region including the small bulge created by the AG pair E) Overall arrangement of PEMV2 PTE and Fab BL3-6 in the crystal structure note loop 2 showed no density and was not modeled in the final solution.

3.2.3 Structural Analysis

The predicted secondary structure of the PEMV2 PTE is shown in Figure 3.2A while our crystal structure derived secondary structure is shown in (Figure 3.2 B). Given the complex architecture of the tertiary structure a new naming system for the helices was required. The P, which stands for paired regions and J which stands for Junctions is the nomenclature used to categorize three way and four-way junctions[29]. Reframing our understanding of the PTE fold using the framework of junctions helps clarify this potentially novel motif. H3 is at the center of the junction with all other helices stacked above or below it[27]. This helix defines the central axis of the junction core. Going by nomenclature standards, H3 therefore becomes P1 because stacked above P1 is P2, in this case S1[27,29]. P2 leads into J2/3 which in this case is the C domain bulge

which then leads into P3, formerly S2[27,29]. So, at the core of the junction H3 is P1, S1 is P2 and S2 is P3[27,29]. There are two other helices or paired regions outside the three-way junction. Following the 3' end of P1 we find P4, formerly H2 and below that P5, formerly H1[27,29]. P4 is connected to P1 through junctions of its own[27,29]. What was referred to in previous literature as the G-domain is now J4/1 as it connects the 5' end strand of P4 to P1[27,29]. On the opposite strand to the G-domain bulge J4/1, we have opposite junction, J1/4, connecting the 3' end strand of H1 to H4[27,29].

The global architecture of the PEMV2 BL3-6 crystal structure shown in Figure 3.2B matches the predicted secondary structure shown in Figure 3.1A[27,28]. Beginning at the 5' end, P5 forms though pairing between residues 2-9 and residues 80-86 in our numbering, with the first G of the construct numbering position 1. P5 ends in an uncommon A-G Watson-Crick pair, unaided by waters or ions (Figure 3.2 D). The exocyclic amine of G9 forms a hydrogen bond with the N1 imine group of A81, while G9's N1 imine forms a hydrogen bond with A81's exocyclic amine. This purine-purine pair widens the helical diameter and leaves the following A80 and C10 residues unpaired and spaced 4.1 angstroms apart (Figure 3.2 D). This space contains no ordered waters or ions therefore we can conclude that these residues are not interacting with each other but remain facing one another through stacking interactions of the helices (Figure 3.2 D).

Stacked above P5, P4 forms between residues 11-16 and 47-79. J4/1, otherwise known as the G-domain bulge in previous literature, was predicted to include eleven residues on the 5' end (residues 16-26) and two residues on the 3' end in J1/4 (residues 72-73) [27,28]. However, in the structure, both of the J1/4 residues engage in pairing interactions with residues at both ends of the predicted J4/1 (Figure 3.2 B). G74 forms a purine-purine Watson-Crick pair with A16 to extend and widen P4 at the top. C73 forms a WC pair with G25, which extends P1 at the bottom and consequently subsumes A26 into P3 as a bulged nucleotide (Figure 3.2 C). These interactions leave

no unpaired nucleotides in J1/4 and effectively truncate the G-domain bulge by three nucleotides. P1 is stacked on P4 with the base plane remaining consistent between the two, however the first pair of P1, C73 G25, are oriented perpendicularly to the final pair of H4 (Figure 3.2 C). The P1 P4 stacking forces the 3' and 5' end of the J4/1 G-domain bulge close together with the backbones of each end crossing over each other.

The remainder of the structure matches the biochemically derived secondary structure aside from the slight elongation of P1. P1 was originally thought to begin at residue 27 and 72 but our structure reveals that P1 actually begins at residue G25 which pairs with C73 (Figure 3.2 C), originally thought to be unpaired in J3/2 (Figure 3.2 A) [27,28]. The top of P1 leads into P2 and P3 which form the top of the T-shaped structure (Figure 3.2 B). P2 encompasses nt 34-38 and 44-48 with loop 1 (L1) forming between 39-53. P3 encompasses nt 53-56 and 61-64 with loop 2 (L2) forming between 57-60. It should be noted that the density for the residues in L2 is missing so these have been excluded from our model (Figure 3.2 E). The 3' end of P2 and the 5' end of P3 are bifurcated by J2/3, otherwise known as the C-domain bulge in previous literature, which encompasses residues 49-52. The J3/1 region contains a single residue U65, as predicted[27,28].

3.2.4 The P1, P2, P3 three-way junction

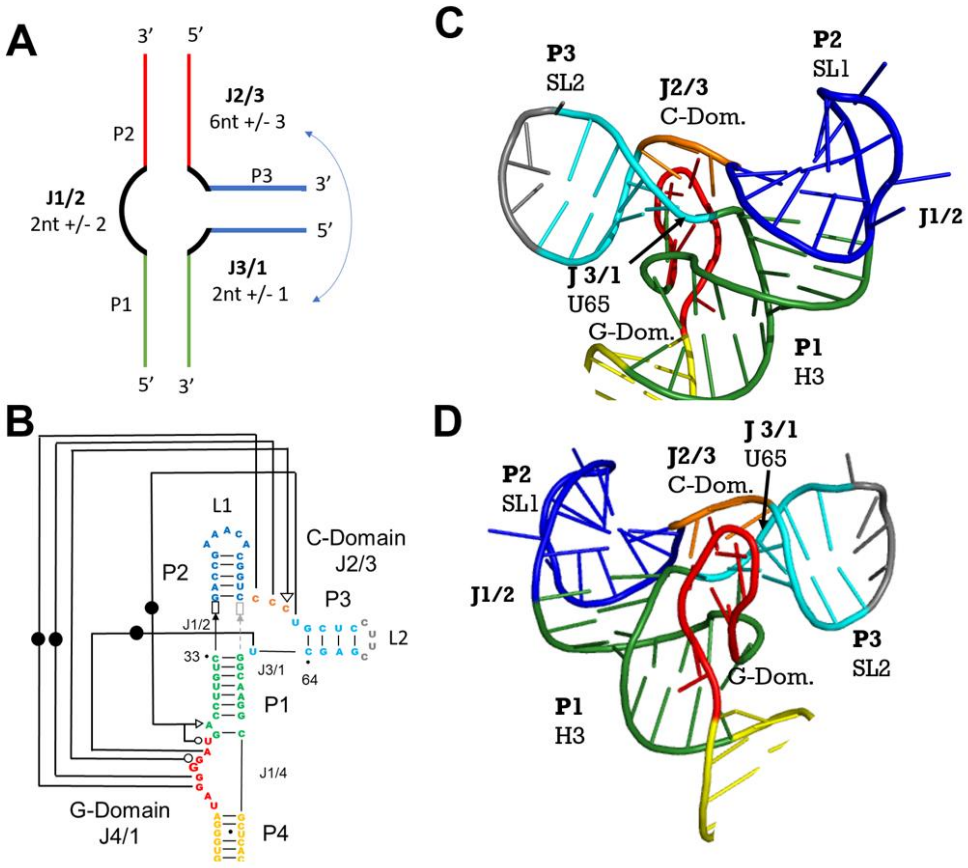


Figure 3.3. A) A type three-way junction diagram illustrating the arrangement of P1, P2 and P3 and the observed lengths of Junctions found in the A type[29]. B) Crystal structure derived secondary structure focused on the long-distance interactions between J4/1 (the G domain) and J2/3 the C domain showing that J2/3 makes long distance interactions outside of the junctions J2/3 would normally interact with in other A type junctions C) Three-dimensional illustration of the three-way junction region with loop 2 included for clarity D) the image shown in C but rotated by 180 degrees.

The lengths of the junctions connecting P1, P2 and P3 and the orientation of the helices define this region as an atypical A type three-way junction[29]. In a three-way junction with the two

longest coaxially stacked paired regions forming P1 and P2 respectively, the J1/2 is typically 0 to 2 residues long (Figure 3.3A). In the case of PEMV2 J1/2 contains no residues (Figure 3.3B). In an A type junction J2/3 is generally 3 to 6 residues long and J3/1 is one or two residues long[29]. In PEMV2 J2/3 is four residues long and J3/1 is one residue long. The orientation of P3 relative to P1 and P2 is also a defining feature of an A type junction[29]. In B type junctions P3 is associated with P1 forming a lambda like conformation and in a C type junction P3 is associated with P2 forming a Y like conformation. The P3 of A type junctions is usually not associated with either helix, and can often be found oriented perpendicularly to the central axis, this is the case for PEMV2 (Figure 3.3C & D). PEMV2's A type junction notably atypical because the junction regions residues do not associate with any part of the junction. Instead J2/3 and J31 make long distance interactions with J4/1, otherwise known as the G-domain (Figure 3.3B, C, D). The arrangement of this junction is essential to forming the eIF4E binding site in the PEMV2 structure as it positions its junction residues to form a stabilizing lattice that allows J2/3, the C-domain, to form an unprecedented parallel strand pairing with J4/1, the G domain.

3.2.5 G- and C-domain interactions form a C-turn structure

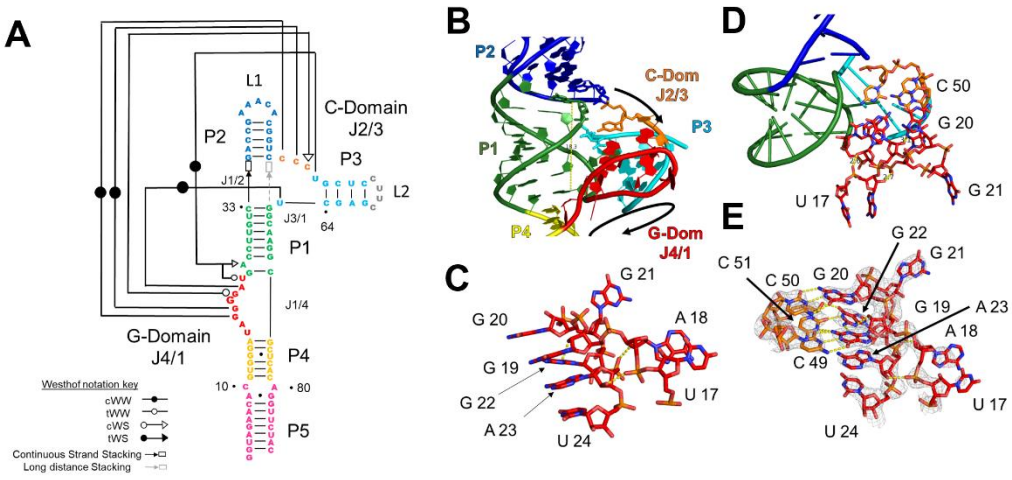


Figure 3.4. A) Crystal structure derived secondary structure with Westhof[31] interaction illustrations. B) G domain ribose zipper shown as cartoon in red and its position relative to the C domain, we propose calling this motif a C-turn structure. C) Ladder like position of the G domain residues each held in place by long distance interactions with the three-way junction, ribose zipper hydrogen bonds shown as dotted yellow lines D) C and G domain positions to highly the solvent exposed residues. E) C and G domains alone shown as sticks with electron density map shown as mesh.

The 5' ends of the C- and G-domain begin 18.3 Å apart because they emerge from opposite ends of the P1, but the arrangement of the three-way junction positions these regions close together so these distal bulges can interact. At the 5' end of the G-domain U17 and A18 are flipped out into solution (Figure 3.4B and E). These residues' base faces lack clear electron density, suggesting disorder in the crystal potentially due to flexibility in the position of these bases. However, the electron density for the backbone atoms of these residues is clearly defined, as illustrated in the electron density map of Figure 3.4E. We observe A18's non bridging phosphate oxygen forming a hydrogen bond with A23's 2'OH. Additionally, A18's 2'OH forms a hydrogen bond with A23's nonbridging phosphate oxygen. These bonds not only restrain the flexibility of the backbone for A18 and U17 but also define J4/1 as a ribose zipper (Figure 3.5C)[23].

Moving through the G-domain from 5’ to 3’, after A18, G19 and G20 form the first two base pairs with the C-domain’s C49 and C50 (Figure 3.4B and E). Instead of forming a standard antiparallel A form helix where G19 pairs with C50 and G20 pairs with C49, we observe unusual parallel cis Watson Crick pairs between G19 and C49 followed by G20 and C50. The parallel pairing of this region may preclude it from meeting the definition of a pseudoknot, therefore we have chosen not to refer to the G- and C-domain structure as a pseudoknot in this work[21]. To achieve the unusual parallel cis Watson Crick pairing, the backbone is flattened and extended creating a wider minor grove into which lower residues pack and stack (Figure 3.4B and E). Together G19-C49 and G20-C50 pairs are nearly adjacent to one another and form the uppermost level of the scaffold structure created in the RNA’s structural core (Figure 3.5A).

Following G20, the G-domain strand inverts on its self, folding back over so that the O4’ of the ribose’s of G20 and G22 point towards one another (Figure 3.4C). This creates an abrupt turn at G21, flipping the G21 base out of the core and into solution (Figure 3.4B, C, E). Taking the three points of the ribose C1’ atoms of G20, G21 and G22 we see that this turn creates a 53-degree angle. Curiously, despite being so highly solvent exposed, the density of G21 is clearly defined (Figure 3.4E). Although its SHAPE sensitivity implied a high degree of flexibility, it is possible that the strain of this turn holds G21 firmly in place away from the core of the PTE element. The order of G21 in the crystal may be additionally aided by a crystal packing interaction with the Fab surface.

Below G21, G22 and C51 tuck back into the space underneath G20 and C49. In Figure 3.5B a top-down view of this region is shown; C50 and C49 G20 and G19 are shown as a transparent layer above G22 A23 C51 and U65 which are shown as opaque. Compared to C50 which faces the opposing G-domain strand, C51 turns inward to the RNA’s structural core such

that its base face is directly below C50. Opposite C51, G22 tilts upward to form a cis WC sugar edge interaction with C51 where G22's keto group and imine group form hydrogen bonds with C51's 2'OH and keto group respectively. G22 and C51 also form additional hydrogen bond interactions with nearby residues. This complex network of hydrogen bonds between the upper and lower levels are shown as dotted yellow lines in Figure 3.5B. We observe G22's exocyclic amine hydrogen bonding to the above G19 N3 imine while its N3 imine hydrogen bonds with G19's 2'OH (Figure 3.4B). On the same base plane, C51's imine bonds with G19's exocyclic amine and C51's exocyclic amine bonds to U65's C4 keto group (Figure 3.5B). However, the interactions which stabilize this unique new turn structure do not stop at the C-domain. Our structure reveals a complex network of multi-base interactions that serve as a stabilizing scaffold structure anchoring the G-domain to the core of the RNA's three-way junction.

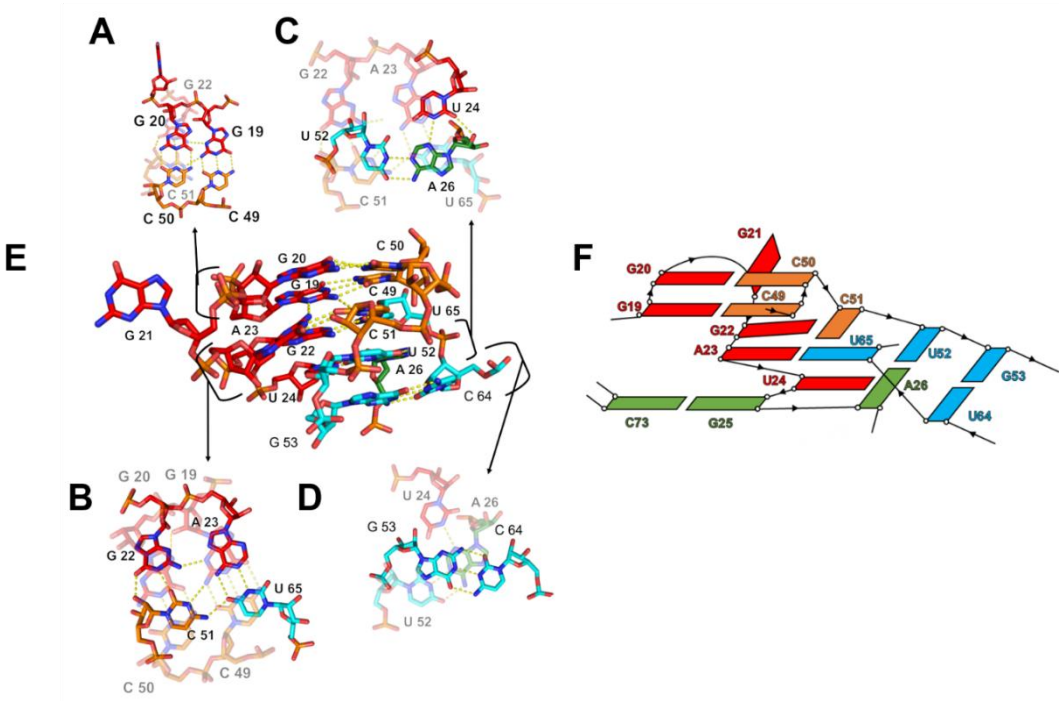


Figure 3.5. The scaffold core's organization A) G20 paired with C50 and G19 paired with C49 shown as opaque with the level below this shown as translucent illustrating the relative positions of G22 and C51 B) G22 paired with C51 SE and A23 paired with U65 in tetra base interaction hydrogen bonds between this tetra base level and the translucent level above it illustrated as dashed lines. The opaque bases in panel A are now illustrated as translucent to retain the relative position information between the two levels. C) The triple base interaction between U52 U24 and A26 shown as opaque with the tetra base interaction above it between G22 A23 C51 and U65 shown as translucent. D) The Di-nucleotide platform at the base of the scaffold between G53 and C64 shown as opaque with the triple base interaction above it shown as translucent. E) The scaffold core formed between the three-way junction and the G domain as viewed from the side illustrating the relative positions of the levels in three dimensions. F) A two-dimensional illustration of the scaffold indicating strand direction, pairing interactions and relative base planes staggered across a 45-degree angle.

3.2.6 Scaffold structure created in the atypical A-type three-way junction long distance interactions.

The arrangement of the C- and G-domains can also be considered as a scaffold structure divided into levels of stacked base planes. Figures 3.5 E and F illustrate the arrangements of the levels we observe in this scaffold core. If we consider the first two G-C pairs of the scaffold the

top level shown as opaque in Figure 3.5 A, the second level below contains G22, A23, C51 and U65 shown as opaque in Figure 3.5 B. If we frame this region as a scaffold, we can conclude that the G23-C51 and A23-U65 pairs form a tetra base interaction stacked below and oriented by interactions with the G19-C49 G20-C50 pairs above it.

U65's involvement in this tetra base network is interesting. We observe U65 stacking below C49 and forming a cis WC interaction with A23 (Figure 3.5 B). In former models U65 was thought to be an unpaired junction between Stem 2 (P3) and H3 (P1). Now we can see how the three-way junction structure has placed these distal residues in close three-dimensional proximity. If we consider U65 as J3/1 of the three-way junction its interactions with the G-domain follow the pattern of interactions observed between J2/3 and the G-domain. In this frame we can say that the all of the junctions' residues pair with residues in the G-domain ribose zipper. Structurally, U65 may serve as an anchor point in this scaffold, but without a measure of the energetic contribution this residue makes to the overall structural stability, we cannot confirm that this is J3/1's role in the structure.

In the layer below the G22 A23 C51 U65 tetra base interaction there is a triple base pair formed between U52, A26, and U24 (Figure 3.5 C). In this level, A26 forms a WC pair with U52, the final residue of J2/3. The final residue of J4/1, U24 forms a Watson-Crick sugar edge interaction with A26. After U24, G25 actually turns away from the J4/1 to begin P1 and pairs with C73 (Figure 3.5 F, Figure 3.2 C). Thereafter, A26 flips back out of P1 to intercalate itself between U65 and U64 to once again pair with U52 and U24. This flipped intercalation of A26 knots the backbone back and forth between the start of P1 and S3 by this base triple interaction flanking either side of G25's pairing in P1.

Beneath the base triple P2 forms a platform for the scaffold stack starting with a simple GC pair between G53 and C64 (Figure 3.5D). This base pair helps orient the base planes of all the layers that stack above it. Determining how this complex network of interactions each contribute to the stability of the PTE fold will require careful investigation and potentially atomic mutations along the hydrogen bond donors and acceptors along each of the bases. However, an initial attempt to disrupt and reestablish the network helps illustrate how essential and complex each of these residue's roles are for PTE fold and function.

3.2.7 Probing the role of the scaffold residues in the PTE structure and eIF4E binding with mutations

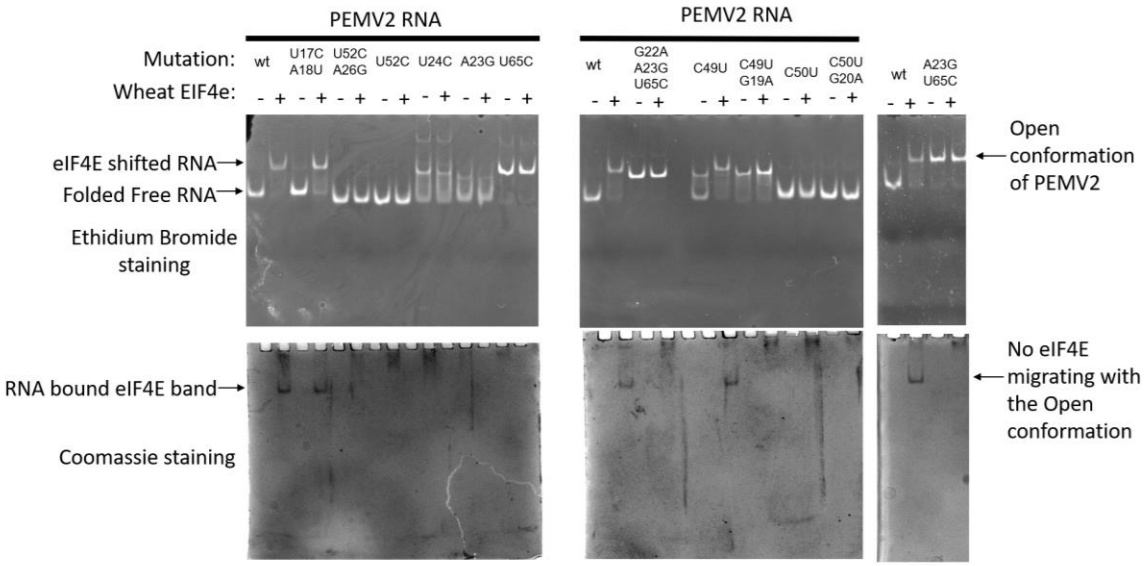


Figure 3.6. Native EMSA with 10 mM MgCl2 effective concentration in the gel and running buffer stained for RNA using Ethidium Bromide and for protein using Coomassie staining. Mutations present in each tested construct listed above their respective lanes. Lanes labeled with a "-" contain only RNA alone while Lanes labeled with a "+" contain 1.5 Molar equivalents of recombinantly expressed wheat eIF4E in addition to the refolded RNA.

PEMV2 PTE function correlates directly with EIF4E binding activity; When refolded PTE is added to the luciferase reporter construct with the 3' and 5' UTRs of the virus enabling translation in wheat germ extract. The affinity the PTE has for eIF4E determines the magnitude of translation inhibition because it competes for the available eIF4E [27,28]. Therefore, we can use EIF4E binding, as detected by a gel shift in an Electro Mobility Shift Assay (EMSA), as a proxy to measure the effect mutations have on PTE function. PEMV2 constructs' scaffold residues were mutated to break and potentially restore the scaffold interaction networks. The effects of these mutations are detected as a shift in the position of the free RNA band in an EMSA which is indicative of an alteration from the native fold, as well as alterations to eIF4E binding capacity. Previous studies have revealed that mutations to crucial residues in the C- and G-domains often eliminate eIF4E binding capacity entirely, instead of simply reducing affinity[27,28]. Therefore, in this assay a 1.5 molar excess of eIF4E was used to generate a quantitative shift in any construct that retains any amount of affinity for the protein.

To determine if the identity of first two residues of the G-domain U17 and A18 contribute to eIF4E binding these residues were mutated to C and U, respectively, in a double mutant construct. This U17C A18U double mutant PEMV2 PTE RNA in the absence of eIF4E had the same electrophoretic mobility as wild type, suggesting that the native fold was not altered by these mutations (Figure 3.6 lane U17CA18U "–"). This mutant also retains eIF4E binding capacity (Figure 6 U17CA18U "+"). These results suggest that the residue identities of U17 and A18 do not play a role in eIF4E recruitment. Nevertheless, interactions with their backbone atoms may help to position G19 to interact with C49.

Judging by electrophoretic mobility, the C50U mutant seems to retain the native fold (Figure 6 lane C50U "-"). However, in the presence of eIF4E there is no observable shift,

indicating that this mutation eliminates eIF4E binding activity (Figure 6 lane C50U "+"). The compensatory G20A mutation also retains the native fold but does not recover eIF4E binding (Figure 6 lane C50UG20A "+"). This result suggests that eIF4E recruitment may be more complex than simply recognizing the native folded PTE structure. G20 may serve an additional role, creating interactions with eIF4E using its solvent exposed major groove face that cannot be created without C50's positioning or as an A. Further experiments will be needed to clarify these less obvious interactions.

To perturb the tetra base interaction, U65 was mutated to a C and its partner A23 was mutated to a G, as both of these mutations should eliminate the same pair (Figure 6 lanes U65C "-", A23G "-"). The U65C mutant runs as a single slower migrating band compared to wild type, presumably representing an "open" conformation with the G-domain released from the core. Interestingly however, an A23G runs as two bands one at the height of the native conformation and one at the height of the open conformation. Neither construct or conformation binds eIF4E, as anticipated (Figure 6 lanes U65C "+", A23G "+"). The U65C mutation cannot be rescued by the corresponding A23G mutation, and yields the same monodisperse open conformer as the U65C point mutation (Figure 6 lanes U65CA23G "-"). These results establish the crucial nature of the U65 A23 pair. When compared to PTE homologs from other viruses this unpaired U65 residue is well retained in each of the sequences and SHAPE derived secondary structure models.

To perturb the base triple at the base of the scaffold structure, U52 was mutated to a C. This mutation would be expected to disrupt the U52 A26 pair. Curiously however this construct migrates marginally faster through the gel compared to wild type, possibly indicating an even greater degree of compaction, or an altered fold (Figure 6 lane U52C "-"). Despite the apparent enhanced compaction, we observe no mobility shift in the presence of eIF4E, suggesting that some

feature required of the eIF4E binding interface has altered even with the compensatory mutation (Figure 6 lane U52C "+"). Much like in the case of U65C, the compensatory A26G mutation yields the same results as the U52C mutation alone (Figure 6 lane U52CA26G "-", "+"). Considering that both constructs run as though they are slightly more compact than wild type, it is possible that allowing the final residue of the C-domain, U25, to exist as a C may allow all four G's from 19 to 22 to pair with the matching C's. In three-dimensional space U52 is near enough to either G21 that it could form a GC pair. This alternate pairing mode would be expected to retain and possibly enhance compaction and eliminate eIF4E binding through sequestration of the critical G21 residue. The loss of the U52 A26 pair may not necessarily eject A26 either even when mutated to a G. As an A or a G, A26 could still be held in place by pi stacking between U65 and U64. Additionally, U24 can form the same sugar edge interactions with either residue.

To disrupt the U24 A26 Watson Crick sugar edge interaction, U24 was mutated to a C. In the EMSA, this mutation drastically alters the electrophoretic signature of the PTE, creating three conformers, one migrating close to the native conformation, one migrating like the open conformation observed with other mutants, and one significantly slower migrating conformation that could involve multimerization (Figure 6 lane U24C "-"). The U24C mutant is unable to bind eIF4E likely due to its altered structure (Figure 6 lane U24C "+"). It is particularly interesting that disrupting different residues within the same base triple has drastically different effects on the fold of the RNA, while each eliminates eIF4E binding activity. These results suggest that the RNA structure and its eIF4E binding function should be considered more separately than initially anticipated.

To probe the role of the first two pairs of the C- and G-domains, each C was mutated to a U individually and the corresponding compensatory G to A mutations were made in double

mutants. In the absence of eIF4E, C49U migrates as two bands, suggesting two conformations (Figure 3.6 C49U "-"). One band has a similar mobility to the wild type native fold and the other runs as the open conformation. Despite the misfolded fraction of the C49U mutant, in the presence of eIF4E the RNA-protein complex migrates as a single band, and Coomassie staining of the gel shows eIF4E co-migrating with the RNA (Figure 6 C49U "+"). Possibly, eIF4E can bind to both conformations and the respective complexes migrate as a single species. Alternatively, the two RNA conformations may interconvert and bind eIF4E only in the compact conformation, thereby shifting the equilibrium between the two states to favor the compact conformation. These observations suggest that C49U may form a favorable interaction with G19 thorough formation of a wobble pair. If this occurs G19 could maintain the correct position to retain its native interactions with G22 and C51 below it. Making the compensatory G19A mutation in the C49U background more strongly favors the open conformation represented by the upper band. This construct and conformation are no longer capable of binding eIF4E. These effects most likely reflect disruption of the hydrogen bonding interactions between the G19 exocyclic amine and C51.

While many of these results were unexpected, they raise interesting questions about the actual mechanism of eIF4E recruitment. eIF4E recognition of G21 is likely dependent on the PTE adopting the solvent exposed G conformation, but additional interactions with the core may also play a role. The PTE core may also be more dynamic and flexible than anticipated and capable of adopting multiple functional arrangements. Collectively, the current mutational analysis supports the functional relevance of the crystal structure and illustrates the specific and complex network of hydrogen bonds at work stabilizing the G-domain's unusual fold. A more in-depth investigation of the PTE biochemistry is needed to relate the crystal structure back to the precise mechanism of eIF4E recruitment.

3.3 Discussion

3.3.1 How the PTE creates the eIF4E binding surface

Our structure shows the PEMV2 PTE's hypermodified G is flipped out into solution, as predicted by SHAPE reactivity and computational modeling. This structure reveals a complex network of interactions with the three-way junction core that pinch and flip the G-domain bulge. The clear and complete density for this G21 residue suggests that this position is well ordered and stable in the crystal. It is possible that the sharp turn at position 21 could keep the exposed G rigid in solution as well. Alternatively, the dense packing of interactions in the scaffold core may leave no space for G21 to dock back into the RNA, forcing it to remain flipped out into solution. G21's 2'OH may also be especially reactive with the SHAPE reagents in part due to its endo sugar pucker conformation, which reduces the steric hinderance of its environment [6,15].

The flipped position of G21 also closely matches predictions made in computational modeling of the PTE structure bound to eIF4E. It has been shown, that mutating either of the two tryptophans (W62 and W108) in the cap-binding pocket of EIF4E drastically reduces EIF4E's affinity for the PTE[27,28]. This observation is consistent with the hypothesis that EIF4E binds the PTE using its cap-binding pocket. It is likely that G21 is the site of eIF4E interactions because eIF4E only protects the G-domain from SHAPE modifications, all other regions of the RNA retain the same reactivity observed in the absence of eIF4E. Currently our collaborators are using this crystal structure with existing biochemical data to model the molecular dynamics of the wheat eIF4E cap binding pocket and PTE G21 interactions. While this work is not complete, initial results converge on a docking model in which the exposed G21 intercalates between the eIF4E's binding pocket tryptophans analogous to eIF4E's interaction with the native M7G ligand.

Mutations that alter the PTE fold to the open conformation abolished eIF4E binding presumably by altering the position or stability of the flipped out G21. Curiously these mutations could not be recovered by compensatory mutations, suggesting the complex network of hydrogen bonds creates site-specific requirements for residue identity. Even constructs that migrate with a degree of compactness analogous to wild-type can lose eIF4E binding capacity as a result of these mutations, indicating that a compact fold, while necessary, is not sufficient to create a productive eIF4E binding site in the G-domain. Therefore, it may be that the exposed G docking into the cap binding pocket is not sufficient for EIF4E binding and additional interactions between the G-domain surface and eIF4E are necessary.

How eIF4E binds its native M^7GpppG structure has been biophysically characterized by measuring affinity and free energies of eIF4E binding various cap analogues[17]. When binding M^7Gppp, π stacking interactions between the base and the two tryptophans above and below the base contribute the majority of the free energy accounting for -4.9 kcal/mol of roughly the total -10.8 kcal/mol of the interaction[17]. This interaction is positioned by additional hydrogen bonds between the binding pocket and the with the guanine WC face. The methylation mark helps to orient the base face and adds additional positive charge character to N7 strengthening the stacking interaction by roughly -3kcal/mol compared to un-modified cap analogs. The total free energy of eIF4E's interaction with M7GpppG structure is -10.8 kcal/mol, the remaining free energy contributions come from amino acid interactions with the three phosphates[17]. Studies demonstrate that this stability is reduced by increasing mono-valent ion concentration and is improved when the cap structure is covalently attached to an RNA oligonucleotide1[20]. Free GTP, GDP, and GMP are poor cap analogues and have binding free energies of -5.808 kcal/mol, -4.950 kcal/mol, and -3.09 kcal/mol respectively, which correspond to the following K_d's 46.7 uM, 204 uM, 5000 uM[17].

PEMV2 PTE binds wheat eIF4E with 58 nM affinity which is even higher than GTP's affinity for eIF4E[27]. If PEMV2's interaction with eIF4E were solely mediated by G21 insertion into the binding pocket, its affinity would likely most closely resemble that of GMP. Therefore, considering the biophysical properties of eIF4E cap-analog interactions, it is likely that PEMV2 creates additional interactions analogous to the triphosphate element of the GTP binding event. These additional interactions could consist of hydrogen bonding or salt bridging between the solvent exposed base faces of the surrounding guanines or interactions between the sugar their phosphate sugar backbones of these G-domain residues. If these unanticipated interactions are present, they may have been unintentionally disturbed by the structural changes induced by the compensatory mutations to the scaffold structure.

To determine if this is in fact the case, the residues of eIF4E that interact with the phosphates of the GTP dinucleotide can be mutated and PEMV2 PTE's affinity for these mutant eIF4E constructs can be measured. If we observe a loss of PEMV2 PTE affinity for eIF4E, we can conclude that these residues are forming additional unanticipated interactions with the G-domain scaffold structure. From there, the hydrogen bonding groups of each solvent exposed base surface could be mutated through the use of artificial base analogues and assayed for their individual contributions to eIF4E binding affinity. Alternatively, NMR profiles of wild type and mutant eIF4E binding labeled PEMV2 constructs may reveal which atoms of the G-domain are forming hydrogen bonding interactions with the eIF4E binding interface through changes in peak shifts in wild type compared to mutant.

The relative position of eIF4E bound to the PTE may also be informative to identify potential additional interaction sites. This could be determined using low resolution Small-angle X-ray scattering (SAXS) envelopes of the complex combined with this high-resolution model of

the RNA and existing crystal structures of eIF4E. Currently, no structures exist of eIF4E bound to an RNA through this unusual single nucleotide interaction, so much remains to be understood how this interaction can be established sufficiently stably to tolerate the circularization necessary to induce translation of the viral mRNA. This presents a very exciting opportunity to study how these viral RNAs evolved to bind eIF4E with such an unusual structural strategy that relies on mimicking the 5' cap structure itself instead of binding the protein through other interactions along its surface.

3.3.2 An overarching PTE fold is detectable in homologues PTE sequences.

Looking at the secondary structure of the PTE homologs of other viruses we can identify a general strategy that PTEs may employ to create this C-turn motif to present the eIF4E binding G21. Here I have hypothesized a definition of this architectural strategy as a series of rules that PTEs may follow to organize the G-domain into the C turn and scaffold. Since the C- and G-domains run in parallel, the 5' end of the G-domain must stretch up across the length of H3 to begin pairing with the 5' end of the C-domain, which stretches down to meet it. Therefore, the length of P1 should be fairly consistent or shorter to accommodate G-domains with fewer unpaired residues at the 5' end. Additionally, the lengths of each of the junctions should be similar if the homologs are folding into the same tertiary arrangement. Many of the secondary structure models for the PTE homologs can be adjusted to match the changes we observe in the PEMV2 model[27]. However, some homolog's secondary structure models required additional rearrangements to create junction and paired region lengths that match PEMV2's atypical A type junction structure. A table comparing the lengths of P1, J1/2, J2/3 and J3/1 of the revised secondary structures of each of the biochemically characterized PTE homologs is shown below (Table 3.2).

Table 3.2: PTE homolog revised A-type junction motif comparison

	P1	J1/2	J2/3	J3/1
PEMV2	8	0	4	1
PMV	8	0	4	1
CMMV	8	0	4	1
TPAV	8	0	4	1
CaRMV*	8*	0	4	1*
PFBV*	8	1	4*	1*
JINRV	7	0	6	1
SCV	7	1	6	3

**Indicates that the length of this region was altered from the predictions made in Wang et al. 2011*[27]

The paired regions and junctions of PTEs belonging to PMV, CMMV, and TPAV easily accommodated the changes applied to the original PTE secondary structure with no additional adjustments needed to match the conditions of the PTE three-way junction (Table 3.2)[27]. For CARMV PTE, the final base pair (C3748 and G3613) of H2 (P4) was eliminated and G3613 was instead made to pair with C3754. Other PTE's required larger rearrangements to make the PEMV2 fold. PFBV was altered by pairing U3718 with A3703 and breaking the C3722-G3742 base pair so that J2/3 now consists of the sequence UCCC and leaving G3781 as the single residue in J3/1.

At the base of P1, the G3695-U3790 pair was broken so that U3790 could pair with A3694. It is impossible to know if these revised secondary structures reflect the actual structures without further data, but these changes do not disagree with the chemical probing sensitivity observed for these residues.

However, three of the biochemically characterized PTE homologs were too different to create a secondary structure that matches PEMV2's three-way junction components, which may suggest that these PTEs adopt a different fold from PEMV2. HCRSV contains an additional bulge in J1/4 not present in any PTE so it was excluded from this analysis. JINRV has a shorter P1 and a longer J2/3 which makes the sites of G- and C-domain interactions more ambiguous. SCV PTE is even more divergent from the PEMV2 structure with a 7 bp long P1, a 1 residue long J1/2, a 6-residue long J2/3 and a 3-residue long J3/1. Each of these changes could substantially alter the alignment of the G- and C-domain pairing residues from the PEMV2 structure model, making this set of interactions difficult to infer from the PEMV2 structure. Therefore HCRSV, JINRV and SCV were excluded from the co-variation analysis that follows to avoid inappropriate speculation.

To determine if the C-turn and scaffold structure could also be conserved in the homologs we looked for co-variation between the interacting positions along the revised secondary structures. We found that the PTEs which could adopt a PEMV2 style three-way junction also show covariation between interacting residues consistent with this PEMV2 model. This includes

the previously un-anticipated parallel base pairing between J4/1 and J2/3 as well as between J4/1 and J3/1.

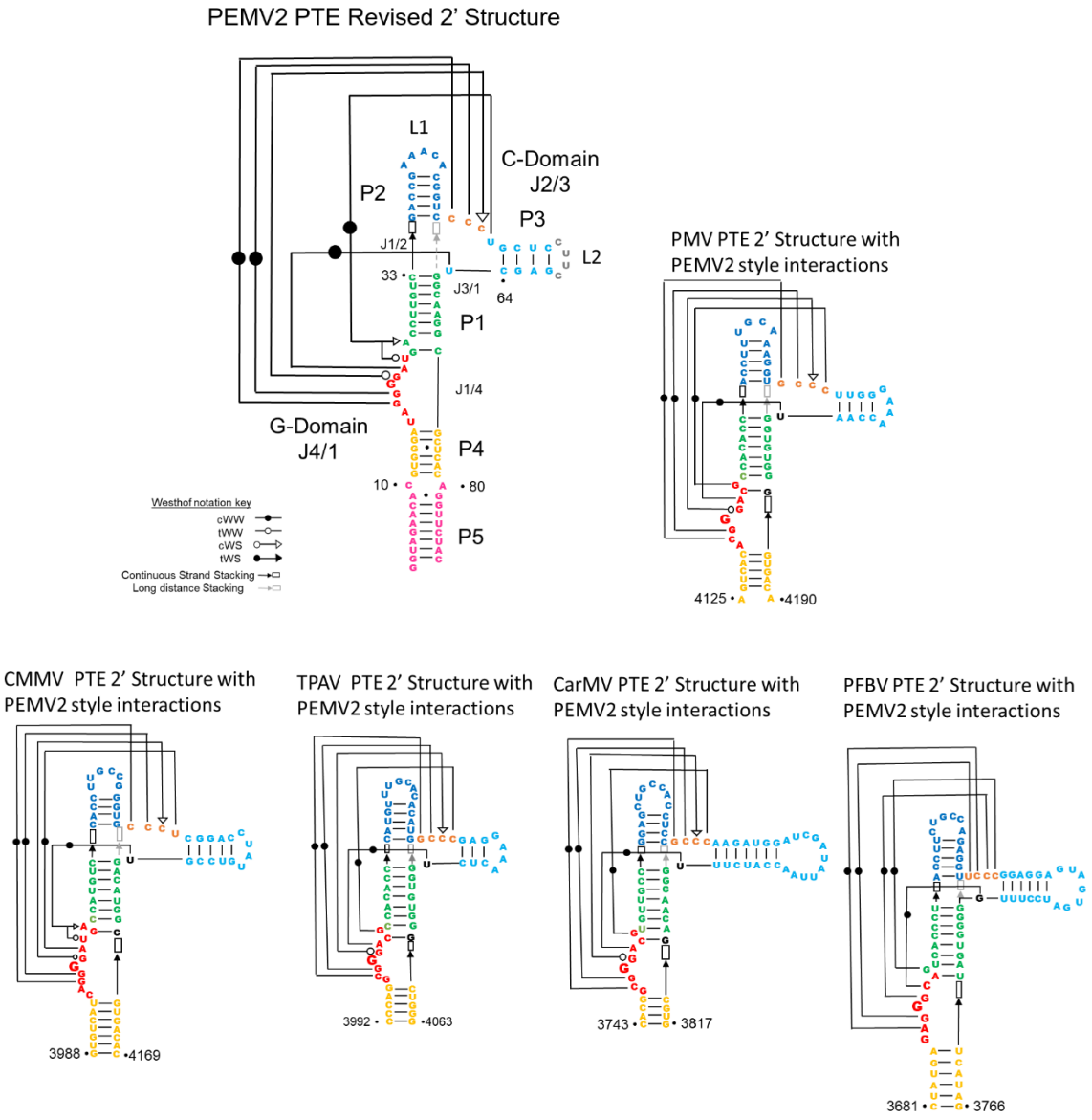


Figure 3.7: PEMV2 long distance interaction network super imposed on the PTE homologs' revised secondary structures.

To create the eIF4E binding site, the central G of the G-domain must be held in place by a sharp turn at either end of the inflection point. In PEMV2 there are two G-domain-C-domain pairs 5' of the EIF4E-binding G and two pairs 3' of the EIF4E binding G. In each of the PTE homologs we see that the first two residues of the C-domain co-vary with the first two residues of the G-domain. These pairs are therefore likely retained in each of the five PTE's capable of forming the same three-way junction as PEMV2. In some cases, the position of the eIF4E binding G had to be adjusted, based on the ambiguity of the band's identity in the SHAPE gels reported by Wang et al. 2011[26]. In these cases, moving the hypermodified G one position more 3' than initial predictions to create two positions 5' of the inflection point. PTEs with longer C-domain bulges like SCV may use different residues in the C-domain to make these pairs depending on where the C-domain is located in three-dimensional space.

Working off of the PEMV2 structure as though it represents a general arrangement of PTEs, the first residue 3' of the bulged G must make an interaction with the third residue of the C-domain. This interaction does not have to be a WC interaction if the residue in the third position of the C-domain turns inward, as C51 does, to form a sugar edge interaction with the first residue 3' of the exposed G. In this position the base pairs are as follows for each homolog forms a GC pair (Figure 3.7). Therefore, the GC WC sugar edge interaction observed in PEMV2 may be present in all of these homologs.

The second residue 3' from the hyper modifiable G should form a WC pair with the J3/1 . These two residues should also co-vary if this interaction is retained. In these positions the base pairs are as follows for each homolog: PMV AU, CMMV AU, TPAV AU, CarMV AU, PFBV CG (Figure 3.7). The fact that each homolog has a standard WC pair at the second residue 3' of the

hyper modifiable G and at J3/1 suggests that the U65 anchor we observe in the PEMV2 structure is a conserved feature of these PTEs.

The final residues of the G-domain knot back and forth between P1 and the scaffold structure. U24 forms a sugar edge interaction with A26 which intercalates into the stack and pairs with the last residue in J2/3. Simultaneously G25 forms the first pair of P1 with the residue originally attributed to J1/4. The PMV, CMMV, TPAV, and JINRV J1/4 residues can each pair with their respective second to last residue of the G-domain, leaving the last residue of the G-domain free to pair with the final residue of J2/3. For PFBV, G3695 can be unpaired from U3750 which can then pair with A3694 and G3695 able to pair with C3721 in J2/3 (Figure 3.7). In CarMV's secondary structure map the bulged G3746 was made to pair with the proximal C3810 which then leaves G3747 and C3748 into the G-domain. This also leaves G3811 free to serve as J1/4 which then pairs with C3754 in P1. Revising the secondary structures of the PTE homologs to match the restraints of the three-way junction and of the C-turn structure reveals a conserved pattern of long-distance pairing interactions. However, this re-organization of the predicted secondary structure of these PTEs is only a hypothesis to present the opportunity to probe for these interactions using mutations and biochemistry.

3.4 Conclusions

Presented here is the first high resolution crystal structure of any 3' CITE. Here we have discovered what is believed to be a new structural motif which we have named a C-turn. While this structure supports the hypothesis that the PEMV2 PTE uses a flipped G in the G domain to bind eIF4E the precise mechanism of binding remains unclear; biochemical mutations suggest that more than the flipped G architecture is required for forming stable eIF4E interactions. Our structure supports the hypothesis for a general architecture through covariation present in PTE

homologs at crucial sequence positions. By forcing a single G to flip out from the core RNA structure this residue may act as a 5' cap structure mimic and bind the EIF4E cap binding pocket. Many of the PTE homologs have the necessary sequences to create a similar three-way junction, scaffold structure and C-turn, so they may adopt a similar fold to present a flipped G for eIF4E binding. However, whether they actually do remains to be determined. Notably, this pattern of conservation would have been impossible to identify with phylogeny and chemical probing data alone. The three-dimensional orientation of the components of the three-way junction would only be detectable using a direct structural determination method like crystallography SAXS or Cryo-EM. Using RNA crystallography to study structured RNAs is therefore a powerful tool to discover new RNA structural motifs and general viral strategies to hijack host cell machinery. Living in a world trapped by a rapidly evolving RNA based virus underscores the importance of experimentally determined three-dimensional RNA structures.

3.5 Methods

3.5.1 RNA Transcription and Purification.

Single stranded DNA templates and primers for PCR and transcription were ordered from IDT encoding the transcription template for each RNA construct with a T7 promoter. Forward primers were ordered matching the T7 promoter region and reverse primers contained a single 2'O-methyl modification at the 3' end to avoid untemplated additions by T7 polymerase. Transcription template DNA was amplified into double stranded DNA using PCR. RNA was transcribed from the purified PCR product using an in vitro transcription reaction as follows 50 pmol mL^{-1} DNA template was incubated for 3 h at 37 °C in buffer containing 40 mM Tris-HCl, pH 8.0, 2 mM spermidine, 10 mM NaCl, 25 mM $MgCl_2$, 10 mM DTT, 40 U mL^{-1} RNase inhibitor, 5 U mL^{-1} thermostable inorganic pyrophosphatase, 5 mM of each NTP, and 50 μg mL^{-1} T7 RNA

polymerase. Reactions were halted by addition of RNase free DNase1 at 5U mL^{-1} and incubation at 37 ^{o}C for 30 minutes. RNA was purified on a 10% denaturing polyacrylamide gel in 0.5x TBE running buffer. The RNA was visualized with UV shadowing, extracted and eluted into 10 mM Tris, pH 8.0, 2 mM EDTA, 300 mM NaCl buffer via overnight incubation at 4 ^{o}C. The eluted RNA was then concentrated and exchanged into double distilled H_2O using a 10K Amicon filter and stored at -80 ^{o}C until further use.

3.5.2 Fab Purification.

The BL3-6 Fab expression vector (available upon request) was transformed into 55244 chemically competent cells (www.atcc.org) and grown on LB plates supplemented with carbenicillin at 100 μg mL^{-1}. Nine colonies from the plates were chosen and inoculated to a starter culture with 100 μg mL^{-1} carbenicillin, which was grown at 30 ^{o}C for 8 hours. Once the starter culture reached an OD 600 of 8, 15mLs of starter culture was used to inoculate 1L of 2×YT media and grown for 24 h at 30 °C. The cells were then pelleted via centrifugation at RT, and the cell pellet was resuspended in 1L of freshly prepared CRAP-Pi media supplemented with 100 ug mL^{-1} carbenicillin. The cells were set to grow for 24 h at 30 °C, harvested via centrifugation at 4 ^{o}C and frozen at -20 ^{o}C. Frozen cell pellets were lysed in PBS buffer supplemented with 0.4 mg mL^{-1} of Lysozyme and 0.01 mg mL^{-1} of DNase I. After 30 minutes PMSF was added to a final concentration of 0.5 mM. After 30 minutes, the cells were centrifuged, 45 min, 12000 rpm, at 4 ^{o}C. Lysate was transferred to new sterile bottles and centrifuged again for 15 minutes, 12000 rpm, at 4 ^{o}C. Supernatant was filtered through 0.45 μm filters into a sterile bottle (Millipore Sigma, www.sigmaaldrich.com), and Fab proteins were purified using the AKTAxpress fast protein liquid chromatography (FPLC) purification system (Amersham, www.gelifesciences.com) as described previously.[37] The lysate in PBS buffer (pH 7.4) was loaded into a protein A column, and the eluted

Fab in 1 M acetic acid was buffer exchanged back into the buffer PBS (pH 7.4) using 30kDa cutoff Amicon filter and loaded into a protein G column. The Fab was eluted from protein G column in 0.1 M glycine (pH 2.7) and then buffer-exchanged into 50 mM NaOAc, 50 mM NaCl buffer (pH 5.5) and loaded into a heparin column. Finally, the eluted Fab in 50 mM NaOAc, 2 M NaCl (pH 5.5) was dialyzed back into 1× PBS (pH 7.4), concentrated, and analyzed by 12% SDS-PAGE using Coomassie Blue R-250 staining for visualization. Aliquots of Fab samples were tested for RNase activity using the RNaseAlert kit (Ambion, www.thermofisher.com). The aliquots of Fab samples were flash frozen in liquid nitrogen and stored at −80 °C until further use.

3.5.3 Wheat eIF4E Expression

The gene for wheat eIF4E (gene bank Z12616.2) was cloned into a pET21a vector to include an N-terminal histidine tag. The resulting plasmid sequence was verified and transformed into BL21 rosetta DE3 *E.coli.* Cells were grown in LB supplemented with 100 μg/mL Carbanicilin at 37 ºC first in a 50 mL small scale culture overnight which was then used to induce a 1L culture. The 1L culture was grown at 37 ºC, until log phase was achieved. Cells were induced for protein expression using 100 mM IPTG and grown for 3 hours before harvesting through centrifugation at 6000 rpm for 10 minutes at 4◦C in 1L batches.

3.5.4 Wheat EIF4E Purification

Cell pellets were lysed using lysis buffer (25mM HEPEs pH 7.6, 100mM KCl, 2mM MgCl2, 10% glycerol), 0.4 mg/mL Lysozyme, 0.1 mg/mL DNAase, 1 mM PMSF, 0.5% Triton-100) and a single freeze thaw cycle followed by an incubation at room temperature for 1 hour with 0.4 mg/mL lysozyme and 0.1 mg/mL DNase. Cell debris were removed from the lysis by a 45 min centrifugation at 12,000 rpm. Clarified lysate was filtered and applied to a Qiagen His trap column,

washed with 5CV running buffer (50mM HEPES pH 7.6 2mM MgCl2 10% Glycerol 10mM Imidazole 100mM KCl pH 7.5) and eluted with running buffer supplemented with 250 mM imidazole. Fractions containing high UV absorbance were pooled and buffer exchanged into running buffer and applied to the Ni+ column once more. The column was then washed with running buffer supplemented with additional 20 mM imidazole and 500 mM NaCl for 30 CV followed by a second wash with 30 mM imidazole and 1 M NaCl for 40CV. Finally, protein was eluted using running buffer supplemented with 250 mM Imidazole. Fractions with a high UV absorbance were collected and pooled, the purified protein was then buffer exchanged into storage buffer containing (50mM HEPES pH 7.6 2mM MgCl2, 10% Glycerol, 100mM KCl pH 7.5) and flash frozen with liquid nitrogen in low concentration aliquots for further use in EMSA assays.

3.5.5 Electrophoretic Mobility Shift Assay (EMSA).

Purified RNA constructs in double distilled H_2O were heated to 90 ^oC for 1 minute, then cooled on ice for 2 minutes then held at room temperature for 3 minutes. Refolding buffer (10mM Tris pH 7.5, 10mM $MgCl_2$, 100mM KCl) was added, and the RNA was then incubated at 50 ^oC for 10 minutes followed by a 5-minute incubation on ice to facilitate refolding.

Refolded RNA was then mixed with either eIF4E storage buffer as a negative control or a 1.5 M ratio of RNase-free eIF4E and incubated on ice for 15 minutes to establish equilibrium state binding. eIF4E RNA complexes were separated by gel electrophoresis in a 10% polyacrylamide gel made in 0.5x TBE buffer supplemented with 11 mM $MgCl_2$. The gel was loaded while running to avoid complex dissociation and was run at 4 ^oC for 1hr. The gel was stained with ethidium bromide and visualized via UV light and photographed then stained in Coomassie and de-stained and photographed a second time.

3.5.6 Crystallization

Purified RNA constructs in double distilled H_2O were heated to 90 ^oC for 1 minute, then cooled on ice for 2 minutes then held at room temperature for 3 minutes. Refolding buffer (10mM Tris pH 7.5, 10mM $MgCl_2$, 100mM KCl) was added, and the RNA was then incubated at 50 ^oC for 10 minutes followed by a 5-minute incubation on ice to facilitate refolding. 480 μg of refolded RNA was mixes with a 1.1 Molar equivalent of Fab BL3-6 in buffer (10mM Tris pH 7.5, 10mM $MgCl_2$, 100mM KCl). The complex was then concentrated to 6 mg/mL final concentration of RNA (80 μL). 100 nL + 100 nL hanging drop crystal trials were set in various crystallization kits and allowed to grow for two to three weeks at room temperature. Crystals grew in (5 mM $MgCl_2$, 50 mM HEPES pH 7.0, 25% PEG 550) and were further optimized with an additive screen resulting in a final condition of (5 mM $MgCl_2$, 50 mM HEPES pH 7.0, 25% PEG 550, 2% Benzamidine hydrochloride). Crystals were then looped and frozen in liquid nitrogen.

3.5.7 Diffraction Data Collection

Diffraction data was collected at APS beam line 24-ID-E. Crystals of PEMV2 and fab were small so data was collected in two 180-degree halves with a pause between the two to take a snap diffraction to confirm that the crystal was still in the beam. The last 300 images from the diffraction set were removed due to radiation damage and the remaining images were merged together into a final data set which was then processed in the Phenix software suite.

3.5.8 Crystallographic Data Processing

An initial molecular replacement solution was generated by searching for the Fab-model bound to the BL3-6 penta loop. The RNA of the model was then built into the resulting electron density map and the model was built and refined iteratively. To combat the accumulating model,

bias a composite omit map was generated with simulated annealing. Anisotropy in the data set was removed using the Xia2 dials server. The data was finally truncated to 2.75 Å out of an abundance of caution to avoid incorporating noise into the highest resolution shells.

3.6 Author Contributions

Contributing Authors: Anna Lewicka, Stacy Jones, Sejal Prachad, Agata Ryzko, Phoebe Rice and Joseph A. Piccirilli

A. L. and S.J. conceived and designed the initial experiments under direction from J.A.P. C.R., S. J. and A. L. each prepared the samples and conducted biochemical and crystallographic experiments and collected x-ray diffraction data across the various years of this project. Ultimately C.R. and A.L. performed the final crystallization trials and data collection together. C.R. cloned, expressed and purified the wheat eIF4E recombinant protein with assistance from her mentees S. P. and A. R.. C. R. solved the crystal structures with P.R.. C.R. collected additional biochemical data and analyzed both the biochemical and crystallographic data, interpreted the results, and wrote the manuscript with J.A.P.

3.7 Works Cited in Chapter 3

1) Batten, J. S., Desvoyes, B., Yamamura, Y., & Scholthof, K. B. G. (2006). A translational enhancer element on the 3′-proximal end of the Panicum mosaic virus genome. FEBS Letters, 580(11), 2591–2597. https://doi.org/10.1016/j.febslet.2006.04.006

2) Nicholson, B. L., Zaslaver, O., Mayberry, L. K., Browning, K. S., & White, K. A. (2013). Tombusvirus Y-Shaped Translational Enhancer Forms a Complex with eIF4F and Can Be Functionally Replaced by Heterologous Translational Enhancers. Journal of Virology, 87(3), 1872–1883. https://doi.org/10.1128/JVI.02711-12

3) Browning, K. S., & Bailey-Serres, J. (2015). Mechanism of Cytoplasmic mRNA Translation. Arabidopsis Book, 13(e0176). https://doi.org/doi:10.1199/tab.0176

4) Du, Z., Alekhina, O. M., Vassilenko, K. S., & Simon, A. E. (2017). Concerted action of two 3′ cap-independent translation enhancers increases the competitive strength of translated viral genomes. Nucleic Acids Research, 45(16), 9558–9572. https://doi.org/10.1093/nar/gkx643

5) Meulewaeter, F., Van Montagu, M., & Cornelissen, M. (1998). Features of the autonomous function of the translational enhancer domain of satellite tobacco necrosis virus. Rna, 4(11), 1347–1356. https://doi.org/10.1017/S135583829898092X Gao, F., Gulay, S. P., Kasprzak, W.,

6) Dinman, J. D., Shapiro, B. A., & Simon, A. E. (2013). The Kissing-Loop T-Shaped Structure Translational Enhancer of Pea Enation Mosaic Virus Can Bind Simultaneously to Ribosomes and a 5' Proximal Hairpin. Journal of Virology, 87(22), 11987–12002. https://doi.org/10.1128/jvi.02005-13

7) Gao, F., Kasprzak, W. K., Szarko, C., Shapiro, B. A., & Simon, A. E. (2014). The 3' Untranslated Region of Pea Enation Mosaic Virus Contains Two T-Shaped, Ribosome-Binding, Cap-Independent Translation Enhancers. Journal of Virology, 88(20), 11696–11712. https://doi.org/10.1128/jvi.01433-14

8) Gherghe, C. M., Mortimer, S. A., Krahn, J. M., Thompson, N. L., & Weeks, K. M. (2008). Slow Conformational Dynamics at C2′-endo Nucleotides in RNA. Journal of the American Chemical Society, 130(28), 8884–8885. https://doi.org/10.1021/JA802691E

9) Huang, H., Suslov, N. B., Li, N.-S., Shelke, S. A., Evans, M. E., Koldobskaya, Y., … Piccirilli, J. A. (2014). A G-quadruplex-containing RNA activates fluorescence in a GFP-like fluorophore. Nature Chemical Biology, 10(8), 686–691. https://doi.org/10.1038/nchembio.1561

10) Javed, T., Shabbir, R., Ali, A., Afzal, I., Zaheer, U., & Gao, S.-J. (2020). Transcription Factors in Plant Stress Responses: Challenges and Potential for Sugarcane Improvement. Plants, 9(4). https://doi.org/10.3390/PLANTS9040491

11) Jones, R. A. C. (2021). Global Plant Virus Disease Pandemics and Epidemics. Plants, 10(2), 1–41. https://doi.org/10.3390/PLANTS10020233

12) Kraft, J. J., Peterson, M. S., Cho, S. K., Wang, Z., Hui, A., Rakotondrafara, A. M., … Miller, W. A. (2019). The 3 ′ untranslated region of a plant viral RNA directs efficient cap-independent translation in plant and mammalian systems. Pathogens, 8(1), 1–19. https://doi.org/10.3390/pathogens8010028

13) Martinez-Salas, E., Francisco-Velilla, R., Fernandez-Chamorro, J., & Embarek, A. M. (2018). Insights into structural and mechanistic features of viral IRES elements. Frontiers in Microbiology, 8(JAN), 1–15. https://doi.org/10.3389/fmicb.2017.02629

14) Miras, M., Sempere, R. N., Kraft, J. J., Miller, W. A., Aranda, M. A., & Truniger, V. (2014a). Interfamilial recombination between viruses led to acquisition of a novel translation-enhancing RNA element that allows resistance breaking. New Phytologist, 202(1), 233–246. https://doi.org/10.1111/nph.12650

15) Miyoshi, H., Suehiro, N., Tomoo, K., Muto, S., Takahashi, T., Tsukamoto, T., … Natsuaki, T. (2006). Binding analyses for the interaction between plant virus genome-linked protein (VPg) and plant translational initiation factors. Biochimie, 88(3–4), 329–340. https://doi.org/10.1016/J.BIOCHI.2005.09.002

16) Mortimer, S. A., & Weeks, K. M. (2009). C2′-endo nucleotides as molecular timers suggested by the folding of an RNA domain. Proceedings of the National Academy of Sciences, 106(37), 15622–15627. https://doi.org/10.1073/PNAS.0901319106

17) Niedzwiecka, A., Marcotrigiano, J., Stepinski, J., Jankowska-anyszka, M., Wyslouch-cieszynska, A., Dadlez, M., … Stolarski, R. (2002). Biophysical Studies of eIF4E Cap-binding Protein : Recognition of mRNA 5 0 Cap Structure and Synthetic Fragments of eIF4G and 4E-BP1 Proteins, 2836(02), 615–635. https://doi.org/10.1016/S0022-2836(02)00328-5

18) Shahzad, R., Jamil, S., Ahmad, S., Nisar, A., Amina, Z., Saleem, S., … Wang, X. (2021). Harnessing the potential of plant transcription factors in developing climate resilient crops to improve global food security: Current and future perspectives. Saudi Journal of Biological Sciences, 28(4), 2323. https://doi.org/10.1016/J.SJBS.2021.01.028

19) Simon, A. E., & Miller, W. A. (2013). 3′ Cap-Independent Translation Enhancers of Plant Viruses. Annual Review of Microbiology, 67, 21–42. https://doi.org/10.1146/annurev-micro-092412-155609

20) Slepenkov, S. V., Darzynkiewicz, E., & Rhoads, R. E. (2006). Stopped-flow kinetic analysis of eIF4E and phosphorylated eIF4E binding to cap analogs and capped oligoribonucleotides: Evidence for a one-step binding mechanism. Journal of Biological Chemistry, 281(21), 14927–14938. https://doi.org/10.1074/jbc.M601653200

21) Staple, D. W., & Butcher, S. E. (2005). Pseudoknots : RNA Structures with Diverse Functions, 3(6), 956–959. https://doi.org/10.1371/journal.pbio.0030213

22) Stupina, V. A., Yuan, X., Meskauskas, A., Dinman, J. D., & Simon, A. E. (2011). Ribosome Binding to a 5' Translational Enhancer Is Altered in the Presence of the 3' Untranslated Region in Cap-Independent Translation of Turnip Crinkle Virus. Journal of Virology, 85(10), 4638–4653. https://doi.org/10.1128/jvi.00005-11

23) Tamura, M., & Holbrook, S. R. (2002). Sequence and Structural Conservation in RNA Ribose Zippers. Journal of Molecular Biology, 320(3), 455–474. https://doi.org/10.1016/S0022-2836(02)00515-6

24) Truniger, V., Miras, M., & Aranda, M. A. (2017). Structural and functional diversity of plant virus 3′-cap-independent translation enhancers (3′-CITEs). Frontiers in Plant Science, 8(November), 1–14. https://doi.org/10.3389/fpls.2017.02047

25) Truniger, V., Nieto, C., González-Ibeas, D., & Aranda, M. (2008). Mechanism of plant eIF4E-mediated resistance against a Carmovirus (Tombusviridae): Cap-independent translation of a viral RNA controlled in cis by an (a)virulence determinant. Plant Journal, 56(5), 716–727. https://doi.org/10.1111/j.1365-313X.2008.03630.x

26) Wang, Z., Kraft, J. J., Hui, A. Y., & Miller, W. A. (2010). Structural plasticity of Barley yellow dwarf virus-like cap-independent translation elements in four genera of plant viral RNAs. Virology, 402(1), 177–186. https://doi.org/10.1016/J.VIROL.2010.03.025

27) Wang, Z., Parisien, M., Scheets, K., & Miller, W. A. (2011). The Cap-Binding Translation Initiation Factor, eIF4E, Binds a Pseudoknot in a Viral Cap-Independent Translation Element. Structure, 19(6), 868–880. https://doi.org/10.1016/J.STR.2011.03.013

28) Wang, Z., Treder, K., & Miller, W. A. (2009). Structure of a viral cap-independent translation element that functions via high affinity binding to the eIF4E subunit of eIF4F. Journal of Biological Chemistry, 284(21), 14189–14202. https://doi.org/10.1074/jbc.M808841200

29) Westhof, E., & Lescoute, A. (2006). Topology of three-way junctions in folded RNAs. RNA Society, 12, 83–93. https://doi.org/10.1261/rna.2208106.In

30) Zuo, X., Wang, J., Yu, P., Eyler, D., Xu, H., Starich, M. R., … Wang, Y.-X. (2010). Solution structure of the cap-independent translational enhancer and ribosome-binding element in the 3’ UTR of turnip crinkle virus. Proceedings of the National Academy of Sciences, 107(4), 1385–1390. https://doi.org/10.1073/pnas.0908140107

31) Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. RNA. 2001 Apr;7(4):499-512. doi: 10.1017/s1355838201002515. PMID: 11345429; PMCID: PMC1370104.

4. Chapter 4 Thesis Conclusions

4.1 PFSE Conclusions

In 2020, the Programmed Frameshifting Element of SARS CoV-2 was a rigorously studied structured RNA element. Simultaneously, many groups attempted to resolve its structure and mechanism using a variety of methods, which makes putting the structure reported in this work into the context of the field at large an interesting exercise illustrating the strengths and limitations of each method.

The high-resolution frameshifting element crystal structure reported here found the PFSE in a 5' end threaded, three stem H-type pseudoknot with all three stems helically stacked on one another in a vertical conformation. S1 and S3 matched previous predictions but S2 was determined to contain only four base pairs with the final two residues forming interactions in S1 instead[10]. Another important observation gleaned from this structure is the fact that L3 begins after A13533 and ends at A13537 and contains three flexible residues in this linear conformation. The helical stack contains two triple base interactions at each helical junction. At the S1 S2 junction G13486's WC face pairs with A13537's 2'OH and the next phosphate group while A13537's WC face pairs with G13493's sugar edge. The circle is closed by G13493's WC face pairing with G13486's Hoogsteen edge. This interaction is enabled by positioning of L1 close to the center of the helix by an unanticipated water mediated U-U pair between U13494 and U13485. At the S1 S3 junction, a triple base forms when A13533 forms a WC sugar edge interaction with G13475 while it forms a wobble pair with U13504 at the start of S1. In SARS CoV-1 A13533's position is a C, which, interestingly, can form the same WC SE interaction with the G. In addition to potentially stabilizing the helical stacking of S3 and S1 this base triple may tie L3 down over the 5' end, creating the treaded conformation observed in all experimentally determined structures of the

PFSE [1,18]. We can compare these findings to the computational predictions which proceeded it to illustrate into the limitations of computational RNA modeling.

Schlick et. al. applied the RNA as a Graph framework to model the PFSE as a pseudoknot to determine the likely impact of mutations on the structure[14]. In this work, they generated a model with S1 and S2 helically stacked and S3 bent at a roughly 90-degree angle, which contained no base triples or major/minor groove interactions[14]. In this RNA graph model, L3 contained only two unpaired residues while L1 contained three unpaired residues[14]. These features are some of the key differences between the literature secondary structure and the PFSE crystal structure reported here and may have favored the bent arrangement[11].

Rangan et. al. applied the energy minimization folding strategy in the Rosetta FARFAR 2 program to fold the PFSE[12]. They report ten clusters of lowest energy PFSE structures. All ten clusters were reported because there was not enough information to differentiate which would be most likely to represent the two conformations identified by force unfolding experiments[12]. Some of the clusters have all three helices vertically stacked and some show S1 and S3 stacked with S2 bent at various orientations away from the helical axis[12]. Further, some models contained interactions between L3 and S1 or S2 of these a portion had a 5' end threaded arrangement, while others did not[12]. Each of the models differed in these finer details but all have similar free energies, which may reflect the dynamic nature of this structure. This highlights the limitations of energy minimization when modeling an RNA with multiple conformations.

Omar et. al. modeled the PFSE using multiple energy minimization programs in parallel SimRNA, Rosetta FARFAR2, RNA composer, RNAvista, MC-Sym, RNA2D3D and V fold[9]. They then performed molecular dynamics (MD) simulations on each of these blind structure predictions [9]. In the molecular dynamics simulations, explicit solvent conditions were modeled in

Amber 18 [9]. The seven folding programs generated seven different PFSE structures with the FARFAR2 and V fold structures most closely resembling the cryo-EM and crystal structures[1,9,18]. The molecular dynamics simulations were successful in predicting the potential presence of triple base pairs at the helical junctions although precisely which residues participated in these triple base and minor grove interactions varied[9]. When the 5' end threading is modeled in these MD simulations, the PFSE adopts a more linear conformation with many interactions forming between L3 and S2[9].

Many of these models share the same inaccuracies, most prominent of which is an elongated S2 which then shifts and truncates L3. These errors were shared across models likely because each used the literature secondary structure as the starting point, which contains these features[10]. Therefore, these small errors may have propagated through each of the tertiary models and contributed to their own inaccuracies compared to experimentally determined structures. For example, a shortened L3 has a substantial influence on the potential orientations of the helices. Without the experimentally determined structures, there would have been no way to determine that S2 was only four residues long instead of the predicted six.

When the 5' end threading feature of the PFSE was first observed in both low resolution cryo-EM structures, this led to a hypothesis that this unusual threaded topology was part of the ribosomal pausing mechanism[18]. It was initially thought that perhaps threaded structure is more energetically challenging to unwind than an unthreaded one and therefore the two predicted conformations could be threaded and unthreaded[18]. To address this hypothesis, unfolding force extension curves were measured for the PFSE by another group. In these experiments, a six-residue long spacer extension of the 5' tail of the PFSE was used to hybridize the DNA handles in atomic microscopy force extension curve experiments. It was found that extending the DNA handles to

pair further and further up the strand shortening the spacer to a single nucleotide altered the ratio of the two conformations, suppressing the conformation with a 30pN unfolding force and favoring the conformation with a 15pN unfolding force[5]. Given that triplex structures in H-type pseudoknots have been measured to contribute roughly 15pN of unfolding force to a structure, we can now postulate that the vertical conformation may represent the 30pN conformation, and the bent conformation may represent the 15pN conformation[3]. To determine if both the bent and linear conformations of the PFSE are present in solution we measured the SAXS profiles of the PFSE BL3-6 construct. Unlike the wild type construct, whose SAXS profiles clearly represented a dimer of PFSEs, the PFSE BL3-6 construct is incapable of dimerizing. The SAXS profile of the PFSE BL3-6 construct suggests the presence of both linear and bent character, further supporting the idea that our linear structure could represent the second PFSE conformation not yet observed experimentally.

The crystal structure reported here helped to fill the gaps between the cryo-EM structures, which observed only a bent threaded conformation, and the biophysical characterization, which identified a second yet unknown conformation of the PFSE [1,18,5]. This structure has provided an alternative explanation to the identity of the second PFSE conformation observed in force unfolding extension curves[4,5]. As the first high resolution coronavirus frameshifting element structure, this structure provides an important avenue to begin more direct experiments, which will guide investigations of the precise mechanism of -1 frameshifting in coronaviruses. When the residues in the S1 S2 triplex have been mutated, and each has been shown to compromise frameshifting and viral replication. However, the residues in the S1 and S3 triplex were not expected to interact and as a result have not yet been explored through mutational analysis[1, 8, 11]. It has also provided a model onto which molecular dynamics simulations may be applied to

determine the site of MTDB binding[6]. Identifying the site of MTDB binding will allow studies in which this ligand can be modified and refined into a pharmaceutical capable of inhibiting SARS CoV replication in the cell[6].

Because of the work described here, the PFSE structure and function relationship can now be probed using a combination of biochemical and structural techniques. Currently our lab is collaborating with other groups that specialize in SAXS and atomic force microscopy to measure the contributions the base triples make to the formation of the bent versus linear conformations in solution. Relating this to the effect mutations to the base triples have on frameshifting frequency will help to characterize their roles. It is likely that these base triples stabilize the vertical conformation and/or regulate conformational switching. Probing the roles of these interactions directly through mutagenesis, frameshifting assays, force unfolding experiments and SAXS in combination may reveal how the conformational dynamics of this structure contribute to the regulation of -1 ribosomal frameshifting in SARS CoV-2. Additionally, other lab members are in the process of crystalizing the SARS CoV-1 programmed frameshifting element to determine the effect that the A to C substitution has on the overall structure. Crystals of this element diffract more poorly than the SARS CoV-2 PFSE, so it may be that the vertical conformation which crystalized for SASR CoV-2 is less stable in this SARS CoV-1 mutant. Further, SAXS profiles of the SARS CoV-1 element are being generated to be compared against with those already collected for SARS CoV-2.

4.2 PEMV2 PTE Conclusions

The PEMV2 PTE is a cap-independent translational enhancer, an RNA element found in the 3' UTR of the Pea Enation Mosaic Virus 2. PEMV2 PTE serves as a structural representative of the PTE class of CITEs. Computational modeling predicted the PEMV2 PTE to adopt a T-

shaped structure with a central pseudoknot[17]. The crystal structure I report here of the PEMV2 PTE element agrees with these predictions broadly, and added finer details to the complexity of this fold as a result of the high resolution of this structure. For example, this structure is the first containing a parallel strand pairing pseudoknot-like structure. Additionally, this structure contains an atypical A-type three-way junction which forms a scaffold structure at its core that stabilizes the C-turn created by the C and G domain pairing.

Specifically, we find that the A type three-way junction formed by P1 stacked beneath P2 and P3 oriented perpendicularly to this axis. The junction positions residues C49, C50 C51 and U52 in the J2/3 C- domain bulge to pair with G19, G20, G22 and A26 respectively with U23 forming a triple base pair with A26 and U52. Meanwhile U65 in J3/1 pairs with A23 forming a lattice of hydrogen bonds in a tetra-base interaction with G22 and C51. these interactions create a scaffold structure that stacks upon a dinucleotide platform created at the base of P3 by residues G53 and C64. The parallel stranded pairing between the C and G domain stabilizes the ribose zipper of the G domain and the sharp turn at G21, which flips this residue out into solution for eIF4E recruitment. Not all PTE homologues can form this base pairing pattern, but among the PTEs whose sequences can form the same A-type three-way junction we find that the base pairing pattern does co-vary, suggesting that this structure could be a general architecture among some PTEs.

Molecular dynamics simulations using the predicted tertiary structure of the PTE and biochemical restraints modeled the hypermodified G of the G domain as flipped out into solution and docked in the cap binding pocket of wheat eIF4E[17]. Our structure indeed identifies this very conformation of the PTE, supporting the previously proposed mechanistic hypothesis. However, our own biochemical assays suggest that the binding mode of the PTE with eIF4E may be more

complicated than docking and π stacking interactions alone. Each mutation intended to disrupt this scaffold structure entirely eliminated eIF4E binding, though not all of them disrupted the PTE fold. The majority of corresponding compensatory mutations which were expected to restore the fold failed to recover eIF4E binding activity, which suggests that eIF4E binding is more complex than simply docking the flipped-out G into the cap binding pocket within eIF4E.

To determine the mechanism of PEMV2 PTE eIF4E binding, we now turn our attention to wheat eIF4E. Mutating the polar residues in and around the binding pocket and measuring their effect on PTE affinity may identify the locations of the additional interactions that stabilize the PTE eIF4E interface. Combined with molecular dynamics simulation on the PTE structure binding eIF4E, this data will provide much needed restraint parameters to most accurately simulate the interaction in solution. The resulting model of the PTE eIF4E complex may identify which atoms along the G and C domains form the additional stabilizing interactions. Simultaneously, we plan to characterize the PTE homologs' structures by mutating the residues we identified as likely to form analogous scaffold structures and measure their effect on the compactness of their folds and their capacity to bind eIF4E. These experiments will help to determine if the revised PTE homolog architectures are, in fact, accurate.

4.3 A larger scope and the significance

At the start of this thesis, I discussed the advances and limitations of computational modeling of RNA structures. The two projects presented in this dissertation both highlight the strengths and limitations of these methods and the important role RNA crystal structures play in the field. Sometimes, we focus specifically on the RNA target at hand and forget to consider unrelated but similarly structured RNAs to guide our investigations. In the case of the PFSE, triplexes have been shown to stabilize an H-type pseudoknot and increase frameshifting efficiency,

as observed by Chen et al, or use bending at helical junctions to adopt the two conformations needed for frameshifting, as observed in the Mouse Sarcoma Virus frame shifting element[3,19]. However, making these apparently unrelated connections can be challenging due to a lack of RNA structure databases and diverse structure classification schemes[14]. Similarly, computational folding algorithms also rely on existing structures to interpret the possible RNA elements, motifs, and non-canonical pairing interactions, which they build into the new models they generate[14]. Other PTE or CITE structures may employ a similar flipped G turn structure to create an eIF4E binding site. However, if this is achieved through the unusual tertiary arrangement of the PEMV2 PTE, it is unlikely blind folding simulations would be able to make such unconventional predictions. The PTE structure provides information about a new motif that can now be used as a template for other three-way junctions in future computational modeling studies. However, neither of these structures would have been able to be crystalized without the extensive structural predictions applied to each of them. The feedback loop between RNA crystallography and computational structure modeling allows both fields to advance in parallel, supported by one another.

Viral RNA crystal structures will continue to reveal new folds, motifs, and structural strategies to manipulate host cell biology. It is imperative that our representative RNA structures broaden and deepen to improve the detection and modeling mechanisms that enable rapid response to emerging pathogenic threats. To draw an analogy to the world beyond biology, hackers are constantly breaking through the latest computer security measures forcing technologies like encryption and two-factor authentication to proliferate and advance. Even then, eventually, those measures will become obsolete as well. Viral RNA elements evolve more rapidly than their protein counterparts because their randomly accumulated point mutations are able to directly alter the

structure and function of the molecule. As we watch the coronavirus mutate through billions of human and animal hosts, it is altering not only its spike protein but other components of its genome. As Covid variants are sequenced from infected patients, the frameshifting element region itself has been found to contain mutations across its structure. Although these mutations are infrequent and generally detrimental to viral replication, eventually a productive set of mutations could emerge that may alter the PFSE structure[8,13]. Similarly, the observation of a resistance breaking CITE being horizontally exchanged between viruses underscores the potential for this viral RNA element to rapidly enhance previously impotent agricultural pathogens[7]. For example, the Maize Cholortic Virus , which infects corn, one of the two most dominant crops in the United States, was discovered to contain a PTE element in its 3' UTR in 2019[2]. Only a subset of existing PTEs are known, so the emergence of a resistance breaking PTE, like the CXTE, is a distinct possibility as more plant viral RNA genomes are analyzed for these elements[16]. To minimize the response time research requires to these viral threats, RNA structural biology must employ more interdisciplinary approaches to studying and characterizing RNA structure and function. RNA crystallography is a critical source of information in this mission and developing technologies, such as chaperone assisted crystallography, is a valuable and effective approach to driving advances in our understanding.

4.4 Works Cited in Chapter 4

1) Bhatt, P. R., Scaiola, A., Loughran, G., Leibundgut, M., Kratzel, A., Meurs, R., … Ban, N. (2021). Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome. Science. https://doi.org/10.1126/science.abf3546

2) Carlno, E. J., Scheets, K., & Miller, W. A. (2020). The RNA of Maize Chlorotic Mottle Virus , an Obligatory Component of Maize Lethal Necrosis Disease , Is Translated via. Journal of Virology, 94(22), 1–28.

3) Chen, G., Chang, K. Y., Chou, M. Y., Bustamante, C., & Tinoco, I. (2009). Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of -1 ribosomal frameshifting. Proceedings of the National Academy of Sciences of the United States of America, 106(31), 12706–12711. https://doi.org/10.1073/pnas.0905046106

4) Halma, M. T. J., Ritchie, D. B., Cappellano, T. R., Neupane, K., & Woodside, M. T. (2019). Complex dynamics under tension in a high-efficiency frameshift stimulatory structure. Proceedings of the National Academy of Sciences of the United States of America, 116(39), 19500–19505. https://doi.org/10.1073/pnas.1905258116

5) Halma, M. T. J., Ritchie, D. B., & Woodside, M. T. (2021). Conformational Shannon Entropy of mRNA Structures from Force Spectroscopy Measurements Predicts the Efficiency of -1 Programmed Ribosomal Frameshift Stimulation. Physical Review Letters, 126(3), 38102. https://doi.org/10.1103/PhysRevLett.126.038102

6) Kelly, J. A., Woodside, M. T., & Dinman, J. D. (2021). Programmed −1 Ribosomal Frameshifting in coronaviruses: A therapeutic target. Virology, 554(October 2020), 75–82. https://doi.org/10.1016/j.virol.2020.12.010

7) Miras, M., Sempere, R. N., Kraft, J. J., Miller, W. A., Aranda, M. A., & Truniger, V. (2014). Interfamilial recombination between viruses led to acquisition of a novel translation-enhancing RNA element that allows resistance breaking. New Phytologist, 202(1), 233–246. https://doi.org/10.1111/nph.12650

8) Neupane, K., Munshi, S., Zhao, M., Ritchie, D. B., Ileperuma, S. M., & Woodside, M. T. (2020). Anti-Frameshifting Ligand Active against SARS Coronavirus-2 Is Resistant to Natural Mutations of the Frameshift-Stimulatory Pseudoknot. Journal of Molecular Biology. https://doi.org/10.1016/j.jmb.2020.09.006

9) Omar, S. I., Zhao, M., Sekar, R. V., Moghadam, S. A., Tuszynski, J. A., & Woodside, M. T. (2021). Modeling the structure of the frameshift-stimulatory pseudoknot in SARS-CoV-2 reveals multiple possible conformers. PLoS Computational Biology, 17(1), 2–10. https://doi.org/10.1371/journal.pcbi.1008603

10) Plant, E. P., Pérez-Alvarado, G. C., Jacobs, J. L., Mukhopadhyay, B., Hennig, M., & Dinman, J. D. (2005). A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. PLoS Biology. https://doi.org/10.1371/journal.pbio.0030172

11) Plant, E. P., Sims, A. C., Baric, R. S., Dinman, J. D., & Taylor, D. R. (2013). Altering SARS coronavirus frameshift efficiency affects genomic and subgenomic RNA production. Viruses. https://doi.org/10.3390/v5010279

12) Rangan, R., Watkins, A., Chacon, J., Kladwang, W., Zheludev, I., Townley, J., … Das, R. (2020). De novo 3D models of SARS-CoV-2 RNA elements and small-molecule-binding RNAs to aid drug discovery. BioRxiv : The Preprint Server for Biology. https://doi.org/10.1101/2020.04.14.041962

13) Ryder, S. P., Morgan, B. R., Coskun, P., Antkowiak, K., & Massi, F. (2021). Analysis of Emerging Variants in Structured Regions of the SARS-CoV-2 Genome. Evolutionary Bioinformatics, 17. https://doi.org/10.1177/11769343211014167

14) Schlick, T., Zhu, Q., Jain, S., & Yan, S. (2020). Structure-Altering Mutations of the SARS-CoV-2 Frame Shifting RNA Element.pdf. Biophysical Journal.

15) Schlick, Tamar, & Pyle, A. M. (2017). Biophysical Perspective Opportunities and Challenges in RNA Structural Modeling and Design. Biophysj, 113(2), 225–234. https://doi.org/10.1016/j.bpj.2016.12.037

16) Truniger, V., Miras, M., & Aranda, M. A. (2017). Structural and functional diversity of plant virus 3′-cap-independent translation enhancers (3′-CITEs). Frontiers in Plant Science, 8(November), 1–14. https://doi.org/10.3389/fpls.2017.02047

17) Wang, Z., Parisien, M., Scheets, K., & Miller, W. A. (2011). The Cap-Binding Translation Initiation Factor, eIF4E, Binds a Pseudoknot in a Viral Cap-Independent Translation Element. Structure, 19(6), 868–880. https://doi.org/10.1016/J.STR.2011.03.013

18) Zhang, K., Zheludev, I. N., Hagey, R. J., Wu, M. T.-P., Haslecker, R.; Hou, Y. J., Kretsch, R., … Das, R. (2020). Cryo-electron Microscopy and Exploratory Antisense Targeting of the 28-kDa.pdf. BioRxiv : The Preprint Server for Biology.

19) Zhou, J., Bean, R. L., Vogt, V. M., & Summers, M. (2007). Solution Structure of the Rous Sarcoma Virus Nucleocapsid Protein: μΨ RNA Packaging Signal Complex. Journal of Molecular Biology, 365(2), 453–467. https://doi.org/10.1016/j.jmb.2006.10.013

Appendix A

A.1 The Orphan yjdF Riboswitch crystallization and Surface Entropy Reduction of the Fab Framework Background

Riboswitches are a class of RNA elements ranging from roughly 50-300 nt in length which regulate translation of bacterial mRNAs through ligand binding events[4]. Phylogenic and homology analysis of bacterial genomes have recently identified numerous riboswitch sequences across bacteria[6]. Even when thousands of representatives exist of a given riboswitch across species the structure function and role of these riboswitches can remain unknown. Orphan riboswitches regulate open reading frames whose protein's identity and function is yet unknown. One approach to identifying these unknown genes is through ligand associations, generally a riboswitch will bind a ligand associated with the function of the protein so that translation of said protein can change in response to its ligand.

The yjdF riboswitch is one such orphan riboswitch[3]. It was first identified in Firmicutes and since 1060 representatives have been identified in Fimicutes, Actinobacteria, Fusobactiera, Spirochates, and Synergistetes. Homology analysis has generated a secondary structure and consensus sequence which was validated by inline probing. This class of riboswitches forms a cloverleaf like structure with four stems emanating from a highly conserved set of single stranded regions. These single stranded regions have a 90% conservation and in the presence of cognate ligand become more protected to inline probing suggesting that this structural architecture is highly conserved. In roughly 40% of the representative sequences an optional paired region follows P1 and precedes P2 this paired region being denoted P1.5. Nucleotides in P4 are less conserved than

other paired regions but contains an 11 nucleotide long stretch that shares complementarity with the downstream ribosome binding site suggesting that yjdF operates though ligand-dependent modulation of the ribosome-binding site access[3].

In line probing was performed on the 108-nucleotide long *B. Subtills* riboswitch in the presence of a screen of FMN riboflavin like ligands. These screens revealed that yjdF was capable of binding a diverse set of compounds with proflavine showing the greatest affinity for yjdF. As the ligand screen broadened a reporter assay was used to determine what compounds could cause yjdF to induce gene expression within a lacZ plasmid in *E. coli*. Only five compounds were capable of inducing gene expression: chelerythrine, harmine, dequalinium, acrifavine, and aminoacridine this nitrogen containing aromatic compounds are known as azzaaromatic compounds. Chelerythrine bound with 2.5 nM affinity while proflavine binds with 6.1nM affinity however many azzaaromatic compounds were found to bind yjdF with similarly high affinities making the natural ligand of this riboswitch impossible to determine. While these compounds are capable of intercalating in DNA and RNA the mode of ligand-yjdF binding occurs with equimolar stoichiometry implying a single specific binding site in yjdF. To determine the mechanism of ligand binding I applied our chaperone assisted RNA crystallography approach to crystalize a representative yjdF riboswitch.

A.2 Results

A.2.1 Construct design and initial crystallization

Mutational analysis has shown that the shape and structure of each of the paired regions and their hairpin loops is required for ligand binding, with the exception of P1.5 and P4. Crystallization constructs were designed for yjdF species for which inline probing had been performed previously.

The yjdF from *R. Gauvreauii* and *L. Plantarum* loop 1.5 and loop 3 were mutated to contain the BL3-6 epitope. Only the R. Gauvreauii loop 1.5 construct was capable of binding Fab. Crystals were readily grown of this construct with wild type BL3-6 but diffraction never improved below 8Å. To address this poor diffraction, I applied a surface entropy reduction strategy to the Fab framework.

A.2.2 The Serine Mutant

The initial work on surface entropy reduction of our Fab framework was carried out by a collaborator (Dr. Jingdong Ye) in the context of Fab2, which binds to the P4-P6 independently folding RNA domain. Ye and coworkers mutated six residues located in three distinct patches to alanine only and serine only[5]. Compared to the wild-type Fab2, this new version of the antibody fragment improved the resolution of the diffraction obtained from crystals of the complex, but the crystals did not diffract beyond four angstrom resolution, precluding structure determination. Extending this work, I applied this initial approach to Fab BL3-6, by mutating the same sites in BL3-6 all to alanine in an A mutant version of BL3-6 and in a second construct these sites were all mutated to serine in an S mutant version of BL3-6. The S mutant crystalized with yjdF bound

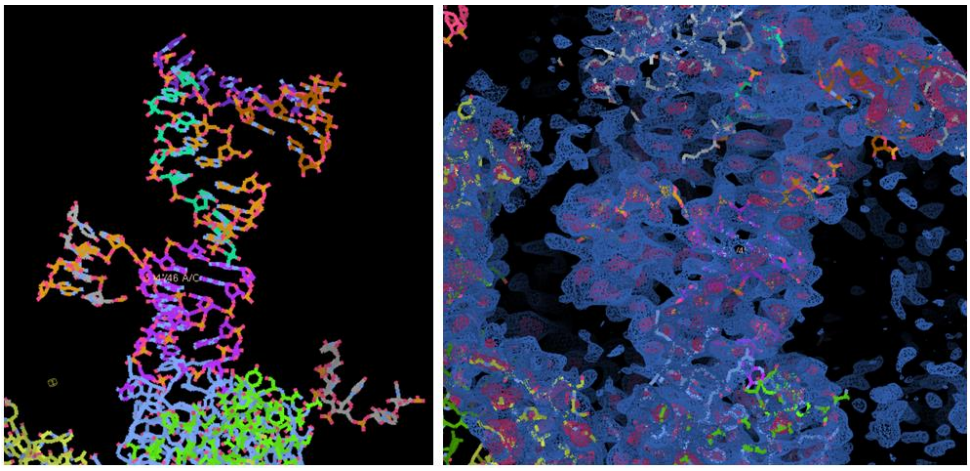


Figure A.1. Partial yjdF structure and corresponding density revealed by density modification

to chelerythrine in two conditions in a Hampton screen: 100 mM HEPEs pH 7.5, 25% PEG 3350 and 200 mM Ammonium Acetate and 50mM Bis-Tris pH 6.5, 25% PEG 3350 and 200mM Ammonium acetate. Crystals from the former condition gave 3.1Å diffraction with a high degree of completeness. Molecular replacement with the Fab-BL3-6 model yielded a solution in the P1 space group but RNA density was very incomplete. Density modification was performed in an attempt to amplify the signal of structured RNA to solvent noise but even after these improvements to the map only 30% of the RNA mass had usable density. To improve the phasing power these crystals were re-grown and soaked in various heavy metals to collect anomalous signal and improve the accuracy of the phases experimentally.

A.2.3 Heavy metal soaking of yjdF S-mutant crystals

Crystals of yjdF and S-mutant BL3-6 and wild type BL3- were grown in larger 2uL drops screening around the initial hit conditions. These crystals were soaked for 24 hours in cryoprotectant which matched the well solution but was supplemented with 5% glycerol and 10mM heavy metal complex from a Hampton screen. The crystals were then looped and shot using a tunable beamline at APS 24-ID-E. Only crystals soaked in Osmium generated a diffraction signal strong enough to calculate experimental phases when combined with molecular replacement. However, the occupancy of the osmium ions was only 10% so the phasing power they provided was minimal and the resulting electron density map was no more complete than the density modified map even after density modification.

A.2.4 The Surface Entropy Reduction Project

The goal of surface entropy reduction to improve a target's likelihood of crystalizing is to mutate larger flexible residues which would normally be entropically costly to restrain in a crystal contact

into patches of short residues like alanine or serine[1]. These patches can serve as sticky surfaces that are particularly amenable to forming crystal contacts with minimal entropic energy expense. To enable a more systematic surface entropy reduction engineering strategy, I analyzed the existing crystal structures of BL3-6 bound to different RNAs to identify surface residues that I could potentially mutate. I used the SERp server to identify sites along both chains that would likely be productive for surface entropy reduction mutations[2]. Comparing the B-factors for each of the Fab-RNA structures the Piccirilli lab has solved, I identified regions of structural flexibility and inconsistency. When visually scanning the surface for protruding residues I also made note of weather any of these residues formed crystal contacts in these structures. Combining these pieces of information, I selected patches that were the best candidates for surface entropy reduction. First, the highly unstructured C terminal tail of the heavy chain BL3-6 was truncated by four residues as in the serine mutant, and all other mutations were installed in this background. Among the selected patches were single sites where a lone glutamine or lysine protruded from what otherwise would be a flat surface; other patches included series of lysine residues along one edge of the Fab that do not engage in crystal contacts. Ultimately, I selected twelve patches to mutate. Previously we found that making many mutations to the Fab framework at once is detrimental to Fab expression; therefore, each patch was mutated and tested independently of other patches. After successfully cloning each of the mutants, Fab expression was verified in a small-scale test. Unexpectedly, all twelve mutants were able to express in bacterial cell culture. After large-scale expression and purification, I set up crystallization trials using these new surface-entropy-reduced antibody fragments and the riboswitch yjdF RNA. Each of the mutant Fabs in complex with yjdF yielded visible crystals at rates similar to wildtype.

I set up broad crystallization trials of the SER mutants in complex with yjdF in the presence of ligand to assesses crystallizability and quality of diffraction. Many of the SER mutants formed crystals in new conditions. Upon optimizing these conditions using an additive screen I found that SER mutants 200, 170, and 150 yielded crystals that diffracted at or below 4Å. Interestingly, the SER Fabs that gave the best diffraction were not the Fabs that crystallized most readily in the broad screens. To further investigate the SER mutants' potential to improve diffraction I set up additive screens around the original S-mutant crystallization conditions (100mM HEPES pH 7.5, 25% PEG 3350, 200mM Ammonium Acetate). I found that SER Fabs 40 and 146 grew multiple large crystals, many of which diffracted to at least 5Å; a few diffracted to 3Å and a 146-crystal diffracted to 2.75Å with strong signal and completeness. Additionally, SER mutants 191 and 170 diffracted to about 4-5Å . These results suggest that SER mutants 200, 170, 150, 40, 146, 141, 191 are worth further investigation with additional RNA targets. In the PEMV2 chapter the structure I report was solved using SER mutant 170 for instance.

In solving the diffraction data set of yjdF bound to 146 I found a molecular replacement solution using the structure of BL3-6 and the RNA hairpin fragment bound to the Fab's CDRs. Iterative rounds of building the RNA off of the fragment followed by refinement yielded increasingly poor Rwork/Rfree ratios, suggesting model bias. To correct for this, I preformed statistical density modification to amplify the signal of the RNA phases. The additional density revealed a clear set of beta sheets. Upon repeating molecular replacement but searching for two Fabs in the asymmetric unit there appeared to be no RNA present in this crystal beyond the small RNA fragment containing the BL3-6 hairpin loop. The RNA degraded under the crystallization conditions except for the AAACA loop, which appeared to be protected by Fab binding. To confirm that the RNA was absent from the crystal I ran the mother liquor of crystallization trays

containing yjdF- chelerythrine bound to 146, wt BL3-6 and a few other BL3-6 variants in a denaturing PAGE. I found that drops of yjdF 146 mother liquor only contained a small RNA fragment and no full length yjdF RNA. I also found that drops of yjdF mother liquor bound to other BL3-6 derivatives contained both full length yjdF and partially degraded yjdF RNA. This suggests that yjdF has degradation sensitive regions and that Fab 146 may have had RNAase contamination.

A.2.5 Iridium hexamine soaking

New crystals of yjdF and Fab BL3-6 were grown in the original 100 mM HEPEs pH 7.5, 25% PEG 3350 and 200 mM Ammonium Acetate condition in 2uL hanging drops. A subset of these crystals were soaked with iridium hexamine following the method established in the Programmed Ribosomal Frameshifting element structure chapter. X-ray diffraction data was collected using the tunable beam line and strong anomalous signal was detected. SAD-MR phasing was performed on a 3.14Å dataset and successfully and found the Fab and 13 atoms of iridium with strong occupancy. Resolve density modification was once again applied to this data set and despite the additional phasing power from the iridium no new RNA density emerged. Similarly building RNA into the density and refining yielded no additional improvements to the electron density map. Interestingly the initial MR-SAD generated electron density map closely resembled the final statistical density modified map from the first yjdF S-mutant BL3-6 data set suggesting that these regions of the RNA are the only portion which are ordered and capable of producing a diffraction pattern. The missing density from both the Iridium-soaked crystals and the native data set is more likely the result of poorly ordered RNA than a lack of phasing accuracy. If the RNA were well ordered the experimentally determined phases contributing to the electron density map should have revealed

even partial new density but not only were the maps not improved, they were nearly identical to the first density modified maps generated with the S-mutant.

A.3 Conclusions

The yjdF riboswitch is too flexible to generate internally ordered crystals. This RNA element may be better suited to studies using Cryo-EM where it can remain in solution.

A.4 Methods

A.4.1 RNA Transcription and Purification.

Single stranded DNA templates and primers for PCR and transcription were ordered from IDT encoding the transcription template for each RNA construct with a T7 promoter. Forward primers were ordered matching the T7 promoter region and reverse primers contained a single 2'O-methyl modification at the 3' end to avoid untemplated additions by T7 polymerase. Transcription template DNA was amplified into double stranded DNA using PCR. RNA was transcribed from the purified PCR product using an in vitro transcription reaction as follows 50 pmol mL^{-1} DNA template was incubated for 3 h at 37 °C in buffer containing 40 mM Tris-HCl, pH 8.0, 2 mM spermidine, 10 mM NaCl, 25 mM $MgCl_2$, 10 mM DTT, 40 U mL^{-1} RNase inhibitor, 5 U mL^{-1} thermostable inorganic pyrophosphatase, 5 mM of each NTP, and 50 μg mL^{-1} T7 RNA polymerase. Reactions were halted by addition of RNase free DNase1 at 5U mL^{-1} and incubation at 37 ^{o}C for 30 minutes. RNA was purified on a 10% denaturing polyacrylamide gel in 0.5x TBE running buffer. The RNA was visualized with UV shadowing, extracted and eluted into 10 mM Tris, pH 8.0, 2 mM EDTA, 300 mM NaCl buffer via overnight incubation at 4 ^{o}C. The eluted RNA was then concentrated and exchanged into double distilled H_2O using a 10K Amicon filter and stored at -80 ^{o}C until further use.

A.4.2 Fab Purification.

The BL3-6 Fab expression vector (available upon request) was transformed into 55244 chemically competent cells (www.atcc.org) and grown on LB plates supplemented with carbenicillin at 100 μg mL^{-1}. Nine colonies from the plates were chosen and inoculated to a starter culture with 100 μg mL^{-1} carbenicillin, which was grown at 30 ºC for 8 hours. Once the starter culture reached an OD 600 of 8, 15mLs of starter culture was used to inoculate 1L of 2×YT media and grown for 24 h at 30 °C. The cells were then pelleted via centrifugation at RT, and the cell pellet was resuspended in 1L of freshly prepared CRAP-Pi media supplemented with 100 ug mL^{-1} carbenicillin. The cells were set to grow for 24 h at 30 °C, harvested via centrifugation at 4 ºC and frozen at -20 ºC. Frozen cell pellets were lysed in PBS buffer supplemented with 0.4 mg mL^{-1} of Lysozyme and 0.01 mg mL^{-1} of DNase I. After 30 minutes PMSF was added to a final concentration of 0.5 mM. After 30 minutes, the cells were centrifuged, 45 min, 12000 rpm, at 4 ºC. Lysate was transferred to new sterile bottles and centrifuged again for 15 minutes, 12000 rpm, at 4 ºC. Supernatant was filtered through 0.45 μm filters into a sterile bottle (Millipore Sigma, www.sigmaaldrich.com), and Fab proteins were purified using the AKTAxpress fast protein liquid chromatography (FPLC) purification system (Amersham, www.gelifesciences.com) as described previously.[37] The lysate in PBS buffer (pH 7.4) was loaded into a protein A column, and the eluted Fab in 1 M acetic acid was buffer exchanged back into the buffer PBS (pH 7.4) using 30kDa cutoff Amicon filter and loaded into a protein G column. The Fab was eluted from protein G column in 0.1 M glycine (pH 2.7) and then buffer-exchanged into 50 mM NaOAc, 50 mM NaCl buffer (pH 5.5) and loaded into a heparin column. Finally, the eluted Fab in 50 mM NaOAc, 2 M NaCl (pH 5.5) was dialyzed back into 1× PBS (pH 7.4), concentrated, and analyzed by 12% SDS-PAGE using Coomassie Blue R-250 staining for visualization. Aliquots of Fab samples were tested for

RNase activity using the RNaseAlert kit (Ambion, www.thermofisher.com). The aliquots of Fab samples were flash frozen in liquid nitrogen and stored at −80 °C until further use.

A.4.3 Crystallization complex formation

Complexes of yjdF and the S-Mutant of BL3-6 were generated by refolding 480ug of the RNA in buffer containing 50 mM Tris-HCl pH 7.5, 100 mM KCl, 10 mM MgCl2 by heating the RNA in water to 90 C for 1 min then adding buffer and refolding the RNA at 50C for 10 mins followed by an incubation at 25C for 5 minutes, after which point 2 molar equivalents of chelerythrine were added to form ligand bound complex. Then 1.1 molar equivalents of Fab BL3-6 wild type or their Surface Entropy reduced mutants were added and the complex was incubated on ice for 30 minutes before concentrating to 6mg/mL RNA concentration and setting hanging drop crystallization trials in pre made screens.

A.5 Works Cited in Appendix A

1) Derewenda, Z. S., & Godzik, A. (2017). The “Sticky Patch” Model of Crystallization and Modification of Proteins for Enhanced Crystallizability. In Methods in molecular biology (Clifton, N.J.) (Vol. 1607, pp. 77–115). https://doi.org/10.1007/978-1-4939-7000-1_4

2) Goldschmidt, L., Eisenberg, D., & Derewenda, Z. S. (2014). Salvage or Recovery of Failed Targets by Mutagenesis to Reduce Surface Entropy. In Methods in molecular biology (Clifton, N.J.) (Vol. 1140, pp. 201–209). https://doi.org/10.1007/978-1-4939-0354-2_16

3) Li, S., Hwang, X. Y., Stav, S., & Breaker, R. R. (2016). The yjdF riboswitch candidate regulates gene expression by binding diverse azaaromatic compounds. RNA (New York, N.Y.), 22(4), 530–541. https://doi.org/10.1261/rna.054890.115

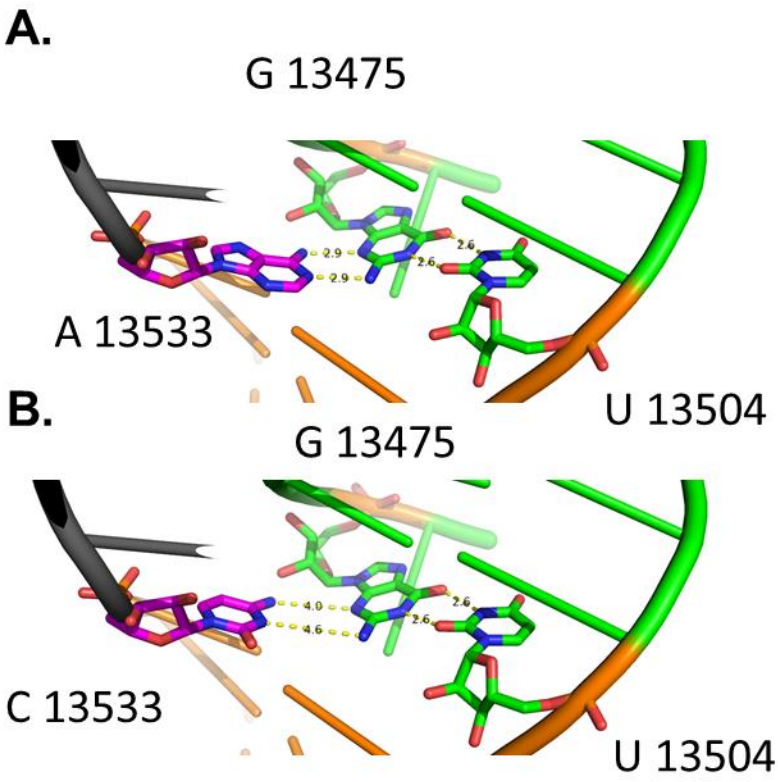
4) Pavlova, N., Kaloudas, D., & Penchovsky, R. (2019). Riboswitch distribution, structure, and function in bacteria. Gene, 708(April), 38–48. https://doi.org/10.1016/j.gene.2019.05.036

5) Ravindran, P. P., Heroux, A., & Ye, J.-D. (2011). Improvement of the crystallizability and expression of an RNA crystallization chaperone. Journal of Biochemistry, 150(5), 535–543. https://doi.org/10.1093/jb/mvr093

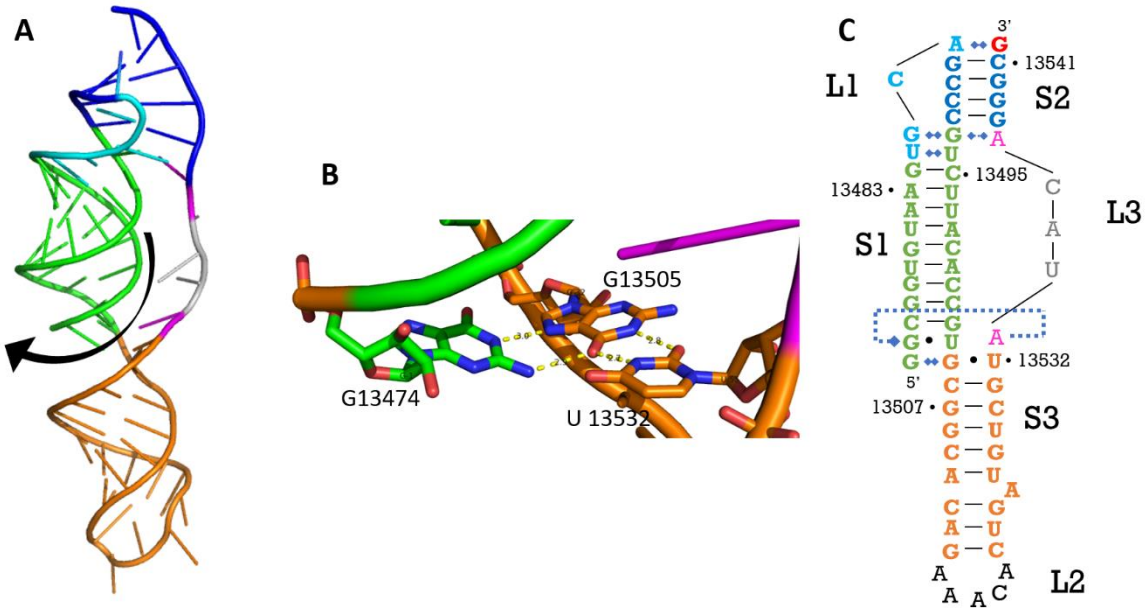
6) Weinberg, Z., Wang, J. X., Bogue, J., Yang, J., Corbino, K., Moy, R. H., & Breaker, R. R. (2010). Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biology, 11(3). https://doi.org/10.1186/gb-2010-11-3-r31

Appendix B

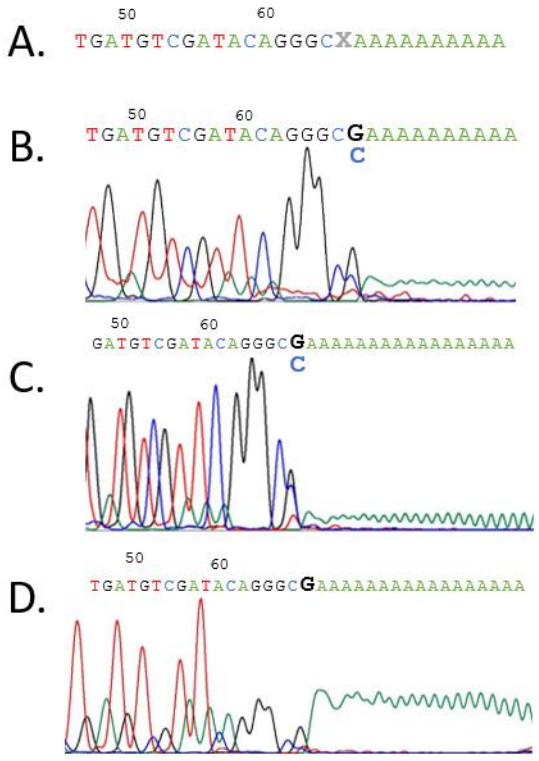
Supplementary Information



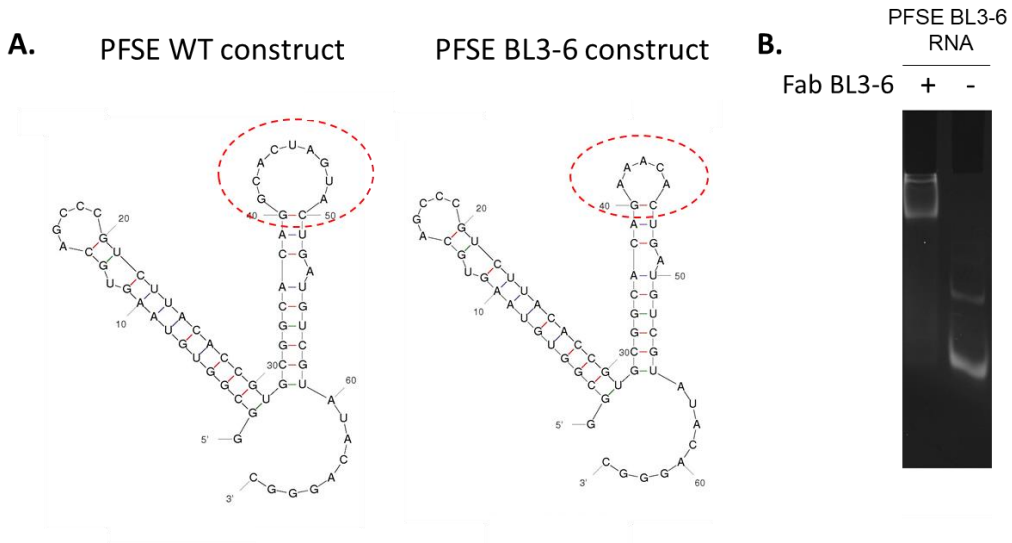
Supplementary Fig. B.1 Comparison of position 13533 as an A and as a C. (A) Close up view of the SARS-CoV-2 PFSE showing position A13533, modeled as magenta-colored sticks, forming hydrogen bonds with G13475, which is paired with U13504 (both green). (B) The image displayed in A but with position 13533 mutated to C to show the potential to form the same hydrogen bonding interactions as (A).



Supplementary Fig. B.2 Observed 5’ end interactions of the PFSE. (A) 5’ end threading of the PFSE illustrated by an arrow. (B) Hydrogen bonding interactions of G13474 with G13505-U13532 wobble pair. (C) Location of panel (B) in relation to the secondary structure map.



Supplementary Fig. B.3 Chromatograms of PFSE RNA sequencing. (A) Sequence of the PFSE BL3-6 RNA with unknown nucleotide in position 65 flanked by poly(A) tail. (B) PFSE BL3-6 RNA sequenced after transcription. (C) PFSE BL3-6 RNA sequenced from the crystallization drop mother liquor. (D) PFSE BL3-6 RNA sequenced from the crystal.



Supplementary Fig. B.4 Construct validation. (A) V-fold[2] predictions of the wild type PFSE sequence and the PFSE BL3-6 sequence demonstrating that mutating Loop 2 to AAACA would not be expected to create new long distance base pairing interactions that could interfere with the native structure. (B) EMSA in 10% native gel demonstrating RNA shift in the presence of Fab BL3-6.

Table B.1 Sequences of RNA Constructs and DNA Oligonucleotides

Construct Name	Sequence
WT PFSE RNA	5’ GGC GGU GUA AGU GCA GCC CGU CUU ACA CCG UGC GGC ACA GGC ACU AGU ACU GAU GUC GUA UAC AGG GC 3’
BL3-6 PFSE	5’ GGC GGU GUA AGU GCA GCC CGU CUU ACA CCG UGC GGC ACA G AAACA CU GAU GUC GUA UAC AGG GC 3’
WT PFSE transcription template DNA oligo	5’ GCG TAA TAC GAC TCA CTA TA GGC GGT GTA AGT GCA GCC CGT CTT ACA CCG TGC GGC ACA GGC ACT AGT ACT GAT GTC GTA TAC AGG GC 3’
WT PFSE transcription template DNA forward primer	5’ GCGTAATACGACTCACTATAGG 3’
WT PFSE transcription template DNA reverse primer	5’ (2’O-Me)-GCCCTGTATACGACATCAG 3’
BL3-6 PFSE transcription template DNA oligo	5’ GCG TAA TAC GAC TCA CTA TA GGC GGT GTA AGT GCA GCC CGT CTT ACA CCG TGC GGC ACA G AAACA CT GAT GTC GTA TAC AGG GC 3’
BL3-6 PFSE transcription template DNA forward primer	5’ GCGTAATACGACTCACTATAGG 3’
BL3-6 PFSE transcription template DNA reverse primer	5’ (2’O-Me)-GCCCTGTATACGACATCAG 3’
Reverse transcription reverse primer (M13 for + 18 T’s)	5’ GTAAAACGACGGCCAGTTTTTTTTTTTTTTTTTT 3’
Reverse transcription forward primer (M13)	5’ GTAAAACGACGGCCAGT 3’
Reverse primer (PFSE specific)	5’ GGCGGTGTAAGTGCAG 3’

Works Cited in Appendix B

(1) Xu, X.; Zhao, P.; Chen, S.-J. Vfold: A Web Server for RNA Structure and Folding Thermodynamics Prediction. *PLoS One* **2014**, 9 (9), e107504. https://doi.org/10.1371/journal.pone.0107504.