BIOCHEMISTRY
including biophysical chemistry & molecular biology

Rapid Report

pubs.acs.org/biochemistry

Quantitative Dimethyl Sulfate Mapping for Automated RNA Secondary Structure Inference

Pablo Cordero,[†] Wipapat Kladwang,[‡] Christopher C. VanLang,[§] and Rhiju Das*[,†,‡,∥]

Departments of [†]Biomedical Informatics, [‡]Biochemistry, [§]Chemical Engineering, and [∥]Physics, Stanford University, Stanford, California 94305, United States

Supporting Information

ABSTRACT: For decades, dimethyl sulfate (DMS) mapping has informed manual modeling of RNA structure in vitro and in vivo. Here, we incorporate DMS data into automated secondary structure inference using an energy minimization framework developed for 2′-OH acylation (SHAPE) mapping. On six noncoding RNAs with crystallographic models, DMS-guided modeling achieves overall false negative and false discovery rates of 9.5% and 11.6%, respectively, comparable to or better than those of SHAPE-guided modeling, and bootstrapping provides straightforward confidence estimates. Integrating DMS−SHAPE data and including 1-cyclohexyl(2-morpholinoethyl) carbodiimide metho-*p*-toluene sulfonate (CMCT) reactivities provide small additional improvements. These results establish DMS mapping, an already routine technique, as a quantitative tool for unbiased RNA secondary structure modeling.

Understanding the many biological functions of RNAs, from genetic regulation to catalysis, requires accurate portraits of the RNAs' folds. Among biochemical tools available for interrogating RNA structure, chemical mapping or "footprinting" uniquely permits rapid characterization of any RNA or ribonucleoprotein system in solution at single-nucleotide resolution (see, e.g., refs 1 and 2). Chemical mapping is being advanced by several groups through new approaches for chemical modification, coupling to high-throughput readouts, rapid data processing, high-throughput mutagenesis, and incorporation into structure prediction algorithms.[3−7]

Perhaps the most widely used RNA chemical probe is dimethyl sulfate (DMS).[8−11] DMS modification of the Watson−Crick edge of adenosines or cytosines (at N1 or N3, respectively) blocks reverse transcription, so that reactivities can be obtained by primer extension at single-nucleotide resolution. Nucleotides that appear to be most strongly protected or reactive to DMS can be inferred to be base-paired or unpaired, respectively. This qualitative or "binary" information can be used for RNA structure modeling by manual or automatic methods.[10,12] More recently developed methods, such as selective 2′-hydroxyl acylation with primer extension (SHAPE),[6] give reactivities that correlate with Watson−Crick base pairing for all nucleotide types, providing more data points than DMS. Indeed, when incorporated into free energy minimization algorithms as energetic bonuses, called pseudoenergies, SHAPE data can recover RNA secondary structures with a high level of accuracy.[11] Further, nonparametric bootstrapping (repeating the algorithms on data sets resampled with replacement) can identify regions with poor confidence.[13] Nevertheless, this pseudoenergy framework has not been leveraged for prior chemical approaches such as DMS mapping, despite the wide use of these data in in vitro, in vivo, and in virio contexts.[9,12,14,15]

We present herein a benchmark of pseudoenergy-guided secondary structure modeling based on DMS data for six noncoding RNAs: unmodified *Escherichia coli* tRNA$^{\mathrm{phe}}$,[16] the P4−P6 domain of the *Tetrahymena* group I ribozyme,[17] *E. coli* 5S rRNA,[12] and three ligand-bound domains from bacterial riboswitches (the *Vibrio vulnificus add* adenine riboswitch,[18] the *Vibrio cholerae* cyclic di-GMP riboswitch,[19] and the *Fusobacterium nucleatum* glycine riboswitch[20]). In all cases, crystallographic data, confirmed by solution analyses with the two-dimensional mutate-and-map approach,[21] have provided "gold-standard" secondary structures (Table S1 of the Supporting Information) for evaluating the method's accuracy. The challenging nature of this benchmark is confirmed by the poor accuracy of the *RNAstructure* algorithm without data (Table 1). These models miss 38% of true helices [false negative rate (FNR)], and 45% of the returned helices are incorrect [false discovery rate (FDR)].

We measured DMS reactivities and estimated errors, inferred from three to eight replicates for each of the six RNAs (Figures S4−S9 and Table S1 of the Supporting Information). Analogous to prior SHAPE studies,[11,13] we incorporated these DMS data into *RNAstructure* by transforming them into pseudoenergies, giving favorable energies or penalties depending on whether paired nucleotides were DMS-protected or reactive, respectively. We tested pseudoenergy frameworks based on both a previous ad hoc formula and an empirically derived statistical potential [inspired by techniques in three-dimensional structure prediction (see Methods and Figure S1 of the Supporting Information)]. The two methods gave consistent secondary structures. Because primer extension primarily reads out DMS reactivity at adenosines and cytosines, we excluded reactivities at other bases when performing structure modeling. DMS-guided modeling of the six ncRNAs gave an FNR of 9.5% and an FDR of 11.6% (Table 1 and Figure 1; see also Table S2 of the Supporting Information), more than 3-fold better than without the data. These error rates

Received: July 2, 2012
Revised: August 12, 2012
Published: August 22, 2012

Table 1. Performance of Free Energy Minimization Guided by Reactivity-Derived Pseudoenergies from DMS and SHAPE Chemical Modifications[a]

		no data		DMS		SHAPE		DMS and SHAPE	
	cryst.	TP	FP	TP	FP	TP	FP	TP	FP
tRNA^{phe}	4	2	3	4	0	4	0	4	0
adenine riboswitch	3	2	3	3	1	3	1	3	1
cdGMP riboswitch	8	6	2	6	0	8	0	8	0
5S rRNA	7	1	9	6	3	6	3	6	3
P4−P6 RNA	11	10	1	10	1	9	1	9	1
glycine riboswitch	9	5	3	9	0	8	0	9	0
total	42	26	21	38	6	38	5	39	5
FNR		38.1%		9.5%		9.5%		7.1%	
FDR		44.7%		11.6%		13.6%		11.4%	
sensitivity		61.9%		90.5%		90.5%		92.9%	
PPV		55.3%		88.4%		86.4%		88.6%	

[a] Abbreviations: TP, true positives; FP, false positives; cryst., number of helices in the crystallographic model; FNR, false negative rate (1 − TP/total); FDR, false discovery rate [FP/(TP + FP)]; sensitivity, 1 − FNR; PPV, positive predictive value (1 − FDR).

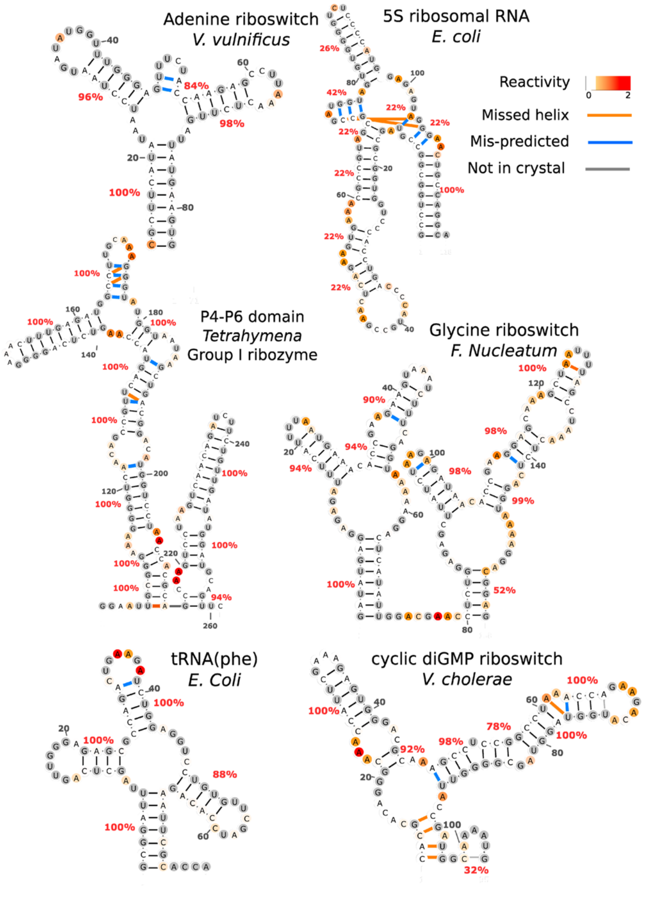


Figure 1. Pseudoenergy-guided secondary structure models using DMS data on six noncoding RNAs. DMS data and secondary structure models for *E. coli* tRNA^{phe}, the P4−P6 domain of the *Tetrahymena* group I ribozyme, *E. coli* 5S rRNA, the *V. vulnificus add* adenine riboswitch, the *V. cholerae* cyclic di-GMP riboswitch, and the *F. nucleatum* glycine riboswitch. Missed base pairs are highlighted with blue lines; mispredicted base pairs are denoted with orange lines. Helix bootstrap confidence values are colored red. G and U nucleotides that do not give DMS signals in primer extension and nucleotides with unavailable reactivities are colored gray.

are lower than those previously achieved by SHAPE-directed modeling (FNR of 17% and FDR of 21% on the same RNAs[13]). Furthermore, the DMS-guided FNR and FDR values are equal to and lower than, respectively, values for SHAPE-based measurements in which primer extension was conducted without deoxyinosine triphosphate (FNR of 9.6% and FDR of 13.6%) to avoid known artifacts.[13]

We were surprised that DMS mapping gave similar or better information content, compared to SHAPE data, as the latter provides reactivities at approximately twice the number of nucleotides per RNA. [Indeed, restricting the algorithm to use SHAPE data at adenines and cytosines gave worse models (see Table S3 of the Supporting Information).] An explanation for our results derives from distinct SHAPE and DMS signatures at nucleotides that are not in Watson−Crick secondary structure but that nevertheless form noncanonical interactions [see, e.g., A37 in the *F. nucleatum* glycine riboswitch (Figure 2A)]. These nucleotides appear to be protected from the SHAPE reaction and thus receive pseudoenergies that incorrectly reward their pairings inside Watson−Crick secondary structure. However, these same nucleotides can expose their Watson−Crick edges to solvent and react strongly with DMS, signifying that they are outside Watson−Crick helices. The DMS-guided modeling can thus return the correct secondary structure in regions where the SHAPE data cannot distinguish Watson−Crick from non-Watson−Crick base pairs (compare panels B and C of Figure 2).

Reactivity histograms (Figure 2D,E) further support the enhanced predictive power of DMS vis-à-vis SHAPE. DMS mapping better distinguishes between nucleotides inside Watson−Crick helices and nucleotides outside helices [see also the receiver operating characteristic curve and quantitation (Figure S2 of the Supporting Information)].

Like SHAPE-guided modeling, DMS-directed structure inference still produces errors (Table 1), e.g., for the central junction of the 5S rRNA (Figure 1). Some of these errors may be resolved through better incorporation of the DMS-derived pseudoenergies at, e.g., isolated, or "singlet", base pairs. Nevertheless, as with SHAPE modeling, these erroneous regions can be pinpointed by estimating helix-by-helix confidence values through nonparametric bootstrapping (Methods of the Supporting Information and ref 13; see also Figure S3 of the Supporting Information). For example, this procedure gives a high degree of confidence (≥90%) at almost all helices in the correctly recovered structure of the glycine riboswitch but low levels of confidence (<50%) throughout the imperfect 5S rRNA DMS model (Figure 1).

For many applications, DMS and SHAPE measurements can be acquired in parallel, so we sought to determine if their combination might improve automated secondary structure inference. Application of both sets of pseudoenergies gave a slight improvement in the algorithm's accuracy (FNR of 7.1% and FDR of 11.4%). In addition, we performed measurements with a reagent that primarily modifies Waston−Crick edges of guanosine and uracil, 1-cyclohexyl(2-morpholinoethyl) carbodiimide metho-*p*-toluene sulfonate (CMCT).[22] Incorporation of these data into *RNAstructure* gave poorer accuracy modeling than the DMS- or SHAPE-guided modeling described above [FNR of 14.3% and FDR or 18.2% (see Table S4 of the Supporting Information)], consistent with weaker discrimination between paired and unpaired residues (Figures S1 and

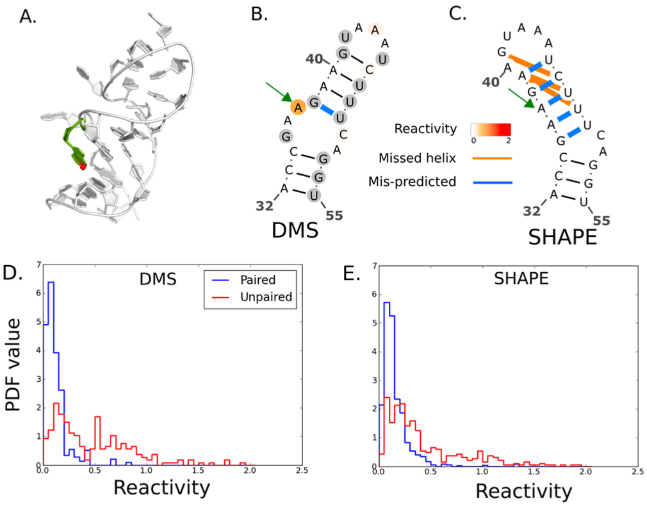


Figure 2. DMS vis-à-vis SHAPE for secondary structure inference. (A) The P3 hairpin of the glycine riboswitch is correctly predicted by DMS-guided modeling, but not by SHAPE. A37 (green) has its Watson−Crick edge exposed, making its N1 atom (red sphere) accessible to DMS modification that guides *RNAstructure* to the correct helix (B). However, A37 (green arrow) is stabilized by local interactions, protecting it from SHAPE modification, resulting in an incorrect SHAPE-predicted helix (C). (D and E) Reactivity histograms for DMS (D) and SHAPE (E) for all chemical mapping data on the six noncoding RNAs.

S2 of the Supporting Information). Integrating CMCT with DMS and/or SHAPE data did not improve accuracy (Table S2 of the Supporting Information). CMCT gives weak reactivities in bases that are unpaired but still stacked (e.g., see ref 23), reducing its information content for discriminating unpaired and paired nucleotides.

The benchmark results presented here establish that chemical mapping with DMS can achieve prediction accuracies comparable to those of the SHAPE protocol using pseudoenergies to guide free energy minimization. DMS has been extensively used both in vitro and in vivo, for time-resolved RNA folding, precise thermodynamic analysis, and mapping RNA−protein interfaces.[9,12,14,15,22] Sophisticated techniques for optimizing the reaction rate and its quenching have been developed.[9,24] Applying automated structure modeling, as demonstrated herein, will allow researchers to better take advantage of this large body of previous work. Furthermore, future studies may find it advantageous to perform both DMS and SHAPE approaches in parallel. Along with bootstrapping,[13] comparison of separate DMS-guided versus SHAPE-guided secondary structure models will permit rapid assessment of systematic errors and thus provide more accurate inferences.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Supporting methods, figures, and model accuracy tables. This material is available free of charge via the Internet at http://pubs.acs.org. Single nucleotide-resolution data are available at http://rmdb.stanford.edu.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rhiju@stanford.edu. Phone: (650) 723-5976. Fax: (650) 723-6783.

Funding

This work is supported by the Burroughs-Wellcome Foundation (CASI to R.D.), a Hewlett-Packard Stanford Graduate Fellowship (to P.C.), and the National Institutes of Health (T32 HG000044 to C.C.V. and R01 GM102519 to R.D.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the authors of *RNAstructure* for making the source code freely available and members of the Das lab for comments on the manuscript.

■ REFERENCES

(1) Black, D. L., and Pinto, A. L. (1989) *Mol. Cell. Biol. 9*, 3350−3359.

(2) Moazed, D., and Noller, H. F. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 3725−3728.

(3) Mitra, S., Shcherbakova, I. V., Altman, R. B., Brenowitz, M., and Laederach, A. (2008) *Nucleic Acids Res. 36*, e63.

(4) Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011) *Bioinformatics 27*, 1798−1805.

(5) Kladwang, W., Cordero, P., and Das, R. (2011) *RNA 17*, 522−534.

(6) Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006) *Nat. Protoc. 1*, 1610−1616.

(7) Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. (2011) *Proc. Natl. Acad. Sci. U.S.A. 108*, 11063−11068.

(8) Peattie, D. A., and Gilbert, W. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 4679−4682.

(9) Tijerina, P., Mohr, S., and Russell, R. (2007) *Nat. Protoc. 2*, 2608−2623.

(10) Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004) *Proc. Natl. Acad. Sci. U.S.A. 101*, 7287−7292.

(11) Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009) *Proc. Natl. Acad. Sci. U.S.A. 106*, 97−102.

(12) Leontis, N. B., and Westhof, E. (1998) *RNA 4*, 1134−1153.

(13) Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) *Biochemistry 50*, 8049−8056.

(14) Lempereur, L., Nicoloso, M., Riehl, N., Ehresmann, C., Ehresmann, B., and Bachellerie, J. P. (1985) *Nucleic Acids Res. 13*, 8339.

(15) Wells, S. E., Hughes, J. M., Igel, A. H., and Ares, M. (2000) *Methods Enzymol. 318*, 479−493.

(16) Byrne, R. T., Konevega, A. L., Rodnina, M. V., and Antson, A. A. (2010) *Nucleic Acids Res. 38*, 4154−4162.

(17) Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R., and Doudna, J. A. (1996) *Science 273*, 1678−1685.

(18) Serganov, A., Yuan, Y.-R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T., Hobartner, C., Micura, R., Breaker, R. R., and Patel, D. J. (2004) *Chem. Biol. 11*, 1729−1741.

(19) Smith, K. D., Lipchock, S. V., Livingston, A. L., Shanahan, C. A., and Strobel, S. A. (2010) *Biochemistry 49*, 7351−7359.

(20) Butler, E. B., Xiong, Y., Wang, J., and Strobel, S. A. (2011) *Chem. Biol. 18*, 293−298.

(21) Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) *Nat. Chem. 3*, 954−962.

(22) Planning, S. (2000) in *Current protocols in nucleic acid chemistry* (Beaucage, S. L., et al., Eds.) Chapter 6, pp 1−21, Wiley, New York.

(23) Sripakdeevong, P., Kladwang, W., and Das, R. (2011) *Proc. Natl. Acad. Sci. U.S.A. 108*, 20573−20578.

(24) Das, R., Karanicolas, J., and Baker, D. (2010) *Nat. Methods 7*, 291−294.

Supporting information

for the manuscript

Quantitative DMS mapping for automated RNA secondary structure inference

Pablo Cordero[1], Wipapat Kladwang[2], Christopher VanLang[3], and Rhiju Das[1,2,4*]
Departments of Biomedical Informatics[1], Biochemistry[2], Chemical Engineering[3], and Physics[4],
Stanford University, Stanford CA 94305

This document contains the following sections:

Supporting Methods

Experimental setup

Chemical mapping experiments were performed using *in vitro* transcribed RNAs from PCR-assembled DNA templates as previously described (*1*). All SHAPE, DMS, and CMCT measurements were performed at least in triplicate using three independent RNA preparations. DNA templates containing a T7 RNA polymerase promoter sequence (TTCTAATACGACTCACTATA) followed by the sequence of interest and a reverse transcription primer binding site (AAAGAAACAACAACAACAAC) were PCR-assembled from oligomers of up to 60 nucleotides in length (Integrated DNA Technologies) with a Phusion DNA polymerase (Finnzymes) and purified with AMPure magnetic beads (Agencourt, Beckman Coulter). Sample concentrations were calculated through UV absorbances on a Nanodrop 100 spectrophotometer, and lengths were verified in 4% agarose gels. *In vitro* RNA transcription was performed as previously described using a T7 RNA polymerase (New England Biolabs) and purified with MagMax magnetic beads (Ambion) or an RNA clean kit (Zymo research); RNA from the two purification methods gave indistinguishable results. RNA concentrations were measured on a Nanodrop 100 spectrophotometer.

Chemical modification was performed in volumes of 20 μL with 1.2 pmols of RNA in 50 mM Na-HEPES (pH 8.0), 10 mM $MgCl_2$, ligand at the desired concentration for riboswitches (see Table S1) and 5 μL of modification reagent [1% dimethyl sulfate (DMS) prepared by mixing 10 μL DMS into 90 μL ethanol, and then 900 μL water; 42 mg/mL 1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate (CMCT); or 24 mg/mL N-methylisatoic anhydride (NMIA)]. Modification reactions were incubated at 24 °C for 15 to 60 minutes depending on the length of the RNA to achieve overall modification rates of less than 30% and then quenched appropriately (adding 5 μL of 0.5 M Na-MES, pH 6.0 for SHAPE and CMCT, or 5 μL of 2-mercaptoethanol for DMS). Quenches also included 1μL of poly(dT) magnetic beads (Ambion) and 0.065 pmols of 5′-rhodamine-green-labeled primer

(AAAAAAAAAAAAAAAAAAAAGTTGTTGTTGTTGTTTCTTT) complementary to the 3′ end of the RNAs used for reverse transcription. The reaction mixtures were purified by magnetic separation, rinsed with 40 μL of 70% ethanol twice, and allowed to air-dry for 10 min while sitting in the magnetic post stand. The magnetic bead mixtures were resuspended in 2.5 μL of deionized water and reverse transcribed by adding a premix solution containing 0.2 μL of SuperScript III (Invitrogen), 1.0 μL of 5× SuperScript First Strand buffer (Invitrogen), 0.4 μL of dNTPs at 10 mM each (dATP, dCTP, dGTP, and dTTP; dITP was not used to reverse-transcribe these RNAs, as it generates signal artifacts in NMIA chemical mapping protocols (*1*)), 0.25 μL of 0.1 M dithiothreitol (DTT), and 0.65 μL of water and incubating at 42 °C for 30 min. RNA was hydrolyzed by adding 5 μL of 0.4 M NaOH and incubating at 90 °C for 3 min. The solutions were neutralized by the addition of 5 μL of an acid quench (2 volumes of 5 M NaCl, 2 volumes of 2 M HCl, and 3 volumes of 3 M sodium acetate) and the resulting fluorescent DNA was purified by magnetic bead separation. The beads were washed with 40 μL of 70% ethanol, air-dried for 5 minutes, and resuspended in 10 μL of a solution containing 0.125 mM Na-EDTA (pH 8.0) and a Texas Red-labeled reference ladder, or in 10 μL of HiDi formamide containing a ROX 350 ladder (Applied Biosystems) spectrally distinct from the rhodamine-green chemical mapping signal. For verifying sequence assignments, reference ladders were created using an analogous protocol without chemical modification and including 2′,3′-dideoxy-GTP in an amount equimolar with dGTP during reverse transcription. The products were separated by capillary electrophoresis on an ABI 3100 or ABI 3700 DNA sequencer.

Analysis of electropherogram traces and quantification of reactivities

Electropherograms were analyzed with the HiTRACE software (*2*). Sequence assignments for bands were obtained by alignment to sequencing lanes with incorporated ddATP, ddCTP, ddGTP, or ddTTP nucleotides. Band intensities were fit as Gaussian peaks and processed through a likelihood-based framework for overmodification correction and background subtraction as defined previously through the *overmod_and_background_correct_logL.m* and *get_average_standard_state.m* HiTRACE scripts (*1*). For DMS or CMCT data, reactivities at guanines/uracils or adenines/cytosines, respectively, are expected to be low. Therefore, when performing likelihood-based background substraction on data for those nucleotides, we used a distribution of the form $p(x) = \exp(F(x))$ where $F(x) = F_{+}(x - x_0)$ if $x > x_0$ and $F(x) = -F_{-}(x - x_0)$ otherwise, with parameter values $F_{-} = F_{+} = 25,\ x_0 = 0$ (for details on this functional form see ref (*1*)). This distribution corresponds to positions with lower expected reactivities, thus attenuating the final reactivity values for those nucleotides.

Final averaged data and errors have been made publicly available in the RNA Mapping Database (http://rmdb.stanford.edu). Accession IDs corresponding to each modifier are:

- For NMIA (SHAPE) modified samples, entries TRNAPH_SHP_0003, TRP4P6_SHP_0004, 5SRRNA_SHP_0003, ADDRSW_SHP_0004, CIDGMP_SHP_0003, and GLYCFN_SHP_0006 were submitted.
- For DMS modified samples, new entries TRNAPH_DMS_0001, TRP4P6_DMS_0001, 5SRRNA_DMS_0001, ADDRSW_DMS_0001, CIDGMP_DMS_0001, and GLYCFN_DMS_0001 were submitted.
- For CMCT modified samples, new entries TRNAPH_CMC_0001, TRP4P6_CMC_0001, 5SRRNA_CMC_0001, ADDRSW_CMC_0001, CIDGMP_CMC_0001, and GLYCFN_CMC_0001 were submitted.

Computational methods

We tested the modeling accuracy of minimum free energy structure calculation with reactivity-derived pseudo-energies added to the scoring function (*3*). The *Fold* executable of the RNAstructure package (version 5.3) was used to infer pseudo-energy-directed secondary structure models. Pseudo-energies were applied once for each nucleotide that forms an edge base pair and twice for each nucleotide that forms an internal base pair. Additionally, non-parametric bootstrap analysis was performed to estimate helix-wise prediction confidence (*1*).

Previous work used an energy-like functional form with two free parameters to calculate pseudo-energies from experimental chemical mapping reactivities that are given as bonuses or penalties to the energy scoring function of a secondary structure prediction algorithm ($\Delta G_i = m \log(S_i + 1) + b$, S_i is the reactivity value at position *i; m* and *b* are free parameters, see ref (*3*)). We also tested a more direct way of expressing the pseudo-energy potential by taking the log-likelihood ratio of a base being unpaired versus paired given a chemical reactivity value:

$$\Delta G = -k_B T \log\left(\frac{P(S_i \mid i \text{ is paired})}{P(S_i \mid i \text{ is unpaired})}\right)$$

Here, T is the temperature and k_B is the Boltzmann constant. The likelihoods for paired and unpaired reactivities were derived from a mixture of two gamma distributions to reactivities of paired and unpaired nucleotides in our non-coding RNA benchmark (see Figure S4). This probabilistic potential is akin to those found in forcefields that include knowledge-based terms, such as the ROSETTA framework for three-dimensional structure modeling (*4*–*6*). In the future, if different reactivity distributions are discovered for different features (e.g., apical loops and interior loops), this framework permits the facile incorporation of that information.

We applied our probabilistic potential to calculate pseudo-energies to guide the free-energy minimization *Fold* program in the RNAstructure package. The performance of the algorithm using this probabilistic potential for SHAPE, DMS, and CMCT reactivities is given in Table S6 and is identical to results obtained for the standard potential with slope (*m*) and intercept (*b*) optimized through grid-search. To test for over-fitting, we performed leave-one-out-validations for each RNA by fitting the probabilistic potential using the mapping data of the other RNAs and re-running the algorithm; validation results were identical to those when using the full data. RNAstructure was modified to allow the DMS and CMCT data to be input through the flags *-dms* and *-cmct*. These options are being made available in release 5.5 of RNAstructure. A structure prediction server that includes these options is available at http://rmdb.stanford.edu/structureserver.

Assessment of accuracy

We evaluated the predictions as defined previously in refs (*1*, *7*): a crystallographic helix was considered correctly recovered if more than 50% of its base pairs were observed in a helix by the computational model; ±1 helix shifts were not considered correct. Modeling errors are expressed as false negative rates (FNR; fraction of crystallographic helices that were predicted to be single-stranded) and false discovery rates (FDR; fraction of predicted helices that were not present in crystallographic models). We also report positive predictive values (PPV) and sensitivities of each approach, and all metrics at the level of individual base pairs rather than helices (see Table S6).

Table S1: RNA systems used in this study

RNA, source	**Solution conditions**[a]	**Replicates**	**Experiments**	**Offset**[b]	**PDB**[c]
tRNAphe, *E. coli*	Standard	SHAPE: 6 DMS: 5 CMCT: 4	SHAPE: 4 DMS: 4 CMCT: 3	-15	**3L0U** 1EHZ
P4-P6 domain, *Tetrahymena* ribozyme	Standard	SHAPE: 11 DMS: 11 CMCT: 4	SHAPE: 5 DMS: 5 CMCT: 3	89	**1GID** 1L8V 1HR2 2R8S
5S rRNA, *E. coli*	Standard	SHAPE: 5 DMS: 5 CMCT: 3	SHAPE: 3 DMS: 4 CMCT: 3	-20	**3OFC** 3OAS 3ORB 2WWQ …
Adenine riboswitch, *V. vulnificus* (*add*)	Standard + 5 mM adenine	SHAPE: 4 DMS: 3 CMCT: 3	SHAPE: 3 DMS: 3 CMCT: 3	-8	**1Y26** 1Y27 2G9C 3GO2 …
c-di-GMP riboswitch, *V. cholerae* (*VC1722)*	Standard + 10 µM cyclic di-guanosine monophosphate	SHAPE: 5 DMS: 3 CMCT: 3	SHAPE: 3 DMS: 2 CMCT: 3	0	**3MXH** 3IWN 3MUV 3MUT …
Glycine riboswitch, *F. nucleatum*	Standard + 10 mM glycine	SHAPE: 16 DMS: 7 CMCT: 9	SHAPE: 4 DMS: 3 CMCT: 3	-10	**3P49**

[a] Standard conditions were: 10 mM $MgCl_2$, 50 mM Na-HEPES, pH 8.0 at 24 °C.

[b] Offset added to the original numbering scheme of the sequence from which this subsequence was taken from.

[c] Boldfaced IDs correspond to the PDB entries from which the sequence for this study was taken. Additional PDB entries correspond to other studies of the same RNA system.

RNA, source	Sequence[d]	Secondary Structure[e]
tRNAphe, *E. coli*	ggaacaaacaaaacaGCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCAaaaccaaagaaacaacaacaacaac	...............(((((((..((((........)))).((((.........)))).....(((((.......)))))))))))).............................
P4-P6 domain, *Tetrahymena* ribozyme	ggccaaaacaacgGAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGCAAAGGGUAUGGUAAUAAGCUGACGGACAUGGUCCUAACCACGCAGCCAAGUCCUAAGUCAACAGAUCUUCUGUUGAUAUGGAUGCAGUUCAaaaccaaaccaaagaaacaacaacaacaac	.................((((((...((((((.....(((.((((.(((..(((((((((....)))))))))..((.......))....)))......)))))))....))))))..)).))))((...((((...(((((((((...)))))))))..))))...)).................................
5S rRNA, *E. coli*	ggaaaggaaagggaaagaaaUGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAUaaaacaaaacaaagaaacaacaacaacaac	.....................(((((((((.....((((((((....(((((((.............))))..)))...))))))).)).((.......((((((((...))))))))......))...)))))))))................................
Adenine riboswitch, *V. vulnificus* (*add*)	ggaaaggaaagggaaagaaaCGCUUCAUAUAAUCCUAAUGAUAUGGUUUGGGAGUUUCUACCAAGAGCCUUAAACUCUUGAUUAUGAAGUGaaaacaaaacaaagaaacaacaacaacaac	....................(((((((((...((((((.........))))))........((((((.......))))))..))))))))).............................
c-di-GMP riboswitch, *V. cholerae* (*VC1722)*	ggaaaaauGUCACGCACAGGGCAAACCAUUCGAAAGAGUGGGACGCAAAGCCUCCGGCCUAAACCAGAAGACAUGGUAGGUAGCGGGGUUACCGAUGGCAAAAUGcauacaaaccaaagaaacaacaacaacaac	..........((((......((...((((((....))))))...))...(((.((((((((..((.........)))))))..))))))...)).))....................................
Glycine riboswitch, *F. nucleatum*	ggacagagagGAUAUGAGGAGAGAUUUCAUUUUAAUGAAACACCGAAGAAGUAAAUCUUUCAGGUAAAAAGGACUCAUAUUGGACGAACCUCUGGAGAGCUUAUCUAAGAGAUAACACCGAAGGAGCAAAGCUAAUUUUAGCCUAAACUCUCAGGUAAAAGGACGGAGaaaacacaacaaagaaacaacaacaacaac	..........((((((((......((((((....)))))).(((....(((.....)))...)))........)))))))).........(((((......(((((.....))))).(((...((((.....(((....)))......))))..)))....)))))..............................

[d] Lowercase subsequences are buffer sequences and primer binding sites used for transcription and reverse transcription. These additional sequences do not interfere with the RNAs’ structures according to ViennaRNA and RNAstructure predictions.

[e] Structure is given in dot-bracket notation. A two-base pair helix for the adenine riboswitch that is not nested is not included in this dot-bracket representation.

Table S2: Base-pair-wise accuracy table for SHAPE and DMS

	Cryst.	no data		DMS		SHAPE		DMS + SHAPE	
		TP	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**
tRNA^{phe}	20	12	12	20	1	20	1	20	1
adenine rbsw.	21	15	10	21	2	21	2	21	2
cyclic di-GMP rbsw.	25	21	5	21	1	25	2	25	1
5S rRNA	34	9	31	32	6	32	6	32	6
P4-P6 RNA	48	44	9	45	6	44	6	44	5
glycine rbsw.	40	23	18	40	2	37	6	40	2
Total	188	124	85	179	18	179	23	182	17
	FNR	34%		4.8%		4.8%		3.2%	
	FDR	40.7%		9.1%		11.4%		8.5%	
	Sensitivity	66%		95.2%		95.2%		96.8%	
	PPV	59.3%		91%		88.6%		91.5%	

Table S3: Using only SHAPE reactivities in adenines and cytosines does not improve structure modeling – To test if the quality of the models given by DMS could be explained by selectively applying pseudo-energies only to adenines and cytosines, we re-ran the *Fold* program only with SHAPE reactivities that fell in adenines and cytosines. The resulting models have worse FDR and FNR than those derived from using DMS or full SHAPE data, confirming that the DMS results could not be explained by applying pseudo-energies to a subset of positions in the RNA. The reported accuracies are helix-wise.

	Cryst.	no data		SHAPE As and Us		SHAPE Gs and Us		SHAPE	
		TP	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**
tRNA^{phe}	4	2	3	4	0	4	0	4	0
adenine rbsw.	3	2	3	3	0	3	0	3	1
cyclic di-GMP rbsw.	8	6	2	5	2	6	1	8	0
5S rRNA	7	1	9	1	7	2	5	6	3
P4-P6 RNA	11	10	1	9	1	8	2	9	1
glycine rbsw.	9	5	3	8	1	8	1	8	1
Total	42	26	21	30	11	31	9	38	6
	FNR	38.1%		28.6%		26.2%		9.5%	
	FDR	44.7%		26.8%		22.5%		13.6%	
	Sensitivity	61.9%		71%		73.8%		90.5%	
	PPV	55.3%		73.2%		77.5%		86.4%	

Table S4: Helix-wise accuracies for inclusion of CMCT data in structure modeling.

	Cryst.	no data		CMCT		CMCT + DMS		CMCT + SHAPE		CMCT + DMS + SHAPE	
		TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
tRNA^{phe}	4	2	3	4	0	4	0	4	0	4	0
adenine rbsw.	3	2	3	3	1	3	1	3	1	3	1
cyclic di-GMP rbsw.	8	6	2	5	2	6	0	6	2	8	0
5S rRNA	7	1	9	6	3	6	3	6	3	6	3
P4-P6 RNA	11	10	1	10	1	10	1	9	1	9	1
glycine rbsw.	9	5	3	8	1	9	0	8	1	9	0
Total	42	26	21	36	8	38	5	36	8	39	5
	FNR	38.1%		14.3%		9.5%		14.3%		7.1%	
	FDR	44.7%		18.2%		11.6%		18.2%		11.4%	
	Sensitivity	61.9%		85.7%		90.5%		85.7%		92.9%	
	PPV	55.3%		81.8%		88.4%		81.8%		88.6%	

Figure S1. A probabilistic potential for pseudoenergy bonuses. (A) Normalized histograms for paired and unpaired reactivities (as defined by the crystallographic model) are retrieved from chemical mapping data. (B) Gamma mixture distributions with two components fitted to the data (dashed lines). The pseudo-energies are calculated as a function of the log-likelihood ratio of paired and unpaired distributions [$\Delta G_i = -k_\beta T$ log(P(S_i | i is paired)/P(S_i | i is unpaired)) for every S_i reactivity at nucleotide i]; error estimates are calculated by a smooth bootstrap procedure.

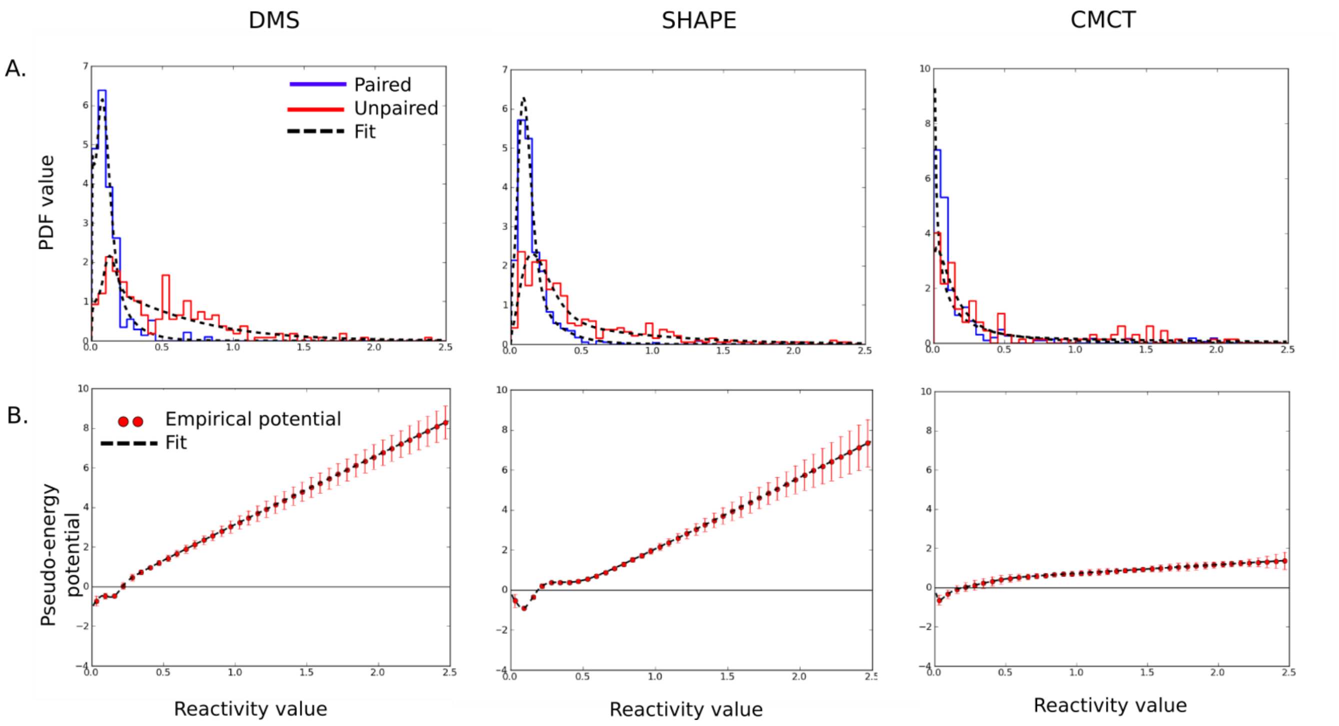


Figure S2. Predictive power of DMS, SHAPE, and CMCT. Receiver operating characteristic curves for predicting unpaired nucleotides given a reactivity threshold. Area under the curve for DMS is 0.86, for SHAPE, 0.83, and for CMCT, 0.74.

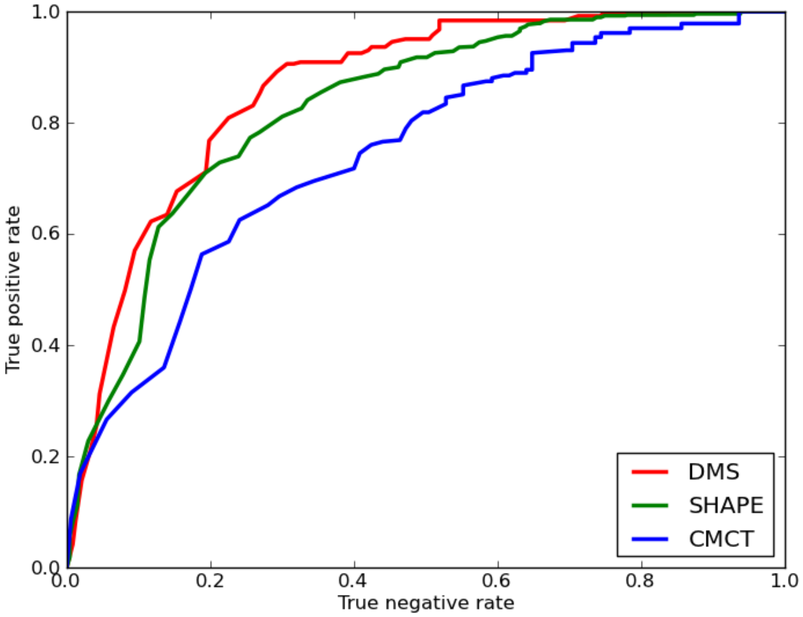


Figure S3. Helix-wise, bootsrap confidence value histograms for DMS, SHAPE, and CMCT models. – Blue histograms are for correctly predicted helices (true positives), red histograms are for incorrectly predicted helices (false positives).

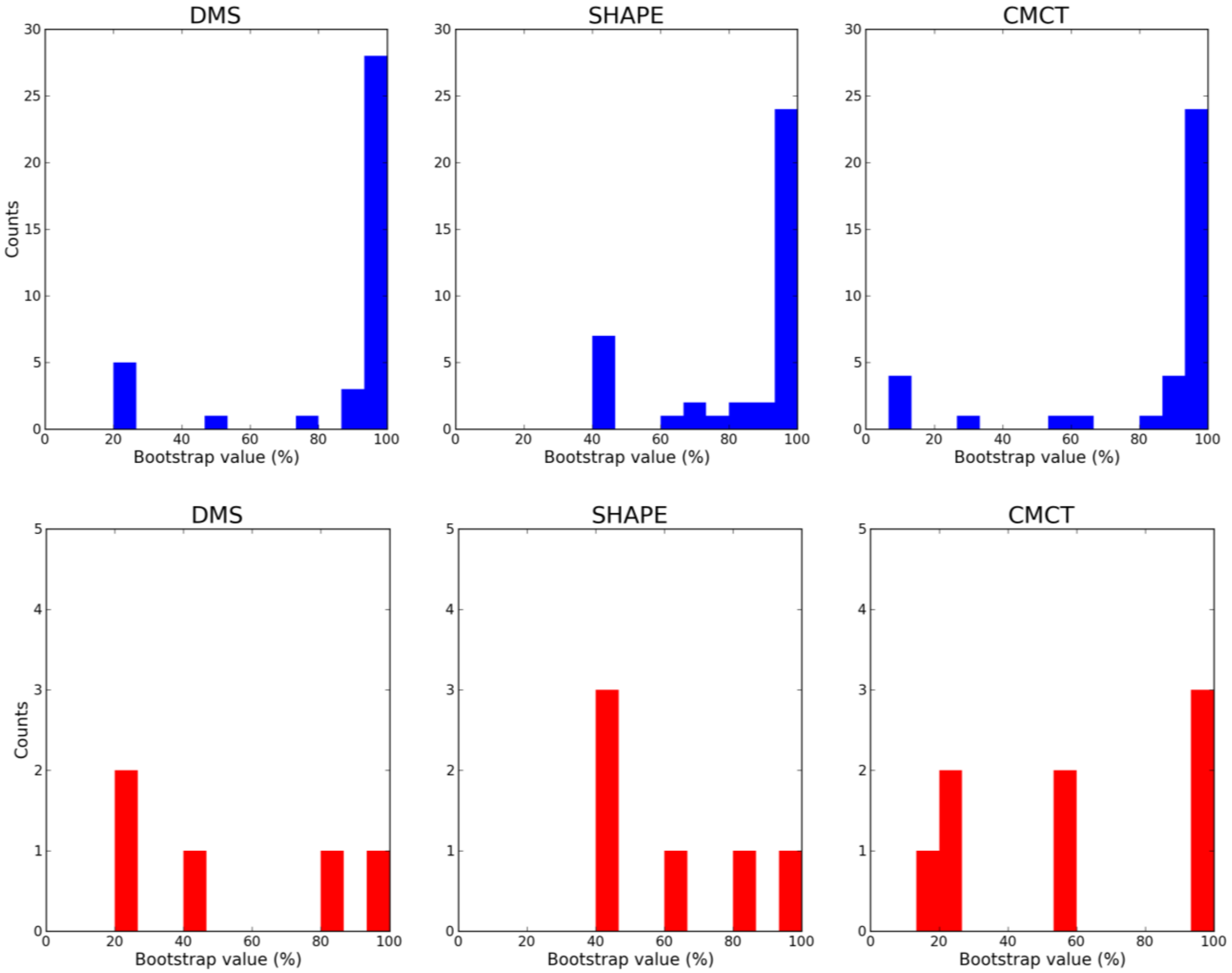


Figure S4. DMS, SHAPE, and CMCT data and pseudo-energy guided models for the *add* adenine riboswitch – DMS and CMCT data at guanines and uracils, and adenines and cytosines, respectively, are marked in gray. Missed base pairs are highlighted in blue lines; mis-predicted base pairs are indicated by orange lines. Helix bootstrap confidence values are shown in red. Secondary structure figures were prepared in VARNA (*8*).

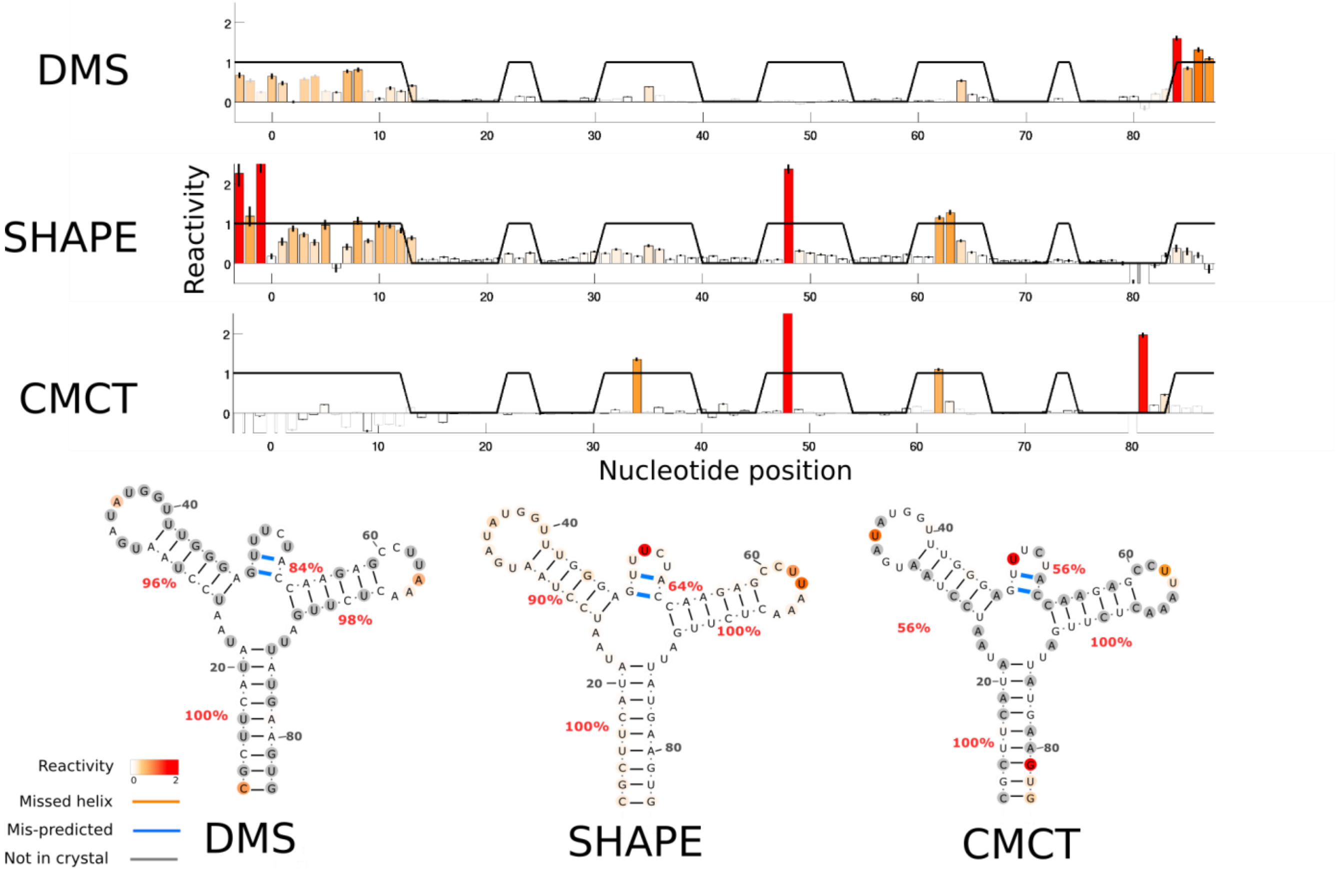


Figure S5. DMS, SHAPE, and CMCT data and pseudo-energy guided models for tRNA^{phe} – DMS and CMCT data at guanines and uracils, and adenines and cytosines, respectively, are marked in gray. Missed base pairs are highlighted in blue lines; mis-predicted base pairs are indicated by orange lines. Helix bootstrap confidence values are shown in red. Secondary structure figures were prepared in VARNA (*8*).

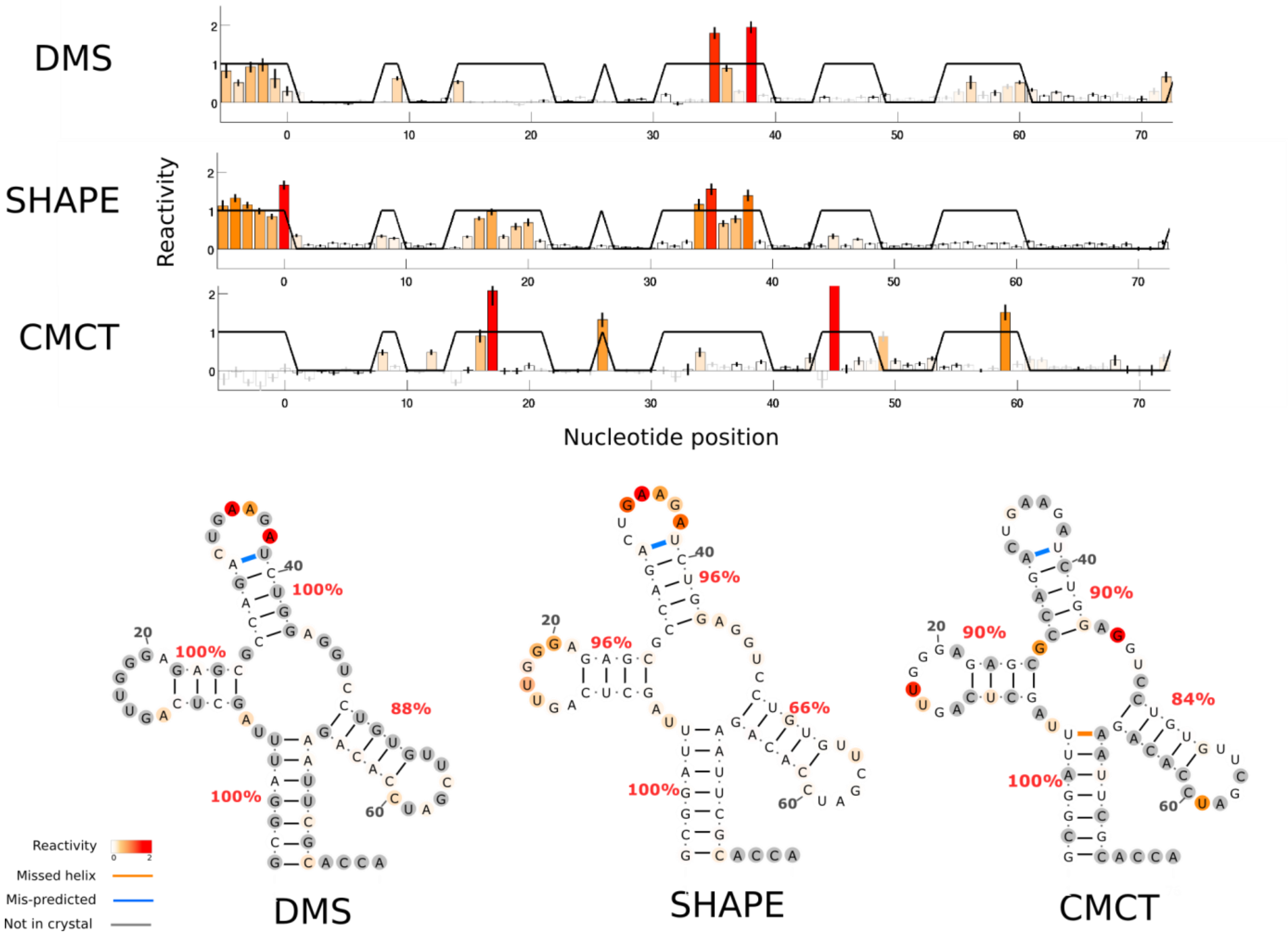


Figure S6. DMS, SHAPE, and CMCT data and pseudo-energy guided models for cyclic di-GMP riboswitch – DMS and CMCT data at guanines and uracils, and adenines and cytosines, respectively, are marked in gray. Missed base pairs are highlighted in blue lines; mis-predicted base pairs are indicated by orange lines. Helix bootstrap confidence values are shown in red. Secondary structure figures were prepared in VARNA (*8*).

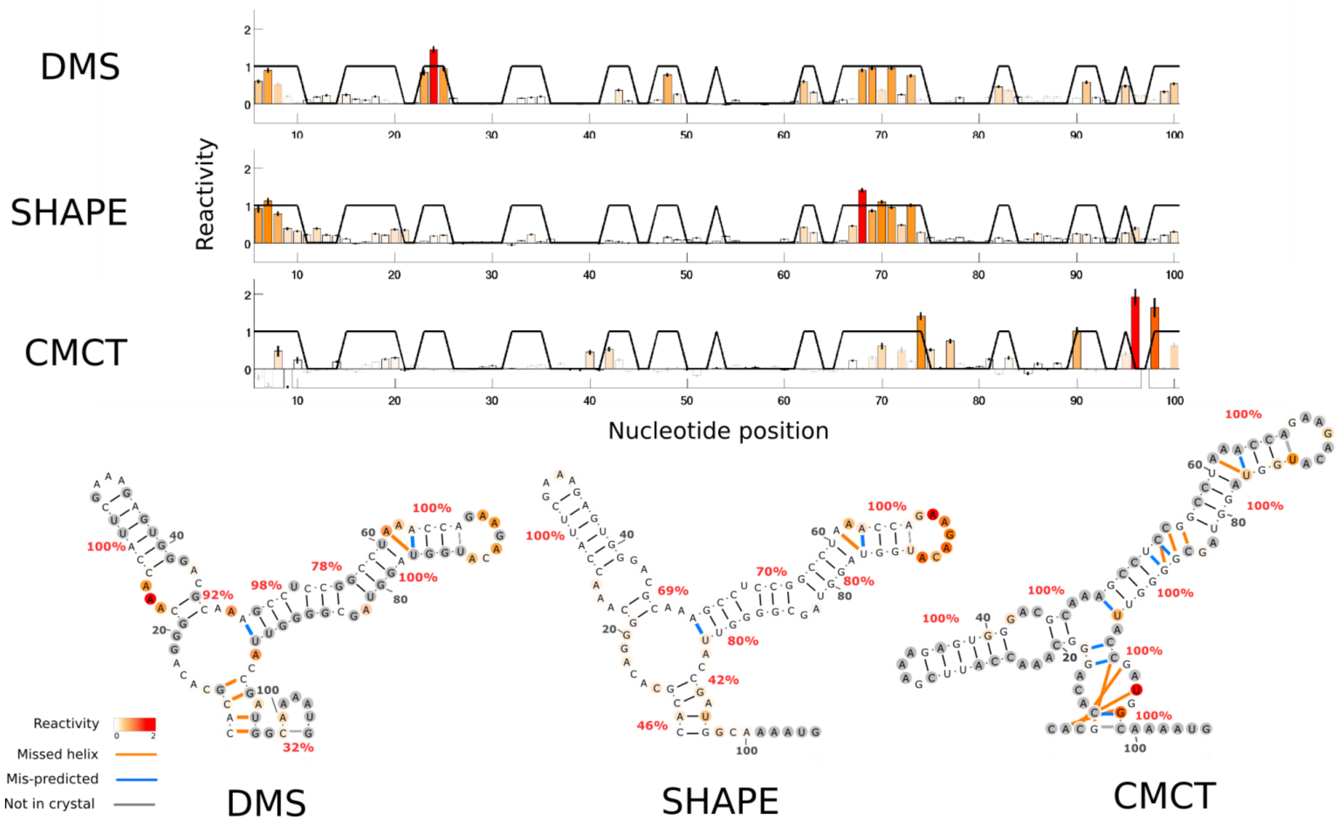


**Figure S7. DMS, SHAPE, and CMCT data and pseudo-energy guided models for the 5S rRNA – **DMS and CMCT data at guanines and uracils, and adenines and cytosines, respectively, are marked in gray. Missed base pairs are highlighted in blue lines; mis-predicted base pairs are indicated by orange lines. Helix bootstrap confidence values are shown in red. Secondary structure figures were prepared in VARNA (*8*).

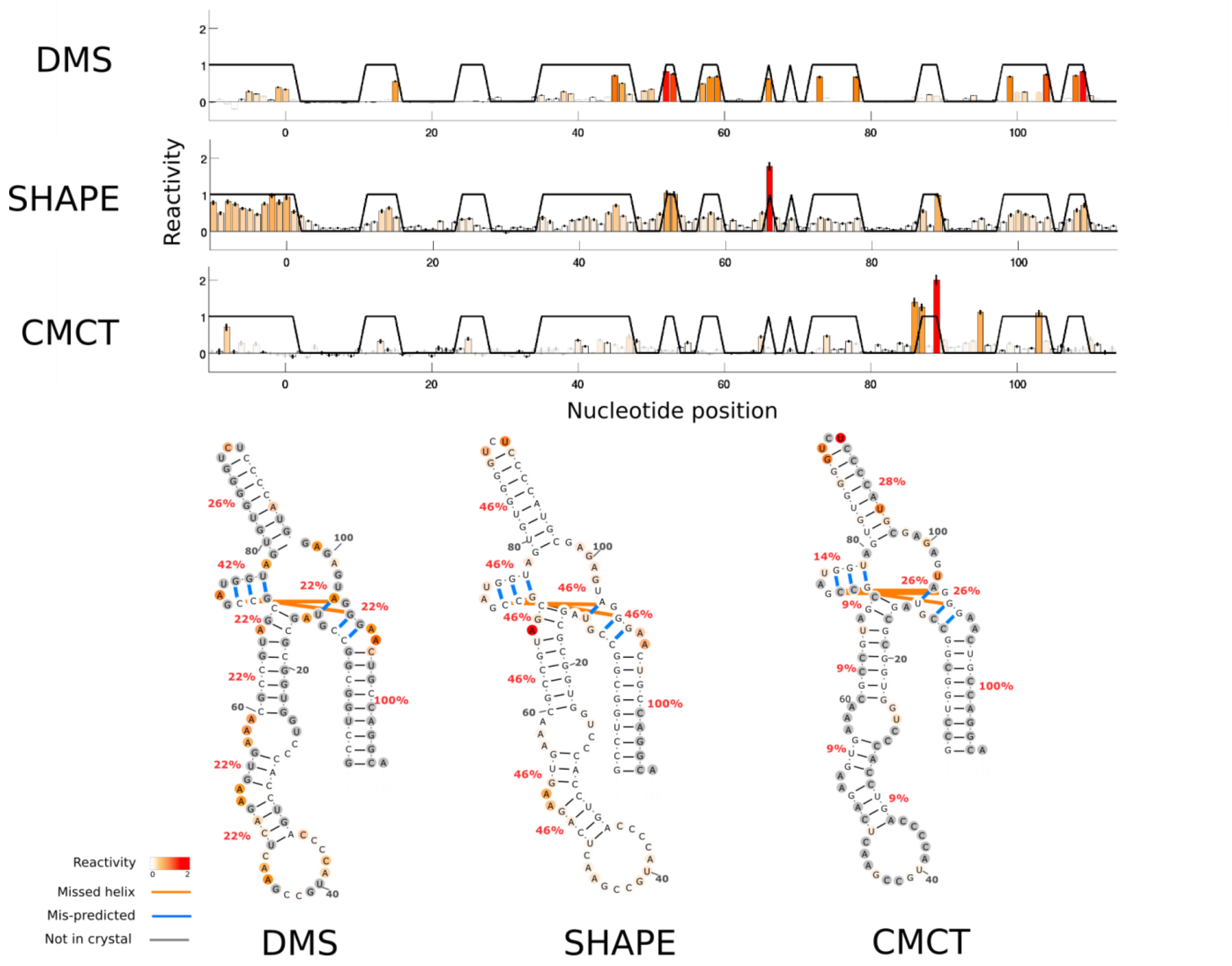


Figure S8. DMS, SHAPE, and CMCT data and pseudo-energy guided models for the P4-P6 domain of the *Tetrahymena* group I ribozyme– DMS and CMCT data at guanines and uracils, and adenines and cytosines, respectively, are marked in gray. Missed base pairs are highlighted in blue lines; mis-predicted base pairs are indicated by orange lines. Helix bootstrap confidence values are shown in red. Secondary structure figures were prepared in VARNA (*8*).

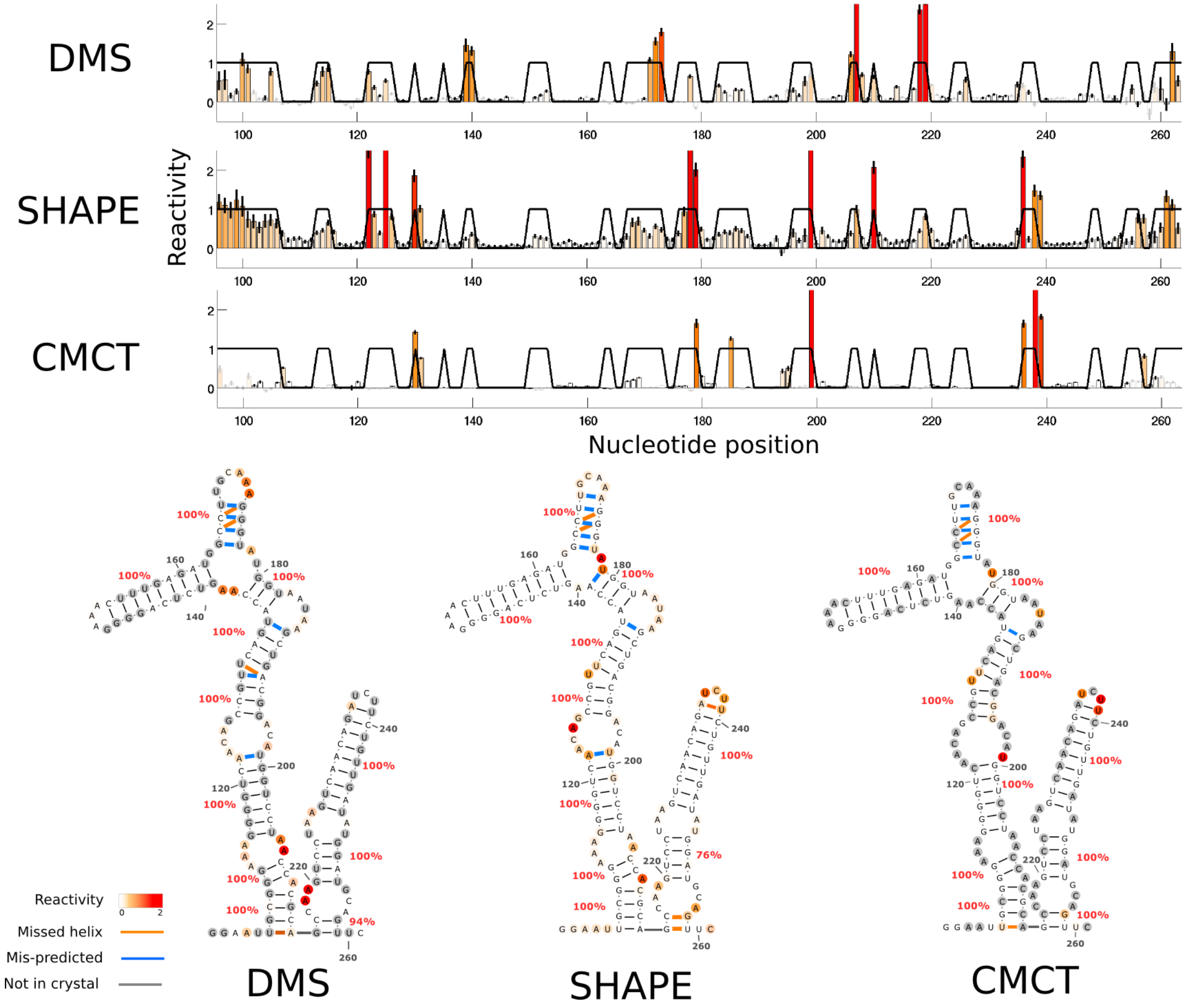
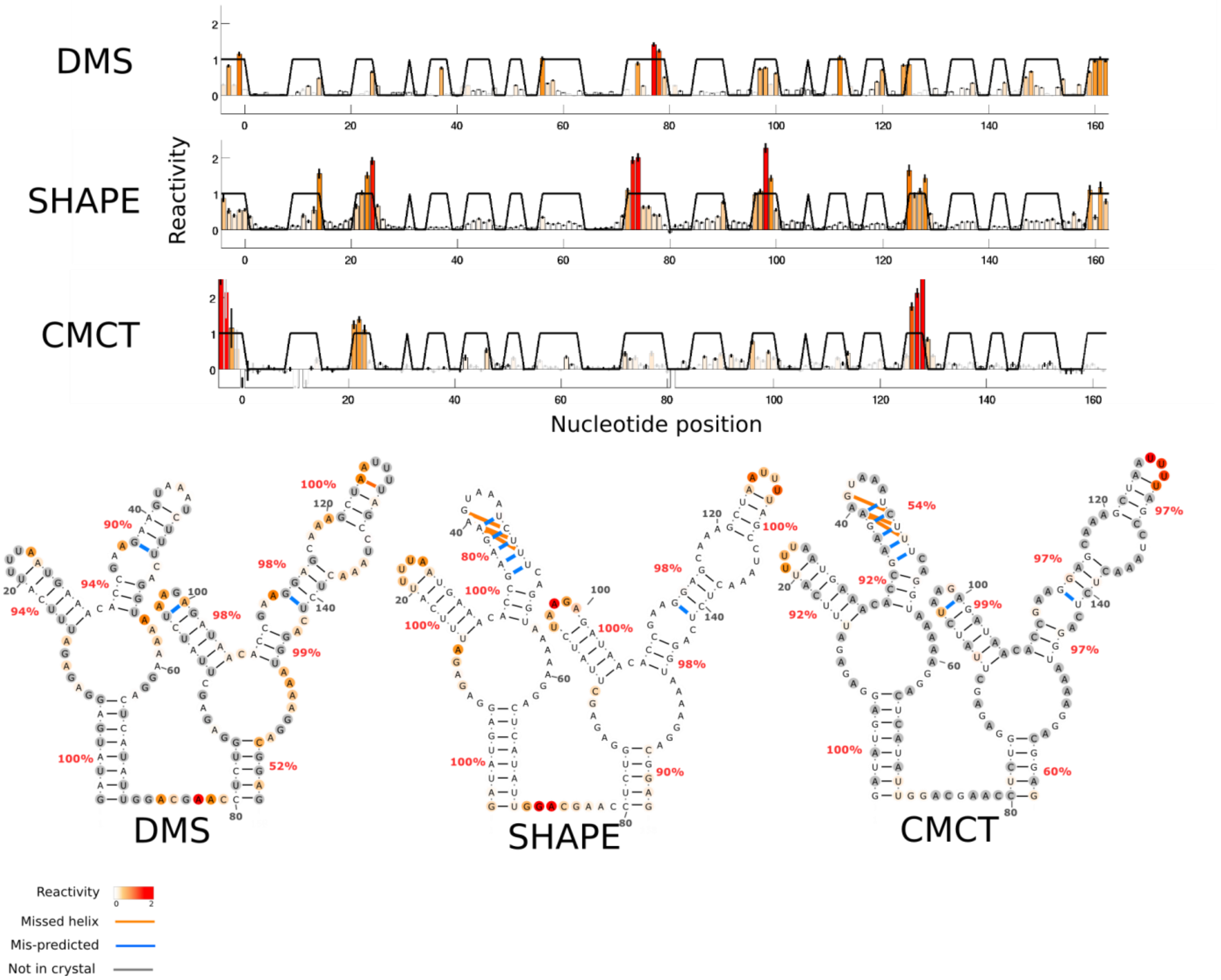


Figure S9. DMS, SHAPE, and CMCT data and pseudo-energy guided models for the *F. nucleatum* glycine riboswitch– DMS and CMCT data at guanines and uracils, and adenines and cytosines, respectively, are marked in gray. Missed base pairs are highlighted in blue lines; mis-predicted base pairs are indicated by orange lines. Helix bootstrap confidence values are shown in red. Secondary structure figures were prepared in VARNA (*8*).



References for supporting information

1. Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) , *Biochemistry 50*, 8049-56.

2. Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011) , *Bioinformatics 27*, 1798-1805.

3. Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009) , *Proc. Natl. Acad. Sci. U. S. A. 106*, 97-102.

4. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004) , *Methods in enzymology 383*, 66-93.

5. Rohl, C. A., Strauss, C. E. M., Chivian, D., and Baker, D. (2004) , *Proteins 55*, 656-77.

6. Das, R., and Baker, D. (2007) , *Proc. Natl. Acad. Sci. U. S. A. 104*, 14664-9.

7. Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) , *Nature chemistry 3*, 954-62.

8. Darty, K., Denise, A., and Ponty, Y. (2009) , *Bioinformatics (Oxford, England) 25*, 1974-5.