BIOCHEMISTRY

Blind prediction of noncanonical RNA structure at atomic accuracy

Andrew M. Watkins,[1] Caleb Geniesse,[1,2] Wipapat Kladwang,[1] Paul Zakrevsky,[3]* Luc Jaeger,[3] Rhiju Das[1,2,4†]

Prediction of RNA structure from nucleotide sequence remains an unsolved grand challenge of biochemistry and requires distinct concepts from protein structure prediction. Despite extensive algorithmic development in recent years, modeling of noncanonical base pairs of new RNA structural motifs has not been achieved in blind challenges. We report a stepwise Monte Carlo (SWM) method with a unique add-and-delete move set that enables predictions of noncanonical base pairs of complex RNA structures. A benchmark of 82 diverse motifs establishes the method's general ability to recover noncanonical pairs ab initio, including multistrand motifs that have been refractory to prior approaches. In a blind challenge, SWM models predicted nucleotide-resolution chemical mapping and compensatory mutagenesis experiments for three in vitro selected tetraloop/receptors with previously unsolved structures (C7.2, C7.10, and R1). As a final test, SWM blindly and correctly predicted all noncanonical pairs of a Zika virus double pseudoknot during a recent community-wide RNA-Puzzle. Stepwise structure formation, as encoded in the SWM method, enables modeling of noncanonical RNA structure in a variety of previously intractable problems.

INTRODUCTION

Significant success in protein modeling has been achieved by assuming that the native conformations of a macromolecule have the lowest free energy and that the free energy function can be approximated by a sum of hydrogen bonding, van der Waals, electrostatic, and solvation terms that extend over angstrom-scale distances. Computational methods that subject large pools of low-resolution protein models to an all-atom Monte Carlo minimization guided by these free energy functions have achieved near–atomic accuracy predictions in the CASP (Critical Assessment of Structure Prediction) community-wide blind trials (*1*). When adapted to RNA structure modeling, analogous methods have consistently achieved nucleotide resolution in the RNA-Puzzle blind trials but have not yet reached atomic accuracy, aside from previously solved motifs that happen to recur in new targets (*2*). A disappointing theme in recent RNA-Puzzle assessments is that the rate of accurate prediction of noncanonical base pairs is typically 20% or lower, even for models with correct global folds (*2*). Without recovery of these noncanonical pairs, RNA computational modeling will not be able to explain evolutionary data, predict molecular partners, or be prospectively tested by compensatory mutagenesis for the myriad biological RNAs that are being discovered at an accelerating pace.

The lag between the protein and RNA modeling fields is partly explained by differences in how protein and RNA molecules fold. Protein structures are largely defined by how α helices and β sheets pack together. As abundant data exist on these regular protein elements and their side-chain interactions, protein models with reasonable accuracy can often be assembled from fragments of previously solved structures. Less regular loops interconnecting α and β elements are less critical for defining protein folds. Those loops are typically not recovered at high accuracy, even in the most exceptional blind predictions (*3*–*5*). By contrast, predictable and the geometrically regular elements of RNA folding are Watson-Crick helices that sequester their side chains and therefore cannot be positioned by direct side-chain interactions. Instead, the RNA loops interconnecting those helices form intricate noncanonical base pairs that define an RNA's global helix arrangement. The RNA structure prediction problem, more so than the protein problem, depends on discovering these irregular loop conformations and their associated noncanonical base pairs ab initio. Unfortunately, discovering the lowest free energy conformations of new noncanonical loop motifs has not generally been tractable because of the vast number of deep, local minima in the all-atom folding free energy landscape of even the smallest such motifs. Essentially all three-dimensional (3D) RNA modeling methods, including MC-Sym/MC-Fold, Rosetta FARFAR, iFoldRNA, SimRNA, and Vfold3D, use coarse-grained modeling stages that allow for smoother conformational search but generally return conformations too inaccurate to be refined to atomic accuracy by Monte Carlo minimization or molecular dynamics refinement (*6*–*10*).

To address this challenge, we have developed Rosetta methods that attempt to remove barriers in conformational search through the addition of residues one at a time rather than through low-resolution coarse-graining or through small perturbations to fully built conformations. We previously described how step-by-step buildup of an RNA structure, enforcing low-energy conformations for each added nucleotide, could lead to atomic accuracy models of irregular single-stranded RNA loops (*11*). The calculation, instantiated in the Rosetta modeling framework, involved a deterministic enumeration over buildup paths, analogous to classic dynamic programming methods developed for canonical RNA secondary structure prediction (*11*, *12*). This enumerative stepwise assembly (SWA) method guaranteed a unique solution for the final conformational ensemble but necessitated large expenditures of computational power. For example, calculations for even small loops of 5 to 7 nucleotides (nt) required tens of thousands of CPU (central processing unit) hours (*11*); junctions involving multiple interacting strands would further increase computational cost to many millions of CPU hours, which is currently prohibitive.

In the hope of reducing this computational expense, we hypothesized that the stepwise addition moves developed for SWA might still

[1] Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA. [2] Biophysics Program, Stanford University, Stanford, CA 94305, USA. [3] Department of Chemistry and Biochemistry, Biomolecular Science and Engineering Program, University of California at Santa Barbara, Santa Barbara, CA 93106, USA. [4] Department of Physics, Stanford University, Stanford, CA 94305, USA.
*Present address: RNA Biology Laboratory, Center for Cancer Research, National Cancer Institute, Frederick, MD 21702, USA.
†Corresponding author. Email: rhiju@stanford.edu

be effective at producing high-accuracy models if implemented as part of a stochastic sampling scheme rather than deterministic enumeration. To test this hypothesis, we have developed stepwise Monte Carlo (SWM), a Monte Carlo optimization method whose primary moves are the stepwise addition moves of SWA. Here, we report that SWM enables significant increases in the computational speed of ab initio structure prediction and describe applications of SWM to previously intractable noncanonical RNA structures. Tests of SWM include stringent blind evaluation through prospective experimental tests and an RNA-Puzzle community-wide structure prediction challenge.

RESULTS

Efficient implementation of SWM

Figure 1 illustrates the SWM procedure, which has been implemented in the Rosetta framework (*13*) and is also freely available through an online ROSIE (Rosetta Online Server that Includes Everyone) server (see Materials and Methods) (*14*). In realistic 3D RNA modeling problems, RNA helices are typically known a priori from secondary structure prediction methods. The main goal is therefore to infer lowest free energy conformations of loops that connect these helices, such as the four nucleotides GCAA closing a hairpin (Fig. 1A) or two strands, each with a single guanosine, in the GG mismatch motif (Fig. 1B). Our previous work (*11*, *15*, *16*) introduced stepwise addition moves that allow building of these nucleotides one at a time, starting from conformations with helices only (Start; Fig. 1, A and B). Conceptually, each addition was proposed to simulate the stepwise formation of well-defined structure from "random coil"–like ensembles (dotted lines; Fig. 1, A and B).

Here, we stochastically carry out these addition moves, choosing random positions on which to prepend or append new nucleotides (Add; Fig. 1, A and B), rather than enumerating these additions at all possible positions [as was implemented previously (*11*, *15*)]. These stochastic moves are accepted if they lower the computed free energy

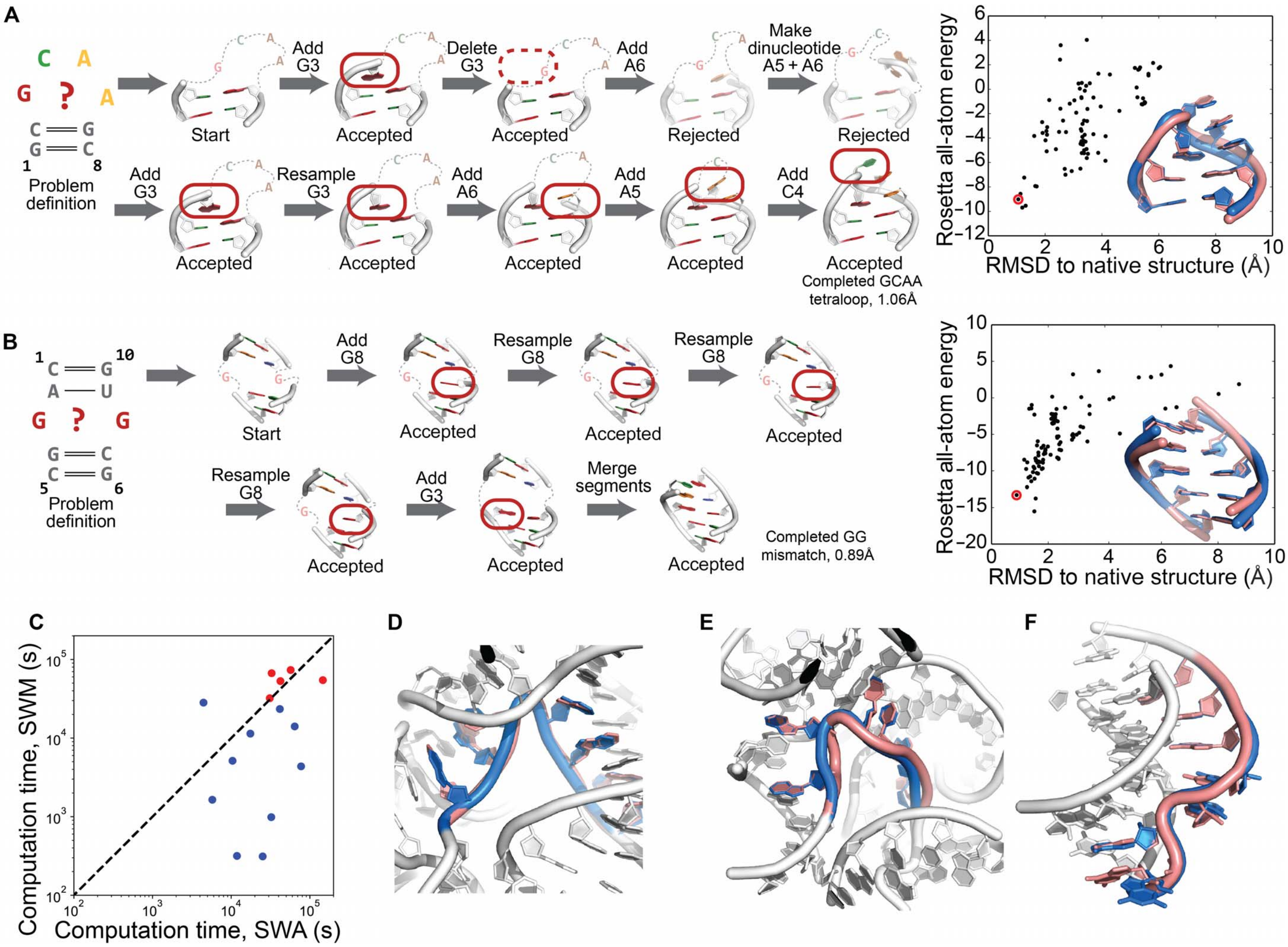


Fig. 1. SWM efficiently searches the complex energy landscapes of noncanonical RNA loops. (**A** and **B**) SWM trajectories solve a GCAA tetraloop [Protein Data Bank (PDB) ID: 1ZIH) (A) and a two-strand GG-mismatch two-way junction (1F5G) (B) in 10 moves or less (left). Final structures achieve low free energies and sub-angstrom RMSD accuracies; numerous such structures appear in simulations involving 100 models (right-hand panels). (**C**) Significantly reduced CPU time is required for convergence of SWM compared to enumeration by SWA (*11*), except for loops drawn from the 23*S* ribosomal RNA (rRNA) (red). (**D** to **F**) SWM models for J2/3 (D) from group II intron (3G78), modeled with the energy function previously used for SWA, and 23*S* rRNA loop (1S72) (E) and L2 loop (F) of viral pseudoknot (1L2X), both modeled with updated Rosetta free energy function, illustrate sub-angstrom recovery of irregular single-stranded loops excised from crystal structures.

of the model or if the free energy increment is lower than a thermal fluctuation energy as set by the Metropolis criterion (*17*). To maintain detailed balance in the Monte Carlo scheme, the moves intermix these additions with deletions of single residues, again chosen randomly and accepted on the basis of the Metropolis criterion (Delete, Fig. 1A). These deletion moves simulate the transient unstructuring of nucleotides at the edges of loops. To allow buildup of multistrand motifs, we developed moves to merge and split independent regions of RNA, such as regions associated with the different helices of multihelix junctions (Merge, Fig. 1B). Last, we allowed resampling of randomly chosen internal degrees of freedom, maintaining chain closure with robotics-inspired kinematic closure algorithms (Resample, Fig. 1A) (*18*, *19*). These resampling moves could not be incorporated into the previous enumerative SWA because of the large number of increased modeling pathways that would need to be enumerated.

Before testing SWM, it was unclear whether a stochastic search might allow for efficient ab initio recovery of RNA loop conformations. In our previous work on enumerative SWA, we posited that "an inability to guarantee exhaustive conformational sampling has precluded the consistent prediction of biomolecular structure at high resolution" (*11*). Nevertheless, in our test cases, SWM did achieve efficient search over the free energy landscape, despite the lack of a guarantee of exhaustive conformational sampling. Figure 1 (A and B) illustrates the recovery of the sub-angstrom accuracy conformations of the GCAA tetraloop and GG mismatch motifs using less than 3 hours of computation on a single modern laptop computer to create 100 models. Furthermore, these runs were "convergent": Different simulations independently achieved the same low-energy configurations repeatedly (numerous models within 2 Å of experimental structures; right panels in Fig. 1, A and B), suggesting highly efficient sampling. For comparison, modeling of these small loops by SWA enumeration required use of many thousands of CPU hours due to the requirement of enumerating multiple loop conformations over multiple buildup pathways.

Ab initio recovery of complex single-stranded RNA loops

After preliminary tests on simple loops, we carried out SWM modeling on a set of 15 single-stranded RNA loops excised from crystal structures, previously used to benchmark the enumerative SWA method (*11*). These loops were specifically chosen because of their irregularity; they each harbor non–A-form backbone conformations, form noncanonical pairs with surrounding residues, and span different helices in functional RNAs. We confirmed that, for nearly all these trans-helix cases, SWM produces conformations that give computed free energies and accuracies as low as those achieved by SWA [median energy gap and root mean squared deviation (RMSD) values; table S1 and Fig. 1, C and D]. In many cases, the computational cost for achieving convergent and accurate modeling was reduced by up to two orders of magnitude (see Supplementary Methods and Fig. 1C). Exceptions to this speed increase were loops that needed to be rebuilt into the 23S ribosomal RNA (rRNA) (red points, Fig. 1C), which featured a particularly large number of surrounding nucleotides, and concomitantly many viable interacting conformations. This observation suggested that SWM would be particularly efficient for ab initio modeling of motifs that primarily form noncanonical interactions within the motif, as is typically the case for RNA junctions and tertiary contacts (see below), but not for the longest ribosomal loops. Furthermore, through its increased speed, SWM allowed us to confirm that recent updates to the Rosetta free energy function (*20*) and estimation of conformational entropy of unstructured segments generally improved modeling accuracy for single-stranded RNA loops (tables S1 and S2). The improvements included rescue of some 23S rRNA loops (Fig. 1E) and solution of a loop from a beet western yellow virus frameshifting pseudoknot that was previously not solvable by SWA (Fig. 1F and table S3) (*11*). Supplementary Text provides a more detailed description of energy function updates and results on this trans-helix loop benchmark.

Ab initio recovery across complex noncanonical motifs

To more broadly evaluate SWM, we expanded the 15 single-stranded loop benchmark to a larger set of 82 complex, multistranded RNA motifs that we encountered in previous RNA-Puzzles and other modeling challenges (table S3 and fig. S1). Because of the efficiency of SWM modeling, we could test a benchmark that was nearly three times larger than our most extensive previous efforts (*8*). Over the entire benchmark, SWM achieved a mean and median RMSD accuracy (over the top five cluster centers) of 2.15 and 1.49 Å and mean and median recovery of non–Watson-Crick pairs of 76 and 96%, respectively (Table 1, table S3, and fig. S2). We observed numerous cases in which the SWM model and experimental structure were nearly indistinguishable by eye (Fig. 2). Examples included two-stranded motifs that required orders of magnitude higher computational expense with the prior enumerative SWA method (*16*), such as the most conserved domain of the signal recognition particle (1.26 Å RMSD, five of five noncanonical pairs recovered; Fig. 2A) and the first RNA-Puzzle challenge, a human thymidylate synthetase mRNA segment (0.96 Å, one noncanonical pair and one extrahelical bulge recovered; Fig. 2B) (*2*). For several test cases, there was experimental evidence that formation of stereotyped atomic structures required flanking helices to be positioned by the broader tertiary context. If the immediately flanking helix context was provided, the median RMSD accuracy and non–Watson-Crick base pair recovery in these cases were excellent (1.19 Å and 100%; Table 1, fig. S2, and table S3), as illustrated by the J5/5a hinge from the P4-P6 domain of the *Tetrahymena* group I intron (0.55 Å RMSD, all four noncanonical pairs and all three extrahelical bulges recovered; Fig. 2C) (*21*).

Perhaps the most striking models were recovered for multi-helix junctions and tertiary contacts, which have largely eluded RNA modeling efforts seeking high resolution (*6*, *8*). SWM achieves high accuracy models for the P2-P3-P6 three-way junction from the Varkud satellite ribozyme, previously missed by all modelers in the RNA-Puzzle 7 challenge (1.13 Å RMSD, three of three noncanonical pairs recovered; Fig. 2D); a highly irregular tertiary contact in a hammerhead ribozyme (1.16 Å RMSD, two of three noncanonical pairs and one extrahelical bulge recovered; Fig. 2E); a complex between a GAAA tetraloop and its 11-nt receptor (0.64 Å RMSD, all four noncanonical pairs recovered when flanking helix context was provided; Fig. 2F); and the tRNA^{phe} T-loop, a loop-loop tertiary contact stabilized by chemical modifications at 5-methyluridine, pseudouridine, and N1-methyladenosine (1.33 Å accuracy when flanking context was provided; Fig. 2G). Motifs without any flanking A-form helices offered particularly stringent tests for ab initio modeling and could also be recovered at high accuracy by SWM, as illustrated by the inosine tetrad–containing quadruplex (2.87 Å RMSD overall, 0.46 Å RMSD if the terminal uracils, which make crystal contacts, are excluded; Fig. 2H). For comparison, we also carried out modeling with Fragment Assembly of RNA with Full Atom Refinement (FARFAR) on these 82 motifs, taking care to remove possible homologs from the method's fragment library to mimic a realistic ab initio prediction scenario (we could not carry out fair comparisons to other

Table 1. Benchmark of SWM compared to previous Rosetta FARFAR over different classes of RNA structure motifs.

Category	Motif properties			Best of five cluster centers			
				RMSD (Å)*		F_{NWC} †,*	
	No. of motifs	Length*	Strands*	SWM	FARFAR	SWM	FARFAR
Single helix or multiple helices with crystallographic context provided							
Trans-helix loop	15	6	1	0.83	3.29	1.00	0.77
Apical loop	4	4.5	1	1.14	2.96	1.00	1.00
Two-way junction	14	7.5	2	0.74	1.15	1.00	1.00
Multi-helix junction	5	11	3	1.91	1.93	0.80	0.33
Tertiary contact	10	8.5	2	1.25	1.78	0.83	0.50
Multiple helices without crystallographic context provided							
Two-way junction	15	7	2	1.59	1.40	1.00	0.55
Multi-helix junction	5	10	3	2.60	3.45	0.40	0.20
Tertiary contact	8	8.5	2	2.89	2.13	0.36	0.20
Non-helix embedded	5	10	4	2.81	4.30	0.80	0.71
Overall	82	7	2	1.49	1.93	0.96	0.67

*Median values reported. Mean values given in tables S3 and S4. †Fraction of non–Watson-Crick pairs from experimental structure observed in computational model.

methods due to the unavailability of similar homolog exclusion options in those methods; Supplementary Methods). SWM strongly outperformed these FARFAR models in terms of recovery of noncanonical pairs ($P < 5 \times 10^{-5}$) and RMSD accuracy ($P < 2 \times 10^{-4}$) (P values are based on Wilcoxon ranked-pairs test, $n = 82$; fig. S3, Table 1, and tables S3 and S4).

In some benchmark cases, SWM did not exhibit near–atomic accuracy recovery and illuminated challenges remaining for computational RNA modeling. While a few discrepancies between SWM models and x-ray structures could be explained by crystallographic interactions (for example, edge nucleotides making crystal contacts; Fig. 3H), most problems were better explained by errors in the energy function. For 9 of the 14 cases in which the SWM modeling RMSD was worse than 3.0 Å (and thus definitively not achieving atomic accuracy), the energy of the lowest free energy SWM model was lower than that of the optimized experimental structure, often by several units of free energy [calibrated here to correspond to $k_{\mathrm{B}}T$ (*20*); table S3]. One clue for the source of this issue came from cases where the fragment-based method (FARFAR) outperformed SWM if assessed by RMSD but not by the fraction of base pairs recovered (Table 1). The existence of these FARFAR models with native-like backbones but incorrect base pairs suggested that conformational preferences implicitly encoded in database fragments in FARFAR might need to be better captured during SWM. One possible route to improving SWM might be to update the RNA torsional potential of the Rosetta free energy function, which currently does not model correlations across most backbone torsions. Results on the hepatitis C virus internal ribosome entry site, the sarcin-ricin loop, and other test cases suggest that a modified torsional potential, as well as inclusion of metal ions, may eventually address these residual problems (fig. S4).

Stringent tests of SWM models for new RNA-RNA tertiary contact motifs

As we began to see significant improvements of modeling accuracy in the 82-motif benchmark, we hypothesized that SWM might be able to predict noncanonical base pairs in motifs that have been refractory to nuclear magnetic resonance and crystallographic analysis. Success in the 11-nt tetraloop/receptor benchmark test case (Fig. 2F), a classic model system and ubiquitous tertiary contact in natural RNAs, encouraged us to model alternative tetraloop/receptor complexes selected for use in RNA engineering. We applied SWM to these complexes, whose structures have not yet been solved experimentally (*2*, *22*), and we designed stringent experimental tests to validate or falsify these models.

A detailed sequence/function analysis previously suggested similarities between the GAAA/C7.2, GAAA/C7.10, and GGAA/R(1) interactions discovered through in vitro evolution (*22*) and the classic GAAA/11-nt receptor, which has been crystallized in numerous contexts. It has not been clear, however, whether this similarity holds at the structural level due to the unavailability of high-resolution structures of the three artificial tetraloop/receptors. For example, prior literature analyses conflicted in the proposal of which C7.2 receptor nucleotides, if any, might form a "platform" (lime, Fig. 3A), analogous to A4-A5 platform in the 11-nt receptor [G4 and A6 with an intervening bulge in the studies of Sripakdeevong *et al.* (*11*) and Costa and Michel (*23*), and G4 and U5 in the study of Geary *et al.* (*22*)]. Similarly, a proposed homology of C9 in R(1) to A8 in the 11-nt receptor (*22*) has not been tested by structure modeling or experiments.

We carried out SWM modeling to explore possible structural homologies of these four receptors. In the SWM runs, the stem and basal G-A sugar-Hoogsteen pair of the GNRA tetraloop and their docking site into the GG/CC stem of the receptor were seeded on

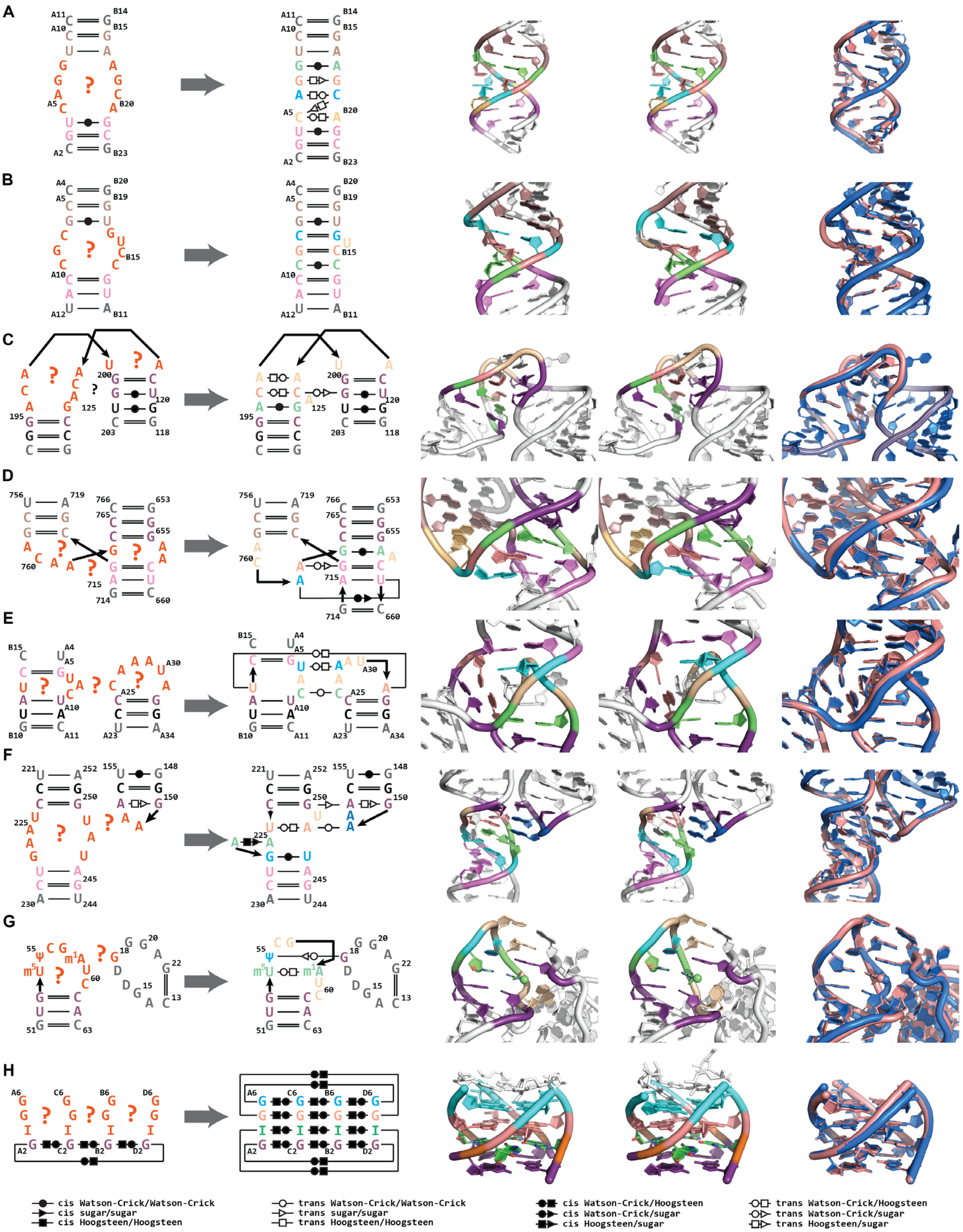

A
B
C
D
E
F
G
H
●— cis Watson-Crick/Watson-Crick
▶— cis sugar/sugar
■— cis Hoogsteen/Hoogsteen
○— trans Watson-Crick/Watson-Crick
▷— trans sugar/sugar
□— trans Hoogsteen/Hoogsteen
●■ cis Watson-Crick/Hoogsteen
●▶ cis Watson-Crick/sugar
■▶ cis Hoogsteen/sugar
○□ trans Watson-Crick/Hoogsteen
○▷ trans Watson-Crick/sugar
□▷ trans Hoogsteen/sugar

Fig. 2. SWM recovers noncanonical base pairs ab initio for complex RNA motifs. From left to right in each panel: 2D diagram with problem definition, 2D diagram with experimental noncanonical base pairs, experimental 3D model, SWM 3D model, and 3D overlay (experimental, marine; SWM model, salmon). (**A** to **H**) Motifs are (A) most conserved domain of human signal-recognition particle (PDB ID: 1LNT); (B) noncanonical junction from human thymidylate synthase regulatory motif, RNA-Puzzle 1 (PDB ID: 3MEI); (C) irregular J5/5a hinge from the P4-P6 domain of the *Tetrahymena* group I self-splicing intron (PDB ID: 2R8S); (D) P2-P3-P6 three-way A-minor junction from the Varkud satellite nucleolytic ribozyme, RNA-Puzzle 7 (PDB ID: 4R4V); (E) tertiary contact stabilizing the *Schistosoma* hammerhead nucleolytic ribozyme (PDB ID: 2OEU); (F) tetraloop/receptor tertiary contact from the P4-P6 domain of the *Tetrahymena* group I self-splicing intron (PDB ID: 2R8S); (G) T-loop/purine interaction from yeast tRNA^{phe} involving three chemically modified nucleotides (PDB ID: 1EHZ); and (H) RNA quadruplex including an inosine tetrad (PDB ID: 2GRB). Colors indicate accurately recovered noncanonical features (pastel colors), accurately recovered extrahelical bulges (wheat with white side chains), flanking helices built de novo (violet), parts of experimental structure used for modeling but allowed to minimize (dark violet), fixed context from experimental structure (black in 2D and white in 3D), and additional helical context not included in modeling (gray in 2D and white in 3D).

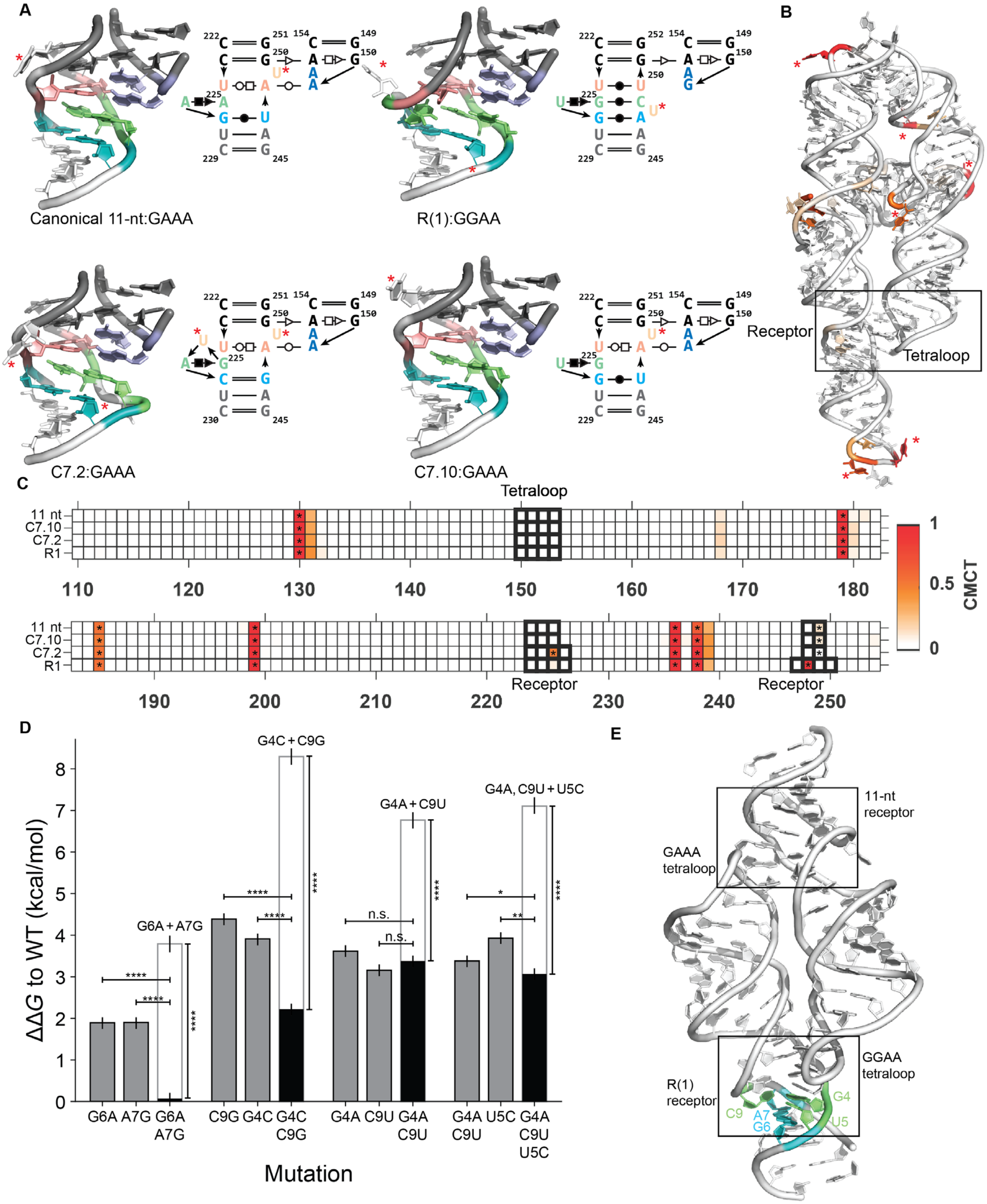


Fig. 3. SWM modeling and prospective experimental tests of previously unsolved tetraloop/receptor motifs. (**A**) Ab initio SWM models for canonical 11-nt tetraloop/receptor motif and alternative motifs discovered through in vitro selection that have resisted crystallization. Lavender, salmon, lime, and teal colorings highlight homologous structural features. During modeling, the bottom flanking helix (white) was allowed to move relative to the top helices of the receptor and tetraloop (gray), which were held fixed. (**B**) Canonical 11-nt tetraloop receptor module from the P4-P6 domain of the *Tetrahymena* group I self-splicing intron (PDB ID: 2R8S). In (A) and (B), red asterisks mark uracil residues predicted to be bulged. (**C**) CMCT mapping of the receptors installed into the P4-P6 domain of the *Tetrahymena* ribozyme (tetraloop and receptor indicated by black boxes) supports the bulged uracils in the predicted models (black asterisks). (**D**) Selective tests of each R(1) receptor base pair by compensatory mutagenesis in tectoRNA dimer. Rescue by double and triple mutants (black bars) was compared to energetic perturbations predicted based on the sum of effects (white bars) of component mutations or, more conservatively, to the single mutants. *$P < 0.05$, **$P < 0.001$, and ****$P < 1 \times 10^{-6}$ (computed by Student's *t* test for difference of means); n.s., not significant. (**E**) Overall 3D model of tectoRNA dimer with SWM model for R(1) receptor. WT, wild-type.

the basis of one crystal structure of the GAAA/11-nt receptor [see considerations of A-minor geometries discussed by Geary *et al.* (*22*)]. The remaining 10 nucleotides and receptor stem were modeled ab initio. SWM modeling of all four of these receptors achieved convergence, with 8 of the top 10 models clustered within 2 Å RMSD of each other. The modeling recovered the known GAAA/11-nt structure at 0.80 Å RMSD and reproduced a previous C7.2 model that involved SWA enumeration of only 3 nucleotides (G4-U5-A6) rather than rebuilding ab initio the complete tetraloop/receptor interaction.

SWM models for all four tetraloop/receptors exhibited not only striking structural homology to each other but also noncanonical features (extrahelical bulges and pairs) that were not anticipated from prior manual modeling efforts (see Fig. 3A, Supplementary Text, and models provided in the Supplementary Materials). Three features were preserved across loops. First, models for all receptors exhibited a docking site for the second nucleotide of the tetraloop (salmon, Fig. 3A). In GAAA/11-nt, GAAA/C7.2, and GAAA/C7.10, where the second tetraloop nucleotide is A, the receptor docking site was predicted to be an adenosine that is part of a Watson-Crick/Hoogsteen U-A pair. In GGAA/R(1), where the second tetraloop nucleotide is G, the receptor docking site was predicted to be U3, part of a noncanonical Watson-Crick U3-U10 pair. Second, SWM models for all receptors exhibited a platform involving two same-strand base-paired nucleotides that stack under the tetraloop (lime, Fig. 3A). The sequence varies, however, between A-A in the 11-nt receptor, G-U in the C7.10 receptor, G-A in the C7.2 receptor [supporting the models by Sripakdeevong *et al.* (*11*) and Costa and Michel (*23*), but not the model in Geary *et al.* (*22*)], and G-U in the R(1) receptor. In the R(1) receptor, C9 was predicted by SWM to form a stabilizing C-G base pair with a platform nucleotide. While C9 was previously proposed to be homologous to A8 in the 11-nt receptor based on mutagenesis data indicating its importance (*22*), the new model also explains prior binding data that implicated C9 as forming core interactions in R(1). Third, all receptors show a noncanonical pair involving Watson-Crick edges, needed to transition between the platform region and the lower stem of the receptor (teal, Fig. 3A). The sequence is a G-A pair in R(1) and a G·U wobble in the others. Overall, given the sequence mapping between receptors revealed by the SWM models, each noncanonical pairing in the naturally occurring GAAA/11-nt structure had a homolog, albeit one that was difficult to predict (and, in some cases, differently predicted) in each of the three non-native tetraloop receptors.

We tested these features using prospective experiments. The SWM models predicted different single uridines to be bulged out of each tetraloop receptor. Reaction to CMCT [*N*-cyclohexyl-*N′*-(2-morpholinoethyl)carbodiimide tosylate], followed by reverse transcription, allows single–nucleotide resolution mapping of unpaired uridines that bulge out of structure and expose their Watson-Crick edges to solution. We therefore installed the tetraloop/receptors into the P4-P6 domain of the *Tetrahymena* ribozyme (Fig. 3B), which also displays other bulged uridines that served as positive controls (asterisks in Fig. 3C). These experiments verified extrahelical bulging of single-nucleotide uridines predicted by SWM at different positions in the different receptors, and disfavored prior manual models (see Fig. 3C and Supplementary Text).

We carried out further prospective experiments to incisively test base pairs newly predicted by SWM modeling. In particular, the R(1) receptor model included numerous unexpected noncanonical features, especially a base triple involving a new Watson-Crick singlet base pair G4-C9 and a dinucleotide platform at G4-U5. These features were stringently evaluated via compensatory mutagenesis. Chemical mapping on the P4-P6 domain confirmed the G4-C9 base pair but was not sensitive enough to test other compensatory mutants (fig. S5). We therefore carried out native gel assembly measurements in a different system, the tectoRNA dimer, which enables precise energetic measurements spanning 5 kcal/mol (Fig. 3, D and E). Observation of energetic disruption by individual mutations and then rescue by compensatory mutants confirmed the predicted interactions of G4-C9, the base triple G4-U5-C9, and noncanonical pair G6-A7 ($P < 0.01$ in all cases; Fig. 3D), as well as other features of the model (see Supplementary Text and fig. S6). Overall, these experimental results falsified bulge predictions and base pairings previously guessed for these tetraloop/receptors (*11*, *22*, *23*) and strongly supported the models predicted by SWM. Our structural inference and mutagenesis-based validation of noncanonical pairs would have been intractable without the SWM-predicted models because of the large number of possible mutant pair and triple combinations that would have to be tested.

Blind prediction of all noncanonical pairs of a community-wide RNA-Puzzle

The community-wide modeling challenge RNA-Puzzle 18 provided an opportunity to further blindly test SWM and to compare it to best efforts of other state-of-the-art algorithms (Fig. 4). This problem was of mixed difficulty. On the one hand, the 71-nt target sequence was readily identified via PDB-BLAST (Basic Local Alignment Search Tool) (*24*) to be a Zika virus RNA homologous to a molecule with a previously solved x-ray structure, an Xrn1 endonuclease-resistant (xrRNA) fragment of Murray Valley Encephalitis virus (PDB ID: 4PQV) (*25*). However, the crystallographic environment of the prior structure disrupted a pseudoknot (between L3 and J1/4; Fig. 4A) expected from sequence alignments so that nearly half of the prior structure could not be trusted as a template for homology modeling. Intermolecular crystal contacts produced an open single-stranded region in the asymmetric unit where the pseudoknot was expected and interleaved regions from separate molecules; the scale of these conformational perturbations was as large as the dimensions of the molecule itself (fig. S7). Further complicating the modeling, two Watson-Crick pairs within stem P3 changed to or from G·U wobble pairs. Moreover, previous literature analysis (*25*) suggested extension of this helix by two further Watson-Crick pairs (U29-A37 and U30-A36), albeit without direct evidence from phylogenetic covariation and in partial conflict with dimethyl sulfate (DMS) probing. Ab initio modeling at a scale inaccessible to the prior enumerative SWA method was necessary for modeling the RNA, and we therefore carried out SWM (see Fig. 4B and Materials and Methods).

The lowest free energy SWM models for RNA-Puzzle 18 converged to a tight ensemble of intricate structures, with one submitted SWM model shown in Fig. 4C. The Watson-Crick pairs U29-A37 and U30-A36 predicted in the literature did not occur in the models. Instead, several other features were consistently observed across the SWM models [colored in Fig. 4, A (right) and C]: coaxial arrangement of the pseudoknot helix (purple) on P3 (light violet); a noncanonical trans Watson-Crick base pair between A37 and U51 stacking under P1 (green); a UA-handle (*26*) formed by U29-A36 (turquoise); and lack of pairing by U30, A35, A52, and A53 (sand and orange). These features were not uniformly present—or not predicted at all—in models created by FARFAR or, as it later turned out, in models submitted by other RNA-Puzzle participants (fig. S8).

The subsequent release of the crystal structure (Fig. 4, D and E) (*27*) confirmed all base pairs predicted by SWM modeling (100%

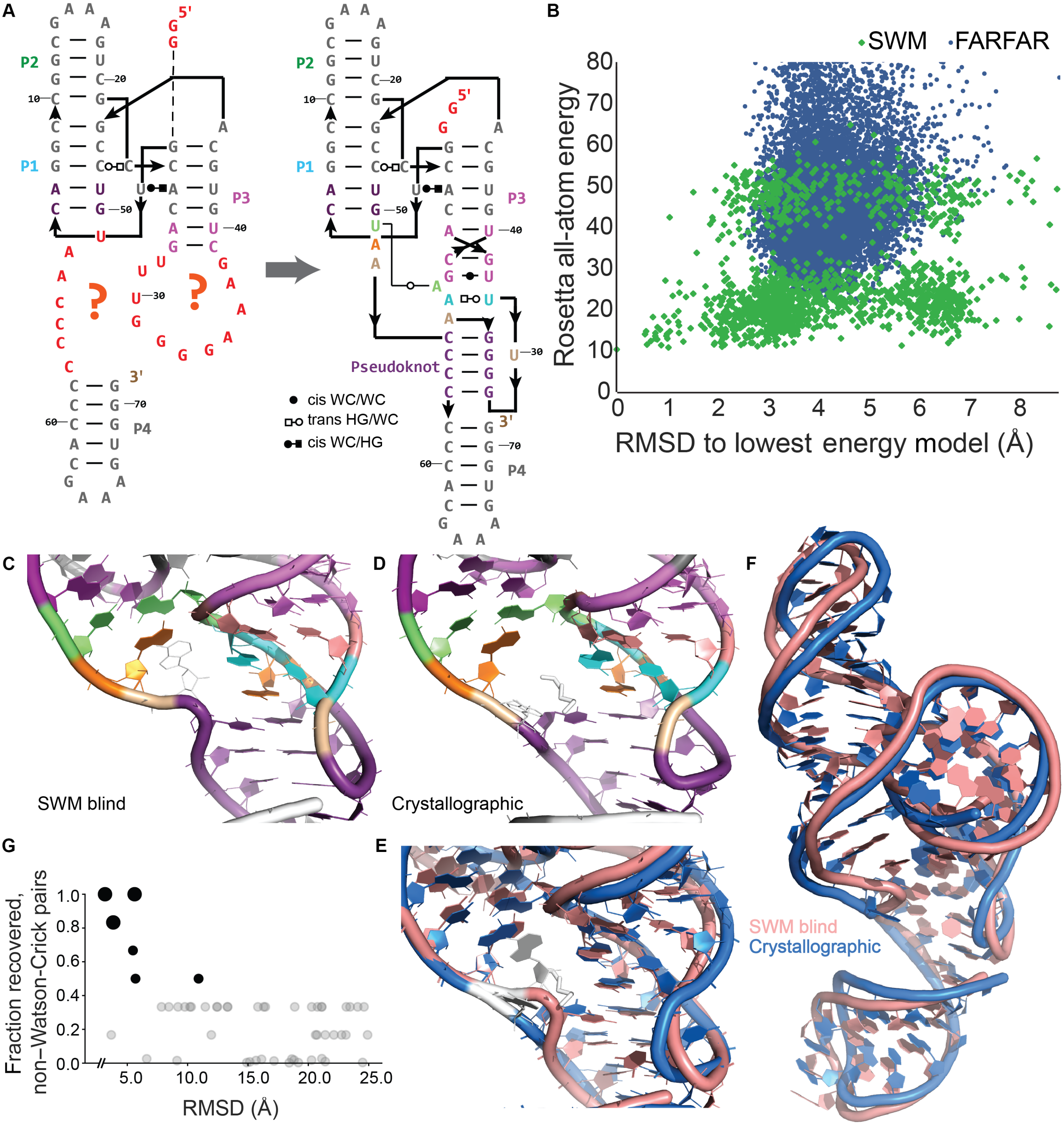


Fig. 4. Blind prediction of a complex RNA tertiary fold during RNA-Puzzle 18. (**A**) Two-dimensional diagram of the RNA-Puzzle 18 (Zika xrRNA) modeling problem, highlighting motifs that needed to be built de novo in red (left) and SWM-predicted pairings (pastel colors; right). WC, Watson-Crick; HG, Hoogsteen. (**B**) Structures discovered by SWM (green) are lower in energy and ~4 Å from models from conventional fragment assembly (FARFAR; blue); note that *x* axis is RMSD to the lowest free energy SWM model, not the experimental structure (unavailable at the time of modeling). (**C** and **D**) Magnified view of noncanonical region built de novo for SWM model submitted for RNA-Puzzle competition (C) and the subsequently released crystal structure (D). (**E**) and (**F**) give overlays in magnified and global views, respectively (SWM, salmon; crystal, marine). (**G**) Fraction of noncanonical base pairs recovered and RMSD to native model obtained by Rosetta modeling (black; larger and smaller symbols are SWM and FARFAR, respectively) and other laboratories (gray) for RNA-Puzzle 18. Points recovering zero noncanonical pairs are given a small vertical perturbation to appear visually distinct.

non–Watson-Crick recovery). The only structural deviation involved A53 (sand, Fig. 4C), which was predicted in SWM models to be unpaired and stacked on neighbor A52 (orange, Fig. 4C). In the crystal, A53 was unpaired but bulged out of the core to form a contact with a crystallographic neighbor, while a 1,6-hexanediol molecule from the crystallization buffer took its place (white sticks, Fig. 4C); this arrangement was noted independently to be a likely crystallographic artifact (*27*). There is striking overall fold agreement (3.08 Å RMSD; and 1.90 Å over just the most difficult noncanonical region, nucleotides 5 and 6, 26 to 40, 49 to 59, and 70 and 71; Fig. 4, C and D), much better than the ~10 Å best-case agreement seen in previous RNA-Puzzles of comparable difficulty (*2*). Furthermore, SWM accurately predicted all noncanonical base pairs ($F_{\mathrm{NWC}} = 1$; Fig. 4G). While one blind model from another method achieved somewhat comparable RMSD to the crystal structure (3.61 Å), it predicted only one of six non–Watson-Crick base pairs (Fig. 4G) and left a "hole" in the central noncanonical region (RMSD of 3.67 Å in that region; fig. S8).

DISCUSSION

We have developed an algorithm for modeling RNA structures called stepwise Monte Carlo (SWM), which uniquely allows for the addition and deletion of residues during modeling guided by the Rosetta all-atom free energy function. The minima of the energy landscape are efficiently traversed by this method, allowing the ab initio recovery of small RNA loop structures in hours of CPU time (Fig. 1). On an extensive benchmark, SWM enables quantitative recovery of noncanonical pairs in cases that include prior RNA-Puzzle motifs, junctions and tertiary contacts involving numerous strands, and motifs without any A-form helices (Fig. 2 and Table 1). We applied SWM to model structures of three previously unsolved tetraloop/receptors and prospectively validated these models through chemical mapping and extensive compensatory mutagenesis (Fig. 3). Last, SWM achieved blind prediction of all noncanonical pairs of a recent RNA-Puzzle, an intricately folded domain of the Zika RNA genome whose pairings were missed by other methods applied by our group and by other modeling groups (Fig. 4). The most striking aspect of the SWM models is the high recovery of noncanonical pairs, which have largely eluded previous algorithms when tested in blind challenges. These results support stepwise nucleotide structure formation as a predictive algorithmic principle for high-resolution RNA structure modeling. We expect SWM to be useful in the ab initio modeling and, if extended to sequence optimization, the discovery of novel motifs for RNA architectonic design (*28*, *29*).

The results above focused on solving individual noncanonical motifs. While these problems arise frequently in real-world modeling (for example, the unsolved tetraloop receptors), most functional RNA structures harbor multiple junctions and tertiary contacts whose folds become dependent on each other through the lever arm–like effects of interconnecting helices. SWM is currently too computationally expensive to simultaneously simulate all motifs and helices in these molecules. It may be necessary to better parallelize the current algorithm to allow concomitant modeling of multiple motifs on multiprocessor computers, as is routine in molecular dynamics simulations (*30*). Alternatively, modeling may benefit from iterating back and forth between high-resolution SWM and complementary low-resolution modeling approaches like MC-Sym/MC-Fold, Rosetta FARFAR, iFoldRNA, SimRNA, and Vfold3D (*6*–*10*), similar to iterative approaches in modeling large proteins (*5*). In addition, we note that SWM relies heavily on the assumed free energy function for folding, and several of our benchmark cases indicate that even the most recently updated Rosetta free energy function is still not accurate when SWM enables deep sampling. Therefore, a critical open question is whether residual free energy function problems might be corrected by improved RNA torsional potentials, treatment of electrostatic effects, or use of energy functions independently developed for biomolecular mechanics and refinement (*5*, *30*, *31*).

MATERIALS AND METHODS

Stepwise Monte Carlo

SWM was implemented in C++ in the Rosetta codebase. The source code and the stepwise executable compiled for different environments are being made available in Rosetta release 3.6 and later releases, free to academic users at www.rosettacommons.org. Full documentation, including example command lines, tutorial videos, and demonstration code, is available at www.rosettacommons.org/docs/latest/application_documentation/stepwise/stepwise_monte_carlo/stepwise.

The full set of benchmark cases, including the 82 central to this work, is available at https://github.com/DasLab/rna_benchmark. The repository contains input files for each benchmark case; scripts for setting up benchmark runs using either SWM or fragment assembly, including automated job submission for multiple cluster job schedulers; and scripts for creating analysis figures and tables. Finally, SWM is available through a web server on ROSIE at http://rosie.rosettacommons.org/stepwise. Supplementary Methods gives detailed descriptions of SWM, SWA, and fragment assembly of RNA with FARFAR modeling, evaluation of RMSD and energetic sampling efficiency, and PDB accession IDs for experimental structures.

Chemical mapping

Chemical mapping was carried out as in the study of Kladwang *et al.* (*32*). Briefly, DNA templates for the P4-P6 RNA were produced through polymerase chain reaction assembly of oligonucleotides of length 60 nucleotides or smaller (Integrated DNA Technologies) using Phusion polymerase (Finnzymes). DNA templates were designed with the T7 RNA polymerase promoter (5′-TTCTAATACGACTCACTATA-3′) at their 5′ ends. A custom reverse transcription primer-binding site (5′-AAAGAAACAACAACAACAAC-3′) was included at the 3′ terminus of each template. RNA transcribed with T7 RNA polymerase (New England Biolabs) was purified using RNAClean XP beads (Beckman Coulter). RNA modification reactions were performed in 20-μl reactions containing 1.2 pmol of RNA. RNAs were incubated with 50 mM Na-Hepes (pH 8.0) at 90°C for 3 min and then cooled to room temperature. $MgCl_2$ at 0 or 10 mM final concentration was then added, followed by incubation at 50°C for 30 min and then room temperature before chemical mapping. Chemical probes were used at the following final concentrations: DMS (0.125%, v/v), CMCT in water (2.6 mg/ml), and 1M7 (1-methyl-7-nitroisatoic anhydride) [1.05 mg/ml in anhydrous dimethyl sulfoxide (DMSO) with final DMSO concentration of 25%]. Chemical probes were allowed to react for 15 min before quenching. 1M7 and CMCT reactions were quenched with 5.0 μl of 0.5 M sodium 2-(*N*-morpholino)ethanesulfonate (Na-MES) (pH 6.0), while the DMS reaction was quenched with 3.0 μl of 3 M NaCl, 5.0 μl of β-mercaptoethanol, 1.5 μl of oligo-dT beads [poly(A)purist, Ambion], and 0.25 μl of a 0.25 μM 5′-FAM-A20-Tail2 primer, which complements the reverse transcription primer-binding site at the RNA 3′ ends. The quench mixture was incubated at room temperature for 15 min, and the purification beads were pulled down with a 96 post magnetic stand and washed with 100 μl of 70% ethanol twice for RNA purification. RNAs were reverse-transcribed with SuperScript III reverse transcriptase at 48°C for 40 min (Life Technologies). RNA template was subsequently hydrolyzed for 3 min at 90°C in 0.2 M NaOH. After pH neutralization, complementary DNA (cDNA) on oligo-dT beads was pulled down by magnetic stand and washed with ethanol as above. cDNAs were eluted into 10 μl of ROX 350 standard ladder in Hi-Di Formamide (Life Technologies) using 1 μl of ROX 350 in 250 μl of Hi-Di Formamide. ABI 3700 sequencers were used for electrophoresis of cDNA. Capillary electrophoresis data were quantitated with HiTRACE (*33*). Data from these P4-P6 RNA experiments have been posted to the RNA Mapping Database (*34*) at the following accession IDs:

TRP4P6_WT2_0000	DMS, CMCT, and 1M7 for GAAA/11-nt (wild type) receptor
TRP4P6_C72_0000	DMS, CMCT, and 1M7 for GAAA/C7.2 receptor
TRP4P6_C7X_0000	DMS, CMCT, and 1M7 for GAAA/C7.10 receptor
TRP4P6_R1J_0000	DMS, CMCT, and 1M7 for GAAA/C7.2 receptor
TRP4P6_R1J_0001	R(1) compensatory mutants tested by 1M7 and DMS

Native gel shift experiments

Gel shifts were performed as previously described (*22*). Briefly, equimolar amounts of each RNA monomer at various concentrations (up to 20 μM final concentration) were mixed in water and denatured at 95°C for 1 min. Mixtures were cooled on ice for 2 min and annealed at 30°C for 5 min before the addition of Mg^{2+} buffer [9 mM tris-borate (pH 8.3) and 15 mM $Mg(OAc)_2$ final concentration]. After 30-min incubation at 30°C, samples were incubated at 10°C for 15 min before native gel analysis [7% (29:1) polyacrylamide gels in Mg^{2+} buffer at 10°C]. One of the monomers contained a fixed amount of 3′ end [^{32}P] pCp-labeled RNA (~0.25 to 0.5 nM final concentration). Monomer and dimer bands were quantified with ImageQuant, and dimer formation was plotted against RNA concentration. K_d's (dissociation constant) were determined as the concentration at which half of the RNA molecules were dimerized and converted to ΔG (relative to 1 M standard state) through the formula $\Delta G = k_B T \ln(K_d/1\ \text{M})$, where k_B is the Boltzmann constant and T is the temperature.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/5/eaar5316/DC1

Supplementary Text

Supplementary Methods

fig. S1. Illustrated descriptions and modeling constraints of all 82 benchmark test cases.

fig. S2. Rosetta free energy versus RMSD summaries of SWM modeling runs for 82 complex RNA motifs.

fig. S3. Comparison of model accuracy between SWM and fragment assembly of RNA with FARFAR over an 82 motif benchmark.

fig. S4. Potential routes to overcome limitations in Rosetta free energy function.

fig. S5. Compensatory mutagenesis of the R(1) receptor read out through chemical mapping.

fig. S6. Comprehensive single mutant analysis of the tetraloop receptor R(1).

fig. S7. Global fold changes between the template viral xrRNA and the Zika xrRNA structure prediction challenge.

fig. S8. Other models of RNA-Puzzle 18 (Zika xrRNA).

table S1. A comparison of the SWA and SWM methods using the same energy function as the original SWA benchmark set of trans-helix single-stranded loops, and SWM results using the updated Rosetta free energy function (SWM*).

table S2. Updates to the Rosetta energy function.

table S3. Detailed performance of the stepwise Monte Carlo algorithm on 82 benchmark cases.

table S4. Detailed performance of the FARFAR algorithm on 82 benchmark cases.

table S5. Measurements of interaction free energy between R(1) mutant tetraloop receptors and GGAA tetraloop.

data file S1. Three-dimensional SWM models canonical 11-nt:GAAA, R(1):GGAA, C7.2:GAAA, and C7.10:GAAA tetraloop/receptors in PDB format.

References (*35*–*87*)

REFERENCES AND NOTES

1. P. Bradley, K. M. S. Misura, D. Baker, Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).
2. Z. Miao, R. W. Adamiak, M. Antczak, R. T. Batey, A. J. Becka, M. Biesiada, M. J. Boniecki, J. M. Bujnicki, S.-J. Chen, C. Y. Cheng, F.-C. Chou, A. R. Ferré-D'Amaré, R. Das, W. K. Dawson, F. Ding, N. V. Dokholyan, S. Dunin-Horkawicz, C. Geniesse, K. Kappel, W. Kladwang, A. Krokhotin, G. E. Łach, F. Major, T. H. Mann, M. Magnus, K. Pachulska-Wieczorek, D. J. Patel, J. A. Piccirilli, M. Popenda, K. J. Purzycka, A. Ren, G. M. Rice, J. Santalucia Jr., J. Sarzynska, M. Szachniuk, A. Tandon, J. J. Trausch, S. Tian, J. Wang, K. M. Weeks, B. Williams II, Y. Xiao, X. Xu, D. Zhang, T. Zok, E. Westhof, RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* **23**, 655–672 (2017).
3. P.-S. Huang, K. Feldmeier, F. Parmeggiani, D. A. F. Velasco, B. Höcker, D. Baker, De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
4. S. Ovchinnikov, D. E. Kim, R. Y.-R. Wang, Y. Liu, F. DiMaio, D. Baker, Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins* **84** (suppl. 1), 67–75 (2016).
5. S. Ovchinnikov, H. Park, D. E. Kim, F. DiMaio, D. Baker, Protein structure prediction using Rosetta in CASP12. *Proteins* **86** (suppl. 1), 113–121 (2018).
6. M. Parisien, F. Major, The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51–55 (2008).
7. A. Krokhotin, K. Houlihan, N. V. Dokholyan, iFoldRNA v2: Folding RNA with constraints. *Bioinformatics* **31**, 2891–2893 (2015).
8. R. Das, J. Karanicolas, D. Baker, Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **7**, 291–294 (2010).
9. C. Zhao, X. Xu, S.-J. Chen, Predicting RNA structure with Vfold. *Methods Mol. Biol.* **1654**, 3–15 (2017).
10. M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K. M. Rother, J. M. Bujnicki, SimRNA: A coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **44**, e63 (2016).
11. P. Sripakdeevong, W. Kladwang, R. Das, An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20573–20578 (2011).
12. S. R. Eddy, How do RNA folding algorithms work? *Nat. Biotechnol.* **22**, 1457–1458 (2004).
13. R. Das, D. Baker, Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008).
14. R. Moretti, S. Lyskov, R. Das, J. Meiler, J. J. Gray, Web-accessible molecular modeling with Rosetta: The Rosetta Online Server that Includes Everyone (ROSIE). *Protein Sci.* **27**, 259–268 (2018).
15. R. Das, Four small puzzles that Rosetta doesn't solve. *PLOS ONE* **6**, e20044 (2011).
16. P. Sripakdeevong, M. Cevec, A. T. Chang, M. C. Erat, M. Ziegeler, Q. Zhao, G. E. Fox, X. Gao, S. D. Kennedy, R. Kierzek, E. P. Nikonowicz, H. Schwalbe, R. K. O. Sigel, D. H. Turner, R. Das, Structure determination of noncanonical RNA motifs guided by ^1H NMR chemical shifts. *Nat. Methods* **11**, 413–416 (2014).
17. J. Ferkinghoff-Borg, in *Bayesian Methods in Structural Bioinformatics*, T. Hamelryck, K. Mardia, J. Ferkinghoff-Borg, Eds. (Springer Berlin Heidelberg, 2012), pp. 49–93.
18. D. J. Mandell, E. A. Coutsias, T. Kortemme, Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6**, 551–552 (2009).
19. F.-C. Chou, P. Sripakdeevong, S. M. Dibrov, T. Hermann, R. Das, Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat. Methods* **10**, 74–76 (2013).
20. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack Jr., R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory. Comput.* **13**, 3031–3048 (2017).
21. A. A. Szewczak, T. R. Cech, An RNA internal loop acts as a hinge to facilitate ribozyme folding and catalysis. *RNA* **3**, 838–849 (1997).
22. C. Geary, S. Baudrey, L. Jaeger, Comprehensive features of natural and in vitro selected GNRA tetraloop-binding receptors. *Nucleic Acids Res.* **36**, 1138–1152 (2008).
23. M. Costa, F. Michel, Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: Comparison with in vivo evolution. *EMBO J.* **16**, 3289–3302 (1997).

24. P. W. Rose, A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman, S. K. Burley, The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45**, D271–D281 (2017).
25. J. S. Kieft, J. L. Rabe, E. G. Chapman, New hypotheses derived from the structure of a flaviviral Xrn1-resistant RNA: Conservation, folding, and host adaptation. *RNA Biol.* **12**, 1169–1177 (2015).
26. L. Jaeger, E. J. Verzemnieks, C. Geary, The UA_handle: A versatile submotif in stable RNA architectures. *Nucleic Acids Res.* **37**, 215–230 (2009).
27. B. M. Akiyama, H. M. Laurence, A. R. Massey, D. A. Costantino, X. Xie, Y. Yang, P.-Y. Shi, J. C. Nix, J. D. Beckham, J. S. Kieft, Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science* **354**, 1148–1152 (2016).
28. C. Geary, A. Chworos, E. Verzemnieks, N. R. Voss, L. Jaeger, Composing RNA nanostructures from a syntax of RNA structural modules. *Nano Lett.* **17**, 7095–7101 (2017).
29. J. D. Yesselman, D. Eiler, E. D. Carlson, A. N. Ooms, W. Kladwang, X. Shi, D. A. Costantino, D. Herschlag, M. C. Jewett, J. S. Kieft, R. Das, Computational design of asymmetric three-dimensional RNA structures and machines. *bioRxiv* **2017**, 223479 (2017).
30. L. Heo, M. Feig, What makes it difficult to refine protein models further via molecular dynamics simulations? *Proteins* **86** (suppl. 1), 177–188 (2018).
31. D. Tan, S. Piana, R. M. Dirks, D. E. Shaw, RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E1346–E1355 (2018).
32. W. Kladwang, T. H. Mann, A. Becka, S. Tian, H. Kim, S. Yoon, R. Das, Standardization of RNA chemical mapping experiments. *Biochemistry* **53**, 3063–3065 (2014).
33. S. Yoon, J. Kim, J. Hum, H. Kim, S. Park, W. Kladwang, R. Das, HiTRACE: High-throughput robust analysis for capillary electrophoresis. *Bioinformatics* **27**, 1798–1805 (2011).
34. J. D. Yesselman, S. Tian, X. Liu, L. Shi, J. B. Li, R. Das, Updates to the RNA mapping database (RMDB), version 2. *Nucleic Acids Res.* **46**, D375–D379 (2017).
35. L. Kinch, S. Yong Shi, Q. Cong, H. Cheng, Y. Liao, N. V. Grishin, CASP9 assessment of free modeling target predictions. *Proteins* **79** (suppl. 10), 59–73 (2011).
36. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
37. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, D. Baker, Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
38. J. M. Blose, M. L. Manni, K. A. Klapec, Y. Stranger-Jones, A. C. Zyra, V. Sim, C. A. Griffith, J. D. Long, M. J. Serra, Non-nearest-neighbor dependence of the stability for RNA bulge loops based on the complete set of group I single-nucleotide bulge loops. *Biochemistry* **46**, 15123–15135 (2007).
39. C. E. Longfellow, R. Kierzek, D. H. Turner, Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry* **29**, 278–285 (1990).
40. P. Thulasi, L. K. Pandya, B. M. Znosko, Thermodynamic characterization of RNA triloops. *Biochemistry* **49**, 9058–9062 (2010).
41. H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets* (World Scientific, ed. 5, 2009), pp. 1015–1024.
42. R. Tarjan, Depth-first search and linear graph algorithms. *SIAM J. Comput.* **1**, 146–160 (1972).
43. T. Xia, J. SantaLucia Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, D. H. Turner, Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).
44. F.-C. Chou, W. Kladwang, K. Kappel, R. Das, Blind tests of RNA nearest-neighbor energy prediction. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8430–8435 (2016).
45. I. W. Davis, L. W. Murray, J. S. Richardson, D. C. Richardson, MOLPROBITY: Structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **32**, W615–W619 (2004).
46. C. Y. Cheng, F.-C. Chou, R. Das, Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol.* **553**, 35–64 (2015).
47. F. L. Murphy, T. R. Cech, GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain. *J. Mol. Biol.* **236**, 49–63 (1994).
48. P. D. Renfrew, E. J. Choi, R. Bonneau, B. Kuhlman, Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. *PLOS ONE* **7**, e32637 (2012).
49. R. Das, Atomic-accuracy prediction of protein loop structures through an RNA-inspired ansatz. *PLOS ONE* **8**, e74830 (2013).
50. J. A. Ippolito, T. A. Steitz, The structure of the HIV-1 RRE high affinity rev binding site at 1.6 Å resolution. *J. Mol. Biol.* **295**, 711–717 (2000).
51. H. Shi, P. B. Moore, The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA* **6**, 1091–1105 (2000).
52. M. E. Burkard, D. H. Turner, NMR structures of $\text{r(GCA}\underline{\text{G}}\text{GC}\underline{\text{G}}\text{UGC)}_2$ and determinants of stability for single guanosine–guanosine base pairs. *Biochemistry* **39**, 11748–11762 (2000).
53. J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, C. E. Kundrot, T. R. Cech, J. A. Doudna, Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* **273**, 1678–1685 (1996).
54. B. Pan, Y. Xiong, K. Shi, J. Deng, M. Sundaralingam, Crystal structure of an RNA purine-rich tetraplex containing adenine tetrads: Implications for specific binding in RNA tetraplexes. *Structure* **11**, 815–823 (2003).
55. M. Egli, G. Minasov, L. Su, A. Rich, Metal ions and flexibility in a viral RNA pseudoknot at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4302–4307 (2002).
56. J. Deng, Y. Xiong, B. Pan, M. Sundaralingam, Structure of an RNA dodecamer containing a fragment from SRP domain IV of *Escherichia coli*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **59**, 1004–1011 (2003).
57. M. Wu, D. H. Turner, Solution structure of (rGCGGACGC)_2 by two-dimensional NMR and the iterative relaxation matrix approach. *Biochemistry* **35**, 9677–9689 (1996).
58. J. E. Wedekind, D. B. McKay, Crystal structure of the leadzyme at 1.8 Å resolution: Metal ion binding and the implications for catalytic mechanism and allo site ion regulation. *Biochemistry* **42**, 9554–9563 (2003).
59. B. Pan, Y. Xiong, K. Shi, M. Sundaralingam, Crystal structure of a bulged RNA tetraplex at 1.1 Å resolution: Implications for a novel binding site in RNA tetraplex. *Structure* **11**, 1423–1430 (2003).
60. C. C. Correll, J. Beneken, M. J. Plantinga, M. Lubbers, Y.-L. Chan, The common and the distinctive features of the bulged-G motif based on a 1.04 Å resolution RNA structure. *Nucleic Acids Res.* **31**, 6806–6818 (2003).
61. D. J. Klein, P. B. Moore, T. A. Steitz, The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.* **340**, 141–177 (2004).
62. M. P. Robertson, H. Igel, R. Baertsch, D. Haussler, M. Ares Jr., W. G. Scott, The structure of a rigorously conserved RNA element within the SARS virus genome. *PLOS Biol.* **3**, e5 (2005).
63. A. Serganov, Y.-R. Yuan, O. Pikovskaya, A. Polonskaia, L. Malinina, A. T. Phan, C. Hobartner, R. Micura, R. R. Breaker, D. J. Patel, Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.* **11**, 1729–1741 (2004).
64. J. SantaLucia Jr., D. H. Turner, Structure of (rGGCGAGCC)_2 in solution from NMR and restrained molecular dynamics. *Biochemistry* **32**, 12612–12623 (1993).
65. F. M. Jucker, H. A. Heus, P. F. Yip, E. H. M. Moors, A. Pardi, A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.* **264**, 968–980 (1996).
66. S. C. Ha, K. Lowenhaupt, A. Rich, Y.-G. Kim, K. K. Kim, Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature* **437**, 1183–1186 (2005).
67. R. K. Montange, R. T. Batey, Structure of the *S*-adenosylmethionine riboswitch regulatory mRNA element. *Nature* **441**, 1172–1175 (2006).
68. B. Pan, K. Shi, M. Sundaralingam, Crystal structure of an RNA quadruplex containing inosine tetrad: Implications for the roles of NH_2 group in purine tetrads. *J. Mol. Biol.* **363**, 451–459 (2006).
69. S. Nozinovic, B. Fürtig, H. R. A. Jonker, C. Richter, H. Schwalbe, High-resolution NMR structure of an RNA model system: The 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.* **38**, 683–694 (2010).
70. Y. V. Lerman, S. D. Kennedy, N. Shankar, M. Parisien, F. Major, D. H. Turner, NMR structure of a 4 × 4 nucleotide RNA internal loop from an R2 retrotransposon: Identification of a three purine–purine sheared pair motif and comparison to MC-SYM predictions. *RNA* **17**, 1664–1677 (2011).
71. S. D. Kennedy, R. Kierzek, D. H. Turner, Novel conformation of an RNA structural switch. *Biochemistry* **51**, 9257–9259 (2012).
72. M. Martick, T.-S. Lee, D. M. York, W. G. Scott, Solvent structure and hammerhead ribozyme catalysis. *Chem. Biol.* **15**, 332–342 (2008).
73. M. P. Robertson, W. G. Scott, The structural basis of ribozyme-catalyzed RNA assembly. *Science* **315**, 1549–1553 (2007).
74. Q. Zhao, Q. Han, C. R. Kissinger, T. Hermann, P. A. Thompson, Structure of hepatitis C virus IRES subdomain IIa. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64**, 436–443 (2008).
75. C. E. Dann III, C. A. Wakeman, C. L. Sieling, S. C. Baker, I. Irnov, W. C. Winkler, Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130**, 878–892 (2007).
76. S. D. Gilbert, R. P. Rambo, D. Van Tyne, R. T. Batey, Structure of the SAM-II riboswitch bound to *S*-adenosylmethionine. *Nat. Struct. Mol. Biol.* **15**, 177–182 (2008).
77. J.-D. Ye, V. Tereshko, J. K. Frederiksen, A. Koide, F. A. Fellouse, S. S. Sidhu, S. Koide, A. A. Kossiakoff, J. A. Piccirilli, Synthetic antibodies for specific recognition and crystallization of structured RNA. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 82–87 (2008).
78. C. C. Correll, B. Freeborn, P. B. Moore, T. A. Steitz, Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell* **91**, 705–712 (1997).
79. A. L. Feig, W. G. Scott, O. C. Uhlenbeck, Inhibition of the hammerhead ribozyme cleavage reaction by site-specific binding of Tb(III). *Science* **279**, 81–84 (1998).
80. S. Thore, C. Frick, N. Ban, Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch. *J. Am. Chem. Soc.* **130**, 8116–8117 (2008).
81. J. Wang, Inclusion of weak high-resolution x-ray data for improvement of a group II intron structure. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 988–1000 (2010).

82. R. T. Byrne, A. L. Konevega, M. V. Rodnina, A. A. Antson, The crystal structure of unmodified tRNA^{Phe} from *Escherichia coli*. *Nucleic Acids Res.* **38**, 4154–4162 (2010).
83. S. Dibrov, J. McLean, T. Hermann, Structure of an RNA dimer of a regulatory element from human thymidylate synthase mRNA. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 97–104 (2011).
84. L. Huang, A. Serganov, D. J. Patel, Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. *Mol. Cell* **40**, 774–786 (2010).
85. N. Safaee, A. M. Noronha, D. Rodionov, G. Kozlov, C. J. Wilds, G. M. Sheldrick, K. Gehring, Structure of the parallel duplex of poly(A) RNA: Evaluation of a 50 year-old prediction. *Angew. Chem. Int. Ed. Engl.* **52**, 10370–10373 (2013).
86. M. Meyer, H. Nielsen, V. Oliéric, P. Roblin, S. D. Johansen, E. Westhof, B. Masquida, Speciation of a group I intron into a lariat capping ribozyme. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7659–7664 (2014).
87. N. B. Suslov, S. DasGupta, H. Huang, J. R. Fuller, D. M. J. Lilley, P. A. Rice, J. A. Piccirilli, Crystal structure of the Varkud satellite ribozyme. *Nat. Chem. Biol.* **11**, 840–846 (2015).

Acknowledgments: We thank F.-C. Chou for early discussions and Stanford Research Computing Center for expert administration of the BioX^3 clusters (supported by NIH 1S10RR02664701) and Sherlock clusters. **Funding:** We acknowledge financial support from the Burroughs Wellcome Fund [CASI (Career Awards at the Scientific Interface) to R.D.], NIH grants R21 CA219847, R35 GM122579, and R21 GM102716 (to R.D.), intramural research grants from the University of California, Santa Barbara Academic Senate (to L.J.), and a RosettaCommons grant. **Author contributions:** A.M.W., C.G., L.J., and R.D. designed the research. A.M.W., C.G., W.K., P.Z., and R.D. carried out the research. A.M.W., L.J., and R.D. wrote the manuscript. All authors reviewed the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data and computer code needed to evaluate the conclusions in this paper are available at links described in Materials and Methods. Additional data related to this paper may be requested from the authors.

Submitted 19 November 2017
Accepted 17 April 2018
Published 25 May 2018
10.1126/sciadv.aar5316

Citation: A. M. Watkins, C. Geniesse, W. Kladwang, P. Zakrevsky, L. Jaeger, R. Das, Blind prediction of noncanonical RNA structure at atomic accuracy. *Sci. Adv.* **4**, eaar5316 (2018).

Science Advances

Blind prediction of noncanonical RNA structure at atomic accuracy

Andrew M. Watkins, Caleb Geniesse, Wipapat Kladwang, Paul Zakrevsky, Luc Jaeger and Rhiju Das

Sci Adv **4** (5), eaar5316.
DOI: 10.1126/sciadv.aar5316

Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/4/5/eaar5316/DC1

Supplementary Materials for

Blind prediction of noncanonical RNA structure at atomic accuracy

Andrew M. Watkins, Caleb Geniesse, Wipapat Kladwang, Paul Zakrevsky, Luc Jaeger, Rhiju Das

Published 25 May 2018, *Sci. Adv.* **4**, eaar5316 (2018)
DOI: 10.1126/sciadv.aar5316

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at advances.sciencemag.org/cgi/content/full/4/5/eaar5316/DC1)

- data file S1. Three-dimensional SWM models canonical 11-nt:GAAA, R(1):GGAA, C7.2:GAAA, and C7.10:GAAA tetraloop/receptors in PDB format.

Supplementary Text

Recovery of enumerative stepwise assembly modeling with stepwise Monte Carlo on trans-helix loops

In previous work, we hypothesized that experimentally observed RNA loop structures could be recovered through the stepwise addition of nucleotides, each placed in conformations making favorable contacts with previously built nucleotides (*11*). However, the previous stepwise assembly (SWA) method involved a computationally expensive enumeration of such conformations. This study presents a Monte Carlo minimization version of the method, called stepwise Monte Carlo (SWM, main text Fig. 1), which chooses stochastically between a variety of possible moves and then accepts or rejects these moves based on the Metropolis criterion.

We tested whether SWM might reproduce at less computational cost the deep energy optimization and atomic accuracy recovery of irregular single-chain loop conformations previously observed with SWA. We revisited a previously curated benchmark of 15 loops with lengths ranging from 4 to 10 nucleotides, taken from ribozymes, riboswitches, and the ribosome (*11*). The loops in this original benchmark were required to connect two distinct helices in a large RNA structure ('trans-helix'), to lack interactions between 5´-most and 3´-most nucleotides, and to exhibit conformations significantly different from those seen in A-form helices. To test the new stochastic SWM method, we created 5000 independent models using the same energy function used by SWA.

Based on the results above, we expected SWM to achieve low-energy models with significantly less computational expense than the enumerative SWA. To evaluate this hypothesis, we noted that for most SWM runs, we observed numerous models similar to the lowest energy cluster. In practical scenarios, we often assess completion of modeling with stochastic methods based on such repeated recovery of similar models from different runs and discontinue computation after observing such replication (*8*). We therefore estimated the minimal cost of achieving convergent runs with SWM, by calculating the frequency of seeing models within 2.0 Å of the lowest energy model over our full 5000-model runs (main text Fig. 1C and table S1). For the majority of cases, SWM gave improved computational efficiency compared to SWA. Indeed, 9 of the 10 loops that were not drawn from the 23S rRNA gave such improvements, in some cases allowing a reduction of computational expense by 2 orders of magnitude. For trans-helix loops drawn from the 23S rRNA, SWM showed similar or worse computational efficiency compared to enumerative SWA (red points, main text Fig. 1C). The ribosomal test cases feature a particularly large number of surrounding nucleotides, and concomitantly many viable interacting conformations, compared to the smaller riboswitch and ribozyme test cases. We hypothesized that as a result of these extensive interactions, stochastic SWM modeling of ribosomal loops may be more likely to be trapped in incorrect minima, compared to SWA enumeration, which retains a large ensemble of conformations at each step of loop buildup and is therefore less sensitive to inaccuracies in the assumed energy function at intermediate steps. This hypothesis predicts that improvements in the energy function might rescue conformational sampling for SWM and is tested below.

To assess the accuracy of SWM vs. SWA, we calculated the all-heavy-atom RMSD of the *ab initio* loop configurations after clustering of the lowest energy models, using the same analysis procedure as in ref. (*11*) (see Supplemental Methods for details). As in that work and previous work in RNA and protein modeling (as well as other areas of modeling, including computer vision), we assess herein the top five models (*8, 11, 35, 36*). Previously, in 10 of the 15 trans-helix loop benchmark cases, one of the five lowest-energy SWA models achieved 1.5 Å RMSD agreement with the crystallographic loop structure, and we reproduced this result with SWA modeling using the current Rosetta database (table S1). We observed that SWM achieved the same 1.5 Å RMSD within the top five cluster centers in 8 of 15 cases (table S1). The somewhat lower accuracy of SWM compared to SWA (8 vs. 10 loops recovered at high resolution) was ameliorated by energy function improvements, as described in the next two sections.

Energy function improvements rapidly tested by SWM

The gains in speed afforded by SWM allowed rapid tests of energy function updates that could not previously be carried out with the enumerative SWA method. Table S2 documents updates in the Rosetta RNA energy function relative to prior work with the enumerative stepwise assembly procedure (*11*). Because stepwise Monte Carlo involves comparison of conformations after simulated structuring and release (addition and deletion) of residues, the work herein made use of a free energy function that estimated not only the free energy of 'structured' residues but also of any 'unstructured' residues that were not explicitly represented in a given working conformation. Since the derived final lowest energy models typically contained most or all of the residues, the fine details of the treatment of 'unstructured residues' did not strongly impact final model selection, but we nevertheless sought to treat them realistically to ensure smooth Monte Carlo behavior at early stages of the simulations.

A prior heuristic entropic bonus for single nucleotide bulges (*rna_bulge*) was replaced with a more general framework to capture the entropic costs of structuring individual nucleotides or closing loops. A term *ref* assigned a fixed cost for instantiating A, C, G, or U, analogous to the amino-acid-wise *ref* term used in protein design (*37*). A term *loop_close* estimated the conformational entropy cost for freezing the two endpoints of each string of *L* nucleotides

$$loop\_close = -k_{\mathrm{B}}T\log(C_{eff}(D)/K_d^{close}) \qquad (1)$$

where the effective concentration of having one chain endpoint a distance *D* from the other endpoint is based on an analytical Gaussian chain model

$$C_{eff}(D) = \frac{1}{(2\pi)^{3/2}\sigma^3}e^{-\frac{D^2}{2\sigma^2}} \qquad (2)$$

Here, $\sigma^2 = L\,\sigma_{nt}^2$, with the Gaussian variance σ_{nt}^2 incurred per loop nucleotide assumed to be (5.0 Å)^2, based on simulations of simple loops. The Gaussian variance is related to the radius-of-gyration R_g of the free Gaussian chain by $R_g = \sqrt{3}\sigma$. The cost of freezing a loop endpoint was expected to reflect the positioning of loop endpoints to an Angstrom-level precision of ~1 Å, so that $\log(K_d^{close}/1\ \text{Å}^3) \sim 0.0$. Indeed, setting this value to –0.29 recovered experimental free energy costs measured for trapping single or multiple nucleotide bulges between helical pairs (3–4 kcal/mol) (*38-40*).

Stepwise Monte Carlo also frequently encounters cases where separate segments of an RNA are instantiated, but the interconnecting loops are not yet instantiated. For example, at the beginning of modeling of a two-way junction, the two starting helices that flank the junction each have known conformations, but their relative location and orientation are unknown and the two loops that interconnect the helices are not explicitly modeled. The same *loop_close* Gaussian chain model in eq. (1)-(2) was integrated to capture the conformational entropy in these configurations compared to conformations where the loops are fully free. For 2-segment internal loops, the expression is

$$C_{eff}(d) = \frac{1}{2(2\pi)^{3/2}\sigma D_1 D_2}\left(e^{-\frac{(D_1-D_2)^2}{2\sigma^2}} - e^{-\frac{(D_1+D_2)^2}{2\sigma^2}}\right) \tag{3}$$

where D_1 and D_2 are distances between loop points of attachment in segments 1 and 2. For 3-segment loops (as occur at the beginning of runs simulating 3-way junctions), the expression is

$$C_{eff}(d) = \frac{1}{16\pi D_1 D_2 D_3}\begin{bmatrix} -\operatorname{erf}(d_1 + d_2 + d_3) + \operatorname{erf}(d_1 + d_2 - d_3) + \\ \operatorname{erf}(d_1 - d_2 + d_3) + \operatorname{erf}(-d_1 + d_2 + d_3) \end{bmatrix} \tag{4}$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}$ is the error function and the values $d_i = D_i/(\sqrt{2}\sigma)$ are rescaled distances. For the general case of *N* segments

$$C_{eff}(d) = \frac{1}{2^n\pi\left(\sqrt{2}\sigma\right)^3 d_1 d_2 \dots d_N}\sum_{\substack{\lambda_1=\pm1,\\ \lambda_2=\pm1,\\ \dots\\ \lambda_N=\pm1}}\begin{bmatrix} \lambda_1\lambda_2 \dots \lambda_N(-1)^{N+1} \times \\ i^{N-3}\operatorname{erfc}(\lambda_1 d_1 + \lambda_2 d_2 + \dots + \lambda_N d_N) \end{bmatrix} \tag{5}$$

where $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$ is the complementary error function, and $i^n\operatorname{erfc}$ refers to the *n*-th integral of complementary error function, defined by $i^n\operatorname{erfc}(x) = \int_x^\infty dx\ i^{n-1}\operatorname{erfc}(x)$. Here, the loops are assumed to form a cycle through the segments, i.e., the loop with 5´ end taking off from segment *i* ends with the 3´ end in the segment *i*+1 (modulo *N*), as occur in N-way junctions. Eqs. (3)-(5) can be derived from, e.g., Fourier transformation of the convolutions of eq. (2) for each loop (*41*), and these calculations are documented in the freely available Rosetta code (`src/core/scoring/func/GaussianChainFunc.cc`). For working conformations that involved multiple loop cycles which were separable (not cycle shared a loop with another cycle), the separate entropy costs were summed into *loop_close*. More complex loop interconnection topologies in which loop cycles are not separable (i.e., at least two cycles share a loop) are not amenable to simple analytical formulae like (1)-(5); fortunately, moves leading to such complex loop arrangements occurred rarely, and could be rapidly identified using Tarjan's algorithm (*42*) and disallowed as valid moves during the stepwise Monte Carlo procedure. We provide a flag

`-allow_complex_loop_graph` to still allow such loop arrangements in cases where they must be modeled (e.g., in RNA puzzle 18, which had a double pseudo-knot structure); in that case *loop_close* is summed over all cycles, including those that are not separable.

Further updates included a torsional potential *RNA11_based_new* derived from a non-redundant, filtered set of empirically observed RNA structures, developed for crystallographic refinement applications (*19*); additional terms captured the entropic cost of bringing two RNA chains together (*intermol*; calibrated based on measurements of helix initiation costs at 1 M standard state(*43*)) and of leaving free ribose moieties and 2´-OH hydroxyls (*free_suite*, *free_2HOprime*) that were not held by structural interactions; an electrostatic term between stacked bases *stack_elec* to supplement Rosetta hydrogen bonding and replace carbon-hydrogen bonds *ch_bond*, described in (*44*); increased repulsion between hydrogens (*enlarge_H_lj_wdepth*) to reduce steric clashes as assessed by MolProbity (*45*) and to disallow over-twisting seen in optimized A-form helix conformations (*2, 46*); and revisions to favor *syn*-G conformations and backbone torsional combinations represented in UUCG tetraloops, a difficult case for Rosetta modeling (*15*) (see Supplemental Methods and table S2). The *ref* weights were set, along with modest re-weighting of previous terms, to match nearest-neighbor stacked pair measurements, as described under 'single-conformation' Rosetta scoring in (*44*). While justified based on other modeling problems and previously tested in energetic modeling challenges for RNA and RNA-protein interfaces ((*44*) and K. Kappel, R. Das, unpublished results), these terms have not been tested in *ab initio* RNA structure modeling scenarios.

Tests of updated energy function in trans-helix loops

On the 15-loop trans-helix benchmark, inclusion of Rosetta energy updates (see previous section; table S2) gave improvements in modeling accuracy (table S1). One of the top five SWM cluster centers achieved 1.5 Å RMSD accuracy in 10 cases, improved from 8 with the prior free energy function and equal to SWA's performance on these loops. This gain in accuracy was also quantitatively reflected in the improved RMSD values among top 5 cluster centers (1.22 Å to 0.91 Å, median values in table S1). For the longest ribosomal RNA loops (nts 2534-2540, 1976-1985, 2003-2012), conformational sampling remained an issue with SWM (poor RMSD values in table S1; energy gaps to experimental structures greater than 0 in table S1; score vs. RMSD plots in fig. S2). However, in other cases, we saw notable improvements, including the shorter 23S rRNA loops that were problematic for SWM with the original energy function (nts 531-549, main text Fig. 1E, table S1), and aloop from a viral pseudoknot that could not previously be solved by SWA (main text Fig. 1F; this recovery required a larger clustering RMSD cutoff; see Methods and table S3). For the only other failed SWM case amongst these trans-helix loops, the HCV IRES IIa loop, we traced the likely problem to neglect of a metal ion binding site (fig. S4).

SWM models unify structural pictures of the class I tetraloop/receptors

GNRA tetraloops (N = any nucleotide, R = purine) are recognized by special receptor sequences throughout natural RNA molecules, and *in vitro* evolution has revealed additional tetraloop/receptor pairs useful for 3D RNA design applications (*22, 23, 47*).

As independent experimental tests, the models make non-trivial predictions for single-nucleotide resolution chemical mapping. Each of the receptor structures presents bulged uridine nucleotides: U9 in the 11-nt receptor and C7.10, U5 and U8 in C7.2, and U8 in R(1). These nucleotides are expected to give strong reactivity to CMCT (*N*-cyclohexyl-*N'*-(2-morpholinoethyl)carbodiimide metho-*p*-toluenesulfonate). In prior work, we confirmed the expected strong chemical reactivity for the 11-nt and C7.2 cases (*11*). Here, we have repeated those measurements as well as additional mapping for the other two cases and find that the predicted uridines and no other nucleotides are strongly reactive, supporting the SWM models (main text Fig. 4D).

An extensive mutational analysis of receptor R(1) was perform to supplement CMCT probing in substantiating the SWM model of the GGAA/R(1) tetraloop/receptor interaction. A tectoRNA heterodimer system (main text Fig. 4E) was used to examine R(1) mutants for their ability to maintain GGAA binding. The tectoRNA dimer is assembled through two loop/receptor interactions. One of these interactions behaves as an anchor for assembly, while the second contains the loop/receptor pair of interest. The high affinity, highly selective GAAA/11nt-motif interaction was use as the anchor interaction for all assays. A set of tectoRNA hairpins that contained mutant R(1) receptors of interest were then tested for GGAA binding within this tectoRNA system. The extent of dimer formation was determined at multiple tectoRNA concentrations by native polyacrylamide gel electrophoresis (PAGE) in the presence of 15 mM $Mg(OAc)_2$, and used to calculate equilibrium dissociation constants for each GGAA/receptor pair. GGAA binding affinities are reported in terms of ΔΔG in reference to the R(1) wildtype receptor at 10 °C, unless otherwise stated in the text.

All point mutations and single nucleotide deletion variants of nucleotide positions 3 through 10 within receptor R(1) were examined for GGAA tetraloop binding. In addition, several double and triple mutations to specific submotifs that were predicted by the SWM model were tested (fig. S9; table S5). All results were consistent with and, in some cases, gave incisive evidence for the R(1) structural model, as follows:

1. *A-minor docking site* The C1:G12 and C2:G11 base pairs were not varied in the current study. Previous variation of these R(1) base pairs, as well as the analogous C1:G11 and C2:G10 base pairs within the 11nt-motif receptor, were shown to be detrimental to tetraloop binding (*22*). The reduction in GGAA affinity for mutations in this CC/GG region of R(1) is consistent with their involvement in the formation of an A-minor interaction with the docked tetraloop, as predicted by the SWM model.
2. *U•U docking site*. The specificity of receptor R(1) for a G at the second position of the tetraloop (G<u>G</u>AA) was predicted by SWM to be defined by a non-canonical U3:U10 WC base pair. Specifically, positions O2 and O2´ of nucleotide U3 form hydrogen bonds with the N2 position of the guanine base at the second position of the tetraloop. A U3C mutation only slightly reduces GGAA affinity by +0.40 kcal mol $^{-1}$, presumably due to a cytosine base being able to maintain contact through position O2 to the N2 position of the loop guanine. Nucleotide U10 also has the ability to form a hydrogen bond contact with the second G of the tetraloop. It is possible for the 2´ OH position of nucleotide U10 to interact with O6 of the second loop guanine. Interaction with the second loop guanine by both U3 and

U10 seems to be due to the narrowed diameter of the receptor as a result of the pyrimidine:pyrimidine (Y:Y) mismatch. This is supported by the U10C point mutant being significantly active, displaying +2.15 kcal mol $^{-1}$ change in affinity. The U10C and U3C mutations are least likely to disrupt the geometry of the wildtype Y:Y mismatch, and would explain why these mutations are less detrimental to GGAA binding than any other point mutation or deletion to the U3:U10 base pair. Although most substitutions of a purine base to either U3 or U10 are unfavorable for GGAA (> 4 kcal mol $^{-1}$), the U3A mutation is more active than other purine substitutions (+3.03 kcal mol $^{-1}$). It may be possible that a U3A mutation is able to maintain one of the two interactions associated with the U3 nucleotide. A U10:A3 transWC:HG bp may be able to maintain interaction through to position O6 of the loop guanine through position N1 of nucleotide A3, but would be likely to distort the backbone geometry of the receptor. However, a U10:A3 cisWC:HG bp may keep the interstrand distance of the 3:10 bp narrow enough to facilitate both 2´ OH interactions (from A3 and U10) with the loop guanine, but would result in loss of the nucleobase-specific hydrogen bond at nucleotide A3.

3. *Dinucleotide platform.* Nucleotides G4 and U5 are predicted to form the dinucleotide platform, analogous to the AA platform seen in the 11nt-motif receptor, upon which the bound GGAA tetraloop stacks. The U5:G4 cisHG:SG base pair appears to be stabilized by formation of a cisWC:WC interaction between G4 and C9, resulting in a triple base pair platform. Any point mutant of G4 or U5 would not be expected to maintain the noncanonical HG:SG base pair of the platform, and such mutants resulted in decreased GGAA affinity by at least + 3.5 kcal mol $^{-1}$. Mutation of C9 would be expected to disrupt its WC edge pairing with G4. Interestingly, the C9U mutant is less detrimental to GGAA binding than other C9 mutations, as the C9U mutant could form a wobble interaction with G4. Point mutation of C9 to either purine base further destabilized GGAA binding, however the C9A mutation was less detrimental than C9G.
4. *Base triple.* Compensatory mutations within the predicted triple base pair were able to partially rescue the drop in GGAA binding observed for individual single point mutations (main text Fig. 4D and Fig. S9). Double mutations to G4 and C9 that maintain the possibility of WC pairing are less detrimental than any single point mutation that disrupts WC pairing between these nucleotides, increasing the confidence of a WC edge interaction between G4 and C9 as predicted by SWM. A triple mutation, substituting the entire G4:U5:C9 triple base pair with a potential A4:C5:U9 triple pair, displayed higher affinity binding than either the U5C point mutant or the G4A:C9U double mutant, substantiating the prediction of a triple base pair between these nucleotide positions. A second triple mutation to A4:A5:U9, which was designed to contain the typical AA platform observed within the 11nt-motif receptor, gave very weak binding (+4.08 kcal mol $^{-1}$ relative to wildtype). However, the U5A mutation contained within this designed triple interaction could also potentially form a WC bp with nucleotide U8, as U8 is otherwise predicted to be in bulge in the SWM model of R(1). This would eliminate formation of the platform, and explain the poor binding affinity observed for this triple. Also, the U5A point mutation is more deleterious (+4.40

kcal mol^{-1}) than any other point mutation to a nucleotide within the triple (G4, U5 or C9).

5. *Closing noncanonical WC pair.* SWM predicted that the R(1) receptor would be closed by a G6:A7 noncanonical WC:WC bp. Maintaining the purine:purine (R:R) mismatch appears to be significant for maintaining the proper geometry of the receptor structure. Point mutations G6U, A7C or A7U, each of which result in formation of a canonical WC or G:U wobble base pair, resulted in a significant loss of binding affinity (+2.71, +4.18 and +3.56 kcal mol $^{-1}$, respectively). On the contrary, individual point mutations G6A and A7G, resulting in an A:A or G:G mismatch, destabilize binding by a more modest +1.89 and +1.90 kcal mol^{-1}, respectively. A compensatory double G6A, A7G mutation to reverse the base pair results in a negligible difference in GGAA affinity compared to the wildtype (+0.07 kcal mol $^{-1}$). Taken together, these results strongly support the formation of a G:A WC:WC mismatch base pair.
6. *Bulged spacer.* As mentioned above, nucleotide U8 is predicted to be in a bulge, and not involved in base pairing interactions with the GGAA tetraloop or other receptor nucleotides. Any point mutation or deletion of this nucleotide results in a decrease in affinity by +2.05-2.55 kcal mol^{-1}. This reduction in affinity is modest compared to other mutations made within the receptor, supporting the SWM prediction that the identity of this nucleotide is not a deciding factor for receptor function. In addition, of all single nucleotide deletions tested with R(1), the U8 deletion exhibits at least 1 kcal mol $^{-1}$ stronger GGAA affinity than any other deletion, signifying it is the least important nucleotide in terms of global receptor structure and function (fig. S9). Interestingly, the second and third strongest binding deletion variants are A7del and C9del. A7 and C9 are directly adjacent in the oligonucleotide strand to U8. A7 and C9 are each predicted to interact in a WC:WC manner with a guanine base in the SWM model. It is possible that upon deletion of either A7 or C9, nucleotide U8 can fill this void and pair with the cognate guanine to partially stabilize the receptor and dampen the diminishment of GGAA binding affinity.

Taking these results in total, the analysis of single, double and triple mutants of R(1) on GGAA tetraloop affinity support the predicted structure of receptor R(1) bound by a GGAA tetraloop, as generated by SWM.

Supplementary Methods

Rosetta structure prediction

Stepwise Monte Carlo (SWM)

Briefly, the Monte Carlo moves in SWM include the addition of single nucleotides to chain termini: these moves involve applying filters for favorable contacts and continuous minimization of motif torsions in a physically realistic all-atom energy function. The sampled torsion combinations and filters exactly match those developed in SWA (*11*). In addition, resample moves sample alternative torsions at specific suite connections at chain termini or internal to chains. Such resampling could not be tested in the previous enumerative SWA scheme due to the explosion in the number of buildup pathways but could be included in SWM due to the ease of including the moves in the stochastic scheme. Last, deletion of single nucleotides at chain termini, followed by torsional minimization of the remaining conformation, is also allowed.

For residues added or resampled with one fixed connection, a 'suite' of torsions is sampled: ε and ζ from residue *i* and α, β, and γ from residue *i+1*. (The δ torsion is sampled indirectly, by optionally sampling both north and south sugar puckers.) By default, α, γ, and ζ are sampled every 20 degrees, while β's 20 degree bins must fall within 100 degrees of 180 and ε torsions are sampled within 20 degrees of a pucker-dependent value. For residues added or resampled with two fixed connections, the prior ε and ζ, the residue's α, and the following residue's α are sampled as above; the remaining torsions are obtained through kinematic closure. To eliminate conformations likely to score poorly, a series of filters is applied. First, for native screens or for 'align' benchmark problems, the pose is discarded if its RMSD falls outside the requested threshold. For two-connection kinematic closure problems, a filter checks for successful chain closure solutions. Structures may be screened to ensure that certain residues are making stacking or pairing interactions, that they are not making steric clashes with structural context omitted from the modeling problem but described as a repulsive grid, and that residues from different 'partitions' (particularly, residues to be built in the modeling problem and input residues) are contacting each other.

For multistranded cases on the 82-motif benchmark, the new tetraloop/receptor structures, and RNA-puzzle 18 (Zika xrRNA) we used the following command-line with the new Rosetta `stepwise` executable, here illustrated for the T-loop modeling challenge `t_loop_modified_fixed:`

```
stepwise.linuxclangrelease -s t_loop_modified_fixed_START1_1ehz.pdb -native
t_loop_modified_fixed_NATIVE_1ehz.pdb -terminal_res  A:52 A:62
-block_stack_above_res  A:62   -block_stack_below_res  A:52   -extra_min_res
A:53 A:61  -fasta t_loop_modified_fixed.fasta -save_times -score:weights
stepwise/rna/rna_res_level_energy4.wts -cycles 2000 -submotif_frequency 0.0
-nstruct 100 -motif_mode -out:file:silent SWM/0/swm_rebuild.out
```

The suffix `.linuxclangrelease` of the executable depends on the platform and compiler used. In this illustrative case, the known crystallographic model of tRNA(phe), including chemical modifications (PDB: 1EHZ) was used as a template, with the coordinates of the T-loop removed. The `–s and –native` flags specify this template structure and the experimental structure that includes the loop. Most benchmark and blind prediction cases involved *de novo* modeling of a noncanonical motif with no flanking helices or other context provided. In those cases, the Rosetta `rna_helix.py` application was used to prepare models of flanking helices (*44*), and these helices were provided to the `stepwise` executable through the `–s` flag. While Rosetta does possess the capacity to incorporate chemical mapping information to guide sampling, these cases were solved based exclusively on the input geometry provided.

Additional parameters include `–terminal_res`, which are all the residues that begin or end chains; `–block_stack_above_res` and `–block_stack_below_res` which place a steric repulsive pseudoatom above and below, respectively, the plane of the nucleobases at the termini of flanking helices where other helix base pairs would typically be present in a full length RNA molecules; `-extra_min_res`, which specifies residues that ought to be minimized along with the built residues during modeling (typically helix residues immediately flanking the loops of interest); and `-motif_mode`, which, if provided, helps to calculate default values for the aforementioned flags, when they are unspecified, and provide automated cross-checks on those flags, when they are specified. The `-fasta` flag provides the sequence of the target molecule, including the motif to be built. The `-save_times` flag turns on recording of computation times per model. The `-score:weights` flag specifies a file of weights that defined the assumed Rosetta energy function (see also table S2 and "Updates to the Rosetta energy function" below). The `-cycles flag` specifies the number of Monte Carlo moves per model. The `-submotif_frequency 0.0` flag turns off an experimental option that allows direct copying of recurrent submotifs (such as UA handles) as a stepwise *submotif* move. The `-nstruct`, and `-out:file:silent` flags give the number of models to create per CPU and the name of a compressed file format ('silent file') used in Rosetta. A slightly different SWM command-line was used for 15 trans-helix single-stranded loops excised from crystal structures to enable fair comparisons to the prior SWA method; it is described in detail in the next section. For RNA-puzzle 18 (Zika xrRNA) modeling, we included the flag `-allow_complex_loop_graph` (see Supplemental Text, "Energy function improvements rapidly tested by SWM").

A few challenges simulated modeling scenarios in which the positions of flanking helices are approximately but not exactly known (test cases including the tag 'align_' in fig. S1). In these test cases, SWM permitted a limited amount of movement of flanking helices with the flag `-align_pdb` and a constraint distance beyond which to penalize structures (generally 4.0 Å), inputted through a `-rmsd_screen` flag. For example, for the benchmark case `t_loop_align`, we used the following command line:

```
stepwise.linuxclangrelease -s t_loop_align_HELIX1.pdb
-native t_loop_align_NATIVE_3l0u.pdb -terminal_res  A:52 A:62
-block_stack_above_res  A:62   -block_stack_below_res  A:52
-extra_min_res  A:53 A:61  -fasta t_loop_align.fasta -save_times
-align_pdb t_loop_align_ALIGN_3l0u_RNA.pdb -rmsd_screen 4.0 -score:weights
stepwise/rna/rna_res_level_energy4.wts -cycles 2000 -submotif_frequency 0.0
-nstruct 100 -motif_mode -out:file:silent SWM/0/swm_rebuild.out
```

Benchmark cases involve modeling chemical modified nucleotides including 2´-OMe cytosine, 2-thiomethyl-N6-isopentenyladenosine, inosine, 5-bromouridine, wybutosine, dimethylguanosine, pseudouridine, 2´-OMe guanosine, 5-methylcytosine, 1-methyladenosine, and 5-methyluridine. To parameterize the geometries of these nucleotides, we began with Rosetta's ideal coordinates for an unmodified nucleotide (i.e., guanosine for 2´-OMe guanosine), and added atoms, bonds, and internal coordinates for the new atoms. Then, we added using pseudo-polymeric methyl and phospho-methyl 'capping' groups, by analogy to the acetyl and N-methyl amidyl caps commonly used to generate peptide units. These structures were optimized in Gaussian09 at the B3LYP level of theory, and the resulting bond lengths and bond angles were employed for subsequent simulation (*48*). Torsional potentials for chemically modified nucleotides were inherited from the most chemically similar unmodified nucleotide.

All of the above helix and template set up and stepwise Monte Carlo modeling are automatically generated by the `setup_stepwise_benchmark.py` script found in the rna_benchmark repository (see main text Materials & Methods), within the scripts/python/benchmark_util directory.

At the end of SWM, some models did not have all residues instantiated. We therefore took every generated structure, compiled into one Rosetta compressed output file (swm_rebuild.out) and filled in missing nucleotides with placeholder residues. The build_full_model executable adds 'repulsive-only' variants of the nucleotides missing from the target sequence (these variants are not allowed to make favorable interactions), and then closes loops involving these placeholder nucleotides using 5000 cycles of fragment assembly at just those loops:

```
build_full_model.linuxclangrelease –in:file:silent swm_rebuild.out
–in:file:fasta t_loop_modified_fixed.fasta –out:file:silent
swm_rebuild_full_model.out –in:file:native
t_loop_modified_fixed_NATIVE_1ehz.pdb
–stepwise:monte_carlo:from_scratch_frequency 0.0 –out:overwrite true
–score:weights stepwise/rna/rna_res_level_energy4.wts –virtualize_built false
–fragment_assembly_mode true –rna:evaluate_base_pairs true
–superimpose_over_all true –allow_complex_loop_graph true
```

To cluster the models resulting from each simulation for final evaluation, we ran the rna_cluster executable with 2 Å RMSD threshold, as in prior work with SWA (*11*) and FARFAR (*8*). An example command-line is as follows:

```
rna_cluster –in:file:silent swm_rebuild_full_model.out –nstruct 100
–cluster:radius 2.0 –out:file:silent
TOP_ENERGY_CLUSTERS/top_energy_clusters.rna_cluster.out –in:file:native
t_loop_modified_fixed_NATIVE_1ehz.pdb
```

The above post-processing command lines are automatically generated by the create_stepwise_benchmark_table.py script found in the rna_benchmark repository (see main text Materials & Methods), within the scripts/python/analysis directory.

For benchmarking, we initially acquired 400 SWM models per motif; if multiple low energy models within 2 Å of each other were not observed, we acquired further models until either the same lowest energy models were seen in independent runs or the energy dropped lower than optimized experimental models, symptomatic of energy function problems. Total CPU-hour expenditure for each model is given in table S3. A CPU here refers to one thread of a 16-thread Intel(R) Xeon(R) CPU E5-2650 (2.00 GHz) processor. The median total run time over the 82 motifs was 4,200 CPU-hours per motif. This level of computational power is typically available to most academic researchers using scientific computing (e.g., less than 24 hours on a cluster of 192 CPU-cores), including resources freely available to researchers through the Xtreme Scientific and Engineering Discovery Environment (XSEDE, https://www.xsede.org). In addition, we estimated the minimum number of CPU hours required to discover one model in the lowest energy cluster center as the sum of total CPU hours required to generate the SWM models divided by the total number of models found in the lowest energy cluster center. The median value of 248 CPU-hours is achievable on a single 16-core processor running for 24 hours.

Stepwise assembly (SWA)

For enumerative stepwise assembly (SWA) on the 15 trans-helix single-stranded loops excised from crystal structures, we ran the following command-line to set up jobs:

```
SWA_DAG/setup_SWA_RNA_dag_job_files.py -s 5P_j12_leadzyme_START1_1nuj_RNA.pdb
-native_pdb 5P_j12_leadzyme_1nuj_RNA.pdb -fasta 5P_j12_leadzyme.fasta
-sample_res 7 8 9 10 -force_field_file
stepwise/rna/rna_loop_hires_04092010.wts
-rna_torsion_potential_folder ps_03242010/ -single_stranded_loop_mode True
-VDW_rep_screen_info 5P_j12_leadzyme_50_ANGSTROM_GRID_1nuj_RNA.pdb
-apply_VDW_rep_delete_matching_res False -tether_jump False
```

In this case, the known crystallographic model of the 5′ J1/2 Leadzyme (PDB: 1NUJ) was used as the template, with the coordinates of the four-nucleotide loop residue removed. This master script was executed on a single CPU, and generates a directed acyclic graph of Rosetta command lines using the executables `swa_rna_main` and `swa_rna_cluster`; 500 separate CPUs were allocated for job distribution by the master script distributed with Rosetta. See ref. (*11*) for details.

To ensure fair comparison to the SWA runs, the SWM modeling for the 15 trans-helix single-stranded loops used a somewhat different command-line than for the broader benchmark:

```
stepwise.linuxclangrelease -s 5P_j12_leadzyme_START1_1nuj_RNA.pdb -native
5P_j12_leadzyme_1nuj_RNA.pdb -fasta 5P_j12_leadzyme.fasta -save_times
-score:weights stepwise/rna/rna_loop_hires_04092010.wts
-score:rna_torsion_potential ps_03242010/     -analytic_etable_evaluation
false -VDW_rep_screen_info 5P_j12_leadzyme_50_ANGSTROM_GRID_1nuj_RNA.pdb
-allow_internal_local_moves false -allow_internal_hinge_moves false
-from_scratch_frequency 0.0 -cycles 5000 -nstruct 10 -allow_split_off false
-allow_skip_bulge false -out:file:silent SWM/0/swm_rebuild.out
```

Most of the SWM flags are same as above. The `-score:rna_torsion_potential` flag provides an RNA torsional potential that directly matches SWA runs (see also table S2 and "Updates to the Rosetta energy function" below). The flag `-analytic_etable_evaluation false` uses an older computation method for van der Waals packing and solvation to allow comparison to SWA. The `-VDW_rep_screen_info` flag allowed modeling of steric repulsion from crystallographic residues that are far from the excised loop and not explicitly included in the run to save computational speed, particularly for ribosomal RNA test cases [see ref. (*11*)]. The `-allow_internal_local_moves false` and `-allow_internal_hinge_moves false` flags turned off stepwise Monte Carlo moves that resample internal coordinates with kinematic closure, to allow direct comparison to SWA (which does not have those moves). The `-from_scratch_frequency 0.0` and `-allow_split_off false` flags turned off stepwise Monte Carlo moves that generate dinucleotide conformations 'from scratch' in previously unbuilt parts of the molecule and that allow splitting of terminal loop regions, again to allow direct comparison to SWA (which does not have those moves). The `-allow_skip_bulge false` turned off moves that built nucleotides separated by bulges (developed for SWA but not tested here in either SWM or SWA due to their computational expense); this flag is unnecessary to specify (the default value of `-allow_skip_bulge` is false) but is described here for completeness.

To match prior SWA work, for table S1 only, we carried out clustering of the 15 trans-helix single-stranded loop cases modeled by SWA and SWM with the executable `swa_rna_main -algorithm_rna_cluster`; the protocol used a tighter cutoff for defining a cluster (0.7 Å instead of 2.0 Å above) and only calculates RMSD over loop residues, as described in detail in ref. (*11*).

To compare computational costs of enumerative SWA to SWM on the 15 trans-helix single-stranded loops excised from crystal structures (main text Fig. 1), we estimated for SWM the number of CPU hours required to discover one model in the lowest energy cluster center as the sum of total CPU hours required to generate each of 5000 SWM models divided by the total number of models found in the lowest energy cluster center. For SWA, we evaluated the computation requirements by summing over the total number of CPU hours recorded for all sub-steps comprising the entire build-up computation. Again, a CPU here refers to one thread of a 16-thread Intel(R) Xeon(R) CPU E5-2650 (2.00 GHz) processor.

Fragment Assembly of RNA with Full Atom Refinement (FARFAR)

As a baseline test in the 82-motif benchmark, we compared SWM to our previously reported fragment-based method, FARFAR. An example command line, for the test case `t_loop_modified_fixed`, was the following:

```
rna_denovo.linuxclangrelease –s t_loop_modified_fixed_HELIX1.pdb
–native t_loop_modified_fixed_NATIVE_1ehz.pdb –terminal_res  A:52 A:62
–block_stack_above_res  A:62   –block_stack_below_res  A:52
–extra_min_res  A:53 A:61 –working_res A:18 A:52-62  –fasta t_loop_align.fasta
-save_times –score:weights stepwise/rna/rna_res_level_energy4.wts
–cycles 20000 –nstruct 1000
–minimize_rna true –fragment_homology_rmsd 1.5
–exclusion_match_type MATCH_YR –chain_connection SET1  A:18 SET2  A:52-62
–out:file:silent FARFAR/0/farna_rebuild.out
```

Most options are the same as for the `rna_denovo` executable. The flag `–minimize_rna true` ensures that, following the low-resolution phase of fragment insertions, the RNA is minimized in the high resolution scoring function. The `–working_res` flag specifies a subset of residues to be excised from a larger structure: for the cases in this benchmark, all residues in the native PDB are specified, but if a larger experimental PDB had been specified, this value of `–working_res` would remove any extraneous residues during FARFAR modeling. The `–chain_connection` flag sets up appropriate kinematic connections for loop-loop contacts within a single RNA chain, in this case clarifying that one strand (residue A:18) must have some base pair (not specified *a priori*) with the other strand (residue A:52-62) in the tertiary contact.

The remaining flags helped simulate an *ab initio* modeling scenario: fragments too similar to the target structure were excluded through a new automated routine for detecting homologies at the fragment level and removing these contaminating homologs from the fragment database prior to the modeling run. The routine took the remodeled loop residues from the native target structure; if a loop contained more than six consecutive residues, we broke it down into 6-mer segments and carried out the following scan for each segment. The routine then searched the FARFAR fragment library for stretches that would align to that native segment within a certain radius (the value of -fragment_homology_rmsd in Å RMSD) and excluded those segments from the library. The -exclusion_match_type flag describes the sequence criterion by which a fragment is deemed homologous: MATCH_YR allows for inexact sequence matching as long as both the possible homolog and the target structure both have purine or both have pyrimidine. For a handful of structures with very short loops, this homolog discovery routine eliminated all available fragments in our database, and so we strengthened the requirement for fragment exclusion to 1.2 Å RMSD. We confirmed that this procedure discovered and successfully eliminated homologous fragments, often several, for recurrent motifs such as the signal-recognition particle. To be conservative, all fragments derived from possible homologs (not just the specific segments of the homologs with close matches) were excluded in FARFAR runs.

As noted above, some benchmark cases involved chemically modified nucleotides. When fragments were inserted into those positions, FARFAR used torsional combinations derived from crystallographic structures with the closest unmodified nucleotide (e.g., cytosine for 2′-OMe cytosine). Clustering of FARFAR models was carried out with rna_cluster, exactly as for SWM.

Evaluation of structure prediction

Accuracy evaluation (RMSD, recovery of non-Watson-Crick pairs)

Root-mean-squared deviation (RMSD) values were computed over all heavy atoms after superimposition to the experimental structure. Nucleotides that were bulged (no stacking or pairing of nucleobases) were automatically identified in the experimental structure and excluded from the superimposition and RMSD calculation. This calculation occurs in the build_full_model executable, described above. [For the 15 trans-helix single-stranded loops excised from crystal structures ('trans-helix-loop' cases), the flag -superimpose_over_all false was applied to ensure RMSDs were calculated only over loop residues, to match prior work (*11*).]

F_{NWC} calculations evaluated the fraction of experimentally observed non-Watson-Crick base pairs recovered by models. An experimental base pair was only considered recovered if both bases' edges (Watson-Crick, sugar, or Hoogsteen) and cis/trans orientation of glycosidic bonds matched between the model and experimental structure. Canonical base pairs (cis Watson-Crick/Watson-Crick interactions of C-G, A-U, and G-U) were excluded from this non-canonical pair recovery metric.

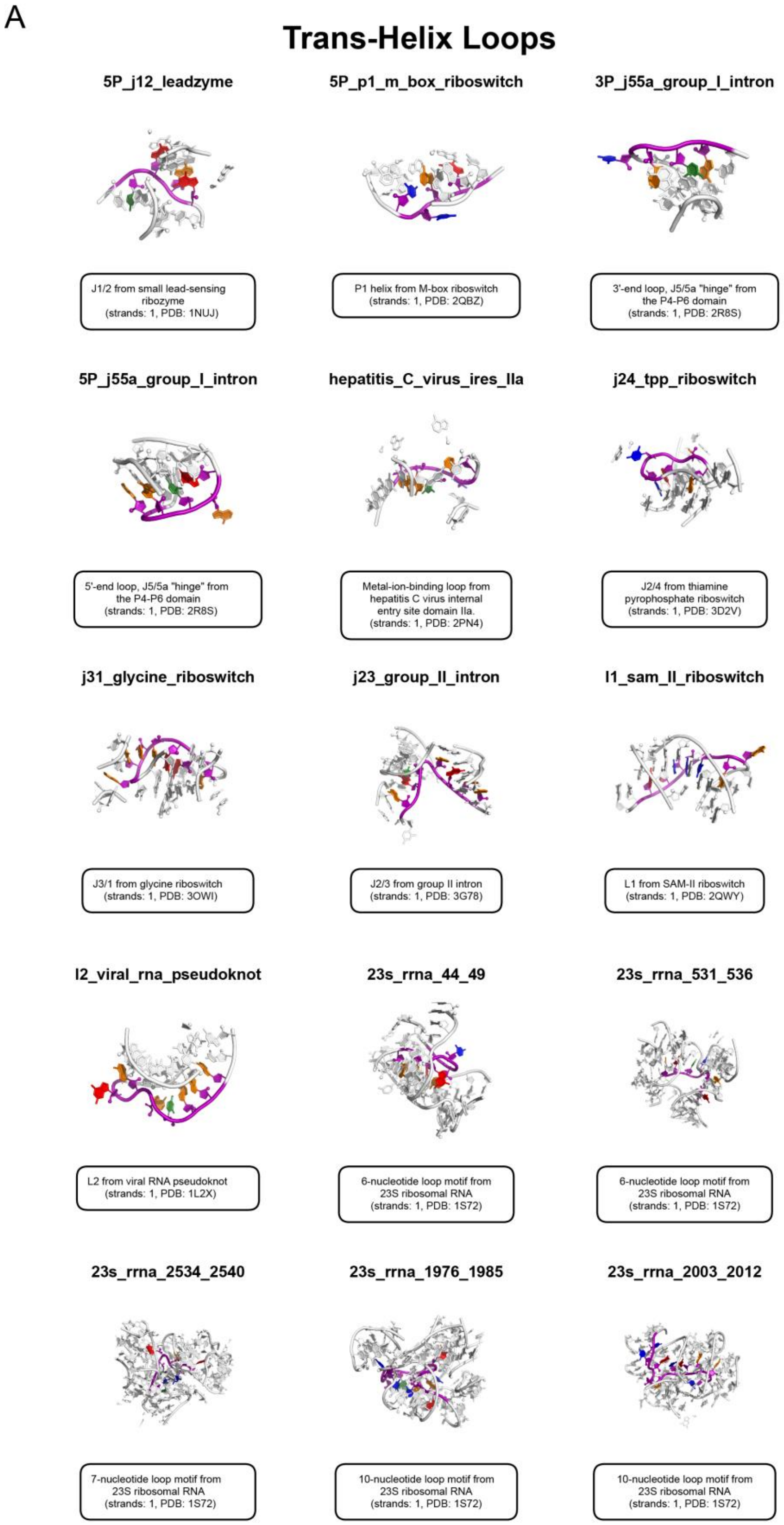
Energy gaps to optimized experimental structures

To assess efficiency of conformational sampling, the all-atom Rosetta energies of crystallographic loops needed to be computed, for comparison to *ab initio* model energies. Generally, experimentally solved RNA structures contain minor steric clashes that are penalized by the Rosetta energy function, and these conformations need to be subjected to local optimization to permit comparison to *de novo* models, with the same bond lengths and angles as used in the modeling (*49*). Such models were obtained using SWM 'native screen' calculations. The same stepwise Monte Carlo command-lines were run but with the -rmsd_screen 2.0 flag, in which coordinate constraints on each atom's distance from its native position penalizing distances greater than 2.0 Å encourage deep sampling near the experimental structure. The runs also included the flag `-add_proposal_density_factor 10000`, which encourages residue additions, leading to faster trajectories. (Since the models are already restrained to be close to the experimental conformation, minimal deletion and resampling is necessary.) For the 15 trans-helix single-stranded loop tests, we also used a 'native SWA' strategy, in which an entire SWA calculation was run, but at each sampling step, models were only carried forward if their backbone RMSD to the crystallographic loop was less than 3.0 Å, following the procedure in ref. (*11*).

Sources of experimental PDB structures

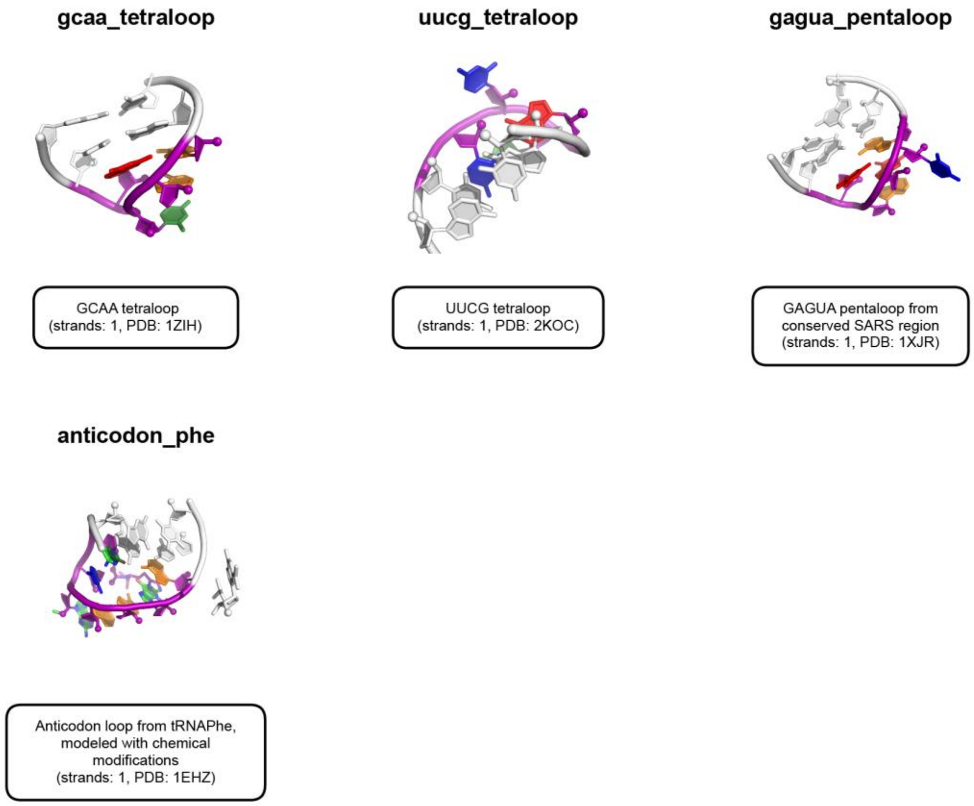
The sources of the experimental structures used in the RNA loop motif and complex multi-stranded motif benchmarks include: Rev response element high affinity site (strands: 2, PDB: 1CSL) (*50*); T loop motif from modified tRNAPhe (strands: 2, PDB: 1EHZ) (*51*); Chemically modified anticodon loop of tRNAPhe (strands: 1, PDB: 1EHZ) (*51*); G(syn)-G(anti) conformation of non-canonical guanosine-guanosine base pair (strands: 2, PDB: 1F5G) (*52*); J4/4a from P4-P6 domain (strands: 2, PDB: 1GID) (*53*); P5b from P4-P6 group I ribozyme domain (strands: 2, PDB: 1GID) (*53*); RNA quadruplex of UGAG, with U substituted with bromouridine (strands: 4, PDB: 1J6S) (*54*); L2 from viral RNA pseudoknot (strands: 1, PDB: 1L2X) (*55*); Conserved domain of human signal-recognition particle (strands: 2, PDB: 1LNT) (*56*); Imino conformation of tandem G-A base pair steps (strands: 2, PDB: 1MIS) (*57*); J1/2 from small lead-sensing ribozyme (strands: 1, PDB: 1NUJ) (*58*); bulged tetraplex (strands: 4, PDB: 1P79) (*59*); Bulged G motif from sarcin/ricin loop (strands: 2, PDB: 1Q9A) (*60*); 10-nucleotide loop motif from 23S ribosomal RNA (strands: 1, PDB: 1S72) (*61*); 6-nucleotide loop motif from 23S ribosomal RNA (strands: 1, PDB: 1S72) (*61*); 7-nucleotide loop motif from 23S ribosomal RNA (strands: 1, PDB: 1S72) (*61*); GAGUA pentaloop from conserved SARS region (strands: 1, PDB: 1XJR) (*62*); L2/L3 from A-riboswitch-adenine complex (strands: 2, PDB: 1Y26) (*63*); Pseudoknot docking interaction A-riboswitch-adenine complex (strands: 2, PDB: 1Y26) (*63*); Sheared conformation of tandem G-A base pair steps (strands: 2, PDB: 1YFV) (*64*, *65*); GCAA tetraloop (strands: 1, PDB: 1ZIH) (*65*); Z-form RNA helix, comprised of C-G base pair steps (strands: 2, PDB: 2ACJ) (*66*); Kink-turn motif derived from SAM-I riboswitch (strands: 2, PDB: 2GIS) (*67*); RNA quadruplex from an inosine-tetrad (strands: 4, PDB: 2GRB) (*68*); UUCG tetraloop (strands: 1, PDB: 2KOC) (*69*); 4-by-4 nucleotide RNA internal loop from an R2 retrotransposon (strands: 2, PDB: 2L8F) (*70*); Major conformation of internal loop from RNA structural switch (strands: 2, PDB: 2LX1) (*71*); Catalytic-like conformation, tertiary interaction in hammerhead ribozyme (strands: 3, PDB: 2OEU) (*72*); Catalytic-like conformation, three-way-junction in hammerhead ribozyme (strands: 3, PDB: 2OEU) (*72*, *73*); Tetraloop-helix interaction in L1 ligase crystal (strands: 1, PDB: 2OIU) (*73*); Metal-ion-binding loop from hepatitis C virus internal entry site domain IIa (strands: 1, PDB: 2PN4) (*74*); P1 helix from M-box riboswitch (strands: 1, PDB: 2QBZ) (*75*); L1 from SAM-II riboswitch (strands: 1, PDB: 2QWY) (*76*); 3'-end loop, J5/5a "hinge" from the P4-P6 domain (strands: 1, PDB: 2R8S) (*77*); 5'-end loop, J5/5a "hinge" from the P4-P6 domain (strands: 1, PDB: 2R8S) (*77*); Canonical GAAA:11-nt tetraloop-receptor module in P4-P6 domain (strands: 3, PDB: 2R8S) (*77*); J5/5a "hinge" from the P4-P6 domain (strands: 2, PDB: 2R8S) (*77*); Receptor motif of canonical 11-nt tetraloop-receptor module in P4-P6 domain (strands: 2, PDB: 2R8S) (*77*, *78*); Loop E motif (strands: 2, PDB: 354D) (*78*); Pre-catalytic conformation, three-way-junction in hammerhead ribozyme (strands: 3, PDB: 359D) (*79*); J2/4 from thiamine pyrophosphate riboswitch (strands: 1, PDB: 3D2V) (*80*); J2/3 from group II intron (strands: 1, PDB: 3G78) (*81*); T loop motif from unmodified tRNAPhe (strands: 2, PDB: 3L0U) (*82*); Alternative conformation of a non-canonical junction from a human thymidylate synthase regulatory motif (strands: 2, PDB: 3MEI) (*83*); Non-canonical junction from a human thymidylate synthase regulatory motif (strands: 2, PDB: 3MEI) (*83*); J3/1 from glycine riboswitch (strands: 1, PDB: 3OWI) (*84*); parallel poly-A helix (strands: 2, PDB: 4JRD) (*85*); P2.1/P5 "kissing" interaction from GIR1 lariat-capping ribozyme (strands: 2, PDB: 4P8Z) (*86*); P2/P9 GAAA docking interaction from GIR1 lariat-capping ribozyme (strands: 3, PDB: 4P8Z) (*86*); VS ribozyme three-way-junction between P2, P3, and P6 (strands: 3, PDB: 4R4V) (*87*); VS ribozyme three-way-junction between P3, P4, and P5 (strands: 3, PDB: 4R4V) (*87*); xrRNA pseudoknot loop L3/S4 (strands: 3, PDB: 5TPY) (*27*); xrRNA three-way-junction (strands: 3, PDB: 5TPY) (*27*). See fig. S1 for illustrated descriptions and modeling constraints.

Supplementary Figures


A
Trans-Helix Loops
5P_j12_leadzyme
J1/2 from small lead-sensing ribozyme (strands: 1, PDB: 1NUJ)
5P_p1_m_box_riboswitch
P1 helix from M-box riboswitch (strands: 1, PDB: 2QBZ)
3P_j55a_group_I_intron
3'-end loop, J5/5a "hinge" from the P4-P6 domain (strands: 1, PDB: 2R8S)
5P_j55a_group_I_intron
5'-end loop, J5/5a "hinge" from the P4-P6 domain (strands: 1, PDB: 2R8S)
hepatitis_C_virus_ires_IIa
Metal-ion-binding loop from hepatitis C virus internal entry site domain IIa. (strands: 1, PDB: 2PN4)
j24_tpp_riboswitch
J2/4 from thiamine pyrophosphate riboswitch (strands: 1, PDB: 3D2V)
j31_glycine_riboswitch
J3/1 from glycine riboswitch (strands: 1, PDB: 3OWI)
j23_group_II_intron
J2/3 from group II intron (strands: 1, PDB: 3G78)
l1_sam_II_riboswitch
L1 from SAM-II riboswitch (strands: 1, PDB: 2QWY)
l2_viral_rna_pseudoknot
L2 from viral RNA pseudoknot (strands: 1, PDB: 1L2X)
23s_rrna_44_49
6-nucleotide loop motif from 23S ribosomal RNA (strands: 1, PDB: 1S72)
23s_rrna_531_536
6-nucleotide loop motif from 23S ribosomal RNA (strands: 1, PDB: 1S72)
23s_rrna_2534_2540
7-nucleotide loop motif from 23S ribosomal RNA (strands: 1, PDB: 1S72)
23s_rrna_1976_1985
10-nucleotide loop motif from 23S ribosomal RNA (strands: 1, PDB: 1S72)
23s_rrna_2003_2012
10-nucleotide loop motif from 23S ribosomal RNA (strands: 1, PDB: 1S72)

B

Apical Loops



gcaa_tetraloop

GCAA tetraloop
(strands: 1, PDB: 1ZIH)

uucg_tetraloop

UUCG tetraloop
(strands: 1, PDB: 2KOC)

gagua_pentaloop

GAGUA pentaloop from
conserved SARS region
(strands: 1, PDB: 1XJR)

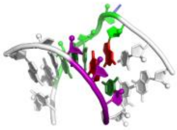
anticodon_phe

Anticodon loop from tRNAPhe,
modeled with chemical
modifications
(strands: 1, PDB: 1EHZ)

C

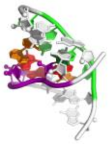
Fixed Two-Way Junctions

puzzle1_alt_fixed



Alternative conformation of a non-canonical junction from a human thymidylate synthase regulatory motif (strands: 2, PDB: 3MEI)

srp_domainIV_fixed

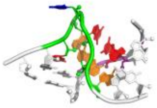


Conserved domain of human signal-recognition particle (strands: 2, PDB: 1LNT)

srl_fixed

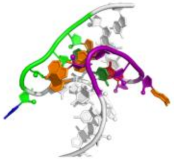
Bulged G motif from sarcin/ricin loop (strands: 2, PDB: 1Q9A)

kink_turn_fixed



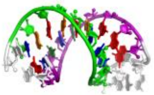
Kink-turn motif derived from SAM-I riboswitch (strands: 2, PDB: 2GIS)

j55a_P4P6_fixed



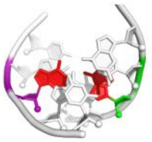
J5/5a "hinge" from the P4-P6 domain (strands: 2, PDB: 2R8S)

P5b_connect



P5b from P4-P6 group I ribozyme domain (strands: 2, PDB: 1GID)

gg_mismatch_fixed



G(syn)-G(anti) conformation of non-canonical guanosine-guanosine base pair (strands: 2, PDB: 1F5G)

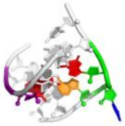
tandem_ga_imino_fixed

Imino conformation of tandem G-A base pair steps (strands: 2, PDB: 1MIS)

tandem_ga_sheared_fixed

Sheared conformation of tandem G-A base pair steps (strands: 2, PDB: 1YFV)

hiv_rre_fixed

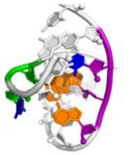


Rev response element high affinity site (strands: 2, PDB: 1CSL)

j44a_p4p6_fixed

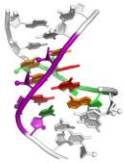
J4/4a from P4-P6 domain (strands: 2, PDB: 1GID)

just_tr_P4P6_fixed



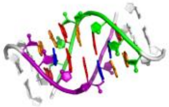
Receptor motif of canonical 11-nt tetraloop-receptor module in P4-P6 domain (strands: 2, PDB: 2R8S)

r2_4x4_fixed



4-by-4 nucleotide RNA internal loop from an R2 retrotransposon (strands: 2, PDB: 2L8F)

loopE_fixed

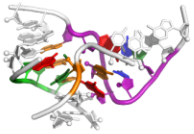


Loop E motif (strands: 2, PDB: 354D)

D

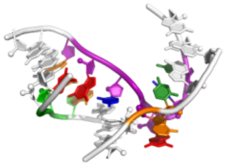
Fixed Three-Way Junctions

hammerhead_3WJ_cat_fixed



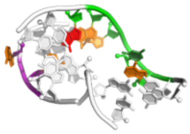
Catalytic-like conformation, three-way-junction in hammerhead ribozyme (strands: 3, PDB: 2OEU)

hammerhead_3WJ_precat_fixed



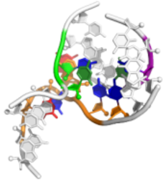
Pre-catalytic conformation, three-way-junction in hammerhead ribozyme (strands: 3, PDB: 359D)

VS_rbzm_P2P3P6_fixed



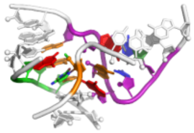
VS ribozyme three-way-junction between P2, P3, and P6 (strands: 3, PDB: 4R4V)

VS_rbzm_P3P4P5_fixed

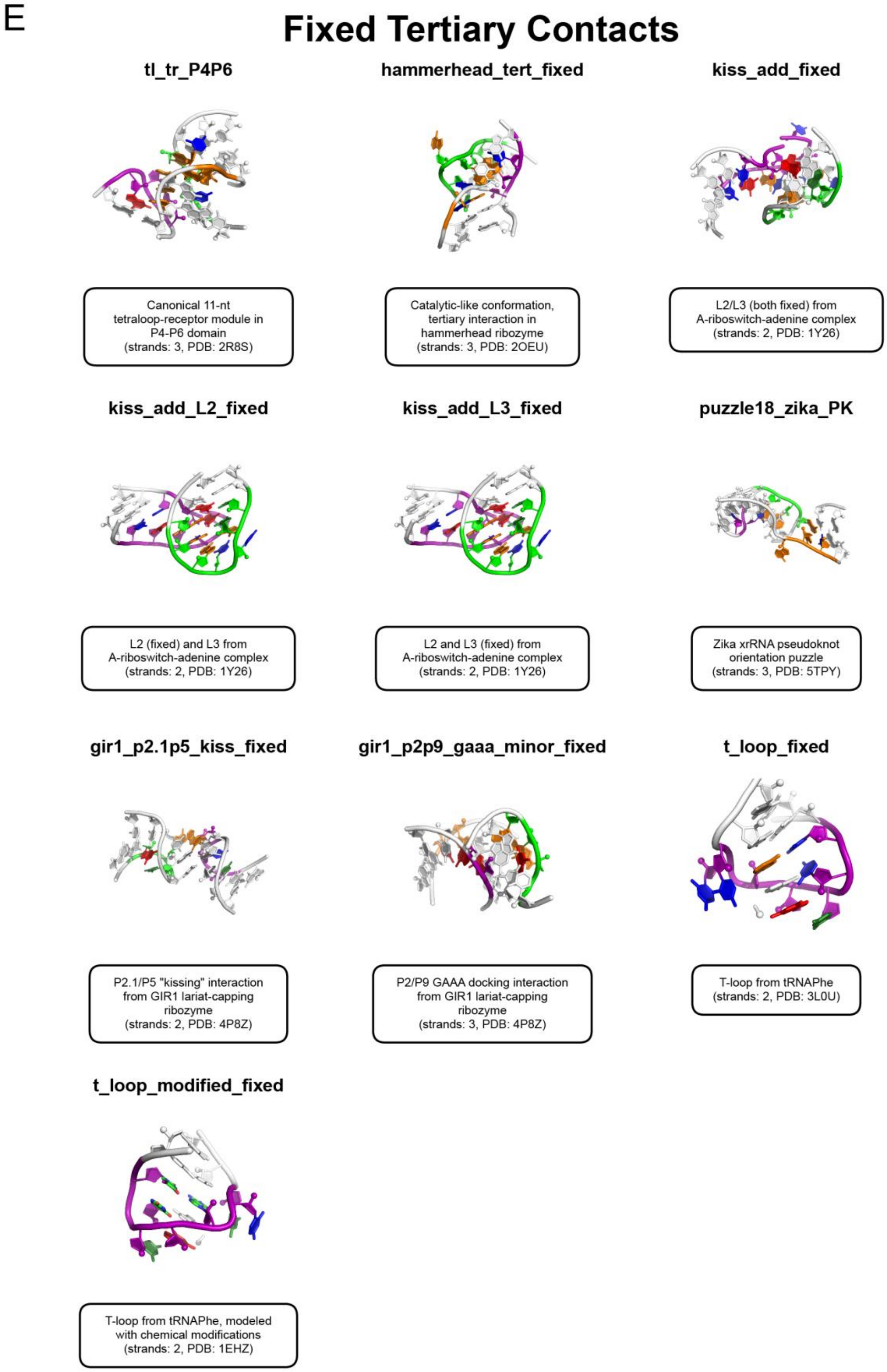


VS ribozyme three-way-junction between P3, P4, and P5 (strands: 3, PDB: 4R4V)

hammerhead_3WJ_cat_OMC_fixed



Catalytic-like conformation, three-way-junction in hammerhead ribozyme, with O-methyl cytosine (strands: 3, PDB: 2OEU)



E

Fixed Tertiary Contacts

tl_tr_P4P6

Canonical 11-nt tetraloop-receptor module in P4-P6 domain (strands: 3, PDB: 2R8S)

hammerhead_tert_fixed

Catalytic-like conformation, tertiary interaction in hammerhead ribozyme (strands: 3, PDB: 2OEU)

kiss_add_fixed

L2/L3 (both fixed) from A-riboswitch-adenine complex (strands: 2, PDB: 1Y26)

kiss_add_L2_fixed

L2 (fixed) and L3 from A-riboswitch-adenine complex (strands: 2, PDB: 1Y26)

kiss_add_L3_fixed

L2 and L3 (fixed) from A-riboswitch-adenine complex (strands: 2, PDB: 1Y26)

puzzle18_zika_PK

Zika xrRNA pseudoknot orientation puzzle (strands: 3, PDB: 5TPY)

gir1_p2.1p5_kiss_fixed

P2.1/P5 "kissing" interaction from GIR1 lariat-capping ribozyme (strands: 2, PDB: 4P8Z)

gir1_p2p9_gaaa_minor_fixed

P2/P9 GAAA docking interaction from GIR1 lariat-capping ribozyme (strands: 3, PDB: 4P8Z)

t_loop_fixed

T-loop from tRNAPhe (strands: 2, PDB: 3L0U)

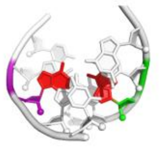
t_loop_modified_fixed

T-loop from tRNAPhe, modeled with chemical modifications (strands: 2, PDB: 1EHZ)

F

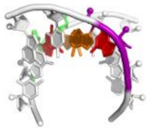
Aligned Two-Way Junctions

gg_mismatch



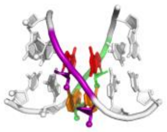
G(syn)-G(anti) conformation of non-canonical guanosine-guanosine base pair (strands: 2, PDB: 1F5G)

tandem_ga_imino



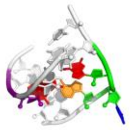
Imino conformation of tandem G-A base pair steps (strands: 2, PDB: 1MIS)

tandem_ga_sheared



Sheared conformation of tandem G-A base pair steps (strands: 2, PDB: 1YFV)

hiv_rre

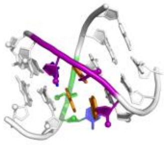


Rev response element high affinity site (strands: 2, PDB: 1CSL)

j44a_p4p6

J4/4a from P4-P6 domain (strands: 2, PDB: 1GID)

just_tr_P4P6

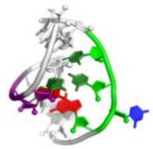


Receptor motif of canonical 11-nt tetraloop-receptor module in P4-P6 domain (strands: 2, PDB: 2R8S)

cg_helix

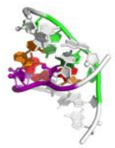
Z-form RNA helix, comprised of C-G base pair steps (strands: 2, PDB: n/a)

puzzle1



Non-canonical junction from a human thymidylate synthase regulatory motif (strands: 2, PDB: 3MEI)

srp_domainIV

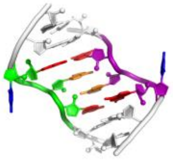


Conserved domain of human signal-recognition particle (strands: 2, PDB: 1LNT)

r2_4x4

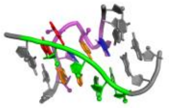
4-by-4 nucleotide RNA internal loop from an R2 retrotransposon (strands: 2, PDB: 2L8F)

gagu_forcesyn_blockstackU



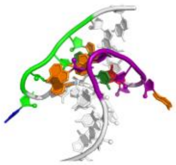
Major conformation of internal loop from RNA structural switch, with stacking blocked on U (strands: 2, PDB: 2LX1)

srl_free_bulgedG



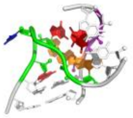
Bulged G motif from sarcin/ricin loop (strands: 2, PDB: 1Q9A)

j55a_P4P6_align



J5/5a "hinge" from the P4-P6 domain (strands: 2, PDB: 2R8S)

kink_turn_align



Kink-turn motif derived from SAM-I riboswitch (strands: 2, PDB: 2GIS)

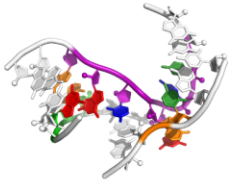
loopE

Loop E motif (strands: 2, PDB: 354D)

G

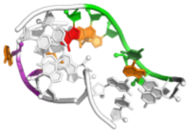
Aligned Three-Way Junctions

hammerhead_3WJ_precat



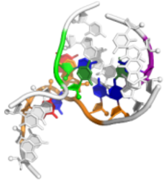
Pre-catalytic conformation, three-way-junction in hammerhead ribozyme (strands: 3, PDB: 359D)

VS_rbzm_P2P3P6_align



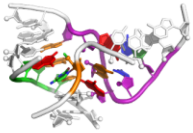
VS ribozyme three-way-junction between P2, P3, and P6 (strands: 3, PDB: 4R4V)

VS_rbzm_P3P4P5_align



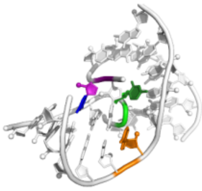
VS ribozyme three-way-junction between P3, P4, and P5 (strands: 3, PDB: 4R4V)

hammerhead_3WJ_cat_OMC_align

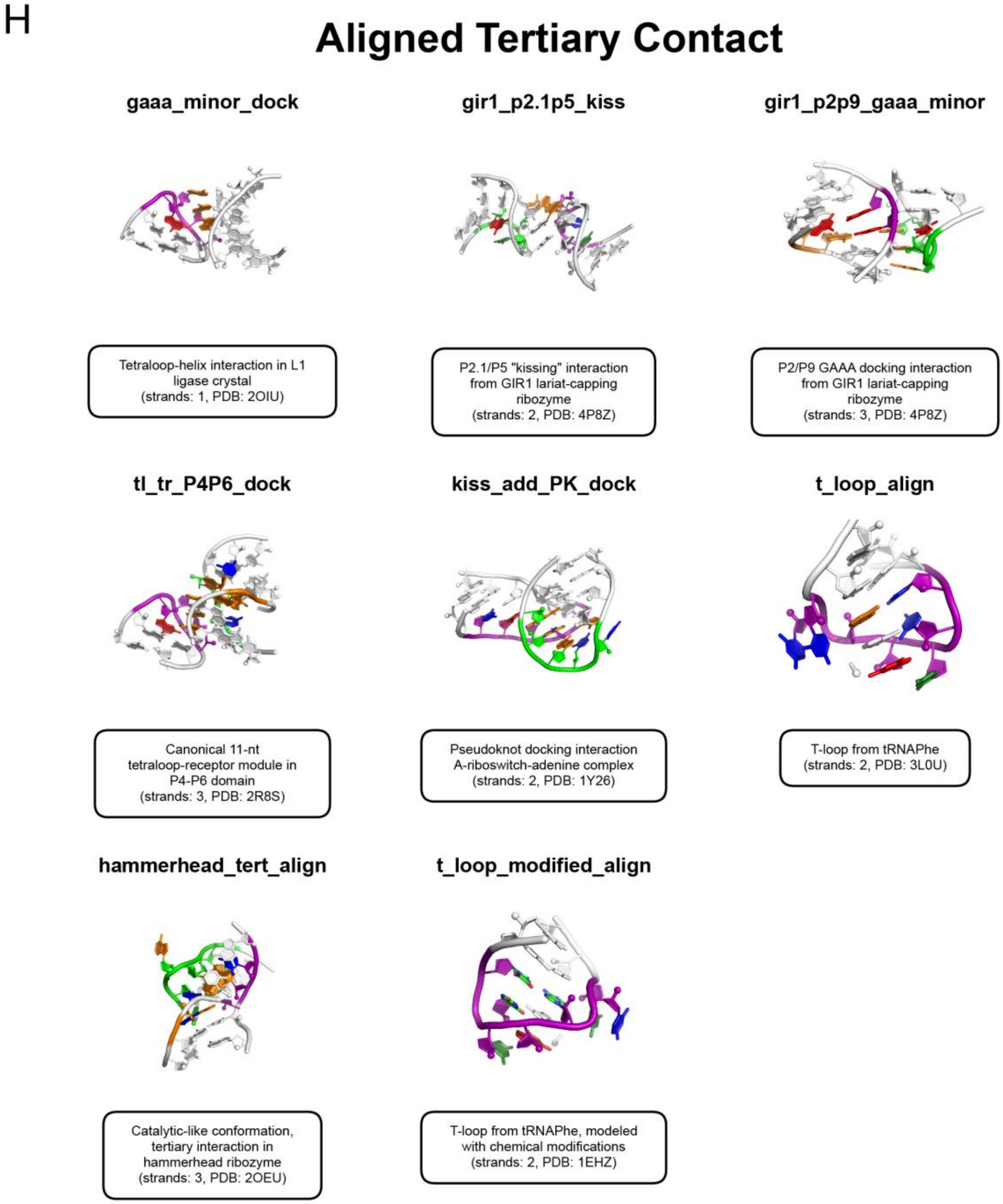


Catalytic-like conformation, three-way-junction in hammerhead ribozyme, with O-methyl cytosine (strands: 3, PDB: 2OEU)

puzzle18_zika_3WJ



A three-way-junction from the Zika xrRNA (strands: 3, PDB: 5TPY)



Aligned Tertiary Contact

gaaa_minor_dock

Tetraloop-helix interaction in L1 ligase crystal (strands: 1, PDB: 2OIU)

gir1_p2.1p5_kiss

P2.1/P5 "kissing" interaction from GIR1 lariat-capping ribozyme (strands: 2, PDB: 4P8Z)

gir1_p2p9_gaaa_minor

P2/P9 GAAA docking interaction from GIR1 lariat-capping ribozyme (strands: 3, PDB: 4P8Z)

tl_tr_P4P6_dock

Canonical 11-nt tetraloop-receptor module in P4-P6 domain (strands: 3, PDB: 2R8S)

kiss_add_PK_dock

Pseudoknot docking interaction A-riboswitch-adenine complex (strands: 2, PDB: 1Y26)

t_loop_align

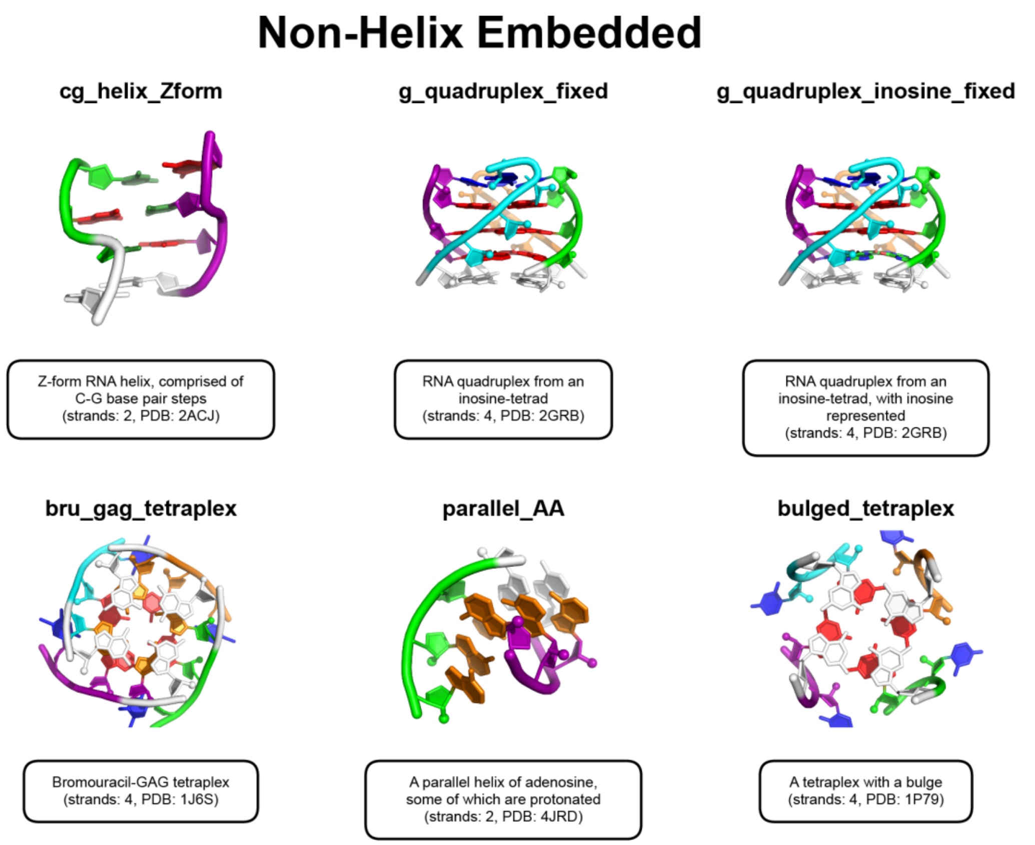
T-loop from tRNAPhe (strands: 2, PDB: 3L0U)

hammerhead_tert_align

Catalytic-like conformation, tertiary interaction in hammerhead ribozyme (strands: 3, PDB: 2OEU)

t_loop_modified_align

T-loop from tRNAPhe, modeled with chemical modifications (strands: 2, PDB: 1EHZ)



Non-Helix Embedded

cg_helix_Zform

Z-form RNA helix, comprised of C-G base pair steps (strands: 2, PDB: 2ACJ)

g_quadruplex_fixed

RNA quadruplex from an inosine-tetrad (strands: 4, PDB: 2GRB)

g_quadruplex_inosine_fixed

RNA quadruplex from an inosine-tetrad, with inosine represented (strands: 4, PDB: 2GRB)

bru_gag_tetraplex

Bromouracil-GAG tetraplex (strands: 4, PDB: 1J6S)

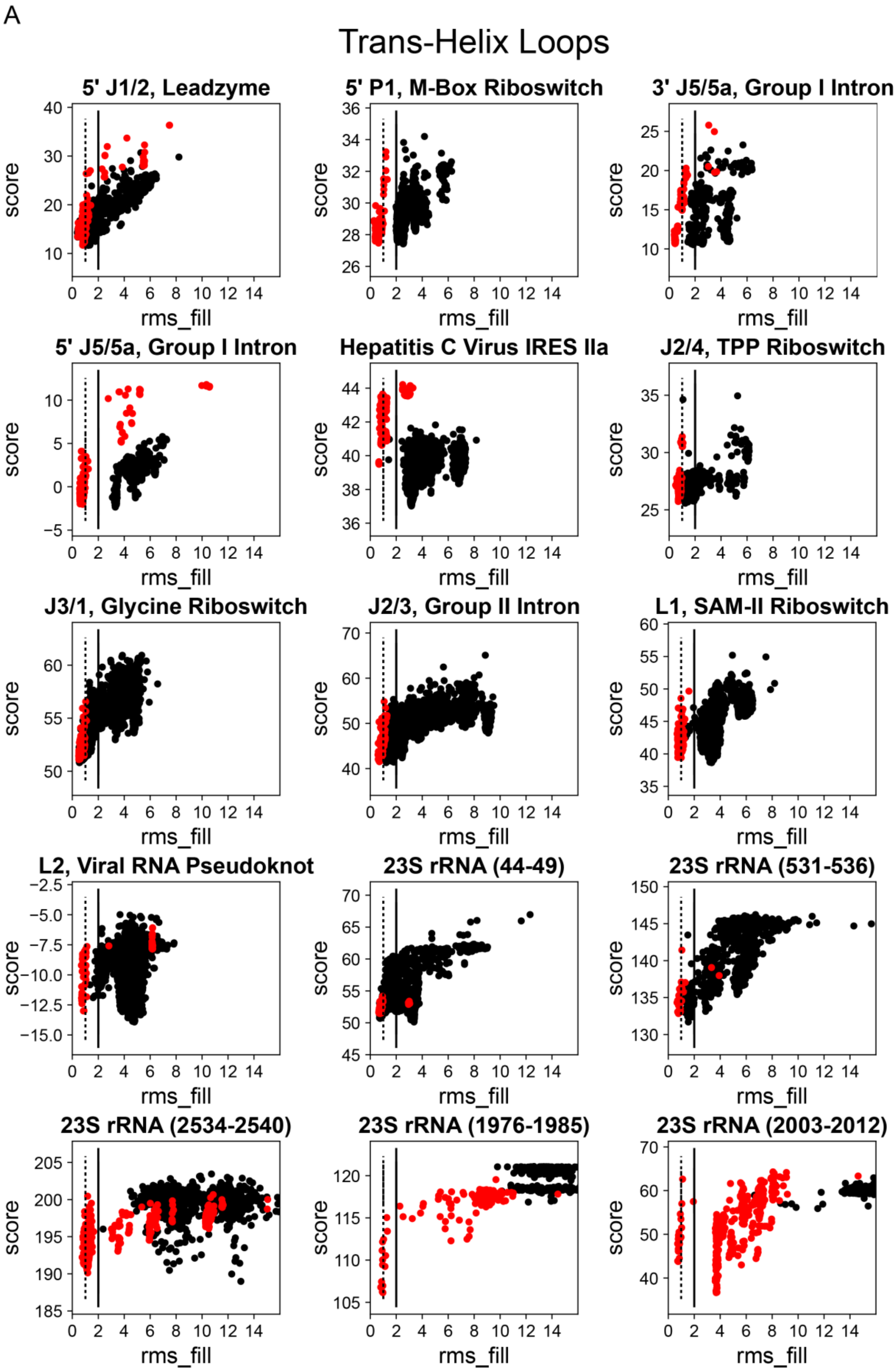
parallel_AA

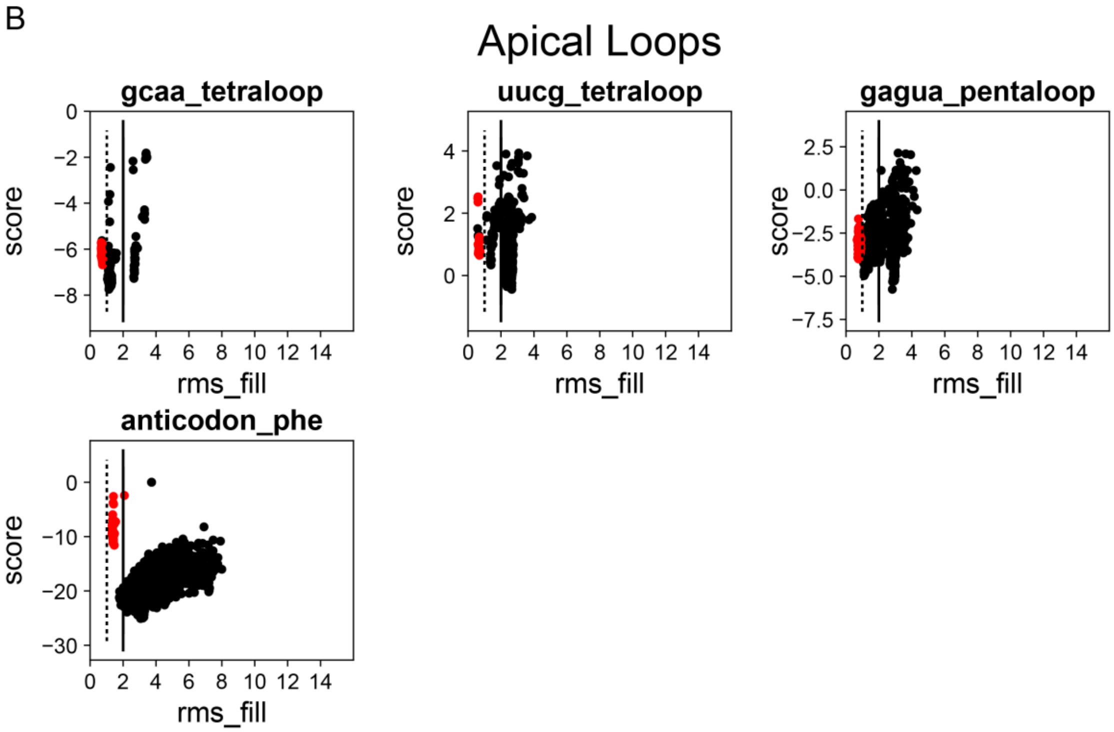
A parallel helix of adenosine, some of which are protonated (strands: 2, PDB: 4JRD)

bulged_tetraplex

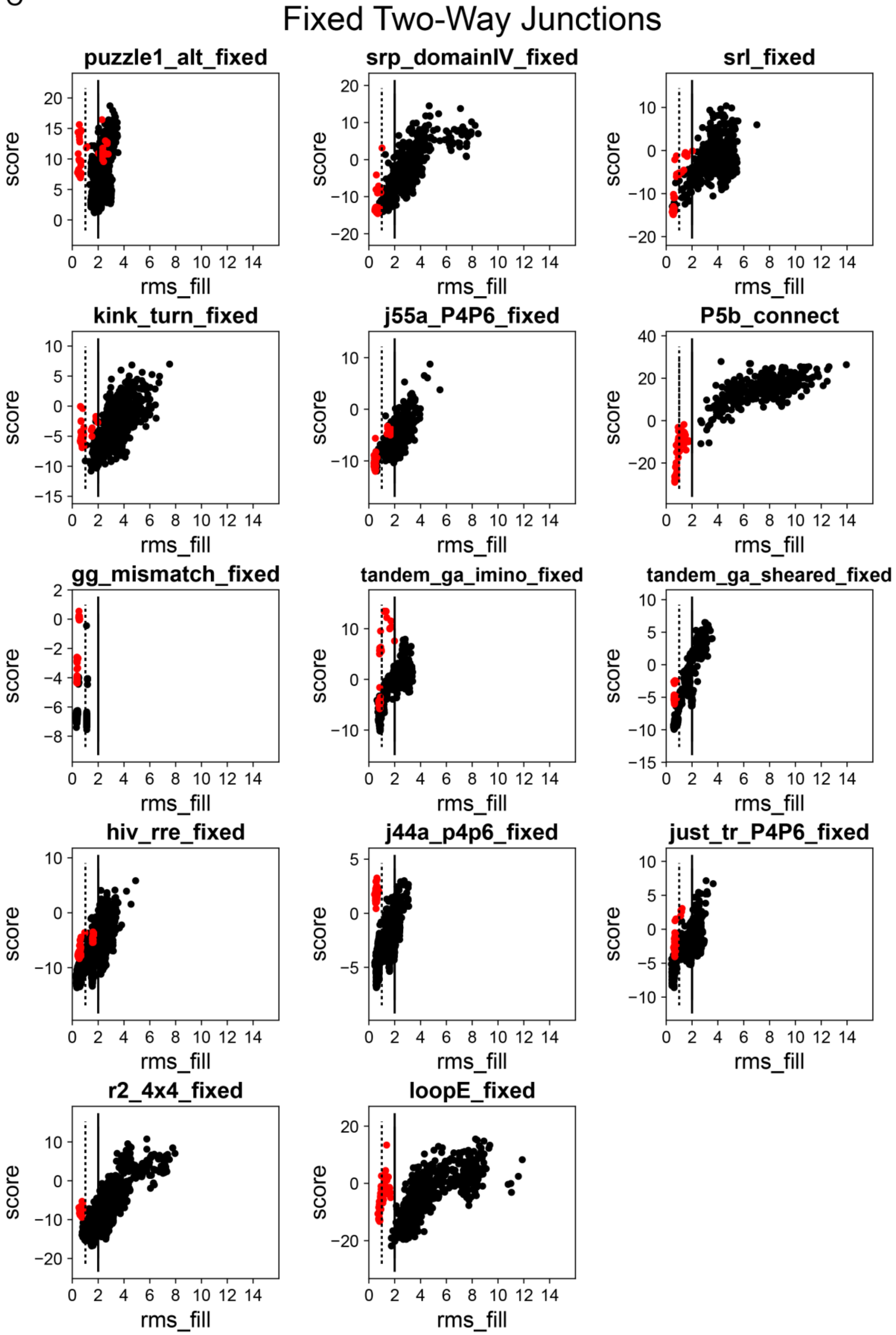
A tetraplex with a bulge (strands: 4, PDB: 1P79)

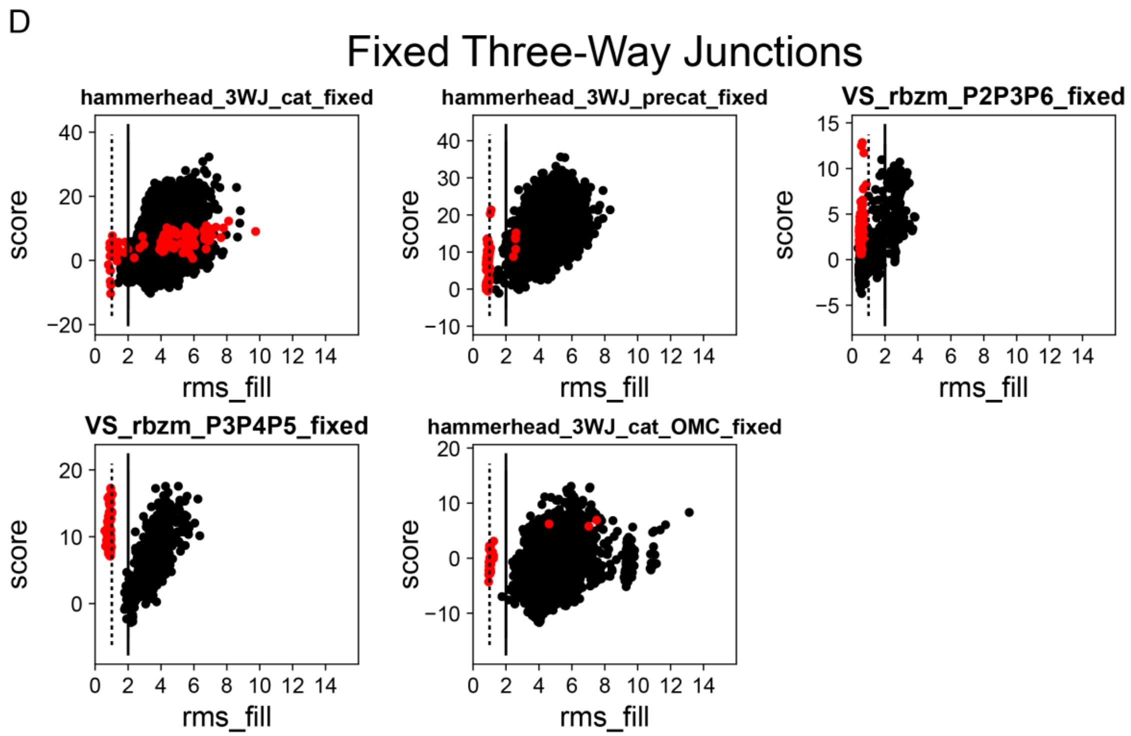
fig. S1. Illustrated descriptions and modeling constraints of all 82 benchmark test cases. Tertiary structure graphics were generated in PyMOL using RiboVis, a PyMOL rendering package developed in-house (see: https://ribokit.stanford.edu/RiboVis) In the tertiary structure graphics, the backbone of sampled nucleotides is colored by chain, while sidechains are colored by identity with A (gold), C (green), G (red), and U (blue); white coloring indicates structural templates supplied as input and held fixed during modeling. Gray coloring indicates structural templates supplied as input but allowed to move relative to other input structures. *Figure is continued on following 8 pages.*

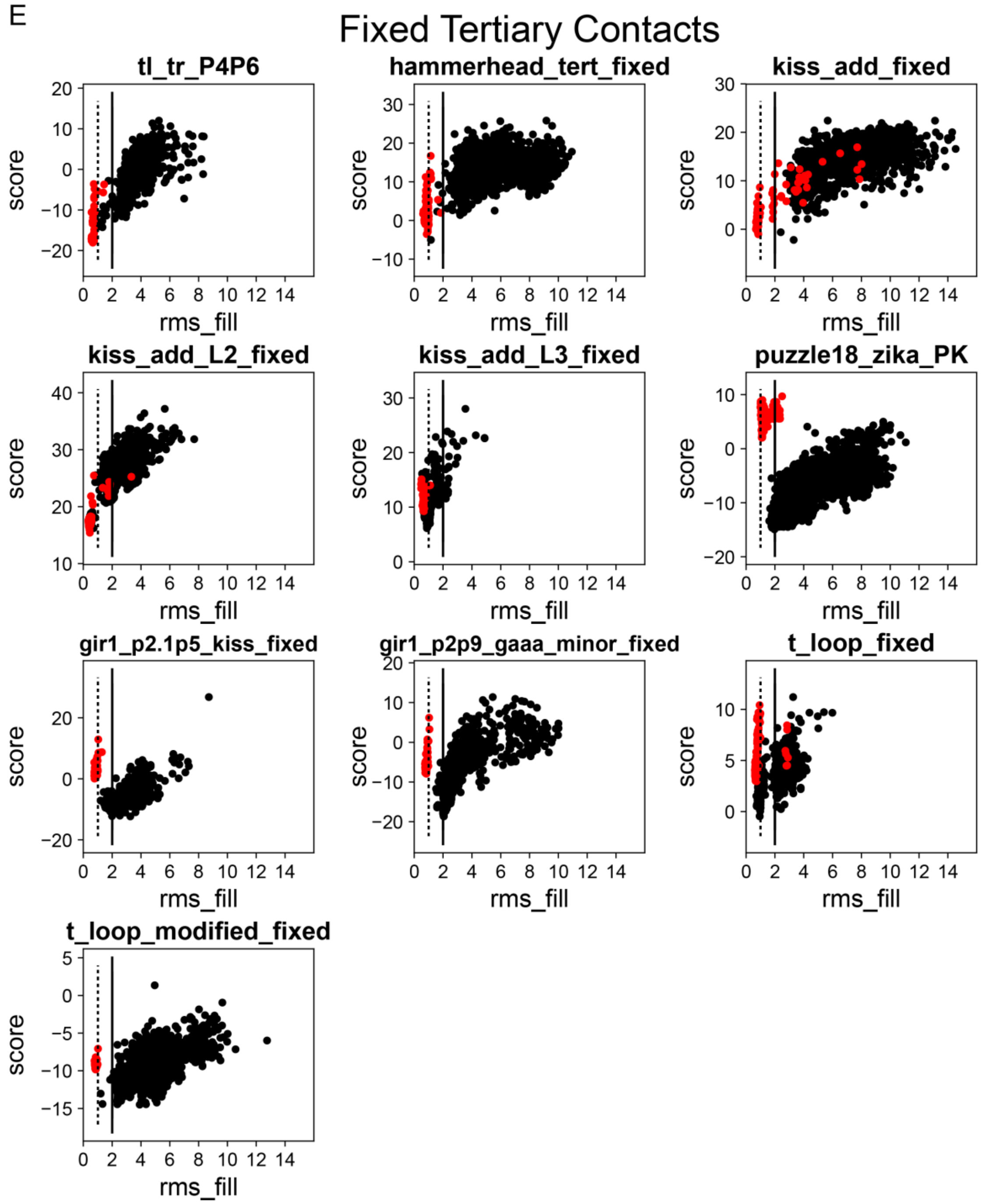

A
Trans-Helix Loops
5' J1/2, Leadzyme
score
10
20
30
40
rms_fill
0 2 4 6 8 10 12 14
5' P1, M-Box Riboswitch
score
26
28
30
32
34
36
rms_fill
0 2 4 6 8 10 12 14
3' J5/5a, Group I Intron
score
10
15
20
25
rms_fill
0 2 4 6 8 10 12 14
5' J5/5a, Group I Intron
score
−5
0
5
10
rms_fill
0 2 4 6 8 10 12 14
Hepatitis C Virus IRES IIa
score
36
38
40
42
44
rms_fill
0 2 4 6 8 10 12 14
J2/4, TPP Riboswitch
score
25
30
35
rms_fill
0 2 4 6 8 10 12 14
J3/1, Glycine Riboswitch
score
50
55
60
rms_fill
0 2 4 6 8 10 12 14
J2/3, Group II Intron
score
40
50
60
70
rms_fill
0 2 4 6 8 10 12 14
L1, SAM-II Riboswitch
score
35
40
45
50
55
60
rms_fill
0 2 4 6 8 10 12 14
L2, Viral RNA Pseudoknot
score
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
rms_fill
0 2 4 6 8 10 12 14
23S rRNA (44-49)
score
45
50
55
60
65
70
rms_fill
0 2 4 6 8 10 12 14
23S rRNA (531-536)
score
130
135
140
145
150
rms_fill
0 2 4 6 8 10 12 14
23S rRNA (2534-2540)
score
185
190
195
200
205
rms_fill
0 2 4 6 8 10 12 14
23S rRNA (1976-1985)
score
105
110
115
120
rms_fill
0 2 4 6 8 10 12 14
23S rRNA (2003-2012)
score
40
50
60
70
rms_fill
0 2 4 6 8 10 12 14


B
Apical Loops
gcaa_tetraloop
score
0
−2
−4
−6
−8
rms_fill
0 2 4 6 8 10 12 14
uucg_tetraloop
score
4
2
0
rms_fill
0 2 4 6 8 10 12 14
gagua_pentaloop
score
2.5
0.0
−2.5
−5.0
−7.5
rms_fill
0 2 4 6 8 10 12 14
anticodon_phe
score
0
−10
−20
−30
rms_fill
0 2 4 6 8 10 12 14

C

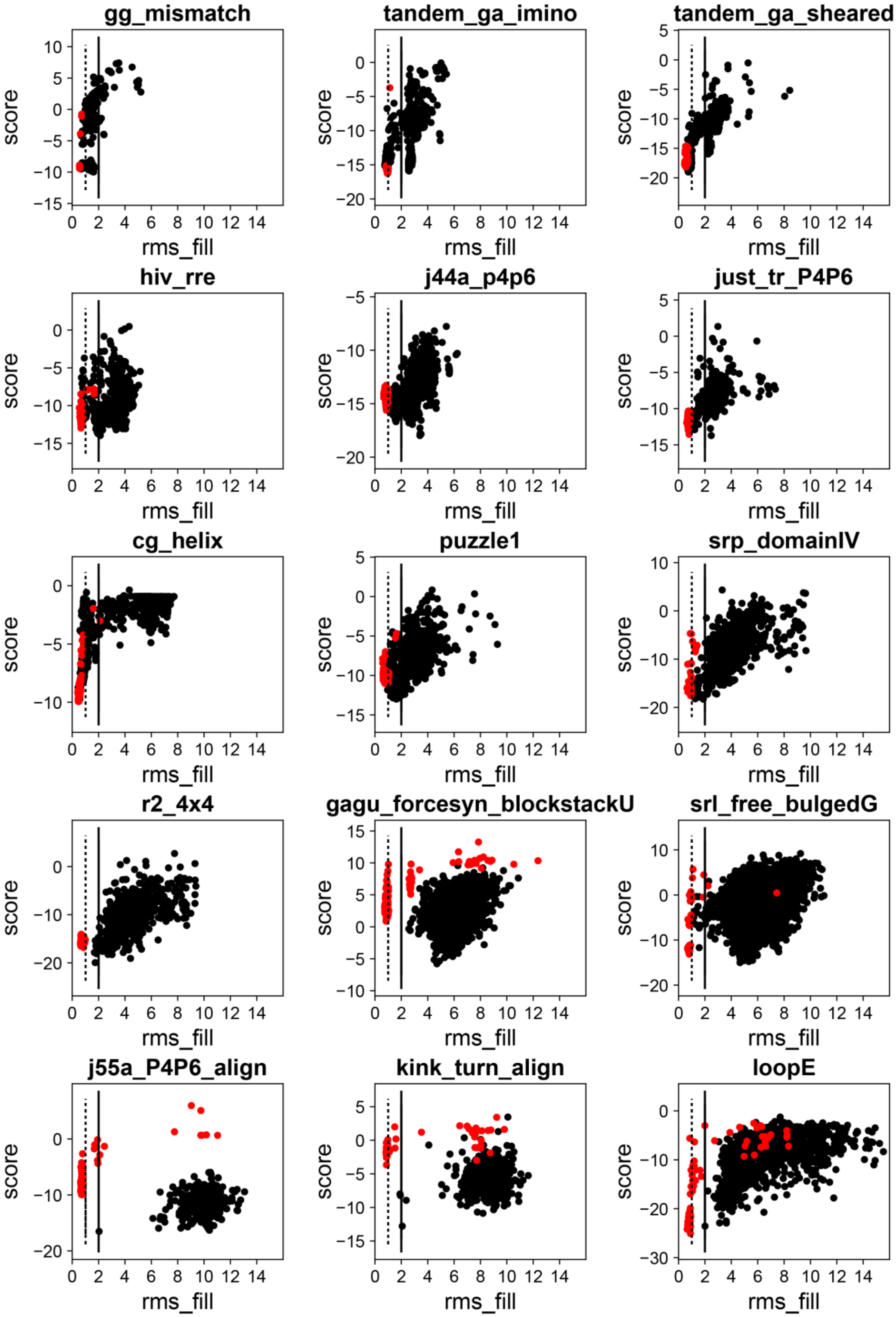

Fixed Two-Way Junctions
puzzle1_alt_fixed
srp_domainIV_fixed
srl_fixed
kink_turn_fixed
j55a_P4P6_fixed
P5b_connect
gg_mismatch_fixed
tandem_ga_imino_fixed
tandem_ga_sheared_fixed
hiv_rre_fixed
j44a_p4p6_fixed
just_tr_P4P6_fixed
r2_4x4_fixed
loopE_fixed
score
rms_fill
0 2 4 6 8 10 12 14

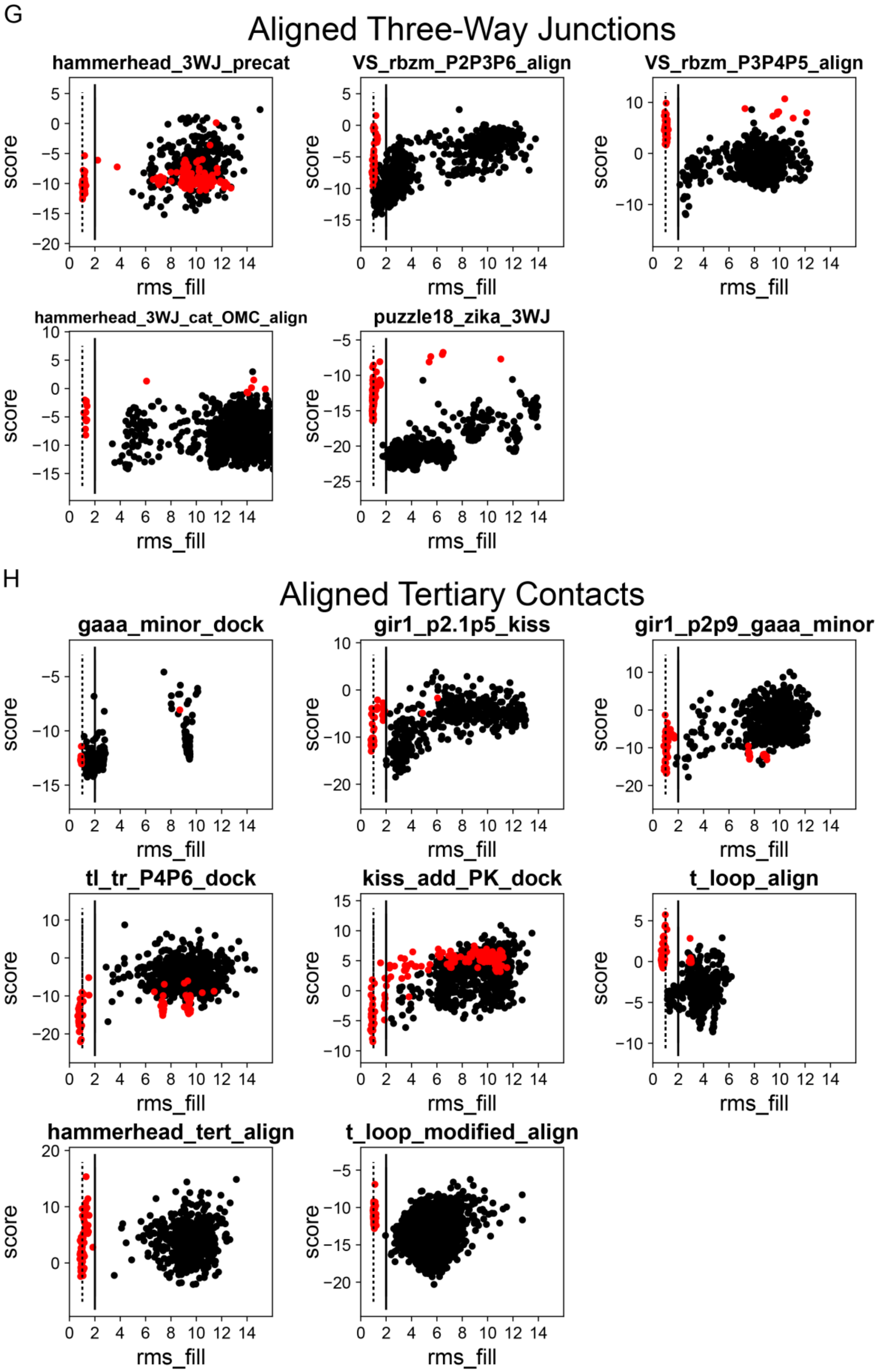

D
Fixed Three-Way Junctions
hammerhead_3WJ_cat_fixed
score
40
20
0
−20
rms_fill
0 2 4 6 8 10 12 14
hammerhead_3WJ_precat_fixed
score
40
30
20
10
0
−10
rms_fill
0 2 4 6 8 10 12 14
VS_rbzm_P2P3P6_fixed
score
15
10
5
0
−5
rms_fill
0 2 4 6 8 10 12 14
VS_rbzm_P3P4P5_fixed
score
20
10
0
rms_fill
0 2 4 6 8 10 12 14
hammerhead_3WJ_cat_OMC_fixed
score
20
10
0
−10
rms_fill
0 2 4 6 8 10 12 14


E
Fixed Tertiary Contacts
tl_tr_P4P6
hammerhead_tert_fixed
kiss_add_fixed
kiss_add_L2_fixed
kiss_add_L3_fixed
puzzle18_zika_PK
gir1_p2.1p5_kiss_fixed
gir1_p2p9_gaaa_minor_fixed
t_loop_fixed
t_loop_modified_fixed
score
rms_fill
0 2 4 6 8 10 12 14
20 10 0 −10 −20
30 20 10 0 −10
30 20 10 0
40 30 20 10
30 20 10 0
10 0 −10 −20
20 0 −20
20 10 0 −10 −20
10 5 0
5 0 −5 −10 −15

F

Aligned Two-Way Junctions




G
Aligned Three-Way Junctions
hammerhead_3WJ_precat
VS_rbzm_P2P3P6_align
VS_rbzm_P3P4P5_align
hammerhead_3WJ_cat_OMC_align
puzzle18_zika_3WJ
H
Aligned Tertiary Contacts
gaaa_minor_dock
gir1_p2.1p5_kiss
gir1_p2p9_gaaa_minor
tl_tr_P4P6_dock
kiss_add_PK_dock
t_loop_align
hammerhead_tert_align
t_loop_modified_align
score
rms_fill
0
2
4
6
8
10
12
14

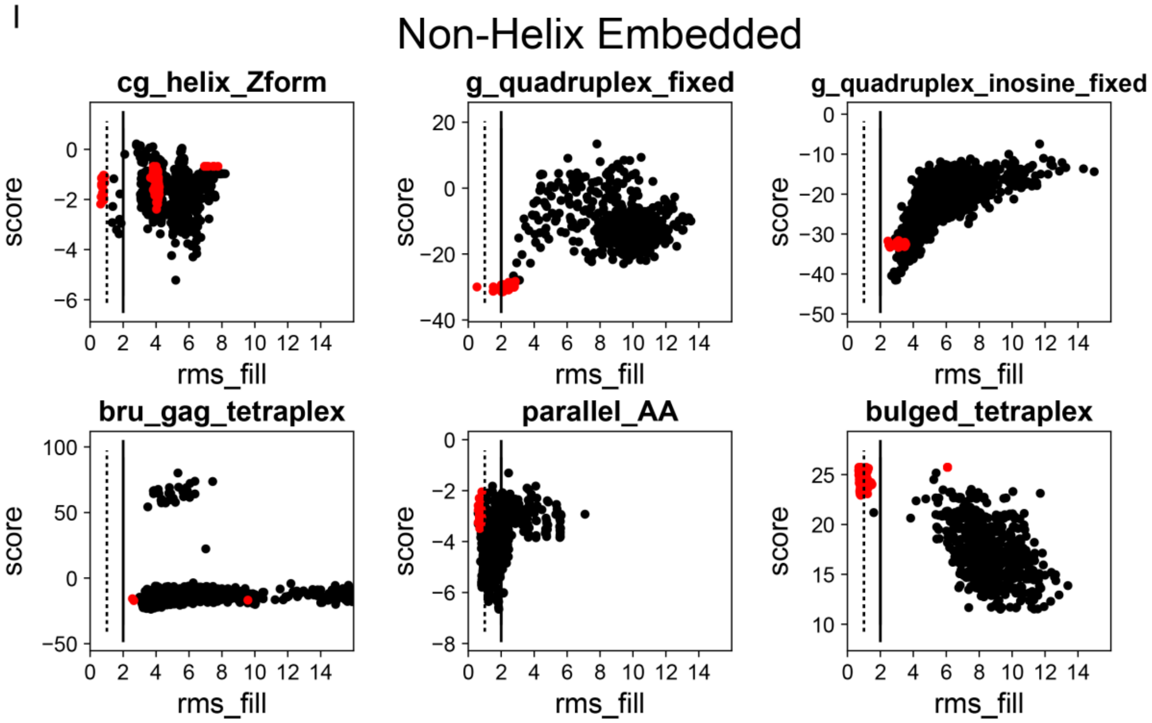


fig. S2. Rosetta free energy versus RMSD summaries of SWM modeling runs for 82 complex RNA motifs. Models generated *ab initio* by SWM (blue) compared to SWM models constructed using information about the native structure and restrained to be within 3.0 Å RMSD of the native conformation (red). Motif definitions are given in fig. S1. *Figure is continued on following 8 pages.*

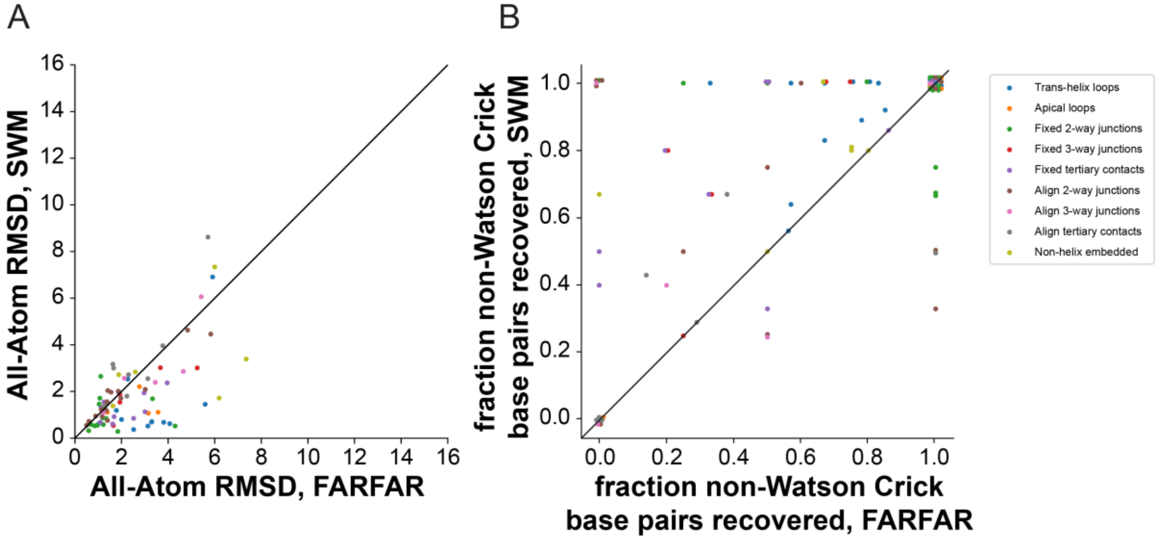


fig. S3. Comparison of model accuracy between SWM and fragment assembly of RNA with FARFAR over an 82 motif benchmark. (**A**) All-heavy-atom root-mean-squared deviation (RMSD) from experimental structure (lower is more accurate), and (**B**) fraction of non-Watson-Crick base pairs recovered (higher is more accurate). In (B), a small positive or negative random jitter value has been applied to points at 0.0 and 1.0, respectively, to resolve overlapping points.

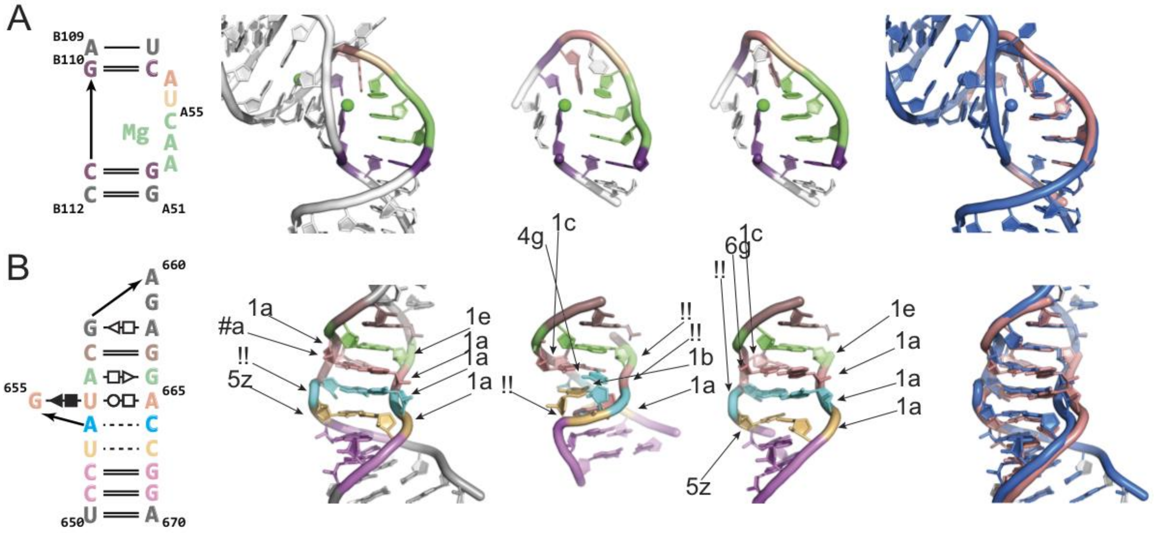


fig. S4. Potential routes to overcome limitations in Rosetta free energy function. Each panel gives, from left to right, secondary structure diagram, the experimental conformation, the best of five SWM cluster centers initially generated, model generated with scoring function or protocol improvements, overlay of native (marine) and improved model (salmon). (**A**) A five-nucleotide loop from the hepatitis C virus internal ribosomal entry site domain IIa that was also a problem case before (*1*) involves metal ion binding (left). Inclusion of the bound metal ion at its crystallographic position enables atomic-accuracy recovery including a previously missed uracil-uracil stack (right). (**B**) Modeling the sarcin-ricin loop with the bottom helix free is a classic prediction challenge that has only been recovered in previous computational studies by including homologous fragments during model assembly (*2*, *62*). SWM results give an incorrect base-pairing pattern and backbone torsional combinations that are not recognized as any of the 53 "rotameric" conformers identified in bioinformatics studies (*63*) ("!!" annotations). The Rosetta RNA torsional potential sums independent one-dimensional potentials based on crystallographic statistics of each of the 7 nucleotide torsions (α, β, γ, δ, ε, ζ, and χ); it has not been fit to recover special combinations of multiple consecutive torsions that recur in natural structures (*63*), in an effort to avoid bias in crystallographic refinement applications (*11*). Instead including a modest bonus of 1 k_BT for these conformers guides SWM to a model with Angstrom accuracy and all correct noncanonical pairs (right).

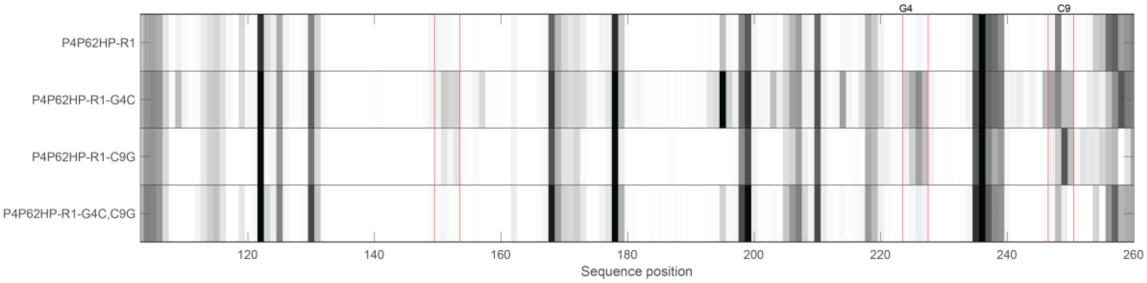


fig. S5. Compensatory mutagenesis of the R(1) receptor read out through chemical mapping. SHAPE (1M7) reactivity data for the P4-P6 RNA with the GGAA/R(1) tetraloop receptor, measured by capillary gel electrophoresis. The G4C and C9G mutations substantially disrupted wild type folding, especially but not exclusively in the tetraloop and receptor regions themselves (between red bars); the wild-type reactivity profile is recovered by the double mutant, supporting a base pair predicted by SWM modeling. Other single and double mutants destabilized the P4-P6 RNA beyond the limit of detection (data available in the RNA Mapping Database under accession code TRP4P6_R1J_0001).

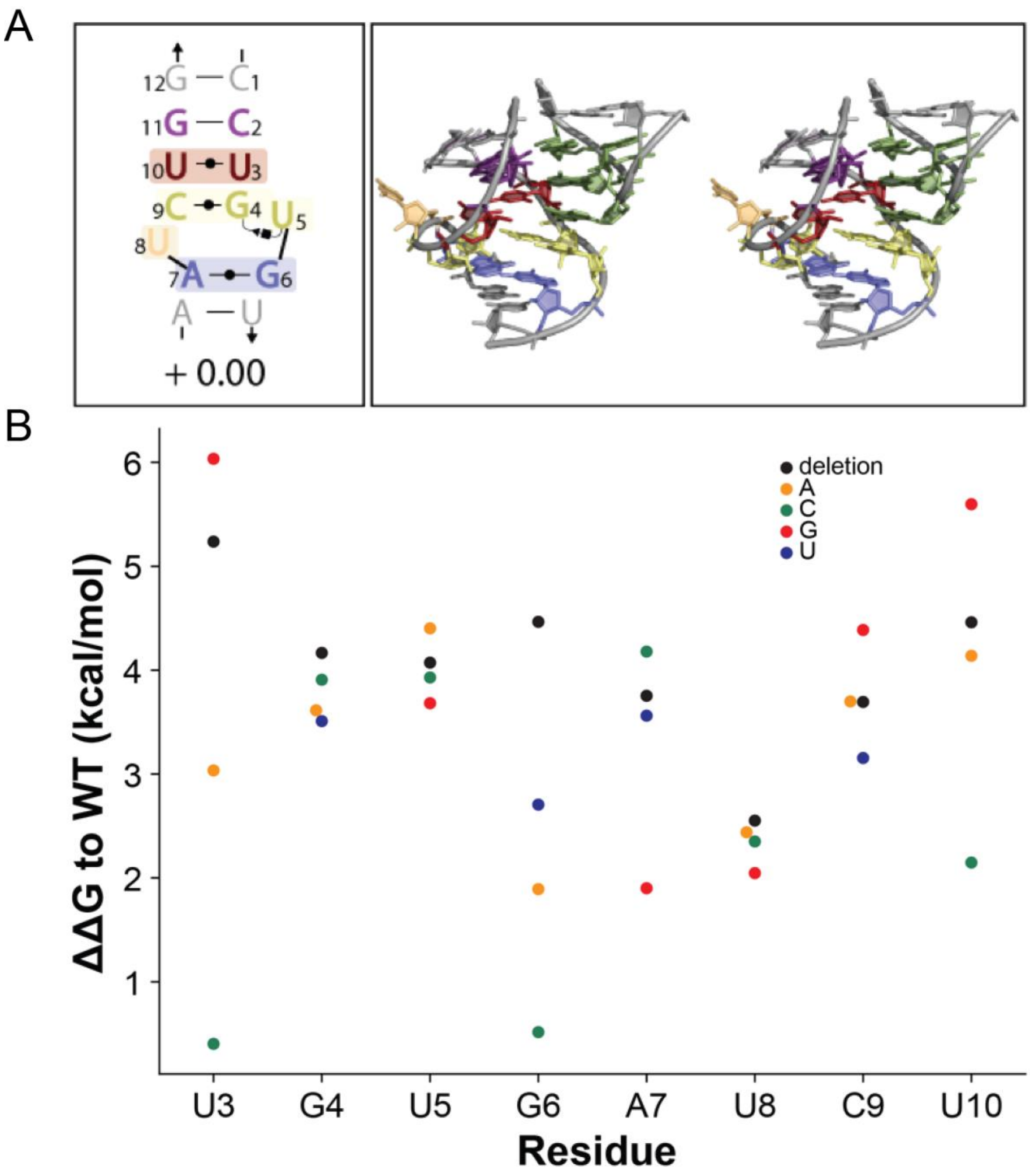


fig. S6. Comprehensive single mutant analysis of the tetraloop receptor R(1). Variant receptors of R(1) were assessed for the their ability to maintain GGAA binding in the context of a tectoRNA heterodimer. Top: The predicted R(1) secondary structure and a stereo image of the GGAA/R(1) SWM model. Non-canonical pairs are denoted using the Leontis-Westhof notation. Bottom: The ΔΔG with respect to the GGAA/R(1) wildtype interaction is plotted for receptor single mutations for each residue.

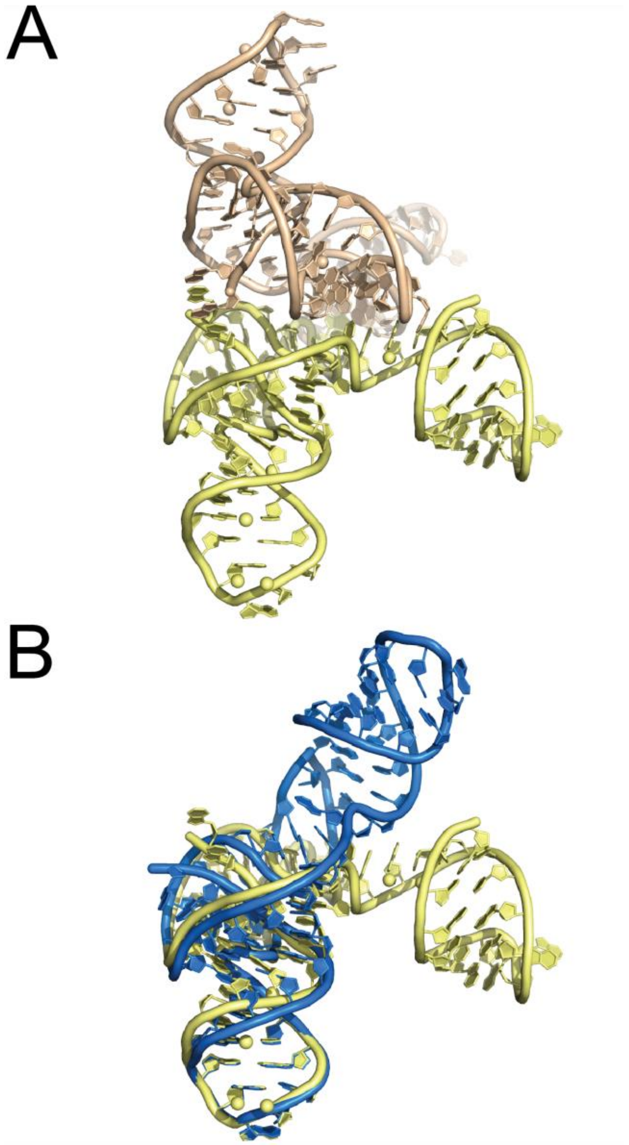


fig. S7. Global fold changes between the template viral xrRNA and the Zika xrRNA structure prediction challenge. (**A**) The Murray Valley Encephalitis Xrn1-resistant RNA (xrRNA) structure available at the time of modeling (PDB: 4PQV) was crystallized in a likely non-biological conformation (*64*): one asymmetric unit (yellow) was splayed out such that the strands predicted to make an intramolecular pseudoknot in fact made an intermolecular stem with a crystallographic neighbor (wheat), and interleaving of 5’ ends of the two molecules (red arrows). (**B**) Global structural rearrangement was confirmed the crystal structure of the Zika target (5TPY) (*61*) (blue), released after blind modeling of the target.

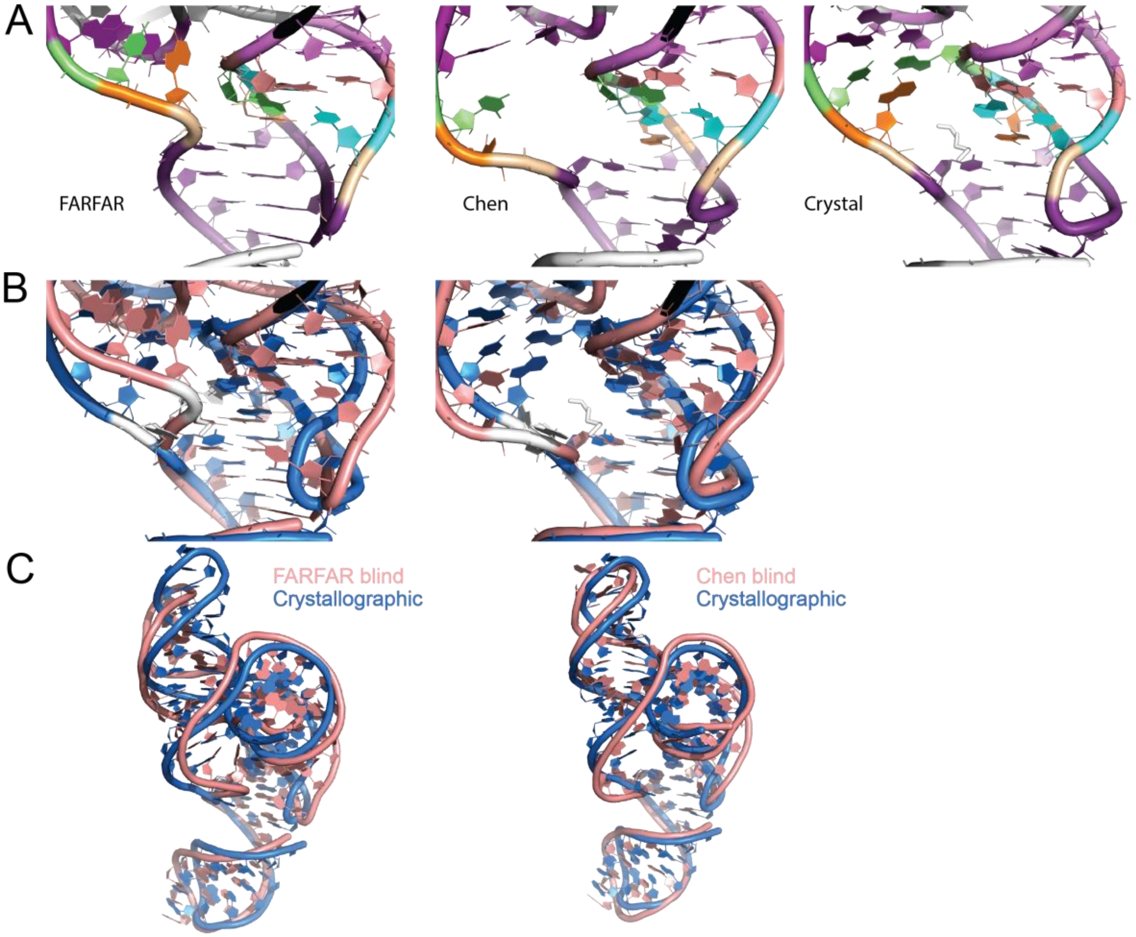


fig. S8. Other models of RNA-Puzzle 18 (Zika xrRNA). (**A**) Best blind model found by fragment assembly of RNA (FARFAR) and best blind model by another group (Chen) compared to crystal structure. Most noncanonical interactions are incorrectly recovered in these blind models, with a large ‘hole’ in the Chen models (middle panel). (**B-C**) overlays of models (salmon) with crystal structure (marine), magnified in this noncanonical region (B) and showing overall global folds in (C).

Supplementary Tables

table S1. A comparison of the SWA and SWM methods using the same energy function as the original SWA benchmark set of trans-helix single-stranded loops, and SWM results using the updated Rosetta free energy function (SWM*).

		Best of 5 Lowest Energy Cluster Centers						Lowest RMSD Model			Lowest Energy Model		
		RMSD (Å)[a]			F_{NWC}			RMSD (Å)			E-Gap (REU)[b]		
Motif	Length	SWA	SWM	SWM*	SWA	SWM	SWM*	SWA	SWM	SWM*	SWA	SWM	SWM*
5' J1/2, Leadzyme	4	0.47	1.21	1.21	1.00	1.00	1.00	0.37	0.38	0.39	-0.08	0.17	-0.04
5' P1, M-Box Riboswitch	4	1.06	1.22	0.63	1.00	1.00	1.00	0.84	0.42	0.36	1.07	-0.01	-0.05
3' J5/5a, Group I Intron	4	0.69	0.56	0.44	1.00	1.00	1.00	0.69	0.46	0.40	-0.72	0.21	-0.14
5' J5/5a, Group I Intron	5	1.11	1.09	0.74	1.00	1.00	1.00	0.71	0.63	0.64	-2.16	-0.08	-0.32
Hepatitis C Virus IRES IIa	5	3.75	7.32	2.96	0.00	0.00	0.00	0.53	2.30	1.33	-4.71	-4.43	-2.47
J2/4, TPP Riboswitch	5	0.97	0.96	0.57	1.00	1.00	1.00	0.74	0.77	0.46	-1.1	-0.34	-0.15
J3/1, Glycine Riboswitch	7	1.15	3.32	0.55	1.00	0.67	1.00	0.63	0.97	0.52	-3.25	-5.35	-0.28
J2/3, Group II Intron	7	0.79	0.83	0.61	1.00	1.00	1.00	0.66	0.72	0.54	-0.6	1.62	-0.20
L1, SAM-II Riboswitch	7	1.07	1.05	0.91	0.78	0.78	0.78	0.72	0.71	0.72	3.07	0.59	-0.75
L2, Viral RNA Pseudoknot	7	5.39	3.35	3.04	0.66	1.00	1.00	1.93	1.40	0.71	-9.01	-6.47	-0.92
23S rRNA (44-49)	6	1.01	2.76	0.70	1.00	1.00	1.00	0.81	1.24	0.69	0.70	-4.05	-1.32
23S rRNA (531-536)	6	1.02	3.48	1.38	0.83	0.83	0.83	0.80	1.77	1.24	-1.76	-5.98	-1.24
23S rRNA (2534-2540)	7	5.71	6.26	7.49	0.77	0.77	0.77	0.78	1.90	2.53	-8.68	-11.55	-1.15
23S rRNA (1976-1985)	10	5.28	5.94	16.32	0.57	0.57	0.57	3.17	3.69	6.20	-5.18	-2.90	9.82
23S rRNA (2003-2012)	10	1.80	1.09	9.96	0.86	0.86	0.64	1.36	1.09	6.23	-5.35	6.63	6.69
Median	6	1.07	1.22	0.91	1	1	1	0.74	0.97	0.69	-1.76	-0.34	-0.28
Mean	6.4	2.12	2.70	3.15	0.83	0.83	0.84	1.01	1.27	1.53	-2.63	-2.16	0.51

[a] RMSD values of best of 5 lowest energy cluster centers for SWM* here in table S1 and in table S3 are slightly different (most strongly in L2 Viral pseudoknot) due to use of legacy clustering method for this table; see Supplemental Methods.

[b] Rosetta Energy Units, arbitrary units in SWA and SWM; calibrated in SWM* so that 1 REU corresponds to 1 k_BT.

table S2. Updates to the Rosetta energy function.

Energy function term	**Used for SWA (rna_loop_hires_04092010.wts)**	**Current work (rna_res_level_energy4.wts)**
Lennard-Jones/dispersion		
fa_atr	0.23	0.21
fa_rep	0.12	0.20
fa_intra_rep	0.0029	0.0029
fa_stack	0.125	0.13
H-bonds and solvation		
lk_nonpolar	0.32	0.25
geom_sol_fast	0.62	0.17
hbond_sr_bb_sc	0.62	0.96
hbond_lr_bb_sc	2.4	0.96
hbond_sc	2.4	0.96
RNA torsion terms		
rna_torsion	2.9	1.0
rna_sugar_close	0.7	0.82
suiteness_bonus	–	1.0
ch_bond	0.42	–
Electrostatics (not captured in H-bonds)		
fa_elec_rna_phos_phos	1.05	1.7
stack_elec	–	0.76
Bonuses/costs for free/instantiated moieties		
rna_bulge	0.45	–
linear_chainbreak	5.0	5.0
intermol	–	1.0
loop_close	–	1.0
free_suite	–	2.0
free_2HOprime	–	1.0
ref	–	1.0
other_pose*	–	1.0
Options and miscellaneous weights		
NO_HB_ENV_DEP	active	active
METHOD_WEIGHTS	–	4.14 (G), 3.58 (A), 2.82 (C), 3.76 (U)
RNA_TORSION_POTENTIAL	ps_03242010	RNA11_based_new
RNA_SYN_G_POTENTIAL_BONUS	–	-1.5
RNA_SUITENESS_BONUS		test/1z_6n_2[_bonus
ENLARGE_H_LJ_WDEPTH	–	active

[*] *other_pose* allows Rosetta to compute energies for conformations in which separately instantiated segments are simulated as separate ‘poses’.

table S3. Detailed performance of the stepwise Monte Carlo algorithm on 82 benchmark cases.

	Motif Properties		Best of Five Lowest Energy Cluster Centers			Lowest RMSD Model			Lowest Energy Model			Benchmark Run Properties		
Motif	Length	Strands	RMSD	n_{NWC}	F_{NWC}	RMSD	n_{NWC}	F_{NWC}	n_{NWC}	F_{NWC}	E-Gap	Time (CPU-hours)	Total Models	Models in best cluster center
Trans-Helix Loop														
5' J1/2, Leadzyme	4	1	1.22	5	1.00	0.39	5	1.00	5	1.00	-0.04	31437	5003	4493
5' P1, M-Box Riboswitch	4	1	0.63	3	1.00	0.36	3	1.00	3	1.00	-0.05	23459	5000	3630
3' J5/5a, Group I Intron	4	1	0.4	4	1.00	0.4	4	1.00	4	1.00	-0.14	25418	5000	3964
5' J5/5a, Group I Intron	5	1	0.74	4	1.00	0.64	4	1.00	4	1.00	-0.32	40199	5007	279
Hepatitis C Virus IRES IIa	5	1	2.56	--	--	1.33	--	--	--	--	-2.47	22859	5000	2889
J2/4, TPP Riboswitch	5	1	0.72	4	1.00	0.46	4	1.00	4	1.00	-0.15	38249	5304	5197
J3/1, Glycine Riboswitch	7	1	0.55	3	1.00	0.52	3	1.00	3	1.00	-0.28	52419	5288	1960
J2/3, Group II Intron	7	1	0.65	7	1.00	0.54	7	1.00	7	1.00	-0.2	52150	5775	4431
L1, SAM-II Riboswitch	7	1	0.83	8	0.89	0.72	8	0.89	8	0.89	-0.75	41575	5103	3293
L2, Viral RNA Pseudoknot	7	1	0.71	6	1.00	0.71	6	1.00	6	1.00	-0.92	36788	5109	1942
23S rRNA (44-49)	6	1	1.25	6	1.00	0.69	6	1.00	6	1.00	-1.32	60186	3927	3713
23S rRNA (531-536)	6	1	1.48	5	0.83	1.24	4	0.67	5	0.83	-1.24	69864	3482	1426
23S rRNA (2534-2540)	7	1	6.94	12	0.92	2.53	13	1.00	10	0.77	-1.15	79686	3502	1
23S rRNA (1976-1985)	10	1	15.27	4	0.57	6.2	4	0.57	4	0.57	9.81	76498	5164	367
23S rRNA (2003-2012)	10	1	9.02	7	0.64	6.23	7	0.64	7	0.64	6.69	79686	5109	33
Median	6.0	1	0.83	5.0	1.00	0.69	4.5	1.00	5.0	1.00	-0.28	41575	5007	2889
Mean	6.3	1	2.86	5.6	0.89	1.53	5.6	0.89	5.4	0.86	0.50	48698	4852	2508
Apical Loop														
gcaa_tetraloop	4	1	1.14	1	1.00	0.71	1	1.00	1	1.00	-1.19	4029	1603	1565
uucg_tetraloop	4	1	1.14	0	0.00	0.58	1	1.00	0	0.00	-1.09	3400	1685	124
gagua_pentaloop	5	1	1.10	1	1.00	0.98	1	1.00	1	1.00	-1.76	3771	1377	92
anticodon_phe	7	1	2.24	2	1.00	2.16	1	0.50	2	1.00	-13.50	6173	1625	66
Median	4.5	1	1.14	1.0	1.00	0.85	1.0	1.00	1.0	1.00	-1.48	3900	1614	108
Mean	5.0	1	1.41	1.0	0.80	1.11	1.0	0.80	1.0	0.80	-4.39	4343	1573	462
Fixed Two-Way Junction														
puzzle1_alt_fixed	7	2	1.72	1	1.00	1.34	1	1.00	1	1.00	-5.95	4157	747	82
srp_domainIV_fixed	8	2	0.90	5	1.00	0.54	4	0.80	5	1.00	0.01	4209	629	41

srl_fixed	9	2	0.56	5	1.00	0.47	5	1.00	5	1.00	0.37	4011	487	14
kink_turn_fixed	9	2	1.45	2	0.67	1.00	3	1.00	2	0.67	-3.69	4476	725	16
j55a_P4P6_fixed	9	2	0.55	4	1.00	0.42	4	1.00	4	1.00	0.07	4118	554	9
P5b_connect	18	2	2.68	2	1.00	2.67	2	1.00	2	1.00	17.90	1943	291	1
gg_mismatch_fixed	2	2	0.32	1	1.00	0.30	1	1.00	0	0.00	-3.18	939	452	64
tandem_ga_imino_fixed	4	2	0.87	2	1.00	0.63	2	1.00	2	1.00	-4.40	2365	863	349
tandem_ga_sheared_fixed	4	2	0.61	2	1.00	0.56	2	1.00	2	1.00	-3.87	2808	868	72
hiv_rre_fixed	5	2	0.35	2	1.00	0.28	2	1.00	2	1.00	-5.28	7477	2299	285
j44a_p4p6_fixed	5	2	0.58	2	1.00	0.46	2	1.00	2	1.00	-7.31	6935	2240	1317
just_tr_P4P6_fixed	5	2	0.61	2	1.00	0.40	2	1.00	2	1.00	-4.56	2419	703	121
r2_4x4_fixed	8	2	1.49	3	0.75	0.75	4	1.00	3	0.75	-7.37	5786	1528	87
loopE_fixed	14	2	1.74	5	0.67	1.74	4	0.67	3	0.50	-8.56	5399	857	1
Median	7.5	2	0.74	2.0	1.00	0.55	2.0	1.00	2.0	1.00	-4.14	4138	747	72
Mean	7.6	2	1.03	2.7	0.90	0.83	2.7	0.93	2.5	0.85	-2.56	4074	988	195
Fixed Multi-Helix Junction														
hammerhead_3WJ_cat_fixed	11	3	3.05	2	0.67	1.41	2	0.67	2	0.67	0.14	116068	12515	16
hammerhead_3WJ_precat_fixed	11	3	1.57	4	0.80	1.35	3	0.60	4	0.80	-0.49	76328	8659	8
VS_rbzm_P2P3P6_fixed	7	3	0.57	3	1.00	0.36	1	0.33	3	1.00	-4.33	4373	1287	363
VS_rbzm_P3P4P5_fixed	10	3	1.91	1	0.25	1.78	0	0.00	2	0.50	-9.94	2413	601	7
hammerhead_3WJ_cat_OMC_fixed	11	3	3.04	3	1.00	1.47	2	0.67	2	0.67	0.40	21165	3595	43
Median	11.0	3	1.91	3.0	0.80	1.41	2.0	0.60	2.0	0.67	-0.49	21165	3595	16
Mean	10.0	3	2.03	2.6	0.72	1.27	1.6	0.44	2.6	0.72	-2.84	44069	5331	87
Fixed Tertiary Contact														
tl_tr_P4P6	9	3	0.64	4	0.80	0.64	4	0.80	4	0.80	0.73	4211	590	6
hammerhead_tert_fixed	10	3	1.16	2	0.67	1.16	2	0.67	2	0.67	-1.65	25254	4376	2
kiss_add_fixed	16	2	2.40	2	0.40	2.40	2	0.40	0	0.00	-0.90	4296	1177	1
kiss_add_L2_fixed	9	2	0.71	7	1.00	0.46	7	1.00	7	1.00	1.12	4037	1214	14
kiss_add_L3_fixed	7	2	0.88	6	0.86	0.55	6	0.86	6	0.86	-2.79	4028	745	679
xrRNA (RNApuzzle-18)	8	3	1.97	1	0.50	1.60	2	1.00	1	0.50	-2.16	21799	2218	25
gir1_p2.1p5_kiss_fixed	13	2	1.99	1	1.00	1.20	1	1.00	0	0.00	-12.40	2366	426	4
gir1_p2p9_gaaa_minor_fixed	8	3	1.58	2	0.33	1.51	1	0.17	2	0.33	-10.75	4245	864	97
t_loop_fixed	7	2	0.95	1	1.00	0.71	1	1.00	1	1.00	-3.44	863	413	37
t_loop_modified_fixed	7	2	1.33	2	1.00	1.20	1	0.50	1	0.50	-5.01	7518	1916	10
Median	8.5	2.0	1.25	2.0	0.83	1.18	2.0	0.83	1.5	0.59	-2.48	4228	1021	12
Mean	9.4	2.4	1.36	2.8	0.74	1.14	2.7	0.74	2.4	0.66	-3.73	7862	1394	88

Two-Way Junction															
gg_mismatch	2	2	0.79	1	1.00	0.65	1	1.00	0	0.00	-0.65	2863	986	851
tandem_ga_imino	4	2	0.98	2	1.00	0.74	2	1.00	2	1.00	-0.04	4232	996	375
tandem_ga_sheared	4	2	0.75	2	1.00	0.50	2	1.00	2	1.00	-1.04	3632	807	503
hiv_rre	5	2	2.12	1	0.50	0.60	2	1.00	1	0.50	-1.21	2135	558	79
j44a_p4p6	5	2	1.59	2	1.00	1.09	2	1.00	1	0.50	-2.52	3931	734	14
just_tr_P4P6	5	2	1.22	2	1.00	1.00	2	1.00	1	0.50	0.03	2103	473	2
cg_helix	6	2	0.58	--	--	0.44	--	--	--	--	0.07	2148	979	361
puzzle1	7	2	0.96	1	1.00	0.72	1	1.00	0	0.00	-2.12	4307	1000	123
srp_domainIV	8	2	1.26	5	1.00	1.22	5	1.00	1	0.20	-0.89	4356	681	8
r2_4x4	8	2	1.74	3	0.75	1.68	3	0.75	3	0.75	-3.21	3969	614	12
gagu_forcesyn_blockstackU	8	2	4.49	2	1.00	2.57	1	0.50	2	1.00	-6.66	4320	1514	5
srl_free_bulgedG	9	2	4.66	2	0.50	0.95	4	1.00	2	0.50	-1.86	70454	13037	15
j55a_P4P6_align	9	2	2.04	1	0.25	2.04	1	0.25	1	0.25	-6.54	857	390	1
kink_turn_align	9	2	2.07	1	0.33	1.90	1	0.33	1	0.33	-9.13	970	510	1
loopE	14	2	2.00	3	0.50	2.00	3	0.50	3	0.50	1.65	4515	1026	1
Median	7.0	2	1.59	2	1.00	1.00	2	1.00	1	0.50	-1.21	3931	807	14
Mean	6.9	2	1.82	2	0.70	1.21	2	0.75	1	0.50	-2.27	7653	1620	157
Multi-Helix Junction															
hammerhead_3WJ_precat	11	3	6.09	2	0.40	4.98	4	0.80	1	0.20	-2.92	1959	338	1
VS_rbzm_P2P3P6_align	7	3	1.13	3	1.00	0.97	2	0.67	3	1.00	-4.59	1983	647	12
VS_rbzm_P3P4P5_align	10	3	2.60	1	0.25	2.13	0	0.00	1	0.25	-13.64	2032	896	3
hammerhead_3WJ_cat_OMC_align	11	3	2.89	3	1.00	2.89	3	1.00	1	0.33	-0.55	12557	4220	3
zika_3WJ (RNApuzzle-18)	3	3	2.42	0	0.00	2.09	0	0.00	0	0.00	0.92	15380	1756	71
Median	10.0	3	2.60	2	0.40	2.13	2	0.67	1	0.25	-2.92	2032	896	3
Mean	8.4	3	3.03	2	0.53	2.61	2	0.53	1	0.35	-4.16	6782	1571	18
Tertiary Contact															
gaaa_minor_dock	4	1	1.41	1	0.50	1.12	1	0.50	1	0.50	-1.07	2087	578	37
gir1_p2.1p5_kiss	13	2	2.75	1	0.50	2.01	1	0.50	1	0.50	-12.40	2001	414	6
gir1_p2p9_gaaa_minor	8	3	1.83	2	0.67	1.83	2	0.67	1	0.33	-0.88	2055	528	2
tl_tr_P4P6_dock	9	3	3.03	2	0.29	2.89	5	0.71	1	0.14	5.49	1976	542	1
kiss_add_PK_dock	12	2	2.58	3	0.43	2.42	1	0.14	1	0.14	2.51	2057	636	1
t_loop_align	7	2	3.20	0	0.00	1.14	1	1.00	0	0.00	-7.82	916	360	11
hammerhead_tert_align	10	3	8.65	0	0.00	3.53	0	0.00	0	0.00	-1.42	867	467	1
t_loop_modified_align	7	2	3.99	0	0.00	1.95	0	0.00	0	0.00	-4.31	7724	2196	3

Median	8.5	2	2.89	1	0.36	1.98	1	0.50	1	0.14	-1.25	2028	535	3
Mean	8.8	2.3	3.43	1	0.33	2.11	1	0.41	1	0.19	-2.49	2460	715	8
Non-Helix Embedded														
cg_helix_Zform	6	2	1.75	2	0.67	1.34	3	1.00	2	0.67	-2.78	2160	1491	1
g_quadruplex_fixed	16	4	2.75	16	0.80	2.75	16	0.80	14	0.70	1.85	2033	475	1
g_quadruplex_inosine_fixed	16	4	2.87	16	0.80	2.68	15	0.75	16	0.80	-7.46	7637	1829	9
bru_gag_tetraplex	12	4	3.42	13	0.81	3.00	13	0.81	13	0.81	-8.47	7498	1836	3
parallel_AA	6	2	1.41	3	1.00	0.88	3	1.00	2	0.67	-0.35	923	1034	195
bulged_tetraplex	8	4	7.37	4	0.50	1.60	8	1.00	4	0.50	-10.71	870	891	3
Median	10.0	4	2.81	9	0.80	2.14	11	0.91	9	0.69	-5.12	2097	1263	3
Mean	10.7	3.3	3.26	9	0.77	2.04	10	0.83	9	0.73	-4.65	3520	1259	35
OVERALL														
Median	7	2	1.49	2	0.96	1.13	2	1.00	2	0.69	-1.28	4222	1013	29
Mean	7.9	2	2.15	3	0.76	1.43	3	0.77	3	0.68	-2.44	15773	2175	560

table S4. Detailed performance of the FARFAR algorithm on 82 benchmark cases.

	Motif Properties		Best of Five Lowest Energy Cluster Centers			Lowest RMSD Model			Lowest Energy Model			Benchmark Run Properties		
Motif	**Length**	**Strands**	**RMSD**	**n_{NWC}**	**F_{NWC}**	**RMSD**	**n_{NWC}**	**F_{NWC}**	**n_{NWC}**	**F_{NWC}**	**E-Gap**	**Time (CPU-hours)**	**Total Models**	**Models in best cluster center**
Trans-Helix Loop														
5' J1/2, Leadzyme	4	1	1.78	4	0.80	0.37	5	1.00	4	0.80	1.90	103	4010	93
5' P1, M -Box Riboswitch	4	1	1.00	2	0.67	0.43	3	1.00	2	0.67	-6.34	93	3904	1421
3' J5/5a, Group I Intron	4	1	2.52	4	1.00	0.73	4	1.00	4	1.00	-0.36	95	3867	1850
5' J5/5a, Group I Intron	5	1	3.30	4	1.00	0.75	4	1.00	1	0.25	5.18	126	4467	22
Hepatitis C Virus IRES IIa	5	1	2.27	--	--	1.91	--	--	--	--	3.73	190	7821	15
J2/4, TPP Riboswitch	5	1	3.29	3	0.75	3.29	3	0.75	2	0.50	7.85	159	4708	204
J3/1, Glycine Riboswitch	7	1	3.43	1	0.33	2.11	3	1.00	1	0.33	9.92	337	9369	22
J2/3, Group II Intron	7	1	3.13	4	0.57	2.90	4	0.57	4	0.57	16.13	372	10608	9
L1, SAM-II Riboswitch	7	1	2.00	7	0.78	1.64	7	0.78	8	0.89	3.19	216	6651	139
L2, Viral RNA Pseudoknot	7	1	3.81	6	1.00	1.31	6	1.00	5	0.83	-0.85	295	9428	443
23S rRNA (44-49)	6	1	1.25	5	0.83	1.09	5	0.83	5	0.83	5.96	303	5017	3131
23S rRNA (531-536)	6	1	5.59	4	0.67	4.26	4	0.67	4	0.67	22.88	397	3896	25
23S rRNA (2534-2540)	7	1	5.91	11	0.85	2.65	12	0.92	11	0.85	16.32	44	637	106
23S rRNA (1976-1985)	10	1	11.44	4	0.57	6.23	4	0.57	4	0.57	32.69	397	4661	2
23S rRNA (2003-2012)	10	1	11.00	6	0.55	7.09	6	0.55	6	0.55	39.97	378	3409	9
Median	6.0	1	3.29	4.0	0.77	1.91	4.0	0.88	4.0	0.67	5.96	216	4661	93
Mean	6.3	1	4.11	4.6	0.74	2.45	5.0	0.80	4.4	0.69	9.77	234	5497	499
Apical Loop														
gcaa_tetraloop	4	1	1.39	1	1.00	0.73	1	1.00	1	1.00	2.49	58	2826	1131
uucg_tetraloop	4	1	3.57	0	0.00	1.61	0	0.00	0	0.00	3.88	39	2001	24
gagua_pentaloop	5	1	3.15	1	1.00	0.73	1	1.00	1	1.00	4.29	35	1626	72
anticodon_phe	7	1	2.77	2	1.00	1.77	2	1.00	2	1.00	-9.01	239	3571	19
Median	4.5	1	2.96	1.0	1.00	1.17	1.0	1.00	1.0	1.00	3.19	49	2414	48
Mean	5.0	1	2.72	1.0	0.80	1.21	1.0	0.80	1.0	0.80	0.41	93	2506	312
Fixed Two-Way Junction														
puzzle1_alt_fixed	7	2	3.33	1	0.50	2.44	1	0.50	1	0.50	11.52	170	5346	4
srp_domainIV_fixed	8	2	1.19	4	0.80	0.61	5	1.00	4	0.80	9.78	47	1580	16

srl_fixed	9	2	0.85	5	1.00	0.72	5	1.00	5	1.00	13.47	49	1550	40
kink_turn_fixed	9	2	1.31	3	1.00	1.14	3	1.00	3	1.00	5.72	36	1184	81
j55a_P4P6_fixed	9	2	4.30	1	0.25	2.16	2	0.50	1	0.25	19.35	50	1577	24
P5b_connect	18	2	1.11	2	1.00	1.02	2	1.00	2	1.00	23.35	106	2907	13
gg_mismatch_fixed	2	2	1.84	0	0.00	1.79	0	0.00	0	0.00	3.61	9	384	105
tandem_ga_imino_fixed	4	2	1.34	2	1.00	0.79	1	0.50	2	1.00	1.07	75	3003	136
tandem_ga_sheared_fixed	4	2	0.69	2	1.00	0.35	2	1.00	2	1.00	1.61	73	2889	2598
hiv_rre_fixed	5	2	0.59	2	1.00	0.59	2	1.00	2	1.00	3.43	60	2300	10
j44a_p4p6_fixed	5	2	0.96	2	1.00	0.96	2	1.00	2	1.00	2.21	44	1517	72
just_tr_P4P6_fixed	5	2	1.22	2	1.00	0.65	2	1.00	2	1.00	4.57	52	1842	333
r2_4x4_fixed	8	2	1.04	4	1.00	0.81	4	1.00	4	1.00	3.14	46	1535	194
loopE_fixed	14	2	1.07	6	1.00	0.81	6	1.00	6	1.00	3.03	61	1625	87
Median	7.5	2	1.15	2.0	1.00	0.81	2.0	1.00	2.0	1.00	4.09	51	1603	77
Mean	7.6	2	1.49	2.6	0.86	1.06	2.6	0.88	2.6	0.86	7.56	63	2089	265
Fixed Multi-Helix Junction														
hammerhead_3WJ_cat_fixed	11	3	3.67	1	0.33	3.19	2	0.67	1	0.33	30.60	195	5603	2
hammerhead_3WJ_precat_fixed	11	3	1.93	1	0.20	1.75	2	0.40	1	0.20	19.45	30	826	6
VS_rbzm_P2P3P6_fixed	7	3	1.65	4	1.00	1.37	2	0.50	4	1.00	6.92	46	1465	169
VS_rbzm_P3P4P5_fixed	10	3	1.93	1	0.25	1.25	1	0.25	1	0.25	4.82	69	2084	9
hammerhead_3WJ_cat_OMC_fixed	11	3	5.25	2	0.67	3.52	1	0.33	2	0.67	10.26	670	19186	4
Median	11.0	3	1.93	1.0	0.33	1.75	2.0	0.40	1.0	0.33	10.26	69	2084	6
Mean	10.0	3	2.89	1.8	0.42	2.22	1.6	0.42	1.8	0.42	14.41	202	5833	38
Fixed Tertiary Contact														
tl_tr_P4P6	9	3	1.62	1	0.20	0.85	3	0.60	2	0.40	17.57	51	1392	28
hammerhead_tert_fixed	10	1	2.33	3	1.00	2.05	2	0.67	3	1.00	17.18	866	14905	2
kiss_add_fixed	16	3	3.40	1	0.20	2.41	0	0.00	0	0.00	17.08	3021	48217	1
kiss_add_L2_fixed	9	2	1.09	7	1.00	0.76	6	0.86	7	1.00	12.17	74	2331	88
kiss_add_L3_fixed	7	2	2.52	6	0.86	2.26	6	0.86	6	0.86	16.61	100	3085	21
xrRNA (RNApuzzle-18)	8	3	2.97	0	0.00	2.63	0	0.00	0	0.00	16.26	56	1292	18
gir1_p2.1p5_kiss_fixed	13	2	1.88	1	0.50	1.39	0	0.00	1	0.50	5.69	110	2633	3
gir1_p2p9_gaaa_minor_fixed	8	3	1.25	3	0.50	0.81	5	0.83	3	0.50	5.26	76	2043	121
t_loop_fixed	7	2	1.68	1	1.00	1.25	1	1.00	0	0.00	7.07	81	3331	164
t_loop_modified_fixed	7	2	1.12	1	0.50	1.09	2	1.00	1	0.50	-2.24	153	5741	28
Median	8.5	2.0	1.78	1.0	0.50	1.32	2.0	0.75	1.5	0.50	14.22	91	2859	25
Mean	9.4	2.3	1.99	2.4	0.53	1.55	2.5	0.63	2.3	0.55	11.26	459	8497	47

Two-Way Junction														
gg_mismatch	2	2	1.40	0	0.00	1.18	0	0.00	0	0.00	7.90	86	3995	87
tandem_ga_imino	4	2	0.89	2	1.00	0.84	2	1.00	0	0.00	6.50	86	3456	582
tandem_ga_sheared	4	2	0.60	2	1.00	0.48	2	1.00	2	1.00	5.91	58	2252	1336
hiv_rre	5	2	3.02	1	0.50	2.38	1	0.50	1	0.50	6.24	93	3494	90
j44a_p4p6	5	2	1.38	2	1.00	1.21	2	1.00	2	1.00	9.25	51	1780	11
just_tr_P4P6	5	2	1.35	2	1.00	1.08	1	0.50	0	0.00	11.75	50	1748	28
cg_helix	6	2	0.50	--	--	0.27	--	--	--	--	3.75	48	3853	3851
puzzle1	7	2	1.17	0	0.00	0.80	1	1.00	1	1.00	8.85	106	3937	501
srp_domainIV	8	2	1.13	3	0.60	1.13	3	0.60	1	0.20	13.00	39	1344	54
r2_4x4	8	2	1.97	2	0.50	1.34	2	0.50	1	0.25	7.53	48	1633	52
gagu_forcesyn_blockstackU	8	4	5.01	0	0.00	3.10	1	0.50	0	0.00	-4.70	71	5782	2
srl_free_bulgedG	9	2	4.84	1	0.25	2.31	3	0.75	1	0.25	6.40	74	2713	92
j55a_P4P6_align	9	2	1.86	2	0.50	1.78	0	0.00	1	0.25	16.89	61	1683	7
kink_turn_align	9	2	1.41	3	1.00	1.23	3	1.00	3	1.00	4.13	47	1387	21
loopE	14	2	1.55	6	1.00	1.18	6	1.00	3	0.50	11.76	62	1743	18
Median	7.0	2	1.40	2.0	0.55	1.18	2.0	0.68	1.0	0.25	7.53	61	2252	54
Mean	6.9	2	1.87	1.9	0.65	1.35	1.9	0.68	1.1	0.43	7.68	65	2720	449
Multi-Helix Junction														
hammerhead_3WJ_precat	11	3	5.42	1	0.20	3.15	3	0.60	1	0.20	13.24	37	969	2
VS_rbzm_P2P3P6_align	7	3	1.22	4	1.00	1.18	4	1.00	4	1.00	13.06	47	1362	17
VS_rbzm_P3P4P5_align	10	3	2.12	2	0.50	1.30	0	0.00	1	0.25	11.00	75	1914	20
hammerhead_3WJ_cat_OMC_align	11	3	4.65	0	0.00	2.38	0	0.00	0	0.00	12.51	804	19287	2
zika_3WJ (RNApuzzle-18)	3	3	3.45	0	0.00	2.67	0	0.00	0	0.00	14.91	399	1292	1
Median	10.0	3	3.45	1.0	0.20	2.67	0.0	0.00	1.0	0.20	13.06	75	1362	2
Mean	8.4	3	3.33	1.4	0.39	2.24	1.4	0.39	1.2	0.39	12.94	272	4965	8
Tertiary Contact														
gaaa_minor_dock	4	1	1.26	2	1.00	0.95	2	1.00	2	1.00	12.63	100	2290	991
gir1_p2.1p5_kiss	13	2	2.30	1	0.50	1.82	1	0.50	1	0.50	25.68	125	3187	4
gir1_p2p9_gaaa_minor	8	1	1.95	2	0.67	1.95	2	0.67	2	0.67	14.29	148	2132	2
tl_tr_P4P6_dock	9	3	1.66	2	0.40	0.97	4	0.80	2	0.40	21.31	134	2350	9
kiss_add_PK_dock	12	5	2.30	0	0.00	1.88	1	0.20	0	0.00	16.45	3037	38834	2
t_loop_align	7	2	1.63	0	0.00	1.63	0	0.00	1	1.00	19.90	96	3148	3
hammerhead_tert_align	10	4	4.01	0	0.00	2.97	0	0.00	0	0.00	44.02	1167	22708	1
t_loop_modified_align	7	2	3.77	0	0.00	2.89	1	0.50	0	0.00	19.49	231	4961	5

Median	8.5	2.0	2.12	0.5	0.20	1.85	1.0	0.50	1.0	0.45	20.61	141	3168	4
Mean	8.8	2.5	2.36	0.9	0.26	1.88	1.4	0.41	1.0	0.30	21.72	630	9951	127
Non-Helix Embedded														
cg_helix_Zform	6	2	6.19	0	0.00	3.32	0	0.00	0	0.00	5.60	119	9715	16
g_quadruplex_fixed	16	4	1.88	16	0.80	1.54	15	0.75	16	0.80	7.80	64	2749	88
g_quadruplex_inosine_fixed	16	4	2.59	15	0.75	1.94	13	0.65	16	0.80	-4.70	92	3892	151
bru_gag_tetraplex	12	4	7.35	12	0.75	3.01	13	0.81	12	0.75	-7.39	326	14221	6470
parallel_AA	6	2	1.63	2	0.67	0.74	3	1.00	3	1.00	1.40	95	8060	381
bulged_tetraplex	8	4	6.00	4	0.50	5.63	4	0.50	4	0.50	2.80	66	4156	8
Median	10.0	4.0	4.30	8.0	0.71	2.48	8.5	0.70	8.0	0.78	2.10	94	6108	120
Mean	10.7	3.3	4.27	8.2	0.70	2.70	8.0	0.69	8.5	0.73	0.92	127	7132	1186
OVERALL														
Median	7.0	2.0	1.93	2.0	0.67	1.36	2.0	0.75	2.0	0.57	7.67	86	2955	25
Mean	7.9	2.1	2.65	2.8	0.64	1.78	3.0	0.68	2.7	0.61	9.95	226	5169	342

table S5. Measurements of interaction free energy between R(1) mutant tetraloop receptors and GGAA tetraloop. A tectoRNA heterodimer system was used to measure differences in loop/receptor affinity between R(1) receptor variants and GGAA tetraloops. Loop/receptor affinities are reported in terms of the change in free energy with respect to the wildtype R(1) receptor's affinity for a GGAA tetraloop [$\Delta G = k_B T \ln(4.4 \text{ nM} / 1 \text{ M}) = –10.83 \text{ kcal mol}^{-1}$]. All ΔΔG values are calculated at 10 °C and expressed in terms of kcal mol^{-1}. Replicates conducted on the C9A mutation reveal measurement variation is generally within 20% or less.

RECEPTOR	ΔΔG GGAA	RECEPTOR	ΔΔG GGAA
R1 WT	0.000		
R1 U3A	+ 3.04	R1 U8A	+ 2.44
R1 U3G	+ 6.03	R1 U8G	+ 2.05
R1 U3C	+ 0.40	R1 U8C	+ 2.35
R1 U3del	+ 5.24	R1 U8del	+ 2.55
R1 G4A	+ 3.61	R1 C9A	+ 3.58
R1 G4U	+ 3.51	R1 C9U	+ 3.15
R1 G4C	+ 3.91	R1 C9G	+ 4.39
R1 G4del	+ 4.17	R1 C9del	+ 3.69
R1 U5A	+ 4.40	R1 U10A	+ 4.13
R1 U5G	+ 3.68	R1 U10G	+ 5.60
R1 U5C	+ 3.93	R1U10C	+ 2.15
R1 U5del	+ 4.07	R1 U10del	+ 4.46
R1 G6A	+ 1.89	R1 4A.5A	+ 3.72
R1 G6U	+ 2.70	R1 4A.9U	+ 3.38
R1 G6C	+ 0.52	R1 4C.9G	+ 2.22
R1 G6del	+ 4.47	R1 4U.9A	+ 3.18
R1 A7U	+ 3.56	R1 4A.5A.9U	+ 4.80
R1 A7G	+ 1.90	R1 4A.5C.9U	+ 3.07
R1 A7C	+ 4.18	R1 6U.7G	+ 3.08
R1 A7del	+ 3.75	R1 6A.7G	+ 0.07