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.
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.](image-url)

(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 23S 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 23S rRNA loop (1572) (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 GpG 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 (1.25 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 multihelix 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 isosine 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
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 ($P < 5 \times 10^{-2}$) and RMSD accuracy ($P < 2 \times 10^{-2}$) ($P$ 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 $k_B T$ (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...
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/Sa 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 × 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 GAAA/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 canonical Watson-Crick U3-U10 pair. In GGAA/R(1), where the second tetraloop nucleotide is G, the platform pairing is predicted to be homologous to A8 in the 11-nt receptor based on mutagenesis data indicating its importance (22), and G-U in the R(1) receptor. While C9 was previously proposed to be homologous in the 11-nt receptor data based on mutagenesis, this new model also explains Watson-Crick pairs newly predicted in the literature did not occur in the models. 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 xRNA) 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.
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 (Finzymes). DNA templates were designed with the T7 RNA polymerase promoter (5′-TTCTAATACGACTCACTATAG-3′) at their 5′ ends. A custom reverse transcription primer-binding site (5′-AAAGAAAACACAACACACACAC-3′) was included at the 3′ terminus of each template. RNA transcribed with T7 RNA polymerase (New England Biolabs) was purified using RNA cleans 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-Hepe (pH 8.0) at 90°C for 3 min and then cooled to room temperature. MgCl2 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 HITTRACE (33). Data from these P4-P6 RNA experiments have been posted to the RNA Mapping Database (34) at the following accession IDs:

| Accession Number | Description                          |
|------------------|--------------------------------------|
| TRP4P6_WT2_0000  | DMS, CMCT, and 1M7 for GAAA/11-nt     |
| TRP4P6_C72_0000  | DMS, CMCT, and 1M7 for GAAA/C7.2      |
| TRP4P6_R1J_0000  | DMS, CMCT, and 1M7 for GAAA/C7.10     |
| 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 MgCl2 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 MOPS (pH 10)]. One of the monomers contained a fixed amount of 3′ end [32P] ppCp-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 $ΔG = k_BT\ln(K_d/1 M)$, 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

**REFERENCES AND NOTES**

1. P. Bradley, K. M. S. Misura, D. Baker, Toward high-resolution de novo structure prediction for small proteins. Science 309, 1668–1671 (2005).
2. Z. Xiao, R. W. Adamiaik, M. Antczak, R. T. Batyev, A. I. Becka, M. Biesiada, M. J. Boniecki, J. M. Bujnicki, S.-J. Chen, C. Y. Cheng, F.-C. Chou, A. R. Ferré-D’Amare, R. Das, W. K. Dawson, F. Ding, N. V. Dokholyan, S. Dunin-Horkavicz, C. Geniesse, K. Kappel, W. Kladwang, A. Krokhintov, G. E. Lach, F. Major, T. H. Mann, M. Magnus, K. Pachula-Wiezorek, D. J. Patel, J. A. Piccirilli, M. Popenda, K. J. Purzycka, A. Ren, G. M. Rice, J. Santalucia Jr., J. Sarzynska, M. Szachniak, A. Tandon, J. J. Trausch, S. Tian, J. Wang, K. M. Weeks, B. Williams Il, 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. Feldmeyer, F. Parmeggiani, D. A. F. Velasco, B. Hörcker, 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. Krokhintov, K. Houilhan, N. V. Dokholyan, iFoldRNA v2: Folding RNA with constraints. Bioinformatics 31, 2891–2893 (2015).
8. R. Das, J. Karanicolias, 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 (2015).
10. M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K. M. Roether, J. M. Bujnicki, SimRNA: A coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic Acids Res. 44, e63 (2016).
11. P. Sripadkeevong, 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. Sripadkeevong, M. C. Erat, M. Ziegeler, Q. Zhao, G. E. Fox, X. Gao, S. D. Kennedy, R. Kierzek, E. P. Nikonovicz, H. Schwalbe, R. K. O. Siegel, D. H. Turner, R. Das, Structure determination of noncanonical RNA motifs guided by H NMR chemical shifts. Nat. Methods 11, 413–416 (2014).
17. J. Feringhoff-Borg, In Bayesian Methods in Structural Bioinformatics, T. Hameleyck, K. Maridja, J. Feringhoff-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. Sripadkeevong, M. S. Dibrov, T. Herrmann, 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. C. Erat, M. Ziegeler, Q. Zhao, G. E. Fox, S. 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).
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. Jareg, 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).
This page contains a list of scientific references, likely from a research article or review, formatted in a way that suggests it is a continuation of the previous page. The references are cited in numerical order, with each entry indicating the authors, title, journal, volume, issue, page numbers, and publication year. The format is consistent with that of academic publications, adhering to the style guide of the journal or conference proceedings. The references span various fields, including biochemistry, molecular biology, and chemistry, with topics ranging from RNA mapping and tertiary structure to protein modeling and structural biology. The text is formatted in a standard way, with authors' names, titles, and publication details clearly listed.
82. R. T. Byrne, A. L. Konevega, M. V. Rodnina, A. A. Antson, The crystal structure of unmodified tRNA\textsuperscript{Phe} from Escherichia coli. \textit{Nucleic Acids Res.} \textbf{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. \textit{Acta Crystallogr. Sect. D Biol. Crystallogr.} \textbf{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. \textit{Mol. Cell} \textbf{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. \textit{Angew. Chem. Int. Ed. Engl.} \textbf{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. \textit{Proc. Natl. Acad. Sci. U.S.A.} \textbf{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. \textit{Nat. Chem. Biol.} \textbf{11}, 840–846 (2015).

Acknowledgments: We thank F.-C. Chou for early discussions and Stanford Research Computing Center for expert administration of the BioX\textsuperscript{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. \textit{Sci. Adv.} \textbf{4}, eaar5316 (2018).