Structure determination of noncanonical RNA motifs guided by ^{1}H NMR chemical shifts

Parin Sripakdeevong[1], Mirko Cevec[2], Andrew T Chang[3], Michèle C Erat[4,5], Melanie Ziegeler[2], Qin Zhao[6], George E Fox[7], Xiaolian Gao[7], Scott D Kennedy[8], Ryszard Kierzek[9], Edward P Nikonowicz[3], Harald Schwalbe[2], Roland K O Sigel[5], Douglas H Turner[10] & Rhiju Das[1,11,12]

Structured noncoding RNAs underlie fundamental cellular processes, but determining their three-dimensional structures remains challenging. We demonstrate that integrating ^{1}H NMR chemical shift data with Rosetta *de novo* modeling can be used to consistently determine high-resolution RNA structures. On a benchmark set of 23 noncanonical RNA motifs, including 11 'blind' targets, chemical-shift Rosetta for RNA (CS-Rosetta-RNA) recovered experimental structures with high accuracy (0.6–2.0 Å all-heavy-atom r.m.s. deviation) in 18 cases.

Noncoding RNA molecules form complex three-dimensional structures that have key roles in a multitude of cellular processes from gene regulation to viral pathogenesis[1]. These RNAs are typically composed of canonical helices interconnected by motifs with intricate noncanonical structures critical for catalysis, binding proteins and higher-order folding. Often comprising a few dozen nucleotides or less, these motifs are compelling targets for solution NMR spectroscopy approaches[2]. Nevertheless, NMR spectroscopy–based characterization of RNA motifs does not always generate sufficient nuclear Overhauser effect (NOE) or other restraints to produce reliable atomic-resolution three-dimensional (3D) models[3–6].

NMR chemical shifts can be an important additional source of structural information for functional macromolecules. In protein studies, backbone chemical shifts are widely used to constrain protein secondary structures and backbone torsions[7], and to refine 3D models[8]. More recently, chemical shift data have been leveraged for *de novo* determination of protein structure[9,10]. Similar tools for RNA are less developed. Chemical shift assignments through NOE spectroscopy and through bond-correlation spectroscopy experiments are standard first steps in NMR spectroscopy of RNA, but chemical shift values are generally not used at the structure-determination stage[2]. Algorithms have been developed to 'back-calculate' non-exchangeable ^{1}H chemical shifts from RNA 3D structure[11,12]. In particular, the Nuchemics[12] program has been used to refine models[13] generated from conventional NMR spectroscopy measurements (NOE, *J*-couplings, residual dipolar couplings) and to determine *de novo* structures of simple helical forms of nucleic acids[14]. A recent study also demonstrated that chemical shift data could be used to stringently constrain RNA molecular dynamics simulations starting from a known structure[15]. This study hypothesized that chemical shift–based modeling without previous knowledge of the structure should be possible, but such *de novo* structure determination has not yet been demonstrated, to our knowledge.

Here we show that assigned ^{1}H chemical shift data provide sufficient information to determine the structures of noncanonical RNA motifs at high resolution, by integrating these data with recent advances in high-resolution RNA *de novo* structure prediction[16,17]. We named the method CS-Rosetta-RNA and extensively benchmarked it on 23 RNA motifs, including 11 motifs for which conventional NMR structural models were unreleased to the public and were kept hidden from the modelers (here referred to as 'blind' targets). CS-Rosetta-RNA is freely available through a web server at http://rosie.rosettacommons.org/rna_denovo.

Methods for prediction of RNA structure by fragment assembly of RNA with full-atom refinement (FARFAR)[16] and stepwise assembly (SWA)[17] have permitted the modeling of RNA motifs that give atomic-resolution agreement to experimentally determined structures in favorable cases[16,17]. However, as in protein studies, inaccuracies in available energy functions preclude high-resolution modeling in many cases[18]. Fortunately, in such cases correct structures are still sampled[17], and even quite sparse experimental data can be used to identify these models with high confidence[10,18]. We illustrate the use of CS-Rosetta-RNA with a complex RNA test motif that was challenging for prior Rosetta approaches, a conserved UUAAGU hexaloop from 16S ribosomal RNA (**Fig. 1a**). Standard Rosetta modeling[16,17] without the use of chemical shift information generated models with atomic-resolution agreement to this hexaloop's crystallographic

[1]Biophysics Program, Stanford University, Stanford, California, USA. [2]Center for Biomolecular Magnetic Resonance, Institute for Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe University Frankfurt, Frankfurt, Germany. [3]Department of Biochemistry and Cell Biology, Rice University, Houston, Texas, USA. [4]Department of Biochemistry, University of Oxford, Oxford, UK. [5]Institute of Inorganic Chemistry, University of Zurich, Zurich, Switzerland. [6]Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA. [7]Department of Biology and Biochemistry, University of Houston, Houston, Texas, USA. [8]Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, New York, USA. [9]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. [10]Department of Chemistry, University of Rochester, Rochester, New York, USA. [11]Department of Biochemistry, Stanford University, Stanford, California, USA. [12]Department of Physics, Stanford University, Stanford, California, USA. Correspondence should be addressed to R.D. (rhiju@stanford.edu).

RECEIVED 17 APRIL 2013; ACCEPTED 6 JANUARY 2014; PUBLISHED ONLINE 2 MARCH 2014; DOI:10.1038/NMETH.2876

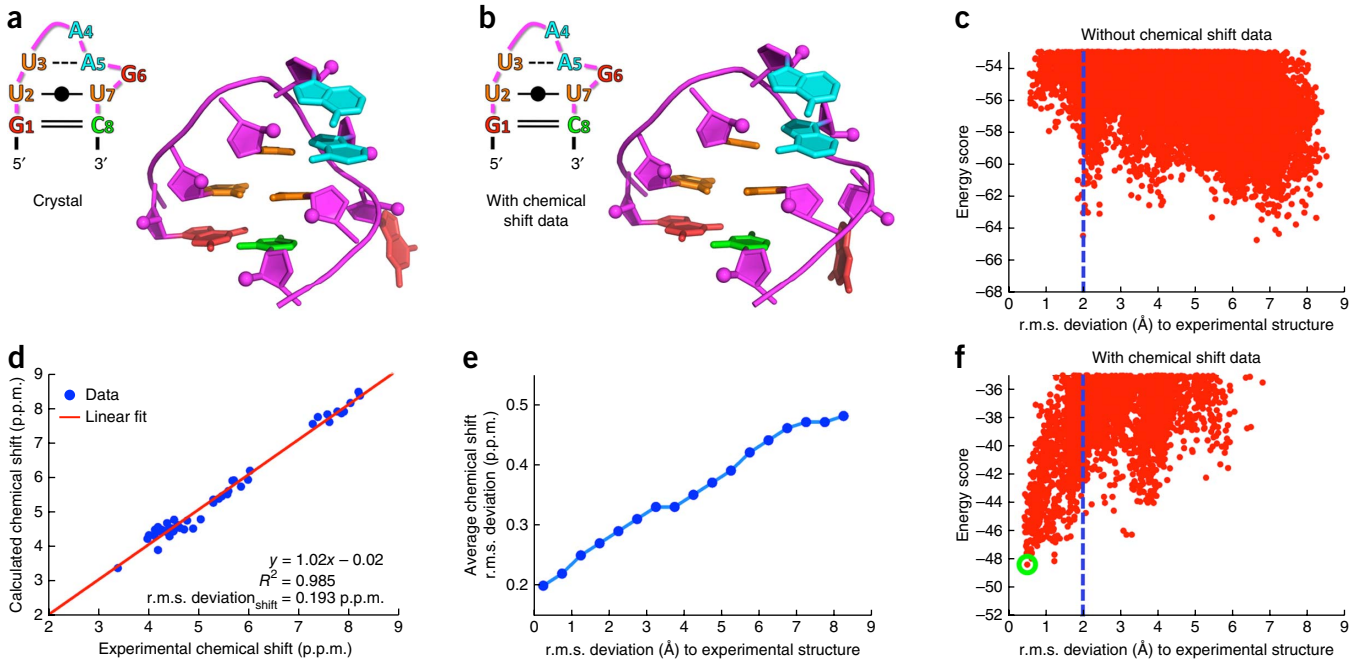


Figure 1 | CS-Rosetta-RNA illustrated on an UUAAGU hairpin. (**a**,**b**) Crystallographic structure (**a**; PDB identifier 1FJG) (**a**) and Rosetta[16,17] near-native (first-ranked) model with a 0.52 Å all-heavy-atom r.m.s. deviation to the crystallographic structure (**b**; r.m.s. deviation was calculated over the entire loop, excluding the flexible G6 extrahelical bulge). Two-dimensional schematics (top left) follow nomenclature in ref. 20. (**c**) Rosetta energy versus r.m.s. deviation to the crystallographic structure for all Rosetta models before the inclusion of the chemical shift pseudo-energy term. Dashed line marks r.m.s. deviation cutoff used herein to evaluate success in high-resolution structure modeling. (**d**) Back-calculated chemical shifts from the Rosetta near-native model versus experimental ^{1}H chemical shift values (first-ranked model; r.m.s. deviation$_{shift}$ = 0.19 p.p.m.). (**e**) Average r.m.s. deviation$_{shift}$ of all Rosetta models in separate bins of 0.5-Å r.m.s. deviation from the crystallographic structure; lines connect bin centers. (**f**) Rosetta energy versus r.m.s. deviation to the crystallographic structure for all Rosetta models after the inclusion of the chemical shift pseudo-energy term. With chemical shift data, the near-native model shown in **b** becomes the lowest energy model overall (green circle).

structure (0.52 Å all-heavy-atom r.m.s. deviation; **Fig. 1a**,**b**), but these models were ranked worse in computed Rosetta energy than non-native models (>5.0 Å r.m.s. deviation; **Fig. 1c**). Nevertheless, the experimentally measured chemical shifts of the non-exchangeable ^{1}H atoms were in strong agreement with the predicted chemical shifts from the near-native models but not from any of the non-native models (**Fig. 1d**,**e** and **Supplementary Figs. 1** and **2**). Supplementing the Rosetta energy function with a chemical–shift–based pseudo-energy score (E_{shift}; Online Methods) then permitted confident discrimination of the near-native models (**Fig. 1f**; see **Supplementary Results** for further discussions on the importance of base and ribose proton chemical shifts for recovering the native structure).

To evaluate the generality and accuracy of CS-Rosetta-RNA, we carried out modeling on a benchmark set of 23 RNA motifs (**Table 1** and Online Methods). First, we applied CS-Rosetta-RNA to a test set of 12 noncanonical motifs for which published chemical shift data as well as structural models derived from NMR data and, in some cases, crystallography data, were available (**Supplementary Table 1**). These RNA motifs included hairpins, internal loops, a three-way junction and a tetraloop-receptor interaction. On average, each data set included 6.0 non-exchangeable ^{1}H chemical shifts per nucleotide (out of 7–8 possible), including both ribose and base protons (**Supplementary Table 1**). We tested CS-Rosetta-RNA on 11 blind RNA targets that were concurrently under investigation in five NMR spectroscopy laboratories. Sequences and assigned chemical shifts for these targets, but no other information, were provided by researchers in these laboratories to the authors of this work carrying out chemical shift–guided modeling. Subsequent comparison of CS-Rosetta-RNA models with structures derived from conventional NMR spectroscopy approaches thus served as rigorous evaluation of blind targets.

Over the entire benchmark set of 23 RNA motifs, CS-Rosetta-RNA returned 18 'success' cases, defined here as cases in which at least one of the five lowest-energy cluster centers achieved better than 2.0 Å all-heavy-atom r.m.s. deviation (r.m.s. deviation values and cluster ranks are provided in **Table 1** and **Supplementary Tables 2** and **3**; energy versus r.m.s. deviation plots are provided in **Supplementary Fig. 3**; PDB files of experimental structures and five lowest-energy cluster centers are provided in **Supplementary Data**). In four of the remaining five cases, structural dynamics in solution precluded high-resolution agreement between the NMR spectroscopy structures and the CS-Rosetta-RNA models (**Supplementary Results**, and **Supplementary Figs. 4** and **5**). CS-Rosetta-RNA performed well on both the test set of known structures (10/12 success cases) and the blind targets (8/11 success cases). 11 of the 23 cases satisfied a more stringent success criterion: the lowest-energy (top-ranked) model was within 1.5 Å all-heavy-atom r.m.s. deviation of the experimental structure (**Table 1**). Lastly, incorporating even sparse data (~1 chemical shift per nucleotide) improved accuracy (**Supplementary Results** and **Supplementary Fig. 6**).

CS-Rosetta-RNA success cases included high-resolution models from diverse sources, such as the most conserved internal loop from the signal recognition particle (SRP) RNA (r.m.s. deviation, 0.81 Å; **Fig. 2a**), a GAAA tetraloop–receptor interaction

Table 1 | The CS-Rosetta-RNA method benchmarked on 23 RNA motifs

Motif name	PDB identifier[a]	Motif size[b]	r.m.s. deviation, lowest-energy model[c,d] (Å)	r.m.s. deviation, top five lowest-energy models[c,e] (Å)
Known structures				
Single G-G mismatch	1F5G	6	0.71	0.71
UUCG tetraloop	2KOC	6	0.84	0.84
Tandem GA-AG mismatch	1MIS	8	1.10	1.10
Tandem UG-UA mismatch	2JSE	8	3.02	2.52
16S rRNA UUAAGU loop	1FJG	8	0.52	0.52
HIV-1 TAR apical loop	1ANR	8	5.86	5.86
$tRNA_i^{Met}$ ASL	1SZY	9	3.89	1.35
Conserved SRP internal loop	1LNT	12	0.81	0.81
R2 retrotransposon 4 × 4 loop	2L8F	12	1.17	1.17
Hepatitis C virus IRES IIa	2PN4	13	3.21	1.48
GAAA tetraloop receptor	2R8S	15	0.68	0.68
Sc.ai5γ three-way junction	2LU0	16	3.66	1.74
Blind targets				
UAAC tetraloop[f]	4A4R	6	0.94	0.94
UCAC tetraloop[f]	4A4S	6	1.00	1.00
UGAC tetraloop[f]	4A4U	6	3.60	1.67
UUAC tetraloop[f]	4A4T	6	1.72	1.72
Chimp HAR1 GAA loop	2LHP	7	2.88	2.88
Human HAR1 GAA loop	2LUB	7	2.26	2.03
GU-UAU internal loop	[g]	9	1.37	1.37
$tRNA^{Gly}$ ASL (cuUCCaa)[h]	2LBL	9	3.28	1.41
$tRNA^{Gly}$ ASL (cuUCCcg)[h]	2LBK	9	3.42	1.94
$tRNA^{Gly}$ ASL (uuGCCaa)[h]	2LBJ	9	3.08	2.93
5′-GAGU–3′-UGAG loop	2LX1	12	1.10	1.10
r.m.s. deviation < 1.50 Å			11/23	14/23
r.m.s. deviation < 2.00 Å			12/23	18/23

Additional information and full motif names provided in **Supplementary Tables 1** and **3**.

[a]PDB identifier of reference experimental structure. [b]Number of nucleotides in the modeled RNA motif. Each motif consists of noncanonical core nucleotides closed by boundary canonical (W.C or G:U wobble) base pairs. [c]All-heavy-atom r.m.s. deviation over all nucleotides, excluding the boundary canonical base pairs after alignment over all nucleotides. Nucleotides found to be extrahelical bulges (both unpaired and unstacked) in the reference experimental structure were excluded from both the alignment and the r.m.s. deviation calculation. [d]All-heavy-atom r.m.s. deviation of the first-ranked (lowest energy) model to the experimental structure. [e]Lowest all-heavy-atom r.m.s. deviation to the experimental structure among the five lowest-energy cluster centers. [f]The four UNAC tetraloops were treated as separate motifs despite adopting similar conformations owing to being blind targets. [g]Experimental structure (unpublished data; M.C. Erat and R.K.O. Sigel) has not yet been deposited into PDB. [h]Sequence of the 7-nucleotide anticodon loop is given in parentheses with the anticodon triplet in upper case.

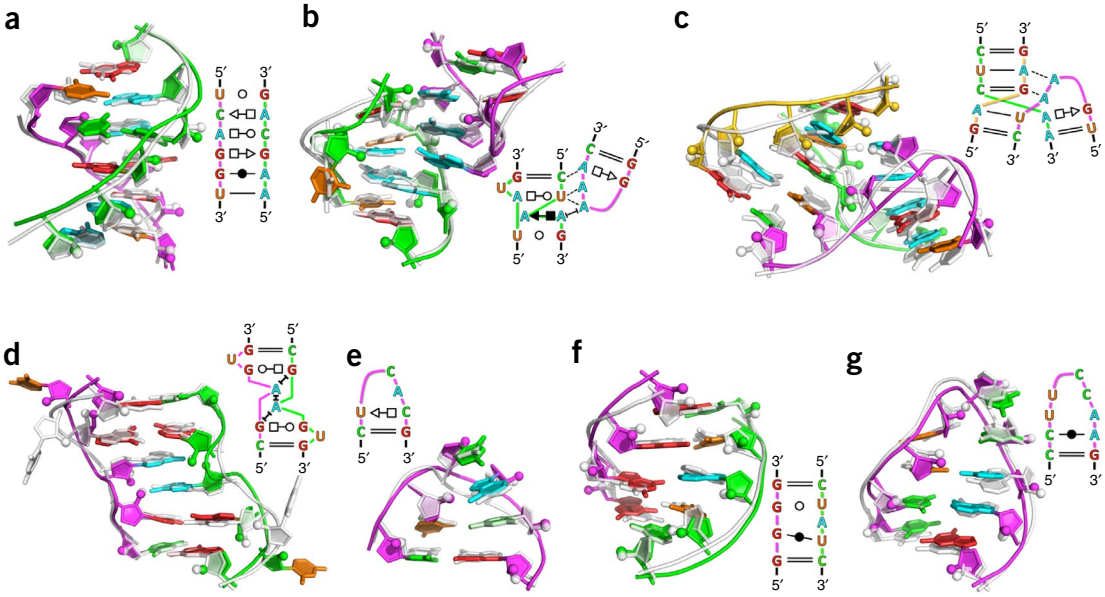


Figure 2 | Comparison of experimental and CS-Rosetta-RNA models for diverse RNA motifs. (**a**–**g**) CS-Rosetta-RNA models (in color) overlaid on the experimental structures (in white) for conserved internal loop from the SRP RNA (**a**; PDB: 1LNT), GAAA tetraloop-receptor tertiary interaction motif (**b**; PDB: 2R8S), three-way junction from yeast mitochondrial group II intron Sc.ai5γ (**c**; PDB: 2LU0), 5′-GAGU-3′–3′-UGAG-5′ self-complementary internal loop (**d**; PDB: 2LX1), UCAC tetraloop (**e**; PDB: 4A4S), 5′-GU-3′–3′-UAU-5′ internal loop from a group II intron (**f**), glycine tRNA(UCC) anticodon stem-loop from *Bacillus subtilis* (**g**; PDB: 2LBL). The r.m.s. deviation values between CS-Rosetta-RNA models (rank of model by energy given in parentheses) and the experimental structure are (**a**) 0.81 Å (ranked first), (**b**) 0.68 Å (first), (**c**) 1.74 Å (fourth), (**d**) 1.10 Å (first), (**e**) 1.00 Å (first), (**f**) 1.37 (first) and (**g**) 1.41 Å (third). The two-dimensional schematics are annotated based on the experimental structure and follow nomenclature in ref. 20.

(r.m.s. deviation, 0.68 Å; **Fig. 2b**), a three-way junction from yeast mitochondrial group II intron Sc.ai5γ (r.m.s. deviation, 1.74 Å; **Fig. 2c**), and both the major and minor conformations of a G:G mismatch (**Supplementary Fig. 7**). Successful blind target cases included predictions for a highly irregular 5′-GAGU-3′–3′-UGAG-5′ self-complementary internal loop that required synthesizing and probing additional constructs to solve by conventional NMR spectroscopy (r.m.s. deviation, 1.10 Å; **Fig. 2d**), all four UNAC tetraloops (where N refers to any nucleotide; **Fig. 2e**), a 5′-GU-3′–3′-UAU-5′ internal loop from a group II intron (r.m.s. deviation, 1.37 Å; **Fig. 2f**) and a CUUCCAA anticodon stem-loop of *Bacillus subtilis* $tRNA^{Gly}$ (r.m.s. deviation, 1.41 Å; **Fig. 2g**).

Several CS-Rosetta-RNA predictions gave strong convergence, as defined by a distinct energy 'funnel': a single dominant conformation and geometrically similar models achieved better energy than all other conformations. In seven benchmark cases, the lowest-energy model gave an energy gap of >3.0 Rosetta units (approximately equal to k_BT, where k_B is the Boltzmann constant and T is temperature, 37 °C) to the next-lowest energy cluster and, in all of these cases, the model achieved better than 1.5 Å r.m.s. deviation to experimental structure (**Supplementary Fig. 8**). This energy gap thus appears to be a hallmark of CS-Rosetta-RNA accuracy (**Supplementary Results**). In one apparent exception, the SRP conserved internal loop, a large energy gap (5.5 Rosetta units) strongly suggested that the CS-Rosetta-RNA prediction should be accurate, but the lowest-energy CS-Rosetta-RNA model disagreed with the experimental NMR spectroscopy models[3] (>2.0 Å r.m.s. deviation; **Supplementary Fig. 9a**,**b**). Additional analysis revealed that the experimental NMR spectroscopy models poorly explained the ^{1}H chemical shift data published in the same study[3] (r.m.s. $\text{deviation}_{\text{shift}}$ = 0.50 p.p.m.) and poorly agreed with subsequently solved crystallographic structures[4,19] (r.m.s. deviation of 2.30 Å to Protein Data Bank (PDB) identifier 1LNT[19]). In contrast, the CS-Rosetta-RNA model gave excellent agreement with the chemical shift data (r.m.s. $\text{deviation}_{\text{shift}}$ = 0.18 p.p.m.) and closely matched the crystallographic structures (r.m.s. deviation of 0.81 Å to PDB identifier 1LNT; **Fig. 2a** and **Supplementary Fig. 9c**,**d**). The SRP motif case supports the use of CS-Rosetta-RNA as a tool to independently cross-validate or remodel NMR spectroscopy–derived structures.

CS-Rosetta-RNA enables confident determination of noncanonical RNA motif structures in a manner fundamentally distinct from prior methods, using independent and far less experimental information. The standard approach[2] of determining NOEs, *J*-couplings and, in some cases, residual dipolar couplings, does not always yield sufficient information to determine an RNA's 3D structure by conventional means, as illustrated by the 5′-GAGU-3′–3′-UGAG-5′ case (**Fig. 2d**; see **Supplementary Note**, and **Supplementary Figs. 10** and **11** for further modeling details of this highly irregular motif). Further integration of *de novo* modeling and NMR methodologies, including the incorporation of ^{13}C, ^{15}N and exchangeable ^{1}H chemical shift data (**Supplementary Results**), may help accelerate determination of RNA structure and eventually help solve currently intractable 3D RNA structures.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank H. Al-Hashimi for suggesting the problem and the Rosetta community for sharing code, and G. Varani and J.D. Puglisi for providing experimental ^{1}H chemical shift data (for HIV-1 TAR apical loop and hepatitis C virus IRES subdomain IIa, respectively). Calculations were carried out on the BioX^2 cluster (US National Science Foundation award CNS-0619926) and Extreme Science and Engineering Discovery Environment (XSEDE) (allocation MCB090153). We acknowledge financial support from a Burroughs Wellcome Career Award at the Scientific Interface (to R.D.), US National Institutes of Health (NIH) grant R21 GM102716 (to R.D.), a C.V. Starr Asia/Pacific Stanford Graduate Fellowship (to P.S.), NIH grant GM73969 (to E.P.N.), NIH grant GM22939 (to D.H.T.), the Swiss National Science Foundation and the European Commission (BIO-NMR, ERC-Starting Grant, Marie Curie Fellowship; to M.C.E. and R.K.O.S.), and the Deutsche Forschungsgemeinschaft (Cluster of Excellence) and the European Commission (Bio-NMR, WeNMR, Marie Curie Fellowship; to M.C., M.Z. and H.S.).

AUTHOR CONTRIBUTIONS

P.S. and R.D. designed the research. P.S. implemented the method, generated the data, analyzed the results, and wrote the paper. R.D. assisted in analyzing the data and writing the paper. M.C., A.T.C., M.C.E., M.Z., Q.Z., G.E.F., X.G., S.D.K., R.K., E.P.N., H.S., R.K.O.S. and D.H.T. provided NMR spectroscopy data for the 11 blind targets and participated in evaluating the blinded trials. All authors discussed the results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Gesteland, R.F., Cech, T. & Atkins, J.F. *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*. Vol. 43 (Cold Spring Harbor Lab Press, 2006).
2. Scott, L.G. & Hennig, M. *Methods Mol. Biol.* **452**, 29–61 (2008).
3. Schmitz, U., James, T.L., Lukavsky, P. & Walter, P. *Nat. Struct. Biol.* **6**, 634–638 (1999).
4. Jovine, L. *et al. Structure* **8**, 527–540 (2000).
5. Nabuurs, S.B., Spronk, C.A.E.M., Vuister, G.W. & Vriend, G. *PLoS Comput. Biol.* **2**, e9 (2006).
6. Tolbert, B.S. *et al. J. Biomol. NMR* **47**, 205–219 (2010).
7. Cornilescu, G., Delaglio, F. & Bax, A. *J. Biomol. NMR* **13**, 289–302 (1999).
8. Clore, G.M. & Gronenborn, A.M. *Proc. Natl. Acad. Sci. USA* **95**, 5891–5898 (1998).
9. Cavalli, A., Salvatella, X., Dobson, C.M. & Vendruscolo, M. *Proc. Natl. Acad. Sci. USA* **104**, 9615–9620 (2007).
10. Shen, Y. *et al. Proc. Natl. Acad. Sci. USA* **105**, 4685–4690 (2008).
11. Case, D.A. *J. Biomol. NMR* **6**, 341–346 (1995).
12. Cromsigt, J.A., Hilbers, C.W. & Wijmenga, S.S. *J. Biomol. NMR* **21**, 11–29 (2001).
13. Girard, F.C., Ottink, O.M., Ampt, K.A., Tessari, M. & Wijmenga, S.S. *Nucleic Acids Res.* **35**, 2800–2811 (2007).
14. van der Werf, R.M., Tessari, M. & Wijmenga, S.S. *J. Biomol. NMR* **56**, 95–112 (2013).
15. Frank, A.T., Horowitz, S., Andricioaei, I. & Al-Hashimi, H.M. *J. Phys. Chem. B* **117**, 2045–2052 (2013).
16. Das, R., Karanicolas, J. & Baker, D. *Nat. Methods* **7**, 291–294 (2010).
17. Sripakdeevong, P., Kladwang, W. & Das, R. *Proc. Natl. Acad. Sci. USA* **108**, 20573–20578 (2011).
18. Fleishman, S.J. & Baker, D. *Cell* **149**, 262–273 (2012).
19. Deng, J., Xiong, Y., Pan, B. & Sundaralingam, M. *Acta Crystallogr. D Biol. Crystallogr.* **59**, 1004–1011 (2003).
20. Leontis, N.B. & Westhof, E. *RNA* **7**, 499–512 (2001).

ONLINE METHODS

Generation of Rosetta models. Two complementary structure modeling methods, FARFAR[16] and SWA[17], were used in parallel to generate the Rosetta models for each motif. SWA models were constructed using a series of recursive building steps, as described previously[17]. Each step involved enumerating several million conformations for each nucleotide, and all step-by-step build-up paths were covered in N^2 building steps, where N is the number of nucleotides in the motif. At the final building steps, all models were finely clustered and a maximum of 10,000 low-energy SWA models were retained. The SWA approach is effective at generating models that are highly optimized with respect to the underlying all-atom energy function but can produce primarily incorrect models when the assumed energy function is inaccurate. Therefore, models were also generated by FARFAR in the Rosetta framework, as described previously[16]; the fragment source was the large ribosomal subunit of *H. marismortui* (PDB: 1JJ2). For each motif, 250,000 FARFAR models were generated; these models were then finely clustered and a maximum of 10,000 low-energy FARFAR models were retained. The SWA and FARFAR models were then combined, which led to ~10,000–20,000 final Rosetta models for each motif. The SWA method was used to model all 23 RNA motifs in the benchmark except for the GAAA tetraloop-receptor interaction and the Sc.ai5γ three-way junction. The FARFAR method was used to model all 23 RNA motifs in the benchmark except for the 5′-GAGU-3′–3′-UGAG-5′ RNA structural switch (**Supplementary Note**). Algorithms and complete documentation are incorporated into Rosetta release 3.5, freely available for academic use.

The total computational costs for the generation of SWA and FARFAR models in term of modern central processing units (CPUs) are as follows. For SWA runs, the computational cost ranged from ~5,000 CPU hours for a 6-nucleotide motif to ~50,000 CPU hours for the 13-nucleotide motif investigated in this work (using Intel Xeon E5345 2.33-GHz CPUs). For FARFAR runs, the computational cost ranged from ~3,000 CPU hours for the 6-nucleotide motif to ~8,000 CPU hours for the 13-nucleotide motif. The majority of the computations for this work were performed on Stanford University's Bio-X^2 cluster, a supercomputer with 2,208 CPUs (Intel Xeon E5345 2.33 GHz). When using 500 CPU (the maximum allocated to each user), it takes less than half a day (of wall-clock time) to perform 5,000 CPU hours of computation and less than 5 d (of wall-clock time) to perform 50,000 CPU hours of computation.

Incorporation of non-exchangeable ^{1}H chemical shifts into structure modeling with CS-Rosetta-RNA. Information from the experimental non-exchangeable ^{1}H chemical shifts was incorporated into the modeling process through the chemical shift pseudo-energy term:

$$E_{\text{shift}} = c \times \sum_i \left(\delta_i^{\text{exp}} - \delta_i^{\text{calc}}\right)^2 \quad (1)$$

where δ_i^{exp} and δ_i^{calc} are, respectively, the experimental and back-calculated chemical shift in p.p.m. units (the index i sums over all experimentally assigned non-exchangeable ^{1}H chemical shifts in the RNA motif), and c is a weighting factor set to 4.0 k_BT/p.p.m.^{2} based on test runs with different motifs. The Nuchemics program[12] was used to back-calculate non-exchangeable ^{1}H chemical shifts. In the 23-RNA-motif benchmark set, only three chemical shift data sets (UUCG tetraloop, chimp human accelerated region 1 (HAR1) GAA loop and human HAR1 GAA loop) included stereospecific assignments of the diastereotopic 1H5′ and 2H5′ protons pair. For the remaining 20 chemical shift data sets, the assignment of 1H5′ and 2H5′ was determined for each model based on which values gave better agreement between the experimental and back-calculated chemical shifts.

Each Rosetta model was refined and rescored under the hybrid all-atom energy:

$$E_{\text{hybrid}} = E_{\text{Rosetta}} + E_{\text{shift}} \quad (2)$$

where E_{Rosetta} is the standard Rosetta all-atom energy function for RNA[16], and E_{shift} is the chemical shift pseudo-energy term. Refinement of the models under the E_{hybrid} all-atom energy function was carried out using continuous minimization in torsional space with the Davidson-Fletcher-Powell algorithm under the Rosetta framework. For this purpose, the Nuchemics[12] algorithm was rewritten inside the Rosetta code base, http://www.rosettacommons.org. After refinement, the models were rescored and reranked under the E_{hybrid} all-atom energy function. Finally, all models were clustered, such that models with pairwise all-heavy-atom r.m.s. deviation below 1.5 Å were grouped. The lowest-energy member of each cluster was designated as the cluster center and the five lowest energy cluster centers were designated the CS-Rosetta-RNA predictions.

Sources of experimental PDB structures used in this study. The sources of the experimental PDB structures used in the 23-RNA benchmark were: (1) single G:G mismatch (PDB: 1F5G[21], PDB: 1F5H[21]); (2) UUCG tetraloop (PDB: 2KOC[22], PDB: 1F7Y[23]); (3) tandem GA:AG mismatch (PDB: 1MIS[24]); (4) tandem UG:UA mismatch (PDB: 2JSE[25]); (5) 16S rRNA UUAAGU (PDB: 1FJG[26], PDB: 1HS2; ref. 27); (6) HIV-1 TAR apical loop (PDB: 1ANR[28]); (7) tRNA$_i^{\text{Met}}$ ASL (PDB: 1SZY[29]); (8) conserved SRP internal loop (PDB: 1LNT[19], PDB: 28SR[3], PDB: 28SP[3]); (9) R2 retrotransposon 4x4 loop (PDB: 2L8F[30]); (10) hepatitis C virus IRES IIa (PDB: 2PN4; ref. 31, PDB: 1P5M[32]); (11) GAAA tetraloop-receptor (PDB: 2R8S[33], PDB: 2ADT[34]); (12) Sc.ai5γ 3-way junction (PDB: 2LU0; ref. 35); (13) UAAC tetraloop (PDB: 4A4R[36]); (14) UCAC tetraloop (PDB: 4A4S[36]); (15) UGAC tetraloop (PDB: 4A4U[36]); (16) UUAC tetraloop (PDB: 4A4T[36]); (17) chimp HAR1 GAA loop (PDB: 2LHP[37]); (18) human HAR1 GAA loop (PDB: 2LUB[37]); (19) GU:UAU internal loop (unpublished data; M.C. Erat and R.K.O. Sigel; not yet deposited into PDB); (20) tRNA$^{\text{Gly}}$ ASL (cuUCCaa) (PDB: 2LBL[38]); (21) tRNA$^{\text{Gly}}$ ASL (cuUCCcg) (PDB: 2LBK[38]); (22) tRNA$^{\text{Gly}}$ ASL (uuGCCaa) (PDB: 2LBJ[38]); and (23) 5′-GAGU–3′-UGAG loop (PDB: 2LX1; ref. 39).

CS-Rosetta-RNA web server. To encourage usage of the CS-Rosetta-RNA method by the general NMR spectroscopy RNA community, a public web server where users can access and submit CS-Rosetta-RNA modeling jobs is made freely available at http://rosie.rosettacommons.org/rna_denovo. Documentation and tutorials on how to submit the modeling jobs are also provided at the website. Owing to computational resource limitations and to ensure short queue time, the web server runs a slightly

modified version of CS-Rosetta-RNA in which models are generated using only the FARFAR method and the maximum number of models per job submission is limited to 50,000.

21. Burkard, M.E. & Turner, D.H. *Biochemistry* **39**, 11748–11762 (2000).
22. Nozinovic, S., Furtig, B., Jonker, H.R., Richter, C. & Schwalbe, H. *Nucleic Acids Res.* **38**, 683–694 (2010).
23. Ennifar, E. *et al. J. Mol. Biol.* **304**, 35–42 (2000).
24. Wu, M. & Turner, D.H. *Biochemistry* **35**, 9677–9689 (1996).
25. Shankar, N. *et al. Biochemistry* **46**, 12665–12678 (2007).
26. Carter, A.P. *et al. Nature* **407**, 340–348 (2000).
27. Zhang, H., Fountain, M.A. & Krugh, T.R. *Biochemistry* **40**, 9879–9886 (2001).
28. Aboul-ela, F., Karn, J. & Varani, G. *Nucleic Acids Res.* **24**, 3974–3981 (1996).
29. Schweisguth, D.C. & Moore, P.B. *J. Mol. Biol.* **267**, 505–519 (1997).
30. Lerman, Y.V. *et al. RNA* **17**, 1664–1677 (2011).
31. Zhao, Q., Han, Q., Kissinger, C.R., Hermann, T. & Thompson, P.A. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 436–443 (2008).
32. Lukavsky, P.J., Kim, I., Otto, G.A. & Puglisi, J.D. *Nat. Struct. Biol.* **10**, 1033–1038 (2003).
33. Ye, J.D. *et al. Proc. Natl. Acad. Sci. USA* **105**, 82–87 (2008).
34. Davis, J.H. *et al. J. Mol. Biol.* **351**, 371–382 (2005).
35. Donghi, D., Pechlaner, M., Finazzo, C., Knobloch, B. & Sigel, R.K. *Nucleic Acids Res.* **41**, 2489–2504 (2013).
36. Zhao, Q. *et al. Biopolymers* **97**, 617–628 (2012).
37. Ziegeler, M., Cevec, M., Richter, C. & Schwalbe, H. *ChemBioChem* **13**, 2100–2112 (2012).
38. Chang, A.T. & Nikonowicz, E.P. *Biochemistry* **51**, 3662–3674 (2012).
39. Kennedy, S.D., Kierzek, R. & Turner, D.H. *Biochemistry* **51**, 9257–9259 (2012).

doi:10.1038/nmeth.2876

Supplementary Information for "Structure determination of noncanonical RNA motifs guided by ^{1}H NMR chemical shifts"

Parin Sripakdeevong[1], Mirko Cevec[2], Andrew T. Chang[3], Michèle C. Erat[4,5], Melanie Ziegeler[2], Qin Zhao[6], George E. Fox[7], Xiaolian Gao[7], Scott D. Kennedy[8], Ryszard Kierzek[9], Edward P. Nikonowicz[3], Harald Schwalbe[2], Roland K. O. Sigel[5], Douglas H. Turner[10], and Rhiju Das[1,11,12]

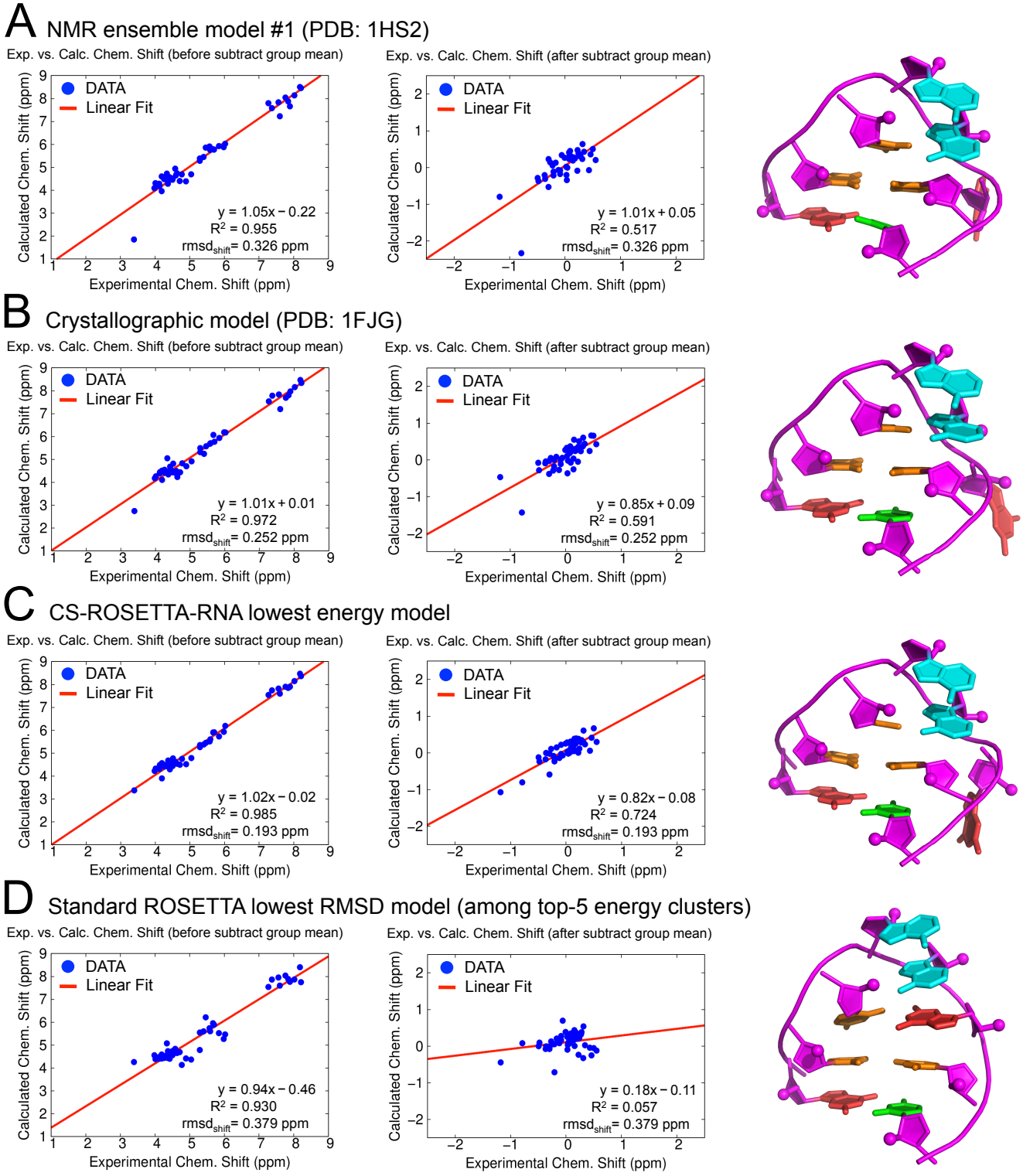
[1]Biophysics Program, Stanford University, Stanford, CA 94305, USA
[2]Center for Biomolecular Magnetic Resonance, Institute for Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe University Frankfurt, Max-von-Laue-Strasse 7, 60438 Frankfurt, Germany
[3]Department of Biochemistry and Cell Biology, Rice University, Houston, TX 77251, USA
[4]Department of Biochemistry, University of Oxford, Oxford OX1 3QU, United Kingdom
[5]Institute of Inorganic Chemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[6]Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA
[7]Department of Biology and Biochemistry, University of Houston, Houston, TX 77004, USA
[8]Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA
[9]Institute of Bioorganic Chemistry, Polish Academy of Sciences, 60-714 Poznan, Noskowskiego 12/14, Poland
[10]Department of Chemistry, University of Rochester, Rochester, NY 14627, USA
[11]Department of Biochemistry, Stanford University, Stanford, CA 94305, USA
[12]Department of Physics, Stanford University, Stanford, CA 94305, USA

To whom correspondence should be addressed. Phone: (650) 723-5976. Fax: (650) 723-6783. E-mail: rhiju@stanford.edu.

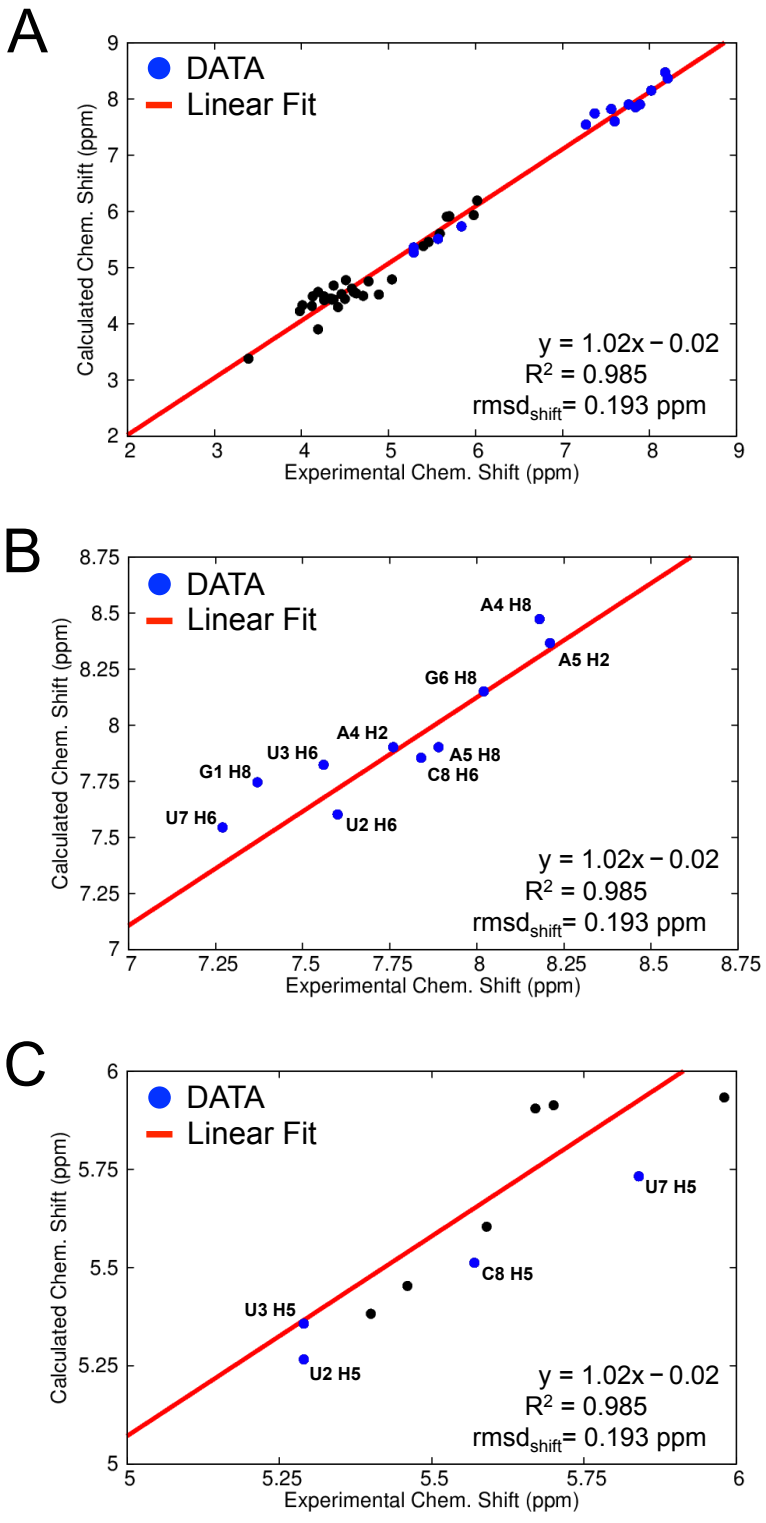
Nature Methods: doi:10.1038/nmeth.2876

Supplementary Item	Title
Supplementary Figure 1	Correlation between back-calculated and experimental ^{1}H chemical shift for the conserved UUAAGU hairpin from the 16S rRNA.
Supplementary Figure 2	Aromatic protons chemical shift correlation plots for the conserved UUAAGU hairpin from the 16S rRNA.
Supplementary Figure 3	Energy vs. all-heavy-atom RMSD to the experimental structure.
Supplementary Figure 4	Four dynamic and/or unstructured motifs from the RNA motif benchmark.
Supplementary Figure 5	Modeling results on the uuGCCaa anticodon stem-loop of *B. subtilis* $tRNA^{Gly}$ by three different models generation methods.
Supplementary Figure 6	Modeling a 13-nucleotide internal loop from hepatitis C virus IRES subdomain IIa using a sparse ^{1}H chemical shift dataset.
Supplementary Figure 7	CS-ROSETTA-RNA modeling of the major and minor conformations of a G:G mismatch.
Supplementary Figure 8	E_{gap} vs. RMSD of the CS-ROSETTA-RNA lowest energy model to the experimental structure.
Supplementary Figure 9	Comparisons between CS-ROSETTA-RNA, NMR and crystallographic models of the most conserved internal loop from the signal recognition particle RNA.
Supplementary Figure 10	CS-ROSETTA-RNA modeling of the 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop.
Supplementary Figure 11	Comparisons between the CS-ROSETTA-RNA and NMR models of the 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop.
Supplementary Table 1	Supplemental information on the RNA motifs benchmark.
Supplementary Table 2	Supplemental Rosetta modeling results (before inclusion of chemical shift data).
Supplementary Table 3	Supplemental Rosetta modeling results (after inclusion of chemical shift data).
Supplementary Results	
Supplementary Note	

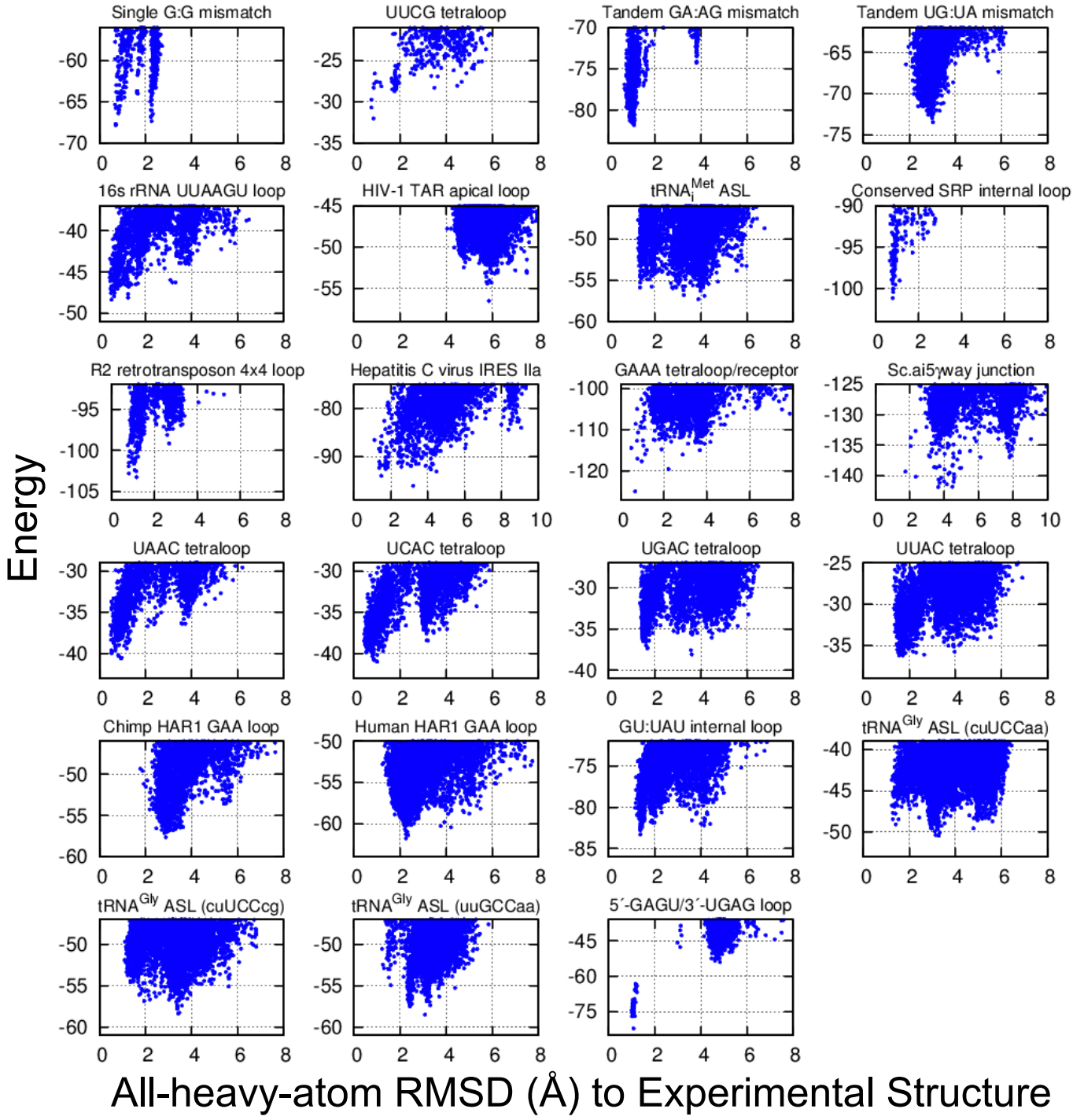
Supplementary Figure 1. Correlation between back-calculated and experimental ^{1}H chemical shifts for the conserved UUAAGU hairpin from the 16S rRNA. The models derived from (A) the NMR study, (B) the crystallographic study, and (C) CS-ROSETTA-RNA all agree to within 1.0 Å rmsd. The chemical shifts back-calculated from all three models also correlate well with the experimentally determined chemical shift ($R^2 > 0.95$; see left column). The correlations remain strong ($R^2 > 0.50$; see middle column), even when variations due to the chemical nature of each proton group (H1′, H2′, H2, H5 and etc.) were removed by subtracting off each proton group's average experimental chemical shift value. (D) In contrast, none of the top-5 standard ROSETTA models (without chemical shift data) adopt the native conformation. Among the top-5 models, model #2 possesses the lowest rmsd (2.0 Å) to the crystal, but even this conformation poorly agrees with the experimental chemical shifts.



Supplementary Figure 2. Aromatic protons chemical shift correlation plots for the conserved UUAAGU hairpin from the 16S rRNA. (A) Correlation between back-calculated and experimental ^{1}H chemical shift for the CS-ROSETTA-RNA with focus on aromatic protons. The aromatic protons (H2, H5, H6, and H8) are colored blue while the remaining non-exchangable ribose protons (H1′, H2′, H3′, H4′, 1H5′, and 2H5′) are colored black. The chemical shifts back-calculated from the CS-ROSETTA-RNA model agree well with the experimentally determined chemical shift over all 46 non-exchangable protons ($rmsd_{shift}$ = 0.193 ppm) and over all 14 aromatic protons ($rmsd_{shift}$ = 0.180 ppm). (B) The chemical shift correlation plot zoomed to visualize H2, H6, and H8 protons chemical shift (between 7.00 ppm and 8.75 ppm). (C) The chemical shift correlation plot zoomed to visualize H5 protons chemical shift (between 5.00 ppm and 6.00 ppm)



Supplementary Figure 3. Energy vs. all-heavy-atom RMSD to the experimental structure. The energy (y-axis) is the sum of the standard Rosetta all-atom energy function for RNA and the chemical shift pseudo-energy term.



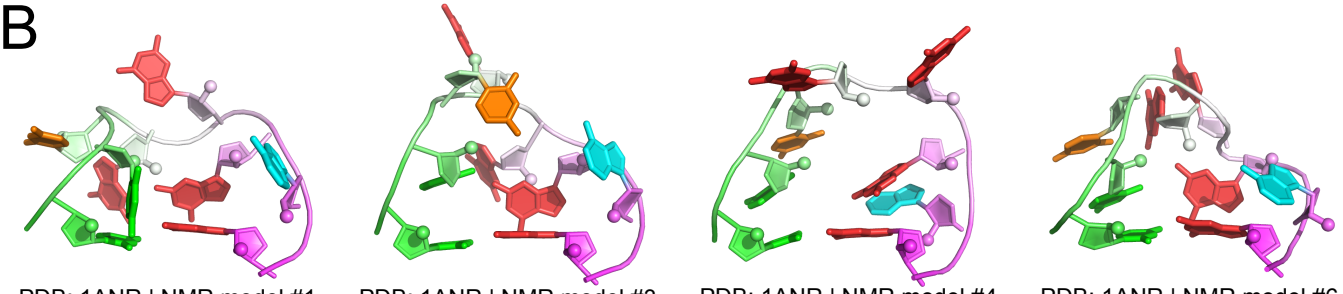
Nature Methods: doi:10.1038/nmeth.2876

Supplementary Figure 4. Four dynamic and/or unstructured motifs from the RNA motif benchmark. (A) The table reports the average rmsd value calculated between every possible pair of models in the NMR-derived ensemble. Inclusion of the well-structured boundary Watson-Crick base pairs in the rmsd calculation lowers the rmsd value. Four structurally diverse models from the NMR-derived ensemble of the (B) HIV-1 TAR apical loop, (C) chimp HAR1F GAA loop, (D) human HAR1F GAA loop, and (E) tandem UG:UA mismatch. In the UG:UA mismatch case, the structural dynamics appear to be localized to a single guanosine nucleotide although the motif is known to be thermodynamically destabilizing and contains no base pairs [Biochemistry. 2007 Nov 6;46(44):12665-78].

A

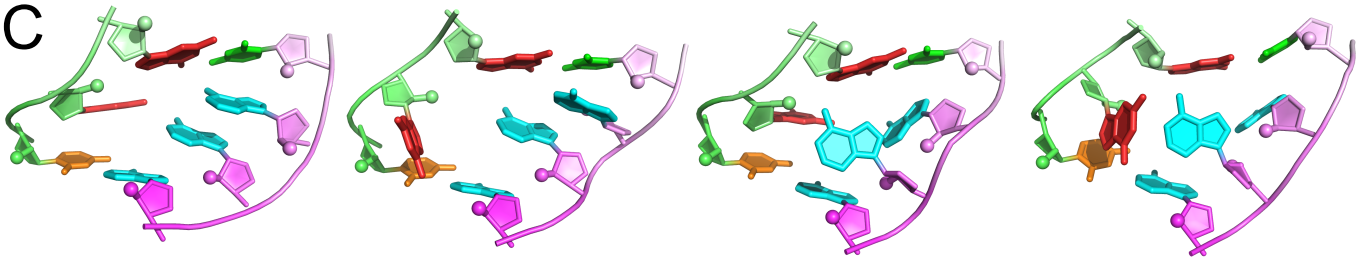
Motif name	NMR PDB ID	Number of models in NMR ensemble	Average pairwise rmsd (Å) of NMR ensemble	
			Include closing W.C. base pairs	Exclude closing W.C. base pairs
HIV-1 TAR apical loop	1ANR	20	4.07	4.42
Chimp HAR1 GAA loop	2LHP	10	2.61	3.33
Human HAR1 GAA loop	2LUB	10	1.98	2.58
Tandem UG:UA mismatch	2JSE	10	1.39	1.77

B



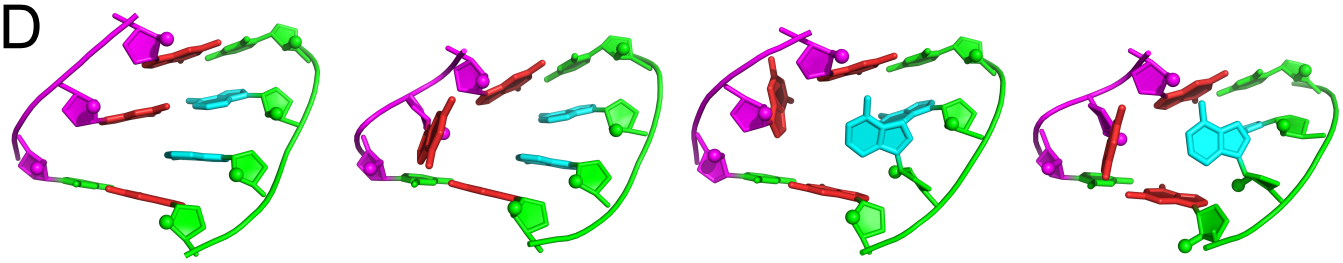
PDB: 1ANR | NMR model #1 PDB: 1ANR | NMR model #3 PDB: 1ANR | NMR model #4 PDB: 1ANR | NMR model #6

C



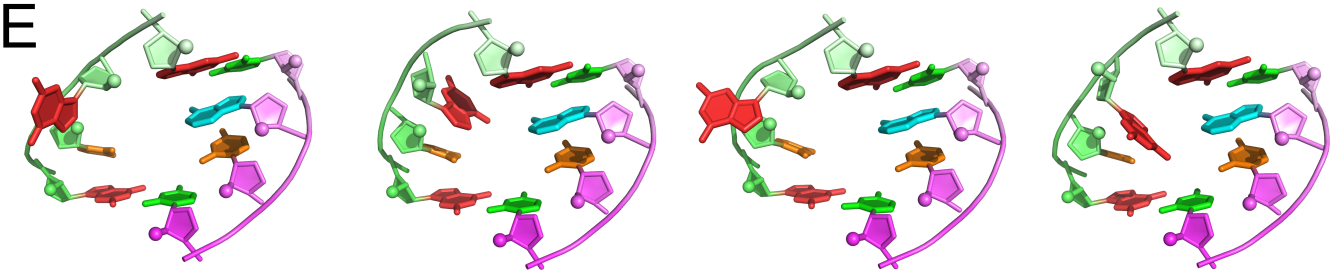
PDB: 2LHP | NMR model #1 PDB: 2LHP | NMR model #5 PDB: 2LHP | NMR model #8 PDB: 2LHP | NMR model #10

D



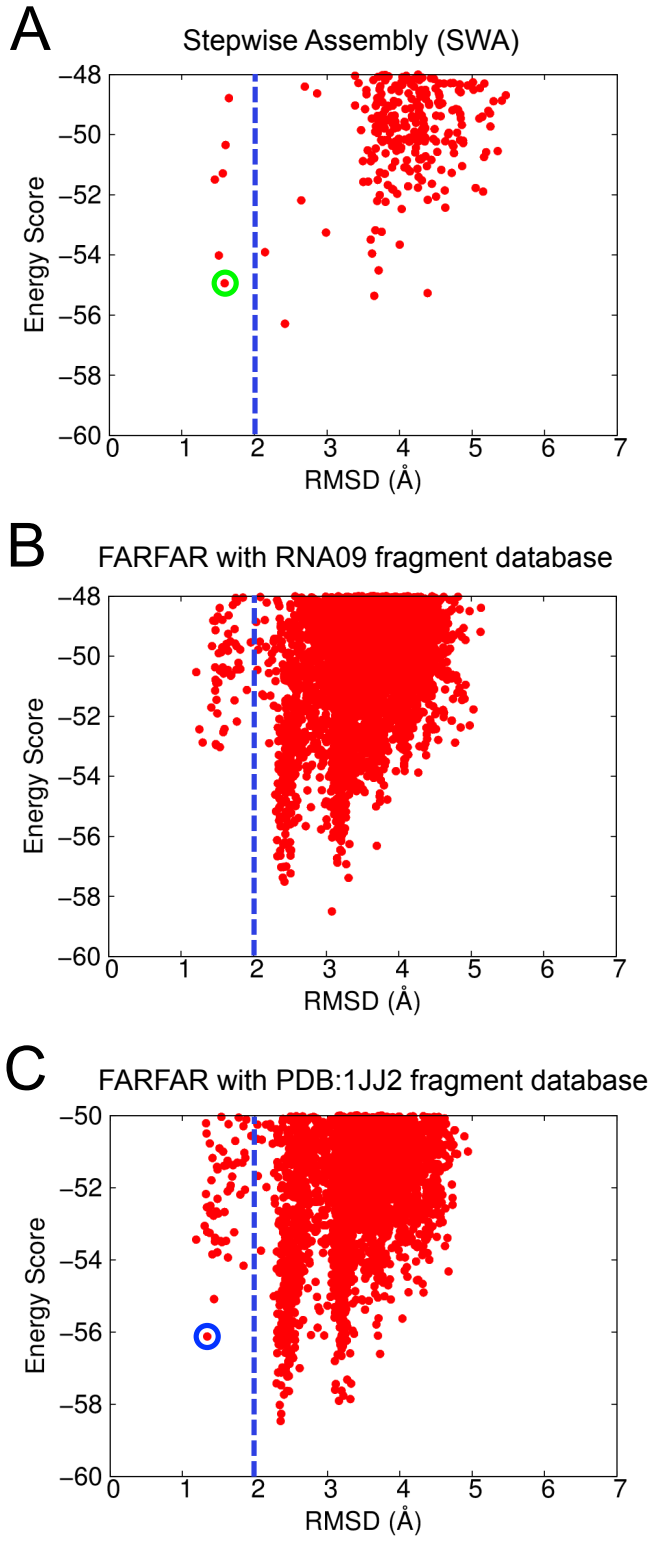
PDB: 2LUB | NMR model #1 PDB: 2LUB | NMR model #3 PDB: 2LUB | NMR model #6 PDB: 2LUB | NMR model #10

E

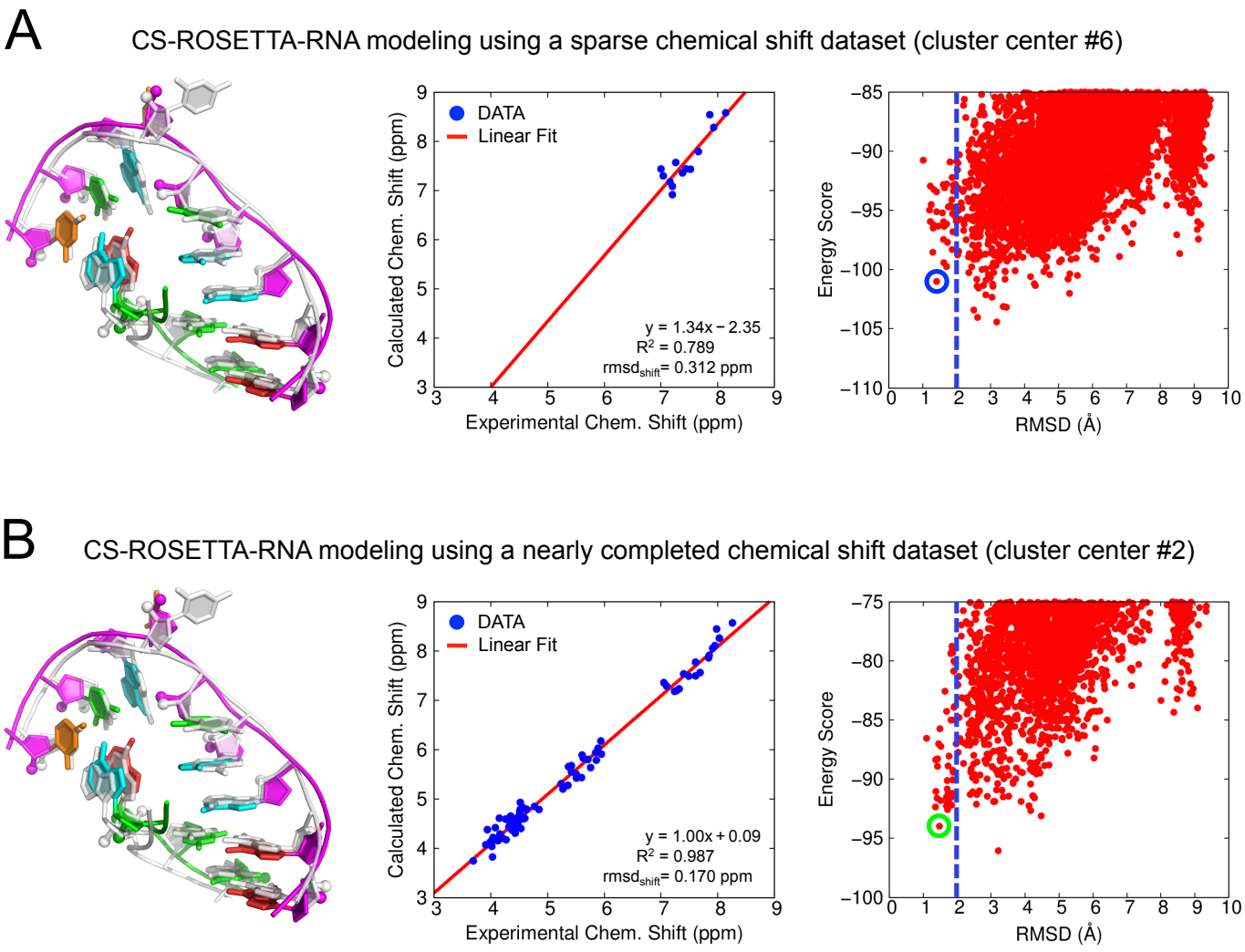


PDB: 2JSE | NMR model #1 PDB: 2JSE | NMR model #2 PDB: 2JSE | NMR model #3 PDB: 2JSE | NMR model #5

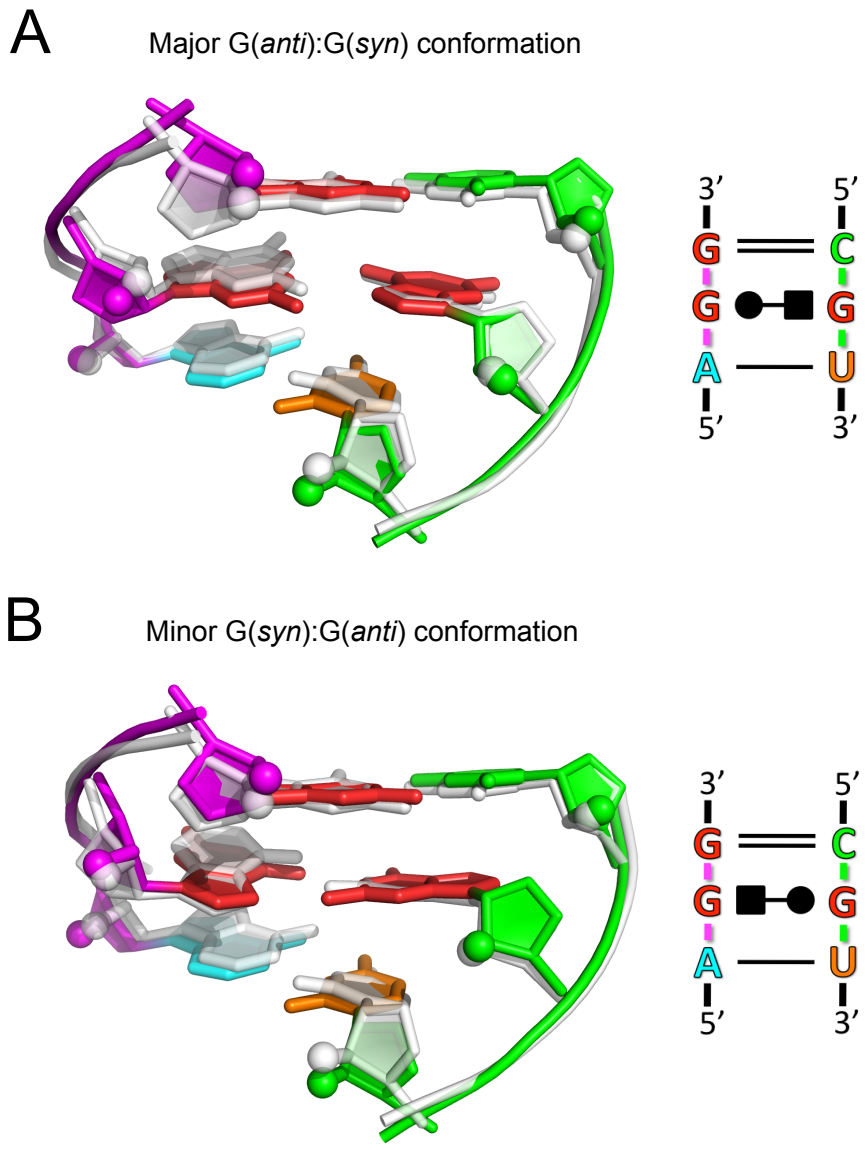
Supplementary Figure 5. Modeling results on the uuGCCaa anticodon stem-loop of *B. subtilis* tRNA^{Gly} by three different models generation methods. During the original blind modeling, we generated models using (A) SWA and (B) FARFAR with RNA09 fragment database (obtained from http://kinemage.biochem.duke.edu/databases/rnadb.php). CS-ROSETTA-RNA prediction using models from both source (A) and (B) did not return a <2.0 Å rmsd structure (to NMR PDB: 2LBJ) among the five lowest energy clusters. However, modeling with (A) SWA alone does return a <2.0 Å rmsd structure among the five lowest energy clusters (green circle). Lastly, modeling with (C) FARFAR but using the PDB:1JJ2 fragment database also returned a <2.0 Å rmsd structure among the five lowest energy clusters (blue circle).



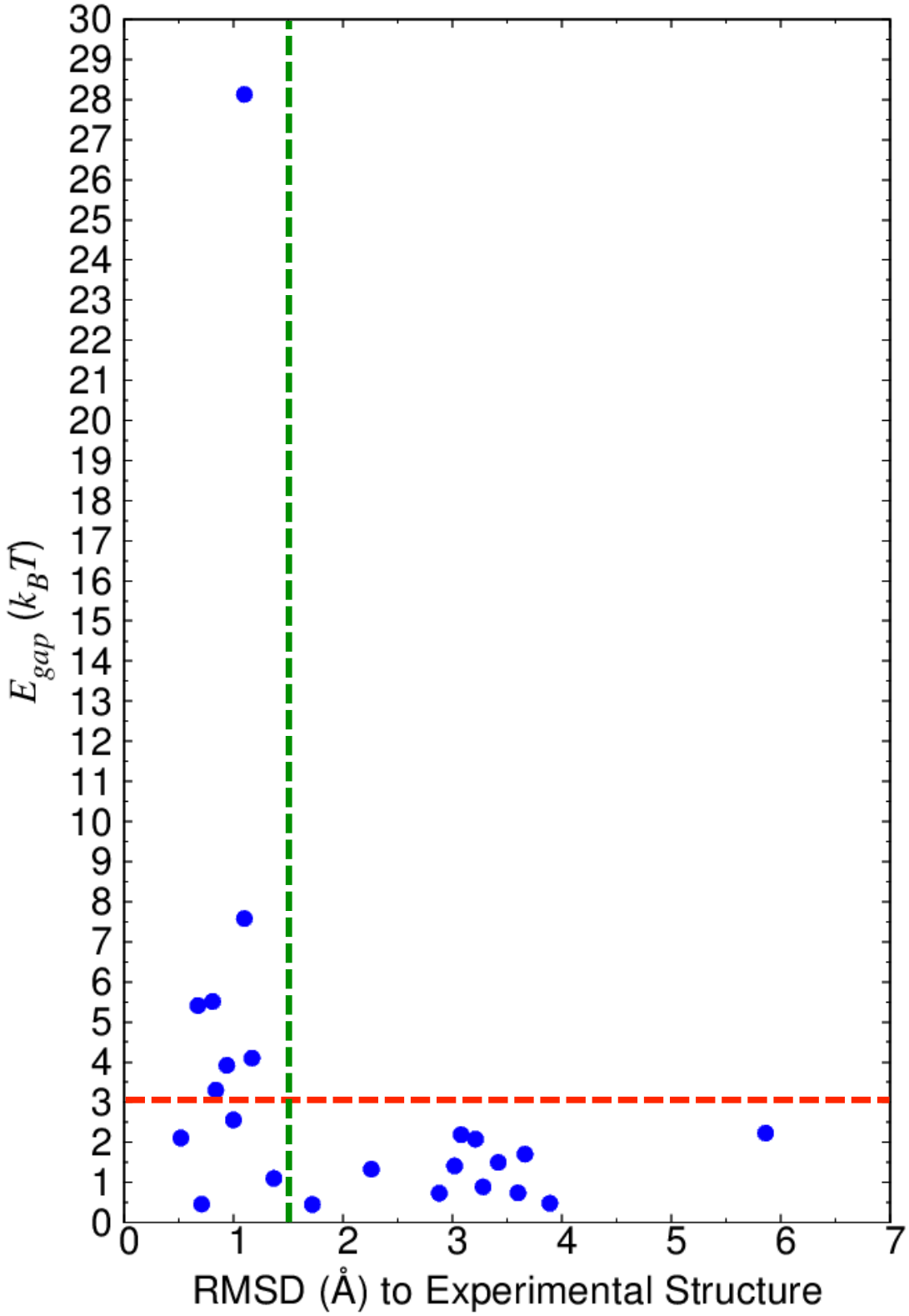
Supplementary Figure 6. Modeling a 13-nucleotide internal loop from hepatitis C virus IRES subdomain IIa using a sparse ^{1}H chemical shift dataset. CS-ROSETTA-RNA modeling results for a 13-nt internal loop from hepatitis C virus IRES subdomain IIa using (A) a sparse experimental chemical shift dataset and, for comparisons, (B) a nearly complete experimental chemical shift dataset. The sparse chemical shift dataset (A) contains only 13 non-exchangeable ^{1}H assignments for the 13-nt internal loop (an average of 1.0 assignments per nucleotide), with only aromatic H2, H6, and H8 assignments and no ribose proton assignments. In contrast, the nearly complete chemical shift dataset (B) contains 93 non-exchangeable ^{1}H assignments for the 13-nt internal loop (an average of 7.2 assignments per nucleotide). In both cases (A) and (B), incorporating non-exchangeable ^{1}H chemical shift data significantly improved the energy discrimination in favor of near-native Rosetta models with CS-ROSETTA-RNA cluster center #6 in case (A) and cluster center #2 in case (B) achieving atomic-accuracy to the experimental crystallographic structure (PDB: 2PN4; the pairwise all-heavy-atom rmsd between the CS-ROSETTA-RNA cluster centers and the experimental structure were 1.41 Å and 1.48 Å, respectively). The left panels display the near-native CS-ROSETTA-RNA cluster centers (shown in color) overlaid on the experimental crystallographic structure (shown in white). The middle panels display the correlation between back-calculated and experimental ^{1}H chemical shift for the CS-ROSETTA-RNA cluster centers. The right panels display plots of the Rosetta energy vs. rmsd to the experimental structure, with the near-native cluster centers highlighted with a blue circle in case (A) and a green circle in case (B). Further details of the modeling results are presented in the ***Modeling a 13-nucleotide internal loop from hepatitis C virus IRES subdomain IIa using a sparse ^{1}H chemical shift dataset*** subsection of Supplementary Results.



Supplementary Figure 7. CS-ROSETTA-RNA modeling of the major and minor conformations of a G:G mismatch. A NMR study on the single G:G mismatch [Biochemistry. 2000 Sep 26;39(38):11748-62] reveals the existence of (A) a major G(*anti*):G(*syn*) conformation (~75% populated; PDB: 1F5G) and (B) a minor G(*syn*):G (*anti*) conformation (~25% populated; PDB: 1F5H). CS-ROSETTA-RNA cluster center #1 (lowest energy) strongly agrees with the major G(*anti*):G(*syn*) conformation (all-heavy-atom rmsd of 0.71 Å). Furthermore, CS-ROSETTA-RNA cluster center #2 (next lowest energy) strongly agrees with minor G(*anti*):G(*syn*) conformation (rmsd of 0.63 Å). The CS-ROSETTA-RNA models (shown in color) are overlaid on the experimental NMR structures (shown in white). The two-dimensional schematics are annotated based on the experimental structure and follow the Leontis and Westhof nomenclature.

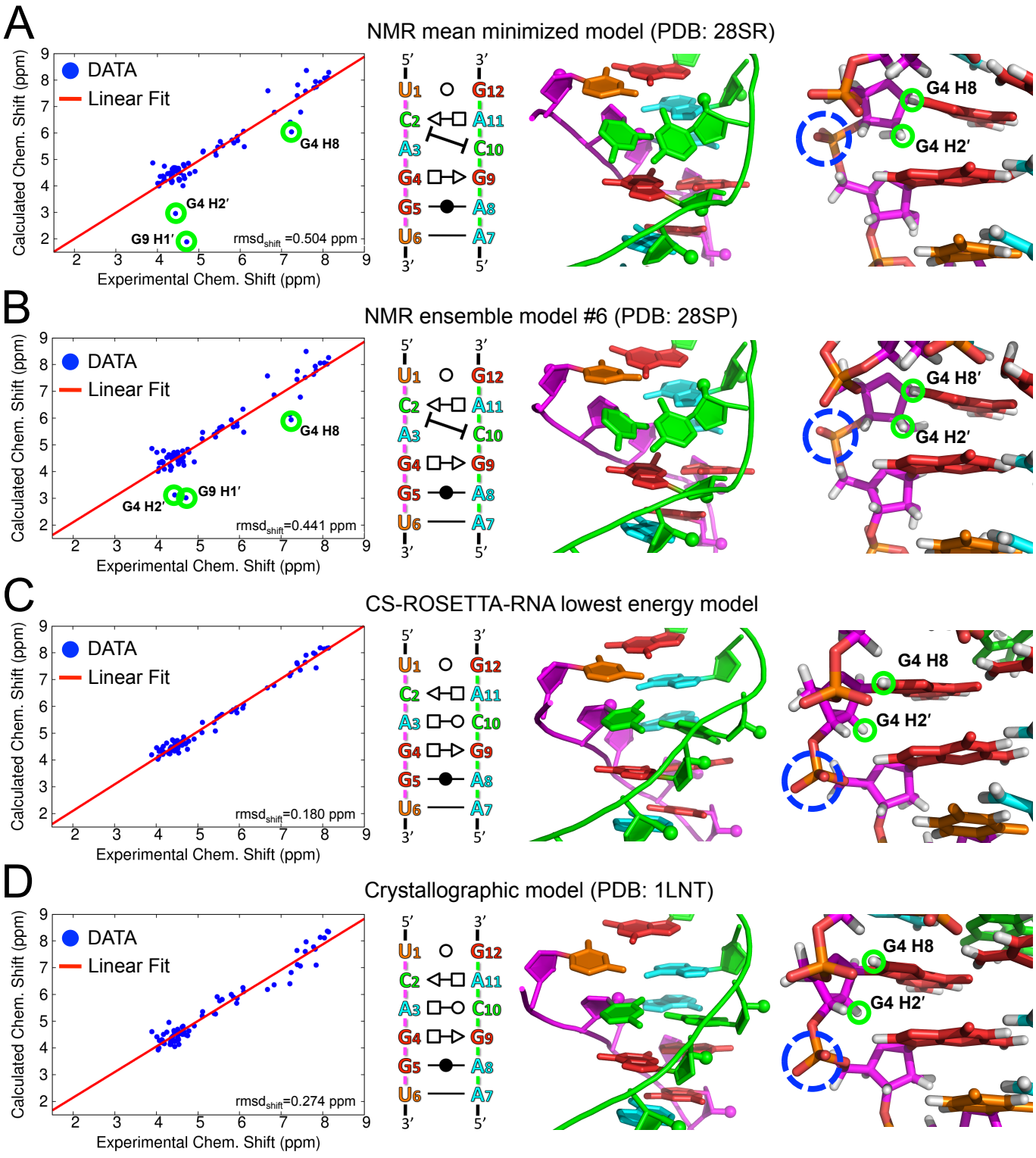


Supplementary Figure 8. E_{gap} vs. RMSD of the CS-ROSETTA-RNA lowest energy model to the experimental structure. An energy gap (E_{gap}) value greater than 3.0 k_BT (red line) is a strong indicator that the lowest energy Rosetta model will achieve atomic-accuracy. In the 23-RNA benchmark, 7 motifs have an E_{gap} value greater than 3.0 k_BT, and the lowest energy CS-ROSETTA-RNA model for all of these 7 cases were found to be within 1.5 Å rmsd (green line) of the experimental structure.

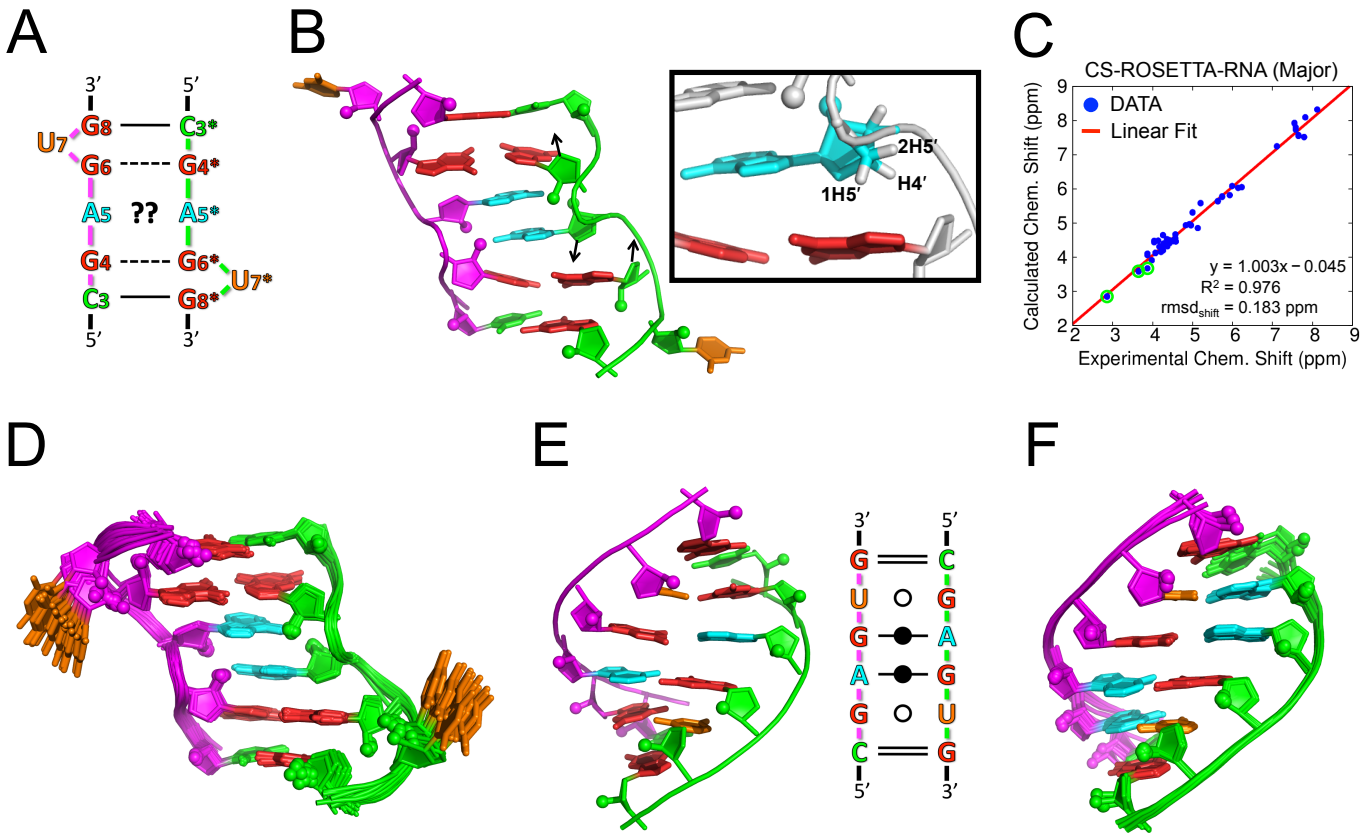


Nature Methods: doi:10.1038/nmeth.2876

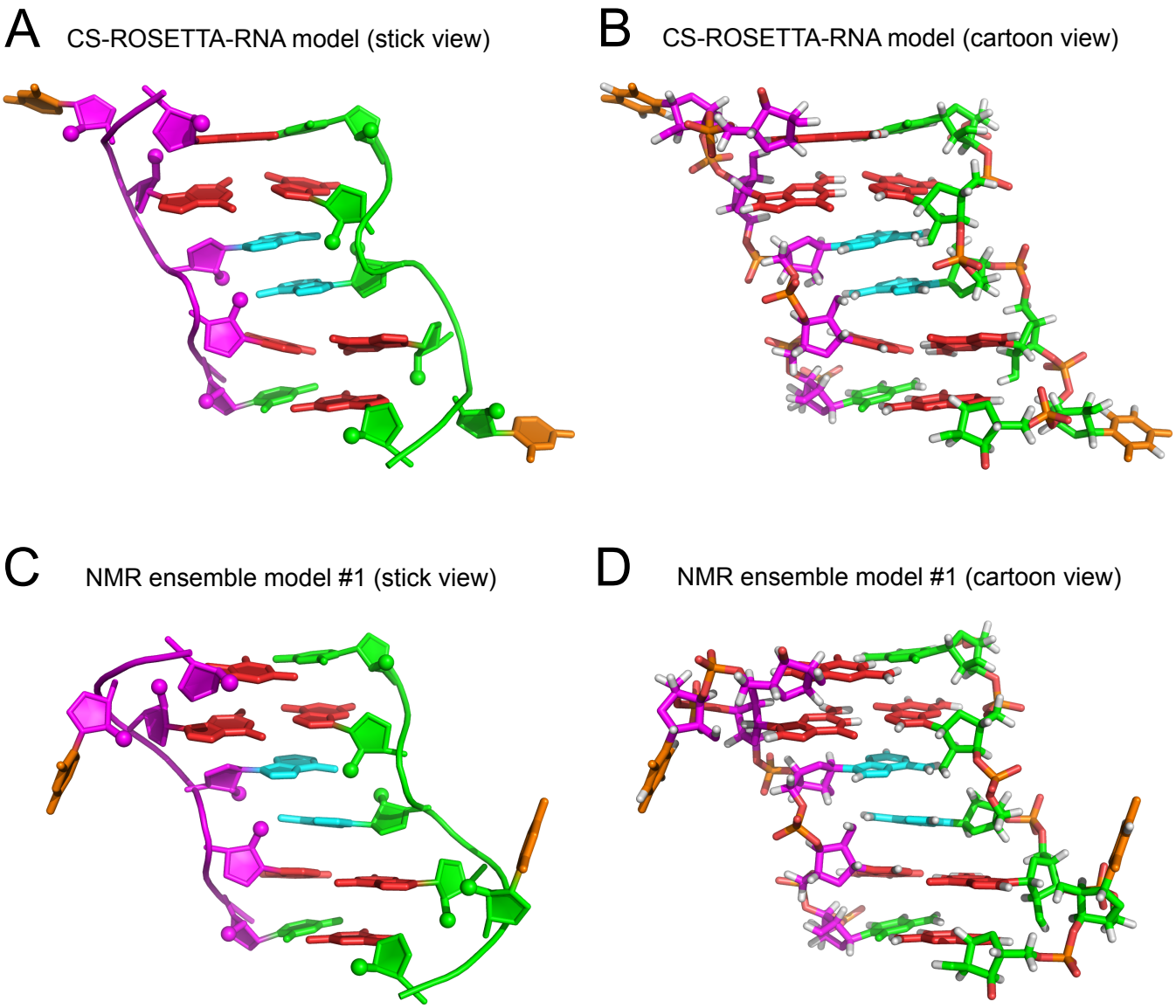
Supplementary Figure 9. Comparisons between CS-ROSETTA-RNA, NMR and crystallographic models of the most conserved internal loop from the signal recognition particle RNA. The NMR study on this motif produced a tightly defined ensemble, where (A) the mean minimized model and (B) the ensemble member with the best $rmsd_{shift}$ both gave back-calculated chemical shifts for the G4 H1′, G4 H8 and G9 H1′ protons that disagreed with the experimental data by at least 1.0 ppm (green circles). Since the experimental ^{1}H chemical shift data and the NMR models originated from the same NMR study, these disagreements could not have been due to differences in experimental conditions (e.g. temperature, salt concentrations and pH). The NMR ensemble also showed disagreements with both (C) the CS-ROSETTA-RNA lowest energy model and (D) the crystallographic model, particularly at the G4-G5 backbone suite (blue dashed circle), the G9-C10 backbone suite, and the C10 nucleobase.



Supplementary Figure 10. CS-ROSETTA-RNA modeling of the 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop. (A) A 2D schematic proposed in the original NMR study [Biochemistry. 2010 Jul 13;49(27):5817-27] and (B) CS-ROSETTA-RNA model of the major conformation of the 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop. In the CS-ROSETTA-RNA model, the backbone of A5 and A5* adopts a highly irregular conformation, which leads to a ~180° inversion in the adenosines' ribose orientation relative to the riboses in the preceding and following nucleotides (G4, G6, G4*, G6*; see direction of arrows at the O4* ribose atoms). This inverted ribose conformation places the H4′, 1H5′ and 2H5′ ribose protons of A5 and A5* directly above the G6 and G6* guanine rings and the resulting ring current effects lead to the large experimentally observed upfield shifts of these protons. (C) Back-calculated chemical shift values give excellent agreement with experimentally determined values ($rmsd_{shift}$ = 0.18 ppm). The upfield shifted H4′, 1H5′ and 2H5′ ribose protons of A5 and A5* are highlighted with green circles. (D) Subsequently determined experimental NMR ensemble of the 5′-GA^{Br}GU-3′/3′-U^{Br}GAG-5′ internal loop (PDB: 2LX1) agrees with the CS-ROSETTA-RNA model at atomic resolution (1.10 Å rmsd between the first member of the NMR ensemble and the CS-ROSETTA-RNA model). (E) CS-ROSETTA-RNA model of excited conformation of the 5′-GAGU-3′/3′-UGAG-5′ sequence adopts a helical conformation with two imino AG base pairs. (F) NMR ensemble of a close sequence variant (5′-AAGU-3′/3′-UGAA-5′; PDB: 2KXZ) supports this two imino AG base pairs helical conformation.



Supplementary Figure 11. Comparisons between the CS-ROSETTA-RNA and NMR models of the 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop. (A-B) The CS-ROSETTA-RNA lowest energy model for the major conformation of the 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop in cartoon and stick view. (C-D) The first model from the NMR ensemble subsequently solved by the Turner and Kennedy group (PDB: 2LX1) in cartoon and stick view. The pairwise all-heavy-atom rmsd value between the CS-ROSETTA-RNA and the NMR model is 1.10 Å.



Supplementary Table 1. Supplemental information on the RNA motifs benchmark

Motif name	Motif properties		Non-exchangable ^{1}H chemical shift data[a]			NMR structure[b]			Crystallographic structure[b]		
	Size	Strands	N_{total}	$N_{per\ nucleotide}$	Source	PDB	nucleotide-segment (chain ID)	rmsd $_{shift}$ (ppm)	PDB	nucleotide-segment (chain ID)	rmsd $_{shift}$ (ppm)
Known Structures											
Single G:G mismatch	6	2	39	6.5	BMRB Entry 4614	1F5G	3-5 (A), 6-8 (B)	0.19	–	–	–
UUCG tetraloop	6	1	46	7.7	BMRB Entry 5705	2KOC	5-10 (A)	0.24	1F7Y	8-13 (B)	0.28
Tandem GA:AG mismatch	8	2	60	7.5	Biochemistry. 1996 Jul 30;35(30):9677-89	1MIS	3-6 (A), 11-14 (B)	0.32	–	–	–
Tandem UG:UA mismatch	8	2	38	4.8	Biochemistry. 2007 Nov 6;46(44):12665-78	2JSE	4-7 (A), 16-19 (A)	0.26	–	–	–
16S rRNA UUAAGU loop	8	1	46	5.8	Biochemistry. 2001 Aug 21;40(33):9879-86	1HS2	3-10 (A)	0.33	**1FJG**	1089-1096 (A)	0.25
HIV-1 TAR apical loop	8	1	56	7.0	Nucleic Acids Res. 1996 Oct 15;24(20):3974-81	1ANR	29-36 (A)	0.28	–	–	–
$tRNA_i^{Met}$ ASL	9	1	67	7.4	J Mol Biol. 1997 Apr 4;267(3):505-19	1SZY	7-15 (A)	0.30	–	–	–
Conserved SRP internal loop	12	2	72	6.0	Nat Struct Biol. 1999 Jul;6(7):634-8	28SR	5-10 (A), 19-24 (A)	0.50	**1LNT**	4-9 (A), 16-21 (B)	0.27
R2 retrotransposon 4x4 loop	12	2	70	5.8	BMRB Entry 17406	2L8F	2-7 (A), 10-15 (A)	0.32	–	–	–
Hepatitis C virus IRES IIa	13	2	93	7.2	Nat Struct Biol. 2003 Dec;10(12):1033-8	1P5M	9-17 (A), 45-48 (A)	0.41	**2PN4**	51-59 (A), 109-112 (B)	0.20
GAAA tetraloop-receptor	15	3	62	4.1	BMRB Entry 6652	2ADT	19-24 (A), 48-51 (B), 78-82 (B)	0.54	**2R8S**	149-154 (R), 223-227 (R), 247-250 (R)	0.32
Sc.ai5γ 3-way junction	16	3	53	3.3	BMRB Entry 18503	2LU0	9-13 (A), 20-25 (A), 36-40 (A)	0.38	–	–	–
Blind Targets											
UAAC tetraloop	6	1	28	4.7	Fox group	4A4R	9-14 (A)	0.31	–	–	–
UCAC tetraloop	6	1	30	5.0	Fox group	4A4S	9-14 (A)	0.25	–	–	–
UGAC tetraloop	6	1	28	4.7	Fox group	4A4U	9-14 (A)	0.39	–	–	–
UUAC tetraloop	6	1	28	4.7	Fox group	4A4T	9-14 (A)	0.50	–	–	–
Chimp HAR1 GAA loop	7	2	38	5.4	Schwalbe group	2LHP	9-11 (A), 26-29 (A)	0.35	–	–	–
Human HAR1 GAA loop	7	2	34	4.9	Schwalbe group	2LUB	9-11 (A), 26-29 (A)	0.35	–	–	–
GU:UAU internal loop	9	2	50	5.6	Sigel group	–[c]	6-9 (A), 30-34 (A)	0.44		–	–
$tRNA^{Gly}$ ASL (cuUCCaa)	9	1	65	7.2	Nikonowicz group	2LBL	5-13 (A)	0.32	–	–	–
$tRNA^{Gly}$ ASL (cuUCCcg)	9	1	59	6.6	Nikonowicz group	2LBK	5-13 (A)	0.19	–	–	–
$tRNA^{Gly}$ ASL (uuGCCaa)	9	1	69	7.7	Nikonowicz group	2LBJ	5-13 (A)	0.37	–	–	–
5′-GAGU/3′-UGAG loop	12	2	90	7.5	Kennedy and Turner group	2LX1	3-8 (A), 14-19 (A)	0.26	–	–	–
Average	9.0	1.6	53.1	6.0	–	–	–	0.34	–	–	0.26

HIV, human immunodeficiency virus; TAR, trans-activation response; $tRNA_i^{Met}$, initiator methionine tRNA; ASL, anticodon stem-loop; SRP, signal recognition particle; IRES, internal ribosome entry site; HAR1, human accelerated region 1; $tRNA^{Gly}$, glycine tRNA.

[a] N_{total} and $N_{per\ nucleotide}$ are, respectively, the number of experimentally assigned non-exchangeable ^{1}H chemical shift data (including H1′, H2′, H3′, H4′, 1H5′ and 2H5′ ribose protons, and H2, H5, H6 and H8 base protons). The chemical shift data were obtained from the BMRB database, from published literature (see citation) and from NMR structure determination studies conducted in parallel with this work (see group name). In two known structure cases, the data were not directly available in the published literature and were graciously provided to us by the authors of the source publication (G. Varani for the HIV-1 TAR apical loop and J. D. Puglisi for the Hepatitis C Virus IRES IIa). In BMRB Entry 5705, correction were made for the switched chemical shift assignments of 1H5′ and 2H5′ for nucleotides C8, G9 and G10.

[b] The NMR structures came from the same source as the experimental non-exchangeable ^{1}H chemical shift data, except for the self-symmetric $GA^{Br}GU$ loop structure which came from a subsequent study by the same group. The first model of the NMR ensemble was used as the experimental reference structure (except in the 3 cases below). In 5 cases with known structure, we found that the motif had also been solved by crystallography. In 4 of the 5 cases, the crystallographic structure provided a better agreement to the chemical shift data (i.e., lower $rmsd_{shift}$ than every member of the NMR ensemble). In these 4 cases (PDB ID highlighted in bold text), the crystallographic structure was used as the experimental reference structure.

[c] The experimental structure has not yet been deposited into the PDB database.

Supplementary Table 2. Supplemental Rosetta modeling results (before inclusion of chemical shift data)

Motif name	Motif properties			Lowest energy model (top-1)						Best of five lowest energy cluster centers (top-5)				
	Size	Strands	PDB[a]	E_{gap}[b] (total)	E_{gap}[b] (chem. shift contribution)	rmsd (Å)	$rmsd_{shift}$ (ppm)	Base-pair recovery[c]	Base-stack recovery[c]	Cluster rank	rmsd (Å)	$rmsd_{shift}$ (ppm)	Base-pair recovery[c]	Base-stack recovery[c]
Known Structures														
Single G:G mismatch	6	2	1F5G	0.33	–	0.72	0.16	1/1	4/4	1	0.72	0.16	1/1	4/4
UUCG tetraloop	6	1	1F7Y	1.54	–	2.38	0.42	0/0	1/2	1	2.38	0.42	0/0	1/2
Tandem GA:AG mismatch	8	2	1MIS	2.35	–	3.84	0.36	0/2	4/6	2	0.99	0.22	2/2	6/6
Tandem UG:UA mismatch	8	2	2JSE	0.80	–	6.36	0.38	0/0	0/3	3	3.43	0.32	0/0	2/3
16S rRNA UUAAGU loop	8	1	1FJG	0.27	–	6.63	0.52	0/1	0/3	2	1.99	0.38	1/1	3/3
HIV-1 TAR apical loop	8	1	1ANR	0.00	–	4.86	0.30	0/0	0/0	1	4.86	0.30	0/0	0/0
$tRNA_i^{Met}$ ASL	9	1	1SZY	0.38	–	5.57	0.32	0/0	2/3	2	4.33	0.25	0/0	2/3
Conserved SRP internal loop	12	2	1LNT	1.80	–	1.89	0.40	1/3	9/10	2	0.91	0.26	3/3	10/10
R2 retrotransposon 4x4 loop	12	2	2L8F	0.06	–	4.68	0.48	0/4	3/10	5	1.39	0.44	2/4	10/10
Hepatitis C virus IRES IIa	13	2	2PN4	2.82	–	6.11	0.35	2/2	5/7	4	5.33	0.35	1/2	5/7
GAAA tetraloop-receptor	15	3	2R8S	1.04	–	3.96	0.49	1/5	6/11	3	3.07	0.61	2/5	5/11
Sc.ai5γ 3-way junction	16	3	2LU0	1.14	–	7.37	0.44	2/3	6/7	4	4.80	0.41	3/3	5/7
Blind Targets														
UAAC tetraloop	6	1	4A4R	0.33	–	0.85	0.33	0/1	3/3	1	0.85	0.33	0/1	3/3
UCAC tetraloop	6	1	4A4S	0.09	–	4.66	0.57	0/1	0/4	2	0.72	0.28	0/1	4/4
UGAC tetraloop	6	1	4A4U	2.11	–	5.94	0.49	0/0	1/2	2	1.60	0.38	0/0	2/2
UUAC tetraloop	6	1	4A4T	0.44	–	5.05	0.51	0/0	0/0	3	1.43	0.39	0/0	0/0
Chimp HAR1 GAA loop	7	2	2LHP	0.52	–	5.04	0.50	0/0	1/2	5	3.78	0.38	0/0	2/2
Human HAR1 GAA loop	7	2	2LUB	0.90	–	4.44	0.48	0/1	2/5	4	2.28	0.27	1/1	4/5
GU:UAU internal loop	9	2	–[d]	0.16	–	4.12	0.25	1/2	3/4	1	4.12	0.25	1/2	3/4
$tRNA^{Gly}$ ASL (cuUCCaa)	9	1	2LBL	0.41	–	5.56	0.29	0/0	2/5	3	5.47	0.33	0/0	2/5
$tRNA^{Gly}$ ASL (cuUCCcg)	9	1	2LBK	0.55	–	5.34	0.35	1/1	1/5	5	5.00	0.29	1/1	1/5
$tRNA^{Gly}$ ASL (uuGCCaa)	9	1	2LBJ	1.23	–	5.02	0.44	1/1	1/3	3	4.11	0.38	1/1	1/3
5′-GAGU/3′-UGAG loop	12	2	2LX1	0.55	–	5.48	0.53	0/2	2/7	3	1.19	0.29	2/2	7/7
Average	9.0	1.6	–	0.86	–	4.60	0.41	0.43/1.30	2.43/4.61	2.70	2.82	0.33	0.91/1.30	3.57/4.61
rmsd < 1.50 Å	–	–	–	–	–	2/23	–	–	–	–	8/23	–	–	–
rmsd < 2.00 Å	–	–	–	–	–	3/23	–	–	–	–	10/23	–	–	–
E_{gap} >3.00 k_BT	–	–	–	0/23	–	–	–	–	–	–	–	–	–	–

[a] PDB ID of reference experimental NMR or crystallographic structure (see Supplementary Table S1 for details).

[b] Energy gap, total (all energy terms) and only chemical shift term contributions (see Supplementary Results for definition). Bold text indicates E_{gap} value greater than 3.0 k_BT.

[c] Number of native base-pairs and native base-stacks correctly recovered by the Rosetta model. Base-pairs and base-stacks are automatically annotated using the program *MC-annotate* [J Mol Biol. 2001 May 18;308(5):919-36]. Base-pairing annotation follows the Leontis and Westhof nomenclature [RNA. 2001 Apr;7(4):499-512] and recovery entails having the correct edge-to-edge interaction (Watson-Crick, Hoogsteen, or Sugar-edge) and local strand orientation (*cis* or *trans*). Both the total native base-pairs counts and correctly recovered base-pairs counts are lowered owing to ambiguities in assignment of bifurcated base-pairs, pairs connected by single hydrogen bonds and pairs that are not completely co-planar. Boundary/closing canonical base-pairs were not counted. Base-stacks are classified as either upward, downward, outward or inward [RNA. 2009 Oct;15(10):1875-85] and recovery entails having the correct base-stacking type.

[d] The experimental structure has not yet been deposited into the PDB database.

Supplementary Table 3. Supplemental Rosetta modeling results (after inclusion of chemical shift data)

Motif name	Motif properties			Lowest energy model (top-1)						Best of five lowest energy cluster centers (top-5)				
	Size	Strands	PDB[a]	E_{gap}[b] (total)	E_{gap}[b] (chem. shift contribution)	rmsd (Å)	$rmsd_{shift}$ (ppm)	Base-pair recovery[c]	Base-stack recovery[c]	Cluster rank	rmsd (Å)	$rmsd_{shift}$ (ppm)	Base-pair recovery[c]	Base-stack recovery[c]
Known Structures														
Single G:G mismatch	6	2	1F5G	0.46	0.32	0.71	0.14	1/1	4/4	1	0.71	0.14	1/1	4/4
UUCG tetraloop	6	1	1F7Y	**3.30**	1.55	0.84	0.18	0/0	2/2	1	0.84	0.18	0/0	2/2
Tandem GA:AG mismatch	8	2	1MIS	**7.58**	7.08	1.10	0.13	2/2	6/6	1	1.10	0.13	2/2	6/6
Tandem UG:UA mismatch	8	2	2JSE	1.41	2.25	3.02	0.12	0/0	2/3	5	2.52	0.15	0/0	3/3
16S rRNA UUAAGU loop	8	1	1FJG	2.11	4.88	0.52	0.19	1/1	3/3	1	0.52	0.19	1/1	3/3
HIV-1 TAR apical loop	8	1	1ANR	2.23	1.05	5.86	0.18	0/0	0/0	1	5.86	0.18	0/0	0/0
$tRNA_i^{Met}$ ASL	9	1	1SZY	0.48	1.18	3.89	0.18	0/0	2/3	3	1.35	0.17	0/0	3/3
Conserved SRP internal loop	12	2	1LNT	**5.51**	2.70	0.81	0.18	3/3	10/10	1	0.81	0.18	3/3	10/10
R2 retrotransposon 4x4 loop	12	2	2L8F	**4.10**	0.46	1.17	0.17	3/4	10/10	1	1.17	0.17	3/4	10/10
Hepatitis C virus IRES IIa	13	2	2PN4	2.08	-3.14	3.21	0.19	2/2	7/7	2	1.48	0.17	2/2	7/7
GAAA tetraloop-receptor	15	3	2R8S	**5.41**	4.40	0.68	0.26	4/5	10/11	1	0.68	0.26	4/5	10/11
Sc.ai5γ 3-way junction	16	3	2LU0	1.71	-1.26	3.66	0.22	3/3	7/7	4	1.74	0.23	3/3	7/7
Blind Targets														
UAAC tetraloop	6	1	4A4R	**3.92**	0.66	0.94	0.24	1/1	3/3	1	0.94	0.24	1/1	3/3
UCAC tetraloop	6	1	4A4S	2.56	1.43	1.00	0.20	1/1	4/4	1	1.00	0.20	1/1	4/4
UGAC tetraloop	6	1	4A4U	0.74	0.32	3.60	0.28	0/0	1/2	2	1.67	0.28	0/0	2/2
UUAC tetraloop	6	1	4A4T	0.45	0.36	1.72	0.27	0/0	0/0	1	1.72	0.27	0/0	0/0
Chimp HAR1 GAA loop	7	2	2LHP	0.73	0.20	2.88	0.18	0/0	2/2	3	2.88	0.21	0/0	1/2
Human HAR1 GAA loop	7	2	2LUB	1.33	0.65	2.26	0.22	1/1	4/5	4	2.03	0.17	0/1	4/5
GU:UAU internal loop	9	2	–[d]	1.10	3.16	1.37	0.16	2/2	4/4	1	1.37	0.16	2/2	4/4
$tRNA^{Gly}$ ASL (cuUCCaa)	9	1	2LBL	0.89	4.02	3.28	0.19	0/0	5/5	3	1.41	0.20	0/0	5/5
$tRNA^{Gly}$ ASL (cuUCCcg)	9	1	2LBK	1.50	-0.23	3.42	0.17	1/1	4/5	2	1.94	0.16	1/1	5/5
$tRNA^{Gly}$ ASL (uuGCCaa)	9	1	2LBJ	2.19	1.51	3.08	0.16	1/1	3/3	3	2.93	0.17	1/1	3/3
5′-GAGU/3′-UGAG loop	12	2	2LX1	**28.13**	19.65	1.10	0.18	2/2	7/7	1	1.10	0.18	2/2	7/7
Average	9.0	1.6	–	3.47	2.31	2.18	0.19	1.22/1.30	4.35/4.61	1.91	1.64	0.19	1.17/1.30	4.48/4.61
rmsd < 1.50 Å	–	–	–	–	–	11/23	–	–	–	–	14/23	–	–	–
rmsd < 2.00 Å	–	–	–	–	–	12/23	–	–	–	–	18/23	–	–	–
E_{gap} >3.00 k_BT	–	–	–	7/23	–	–	–	–	–	–	–	–	–	–

[a] PDB ID of reference experimental NMR or crystallographic structure (see Supplementary Table S1 for details).

[b] Energy gap, total (all energy terms) and only chemical shift term contributions (see Supplementary Results for definition). Bold text indicates E_{gap} value greater than 3.0 k_BT.

[c] Number of native base-pairs and native base-stacks correctly recovered by the Rosetta model. Base-pairs and base-stacks are automatically annotated using the program *MC-annotate* [J Mol Biol. 2001 May 18;308(5):919-36]. Base-pairing annotation follows the Leontis and Westhof nomenclature [RNA. 2001 Apr;7(4):499-512] and recovery entails having the correct edge-to-edge interaction (Watson-Crick, Hoogsteen, or Sugar-edge) and local strand orientation (*cis* or *trans*). Both the total native base-pairs counts and correctly recovered base-pairs counts are lowered owing to ambiguities in assignment of bifurcated base-pairs, pairs connected by single hydrogen bonds and pairs that are not completely co-planar. Boundary/closing canonical base-pairs were not counted. Base-stacks are classified as either upward, downward, outward or inward [RNA. 2009 Oct;15(10):1875-85] and recovery entails having the correct base-stacking type.

[d] The experimental structure has not yet been deposited into the PDB database.

Supplementary Results

Dependencies of non-exchangeable ^{1}H chemical shift values on RNA 3D conformation

The chemical shift value of each atom in a RNA molecule is the sum two components: a conformation-independent component (due to the covalent bonding configuration of the molecule), and a conformation-dependent component (due to the three-dimensional structure of the molecule). It is the conformation-dependent component of the chemical shift that provides 3D structural information about each atom's local environment, and causes the variations in the chemical shifts between two protons of the same type (e.g. two H8 protons on different guanosine nucleotide). For non-exchangeable protons in RNA, the conformation-dependent component of the chemical shift can be adequately explained by the following three terms[1-4]:

1. Ring current effect: All nucleobases in RNA are aromatic and a ring current is induced in the delocalized π-electrons due to the externally applied magnetic field NMR experiment. A proton located above or below an aromatic nucleobase will be in a strong shielding zone leading to a decrease in the proton's chemical shift value (i.e. shifted upfield)[2]. Conversely, a proton located adjacent to and in the plane of the nucleotide will experience a de-shielding effect leading to an increase in the proton's chemical shift value (i.e. shifted downfield)[2].

2. Local (atomic) magnetic anisotropy contribution: The ring current effect can be viewed as the magnetic anisotropy contribution due to the aromatic property of the entire nucleobase. However, there is also a second contribution from the 'local' magnetic anisotropy of the individual atoms. That is, the chemical shift of each proton is affected by the local magnetic dipoles of neighboring atoms. Studies have found the magnetic anisotropy contributions from RNA backbone atoms (ribose and phosphate) to be small[2,4]. The nucleobase atoms are therefore the main local magnetic anisotropy contributors and their overall effect on the chemical shift value has been shown to have the same sign as the ring current effects in almost all regions of space[2].

3. Electric field effect: Electric fields can cause polarization in the electron density along the chemical bonds leading to a change in the atom's chemical shift value[2]. However, the electric field effect on the chemical shift of non-exchangeable protons has been found to be negligible[3,4]. The electric field effect, however, does have a more important role on the chemical shift of exchangeable protons (e.g. imino, amino, and hydroxyl protons), especially those involved in hydrogen bonds[2].

Note that the conformation-dependent component of the chemical shift is determined in the same manner for all non-exchangeable proton types. This means that if a base proton (e.g. H8) and a ribose proton (e.g. H1′) were put in the same environment (i.e. in the exact same position relative to all the neighboring atoms), the resulting conformation-dependent component of the chemical shift will be almost exactly the same for the two protons[2].

Importance of base and ribose ^{1}H chemical shift data for recovering the native RNA structure

The chemical shifts of both base and ribose protons are dependent on the specific 3D conformation of the RNA molecule (please see the above ***Dependencies of non-exchangeable ^{1}H chemical shift values on RNA 3D conformation*** subsection for details). Hence, in principle, the chemical shifts of both base and ribose protons can provide important structural information for discriminating and recovering the native (correct) RNA structure. Early on in our investigation, we hypothesized that the chemical shifts of base protons might play a more important role given that RNA nucleobases tend to pair and stack with each other. Different base-pairing and base-stacking patterns should give rise to differences in the chemical shift values of the base protons (due to the base protons being shielded or de-shielded by a different amount in each cases). Furthermore, if a nucleotide is an extra-helical bulge (i.e. flipped-out of the helix), then its base protons can provide a downfield chemical shift signature (due to the lack of shielding effect from neighboring nucleobases). In contrast, we hypothesized that chemical shifts of ribose protons (especially H3′, H4′, 1H5′ and 2H5′) might play a less important role given that these protons tend to be positioned near the exterior of the RNA structure, further away from the neighboring nucleobases.

However, our experience from modeling the various RNA motifs in this work indicates that in certain cases, the ribose protons in fact provide the most important chemical shifts for discriminating and recovering the native (correct) structure. For example, in the major conformation of the 5′-GAGU-3′/3′-UGAG-5′ internal loop, the backbone of A5 adopts a highly irregular conformation which positions the H4′, 1H5′ and 2H5′ protons right above the G6 nucleobase (see Supplementary Fig. 10B). The large upfield chemical shifts of these protons provide a unique chemical shift signature that strongly discriminates in favor of the native conformation (see Supplementary Fig. 10C). Another example is the UUAAGU hexaloop from 16S ribosomal RNA. In the native conformation of this hexaloop, the H4′ atom of the U7 nucleotide is positioned directly below the A5 nucleobase (see main text Figs. 1A-B). This geometry leads to a substantial upfield chemical shift of the proton at 3.39 ppm, more than six standard deviations upfield of the average chemical shift value of all RNA uracil H4′ atoms deposited in the Biological Magnetic Resonance Data Bank[5] (4.36 ± 0.16 ppm; see main text Fig. 1C). Lastly, we note that aside from recovering the native structure, chemical shifts can also be use as evidence to disfavor certain non-native structures (see Supplementary Fig. 9).

Thus all non-exchangeable protons (both base and ribose protons) can provide important structural information for discriminating and recovering the native (correct) RNA structure.

Poor accuracy cases

Despite generally giving high-resolution models, CS-ROSETTA-RNA returned 5 of 23 cases with poorer than 2.0 Å rmsd accuracy. In one case (uuGCCaa anticodon stem-loop of *B. subtilis* $tRNA^{Gly}$), either FARFAR modeling using a different fragment database or SWA modeling alone returned < 2.0 Å rmsd structures among the five lowest energy clusters (Supplementary Fig. 5), suggesting that this motif would be recoverable with different sampling schemes. For the other four failure cases, examination of the original experimental NMR models revealed potential complications for structure modeling. These cases

include a poorly structured apical loop of the HIV-1 TAR RNA (PDB: 1ANR), two HAR1 internal loops with multiple nucleotides shown to be flipping in and out of the helix in the NMR ensemble (PDB: 2LHP and 2LUB), and a tandem UG:UA mismatch that contains no hydrogen bonds (PDB: 2JSE). Each of these cases presented large deviations in the coordinates of the experimental NMR ensemble (>2.5 Å average pairwise rmsd for HIV-1 TAR apical loop and the two HAR1 internal loops) or more localized dynamics (tandem UG:UA mismatch). The NMR ensembles' average pairwise rmsd values are presented in Supplementary Figure 4A. The diverse structural conformations adopted by each of these four RNA motifs in the NMR ensembles are presented in Supplementary Figures 4B-E. Such structural dynamics in solution precluded high-resolution agreement between these structures and the CS-ROSETTA-RNA models since the Rosetta modeling approach optimizes for energetically favorable well-structured conformations. Interestingly, a separate benchmark case involving a heterogeneous ensemble was successfully recovered by CS-ROSETTA-RNA: models for two populated conformations of the G:G mismatch were attained and agreed with the experimental NMR ensemble[6] (Supplementary Fig. 7). Thus, overall, our benchmark results demonstrate the general applicability of CS-ROSETTA-RNA for high resolution RNA structure determination but highlight limitations in cases where the RNA conformation is dynamic or unstructured.

Modeling a 13-nucleotide internal loop from hepatitis C virus IRES subdomain IIa using a sparse ^{1}H chemical shift dataset

As a test of how much chemical shift data are needed to improve modeling accuracy, we modeled a 13-nt internal loop from the hepatitis C virus IRES subdomain IIa using a sparse chemical shift dataset that is missing a large majority of its proton assignments.

In the main text, we have presented the CS-ROSETTA-RNA modeling results on this 13-nt internal loop using an nearly complete experimental chemical shift dataset obtain from a NMR study[7] of the hepatitis C virus IRES subdomain IIa RNA construct. This nearly complete chemical shift dataset contains 93 non-exchangeable ^{1}H assignments for the 13-nt internal loop (an average of 7.2 assignments per nucleotide). However, the same NMR study[7] also characterized the 13-nt internal loop in the context of a larger RNA construct encompassing the entire hepatitis C virus IRES domain II. The NMR study[7] indicated that this 13-nt internal loop adopted the same conformation when investigated as part of the subdomain IIa or the entire domain II. However, due to the large size of the entire domain II, severe spectral overlap from ^{1}H resonances prevented unambiguous assignments for a large majority of the proton chemical shifts. This resulted in a sparse chemical shift dataset containing only 13 non-exchangeable ^{1}H assignments for the 13-nt internal loop (an average of 1.0 assignments per nucleotide). Furthermore, the sparse chemical shift dataset contained only aromatic H2, H6, and H8 assignments and no ribose proton assignments.

This sparse chemical shift dataset provided us with a realistic case study to investigate the sensitivity and robustness of the CS-ROSETTA-RNA method toward missing chemical shift assignments. We found that incorporating even this sparse chemical shift dataset substantially improved the energetic discrimination in favor of near-native Rosetta models with CS-ROSETTA-RNA cluster center #6

achieving atomic-accuracy to the experimental conformation (PDB: 2PN4; 1.41 Å pairwise all-heavy-atom rmsd; see Supplementary Fig. 6A). In contrast, standard Rosetta *de novo* modeling without chemical shift data produced no model (among the 20 lowest energy cluster centers) that is within 5.0 Å rmsd of the experimental conformation. We note however, that despite the substantial improvement in the energetic discrimination, the near-native CS-ROSETTA-RNA model was ranked as cluster center #6 and fell just outside the five lowest energy cluster centers (see modeling success criteria presented in the main text). Not surprisingly, using the nearly complete chemical shift dataset led to even better energetic discrimination with CS-ROSETTA-RNA cluster center #2 achieving atomic-accuracy to the experimental conformation (1.48 Å pairwise all-heavy-atom rmsd; see main text Table 1 and Supplementary Fig. 6B).

These results suggested to us that while incorporating sparse chemical shift dataset (~1 non-exchangeable ^{1}H assignments per nucleotide) can substantially improve the energetic discrimination, the resulting near-native models might not be reliably ranked as one of the five lowest energy cluster centers. However, we also note that it is not necessary to use a nearly complete chemical shift dataset (7-8 non-exchangeable ^{1}H assignments per nucleotide) in order to achieve reliable modeling as many of the successful CS-ROSETTA-RNA modeling cases in the 23-RNA benchmark used chemical shift datasets with fewer than 6.0 non-exchangeable ^{1}H assignments per nucleotide (see Supplementary Table 1). Furthermore, the Sc.ai5γ three-way junction case was successfully modeled using a chemical shift dataset with just 3.3 non-exchangeable ^{1}H assignments per nucleotide (see Supplementary Table 1).

A criterion for confidence prediction

We discovered that the energy gap, between the lowest energy model and the next lowest energy model that is structurally distinct from the former (at least 1.5 Å rmsd separation), could be used as a metric to assess the confidence of the CS-ROSETTA-RNA modeling. A large energy gap indicates the presence of a single dominant lowest energy conformation with much better energy than all others. Conversely, a small energy gap indicates the presence of multiple structurally distinct conformations with similar energies. As a metric to assess the confidence level of future predictions, we defined the E_{gap}:

$$E_{gap} = \left| E_{lowest} - E_{next} \right|$$

where E_{lowest} is the Rosetta energy (including chemical shift pseudo-energy) of the lowest energy model and E_{next} is the Rosetta energy (including chemical shift pseudo-energy) of the next lowest energy model that is at least 1.5 Å all-heavy-atom rmsd from the former. Based on the results from the 23-RNA benchmark, an E_{gap} value greater than 3.0 k_BT is a strong indicator that the lowest energy Rosetta model is accurate. Seven cases from the 23-RNA benchmark were found to have an E_{gap} value greater than 3.0 k_BT: (1) UUCG tetraloop, (2) Tandem GA:AG mismatch, (3) Conserved SRP internal loop, (4) R2 retrotransposon 4x4 loop, (5) GAAA tetraloop-receptor, (6) UAAC tetraloop, and (7) 5′-GAGU/3′-UGAG loop (see Supplementary Table 3 for details). For all 7 cases, the lowest energy model was within atomic-accuracy of the experimental structure (under 1.5 Å all-heavy-atom rmsd; see Supplementary Fig. 8 and Supplementary Table 3). The criterion was applicable to motifs in both the known structure test set and the blind target set. Interestingly, the same criterion was not applicable to the benchmark

results prior to the addition of the chemical shift pseudo-energy term (see Supplementary Table 2). Thus, inclusion of the experimental chemical shift data was critical for both achieving accurate predictions and assessing their confidence.

Contributions of chemical shift pseudo-energy and other energy terms to confidence prediction

To further evaluate the importance of chemical shift data for confident CS-ROSETTA-RNA modeling, we investigated the relative contributions of the chemical shift pseudo-energy and other Rosetta energy terms to the energy gap. We focus on the 7 high confidence benchmark cases with E_{gap} value greater than 3.0 k_BT (for a complete list of the 7 cases, see ***A criterion for confidence prediction*** subsection). The average E_{gap} value among these 7 cases, totaled over all energy terms, was 8.28 k_BT. We found that the chemical shift pseudo-energy term alone contributes a sizable fraction of the overall energy gap [please see the E_{gap} (chem. shift contribution) column of Supplementary Table 3 for details]. On average among the 7 cases, the chemical shift pseudo-energy term contribution to the energy gap is 3.06 k_BT or 37% of the overall energy gap. For comparison, the Rosetta hydrogen-bonding energy terms contribution is 1.66 k_BT or 20% of the overall energy gap. The π-π stacking energy term contribution is 0.87 k_BT or 10% of the overall energy gap.

Blind modeling of a highly irregular internal loop and its excited conformation

Perhaps the most striking CS-ROSETTA-RNA result was the blind prediction of a self-complementary 4x4 internal loop with the sequence 5′-GA^{Br}GU-3′/3′-U^{Br}GAG-5′ (^{Br}G, denoting 8-bromoguanosine, was introduced to stabilize the major of two conformations observed in the RNA). The original NMR study[8] of this motif provided a descriptive 2D schematic (reproduced in Supplementary Fig. 10A) but did not produce sufficient restraints to generate an atomic-detail three-dimensional structure due to ambiguity in NOE assignments. In contrast, CS-ROSETTA-RNA was able to produce a 3D model (Supplementary Fig. 10B) with a large energy gap (28.1 k_BT), giving strong confidence in the model's accuracy. Further supporting its accuracy, the CS-ROSETTA-RNA model recovered information presented in the NMR study but set aside during our structure modeling, including two G-G *trans* Watson-Crick/Hoogsteen base pairs, the two-fold rotational symmetry of the structure, and C2′-endo ribose conformations at six nucleotides (see ***CS-ROSETTA-RNA modeling and refinement of the 4x4 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop*** subsection of Supplementary Note). Finally, chemical shift data strongly supported an unusual feature of the CS-ROSETTA-RNA model, previously unseen in the database of RNA structures [as assessed by FR3D[9]; see next subsection]: the nucleobases of the two central adenosines were stacked on each other, formed no base pairs, and formed hydrogen bonds to the 2′-oxygens in the opposing strand riboses (via both the W.C. and Hoogsteen edges). To accommodate these interactions, the adenosines adopted an irregular backbone conformation, involving ~180° inversions in their ribose orientation relative to the riboses in the preceding and following nucleotides. The resulting geometry positioned the adenosines' H4′, 1H5′, and 2H5′ atoms directly above the neighboring guanosine nucleobases, leading to strongly predicted upfield shifts that agreed well with the experimentally measured chemical shifts ($rmsd_{shift}$ = 0.18 ppm; Supplementary Fig. 10C). Subsequently,

an additional series of NMR experiments with non-self-complementary duplexes allowed for the assignment of previously ambiguous NOEs[10]. The resulting conventionally determined NMR structure of the 5′-GA^{Br}GU-3′/3′-U^{Br}GAG-5′ internal loop confirmed the accuracy of the CS-ROSETTA-RNA model at atomic resolution (PDB: 2LX1; rmsd of 1.10 Å; see Supplementary Fig. 11).

The 5′-GAGU-3′/3′-UGAG-5′ sequence motif also gave evidence for an 'excited' conformation in the form of lower intensity spectral peaks (~30% population)[8], and partial chemical shifts for this state could be measured. CS-ROSETTA-RNA modeling guided by the excited state data gave a well-converged model (energy gap of 3.5 k_BT) that adopts a helical conformation with two imino (cWW) AG base pairs (Supplementary Fig. 10E). This model indeed agreed at high accuracy (1.11 Å) with the conventionally determined NMR structures of several close sequence variants of the motif[8] (Supplementary Fig. 10F).

FR3D search on the 4x4 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop

In the CS-ROSETTA-RNA model of the major conformation of the 4x4 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop, the backbone of A5 and A5* adopted a highly irregular conformation, which led to a ~180° inversion in the adenosines' ribose orientation relative to the riboses in the preceding and following nucleotides (G4, G6, G4*, G6*; see Supplementary Fig. 10B). We used the FR3D program[9] to determine whether this conformational arrangement had been previously observed. The FR3D search, which relies on geometric comparisons of the base conformations, was carried out on a non-redundant list of 654 RNA-containing PDB structures with resolution better than 4 Å (downloaded from the FR3D website on June 02, 2012). The atomic coordinates of the G4-A5-G6 tri-nucleotide from the CS-ROSETTA-RNA model were used as the query conformation and FR3D was used to search for any geometrically similar consecutive tri-nucleotide conformation in the non-redundant PDB database. FR3D found no matching candidate with geometric discrepancy below the 0.50 Å default cutoff value. The candidate with the lowest geometric discrepancy (0.62 Å), the G1297-U1298-A1299 tri-nucleotide of the 16S rRNA of *E. coli* (PDB: 2AW7), does not display the aforementioned ~180° inverted ribose conformational arrangement.

Lastly, a manual search reveals a potentially similar 180° inverted ribose conformation at the G1426-A1427-G1428 tri-nucleotide of the 23S rRNA of *E. coli* (PDB: 1VS6). However, all bases in this prior structure were in the *anti* glycosidic configuration, distinct from the *syn* G configurations exhibited in the 5′-GAGU-3′/3′-UGAG-5′ CS-ROSETTA-RNA model.

Current and future uses of ^{13}C and ^{15}N chemical shifts for RNA structural modeling

Relationships between ^{13}C chemical shift values and RNA structural conformations have been investigated by various research groups (see e.g., refs.[11-14]). In particular, the studies by Ebrahimi et al.[11] and more recently by Ohlenschläger et al.[13] have demonstrated that the chemical shift values of the ribose carbon atoms (C1′, C2′, C3′, C4′ and C5′) can be used to determine the ribose puckering [C2′ endo (south) vs. C3′ endo (north)] of RNA nucleotides. Ebrahimi et al.[11] also proposed a relationship between

the ribose carbon atoms chemical shift values and the γ backbone torsion angle, but a more recent study by Ohlenschläger et al.[13] did not support this relationship. A potential limitation to the use of ^{13}C chemical shifts is that RNA ^{13}C chemical shift data are often improperly referenced or contains inconsistencies, although this issue has now been addressed in a recent study by Aeschbacher and co-authors[15]. Lastly, in comparison to ^{1}H and ^{13}C chemical shifts, their have been very limited investigation into the connection between ^{15}N chemical shift values and RNA structural conformations[2].

The CS-ROSETTA-RNA framework would allow for the incorporation of ^{13}C and ^{15}N chemical shift data as pseudo-energy terms in a similar manner to how the non-exchangeable ^{1}H chemical shift data were incorporated as outlined in the main text and the Online Methods section. However, a program that can accurately back-calculate ^{13}C and ^{15}N chemical shift from RNA 3D structure will be required for this to happen. At the time that this study was conceived and carried-out, we searched the literature for such a program but found that none existed. After this study was completed and the manuscript is under revision after it has been submitted for publication, a program to back-calculate ^{13}C chemical shift from RNA 3D structure called RAMSEY became available[16]. The RAMSEY program employs a random forest (machine learning) approach to predict the chemical shifts of protonated ^{13}C atoms. We look forward to investigating the accuracy and robustness of the RAMSEY program for ^{13}C chemical shifts back-calculation, and the possible future incorporation of ^{13}C chemical shift data into CS-ROSETTA-RNA through RAMSEY.

Current and future uses of exchangeable ^{1}H chemical shifts for RNA structural modeling

Chemical shifts of exchangeable protons (those attached to nitrogen and oxygen) can provide important structural data about the macromolecule[17]. In RNA, the exchangeable protons include the H1 imino proton of guanine; the H3 imino proton of uracil; the H21 and H22 amino protons of guanine; the H41 and H42 amino protons of cytosine; H61 and H62 amino protons of adenine; and the HO2′ ribose proton.

Imino protons, in particular, can be used to help establish Watson-Crick and G:U wobble base pairing patterns. H1 imino protons of guanine give chemical shift values in the 12 – 13.5 ppm range when involved in a Watson-Crick base-pair, and in the 10 – 12 ppm range when involved in a G:U wobble base-pair[17]. Similarly, H3 imino protons of uracil give chemical shift values in the 13 – 15 ppm range when involved in a Watson-Crick base-pair and in the 11 – 12 ppm range when involved in a G:U wobble base-pair[17].

For our blind modeling targets (see main text Table 1), sequences and assigned chemical shifts for these targets, but no other information (e.g. secondary structure), were made available for chemical-shift-guided modeling. To establish the secondary structure of these RNAs, we utilized the mfold web server[18]. Due to the small size of the RNA molecules (few dozen nucleotides at most), we were able to unambiguously establish the correct secondary structure for all blind target cases. However, for longer RNA molecules, imino protons chemical shift values may be used to help establish the optimal secondary structure over the suboptimal ones (see e.g. RNA-PAIRS[19]).

Lastly, due to the strong dependence on its local environment (including factors such as ring current effect, local magnetic anisotropy, and electric field effect), exchangeable ^{1}H chemical shifts

should provide structural information that both supports and complements the information available in non-exchangeable ^{1}H chemical shifts. Currently, there exist no program that can accurately back-calculate exchangeable ^{1}H chemical shifts from RNA 3D structure[2] (e.g. in the same way that the NUCHEMICS program[4] can accurately back-calculate non-exchangeable ^{1}H chemical shifts from RNA 3D structure). However, given the recent increased interest in the use of chemical shifts for RNA structural modeling[20,21], we anticipate that such a program might become available in the near future. Once available, exchangeable ^{1}H chemical shift data can be incorporated as a pseudo-energy term under the CS-ROSETTA-RNA framework in a similar manner to how the non-exchangeable ^{1}H chemical shift data was incorporated as outlined in the main text and the Online Methods section.

Supplementary Note

Updates to the standard Rosetta all-atom energy function for RNA

Minor updates were made to the Rosetta all-atom energy function for RNA[22]. The energy unit reported herein is in Rosetta units (RU), which is used internally by the Rosetta program to store evaluated energy values. Comparisons to RNA Watson-Crick helix thermodynamic parameters[23] indicate that 1 Rosetta unit (RU) is approximately equal to 1 k_BT[22] (note that the modeling results in this study do not depend on the absolute scale of this energy unit). Compared to the Rosetta energy function used in the prior published work[24], two minor updates were made:

a. *Glycosidic (χ) torsional potential*: In the prior implementation of the glycosidic torsional potential, the glycosidic torsion at the *syn* minima (~69° if north pucker; ~70° if south pucker) were penalized by 2.2 RU relative to conformations at the *anti* minima (199° if north pucker; ~237° if south pucker) based on the lower frequency of *syn* glycosidic conformation found in the crystallographic structure of the large ribosomal subunit of *H. marismortuii* (PDB: 1JJ2). Prior to this work, in separate studies on RNA loop modeling[24] and *ab initio* motif structure prediction (unpublished results), we observed poor recovery of *syn* guanosines in noncanonical motifs. Hence, we have modified the glycosidic torsional potential to remove the energy penalty of the *syn* conformation for guanosine nucleotide.
b. *Intra-nucleotide energetic terms:* For historical reasons, the energetic contributions from the standard Rosetta all-atom energy terms (van der Waals interactions, hydrogen bonds and solvation effects) were omitted for atom-pairs belonging to the same nucleotide (the exception being a very weak van der Waals interactions repulsion term). These energy terms have been reintroduced here for all base-phosphate intra-nucleotide atom-pairs, including those forming potential hydrogen bonds. The energetic contribution from the standard Rosetta energy terms between intra-nucleotide base-ribose and ribose-phosphate atom-pairs are still presently omitted and instead captured through existing torsional potential terms[22].

Modeling of (primary and secondary) canonical base pairs

The primary focus of this study is to investigate whether noncanonical structural elements of RNAs can be accurately modeled through the incorporation of NMR ^{1}H chemical shift data. Nevertheless, these noncanonical core nucleotides are closed by boundary canonical (Watson-Crick and G:U wobble) base pairs, which also need to be explicitly modeled as well. However, given that (1) the structural modeling of the canonical base pairs is not the primary focus of this study and (2) the canonical base-pairing patterns can be determined via secondary structure prediction algorithms (e.g. mfold[18]), we decided to directly input the canonical base-pairing patterns as constraints during the structural modeling process (as was done in a previous study[22]).

Each motif in the 23-RNA benchmark consists of noncanonical core nucleotides closed by boundary canonical base pairs (see schematic below). As part of the motif definition, one (primary) canonical base pair is typically included at each helical boundary. These primary canonical base pairs

counts toward the motif size reported in Table 1 and are also explicitly represented in the schematics shown in Figures 1 and 2. Furthermore, the assigned ^{1}H chemical shift data of these primary canonical base pairs are also included when computing $rmsd_{shift}$ and the chemical shift pseudo-energy (E_{shift}) (see Supplementary Data for actual chemical shift data files used in this study).

During the modeling process, we also model an additional (secondary) canonical base pair at each helical boundary. The 2D schematic below illustrates the relative positioning of the core noncanconical nucleotides (black), the primary canonical base pairs (blue), and the secondary canonical base pairs (red) for the tandem GA:AG mismatch motif case:

```
5′-CGGACG-3′
  ||**||
3′-GCAGGC-5′
```

The secondary base pairs were explicitly modeled in this study for the following two reasons:

i. Improved accuracy of back-calculated chemical shifts: The ring current effect and the local magnetic anisotropy contributions of the secondary base pairs were explicitly included when back-calculating the chemical shift value of the protons found in the primary canonical base pairs and the noncanonical core nucleotides. Since the chemical shift value of each proton is highly dependent on the positions and orientations of nearby nucleobases[4], explicitly modeling the secondary canonical base pairs allowed for improved agreement between the back-calculated and the experimental chemical shifts. Lastly, note that the assigned ^{1}H chemical shift data of the secondary canonical base pairs themselves were not included when computing $rmsd_{shift}$ and the chemical shift pseudo-energy (E_{shift}).

ii. Realistic steric occlusion: The atoms in the secondary canonical base pairs fill up the 3D space next to the primary canonical base pairs. In the absence of the secondary canonical base pairs, we occasionally observed CS-ROSETTA-RNA models where the noncanonical core nucleotides adopt unrealistic conformations that occupy the voided space that would otherwise be occupied by the secondary canonical base pairs. Explicitly modeling the secondary canonical base pairs sterically occludes these unrealistic CS-ROSETTA-RNA conformations and provides a more realistic representation of the sterically allowed conformational space.

Finally, please note that we have included these secondary canonical base pairs in all of the PDB files provided in Supplementary Data.

CS-ROSETTA-RNA modeling and refinement of the 4x4 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop

Among the 11 blind motifs in the 23-RNA benchmark, the 4x4 5′-GA^{Br}GU-3′/3′-U^{Br}GAG-5′ self-complementary internal loop (^{Br}G, denoting 8-bromoguanosine, was introduced to stabilize the major of

two conformations observed in the RNA) was unique in that an initial NMR study on the motif was already published prior to the CS-ROSETTA-RNA modeling[8]. This prior study did not produce sufficient restraints to generate an atomic-detail three-dimensional structure, but a descriptive 2D schematic was proposed (see Supplementary Fig. 10A). We therefore took the following approach to the modeling process. First, half of the features proposed by the 2D schematic were incorporated as constraints during the modeling process, and the other half were left out and subsequently used to validate the accuracy of the resulting model. The features incorporated as constraints during the modeling process were:

i). G4, G4*, G6 and G6* were assumed to adopt the *syn glycosidic* conformation.

ii). G4-G6* and G4*-G6 were assumed to form base pairs (but the exact base-pairing pattern was not specified).

iii). U7 and U7* were assumed to be extra-helical bulges.

In this case, only Stepwise Assembly[24] (SWA) modeling was performed since the Fragment Assembly with Full-Atom Refinement[22] (FARFAR) method did not have a framework to support incorporating the constraints. Additional features proposed in the 2D schematic that were not used as modeling constraints were:

i). The exact base-pairing pattern (*trans* Watson-Crick/Hoogsteen) of G4-G6* and G4*-G6.

ii). Two-fold rotational symmetry of the structure.

iii.) 2′-endo ribose conformations at G4, G4*, A5, A5*, G6, and G6*.

The CS-ROSETTA-RNA model was able to independently recover all of these additional features.

The original blind modeling of the major conformation of the 5′-GAGU-3′/3′-UGAG-5′ self-complementary internal loop was also completed during the early development stage of the CS-ROSETTA-RNA method. At the time of the original modeling, the *intra-nucleotide energetic terms* update to the Rosetta all-atom energy function for RNA (see above) was not yet implemented. The lowest energy model obtained from this original CS-ROSETTA-RNA modeling (referred to here as the original CS-ROSETTA-RNA model) thus contained steric clashes between the amino atoms in the base and the phosphate group atoms of both the G4 and G4* nucleotides. Once the *intra-nucleotide energetic terms* update to the Rosetta all-atom energy function for RNA was implemented, the original CS-ROSETTA-RNA model was refined using the following protocol: Stepwise Assembly (SWA) modeling was carried out again on the RNA motif, but now focused only on conformations near the original CS-ROSETTA-RNA model; a filter was imposed at each SWA building step requiring models to be with 3.0 Å rmsd of the original CS-ROSETTA-RNA model. The resulting refined CS-ROSETTA-RNA model generated from this procedure closely resembles the original CS-ROSETTA-RNA model (1.20 Å pairwise all-heavy-atom rmsd) and retained all the base-pairing and base-stacking features. The refined model also removed steric clashes between the amino atoms in the base and the phosphate group atoms of both the G4 and G4* nucleotides, as expected, and now contains an energetically favorable hydrogen bond between the 1H2 amino proton and the OP1 phosphate oxygen at these nucleotides.

The refined 5′-GAGU-3′/3′-UGAG-5′ CS-ROSETTA-RNA model was used to generate all of the modeling results presented in this article (e.g. the 3D models in the figures, the back-calculated chemical shifts, the rmsd calculations). The only exception is the E_{gap} result (28.1 k_BT), which was based on the original CS-ROSETTA-RNA model. Due to the close similarity between the original and refined CS-ROSETTA-RNA models (1.20 Å pairwise all-heavy-atom rmsd), we could have based our analysis on either model and arrive at essentially the same results and conclusions presented in the article. Both the original and refined CS-ROSETTA-RNA models were within atomic-accuracy of the experimental NMR conformation (PDB: 2LX1; the rmsds to the experimental conformation values were, respectively, 1.07 Å and 1.10 Å). Both the original and refined CS-ROSETTA-RNA models also gave back-calculated chemical shifts that strongly agree with the experimental non-exchangeable ^{1}H chemical shift data (the $rmsd_{shift}$ values were, respectively, 0.22 ppm and 0.18 ppm).

References

1. Prado, F.R. & Giessner-Prettre, C. Parameters for the calculation of the ring current and atomic magnetic anisotropy contributions to magnetic shielding constants: Nucleic acid bases and intercalating agents. *Journal of Molecular Structure: THEOCHEM* **76**, 81-92 (1981).
2. Giessner-Prettre, C. & Pullman, B. Quantum mechanical calculations of NMR chemical shifts in nucleic acids. *Q Rev Biophys* **20**, 113-172 (1987).
3. Wijmenga, S.S., Kruithof, M. & Hilbers, C.W. Analysis of (1)H chemical shifts in DNA: Assessment of the reliability of (1)H chemical shift calculations for use in structure refinement. *J Biomol NMR* **10**, 337-350 (1997).
4. Cromsigt, J.A., Hilbers, C.W. & Wijmenga, S.S. Prediction of proton chemical shifts in RNA. Their use in structure refinement and validation. *J Biomol NMR* **21**, 11-29 (2001).
5. Ulrich, E.L. et al. BioMagResBank. *Nucleic Acids Res* **36**, D402-408 (2008).
6. Burkard, M.E. & Turner, D.H. NMR structures of r(GCAGGCGUGC)2 and determinants of stability for single guanosine-guanosine base pairs. *Biochemistry* **39**, 11748-11762 (2000).
7. Lukavsky, P.J., Kim, I., Otto, G.A. & Puglisi, J.D. Structure of HCV IRES domain II determined by NMR. *Nat Struct Biol* **10**, 1033-1038 (2003).
8. Hammond, N.B., Tolbert, B.S., Kierzek, R., Turner, D.H. & Kennedy, S.D. RNA internal loops with tandem AG pairs: the structure of the 5'GAGU/3'UGAG loop can be dramatically different from others, including 5'AAGU/3'UGAA. *Biochemistry* **49**, 5817-5827 (2010).
9. Sarver, M., Zirbel, C., Stombaugh, J., Mokdad, A. & Leontis, N. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of Mathematical Biology* **56**, 215-252 (2008).
10. Kennedy, S.D., Kierzek, R. & Turner, D.H. Novel conformation of an RNA structural switch. *Biochemistry* **51**, 9257-9259 (2012).
11. Ebrahimi, M., Rossi, P., Rogers, C. & Harbison, G.S. Dependence of 13C NMR chemical shifts on conformations of rna nucleosides and nucleotides. *J Magn Reson* **150**, 1-9 (2001).
12. Fares, C., Amata, I. & Carlomagno, T. 13C-detection in RNA bases: revealing structure-chemical shift relationships. *J Am Chem Soc* **129**, 15814-15823 (2007).
13. Ohlenschlager, O., Haumann, S., Ramachandran, R. & Gorlach, M. Conformational signatures of 13C chemical shifts in RNA ribose. *J Biomol NMR* **42**, 139-142 (2008).
14. Suardiaz, R., Sahakyan, A.B. & Vendruscolo, M. A geometrical parametrization of C1'-C5' RNA ribose chemical shifts calculated by density functional theory. *J Chem Phys* **139**, 034101 (2013).
15. Aeschbacher, T., Schubert, M. & Allain, F.H. A procedure to validate and correct the 13C chemical shift calibration of RNA datasets. *J Biomol NMR* **52**, 179-190 (2012).
16. Frank, A.T., Bae, S.H. & Stelzer, A.C. Prediction of RNA (1)h and (13)c chemical shifts: a structure based approach. *J Phys Chem B* **117**, 13497-13506 (2013).
17. Furtig, B., Richter, C., Wohnert, J. & Schwalbe, H. NMR spectroscopy of RNA. *Chembiochem* **4**, 936-962 (2003).
18. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).
19. Bahrami, A., Clos, L.J., 2nd, Markley, J.L., Butcher, S.E. & Eghbalnia, H.R. RNA-PAIRS: RNA probabilistic assignment of imino resonance shifts. *J Biomol NMR* **52**, 289-302 (2012).
20. van der Werf, R.M., Tessari, M. & Wijmenga, S.S. Nucleic acid helix structure determination from NMR proton chemical shifts. *J Biomol NMR* (2013).
21. Frank, A.T., Horowitz, S., Andricioaei, I. & Al-Hashimi, H.M. Utility of (1)h NMR chemical shifts in determining RNA structure and dynamics. *J Phys Chem B* **117**, 2045-2052 (2013).
22. Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical

RNA structure. *Nat Methods* **7**, 291-294 (2010).
23. Xia, T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719-14735 (1998).
24. Sripakdeevong, P., Kladwang, W. & Das, R. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proc Natl Acad Sci U S A* **108**, 20573-20578 (2011).