De novo 3D models of SARS-CoV-2 RNA elements and small-molecule-binding RNAs to guide drug discovery

Ramya Rangan[1*], Andrew M. Watkins[2*], Wipapat Kladwang[2], Rhiju Das[1,2,3†]

[1]Biophysics Program, Stanford University, Stanford CA 94305
[2]Department of Biochemistry, Stanford University School of Medicine, Stanford CA 94305
[3]Department of Physics, Stanford University, Stanford CA 94305

* Equally contributing lead authors.
† Corresponding author: rhiju@stanford.edu

Abstract

The rapid spread of COVID-19 motivates development of antivirals targeting conserved molecular machinery of the SARS-CoV-2 virus. The SARS-CoV-2 genome includes conserved RNA elements that offer potential targets for RNA-targeting small-molecule drugs, but 3D structures of most of these elements have not been experimentally characterized. Here, we provide a dataset called 'FARFAR2-SARS-CoV-2', a collection of 3D coordinates modeled using Rosetta's FARFAR2 algorithm, including de novo models for thirteen RNA elements in SARS-CoV-2 and homology models for a fourteenth. These elements comprise SL1, SL2, SL3, SL4, SL5, putative SL6 and SL7 in the extended 5′ UTR, as well as the entire extended 5′ UTR; the frameshifting element (FSE) from the SARS-CoV-2 ORF1a/b gene and a putative dimer of FSE; and the extended pseudoknot, hypervariable region, and the s2m of the 3′ UTR, as well as the entire 3′ UTR. For five of these elements (SL1, SL2, SL3, FSE, s2m), convergence of lowest predicted energy structures supports their accuracy in capturing low energy states that might be targeted for small molecule binding. To aid efforts to discover small molecule RNA binders guided by computational models, we provide a second benchmarking dataset called 'FARFAR2-Apo-Riboswitch', which consists of similarly prepared Rosetta-FARFAR2 models for RNA riboswitch aptamer regions that bind small molecules. Both datasets include up to 400 3D models for each RNA element, which may facilitate drug discovery approaches targeting dynamic ensembles of low-energy excited states of RNA molecules.

Introduction

The COVID19 outbreak has rapidly spread through the world, presenting an urgent need for therapeutics targeting the betacoronavirus SARS-CoV-2. RNA-targeting antivirals have potential to be effective against SARS-CoV-2, as the virus's RNA genome harbors conserved regions predicted to have stable secondary structures[1] that have been shown to be essential for the life cycle of related betacoronaviruses.[2] Efforts to identify small molecules that target stereotyped 3D RNA folds have advanced over recent years,[3] making RNA structures like those in SARS-CoV-2 potentially attractive targets for small molecule drugs.

Several RNA regions in betacoronavirus genomes, including the 5′ UTR, the frameshifting element (FSE), and 3′ UTR, feature RNA structures with likely functional importance. These regions include a series of five conserved stem loops in the 5′ UTR, the frameshifting element along with a proposed dimerized state, a pseudoknot in the 3′ UTR proposed to form two structures, and the hypervariable region in the 3′ UTR, which includes an absolutely conserved octanucleotide and the stem-loop II-like motif ('s2m'). An NMR structure of stem-loop 2 in the 5′ UTR has been solved, adopting a canonical CUYG tetraloop fold.[4] A crystal structure for s2m in the 3′ UTR has been solved for the original SARS virus, SARS-CoV-1.[5] With nearly complete sequence conservation compared to the SARS-CoV-2 reference genome, this structure is expected to be maintained in SARS-CoV-2. Beyond these regions, however, 3D structures for structural domains in the 5′ UTR, FSE, and 3′ UTR have not been solved.

In advance of detailed experimental structural characterization, computational predictions for the 3D structural conformations adopted by conserved RNA elements may aid the search for RNA-targeting antivirals. Representative conformations from these RNA molecules' structural ensembles can serve

as starting points for virtual screening of small-molecules drug candidates. For example, in prior work by Stelzer, et al.,[6] virtual screening of a library of compounds against an ensemble of modeled RNA structures led to the de novo discovery of a set of small molecules that bound a structured element in HIV-1 (the transactivation response element, TAR) with high affinity. Such work motivates our modeling of not just a single 'native' structure but an ensemble of states for SARS-CoV-2 RNA regions. As with HIV-1 TAR, many of the SARS-CoV-2 elements are unlikely to adopt a single conformation but instead may sample conformations from a heterogeneous ensemble. Furthermore, transitions among these conformations may be implicated in the viral life cycle, as they change long-range contacts with other RNA elements or form interactions with viral and host proteins at different steps of replication, translation, and packaging. A possible therapeutic strategy is therefore to find drugs that stabilize an RNA element in a particular conformation incompatible with conformational changes and/or changing interactions with biological partners at different stages of the complete viral replication cycle. Consistent with this hypothesis, prior genetic selection and mutagenesis experiments stabilizing single folds for stem loops in the 5′ UTR and the pseudoknot in the 3′ UTR demonstrate that changes to these RNA elements' structural ensembles can prove lethal for viral replication.[7-9]

Here, we provide de novo modeled structure ensembles for key RNA elements in the SARS-CoV-2 genome obtained from Rosetta's protocol for Fragment Assembly of RNA with Full-Atom Refinement, version 2 (FARFAR2).[10] These structures include de novo models for stem loops 1 to 7 (SL1-7) in the extended 5′ UTR, the frameshifting element and its dimerized form, the 3′ UTR pseudoknot, and the 3′ UTR hypervariable region, along with homology models of SL2 and s2m. The use of Rosetta FARFAR2 is motivated by extensive testing: FARFAR2 has been benchmarked on all community-wide RNA-Puzzle modeling challenges to date,[11-13] achieving accurate prediction of complex 3D RNA folds for ligand-binding riboswitches and aptamers, and producing models with 3–14 Å RMSD across six additional recent blind modeling challenges.[10] In addition to providing structural ensembles for SARS-CoV-2 RNA elements, we provide analogous FARFAR2 de novo and homology models for 10 riboswitch aptamers, in the hopes of providing a benchmark dataset for virtual screening approaches that make use of computational RNA models.

Results

Using Rosetta-FARFAR2 to model ensembles of SARS-CoV-2 structures: updates and caveats

We generated ensembles for SARS-CoV-2 RNA elements using Rosetta's FARFAR2 protocol, a collection of models we term the **FARFAR2-SARS-CoV-2** dataset. Beginning with a sequence and secondary structure, FARFAR2 generates models through Monte Carlo substitutions of 3-residue fragments sampled from previously solved RNA structures, followed by refinement in a high-resolution physics-based free energy function, which models hydrogen bonding, solvation effects, nucleobase stacking, torsional preferences and other physical forces known to impact macromolecules structure.[10]

FARFAR2 3D modeling requires an input secondary structure. Here, most secondary structures for SARS-CoV-2 RNA elements were obtained from homologies to secondary structures previously probed in related betacoronaviruses, with the exception of stem loops 6-7 in the extended 5′ UTR,

which were generated from computational secondary structure modeling. We emphasize that these 3D model collections reflect a single secondary structure for each RNA element (or two secondary structures for FSE; see below). Future experimental work that elucidates additional secondary structures or tertiary contacts may prompt further 3D modeling of alternative states, but such wet lab experiments are currently not possible in our laboratory due to pandemic-related shutdowns.

For systems larger than 50 nucleotides, we generated large model sets and clustered the 400 lowest energy structures with a 5 Å RMSD clustering radius, the procedure used for similarly sized systems in our recent FARFAR2 modeling benchmark. (Here and below, RMSD was computed as the all-heavy-atom RMSD between two models, as in all recent Rosetta work on RNA modeling.) For smaller systems, we clustered the 400 lowest energy structures with a 2 Å RMSD clustering radius, analogous to the procedure used for similarly sized systems in the FARFAR2 study. We note here that the clustering method follows the protocol used in the FARFAR2 and other Rosetta RNA studies, where cluster centers are drawn from the very lowest energy models observed (see Methods). We make available models across up to 10 clusters in the resulting FARFAR2-SARS-CoV-2 data set to help efforts in virtual screening that take advantage of ensembles.

We note that these models and their frequency in clusters are not necessarily an accurate representation of the thermodynamic ensemble obtained by the RNA due to biases in Rosetta FARFAR2 sampling and inaccuracies in the Rosetta all-atom free energy function. Nevertheless, the models offer a starting point of physically realistic conformations for virtual ligand screening and more sophisticated approaches to thermodynamic ensemble modeling.

In Figures 1-4, we focus on simulations that were sufficiently converged to produce multiple-occupancy clusters, a hallmark of FARFAR2 sampling being able to discover lowest energy states, as evaluated in Rosetta's all-atom energy function. Other runs with poorer convergence are provided in Supplementary Figures 1-3; these cases would need more computer power to discover lowest energy states. In Table 1, we report the E-gap: the difference in Rosetta energy units (R.E.U.) for best-scoring models in each cluster compared to the top-scoring model in the simulation overall. Rosetta energy functions have been fit such that R.E.U. estimate energies in kcal/mol,[14] so E-gap values similar to or smaller than 3.0 indicate structures that are predicted to make up a significant fraction of the ground state ensemble and that may be trapped by small molecule drugs without a major cost in binding affinity.

For each simulation, we additionally report a convergence estimate in Table 1, estimated as the mean pairwise RMSD of the top 10 cluster centers predicted by FARFAR2. Prior work aiming at accurate prediction of single native crystal structures has demonstrated that convergence is a predictor for modeling accuracy,[10, 15-16] with prior tests suggesting that models that have 7.5 Å convergence or lower have mean single-structure prediction accuracy of at least 10 Å, and models with 5 Å convergence or lower have single-structure prediction accuracy of at least 8 Å.[10] In this work, we are however not assuming that the RNA targets form a single 'native' structure. Instead, we take this convergence measure as a proxy for whether sampling may have been adequate to generate a useful model set. As a more direct measure of the thoroughness of sampling, we also present the occupancies of each of the top 10 clusters. Conformations sampled repeatedly in independent Rosetta-FARFAR2 runs (cluster membership greater than 1) indicate some level of convergence in sampling and those conformations are more likely to be realistic low-energy

structures. In Figs. 1-4 below, we show cluster members as a cloud of translucent structures behind each cluster's lowest energy conformation to visually convey the level of convergence; lack of such a cloud indicates a 'singlet' in which the cluster involves only one member. We include up to 50 representative top-scoring models in each cluster as the model collection for each RNA element, with structures available in the Github repository: https://github.com/DasLab/FARFAR2-SARS-CoV-2.

With these caveats and metrics of convergence in mind, a description of models from the extended 5′ UTR, frameshifting element, and 3′ UTR follows.

Models of SARS-CoV-2 extended 5´ UTR

Fig. 1 presents models for the stem loops that make up the extended 5´ UTR. Also called the 5′-proximal region, this region extends the 5´ UTR by ~100 residues to bracket potential structures that involve the beginning of the coding region. The assumed secondary structure depicted in Fig. 1A is largely based on previous dissection of betacoronavirus secondary structures by several groups.[2, 17] More specifically, secondary structures for SL1-5 in the 5' UTR are based on homology to prior betacoronaviruses, where these conserved stems have been confirmed through genetic experiments and sequence alignments in related betacoronaviruses. (We note for non-coronavirus researchers here that SL1, SL2, SL4, and SL5 have also been termed SLI, SLII, SLIII, and SLIV in an important set of studies.[18-20])

Some of these stems have structural preferences that have proven critical to the viral life cycle based on genetic experiments in other betacoronaviruses; these preferences have the potential to be altered through the binding of small-molecule drugs. Prior genetic selection experiments have demonstrated a preference for mutations that destabilize SL1. The lower part of this stem must unpair to allow for the formation of a long-range RNA contact between the 5′ and 3′ UTRs.[9] The stem must also presumably unfold to enable cap-dependent initiation of translation by the human ribosome. The loop in SL2 has sequence features across betacoronaviruses consistent with a U-turn conformation, and mutations that disrupt this structure are not viable, leading to a loss of sub-genomic RNA synthesis.[7] SL3 presents the transcription regulation sequence of the leader (TRS-L), which must be available to base-pair with TRS-B binding partners in the negative-strand viral genome to facilitate sub-genomic RNA synthesis.[21] SL4 has been shown to be functionally important, and also harbors an upstream open reading frame (uORF) across many betacoronaviruses. Here computational secondary structure modeling returned an additional stem immediately 3′ of SL4, and we have modeled them together as SL4ab. SL5 is a well-established domain that, in SARS-related viruses, has a long stem elaborated with a 4-way junction; this element has been proposed to harbor packaging signals, and it harbors the AUG start codon for the genome's first gene product, the ORF1a/b polyprotein. Stems SL6 and SL7, downstream of the start codon, were modeled through computational secondary structure prediction. SL6 and SL7, computationally predicted to occur downstream of the AUG start codon, are included as potentially analogous to stems discovered to be important in bovine coronavirus but not yet established in SARS-related viruses.[22]

We first produced models for the full extended 5′ UTR (over 10,000 FARFAR2 models), with the top-scoring structure depicted in Supplementary Figure 1. With the current level of sampling, this and

other lowest energy structures appear only once amongst our models (Table 1), reducing any confidence that these models capture lowest energy conformations in reality. A construct of this size would require substantially more models to achieve sufficient modeling convergence. Nevertheless, the top-scoring models suggest the potential for compact RNA structures for the 5´ UTR mediated by potential tertiary contacts between stem-loops. While such tertiary contacts are of potential interest, significantly more computational power would need to be expended to have confidence that such collapsed states are well-defined low energy states and could act as therapeutic targets. We therefore turned to smaller segments of the extended 5´ UTR for which Rosetta-FARFAR2 had a reasonable prospect of achieving convergent models.

For each stem-loop in the 5′ UTR and initial stretch of the ORF1a/b coding region, we generated at least 200,000 FARFAR2 models. In Fig. 1B-G, we depict clusters for each 5′ UTR stem-loop from simulations that produced multiple-occupancy clusters, including SL1, SL2, SL3, SL4ab, SL6, and SL7. As expected for these smaller RNA segments, all these stem-loops had excellent modeling convergence. Most clusters had occupancies of greater than 1, indicating that numerous independent de novo modeling trajectories resulted in conformations similar to within 2 Å RMSD. Furthermore, the mean pairwise RMSD of 10 lowest energy cluster centers (Table 1) approached 5 Å or better, suggesting that if a single dominant structure exists for these elements, our average model accuracy would be around 8 Å RMSD or better. Nevertheless, SL1-3 have four or more clusters with E-gap values less than 1 R.E.U., suggesting the presence of many distinct structures at a 2 Å clustering radius that may be trapped by a small molecule without substantial energetic penalty.

In each of these RNA elements, stereotyped configurations of apical loops are modeled by Rosetta-FARFAR2 (Figs. 1B-G). However, it is important to note that our modeling of SL2 does not recover structures with the atomic level features of the previously reported canonical CUYG tetraloop, whose structure was determined by NMR for the SARS-CoV-1 sequence[4] (Fig. 1C). We present a homology model of SL2 based on the SARS-CoV-1 sequence as an additional cluster in our data set (Fig. 1C, see Methods). Our de novo FARFAR2 models approach 3.1 Å RMSD to this homology-directed model, consistent with our convergence estimate above. Nevertheless, there is still an atomic level discrepancy between the NMR structure and FARFAR2 clusters. This difference likely reflects limitations in Rosetta's high-resolution energy function, with the E-gap value for the homology-directed SL2 structure over 20 R.E.U. worse than the top-scoring de novo model. The primary issue is that the Rosetta energy function assigns un-stacked base pairs, as in the C and G nucleotides of the NMR structure of the CUYG tetraloop, a substantial desolvation penalty without stacking bonuses to compensate. The fine structural discrepancy may additionally reflect limitations in Rosetta FARFAR2 sampling, which might be resolved by bringing to bear more computational expensive sampling methods like stepwise Monte Carlo (SWM).[23]

The SL5 element is a long stem loop in all betacoronaviruses whose tip has been elaborated into a 4-way junction in SARS-CoV-2 and related subgroups.[2] Due to the larger size and complexity of SL5, our 3D models for this domain did not converge sufficiently, i.e., each of the top 10 lowest energy models were 'singlets' with no other conformations discovered within 5 Å RMSD among the top 400 lowest energy models (Table 1). Nevertheless, the modeling did suggest the potential for drug-binding pockets between helices that are brought into proximity by the four-way junction (Fig. 2A). It was of interest to test whether the secondary structure of SL5 might be well-defined even if

our 3D modeling did not converge. To test the secondary structure of this stem-loop, we collected DMS and SHAPE data for this construct (see Methods). Secondary structure predictions from RNAstructure[24] guided by these DMS and SHAPE reactivity data (Fig. 2B) recovered the same secondary structure developed from homology modeling to SL5 elements in other betacoronaviruses, as depicted in Fig. 1A, supporting the accuracy of this model as the dominant secondary structure.

Models of SARS-CoV-2 frameshifting element

Fig. 3 presents Rosetta-FARFAR2 models for the SARS-CoV-2 frameshifting element (FSE). In SARS-CoV-1, the FSE pseudoknot structure and its dimerization domain have been shown to be critical for a (-1) ribosomal frameshifting event that leads to the production of ORF1a and ORF1b proteins from the same genomic region (Fig. 3A). For this reason, we present models for both the monomer frameshifting element along with a dimerized system. We generated over 200,000 FARFAR2 models for the monomeric frameshifting element, with 100,000 models separately modeled with each of two similar secondary structures reported in the literature, which differ by a single base pair (Fig. 3).[1, 25] The frameshifting element simulation reached 14 Å convergence and yielded six clusters with E-gap values below 8 R.E.U., with five of these clusters having more than one cluster member at a 5 Å RMSD clustering radius.

Despite the presence of various multiple-occupancy clusters, compared to the smaller stem-loops modeled in the 5′ UTR, the frameshifting element models were less converged. The majority of variation between top models arises from different orientations of the 5′ "slippery sequence," which forms a weak stem-loop of variable orientation in Rosetta-FARFAR2 simulations. The convergence of the pseudoknot region excluding this 5´ slippery sequence was tighter, 8.7 Å, suggesting that low-energy conformations of this pseudoknot element are captured by the current models. Models with both of the assumed secondary structures produced low energies; they contribute to distinct but similar clusters. We additionally generated over 20,000 FARFAR2 models for the dimerized frameshifting element. Modeling of the dimerized frameshifting element did not converge on a well-defined 3D structure (each of the 10 lowest energy conformations were 'singlets' with no other conformations discovered within 5 Å RMSD; Table 1 and Supplementary Figure 2).

Models of SARS-CoV-2 3´ UTR

Fig. 4 presents models for structured regions of the 3′ UTR. The 3′ UTR includes a switch-like pseudoknot element on the 5′ end, a hypervariable region, and the stem-loop II-like motif (Fig. 4A); all these regions have secondary structures built based on homology to other betacoronaviruses.[2, 17] The 3′ UTR pseudoknot along with its mutually exclusive stem-loop structure have suggested functional roles in viral RNA synthesis, with mutations that destabilize either the pseudoknot or the stem-loop structure proving inviable in related betacoronaviruses.[8] The structure of the stem-loop II-like motif resembles that of an rRNA loop, leading to its proposed role in recruiting host translation machinery.[5]

We generated over 10,000 models for the full 3´ UTR. Analogous to the models generated for the extended 5′ UTR, this simulation was far from reaching convergence (Table 1), but suggested the possibility of stereotyped RNA 3D structure, perhaps recognized by factors such as the virus

replicase (Supplementary Figure 3). We additionally generated over 20,000 models for two large subregions of the 3´ UTR, the hyper-variable region and a region including the 3′ UTR pseudoknot (Supplementary Figure 3). Here we built models for the 3′ UTR pseudoknot based on the secondary structure forming the pseudoknot P1Pk as depicted in Fig. 4A. Future simulations of this region should also include models based on the alternative secondary structure for this switch-like element forming a stem extending P0b. In both cases, we were not able to observe convergence; none of the top 10 lowest energy conformations were similar to each other or to any of the other top 400 lowest energy conformations, as evaluated by RMSD with cutoff 5 Å. This was expected: with nearly 150 nucleotides in the 3′ UTR pseudoknot and with the hyper-variable region having limited homology to prior structures, these two modeling cases were the largest sub-domains modeled in this report, and likely will remain challenging to model even with more computational power. Modeling of smaller sub-domains of the 3′ UTR and/or acquisition and incorporation of experimental data will likely be critical to model reasonable starting ensembles for virtual docking algorithms. As a preview of such modeling, we used the crystal structure of the SARS stem-loop II-like motif (s2m) to build over 200,000 homology models for the SARS-CoV-2 s2m with FARFAR2 (Fig. 4B, see Methods).[5, 26] With near-identical sequences, the SARS s2m template crystal structure is already a near-complete model for the SARS-Cov-2 domain, such that the FARFAR2 homology models for this region are highly converged (0.21 Å).

Models of riboswitch aptamers as a benchmark for virtual drug screening methods

While there are not yet extended benchmarks for using FARFAR2 model ensembles for small molecule virtual screening, recent RNA-puzzles blind prediction trials have shown that FARFAR2 models in combination with conservation information enables manual identification of ligand binding sites in 3D models of bacterial 'riboswitch' aptamers at nucleotide resolution.[11-13] To guide use of FARFAR2 models for virtual screening, we have therefore collected a data set termed **FARFAR2-Apo-Riboswitch**, containing models of RNA elements that are known to bind small molecules, depicted in Fig. 5. These targets include binders for diverse ligands, from S-adenosyl methionine (SAM) to glycine, cobalamin, 5-hydroxytryptophan, the cyclic dinucleotides c-di-AMP (the ydaO riboswitch), the alarmone nucleotide ZMP (AICAR monophosphate), glutamine, and guanidinium. Each RNA was modeled without ligands present during fragment assembly or refinement, to mimic the protocols that would be used in virtual drug screening in which modeling of *apo* RNA structures are used for computational docking of ligands.

Three of these model sets for SAM-I, SAM-I/IV, and SAM-IV made use of homology to previous riboswitch structures' ligand binding sites (Homology, Figs. 5A-C), for historical reasons: the actual RNA-Puzzles challenges (or in the case of SAM-IV, an 'unknown RFAM' challenge for the RNA-Puzzles, involving tests based on cryoEM[27]) were posed at times in which crystal structures of riboswitch aptamers with homologous SAM binding sites were available. These model sets should therefore serve as 'positive controls' for virtual drug screening protocols, which should unambiguously identify homology-guided SAM binding sites as good aptamers for SAM.

The modeling also included riboswitch aptamer cases (De novo, Figs. 5D-J) in which the ligand binding sites were not modeled by homology, in closer analogy to virtual screening approaches that might make use of the FARFAR2-SARS-CoV-2 models. These models include cases such as the ydaO riboswitch, where modeling did not achieve a model closer than 10.0 Å RMSD to the RNA

crystallized with two cyclic-diAMP ligands, perhaps owing to the large ligand and the substantial degree to which contacts with that ligand may organize the crystallized conformation. It will be interesting to see if these models still allow recognition of small molecule binding sites by computational methods. We provide in Table 2 additional metrics for these model sets, including the same RMSD convergence estimates, cluster occupancy, and E-gap numbers as for our FARFAR2-SARS-CoV-2 models as well as RMSD to experimentally determined ligand-bound structures. It will be interesting to see if these metrics give any correlation with the ability of virtual screening methods to discover known native ligands for these RNA elements, which will be useful in evaluating the likelihood of success of such methods in discovering molecules for the FARFAR2-SARS-CoV-2 models.

Discussion

We provide a collection of 3D models for elements comprising the extended 5′ UTR, frameshifting element, and 3′ UTR of the SARS-CoV-2 RNA genome. We hope that this FARFAR2-SARS-CoV-2 dataset provides a starting point for virtual screening approaches seeking compounds that stabilize individual SARS-CoV-2 RNA conformations, preventing access to alternative conformations required for viral replication. Models for the 5′ UTR SL1-4 and the frameshifting element appear to be especially promising candidates for small-molecule drug discovery. Modeling of these elements gave ensembles converging sufficiently to produce multiple representatives in clusters with 5 Å RMSD, and these elements all harbor sequences conserved across betacoronaviruses and structures having documented functional roles in the replication and/or translation of betacoronavirus genomes.

The structural ensembles presented here have a number of limitations. First, these ensembles are not true thermodynamic ensembles in that the structure occupancies do not necessarily reflect the underlying probabilities of occurrence for each conformational state. Additionally, some of the simulation ensembles described here did not achieve sufficient convergence to provide confidence in the resulting models (3′ UTR hypervariable region, 3′ UTR pseudoknot) – that is, independent modeling runs did not converge to similar low-energy structures. It is possible that these RNA elements do not have well-defined 3D structures in solution unless bound tightly to partners such as the replicase complex. Alternatively or in addition, our modeling methods and currently available computational power are not well-suited to regions of this size. While additional sampling may alleviate this problem, these regions have more de novo modeled positions than most prior FARFAR2 benchmark cases and may remain challenging for current de novo RNA modeling approaches.

Given these limitations in de novo modeling, we felt that it was important to provide analogous models of RNAs of known structure. Virtual screening approaches appear poised to make good use of computational models of RNA, but have so far, to our knowledge, not been used on de novo predicted models. To provide benchmark structural ensembles for such efforts, we have therefore used the same Rosetta-FARFAR2 modeling method and model selection procedure for a variety of RNA riboswitches with known small-molecule ligands. We hope this FARFAR2-Apo-riboswitch data set provides an opportunity for testing the accuracy of virtual screening approaches that use FARFAR2 model sets. With the current need for SARS-CoV-2 antiviral discovery, we believe this is

an opportune time to explore and evaluate new approaches for virtual screening of small-molecule drug candidates that target structured RNA.

Methods

Rosetta modeling

For each target, FARFAR2 simulations were executed as previously described,[10] providing sequence and secondary structure information to the *rna_denovo* application provided with Rosetta 3.12. Clustering was achieved via the *rna_cluster* application, examining the top 400 structures by energy with a cluster radius of 5.0 or 2.0 Å depending on the structural element as described above. Specifically, the rna_cluster application iterates through unclustered structures from best to worst energy, either assigning them to an existing cluster center (if the all-heavy-atom RMSD of the model is within the clustering radius of the cluster center) or starting a new cluster. Homology modeling with FARFAR2 for the 5′ UTR SL2 and the 3′ UTR stem-loop II-like motif (s2m) was carried out using the approach outlined in ref.[26] For the 5′ UTR SL2, PDB ID 2L6I[4] was used as a template for positions 45 to 59. For the 3′ UTR s2m, PDB ID 1XJR[5] was used as a template for positions -144 to -104; here, nucleotide numbering maps to the 3′ UTR secondary structure in Fig. 4.

Chemical reactivity experiments

The DNA template for stem-loop 5 RNA was amplified from a gBlock sequence for the extended 5′ UTR (primers and gBlock sequences ordered from Integrated DNA Technologies, sequences in Table S1). The primers were designed to add a Phi2.5 promoter sequence (TTCTAATACGACTCACTATT) at the amplicon's 5' end and a 20 base-pair Tail2 sequence (AAAGAAACAACAACAACAAC) at its 3' end. The PCR reaction contained 5 ng of gBlock DNA template, 2 µM of forward and reverse primer, 0.2 mM of dNTPs, 2 units of DNA Phusion DNA polymerase, and 1X of HF buffer. The reaction was first denatured at 98 °C for 30 s. Then for 35 cycles, the sample was denatured at 98 °C for 10 sec, annealed at 64 °C for 30 sec, and extended at 72 C for 30 °C. This was followed by an incubation at 72 °C for 10 min for a final extension. Assembly products were verified for size via agarose gel electrophoresis and subsequently purified using Agencourt RNAClean XP beads. Purified DNA was quantified via NanoDrop (Thermo Scientific) and 8 pmol of purified DNA was then used for in vitro transcription with T7 RNA polymerase (New England Biolabs Inc.). The resulting RNA was purified with Agencourt RNAClean XP beads supplemented with an additional 12% of PEG-8000 and quantified via NanoDrop.

For RNA modification, 1.2 pmol of RNA was denatured in 50 mM Na-HEPES pH 8.0 at 90 °C for 3 minutes and cooled at room temperature for 10 minutes. The RNA was then folded with the addition of $MgCl_2$ to a final concentration of 10 mM in 15 µL and incubated at 50 °C for 30 minutes, then left at room temperature for 10 minutes. For chemical modification of folded RNA, fresh working stocks of dimethyl sulfate (DMS) and 1-methyl-7-nitroisatoic anhydride (1M7) were prepared. For DMS, 1 µL of DMS was mixed with 99 µL of 100% EtOH. The 100 µL solution was then added to 100 µL RNase free H_2O for a final volume of 200 µL. For 1M7, 4.24 mg of 1M7 was dissolved in 1 mL of anhydrous DMSO. For a no-modification control reaction, 5 µL of Rnase free H_2O was added to 15 µL of folded RNA. Samples were incubated at room temperature for 15 minutes. Then, 5 µL of 5 M

NaCl, 1.5 µL of oligo-dT Poly(A)Purist MAG beads (Ambion), and 0.065 pmol of 5' Fluorescein (FAM)-labeled Tail2-A20 primer were added (sequence in Table S1), and the solution was mixed and incubated for 15 minutes. The magnetic beads were then pulled down by placing the mixture on a 96-post magnetic stand, washed twice with 100 µL of 70% EtOH, and air dried for 10 minutes before being resuspended in 2.5 µL RNase free H_2O.

For cDNA synthesis, 2.5 µL resuspension of purified, polyA magnetic beads carrying chemically modified RNA was mixed with 2.5 µL of reverse transcription premix with SuperScript-III (Thermo Fisher). The reaction was incubated at 48 °C for 45 minutes. The RNA was then degraded by adding 5 µL of 0.4 M NaOH and incubating the mixture at 90 ℃ for 3 minutes. The degradation reaction was placed on ice and quickly quenched by the addition of 2 µL of an acid quench solution (1.4 M NaCl, 0.6 M HCl, and 1.3 M NaOAc). Bead-bound, FAM labeled cDNA was purified by magnetic bead separation, washed twice with 100 µL of 70% EtOH, and air-dried for 10 minutes. To elute the bound cDNA, the magnetic beads were resuspended in 10.0625 µL ROX/Hi-Di (0.0625 µL of ROX 350 ladder [Applied Biosystems] in 10 µL of Hi-Di formamide [Applied Biosystems]) and incubated at room temperature for 20 minutes. The resulting eluate was loaded onto capillary electrophoresis sequencers (ABI-3100 or ABI-3730) either on a local machine or through capillary electrophoresis (CE) services rendered by ELIM Biopharmaceuticals.

CE data was analyzed using the HiTRACE 2.0 package (https://github.com/ribokit/HiTRACE)[28], following the recommended steps for sequence assignment, peak fitting, background subtraction of the no-modification control, correction for signal attenuation, and reactivity profile normalization. Secondary structures were predicted using the Biers (https://github.com/ribokit/Biers) wrapper for RNAStructure[24] with default settings, making one prediction guided by the normalized DMS reactivity profile and another guided by the normalized 1M7 reactivity profile.

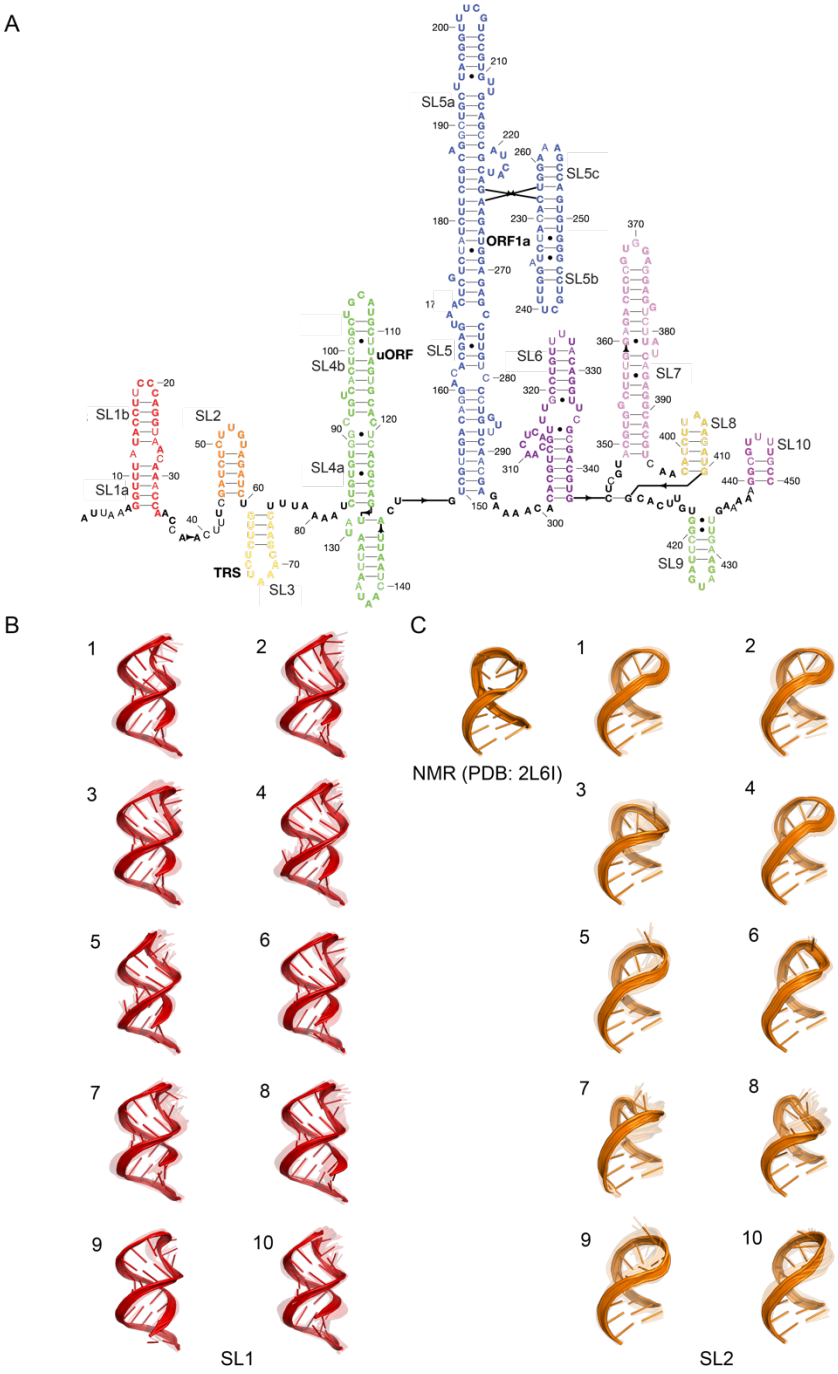
Data availability

The supplementary file includes depictions of top 10 clusters for the full extended 5′ UTR, frameshift element dimer, full 3′ UTR, hypervariable region, and an extended 3′ UTR pseudoknot construct. FARFAR2-SARS-CoV-2 models are included at https://github.com/DasLab/FARFAR2-SARS-CoV-2. FARFAR2-Apo-Riboswitch models are included at https://github.com/DasLab/FARFAR2-Apo-Riboswitch.

Acknowledgments

We thank I. Zheludev for extensive discussions of RNA elements to prioritize for structural characterization. We thank Sergey Lyskov and Jeff Gray for expedited access to the ROSIE cluster to pilot Rosetta modeling efforts. We acknowledge funding from the National Science Foundation Graduate Research Fellowship Program under grant no. 1650114 and National Institutes of Health (R21 CA219847 and R35 GM122579). The computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

Figures



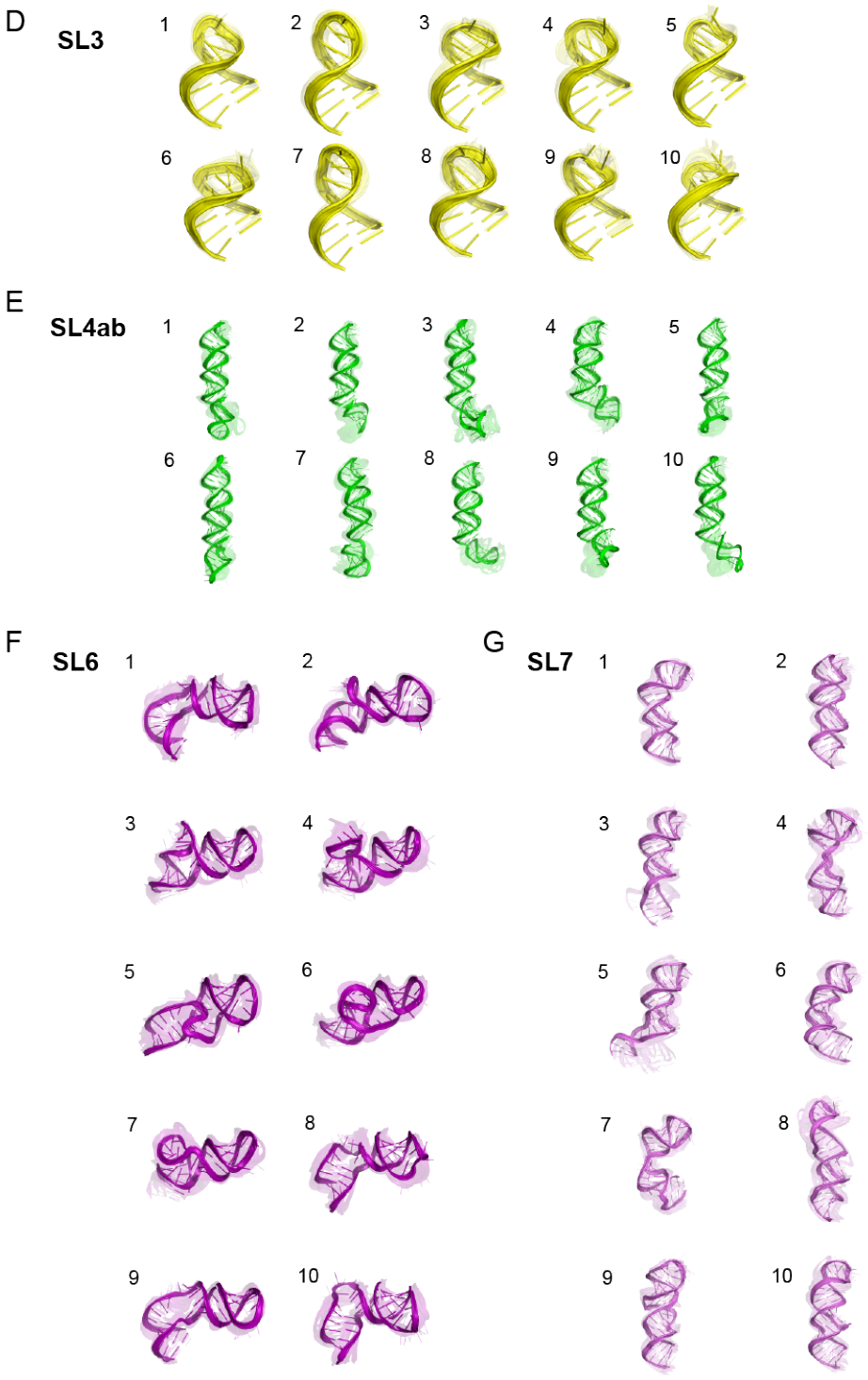


Figure 1. A) 5′ UTR secondary structure, with each modeled stem colored. In bold are positions that are completely conserved across a set of SARS-related virus sequences.[1] B-G) Top 10 clusters for SL1, SL2, SL3, SL4ab, SL6, and SL7. The top-scoring cluster member in each case is depicted with solid colors, and the top cluster members (up to 10) are depicted as transparent structures.

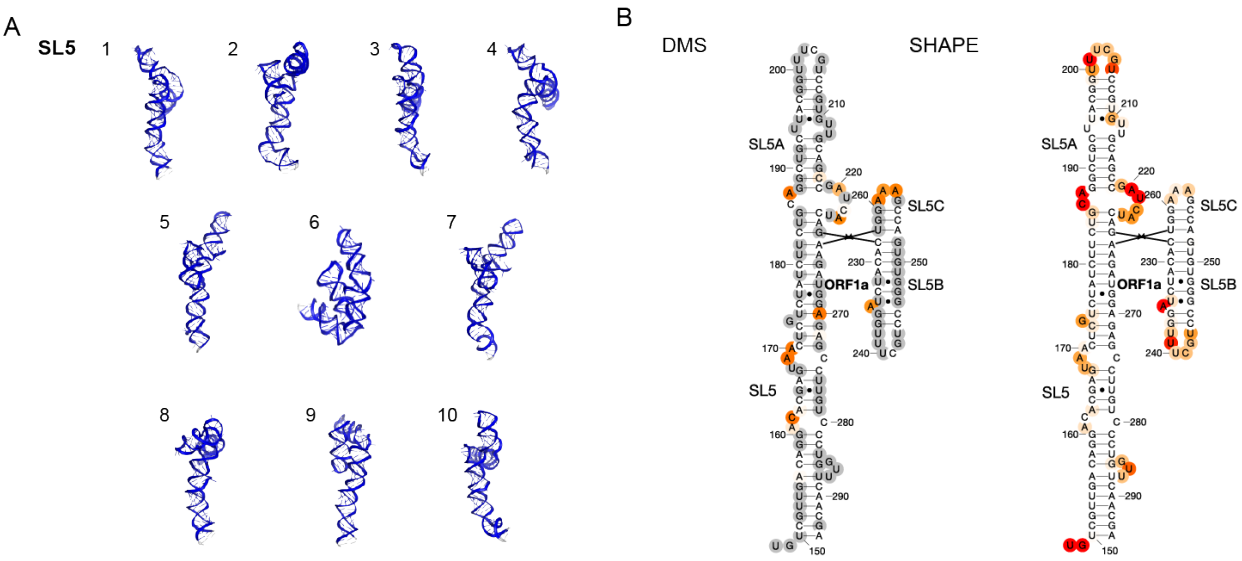


Figure 2. A) Top 10 clusters for SL5 from FARFAR2 modeling. 3D Modeling did not converge, with only one member per cluster with a 5 Å clustering radius. B) Secondary structures for the SARS-CoV-2 SL5 predicted from RNAstructure[24] guided by DMS and SHAPE reactivity data validate the SL5 secondary structure modeled based on homology to literature models (Fig. 1A). DMS and SHAPE reactivity are colored for each position, with red indicating high reactivity, white indicating low reactivity, and grey indicating positions that cannot be modified.

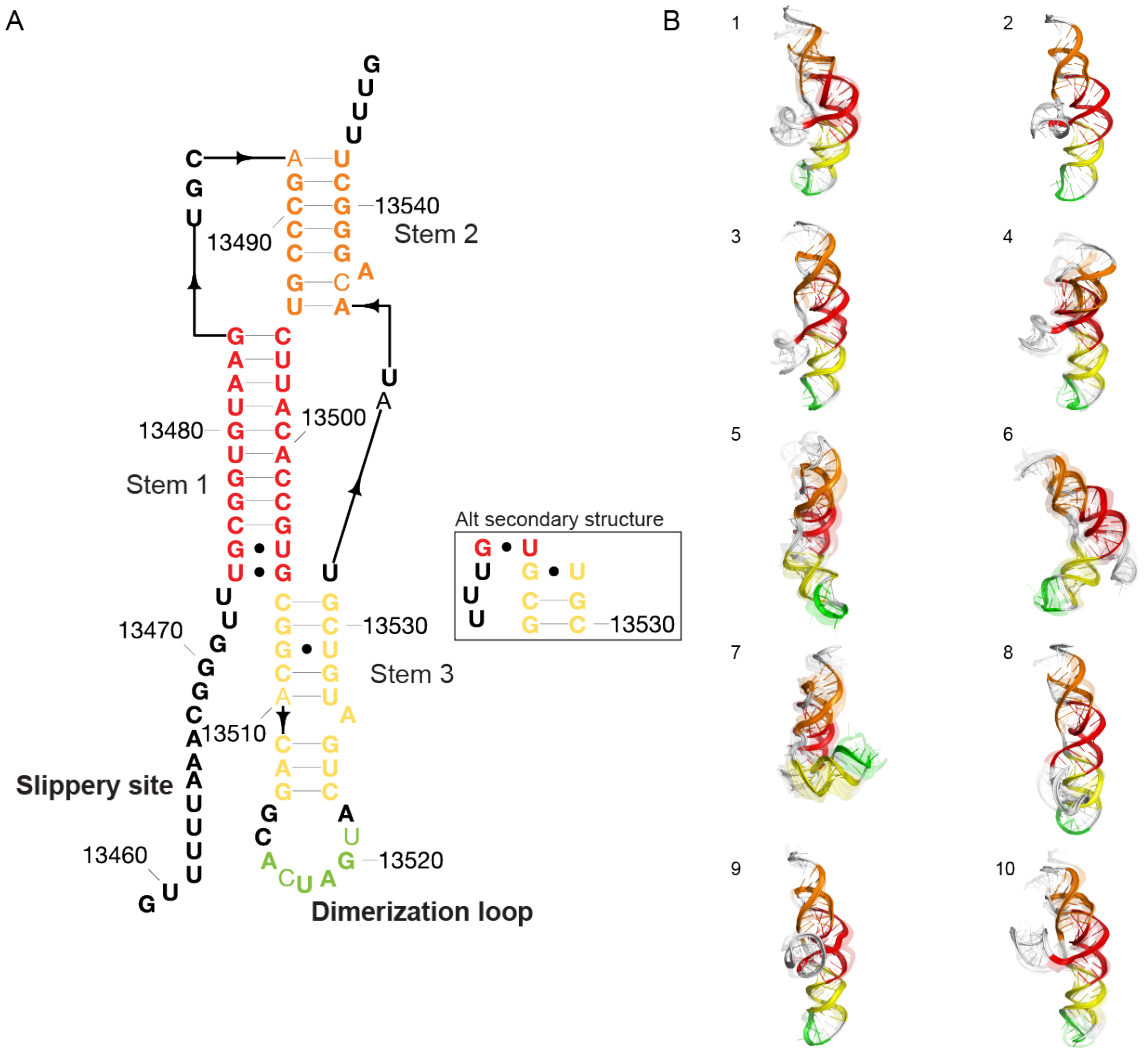


Figure 3. A) Frameshifting element secondary structure, depicting two secondary structures used for FARFAR2 modeling. In bold are positions that are completely conserved across a set of SARS-related virus sequences.[1] B) Top 10 frameshifting element clusters. The top-scoring cluster member in each case is depicted with solid colors, and the top cluster members (up to 10) are depicted as transparent structures.

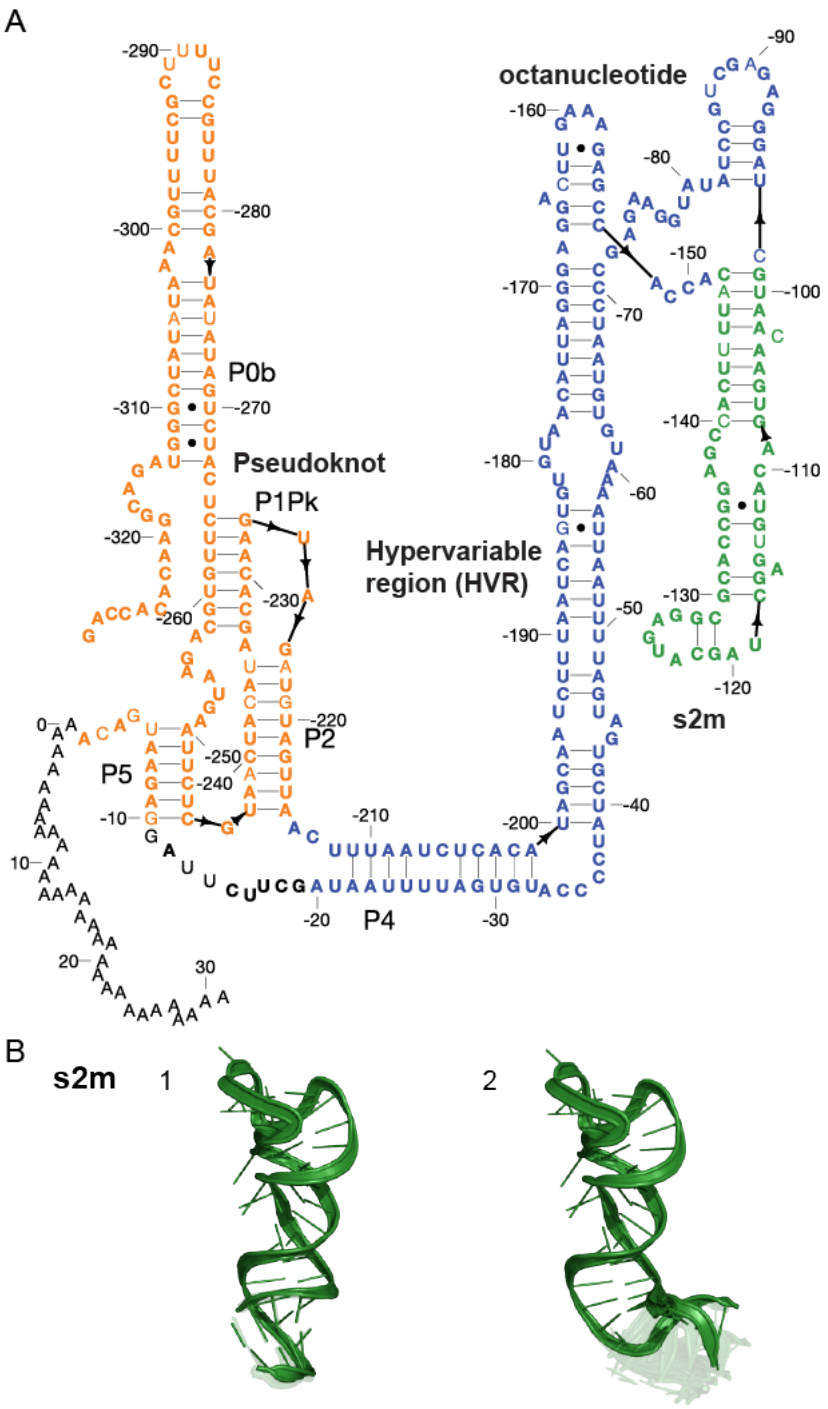
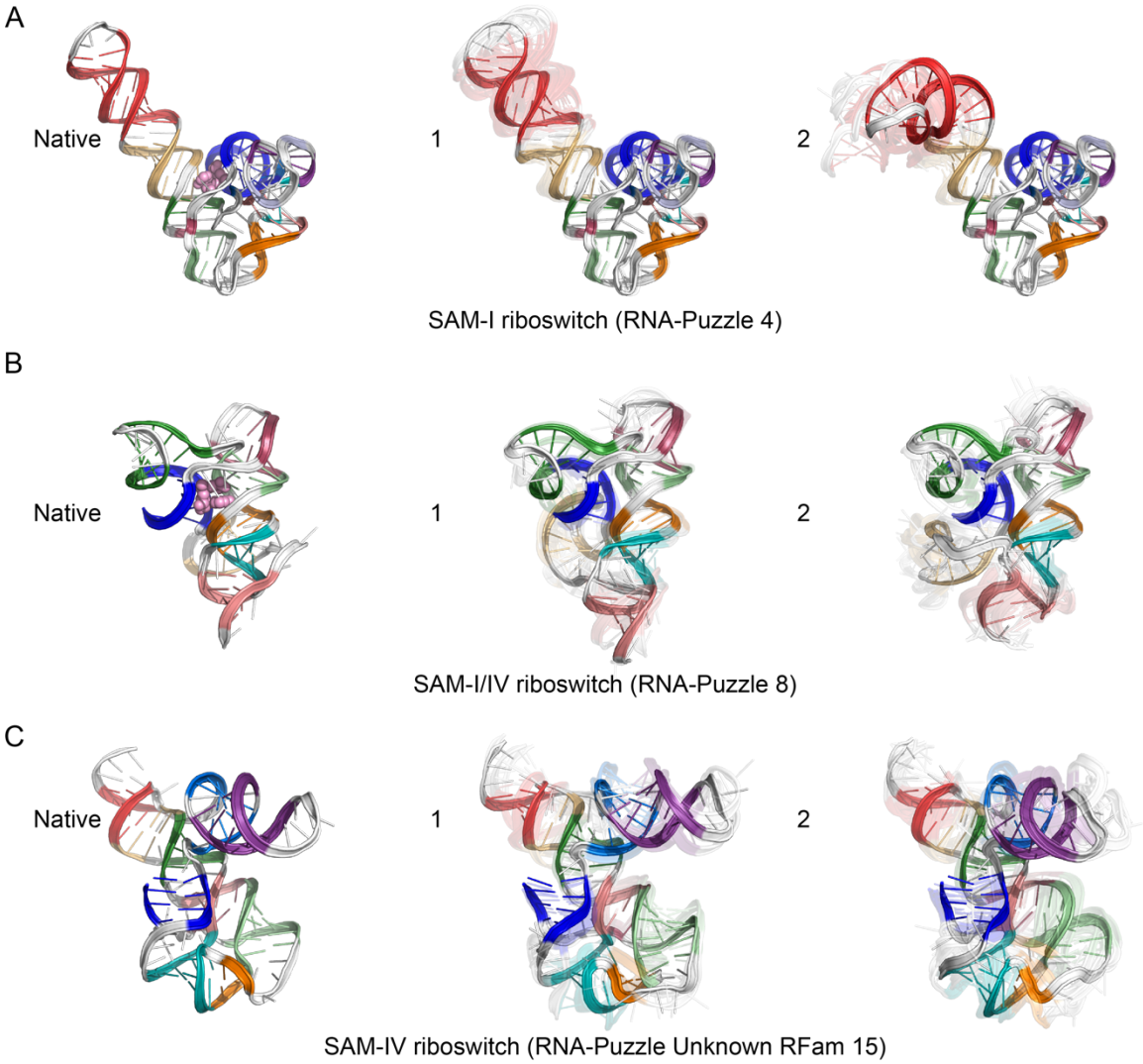
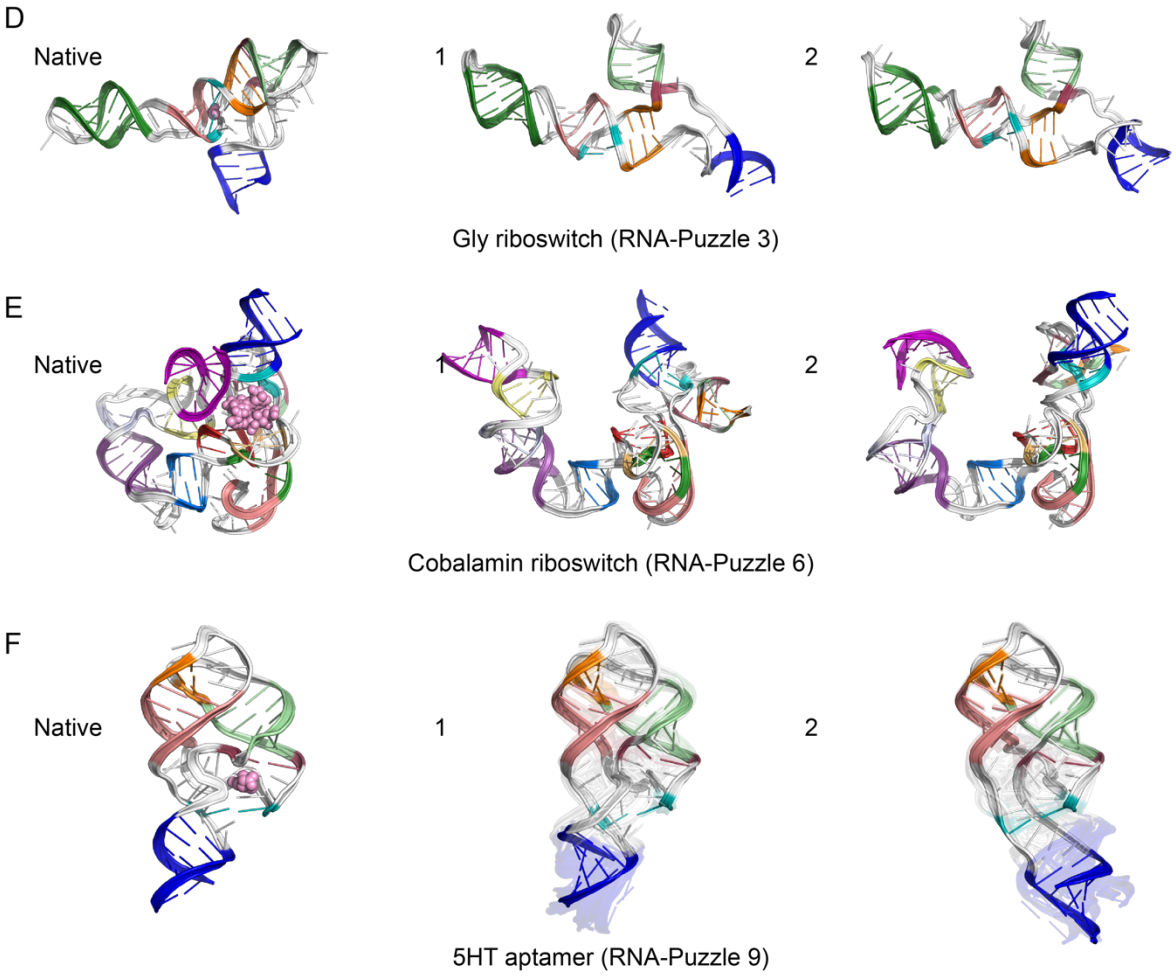


Figure 4. A) 3′ UTR secondary structure, colored by domain. In bold are positions that are completely conserved across a set of SARS-related virus sequences.[1] B) Top 2 clusters for the stem-loop II-like motif, with models built based on homology to template structure 1XJR[5]. The top-scoring cluster member in each case is depicted with solid colors, and the top cluster members (up to 10) are depicted as transparent structures.

HOMOLOGY



DE NOVO



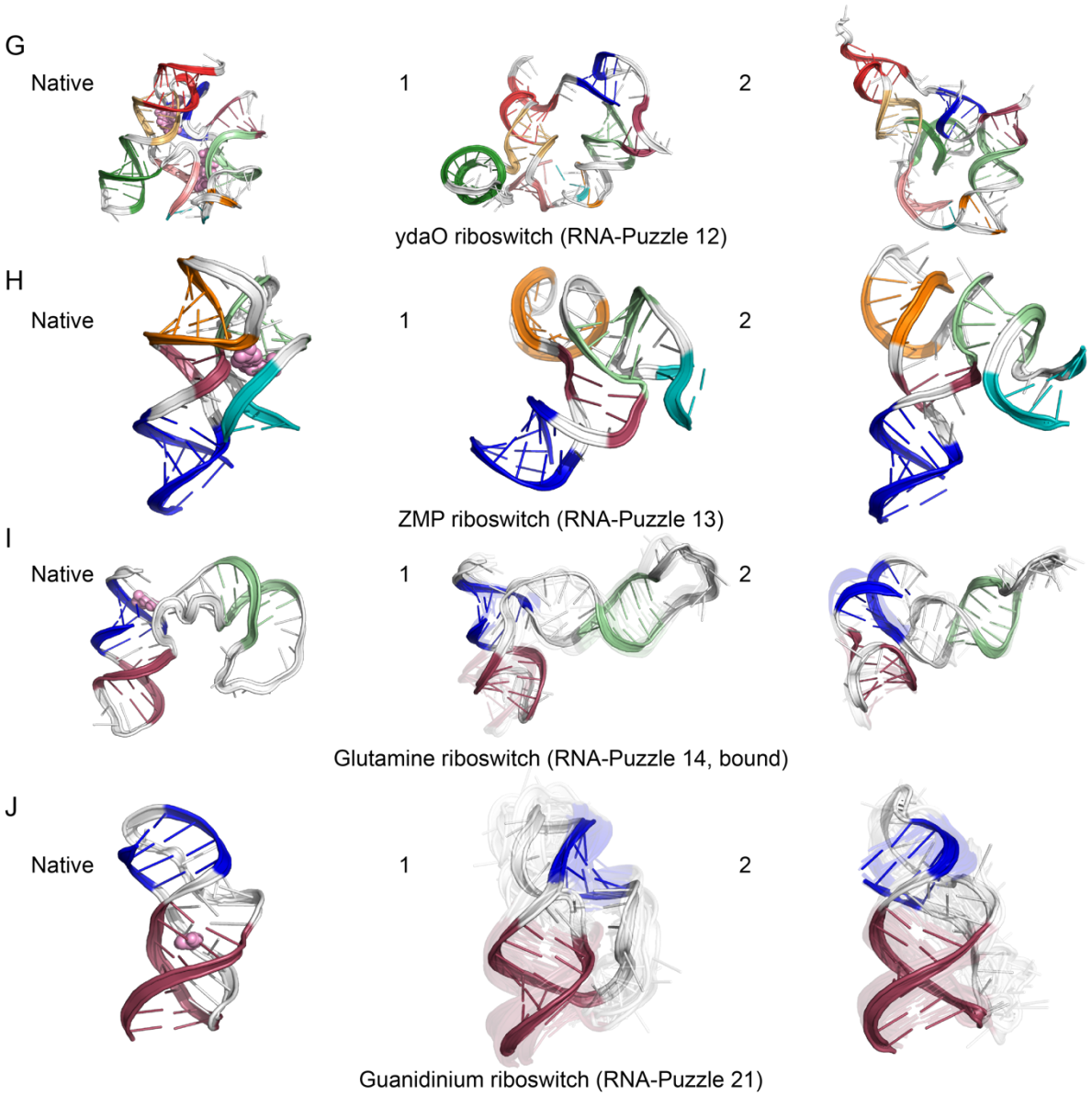


Figure 5. Top two clusters for the following small-molecule binding RNA riboswitches and aptamers, along with their native structure (left in each panel, with ligand in pink spheres), beginning with three template-guided modeling challenges: A) SAM-I riboswitch, B) SAM-I/IV riboswitch, C) SAM-IV riboswitch, D) glycine riboswitch, E) cobalamin riboswitch, F) 5HT riboswitch, G) cyclic diGMP ydaO riboswitch , H) ZMP riboswitch, I) glutamine riboswitch, J) guanidinium riboswitch. The top-scoring cluster member in each case is depicted with solid colors, and the top cluster members (up to 10) are depicted as transparent structures.

Table 1. FARFAR2-SARS-CoV-2 models

System/cluster	Length	Models generated	Model convergence (Å)[a]	E-gap to lowest energy model (REU)[b]	Cluster occupancy[c]
SARS-CoV-2 extended 5′ UTR (1–450)					
5′ UTR/1	450	10000	45.47	0.0	1
5′ UTR/2				5.04	1
5′ UTR/3				12.78	1
5′ UTR/4				18.77	1
5′ UTR/5				19.00	1
5′ UTR/6				28.19	1
5′ UTR/7				28.93	1
5′ UTR/8				33.19	1
5′ UTR/9				24.71	1
5′ UTR/10				35.64	1
5′ UTR stem-loop 1 (7–33)					
SL1/1	27	200000	1.41	0.0	6
SL1/2				0.07	16
SL1/3				0.25	45
SL1/4				0.91	13
SL1/5				0.97	6
SL1/6				1.02	18
SL1/7				1.04	9
SL1/8				1.28	14
SL1/9				1.56	4
SL1/10				1.57	12
5′ UTR stem-loop 2 (45–59)					
SL2/1	15	200000	1.18	0.0	52
SL2/2				0.73	44
SL2/3				0.95	78

SL2/4				0.98	56
SL2/5				1.08	15
SL2/6				1.13	207
SL2/7				1.15	26
SL2/8				1.38	29
SL2/9				1.40	13
SL2/10				1.48	41
SL2/Homology				20.2	27
5′ UTR stem-loop 3 (61–75)					
SL3/1	15	200000	2.10	0.0	7
SL3/2				0.10	45
SL3/3				0.61	10
SL3/4				0.78	69
SL3/5				1.31	70
SL3/6				1.55	6
SL3/7				1.89	84
SL3/8				1.93	49
SL3/9				2.10	35
SL3/10				2.23	18
5′ UTR stem-loop 4ab (84–146)					
SL4ab/1	63	143897	3.53	0.0	645
SL4ab/2				5.28	377
SL4ab/3				5.79	21
SL4ab/4				8.29	17
SL4ab/5				8.79	202
SL4ab/6				9.07	560
SL4ab/7				11.76	15
SL4ab/8				11.83	25
SL4ab/9				12.02	177
SL4ab/10				12.58	742
5′ UTR stem-loop 5 (148–295)					
SL5/1	148	75861	20.86	0.0	1

SL5/2				2.14	1
SL5/3				9.19	1
SL5/4				10.96	1
SL5/5				11.60	1
SL5/6				11.78	1
SL5/7				13.15	1
SL5/8				13.99	1
SL5/9				15.46	1
SL5/10				15.56	1
Suspected stem-loop 6 (302–343)					
SL6/1	42	200000	5.02	0.0	162
SL6/2				4.22	537
SL6/3				4.97	27
SL6/4				5.32	12
SL6/5				5.37	119
SL6/6				5.38	193
SL6/7				6.03	15
SL6/8				6.42	47
SL6/9				6.64	26
SL6/10				6.86	148
Suspected stem-loop 7 (349–394)					
SL7/1	46	200000	3.33	0.0	256
SL7/2				1.35	292
SL7/3				12.24	726
SL7/4				14.60	15
SL7/5				17.50	46
SL7/6				20.22	2449
SL7/7				23.12	9
SL7/8				23.61	20
SL7/9				23.88	43
SL7/10				25.41	331
Frameshifting element (13459–13546)					

FSE/1	88	229610	14.00	0.0	3
FSE/2				3.23	1
FSE/3				4.45	2
FSE/4				5.44	3
FSE/5				6.22	6
FSE/6				7.81	6
FSE/7				8.06	16
FSE/8				8.73	2
FSE/9				8.88	2
FSE/10				9.65	4
Suspected frameshifting element dimer (13459–13546 x 2)					
FSE Dimer/1	176	23066	21.30	0.0	1
FSE Dimer/2				6.37	1
FSE Dimer/3				6.79	1
FSE Dimer/4				10.02	1
FSE Dimer/5				10.11	1
FSE Dimer/6				10.22	1
FSE Dimer/7				10.31	1
FSE Dimer/8				11.55	1
FSE Dimer/9				12.06	1
FSE Dimer/10				12.45	1
SARS-CoV-2 3′ UTR beginning with bulged hairpin (-361 – -1)					
3′ UTR/1	361	11430	41.16	0.0	1
3′ UTR/2				4.07	1
3′ UTR/3				26.33	1
3′ UTR/4				32.67	1
3′ UTR/5				42.10	1
3′ UTR/6				43.14	1
3′ UTR/7				46.96	1
3′ UTR/8				47.55	1
3′ UTR/9				48.04	1
3′ UTR/10				48.28	1

3´ UTR hypervariable region (-213 – -20)					
HVR/1	194	28029	29.33	0.0	1
HVR/2				2.49	1
HVR/3				3.60	1
HVR/4				4.16	1
HVR/5				5.81	1
HVR/6				6.79	1
HVR/7				7.63	1
HVR/8				7.77	1
HVR/9				9.06	1
HVR/10				11.08	1
3´ UTR pseudoknot (-329 – -210; -23 – 4)					
Pseudoknot/1	158	21512	22.10	0.0	1
Pseudoknot/2				7.46	1
Pseudoknot/3				8.50	1
Pseudoknot/4				8.75	1
Pseudoknot/5				9.98	1
Pseudoknot/6				10.14	1
Pseudoknot/7				10.31	1
Pseudoknot/8				11.81	1
Pseudoknot/9				13.78	1
Pseudoknot/10				13.99	1
3´ UTR stem-loop II-like motif, homology modeled from PDB ID: 1XJR[5] (-148 – -99)					
S2M/1	50	200000	0.21	0.0	1999777
S2M/2				11.38	23

[a]Mean pairwise all-heavy-atom RMSD between 10 lowest energy models discovered.
[b]Rosetta all-atom free energy gap of cluster's lowest energy model compared to lowest energy model discovered in run. REU = Rosetta energy units, calibrated so that 1.0 corresponds approximately to 1 k_BT.
[c]Number of models that appear in each cluster. Clustering was carried out on top 400 models ranked by Rosetta all-atom free energy, based on 5.0 Å threshold, except for small RNAs (SL1-4, SL6-7), where 2.0 Å threshold was applied.

Table 2. FARFAR2-Apo-Riboswitch models

System/cluster	Length	Models generated	Model convergence (Å)[a]	E-gap to lowest energy model (REU)[b]	Cluster occupancy[c]	RMSD to experimental structure with ligand bound
SAM-I riboswitch, RNA-Puzzle 4. PDB ID: 3V7E[29]						
SAM-I/1	126	5768	8.67	0.0	99	2.52
SAM-I/2				6.31	6	10.76
SAM-I/3				8.03	211	8.23
SAM-I/4				18.34	2806	9.68
SAM-I/5				25.90	362	9.72
SAM-I/6				26.80	22	14.83
SAM-I/7				28.76	1	17.10
SAM-I/8				29.06	282	11.90
SAM-I/9				39.12	140	8.07
SAM-I/10				39.77	588	11.33
SAM-I/IV riboswitch, RNA-Puzzle 8. PDB ID: 4L81[30]						
SAM-I/IV/1	96	33086	11.12	0.0	4	8.83
SAM-I/IV/2				1.78	7	11.18
SAM-I/IV/3				2.75	19	10.98
SAM-I/IV/4				3.65	15	5.23
SAM-I/IV/5				5.16	4	9.62
SAM-I/IV/6				5.99	2	12.25
SAM-I/IV/7				6.85	1	7.69
SAM-I/IV/8				8.78	2	7.44
SAM-I/IV/9				9.07	15	5.36
SAM-I/IV/10				9.21	1	14.13
SAM-IV riboswitch, RNA-Puzzle Unknown Rfam 15. PDB ID: 6UET[27]						
SAM-IV/1	119	10828	8.69	0.0	4	7.61
SAM-IV/2				2.16	6	3.70
SAM-IV/3				8.44	3	6.04
SAM-IV/4				10.13	8	11.37

SAM-IV/5				10.15	1	9.55
SAM-IV/6				12.88	3	7.03
SAM-IV/7				14.40	3	8.68
SAM-IV/8				14.72	1	8.51
SAM-IV/9				17.51	2	8.01
SAM-IV/10				17.94	2	6.28
Glycine riboswitch, RNA-Puzzle 3. PDB ID: 3OXE[31]						
Gly/1	84	33442	12.98	0.0	1	17.96
Gly/2				1.53	1	17.86
Gly/3				3.87	1	21.51
Gly/4				7.46	1	12.41
Gly/5				7.76	1	14.54
Gly/6				7.79	1	18.93
Gly/7				8.20	1	16.66
Gly/8				10.03	1	19.75
Gly/9				10.71	1	15.60
Gly/10				10.89	1	17.17
Cobalamin riboswitch, RNA-Puzzle 6. PDB ID: 4GXY[32]						
Cobalamin/1	158	28859	16.33	0.0	1	20.39
Cobalamin/2				0.47	1	13.08
Cobalamin/3				2.41	1	24.45
Cobalamin/4				4.14	1	19.51
Cobalamin/5				4.84	1	17.94
Cobalamin/6				6.10	1	24.51
Cobalamin/7				7.48	1	19.98
Cobalamin/8				9.07	1	25.93
Cobalamin/9				9.74	1	19.14
Cobalamin/10				11.56	1	16.42
5HT riboswitch, RNA-Puzzle 9. PDB ID: 5KPY[33]						
5HT/1	71	18660	9.10	0.0	44	4.56
5HT/2				1.40	27	10.12
5HT/3				2.43	32	6.97

5HT/4				7.14	12	6.54
5HT/5				7.91	32	5.29
5HT/6				9.40	9	6.99
5HT/7				9.82	16	9.59
5HT/8				9.87	10	9.37
5HT/9				10.00	21	6.92
5HT/10				10.44	6	10.81
ydaO riboswitch, RNA-Puzzle 12. PDB ID: 4QLM[34]						
ydaO/1	117	35506	17.16	0.0	1	23.35
ydaO/2				1.66	1	16.62
ydaO/3				4.02	1	17.26
ydaO/4				5.53	1	18.62
ydaO/5				7.20	1	14.28
ydaO/6				7.21	1	18.13
ydaO/7				7.40	1	16.90
ydaO/8				7.75	1	13.32
ydaO/9				8.13	1	16.23
ydaO/10				9.54	1	19.93
ZMP riboswitch, RNA-Puzzle 13. PDB ID: 4XW7[35]						
ZMP/1	60	20297	9.82	0.0	1	16.19
ZMP/2				1.072	1	10.97
ZMP/3				2.07	1	13.30
ZMP/4				2.79	1	16.25
ZMP/5				3.88	2	11.43
ZMP/6				4.33	3	13.83
ZMP/7				6.29	1	7.13
ZMP/8				7.26	2	14.76
ZMP/9				7.80	1	10.62
ZMP/10				8.03	1	12.43
Glutamine riboswitch (bound), RNA-Puzzle 14. PDB ID: 5DDP[36]						
Gln (bound)/1	61	24531	13.80	0.0	3	10.93
Gln (bound)/2				1.34	2	11.71

Gln (bound)/3				1.45	1	6.88
Gln (bound)/4				5.31	5	11.96
Gln (bound)/5				6.22	2	10.85
Gln (bound)/6				7.27	1	10.28
Gln (bound)/7				7.78	1	16.11
Gln (bound)/8				8.28	8	11.33
Gln (bound)/9				9.24	2	13.41
Gln (bound)/10				9.64	1	10.02
Guanidinium riboswitch, RNA-Puzzle 21. PDB ID: 5NWQ[37]						
Guanidine/1	41	48146	8.24	0.0	21	11.28
Guanidine/2				1.98	10	6.90
Guanidine/3				7.02	62	6.94
Guanidine/4				7.06	24	6.04
Guanidine/5				7.50	11	7.38
Guanidine/6				8.66	48	6.60
Guanidine/7				8.94	10	8.01
Guanidine/8				9.02	16	6.09
Guanidine/9				9.34	19	10.09
Guanidine/10				9.77	6	9.05

[a]Mean pairwise all-heavy-atom RMSD between 10 lowest energy models discovered.
[b]Rosetta all-atom energy gap of cluster's lowest energy model compared to lowest energy model discovered in run. REU = Rosetta energy units, calibrated so that 1.0 corresponds approximately to 1 k_BT.
[c]Number of models that appear in each cluster. Clustering was carried out on top 400 models ranked by Rosetta all-atom energy, based on a 5.0 Å threshold.

References

1. Rangan, R.; Zheludev, I. N.; Das, R., RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses. *bioRxiv* **2020**.
2. Chen, S.-C.; Olsthoorn, R. C. L., Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. *Virology* **2010,** *401* (1), 29-41.
3. Costales, M. G.; Childs-Disney, J. L.; Haniff, H. S.; Disney, M. D., How We Think about Targeting RNA with Small Molecules. *J. Med. Chem.* **2020**.
4. Lee, C. W.; Li, L.; Giedroc, D. P., The solution structure of coronaviral stem-loop 2 (SL2) reveals a canonical CUYG tetraloop fold. *FEBS Lett.* **2011,** *585* (7), 1049-1053.
5. Robertson, M. P.; Igel, H.; Baertsch, R.; Haussler, D.; Ares, M., Jr.; Scott, W. G., The structure of a rigorously conserved RNA element within the SARS virus genome. *PLoS Biol.* **2005,** *3* (1), e5.
6. Stelzer, A. C.; Frank, A. T.; Kratz, J. D.; Swanson, M. D.; Gonzalez-Hernandez, M. J.; Lee, J.; Andricioaei, I.; Markovitz, D. M.; Al-Hashimi, H. M., Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. *Nature Chemical Biology* **2011,** *7* (8), 553-559.
7. Liu, P.; Li, L.; Millership, J. J.; Kang, H.; Leibowitz, J. L.; Giedroc, D. P., A U-turn motif-containing stem-loop in the coronavirus 5' untranslated region plays a functional role in replication. *RNA* **2007,** *13* (5), 763-780.
8. Goebel, S. J.; Hsue, B.; Dombrowski, T. F.; Masters, P. S., Characterization of the RNA components of a putative molecular switch in the 3' untranslated region of the murine coronavirus genome. *J. Virol.* **2004,** *78* (2), 669-682.
9. Li, L.; Kang, H.; Liu, P.; Makkinje, N.; Williamson, S. T.; Leibowitz, J. L.; Giedroc, D. P., Structural lability in stem-loop 1 drives a 5' UTR-3' UTR interaction in coronavirus replication. *J. Mol. Biol.* **2008,** *377* (3), 790-803.
10. Watkins, A. M.; Das, R., FARFAR2: Improved de novo Rosetta prediction of complex global RNA folds. *bioRxiv* **2019**.
11. Miao, Z.; Adamiak, R. W.; Blanchet, M.-F.; Boniecki, M.; Bujnicki, J. M.; Chen, S.-J.; Cheng, C.; Chojnowski, G.; Chou, F.-C.; Cordero, P.; Cruz, J. A.; Ferré-D'Amaré, A. R.; Das, R.; Ding, F.; Dokholyan, N. V.; Dunin-Horkawicz, S.; Kladwang, W.; Krokhotin, A.; Lach, G.; Magnus, M.; Major, F.; Mann, T. H.; Masquida, B.; Matelska, D.; Meyer, M.; Peselis, A.; Popenda, M.; Purzycka, K. J.; Serganov, A.; Stasiewicz, J.; Szachniuk, M.; Tandon, A.; Tian, S.; Wang, J.; Xiao, Y.; Xu, X.; Zhang, J.; Zhao, P.; Zok, T.; Westhof, E., RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **2015,** *21* (6), 1066-1084.
12. Miao, Z.; Adamiak, R. W.; Antczak, M.; Batey, R. T.; Becka, A. J.; Biesiada, M.; Boniecki, M. J.; Bujnicki, J. M.; Chen, S.-J.; Cheng, C. Y.; Chou, F.-C.; Ferré-D'Amaré, A. R.; Das, R.; Dawson, W. K.; Ding, F.; Dokholyan, N. V.; Dunin-Horkawicz, S.; Geniesse, C.; Kappel, K.; Kladwang, W.; Krokhotin, A.; Łach, G. E.; Major, F.; Mann, T. H.; Magnus, M.; Pachulska-Wieczorek, K.; Patel, D. J.; Piccirilli, J. A.; Popenda, M.; Purzycka, K. J.; Ren, A.; Rice, G. M.; Santalucia, J., Jr.; Sarzynska, J.; Szachniuk, M.; Tandon, A.; Trausch, J. J.; Tian, S.; Wang, J.; Weeks, K. M.; Williams, B., 2nd; Xiao, Y.; Xu, X.; Zhang, D.; Zok, T.; Westhof, E., RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* **2017,** *23* (5), 655-672.
13. Cruz, J. A.; Blanchet, M.-F.; Boniecki, M.; Bujnicki, J. M.; Chen, S.-J.; Cao, S.; Das, R.; Ding, F.; Dokholyan, N. V.; Flores, S. C.; Huang, L.; Lavender, C. A.; Lisi, V.; Major, F.; Mikolajczak, K.; Patel, D. J.; Philips, A.; Puton, T.; Santalucia, J.; Sijenyi, F.; Hermann, T.; Rother, K.; Rother, M.; Serganov, A.; Skorupski, M.; Soltysinski, T.; Sripakdeevong, P.; Tuszynska, I.; Weeks, K. M.; Waldsich, C.; Wildauer, M.; Leontis, N. B.; Westhof, E., RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* **2012,** *18* (4), 610-625.
14. Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L., Jr.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J., The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017,** *13* (6), 3031-3048.

15. Kappel, K.; Liu, S.; Larsen, K. P.; Skiniotis, G.; Puglisi, E. V.; Puglisi, J. D.; Zhou, Z. H.; Zhao, R.; Das, R., De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes. *Nat. Methods* **2018,** *15* (11), 947-954.
16. Kappel, K.; Zhang, K.; Su, Z.; Kladwang, W.; Li, S.; Pintilie, G.; Topkar, V. V.; Rangan, R.; Zheludev, I. N.; Watkins, A. M.; Yesselman, J. D.; Chiu, W.; Das, R., Ribosolve: Rapid determination of three-dimensional RNA-only structures. *bioRxiv* **2019**.
17. Yang, D.; Leibowitz, J. L., The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res.* **2015,** *206*, 120-133.
18. Chang, R. Y.; Hofmann, M. A.; Sethna, P. B.; Brian, D. A., A cis-acting function for the coronavirus leader in defective interfering RNA replication. . *J. Virol.* **1994,** *68* (12), 8223–8231.
19. Raman, S.; Bouma, P.; Williams, G. D.; Brian, D. A., Stem-loop III in the 5' untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. *J Virol* **2003,** *77* (12), 6720-30.
20. Raman, S.; Brian, D. A., Stem-loop IV in the 5' untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. *J Virol* **2005,** *79* (19), 12434-46.
21. van den Born, E.; Posthuma, C. C.; Gultyaev, A. P.; Snijder, E. J., Discontinuous subgenomic RNA synthesis in arteriviruses is guided by an RNA hairpin structure located in the genomic leader region. *J. Virol.* **2005,** *79* (10), 6312-6324.
22. Brown, C. G.; Nixon, K. S.; Senanayake, S. D.; Brian, D. A., An RNA stem-loop within the bovine coronavirus nsp1 coding region is a cis-acting element in defective interfering RNA replication. *J Virol* **2007,** *81* (14), 7716-24.
23. Watkins, A. M.; Geniesse, C.; Kladwang, W.; Zakrevsky, P.; Jaeger, L.; Das, R., Blind prediction of noncanonical RNA structure at atomic accuracy. *Sci Adv* **2018,** *4* (5), eaar5316.
24. Reuter, J. S.; Mathews, D. H., RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **2010,** *11*, 129.
25. Kelly, J. A.; Dinman, J. D., Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS-CoV-2. *bioRxiv* **2020**.
26. Watkins, A. M.; Rangan, R.; Das, R., Using Rosetta for RNA homology modeling. *Methods Enzymol.* **2019,** *623*, 177-207.
27. Zhang, K.; Li, S.; Kappel, K.; Pintilie, G.; Su, Z.; Mou, T. C.; Schmid, M. F.; Das, R.; Chiu, W., Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 A resolution. *Nat Commun* **2019,** *10* (1), 5511.
28. Yoon, S.; Kim, J.; Hum, J.; Kim, H.; Park, S.; Kladwang, W.; Das, R., HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics* **2011,** *27* (13), 1798-805.
29. Baird, N. J.; Zhang, J.; Hamma, T.; Ferre-D'Amare, A. R., YbxF and YlxQ are bacterial homologs of L7Ae and bind K-turns but not K-loops. *RNA* **2012,** *18* (4), 759-70.
30. Trausch, J. J.; Xu, Z.; Edwards, A. L.; Reyes, F. E.; Ross, P. E.; Knight, R.; Batey, R. T., Structural basis for diversity in the SAM clan of riboswitches. *Proc Natl Acad Sci U S A* **2014,** *111* (18), 6624-9.
31. Huang, L.; Serganov, A.; Patel, D. J., Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. *Mol Cell* **2010,** *40* (5), 774-86.
32. Peselis, A.; Serganov, A., Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch. *Nat Struct Mol Biol* **2012,** *19* (11), 1182-4.
33. Porter, E. B.; Polaski, J. T.; Morck, M. M.; Batey, R. T., Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors. *Nat Chem Biol* **2017,** *13* (3), 295-301.
34. Ren, A.; Patel, D. J., c-di-AMP binds the ydaO riboswitch in two pseudo-symmetry-related pockets. *Nat Chem Biol* **2014,** *10* (9), 780-6.
35. Trausch, J. J.; Marcano-Velazquez, J. G.; Matyjasik, M. M.; Batey, R. T., Metal Ion-Mediated Nucleobase Recognition by the ZTP Riboswitch. *Chem Biol* **2015,** *22* (7), 829-37.
36. Ren, A.; Xue, Y.; Peselis, A.; Serganov, A.; Al-Hashimi, H. M.; Patel, D. J., Structural and Dynamic Basis for Low-Affinity, High-Selectivity Binding of L-Glutamine by the Glutamine Riboswitch. *Cell Rep* **2015,** *13* (9), 1800-13.

37. Huang, L.; Wang, J.; Wilson, T. J.; Lilley, D. M. J., Structure of the Guanidine III Riboswitch. *Cell Chem Biol* **2017,** *24* (11), 1407-1415 e2.