Biophysical *Journal*
Article

Biophysical Society

Structure-Altering Mutations of the SARS-CoV-2 Frameshifting RNA Element

Tamar Schlick,[1,2,3,*] Qiyao Zhu,[1,2] Swati Jain,[1] and Shuting Yan[1]
[1]Department of Chemistry and [2]Courant Institute of Mathematical Sciences, New York University, New York, New York; and [3]NYU-ECNU Center for Computational Chemistry, NYU Shanghai, Shanghai, P. R. China

ABSTRACT   With the rapid rate of COVID-19 infections and deaths, treatments and cures besides hand washing, social distancing, masks, isolation, and quarantines are urgently needed. The treatments and vaccines rely on the basic biophysics of the complex viral apparatus. Although proteins are serving as main drug and vaccine targets, therapeutic approaches targeting the 30,000 nucleotide RNA viral genome form important complementary approaches. Indeed, the high conservation of the viral genome, its close evolutionary relationship to other viruses, and the rise of gene editing and RNA-based vaccines all argue for a focus on the RNA agent itself. One of the key steps in the viral replication cycle inside host cells is the ribosomal frameshifting required for translation of overlapping open reading frames. The RNA frameshifting element (FSE), one of three highly conserved regions of coronaviruses, is believed to include a pseudoknot considered essential for this ribosomal switching. In this work, we apply our graph-theory-based framework for representing RNA secondary structures, "RAG (or RNA-As-Graphs)," to alter key structural features of the FSE of the SARS-CoV-2 virus. Specifically, using RAG machinery of genetic algorithms for inverse folding adapted for RNA structures with pseudoknots, we computationally predict minimal mutations that destroy a structurally important stem and/or the pseudoknot of the FSE, potentially dismantling the virus against translation of the polyproteins. Our microsecond molecular dynamics simulations of mutant structures indicate relatively stable secondary structures. These findings not only advance our computational design of RNAs containing pseudoknots, they pinpoint key residues of the SARS-CoV-2 virus as targets for antiviral drugs and gene editing approaches.

SIGNIFICANCE   Since the outbreak of COVID-19, numerous projects have been launched to discover drugs and vaccines. Compared with protein-focused approaches, targeting the RNA genome, especially highly conserved crucial regions, can destruct the virus life cycle more fundamentally and avoid complications of viral mutations. We choose to target the small frameshifting element (FSE) embedded in the open reading frame 1a,b of SARS-CoV-2. This FSE is essential for translating overlapping reading frames and thus controlling the viral protein synthesis pathway. By applying graph-theory-based computational algorithms, we identify structurally crucial residues in the FSE as potential targets for antiviral drugs and gene editing.

INTRODUCTION

The novel coronavirus SARS-CoV-2, the agent of the COVID-19 pandemic that has upended our world in a very short time, poses an international health emergency of unprecedented proportions. With rapid transmission and high death rates, this virus threatens the lives and livelihood of billions of people worldwide. Scientists have been quick to rise to the task to investigate collaboratively many viable paths of defense. These include development of better techniques for detection and tracing of infection, identification of viable compounds that mitigate the virus's harm, and development of vaccines.

Some clues into the complex virus apparatus involved in the infection process are available from highly related cousins of this pathogenic coronavirus. From the SARS (SARS-CoV) and MERS (MERS-CoV) viral outbreaks, we have a reasonable mechanistic understanding of the SARS-CoV-2 virus life cycle and the infection process. This includes the cellular invasion by the viral spike (or S) protein and its binding to the ACE2 cellular receptors and the triad of main proteins essential for producing a complete viral particle (envelope, membrane, and nucleocapsid) (1–3). In addition, scientific advances since the SARS and

MERS outbreaks, along with accumulating data about the disease since its emergence in late 2019 from China, provide insights into the virology, pathology, and empirical treatments of such infections (1–3).

From this information and knowledge of other viruses such as HIV-AIDS, Ebola, influenza, hepatitis C, or malaria (4,5), we have suites of antiviral and other compounds that can potentially mitigate the damage of the virus on the body, including remdesivir, tocilizumab, favipiravir, galidesivir, brilacidin, and dexamethasone (5). In particular, remdesivir, a viral-specific RNA dependent RNA polymerase inhibitor, which resembles the structure of an HIV reverse-transcriptase inhibitor, has shown promise for reducing recovery time of patients in acute stages; the steroid dexamethasone, used widely to fight serious inflammatory reactions, was shown to reduce COVID-19 deaths. Drugs that block spike, membrane, or envelope proteins on the viral shell of coronaviruses are actively being sought, and some are being investigated in clinical trials.

With rapid deaths, as well as second and third waves of infection already seen in some countries, scientific data are critically needed to help guide clinical efforts and offer genomic, bioinformatic, and biophysical insights into effective disease treatment approaches, both short and long term. Because a vaccine is not imminent and new coronavirus epidemics can be anticipated, a better understanding of the viral machinery and potential therapeutic avenues that are not currently widely explored are highly desirable.

SARS-CoV-2, like other coronaviruses, has an $\sim$30,000 base RNA genome. The viral RNA agent hijacks the host cells' ribosome machinery to replicate and assemble itself by synthesizing a suite of viral proteins and thus continue to invade host organs. An attractive alternative line of research to protein-targeting approaches involves compounds that degrade and thus disarm the main workhorse of the virus, the RNA genome (6,7). RNA-targeting approaches offer therapeutic potentials because of the high sequence and structure conservation of the untranslated regions of the viral genomes; they nonetheless pose physiochemical challenges because of their more complex structures. With rapid emergence of CRISPR technology, however, such approaches are becoming more viable and valuable. Indeed, CRISPR/Cas13d technology may be used in connection with a guided-RNA delivery system to edit the SARS-CoV-2 viral genome in strategic locations such as the open reading frame ORF1a,b and the spike gene of the virus, with the aim of disrupting key replication or invasion processes (7). Two vaccine candidates based on the RNA and bankrolled by the US government are under advanced clinical trials (8): an mRNA vaccine by Moderna and an adenovirus vector by Johnson & Johnson.

One of the critical steps in viral replication is the translation of the overlapping, but shifted, ORF1a,b gene region, which codes for the unstructured polypeptides that start the cascade of viral protein synthesis. This is achieved via the process of −1 programmed ribosomal frameshifting, a mechanism to stall ribosomal translation utilized by many viruses to handle overlapping reading frames (9). The frameshifting element (FSE) is a small region (less than 100 nucleotides) in the middle of the ORF1a,b gene region; it is believed that such mechanisms rely on specific structural modules and/or associated structural transitions (10). In the SARS-CoV virus, this FSE was identified as a three-stem structure, in which stems 1 and 2 intertwine via a pseudoknot or an intertwined basepair region (11,12). Drug screening for this structure had suggested a potential drug to a central loop region (10,13), and recent research suggests its promise for SARS-CoV-2 as well (14).

Here, we expand upon these insights for SARS-CoV-2 applications using our graph-theoretic approach, "RAG" (RNA-As-Graphs), which uses tree and dual graphs to represent, study, predict, and design RNA secondary (2D) structures (15–20). Our dual graphs represent RNA stems and helices as graph vertices and loops as edges. These graph objects can handle RNA pseudoknot motifs. By focusing on the overall connectivity or shape of the RNA 2D structural elements, RAG substantially simplifies the RNA structure, making it insensitive to the number of residues in stems and loop regions. Because our approach is general, similar techniques can be applied to other regions of the viral genome.

Here, we investigate structural features of the FSE of SARS-CoV-2 and predict minimal mutations that destroy key structural features thought to be critical to its function, thus providing targets for antiviral or gene therapy. Specifically, we use our inverse-folding procedure, RAG-IF, which utilizes a genetic algorithm to systematically mutate RNA sequences to fold onto a target graph motif (21). RAG-IF relies on our computational pipeline to design RNA sequences for target graphs from constituent building blocks (22) based on fragment assembly of subgraphs (23,24) using a library of subgraphs and corresponding three-dimensional (3D) structures of solved RNAs (25). Such RNA design is a computationally efficient approach that can generate hundreds of candidate sequences and then sort them for minimal mutations. The proposed candidates from our original design pipeline have already shown promise through experimental chemical reactivity testing (SHAPE) (24).

We first identify the SARS-CoV-2 FSE region (based on the SARS-CoV FSE sequence) and obtain its consensus structure with eight different secondary structure prediction programs; we represent this region as dual graph 3_6, three stems with a pseudoknot intertwining stems 1 and 2 (Fig. 1, A–C). Next, we use RAG-IF (21), modified here to handle pseudoknots, to identify critical residues that can transform the FSE using minimal mutations to closely related graph motifs that correspond to other known RNA structures chosen from our dual-graph motif atlas (26). Intriguingly, we find that mutating only two to three critical residues in the
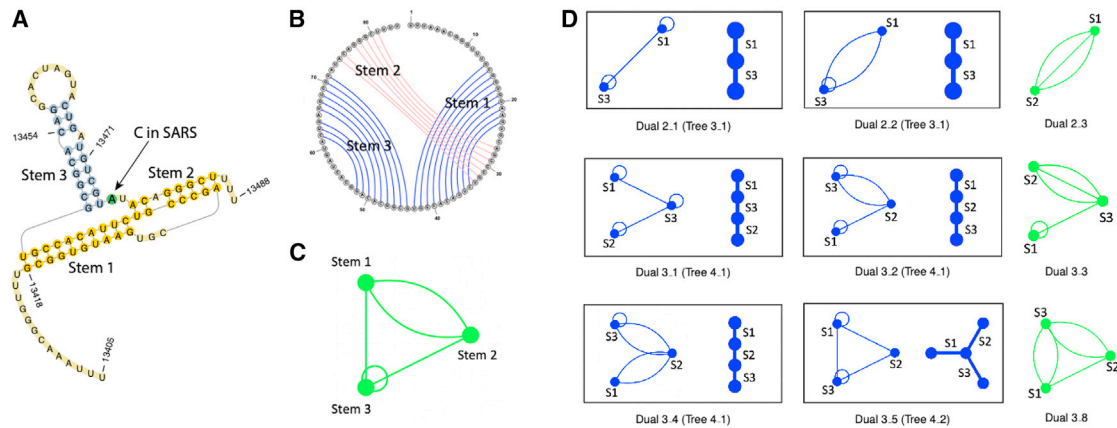
FIGURE 1 Consensus structure of the SARS-CoV-2 frameshifting element (FSE) and related dual-graph targets. (*A*) 2D structure of the FSE, a three-stem pseudoknot motif, is shown. (*B*) Circular plot of the FSE 2D structure, in which the pseudoknot corresponds to the crossing of the two basepairs, is given. (*C*) Associated dual-graph motif 3_6 for the FSE 2D structure is shown. (*D*) Dual graphs with two or three vertices related to the FSE 3_6 motif are shown. In dual graphs, stems in RNA secondary structures are vertices, whereas junctions, bulges, and loops are edges (15). The targets containing pseudoknots are shown in green. Those not containing pseudoknots are also represented as tree graphs, in which stems denote edges and junctions, bulges, and loops are vertices. To see this figure in color, go online.

FSE can alter its structure dramatically. We also assess the stability of these mutants by compensatory mutations using the same inverse-folding techniques; this essentially provides a check on our predictions and a measure of robustness, i.e., how easily "reversible" these mutations are. Next, we report emerging structural insights from microsecond molecular dynamics (MD) simulations of the wild-type and mutants as additional measures of structural stability. We conclude with a summary and discussion of our findings, including a figure with key drug-binding or gene-editing target residues emerging from this work, implications for COVID-19 therapy, and future plans.

MATERIALS AND METHODS

Identification of the SARS-CoV-2 FSE

The SARS-CoV FSE has been identified as the slippery site UUUAAAC and a downstream three-stem pseudoknot-containing structure (12,27). Taking this information as a guide, we used RNAMotif (28) to search for the same nucleotide pattern (bracketed by 5′-UUUAAAC, an H-type pseudoknot, and UUU-3′) in SARS-CoV-2 complete RNA genome (GenBank: MT_246482). This led to the identification of an 84-nucleotide sequence in the viral genome of SARS-CoV-2, with a single-nucleotide substitution as compared with SARS-CoV FSE (Fig. 1 A).

Secondary structure prediction packages

We use eight RNA 2D structure prediction packages that handle pseudoknots to predict the 2D structures of 84-residue and 77-residue (without the slippery site) FSE of SARS-CoV-2: PKNOTS (29), NUPACK (30), IPknot (31), Hotknot (32), Vfold2D (33), SPOT-RNA (34), ProbKnot (35), and vsFold5 (36). PKNOTS and NUPACK were installed locally, and webservers were used for the others. The structures were first predicted using the default parameter sets for all eight packages; if the default set did not yield the expected three-stem pseudoknot structure, other parameter sets available on the webservers were also utilized.

Figs. S1 and S2 show the minimal free-energy structures predicted by the eight programs, using the default or alternative parameter set, for both the 84-residue and 77-residue FSE.

RAG dual graphs

In RNA 2D structures, double-stranded (basepaired) regions are denoted stems or helices, and single-stranded regions are loops (15–17). The rules for representing 2D structures as dual graphs are as follows:

1) Stems with at least two basepairs are represented as vertices.
2) Each single strand in bulges, internal loops, and junctions is represented as an edge that connects vertices. Bulges or internal loops with only one residue on each strand are ignored.
3) Hairpin loops are represented as self edges.
4) Unpaired residues at the 5′ and 3′ ends are ignored.

Fig. 1, A and C show the 2D structure of the SARS-CoV-2 FSE and its corresponding dual graph, 3_6. Our dual-graph enumeration algorithm (26) uses graph theory to describe all possible dual graphs for two to nine vertices. There are 110,667 dual graphs in our current dual-graph library, and we show those with two to five vertices in Fig. S3.

RAG-IF for minimal mutations

We use our inverse-folding protocol RAG-IF (21) to mutate the 77-residue FSE sequence to fold onto closely related dual-graph motifs. Our RAG-IF program defines the inverse folding with mutations section of our computational pipeline for novel RNA design (21,24). For a target tree graph (i.e., no pseudoknots), our original design pipeline generates a pool of RNA sequences, some of which fold onto the target graph, as determined by two 2D structure prediction packages. For the unsuccessful sequences, RAG-IF determines minimal mutations that lead these sequences to fold onto the target topology.

RAG-IF has three steps: 1) identify mutation regions and target 2D structures, 2) produce candidate sequences by mutations using a genetic algorithm (GA), and 3) optimize the mutated sequence pool by sorting and retaining only minimal or essential mutations that fold onto the target graph.

For tree graphs, all three steps are automated. However, to use RAG-IF for dual graphs capable of handling pseudoknots, we manually choose the target residues for mutation (step 1), modify aspects of the GA for

Schlick et al.

computational efficiency, and change the 2D structure prediction packages to handle pseudoknots for steps 2 and 3.

Selection of mutation regions

We compare the wild-type FSE graph and the target graph to identify the smallest possible mutation regions required for the transformation. Specifically, we focus on breaking the FSE pseudoknot or stem 2. Here, we describe how this is accomplished for our four most successful target graphs, 3_5, 3_2, 2_1, and 3_3 (shown in Fig. 1 D).

Note that these targets are topologically defined. Thus, changes in the length of the stem or the loop do not change the target dual graph as long as the number of stems and their connectivity remain the same.

Target dual graph 3_5 is a three-way junction, without any pseudoknots, and it corresponds to RNAs like the thi-box riboswitch. Two of the three stems of the FSE (stems 1 and 3) are also present in the 3_5 graph; hence, they are kept intact in the transformation. To break the FSE pseudoknot or stem 2, we seek residues at the 3′ end of the FSE to form a new stem with the 5′ end (shown in Fig. 2 A). Hence, we select residues 71–77, part of stem 2 and the 3′ end, as targets for mutation.

Target graph 3_2 has three stems connected by a single-stranded region and an internal loop, and it includes the original stem 3. This graph corresponds to RNAs like U6 small nuclear RNA. For this transformation, we use the 5′ end and hairpin loop of stem 1 to form a new stem. To pair the residues involved, we select the 3′ strand of the original stem 1, residues 27–33, as the mutation region shown in Fig. 3 A.

Target graph 2_1, corresponding to RNAs such as the cleaved hammerhead ribozyme, has two stems and a connecting single-stranded region. For this transformation, we aim to keep stems 1 and 3 intact but destroy stem 2 while avoiding two other potential pseudoknots shown in Fig. 4 A. Hence, we select residues 22–24 and 70–74 of stem 2 and the 3′ end residues 75–77 for mutation.

Target graph 3_3 has three stems and a structurally different pseudoknot. It corresponds to RNAs like the internal ribosome entry site (Protein Data Bank, PDB: 2IL9). To break the FSE pseudoknot, we want the 3′ strand of stem 2 to bind with the hairpin loop of stem 3 to form a new pseudoknot. However, to avoid a another potential pseudoknot (shown in Fig. 5 A), we select residues 22–24 and the 3′ strand of stem 2 (residues 68–73) as mutation regions.

Genetic algorithm application

GAs mimic evolution in nature. A population with a defined fitness undergoes random mutation, crossover, and selection, and those with high fitness are retained. In our case, the individuals in the population are RNA sequences, and the fitness function is the number of residues that have the same 2D structure in both the target (provided for fitness calculation) and the predicted structure, as determined by the Hamming distance. Based on prediction performance on the 77-residue FSE and computational complexity (see below), we choose NUPACK for our GA. The initial population is obtained by randomly assigning nucleotides to the mutation region in the RNA sequence. This population is then subject to 100 iterations of random mutation, crossover, selection, and nomination (see (21) for details).

For minimal mutation search, we use a population of 100 sequences (and 10 sequence replacement in the selection step). The GA terminates if 200 high-fitness sequences are nominated or the execution time exceeds 6 h. The nominated sequences are further screened by prediction program PKNOTS, and only those that satisfactorily fold onto the target graph by both NUPACK and PKNOTS are retained.

For compensatory mutation search, the starting sequences are the mutants, the mutation regions are the same, and the target folding is the 2D structure predicted by NUPACK for the wild-type 77-residue FSE. Expanded RAG-IF runs (500 sequences) are used for targets 2_1 and 3_3, with 50 sequence replacement instead of 10.

Sequence pool optimization

During mutation optimization, we remove any mutations that do not change the target fold (dual graph) of the sequence. The remaining mutations are deemed as essential or minimal (21).

Execution time

RAG-IF runs were performed on a 13-inch MacBook Pro (Apple, Cupertino, CA) with 3.1 GHz dual-core Intel Core i5 processor with 8 GB random-access memory. Our RAG-IF code was modified so as to perform in parallel all sequence mutation searches and 2D in silico folding for four candidate sequences at a time. A typical RAG-IF run takes ~5 h for this RNA, and around 50% is spent on producing high-fitness sequences,
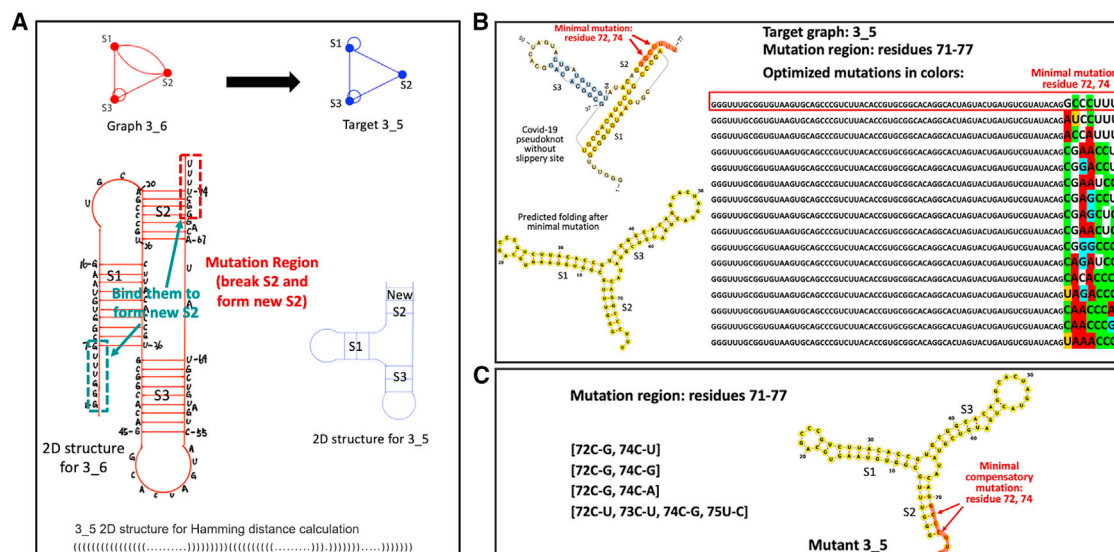

FIGURE 2 Mutation results for the three-way junction target graph 3_5. (A) The selected mutation region is shown. (B) Minimal mutations are shown; mutation candidate region is highlighted in red. Point mutations are colored according to the nucleotides. One top mutation is selected, whose predicted secondary structure is shown. (C) Minimal compensatory mutations are shown. To see this figure in color, go online.
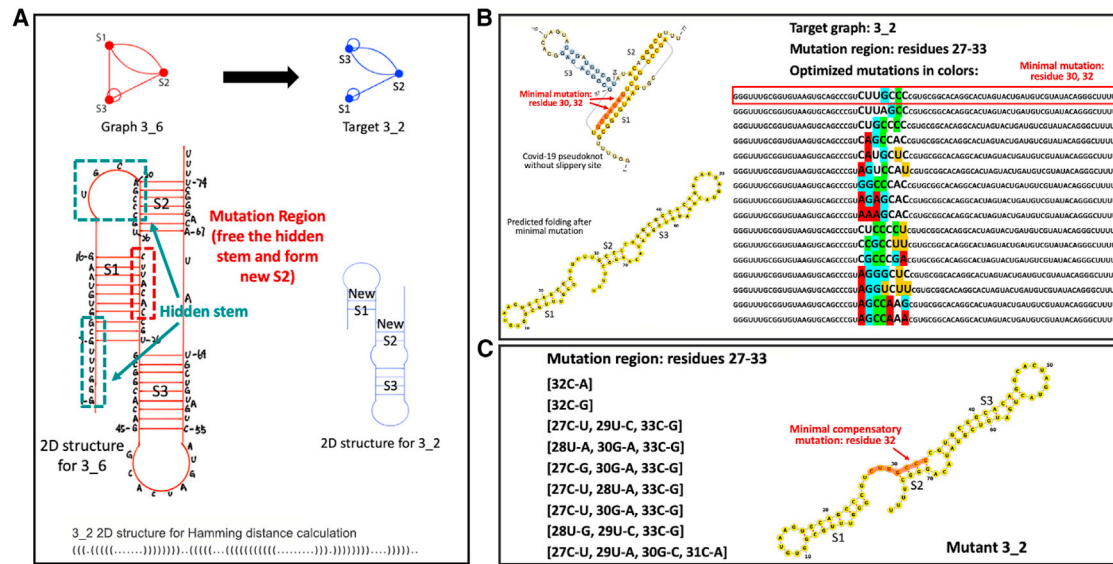
FIGURE 3 Mutation results for the three-stem with internal loop target graph 3_2. (*A*) The selected mutation region is shown. (*B*) Minimal mutations are shown; mutation candidate region is highlighted in red. Point mutations are colored according to the nucleotides. One top mutation is selected, whose predicted secondary structure is shown. (*C*) Minimal compensatory mutations are shown. To see this figure in color, go online.

20% on additional screening by PKNOTS, and 30% on the optimization. Within RAG-IF, >99% of RAG-IF execution time is spent on the 2D structure predictions. We estimate that each NUPACK folding is ~3 s, whereas PKNOTS requires 27 s.

Note that for small mutation regions and small number of minimal mutations, exhaustive enumeration is a feasible alternative approach. However, in general, for a mutation region of n residues and a minimal mutation of k residues, in silico folding of $O(n^k)$ sequences is required. Beyond four mutants, exhaustive enumeration is generally impractical. It is also difficult to estimate the number of minimal mutations beforehand, and hence, our GA approach (followed by mutation optimization) is more generally suitable. An added advantage of our automated optimization approach is that

at the end of a single run, we obtain a large pool of candidate sequences with increasing number of minimal mutations. These data would be useful for further applications.

MD simulation details

The 3D structures of the 77-residue wild-type FSE, along with three mutants—72C-74C, 30G-32C, and 23G-73A—were predicted using four structure prediction webservers: RNAComposer ([37]), Vfold3D ([38]), SimRNA ([39]), and 3dRNA ([40]). We discard any structure that did not correspond to the intended dual-graph motif (3_6 for the wild-type, 3_5
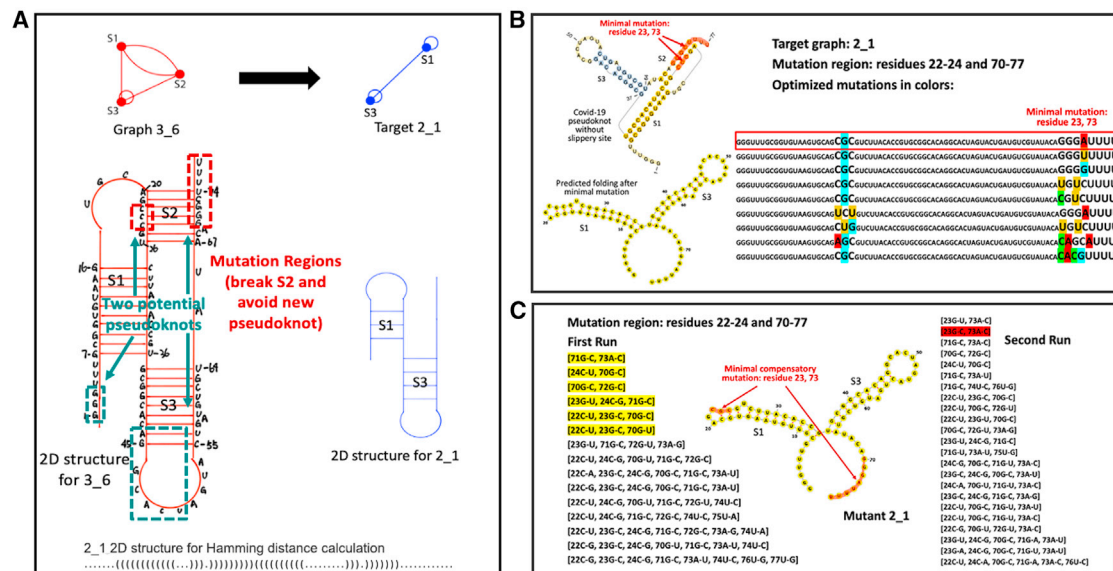

FIGURE 4 Mutation results for the two-stem target graph 2_1. (*A*) The selected mutation region is shown. (*B*) Minimal mutations are shown; mutation candidate region is highlighted in red. Point mutations are colored according to the nucleotides. One top mutation is selected, whose predicted secondary structure is shown. (*C*) Minimal compensatory mutations are shown. To see this figure in color, go online.
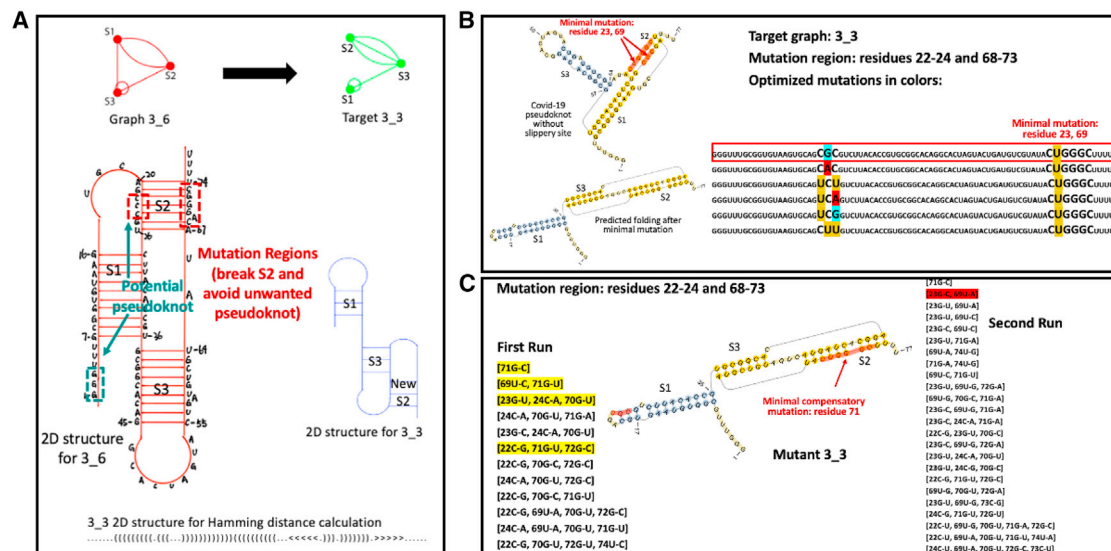
FIGURE 5 Mutation results for the three-stem with different pseudoknot target graph 3_3. (*A*) The selected mutation region is shown. (*B*) Minimal mutations are shown; mutation candidate region is highlighted in red. Point mutations are colored according to the nucleotides. One top mutation is selected, whose predicted secondary structure is shown. (*C*) Minimal compensatory mutations are shown. To see this figure in color, go online.

for 72C-74C mutant, 3_2 for 30G-32C mutant, and 2_1 for 23G-73A mutant) based on the 2D structures annotated by either DSSR (41) or RNA-View (42). Structure validation was performed by MolProbity (43) (using nuclear hydrogen positions), and the structure with least steric clashes and best geometry was selected for the subsequent MD simulations. For the wild-type, as well as the three mutants, the selected structure was predicted by RNAComposer.

MD simulations were performed using Gromacs 2020.3 (44) with the Amber OL3 force field (45). The systems were solvated with TIP3P water molecules in the cubic box whose boundaries extended at least 10 Å from any RNA atom (46). Because of unavailable information on specific locations of Mg^{2+} ions, we placed sodium ions randomly for charge neutralization, with additional Na^+ and Cl^- ions added to achieve a 0.1 M bulk concentration. The total number of atoms for the simulations of the wild-type, 72C-74C mutant, 30G-32C mutant, and 23G-73A mutant are 119,671, 156,619, 162,175, and 151,076, respectively.

The systems were energy minimized via steepest descent and equilibrated while fixing the RNA. Simulations were run with a time step of 2 fs and a SHAKE-like LINCS algorithm (47) with constraints on all bonds. The equilibration was performed for 100 ps in the NVT ensemble (300 K) and then 100 ps in the NPT ensemble (300 K and 1 bar). Production runs were performed for 1 μs under NPT. The RNA and ionic solvent were independently coupled to external heat baths with a relaxation time of 0.1 ps. The particle mesh Ewald method (48) was used to treat long-range electrostatics. The short-range electrostatic cutoff is set at 10 Å, and the van der Waals cutoff is set at 10 Å.

Our MD simulations were performed on dedicated compute nodes provided by the New York University High-Performance Computing facilities. Each compute node is equipped with two Intel Xeon E5-2690v4 2.6 GHz CPUs ("Broadwell," 14 cores/socket, 28 cores/node) and 125 GB memory, EDR InfiniBand with 20:16 oversubscription. Each simulation was performed with seven to eight nodes (i.e., 196–224 cores), so that 1-μs simulations complete in 7–10 days.

The trajectories were analyzed using Gromacs, and structure visualization was prepared by PyMOL (49) and VMD (50). Root mean-square deviations (RMSDs) and root mean-square fluctuations (RMSFs) were calculated with reference to the NPT equilibrated structure using all heavy RNA atoms. Cluster analysis was conducted via Gromos method with the RMSD cutoff of 3.5 Å on RNA non-H backbone atoms. Principal component analysis (PCA) was performed using all heavy atoms in the RNA. Structures used for clustering and PCA analyses were taken every 250 ps from 500 ns to 1 μs.

RESULTS

Secondary structure and dual graph for SARS-CoV-2 FSE

Based on genomic sequencing, the SARS-CoV-2 genome is 89% similar to SARS-CoV and 50% to MERS-CoV. From the study of earlier RNA viruses, the FSE consists of two components: a heptamer of the form XXXYYYZ, known as the slippery site, where the ribosome "slips" to an overlapping reading frame; and a downstream mRNA element, usually a pseudoknot (9,51). Specifically, the SARS-CoV FSE has been identified as the slippery site UUUAAAC and a downstream three-stem pseudoknot-containing structure (12,27). Because of strong sequence conservation, it is presumed that this similar FSE motif would also exist in SARS-CoV-2. We searched for this FSE motif (see Identification of the SARS-CoV-2 FSE) and identified the same 84-nucleotide sequence in the viral genome of SARS-CoV-2 (residues 13,405–13,488, analogous to the SARS-CoV residues 13,392–13,475), with a mere single-nucleotide substitution (A in SARS-CoV-2 instead of C in SARS-CoV in position 72 of this sequence), as shown in Fig. 1 A.

To assess the 2D structure of the SARS-CoV-2 FSE and to select software for our mutational analysis, we applied eight 2D-structure prediction packages that can handle pseudoknots. See Materials and Methods.

Out of these eight packages, three programs predicted a three-stem pseudoknot structure (Fig. S1). Because the remaining programs predicted additional stems or pseudoknots that involved the heptamer slippery sequence, we applied the eight programs again to the 77-residue system without the slippery site. Now, five of the eight programs predict a three-stem pseudoknot structure (Fig. S2).

Our resulting consensus three-stemmed pseudoknot structure in Fig. 1 *A* shows the stem regions that lead to an intertwined structure in three-dimensional space. The corresponding circular plot shows the base pairing. Using our rules to represent RNA 2D structure for RAG dual graphs (RAG Dual Graphs), this FSE structure corresponds to the dual graph 3_6 from our recently enumerated atlas of dual graphs up to nine vertices (26). The three stems correspond to three vertices, and the loops correspond to connecting edges. This 3_6 fold motif also corresponds to multiple riboswitch structures of S-adenosylmethionine (PDB: 3NPN and 3NPQ) and fluoride riboswitches (PDB: 3VRS, 4EN5, 4ENA, 4ENB, and 4ENC) (26).

Although it is not conclusively known what the FSE structure is in SARS-CoV-2 and whether this region's folding may be affected by neighboring sequences and its binding to the nucleocapsid protein or due to the helicase activity of the ribosome (52,53) (see Discussion), other recent works have used various computational tools and experimental input to suggest its structure (14,54).

Because our dual graphs are independent of the number of residues in stems and loops, the graph motif 3_6 corresponds to every three-stem pseudoknot structure predicted by the above programs for both the 84-residue and 77-residue FSE. This insensitivity of our graph-based approach to small differences in base pairing annotations makes its application robust. We use the 77-residue FSE and software packages NUPACK and PKNOTS for our computational mutational analysis.

Identification of minimal mutations to destroy Stem 2 and/or pseudoknot

The SARS-CoV-2 FSE structure and associated transitions are thought to be key for ribosomal frameshifting and the subsequent translation of the ORF1a,b region of the viral genome. Any disruption of key structural features would possibly inhibit or hamper this translation. Of course, mutations that affect tertiary interactions are also known to alter frameshifting efficiency in other viruses (55–58). Hence, we seek crucial FSE nucleotides, especially those that destroy stem 2 (Fig. 1) and the pseudoknot, that can potentially serve as gene therapy or antiviral drug targets.

Identifying these crucial residues is challenging, even in a relatively small sized RNA like the FSE. Fortunately, our RAG representations allow us to simplify this problem. We formulate our goal here as identifying minimal mutations to transform the 3_6 dual-graph motif of the FSE into nine other dual graphs, of equal or smaller size, that correspond to 2D structures of known RNAs (26). These nine targets are shown in Fig. 1 *D*. Among these, six graphs are pseudoknot free (*blue*), and the remaining three contain pseudoknots (*green*).

We use our inverse-folding protocol RAG-IF (21) to mutate the FSE sequence, as described in Materials and Methods. After the minimal mutations are identified, we transform the top mutated sequences back to the original FSE graph 3_6 as a double check on our RAG-IF protocol and to assess the role of selected residues in mutant stability. Recovering the original FSE sequence and determining whether other compensatory mutations can accomplish structural transformations would be an interesting added finding.

Below, we describe minimal mutation and compensatory mutation results for four of the most successful targets, 3_5, 3_2, 2_1, and 3_3, in turn (see Figs. 2, 3, 4, and 5). Summary Fig. 6 *A* shows residues involved.

Mutations for target graph 3_5, the three-way junction

To transform 3_6 of the FSE into target 3_5 (Fig. 2 *B*), we keep stems 1 and 3 intact and select residues 71–77 (highlighted in *red*) in stem 2 for mutation.
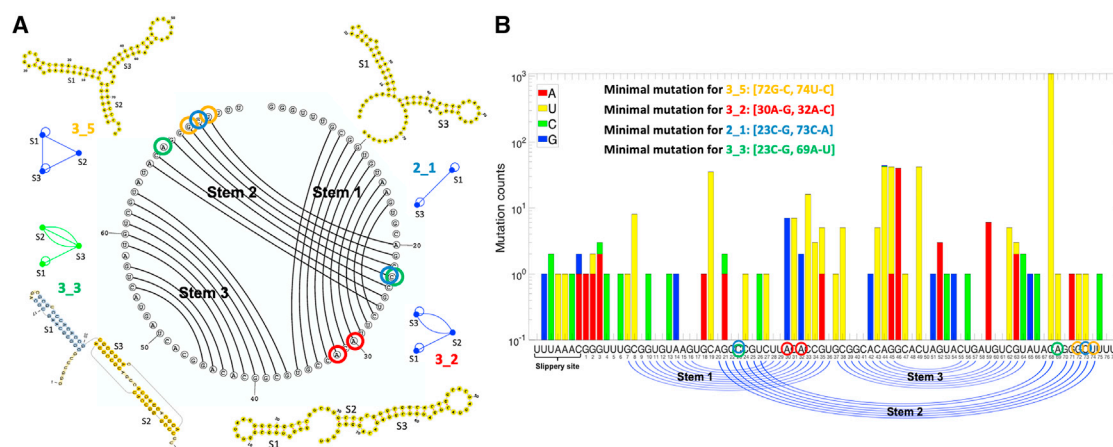

FIGURE 6 FSE residue analyses. (*A*) Minimal mutation residues are circled for the different mutants: yellow for 3_5, red for 3_2, blue for 2_1, and green for 3_3. (*B*) FSE mutation findings for COVID-19 variants from the GISAID database (59) as of August 10, 2020 are shown. To see this figure in color, go online.

RAG-IF GA produces 115 sequences that fold onto the target graph 3_5 by both NUPACK and PKNOTS, with number of mutations ranging from 3 to 7 (Table S1). After mutation optimization to remove nonessential mutations, we obtain 16 unique sequences with two to seven minimal mutations. Significantly, [72G-C, 74U-C] destroys the corresponding basepairs in stem 2 of the FSE and strengthens the new stem of 3_5 with G-C basepairs.

The same outcome for this transformation of 3_6 to 3_5 using [72G-C, 74U-C] comes from two additional prediction programs (see Fig. S4). Thus, we consider this the top mutant choice for target 3_5. Its predicted 2D structure is shown in Fig. 2 B.

Transforming this mutant [72G-C, 74U-C] back to the 3_6 graph can be accomplished by two double mutants besides [72C-G, 74C-U]: [72C-G, 74C-G] and [72C-G, 74C-A] (Fig. 2 C).

Mutations for target graph 3_2, three stems with internal loop

Next, we transform the FSE into target graph 3_2 (Fig. 3 B). To construct 3_2 from 3_6, we keep stem 3 intact and select the mutation region as residues 27–33.

RAG-IF GA generates 64 candidate sequences that fold onto the target 3_2 by both NUPACK and PKNOTS, with mutations ranging from 3 to 7 (see Table S1). After mutation optimization, we obtain 16 unique sequences, with two to six minimal mutations. Two of the 16 optimized sequences require only two mutations: either residues 30 (from A to G) or 31 (from C to G) with residue 32 (from A to C).

Both sequences mutate key residues to destroy stem 1 of the FSE, along with the hairpin loop that forms the pseudoknot (stem 2), and strengthen the newly formed stems with G-C basepairs. The predicted structure for mutation [30A-G, 32A-C] (Fig. 3 B) is confirmed by another prediction program (Fig. S4).

When transforming the mutant [30A-G, 32A-C] back to the 3_6 graph, we obtain 55 candidate sequences before optimization from the GA. Besides recovering [30G-A, 32C-A], we find two other single-point mutations, [32C-A] and [32C-G], that recover the graph 3_6. In fact, 48 of the 55 sequences mutate residue 32. After optimization, we obtain nine unique sequences (Fig. 3 C).

Interestingly, most of the mutations from 44 of the 55 candidates are deemed nonessential, and only [32C-A] or [32C-G] are retained. This includes the reverse mutation [30G-A, 32C-A]. Mutating residue 32 breaks the middle basepair in a relatively small stem in the 3_2 mutant, thus destabilizing the whole structure. If residue 32 is not mutated, at least three minimal mutations are required, rendering residue 32 as one of the key residues for structural transformations.

Mutations for target graph 2_1, two stems

To transform 3_6 of the FSE into 2_1, we aim to destroy stem 2 and select residues 22–24 and 70–77 as mutation candidate regions.

RAG-IF GA produces 221 candidate sequences that fold onto the target graph 2_1 by both NUPACK and PKNOTS; the number of mutations ranges from 6 to 11 (Table S1). After mutation optimization, we retain nine unique sequences (shown in Fig. 4 B) with two to five minimal mutations. The three sequences with only two mutations involve residues 23 (from C to G) and 73 (from C to A, G, or U). Because stem 2 in the FSE is mainly formed by four G-C basepairs, mutating 23 and 73 breaks two of them, which destabilizes the stem.

The mutant RNA [23C-G, 73C-A] sequence also folds onto our target graph 2_1 by two additional prediction programs (Fig. S4). The predicted 2D structure, shown in Fig. 4 B, has stem 1 longer than that of the original FSE.

Our reverse transformation of the mutant [23C-G, 73C-A] to 3_6 generates 23 unique sequences after an expanded pool (see Materials and Methods), recovering [23G-C, 73A-C] and finding many others (Fig. 4 C).

Mutations for target graph 3_3, three stems with different pseudoknot

To transform 3_6 of the FSE into 3_3 (Fig. 5 B), we destroy stem 2 to form the new pseudoknot by mutating residues 22–24 and 68–73, the two strands of stem 2.

RAG-IF GA produces 200 candidate sequences that fold onto the target 3_3 by both NUPACK and PKNOTS, with mutations ranging from 4 to 9. After optimization, six unique sequences emerge with only two to three minimal mutations. Two sequences that require two mutations involve residues 23 (from C to G or A) and 69 (from A to U).

Interestingly, all unique sequences require mutating residue 69 to U, which is a bulge in the FSE stem 2. Although mutating residue 69 to U creates an A-U basepair for the new pseudoknot, residue 69 likely has a more important stabilizing role for the original stem 2. The RNA sequence with mutation [23C-G, 69A-U] in Fig. 5 B is supported by another prediction program (Fig. S4).

Reverse transformation of mutant [23C-G, 69A-U] onto graph 3_6 with a large pool of sequences generates 24 unique sequences, eight of which require only two mutations (Fig. 5 C). Four of these sequences mutate the same residues 23 and 69, including the mutation that recovers the wild-type FSE [23G-C, 69U-A] (highlighted in *red*).

In addition, one single-point mutation [71G-C] is also found. Residue 71 is not one of the original mutation residues, but it is basepaired with residue 23 in the wild-type FSE pseudoknot. Therefore, the compensatory mutation [71G-C] restores this G-C basepair and hence the original pseudoknot.

Summary of minimal mutations

Transformations of the FSE into our four target RNAs utilize unpaired residues to form new stems, often keeping stem 3 intact and focusing on stem 2 (Fig. 6 A). Residue 23 is mutated from C to G in the top minimal mutation

for both 2_1 and 3_3. This mutation breaks the G-C basepair in the original stem 2 and elongates stem 1 by forming a new G-C basepair.

In the reverse transformations back to 3_6, the original mutations are recovered by RAG-IF, and single mutations are also found in the case of 3_2 and 3_3. Because RAG-IF aims to recover the original graph 3_6, not the original FSE sequence, compensatory mutations that require fewer residues than the forward mutations emerge. In 3_5, all compensatory mutations require the original two residues, so we consider this design the most stable one.

For other target graphs shown in Fig. 1 *D*, such as 3_1 and 3_8, at least four residues are required for minimal mutations, and for graph 2_2, 14 residues are required (see Table S2).

MD analysis of wild-type and mutant systems

To assess the stability of the wild-type FSE of SARS-CoV-2 and the top three mutants predicted by mutational analysis (3_5 [72G-C, 74U-C] mutant, 3_2 [30A-G, 32A-C] mutant, and 2_1 [23C-G, 73C-A] mutant), we subject the solvated systems to MD simulations of 1 μs. To develop initial tertiary structures for these systems, we use four 3D structure prediction webservers: RNAComposer (37), Vfold3D (38), SimRNA (39), and 3dRNA (40). We select one structure each for the wild-type and the three mutants for MD simulation (Materials and Methods).

RMSD over the 1-μs simulations is a measure of the conformational stability of the predicted structures (Fig. 7). Because predicted (and not experimentally determined) 3D structures are used for the simulations, high RMSD-values (10–15 Å here) with reference to the starting structures are expected. However, the simulation settles down within the first 250–300 ns (with the 3_2 mutant taking the longest), indicating that the wild-type and the three mutant systems have reached a relatively stable state. The residues that fluctuate most are the unpaired residues in the loop regions, as indicated by the RMSF plots in Fig. 7. The 3_2 mutant is again the exception here, likely because of the time taken to reach relative stability.

To study changes in 2D structures of the wild-type and mutants, we clustered the conformations between 500 ns and 1 μs (see Materials and Methods). The dominant cluster for the wild-type contains roughly 73% of the total conformations, indicating the stability of the system. The values are 50–60% for the 3_5 and 2_1 mutants and around 23% for the 3_2 mutant, suggesting that order of increased flexibility. In addition, the representatives of all top clusters (more than 50 conformations) for the wild-type, the 3_5 and 2_1 mutant, and all but one top cluster for the 3_2 mutant have the intended dual-graph motif. Thus, even though small changes in 2D structure occur over the course of the simulation (see Fig. S5), our dual-graph motifs and predicted mutations have stable structures.

We also dissect the dominant motions of the wild-type and the 3_5 mutant with PCA (see Fig. 8). The wild-type system has global bending of loop 3 as PC1 and stem twisting for PC2 and PC3. The main motions of the 3_5 mutant are bending of terminal loops and twisting of the stems. It is possible that these motions are involved in large-scale structural transitions, but more extensive simulations with possibly enhanced sampling techniques are needed.

Our electrostatic surface (generated using the CHARMM-GUI PBEQ solver at http://www.charmm-gui.org/?doc=input/pbeqsolver) analysis of the wild-type system reveals some positively charged regions along the mostly negative surface, and some grooves (Fig. 9). Some of these positively charged regions and grooves are found near residues 75–77 at the 3′ end of the FSE, residues 24–25 at the juncture of stems 1 and 2, residue 66 at the juncture of stems 2 and 3, basepair 10–33 in stem 1, residues three to five at the 5′ end of the FSE, and residues 47–53 near the hairpin loop of stem 3. Interestingly, some of these regions coincide with the crucial residues identified by our mutational analysis. These regions provide candidate targets for gene editing, as well as in vitro screening of active compounds.

DISCUSSION

We have proposed motif-altering minimal mutations of the SARS-CoV-2 FSE by computational design. In the key mutations highlighted, only two residue changes can accomplish a dramatic change in structure, namely destruction of stem 2 and the pseudoknot. That the graph motif of the wild-type FSE can be recovered by compensatory mutations using our reverse design, in some cases with only one residue change, is also promising. Our MD simulations suggest overall stability of the wild-type and mutant system 2D structures. These tools developed to design in silico RNAs containing pseudoknots are widely applicable to other design problems. The GA in RAF-IF for inverse folding is very successful at generating a large number of sequences quickly and sorting out the sequence pools to obtain minimal mutations.

Besides these computational advances, the results produced here immediately suggest key residues as drug, as well as gene editing, targets for COVID-19 therapy. Already, for the SARS FSE, the small molecule 1,4-diazepane derivative 10 (MTDB) has been shown to inhibit this frameshifting process (10,13). Recent in vitro MTDB binding experiments to the SARS-CoV-2 FSE demonstrated a similar inhibition effect (14).

This lead, together with our identified mutation sites and positively charged regions of the RNA identified by electrostatic surface analysis of the FSE, is shown in Fig. 9. Underway are in silico screening for antiviral compounds targeting the FSE with these residues. Viral-RNA-based
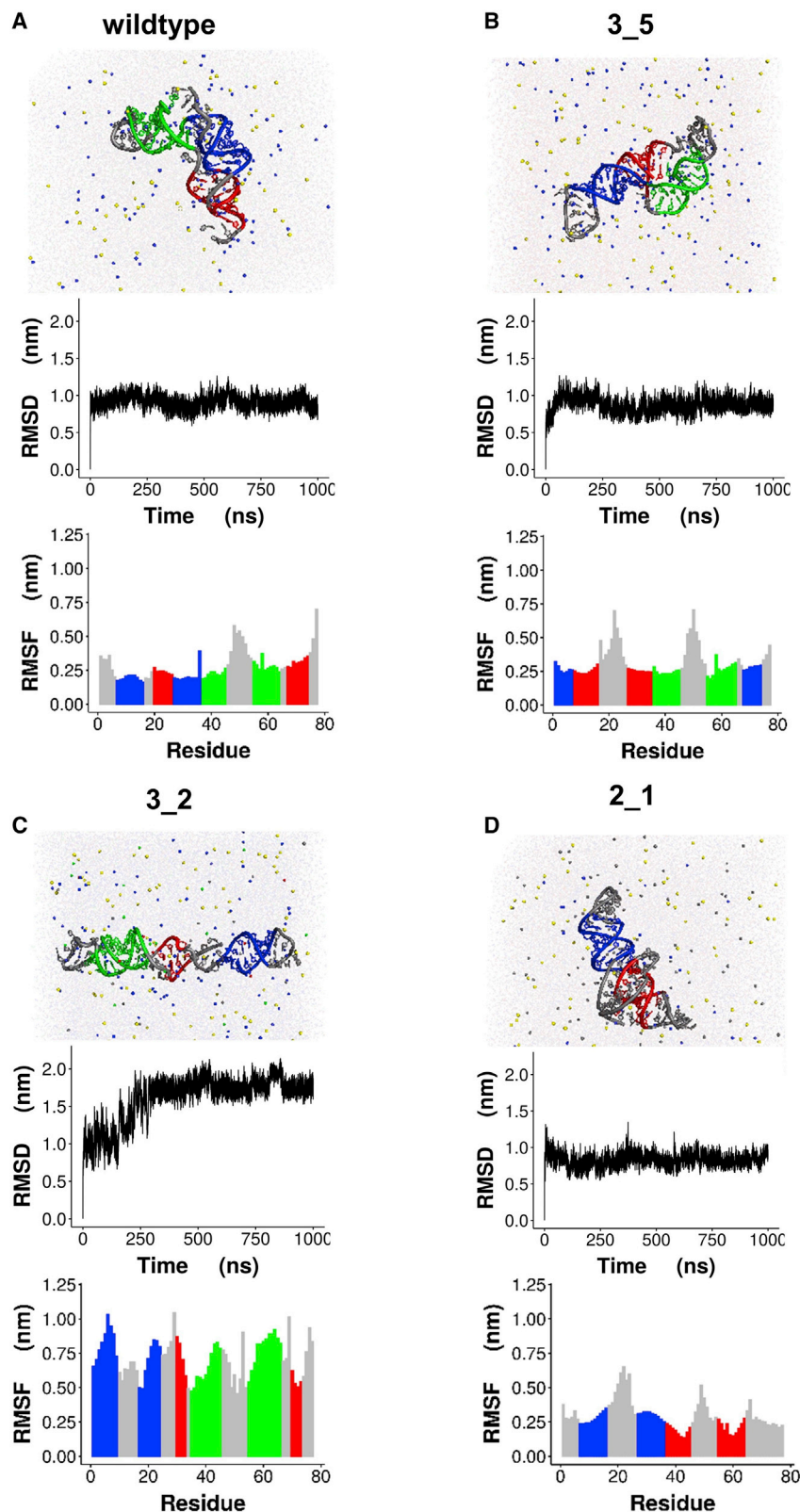
Schlick et al.

FIGURE 7 MD simulations of the wild-type and selected mutant systems using predicted models as initial structures. Snapshots of the wild-type and mutant systems at 1 μs along with RMSD and RMSF are presented for (*A*) wild-type, (*B*) 3_5 mutant, (*C*) 3_2 mutant, and (*D*) 2_1 mutant. The screenshots and RMSF plots are color coded for each RNA: (*A*) residues 7–16 and 27–36 in blue, residues 20–26 and 67–74 in red, and residues 37–45 and 55–64 in green; (*B*) residues 1–7 and 68–74 in blue, residues 27–35 and 8–16 in red, and residues 36–45 and 55–65 in green; (*C*) residues 1–9 and 17–24 in blue, residues 30–33 and 70–73 in red, and residues 35–45 and 55–66 in green; and (*D*) residues 7–16 and 27–36 in blue and residues 37–45 and 55–64 in red. To see this figure in color, go online.

therapies that utilize CRISPR-like systems delivered through small-molecule vectors are also becoming more feasible.

Recently, a complete 2D structure prediction of SARS-CoV-2 genome was reported using in vivo SHAPE-MaP data (54). Two different FSE structures were predicted:
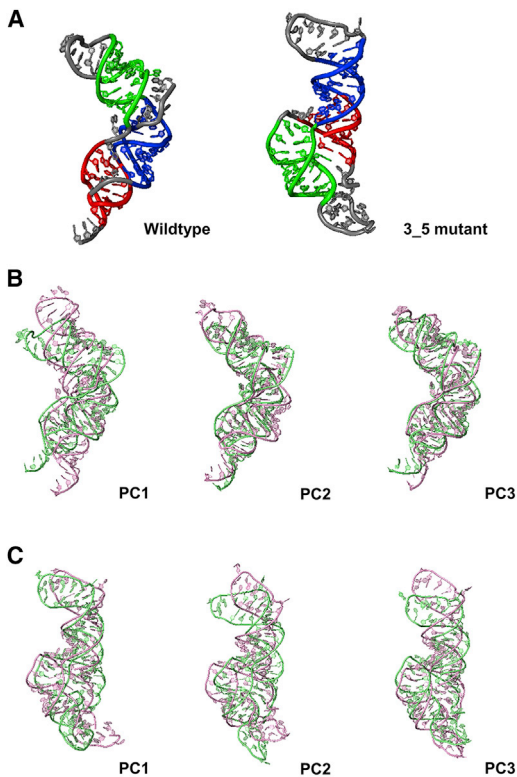
FIGURE 8 Dominant structure and motions of the wild-type and three-way junction (3_5) mutant from MD simulations. (A) The largest cluster center structures from last 500 ns in the 1-μs MD trajectories of the wild-type and the 3_5 mutant are given. (B and C) Motions of (B) the wild-type and (C) the 3_5 mutant by top three principal components (PCs) are shown. Screenshots were extracted as two extreme structures on the selected PC. To see this figure in color, go online.

one consistent with our three-stem structure and one forming a different stem 3 with downstream sequence. The two possibilities were considered commensurate with the conformational flexibility of the pseudoknot. The recent cryo-electron microscopy structure of the SARS-CoV-2 FSE (60) also describes the three-stem pseudoknot consis-

tent with ours, although with some differences in length of stems and loops. More thorough experiments are needed to confirm the FSE structure and examine its conformational flexibility. Note that our graph representations are robust and insensitive to sizes of loops and stems; only the overall topology matters.

As is well appreciated, various secondary structure prediction programs use different algorithms and often produce different predictions. For the 77-residue system, five of the eight prediction programs yield the three-stem pseudoknot structure (Fig. 1). However, when the slippery site is added, only two of the five programs preserve the pseudoknot structure. Larger systems yield more diverse structures. The computational time for the different programs also varies significantly. For example, PKNOTS takes ~30 s for predicting the structure of 77 nucleotides, but with a time complexity $O(n^6)$, where n is the number of residues, large systems are too costly to handle. Programs like IPknot have a more favored size scaling, and replacing PKNOTS with IPknot in our RAG-IF pipeline will allow us to extend our design pipeline to longer RNAs. The use of our graph-based tools is critical here for mutational transformations because it provides us with a robust coarse-grained approach insensitive to small differences in 2D structures.

Clearly, the overall fold of the FSE may depend on neighboring residues and on complexes with the nucleocapsid protein and the ribosome in the cellular context. In Fig. 10, we show predicted structures for the wild-type and three-way junction mutant 3_5 in the context of different flanking sequences. Differences result as a function of sequence length and from application of different programs. Interestingly, in some cases, a large sequence can recover our 77-nt consensus three-stem pseudoknot structure better than a shorter construct. Thus, in vivo imaging and other experiments like SHAPE reactivity probing (61) would be important for understanding
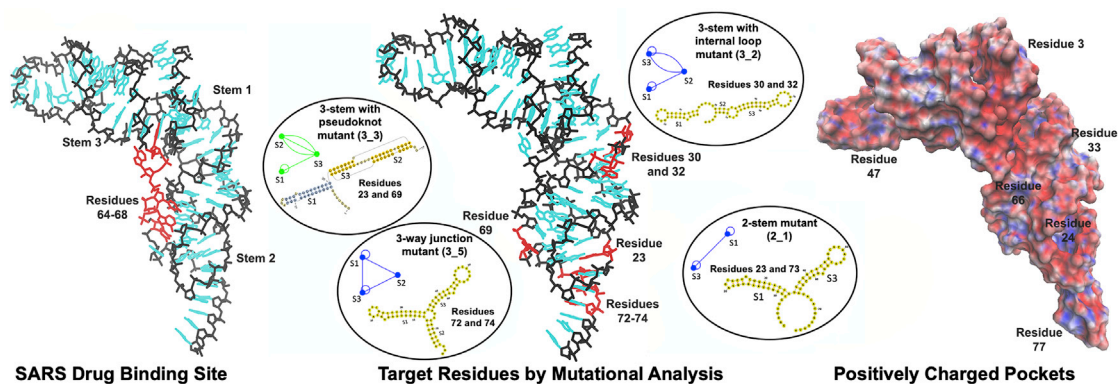

FIGURE 9 Target residues for drugs or gene editing emerging from this work. Residues corresponding to previously identified drug-binding residues (10,13), minimal mutations, and positively charged pockets that arose out of our electrostatic analysis are shown for the FSE (structure corresponding to the most populated cluster representative from MD) of the SARS-CoV-2 viral genome. To see this figure in color, go online.
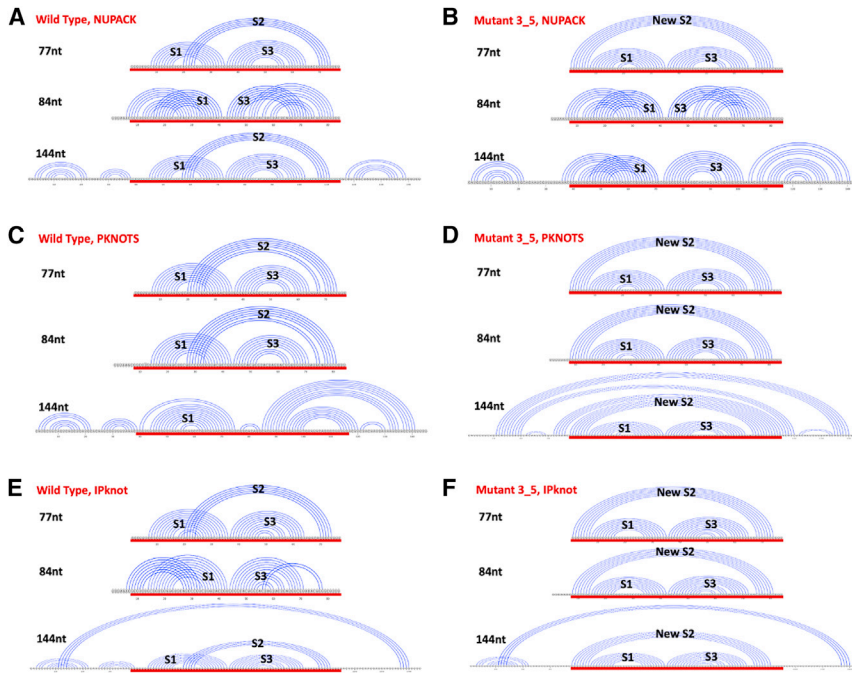
FIGURE 10 Predicted structures for the wild-type and mutant 3_5 using different sequence lengths. The 84-nt sequence contains the slippery site, and the 144-nt sequence adds 30 nt to both ends of the 84-nt sequence. The sequences (aligned with common 77-nt subsequence, *red*) are shown for (*A*) wild-type FSE/NUPACK (30), (*B*) three-way junction mutant 3_5/NUPACK, (*C*) wild-type/PKNOTS (29), (*D*) 3_5/PKNOTS, (*E*) wild-type/IPknot (31), and (*F*) 3_5/IPknot. To see this figure in color, go online.

the contextual folding of this important region of the virus explored in this work in isolation.

Because accumulating data show that SARS-CoV-2 is actively acquiring new mutations that allow it to escape known protein-targeting drugs, alternative therapeutic strategies that target the RNA viral genome, rather than viral shell or auxiliary proteins, present a flexible and efficient approach to counter new infections. By examining the 73,739 aligned viral sequences available in the Global Initiative on Sharing All Influenza Data (GISAID) database (59) as of August 10, 2020, in Fig. 6 *B*, we found that only 1316 sequences (1.78%) have mutations in the 84-nt FSE region. Among these, 1268 contain a single-point mutation, underscoring the high conservation of the FSE region. These mutations are shown along with the residues we identified for top minimal mutations in Fig. 6. Our mutation residues 23 and 74 are not among current variants, but residues 69, 72, and 73 occur with one mutation to U each; residue 32 is mutated twice to G, and residue 30 is mutated seven times to G. This generally high sequence conservation of the FSE and stability to mutations suggests it is a good drug target.

Indeed, the small-molecule MTDB binding to six FSE natural point mutations suggests insensitivity to natural mutations (62). The mutational transformations we reported here provide insights for further in silico and ultimately in vitro drug screening to find other therapeutically effective drugs. In general, understanding the structure of the FSE and other gene regions of coronaviruses is important for helping address future viral threats that inevitably may arise. Such knowledge and methodologies are also transferable to other coronaviruses or other viruses that could pose threats to human health in the future.

Further SARS-CoV-2 viral genome variant analyses are also important for drug discovery and for identifying the spread of COVID-19 around the globe. Initially, it was believed that different strains are not related to virulence or infectivity, but theories are evolving as associations of different genomic mutations with different levels of infectivity are being debated (e.g., (63)). As more data on associations of viral mutants and disease outcome accumulate, further analysis on the viral genome will be invaluable.

In closing, molecular modeling and computational experiments such as described here have an important role to play in this fight against the small RNA enemy that has upended our modern world as never before since World War II. The continued interaction and participation of the biophysics, high-performance computing, and bioinformatics and computational biology communities will be invaluable in this quest. An improved understanding of the complex structural biophysics of the RNA viral genome will undoubtedly be useful and relevant to other coronaviruses, whose potential to emerge in the future cannot be understated.

SUPPORTING MATERIAL

AUTHOR CONTRIBUTIONS

T.S., Q.Z., S.J., and S.Y. designed the research. T.S., Q.Z., S.J., and S.Y. performed the computational study and analyzed the results. T.S., Q.Z., S.J., and S.Y. wrote the article.

ACKNOWLEDGMENTS

REFERENCES

1. Xie, M., and Q. Chen. 2020. Insight into 2019 novel coronavirus - an updated interim review and lessons from SARS-CoV and MERS-CoV. *Int. J. Infect. Dis.* 94:119–124.

2. Xu, Z., L. Shi, …, F. S. Wang. 2020. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir. Med.* 8:420–422.

3. Manickum, T. 2020. A 3-month preliminary epidemiological assessment of some factors affecting the human-to-human transmission of causal agent SARS-CoV-2 virus for infection rates of COVID-19. *Int. J. Pub. Heal. Saf.* 5:186.

4. Kebede, T., D. Kumar, and P. Sharma. 2020. Potential drug options for treatment process of COVID-19: a review. *Coronaviruses.* 1:1–7.

5. Cascella, M., M. Rajnik, …, R. Di Napoli. 2020. Features, Evaluation and Treatment Coronavirus (COVID-19). StatPearls. StatPearls Publishing, Treasure Island, FL.

6. Hermann, T. 2016. Small molecules targeting viral RNA. *Wiley Interdiscip. Rev. RNA.* 7:726–743.

7. Nguyen, T. M., Y. Zhang, and P. P. Pandolfi. 2020. Virus against virus: a potential treatment for 2019-nCov (SARS-CoV-2) and other RNA viruses. *Cell Res.* 30:189–190.

8. Jackson, L. A., E. J. Anderson, …, J. H. Beigel; mRNA-1273 Study Group. 2020. An mRNA vaccine against SARS-CoV-2 - preliminary report. *New Eng. J. Med*, NEJMoa2022483. Published online July 14, 2020.

9. Brierley, I. 1995. Ribosomal frameshifting viral RNAs. *J. Gen. Virol.* 76:1885–1892.

10. Ritchie, D. B., J. Soong, …, M. T. Woodside. 2014. Anti-frameshifting ligand reduces the conformational plasticity of the SARS virus pseudoknot. *J. Am. Chem. Soc.* 136:2196–2199.

11. Dinman, J. D., M. J. Ruiz-Echevarria, and S. W. Peltz. 1998. Translating old drugs into new treatments: ribosomal frameshifting as a target for antiviral agents. *Trends Biotechnol.* 16:190–196.

12. Plant, E. P., G. C. Pérez-Alvarado, …, J. D. Dinman. 2005. A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.* 3:e172.

13. Park, S.-J., Y.-G. Kim, and H.-J. Park. 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.* 133:10094–10100.

14. Kelly, J. A., A. N. Olson, …, J. D. Dinman. 2020. Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *J. Biol. Chem.* 295:10741–10748.

15. Gan, H. H., S. Pasquali, and T. Schlick. 2003. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res.* 31:2926–2943.

16. Pasquali, S., H. H. Gan, and T. Schlick. 2005. Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. *Nucleic Acids Res.* 33:1384–1398.

17. Gevertz, J., H. H. Gan, and T. Schlick. 2005. In vitro RNA random pools are not structurally diverse: a computational analysis. *RNA.* 11:853–863.

18. Schlick, T. 2018. Adventures with RNA graphs. *Methods.* 143:16–33.

19. Bayrak, C. S., N. Kim, and T. Schlick. 2017. Using sequence signatures and kink-turn motifs in knowledge-based statistical potentials for RNA structure prediction. *Nucleic Acids Res.* 45:5414–5422.

20. Meng, G., M. Tariq, …, T. Schlick. 2020. RAG-Web: RNA structure prediction/design using RNA-As-Graphs. *Bioinformatics.* 36:647–648.

21. Jain, S., Y. Tao, and T. Schlick. 2020. Inverse folding with RNA-As-Graphs produces a large pool of candidate sequences with target topologies. *J. Struct. Biol.* 209:107438.

22. Kim, N., C. Laing, …, T. Schlick. 2014. Graph-based sampling for approximating global helical topologies of RNA. *Proc. Natl. Acad. Sci. USA.* 111:4079–4084.

23. Jain, S., and T. Schlick. 2017. F-RAG: generating atomic coordinates from RNA graphs by fragment assembly. *J. Mol. Biol.* 429:3587–3605.

24. Jain, S., A. Laederach, …, T. Schlick. 2018. A pipeline for computational design of novel RNA-like topologies. *Nucleic Acids Res.* 46:7040–7051.

25. Zahran, M., C. Sevim Bayrak, …, T. Schlick. 2015. RAG-3D: a search tool for RNA 3D substructures. *Nucleic Acids Res.* 43:9474–9488.

26. Jain, S., S. Saju, …, T. Schlick. 2019. An extended dual graph library and partitioning algorithm applicable to pseudoknotted RNA structures. *Methods.* 162–163:74–84.

27. Thiel, V., K. A. Ivanov, …, J. Ziebuhr. 2003. Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* 84:2305–2315.

28. Macke, T. J., D. J. Ecker, …, R. Sampath. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* 29:4724–4735.

29. Rivas, E., and S. R. Eddy. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285:2053–2068.

30. Dirks, R. M., and N. A. Pierce. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* 24:1664–1677.

31. Sato, K., Y. Kato, …, K. Asai. 2011. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics.* 27:i85–i93.

32. Andronescu, M. S., C. Pop, and A. E. Condon. 2010. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA.* 16:26–42.

33. Xu, X., P. Zhao, and S.-J. Chen. 2014. Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLoS One.* 9:e107504.

34. Singh, J., J. Hanson, …, Y. Zhou. 2019. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* 10:5407.

35. Bellaousov, S., and D. H. Mathews. 2010. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA.* 16:1870–1880.

36. Dawson, W. K., K. Fujiwara, and G. Kawai. 2007. Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. *PLoS One.* 2:e905.

37. Popenda, M., M. Szachniuk, …, R. W. Adamiak. 2012. Automated 3D structure composition for large RNAs. *Nucleic Acids Res.* 40:e112.

38. Xu, X., and S.-J. Chen. 2018. Hierarchical assembly of RNA three-dimensional structures based on loop templates. *J. Phys. Chem. B.* 122:5327–5335.

39. Magnus, M., M. J. Boniecki, …, J. M. Bujnicki. 2016. SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Res.* 44:W315–W319.

40. Wang, J., J. Wang, …, Y. Xiao. 2019. 3dRNA v2.0: an updated web server for RNA 3D structure prediction. *Int. J. Mol. Sci.* 20:4116.

41. Lu, X. J., and W. K. Olson. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* 31:5108–5121.

42. Yang, H., F. Jossinet, …, E. Westhof. 2003. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.* 31:3450–3460.

43. Williams, C. J., J. J. Headd, …, D. C. Richardson. 2018. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* 27:293–315.

44. Abraham, M. J., T. Murtola, …, E. Lindahl. 2015. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX.* 1–2:19–25.

45. Zgarbová, M., M. Otyepka, …, P. Jurečka. 2011. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem.* 7:2886–2902.

46. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

47. Hess, B., H. Bekker, …, J. G. E. M. Fraaije. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.

48. Essmann, U., L. Perera, …, L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577–8593.

49. Schrödinger, LLC. 2015. The PyMOL molecular graphics system, version 1.8.

50. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–28..

51. Brierley, I., P. Digard, and S. C. Inglis. 1989. Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. *Cell.* 57:537–547.

52. Zhang, Y., S. Hong, …, C. M. Dunham. 2018. Alternative mode of E-site tRNA binding in the presence of a downstream mRNA stem loop at the entrance channel. *Structure.* 26:437–445.e3.

53. Amiri, H., and H. F. Noller. 2019. Structural evidence for product stabilization by the ribosomal mRNA helicase. *RNA.* 25:364–375.

54. Huston, N. C., H. Wan, …, A. M. Pyle. 2020. Comprehensive in-vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *bioRxiv* https://doi.org/10.1101/2020.07.10.197079.

55. Liphardt, J., S. Napthine, …, I. Brierley. 1999. Evidence for an RNA pseudoknot loop-helix interaction essential for efficient -1 ribosomal frameshifting. *J. Mol. Biol.* 288:321–335.

56. Kim, Y. G., L. Su, …, A. Rich. 1999. Specific mutations in a viral RNA pseudoknot drastically change ribosomal frameshifting efficiency. *Proc. Natl. Acad. Sci. USA.* 96:14234–14239.

57. Chen, G., K.-Y. Chang, …, I. Tinoco, Jr. 2009. Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of -1 ribosomal frameshifting. *Proc. Natl. Acad. Sci. USA.* 106:12706–12711.

58. Puah, R. Y., H. Jia, …, G. Chen. 2018. Selective binding to mRNA duplex regions by chemically modified peptide nucleic acids stimulates ribosomal frameshifting. *Biochemistry.* 57:149–159.

59. Elbe, S., and G. Buckland-Merrett. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* 1:33–46.

60. Zhang, K., I. N. Zheludev, …, R. Das. 2020. Cryo-electron microscopy and exploratory antisense targeting of the 28-kDa Frameshift stimulation element from the SARS-CoV-2 RNA genome. *bioRxiv* https://doi.org/10.1101/2020.07.18.209270.

61. Siegfried, N. A., S. Busan, …, K. M. Weeks. 2014. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods.* 11:959–965.

62. Neupane, K., S. Munshi, …, M. T. Woodside. 2020. Anti-frameshifting ligand active against SARS coronavirus-2 is resistant to natural mutations of the frameshift-stimulatory pseudoknot. *bioRxiv* https://doi.org/10.1101/2020.06.29.178707.

63. Korber, B., W. M. Fischer, …, D. C. Montefiori; Sheffield COVID-19 Genomics Group. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell.* 182:812–827.e19.

Supplemental Information

Structure-Altering Mutations of the SARS-CoV-2 Frameshifting RNA Element

Tamar Schlick, Qiyao Zhu, Swati Jain, and Shuting Yan

Graph	Regions	Candidates	Mutation results		
			Number of mutations	Before optim.	After optim.
3_5	71-77	115	2	0	56
			3	3	18
			4	13	0
			5	34	32
			6	45	5
			7	20	4
3_2	27-33	64	2	0	16
			3	2	28
			4	13	13
			5	16	4
			6	27	3
			7	6	0
2_1	22-24 and 70-77	221	2	0	212
			3	0	4
			4	0	2
			5	0	3
			6	13	0
			7	25	0
			8	69	0
			9	76	0
			10	37	0
			11	1	0
3_3	22-24 and 68-73	200	2	0	84
			3	0	116
			4	5	0
			5	21	0
			6	54	0
			7	73	0
			8	44	0
			9	3	0

Table S1: Mutation results before and after optimization for target graphs 3_5, 3_2, 2_1 and 3_3. We show the number of candidate sequences that fold onto the target graph by both prediction programs, and for these candidates, we count the number of point mutations involved before and after optimization.

Graph	Regions	Number of mut.	Minimal mutations
2_2	1-21	14	[1G-A, 2G-A, 3G-A, 7G-A, 10G-C, 11U-C, 12G-C, 13U-C, 16G-A, 18G-C, 19C-U, 20A-U, 21G-C, 15A-U]
3_1	22-24 and 66-77	4-12	[22C-U, 24C-G, 73C-A, 75U-G]
3_8	11-15 and 32-33	4-5	[12G-U, 13U-C, 14A-U, 33C-U] [11U-A, 13U-C, 14A-U, 32A-U]

Table S2: Optimized mutation results for target graphs 2_2, 3_1 and 3_8. The number of mutations required after optimization is given, and the mutations with fewest mutated residues are given.
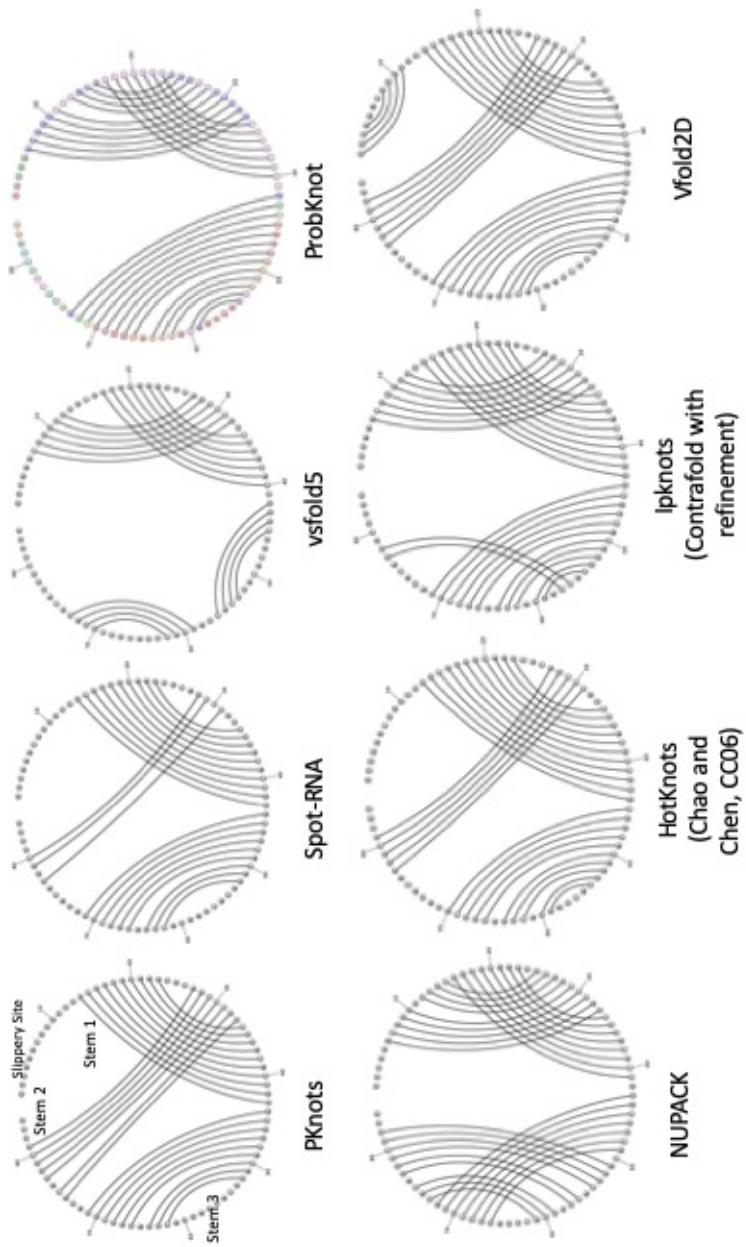
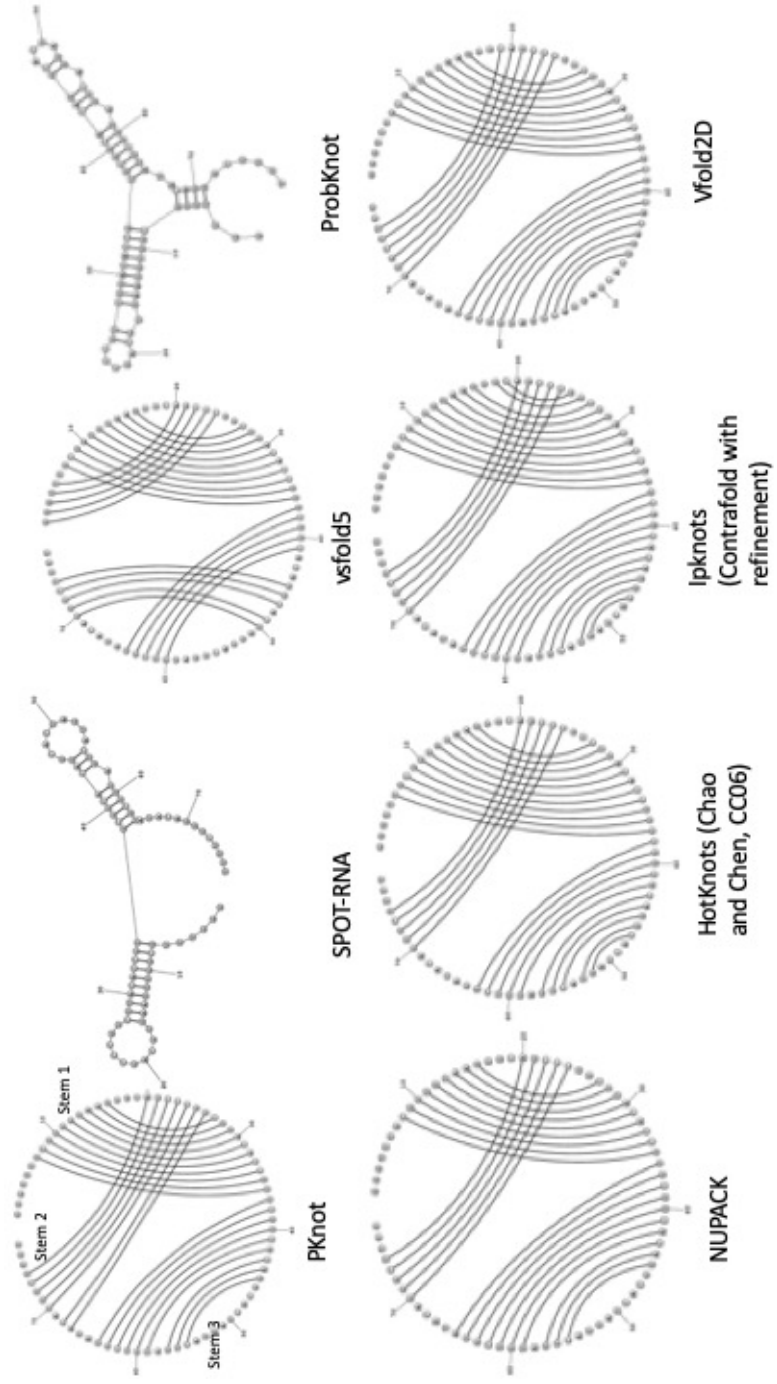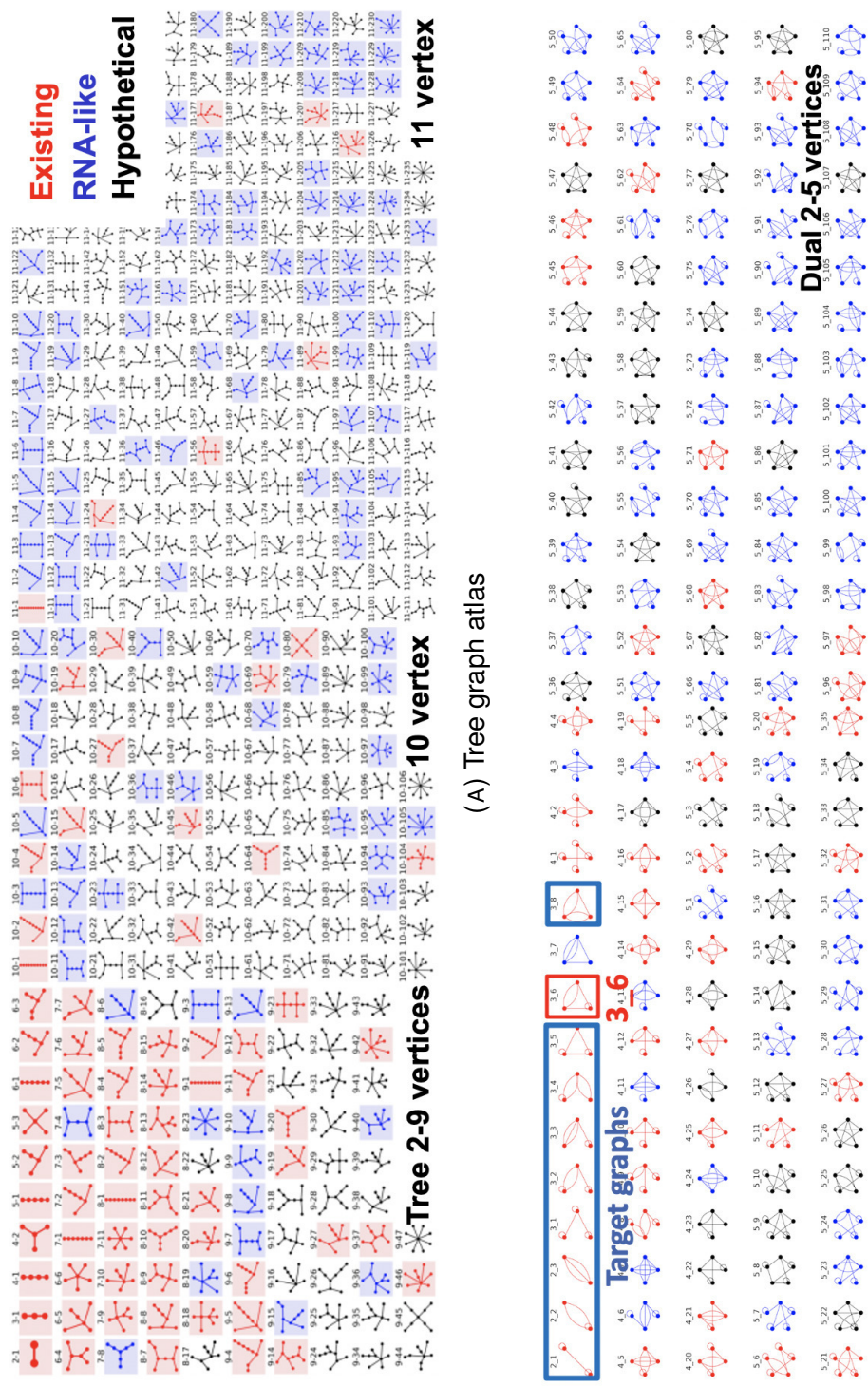Figure S1: Secondary structures predicted by eight software packages for SARS-CoV-2 84-residue FSE.

Figure S2: Secondary structures predicted by eight software packages for SARS-CoV-2 77-residue FSE.

(A) Tree graph atlas

(B) Dual graph atlas

Figure S3: RAG atlas segments for tree and dual graphs with frame-shifting element and other target mutations identified.
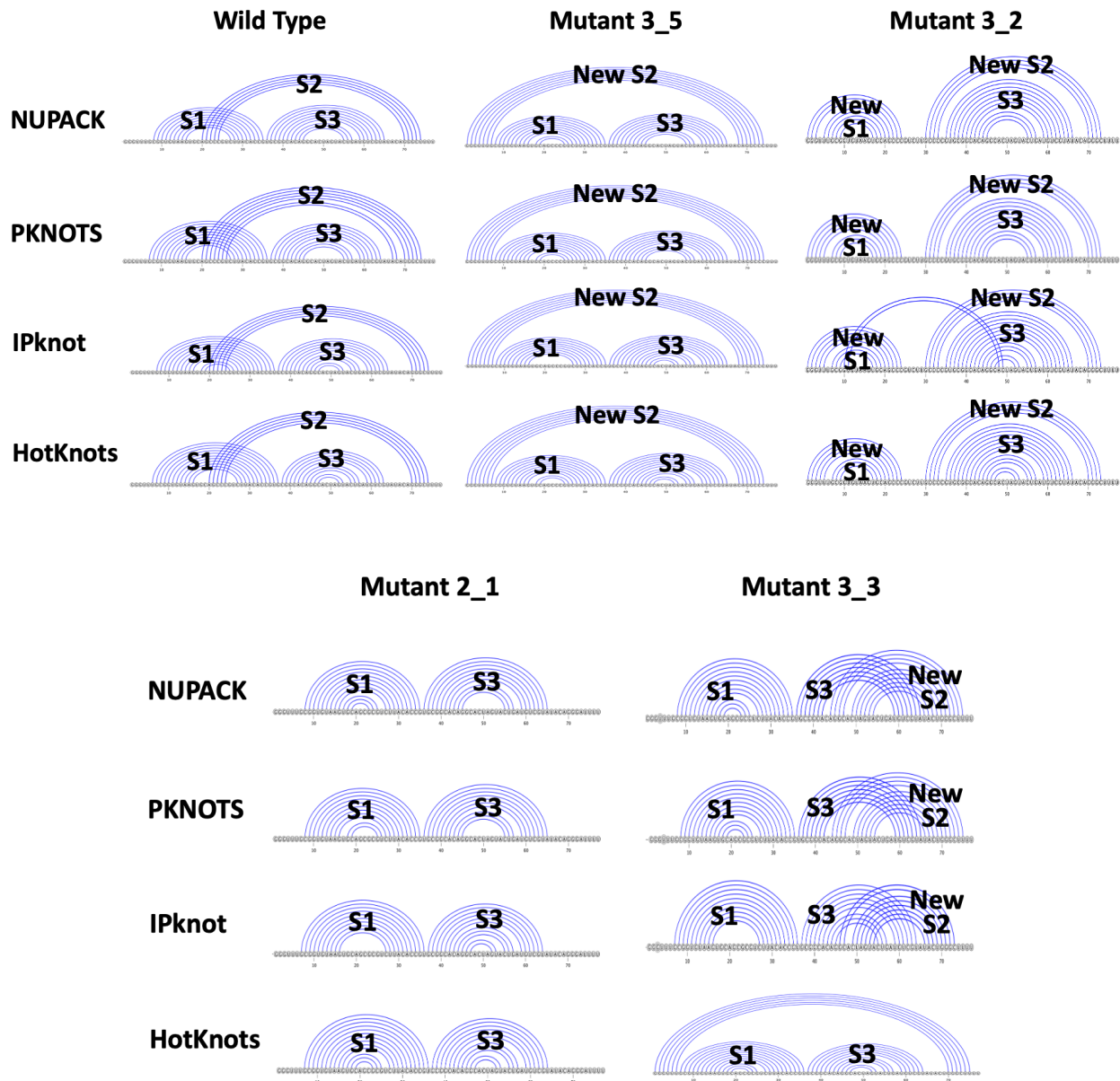
Figure S4: Predicted structures for the wild type and 4 mutant 3_5, 3_2, 2_1 and 3_3 using different prediction programs. Prediction programs are NUPACK, PKNOTS, IPknot (CON-TRAfold model with refinement) and HotKnots (Chao & Chen model, CC06 parameters).
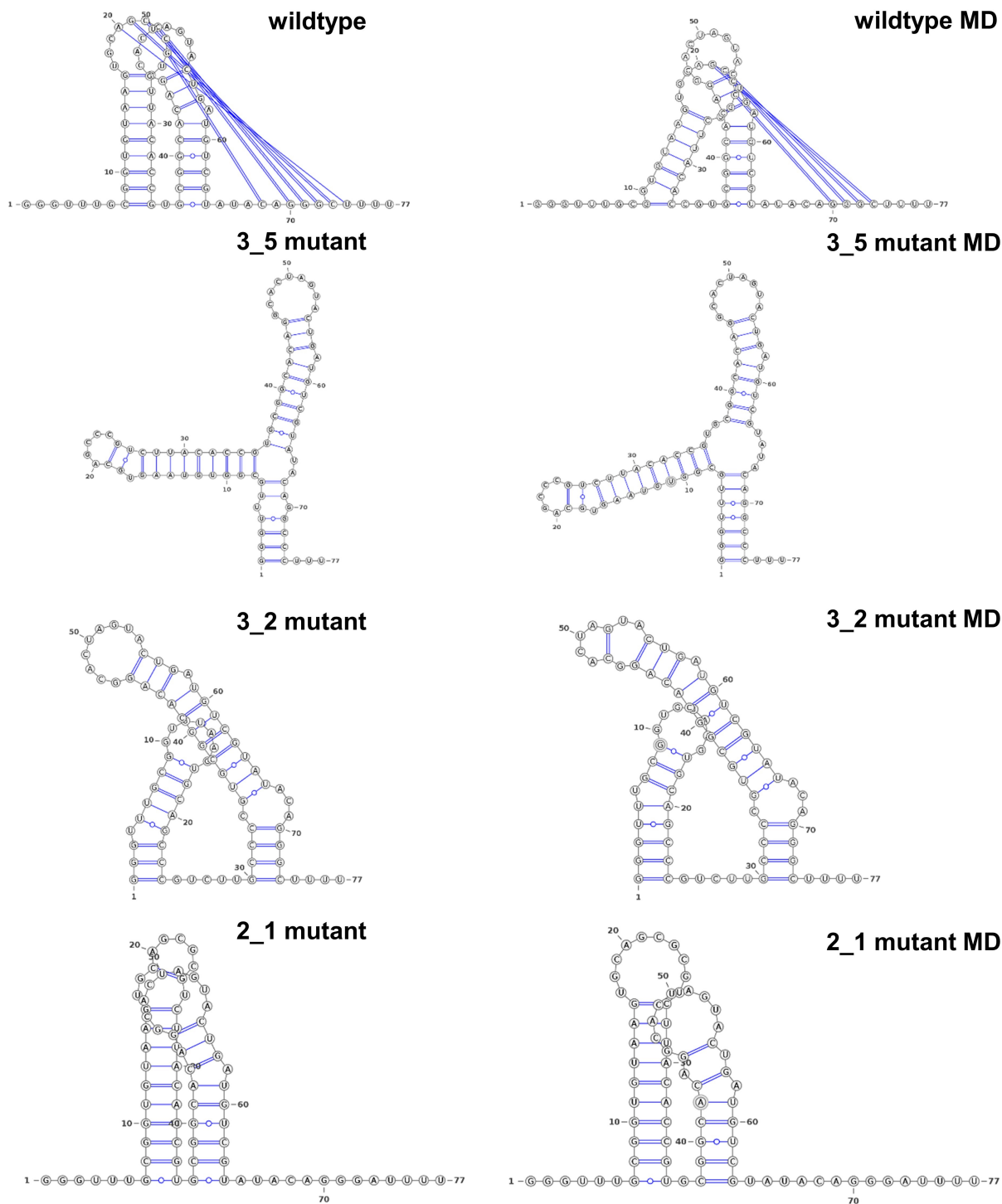
Figure S5: Secondary structures of the starting structure and the representative of the dominant cluster in the 1 microsecond MD simulations of the wildtype FSE and the 3_5, 3_2, and 2_1 mutants.