Understanding the Errors of SHAPE-Directed RNA Structure Modeling

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

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

Supporting Information

ABSTRACT: Single-nucleotide-resolution chemical mapping for structured RNA is being rapidly advanced by new chemistries, faster readouts, and coupling to computational algorithms. Recent tests have shown that selective 2′-hydroxyl acylation by primer extension (SHAPE) can give near-zero error rates (0–2%) in modeling the helices of RNA secondary structure. Here, we benchmark the method using six molecules for which crystallographic data are available: tRNA(phe) and 5S rRNA from Escherichia coli, the P4–P6 domain of the Tetrahymena group I ribozyme, and ligand-bound domains from riboswitches for adenine, cyclic di-GMP, and glycine. SHAPE-directed modeling of these highly structured RNAs gave an overall false negative rate (FNR) of 17% and a false discovery rate (FDR) of 21%, with at least one helix prediction error in five of the six cases. Extensive variations of data processing, normalization, and modeling parameters did not significantly mitigate modeling errors. Only one variation, filtering out data collected with deoxyinosine triphosphate during primer extension, gave a modest improvement (FNR = 12%, and FDR = 14%). The residual structure modeling errors are explained by the insufficient information content of these RNAs’ SHAPE data, as evaluated by a nonparametric bootstrapping analysis. Beyond these benchmark cases, bootstrapping suggests a low level of confidence (<50%) in the majority of helices in a previously proposed SHAPE-directed model for the HIV-1 RNA genome. Thus, SHAPE-directed RNA modeling is not always unambiguous, and helix-by-helix confidence estimates, as described herein, may be critical for interpreting results from this powerful methodology.

The continuing discoveries of new classes of RNA enzymes, switches, and ribonucleoprotein assemblies provide complex challenges for structural and mechanistic dissection (see, e.g., refs 1–4). While crystallographic, spectroscopic, and phylogenetic analyses have led to a deeper understanding of several key model systems, the throughput or applicability of these methods is limited, especially for noncoding RNAs that switch between multiple states in their functional cycles.5–8 In recent years, several laboratories have revisited a widely applicable chemical approach for attaining nucleotide-resolution RNA structural information, variously called “footprinting” or “chemical structure mapping”. Recent advances have included novel chemical modification strategies, faster data analysis software, accelerated readouts via capillary electrophoresis, and multiplexed purification by magnetic beads.9–14

Despite these advances, chemical mapping data are not expected to generally give structure models accurate at nucleotide resolution. To a first approximation, the protection of an RNA nucleotide from chemical modification indicates that it forms some interaction with a partner elsewhere in the system. However, these data by themselves do not provide enough information to define the interaction partner. Instead, the mapping data can be used to test, refine, or guide structure hypotheses derived from manual inspection or automated algorithms.15–17 The accuracy of this approach is necessarily limited by uncertainties in the modeling, including incomplete treatment of noncanonical base pairs, base–backbone interactions, and pseudoknotted folds,17 and imperfect correlations of chemical modification rates to structural features. Indeed, there are notable historical examples of chemical data giving misleading structural suggestions, including blind modeling work on tRNA18,19 and SS rRNA.20,21

It was therefore exciting when recent studies of 2′-OH acylation (the SHAPE method) coupled to the RNAstructure algorithm reported secondary structure inference with unprecedented sensitivity (98–100% helix recovery).17 The work acknowledged several uncertainties. Measurements were taken on ribosomal RNA without protein partners, which may not form the same structures as crystallized protein-bound complexes. For other test cases, the assumed experimental structures were derived from phylogenetic analysis (P546 domain from the bI3 group I intron), NMR data (HCV IRES), or crystals of constructs with modifications not present in the SHAPE-probed constructs (tRNAAsp). A “gold-standard” benchmark of SHAPE-directed secondary structure inference on RNAs with corresponding crystallographic models remains unavailable. We present herein...
SHAPE data, secondary structure inference, and analysis of systematic and statistical errors for six such RNAs containing a total of 661 nucleotides and 42 helices. Our results provide a rigorous appraisal of the strengths and limitations of this promising chemical–computational technology.

**EXPERIMENTAL PROCEDURES**

**Preparation of Model RNAs.** The DNA templates for each RNA (Table S1 of the Supporting Information) consisted of the 20-nucleotide T7 RNA polymerase promoter sequence (TTCTAATACGACTCACTATAG) followed by the desired sequence. Double-stranded templates were prepared by polymerase chain reaction assembly of DNA oligomers up to 60 nucleotides in length (IDT, Integrated DNA Technologies, Coralville, IA) with Phusion DNA polymerase (Finnzymes) and purified with AMPure magnetic beads (Agencourt, Beckman Coulter) following the manufacturer’s instructions. Sample concentrations were estimated on the basis of UV absorbance at 260 nm measured on Nanodrop 100 or 8000 spectrophotometers. Verification of template length was accomplished by electrophoresis of all samples and 10 and 20 bp ladder length standards (Fermentas) in 4% agarose gels (containing 0.5 mg/mL ethidium bromide) and 1× TBE (100 mM Tris, 83 mM boric acid, and 1 mM disodium EDTA).

In vitro RNA transcription reactions were conducted in 40 μL volumes with 10 pmol of DNA template, 20 units of T7 RNA polymerase (New England Biolabs), 40 mM Tris-HCl (pH 8.1), 25 mM MgCl2, 2 mM spermidine, ATP, CTP, GTP, and UTP (1 mM each), 4% polyethylene glycol 1200, and 0.01% Triton X-100. Reaction mixtures were incubated at 37 °C for 4 h and monitored by electrophoresis of all samples along with 100–1000-nucleotide RNA length standards (RiboRuler, Fermentas) in 4% agarose gels [1.1% formaldehyde; run in 1× TAE (40 mM Tris, 20 mM acetic acid, and 1 mM disodium EDTA)], stained with SYBR Green II RNA gel stain (Invitrogen) following the manufacturer’s instructions. RNA samples were purified with MagMax magnetic beads (Ambion), following the manufacturer’s instructions, and concentrations were measured by absorbance at 260 nm on Nanodrop 100 or 8000 spectrophotometers.

**Chemical Probing Measurements.** Chemical modification reaction mixtures consisted of 1.2 pmol of RNA in a volume of 20 μL with 50 mM Na-HEPES (pH 8.0) and 10 mM MgCl2 and/or ligand at the desired concentration (see Table S1 of the Supporting Information), and 5 μL of SHAPE modification reagent. The modification reagent was 24 mg/mL N-methylisatoic anhydride (NMIA) freshly dissolved in anhydrous dimethyl sulfoxide (DMSO). The reaction mixtures were incubated at 24 °C for 15–60 min, with shorter modification times for the longer RNAs to maintain overall modification rates of <30%. In control reactions (for background measurements), 5 μL of deionized water or DMSO was added instead of modification reagent, and the mixture was incubated for the same amount of time. For experiments testing DMSO effects, higher concentrations of NMIA in DMSO were prepared and 2 μL of the modification reagent was added to the 20 μL reaction mixture. Reactions were quenched with a premixed solution of 5 μL of 0.5 M Na-MES (pH 6.0), 3 μL of 5 M NaCl, 1.5 μL of oligo-dT beads [poly(A) purist (Ambion)], and 0.25 μL of 0.5 μM 5’-rhodamine-green-labeled primer (AAAAAAGTTGGTTGGTTTTTCTTT) complementary to the 3’ end of the RNAs (also used in our previous studies15,14), and 0.05 μL of a 0.5 μM Alexa-555-labeled oligonucleotide (used to verify normalization). 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 remaining on a 96-post magnetic stand. The magnetic bead mixtures were resuspended in 2.5 μL of deionized water.

The resulting mixtures of modified RNAs and primers bound to magnetic beads were reverse transcribed by the addition of a premixed 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, and dTTP, with either dGTP or dITP22), 0.25 μL of 0.1 M DTT, and 0.65 μL of water. The reaction mixtures (total volume of 5 μL) were incubated at 42 °C for 30 min. RNA was degraded by the addition of 5 μL of 0.4 M NaOH and incubation 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). Fluorescent DNA products were purified by magnetic bead separation, rinsed twice with 40 μL of 70% ethanol, and air-dried for 5 min. The reverse transcription products, along with magnetic beads, were resuspended in 10 μL of a solution containing 0.125 mM Na-EDTA (pH 8.0) and a Texas Red-labeled reference ladder (whose fluorescence is spectrally separated from the rhodamine green-labeled products). The products were separated by capillary electrophoresis on an ABI 3100 or ABI 3700 DNA sequencer. Reference ladders were created using an analogous protocol without chemical modification and the addition of, for example, 2,3-dideoxy-TTP in an amount equimolar with respect to the amount of dTTP in the reverse transcriptase reaction.

The HiTRACE software23,24 was used to analyze the electropherograms. Briefly, traces were aligned by automatically shifting and scaling the time coordinate, based on cross correlation of the Texas Red reference ladder loaded with all samples. Sequence assignments for bands, verified by comparison to sequencing ladders, permitted the automated peak fitting of the traces to Gaussians.

**Likelihood-Based Processing of SHAPE Data.** Quantified SHAPE data were corrected for attenuation of longer reverse transcriptase products due to chemical modification, normalized, and background-subtracted. Rather than using an approximate exponential correction and background scaling,25 we used a likelihood framework to determine the final, corrected SHAPE reactivities (see also ref 26). Furthermore, a likelihood-derived analysis was implemented to average replicate SHAPE data sets across several experiments. Both of these procedures are described in detail in the methods of the Supporting Information. The algorithms are available in the functions `overmod_and_background_correct`, `log_L_m`, and `get_average_standard_state.m` within the freely available HiTRACE software package.24 Final averaged data and errors have been made made publicly available in the Stanford RNA Mapping Database (http://rmdb.stanford.edu). The accession IDs are TRNAPH_SHP_0001, TRP4P6_SHP_0001, SSRRNA_SHP_0001, ADRSW_SHP_0001, CIDGMP_SHP_0001, and GLYCFN_SHP_0001.

**Computational Modeling.** The Fold executable of the RNAstructure package (version 5.3) was used to infer SHAPE-directed secondary structures. The entire RNA sequences (Table S1 of the Supporting Information), including added flanking sequences, were used for all calculations. The flag “−T 297.15” set the temperature to match our experimental conditions (24 °C). The flags “−sh”, “−sm”, and “−si” were used
to input the SHAPE data file, slope \( m \), and intercept \( b \). The latter parameters define the pseudoenergy formula \( \Delta G_i = m \log(S_i + 1) + b \), where \( S_i \) is the SHAPE reactivity. In the RNAstructure implementation, these pseudoenergies are applied to each nucleotide that forms an edge base pair and doubly applied to each nucleotide that forms an internal base pair. Boltzmann probability calculations used the partition executable with the same flags.

Nonparametric bootstrapping analysis was conducted as follows. Given normalized SHAPE data \( S_i \) for nucleotides \( i = 1, 2, \ldots, N \), a bootstrap replicate was generated by choosing \( N \) random indices \( i' \) from 1 to \( N \), with replacement \(^{27,28,50} \) (i.e., some nucleotide positions are not represented, and some are present in multiple copies; for the latter, SHAPE pseudoenergies were scaled proportionally). The resulting data sets \( S_i' \) contained the same number of data points and carried any systematic errors present in the original data set. Secondary structure models directed by these data were analyzed in MATLAB to assess the frequency of each base pair arising in the replicates; the maximum bootstrap value across the base pairs of each helix was taken as the bootstrap value for the helix. The bootstrapping analysis is being made available on an automated server at http://rmdb.stanford.edu/structureserver.

Additional calculations were conducted with the fold() routine of the ViennaRNA package (version 1.8.4, equivalent to the “RNAfold” command lines) \(^{29} \) extended to accept SHAPE data and calculate pseudoenergies with the same formula used in RNAstructure; calculations were facilitated through Python bindings available through the software’s convenient SWIG (Simplified Wrapper and Interface Generator) interface. Secondary structure figures were prepared with VARNA.\(^{30} \)

**Assessment of Accuracy.** A crystallographic helix was considered correctly recovered if more than 50% of its base pairs were observed in a helix by the computational model. (In practice, 34 of 35 such helices retained all crystallographic base pairs.) Note that, unlike prior work, helix slips of \( \pm 1 \) were not considered correct [i.e., the pairing \((ij)\) was not allowed to match the pairing \((ij-1)\) or \((ij+1)\)].

### RESULTS

**Accuracy of Modeling without Experimental Data.** The benchmark herein (Table S1 of Supporting Information) collects a diverse set of noncoding RNA domains, containing two classic RNA folding model systems, unmodified tRNA\(^{\text{K125e}}\) from *Escherichia coli*\(^{33} \) and the P4–P6 domain of the *Tetrahymena* group I ribozyme;\(^{32} \) a functional RNA that has been a frequent test case for modeling algorithms, *E. coli* 5S rRNA;\(^{15,16,20,21} \) and three ligand-bound domains from bacterial riboswitches for adenine, cyclic di-GMP, and glycine;\(^{33–39} \) For the last RNA (glycine riboswitch from *Fusobacterium nucleatum*), crystallographic data were not available at the time of modeling but released at the time of submission of the manuscript; it served as a blind test within our benchmark.

As a control, we first applied the RNAstructure\(^{15,16} \) algorithm *Fold* without any experimental data to the benchmark set (Figure S1 of the Supporting Information). Here and below, we discuss modeling errors in terms of the false negative rate (FNR; fraction of crystallographic helices that were missed) and false discovery rate (FDR; fraction of predicted helices that were incorrect). The values are summarized, along with the related statistics of sensitivity and positive predicted value, in Table 1. To highlight features of the RNAs’ global folds, we present results in terms of helices rather than individual base pairs. For the sake of completeness, FNR, FDR, sensitivity, and positive predictive values at the base pair level are also compiled in Table S2 of the Supporting Information.

Without any data, the RNAstructure algorithm missed 16 of 42 helices, giving an FNR of 16/42 (38%). The models mispredicted an additional 21 helices, giving an FDR of 21/42 (5%) (Table 1). These error rates are significantly worse than their ideal values (0%) and confirm the known inaccuracy of current secondary structure prediction methods without experimental guidance (see, e.g., refs 16 and 17).

**Accuracy of Modeling with SHAPE Data.** We then acquired SHAPE data for each RNA in 50 mM Na-HEPES (pH 8.0), 10 mM MgCl\(_2\), and saturating concentrations of ligand (for the three riboswitch domains), using the modification reagent N-methylisatoic anhydride (NMIA). Quantitation of data for each RNA involved correction for attenuation of long products, background subtraction, and averaging of 12–28 replicates (Table S1 of the Supporting Information) guided by a likelihood framework (methods). The data were in excellent agreement with the expected structures [Figures 1 and 2 (left panels)]. Strong SHAPE reactivities occur mainly at nucleotides that are outside Watson–Crick helices observed in crystallographic models. On the basis of prior work,\(^{17} \) we expected that inclusion of these data as a pseudoenergy term in the RNAstructure algorithm would substantially improve the accuracy of computational models, with a helix-level FNR as low as 0–2%. The improvement was indeed significant, but not to the extent expected (Figure 2, right panels; Table 1). The FNR decreased from 38 to 17% (missing 7 of 42 helices), and the FDR decreased from 43 to 21% (misprediction of 9 helices). In five of the six RNAs, the calculations failed to recover all the crystallographic helices.

**Evaluating Sources of Systematic Error.** The results described above give a somewhat less optimistic picture of SHAPE-directed modeling than previously published measurements.\(^{17} \)

---

### Table 1. Accuracy of Secondary Structure Recovery by RNAstructure with and without SHAPE Data

<table>
<thead>
<tr>
<th>RNA</th>
<th>no. of nucleotides</th>
<th>Cryst</th>
<th>TP</th>
<th>FP</th>
<th>TP</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNAstructure with SHAPE</td>
<td>no. of helices</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tRNA^{\text{K125e}}</td>
<td>76</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P4–P6 RNA</td>
<td>158</td>
<td>11</td>
<td>10</td>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>5S rRNA</td>
<td>118</td>
<td>7</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>adenine riboswitch</td>
<td>71</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cyclic di-GMP ribosw.</td>
<td>80</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>glycine riboswitch</td>
<td>158</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>total</td>
<td>661</td>
<td>42</td>
<td>26</td>
<td>21</td>
<td>35</td>
<td>9</td>
</tr>
</tbody>
</table>

False negative rate = 38.1% 16.7%
False discovery rate = 44.7% 20.5%
Sensitivity = 61.9% 83.3%
Positive predictive value = 55.3% 79.5%

Abbreviations: Cryst, number of helices in the crystallographic model; TP, true positives; FP, false positives. \( a \) False negative rate = 1 – TP/Cryst. \( b \) False discovery rate = FP/(TP + FP). \( c \) Sensitivity = (1 − false negative rate) = TP/Cryst. \( d \) Positive predictive value = (1 − false discovery rate) = TP/(TP + FP).
The differences between SHAPE benchmarks can be most simply ascribed to different test RNAs. Nevertheless, we investigated several other possible systematic explanations for the error rates (FNR and FDR of 17 and 21%, respectively) in our test set. First, we used herein an evaluation scheme to define helix recovery more stringent than those used in previous work,

which permitted helix register slips of ±1 (see Experimental Procedures). Using those less stringent criteria gave similar FNR and FDR values of 14 and 18%, respectively. Second, we checked for experimental artifacts. Filtering out nucleotides whose SHAPE pseudoenergy errors exceeded 0.4 kcal/mol gave similar FNR and FDR values [14 and 18%, respectively (Table 2)]. Third, to test the quality of our lab’s experimental procedures and data processing, we acquired SHAPE measurements on an RNA with a previously published SHAPE-directed model, the hepatitis C virus internal ribosomal entry site domain II. Using those less stringent criteria gave similar FNR and FDR values of 14 and 18%, respectively. Fourth, we acquired SHAPE data for all the RNAs with the newly developed 2′-OH acylating reagent 1-methyl-7-nitroisatoic anhydride (1M7),

the FNR and FDR for models based on these data were identical to the measurements with the more widely used NMIA (Table 2).

Fifth, model accuracy might be unduly sensitive to the highest or lowest reactivities in the SHAPE data. However, capping “outliers” (see the methods of the Supporting Information); changing the cutoffs for capping; removing outliers; only including high-reactivity data; and excluding SHAPE data for nucleotides near the 5′ and 3′ ends of the RNA did not improve the accuracy (Table 2). Seventh, the pseudoenergy for base pairing is derived from SHAPE data by a logarithmic formula

Optimizing parameters 
m and 
b did not affect the FNR and improved the FDR only slightly [from 21 to 18% (Table 2)]. Eighth, choices in normalizing SHAPE data can affect the modeling, but varying the normalization by factors between 0.5- and 2-fold did not significantly improve the accuracy (Table 2). Ninth, we explored whether energy inaccuracies stem from RNAstructure’s thermodynamic parameters, SHAPE data, or both. Comparing energies of crystallographic and model structures indicated that both thermodynamic and SHAPE energies are imbalanced to favor incorrect models [by averages of 1.7 and 1.3 kcal/mol, respectively (Table S3 of the Supporting Information)]. Additionally, shifting the Boltzmann weight balances by increasing the modeling temperature from 24 to 37 °C did not change the error rates (Table 2). Tenth, we additionally tested for algorithm biases by recomputing models in ViennaRNA

rather than RNAstructure, but overall, the FNR and FDR both increased [to 26 and 28%, respectively (Table 2)].

Evidence against Crystal versus Solution Structure Discrepancies. Having found no straightforward explanation for SHAPE-directed modeling errors from systematic errors in experimental data acquisition, data processing, or modeling protocols, we investigated whether there might be differences between these RNAs’ secondary structures in available crystals and under our experimental solution conditions, as occurred in prior work with extracted rRNA.

Several lines of evidence disfavor this hypothesis in our cases. For tRNA(phe), the P4−P6 domain, the 5S rRNA, and the purine and cyclic di-GMP riboswitch, independent crystallographic models of several variants indicate that the RNAs’ secondary structures agree with phylogenetic analysis and are furthermore robust to different conditions, binding partners, and crystallographic contexts (Table S1 of the Supporting Information). In addition, while flanking sequences added to constructs (Table S1 of the Supporting Information) might disrupt the target domains, we designed these sequences to prevent such pairings and checked this lack of pairings by calculations with and without SHAPE data (Figure S1 of the Supporting Information and Figure 2).

Misfolding to kinetically trapped secondary or tertiary structures could lead to differences in solution chemical mapping data compared to those expected from crystallographic structures. To test this possibility, we acquired data for the RNAs after incubating them in 10 mM Na-MES (pH 6.0) and 10 mM MgCl₂ for 30 min at 50 °C (refolding conditions developed for large ribosomes)

the resulting reactivities were indistinguishable from those of RNAs without the refolding treatment (see, e.g., Figure S3 of the Supporting Information for tRNA data). Similarly, we tested for adverse effects of dimethyl sulfoxide (DMSO), used to solubilize the
SHAPE reagent) by repeating measurements at lower DMSO concentrations (10% vs 25%); SHAPE data were indistinguishable under the two conditions (Figure S3 of the Supporting Information gives tRNA data).

In addition to these results disfavoring differences in crystal and solution structures, our solution measurements gave positive evidence of the RNAs folding into the correct tertiary conformations. The P4–P6 domain and the 5S rRNA gave changes in their metal core and loop E regions, respectively, upon addition of Mg²⁺, as expected from prior biophysical analysis (e.g., refs 45–48). The three riboswitches gave SHAPE changes with and without their ligands (Figure S4 of the Supporting Information). Most strongly, we have subjected each of these RNAs to the mutate-and-map method, a two-dimensional extension of chemical mapping, and observed near-complete recovery of the crystallographic helices (98% sensitivity), indicating that the dominant solution structure matches the structure determined by crystallography.

Assessing Information Content and Confidence by Bootstrapping. A final explanation for the errors of SHAPE-directed structure models could be that the experimental data have insufficient information content to define the secondary structure. That is, the data, while accurately reflecting each RNA’s solution conformation, are also consistent with incorrect secondary structures with similar calculated energy. Indeed, the minimum energy model can be highly sensitive to small changes in the SHAPE data (see the tRNA example in Figure S5a–c of the Supporting Information), and in some cases, the incorrect lowest-energy SHAPE-directed model is within 1 kcal/mol of the crystallographic model [see tRNA³leu and the cyclic di-GMP riboswitch (Table S3 of the Supporting Information)]. Unfortunately, quantitatively interpreting energy differences between models [as well as partition function-based base pair probabilities, which are skewed to high values (see Figure S5a of the Supporting Information)] is currently complicated by the nonphysical nature of the SHAPE pseudoenergies. For example, a useful confidence value should be a good approximation of the actual modeling accuracy. In contrast, the mean base pair probability value over all predicted helices is 88%, suggesting a false discovery rate of 12% (100% − 88%), underestimating the actual error rate of 21%.

Figure 2. Crystallographic (left) and SHAPE-directed (right) secondary structure models for a benchmark of noncoding RNAs. SHAPE reactivities are shown as colors on bases and match colors in Figure 1. Cyan lines mark incorrect base pairs. Orange lines mark crystallographic base pairs missing in each model. Gray lines mark base pairs in regions outside the crystallized construct. Helix confidence estimates from bootstrap analyses are given as red percentages. For the sake of clarity, flanking sequences (see Table S1 of the Supporting Information) are not shown.
Table 2. Effects of Variations of Data Processing or Modeling on the Accuracy of SHAPE-Directed Secondary Structure Modeling

<table>
<thead>
<tr>
<th>variation in modeling(\mathrm{a})</th>
<th>TP(\mathrm{b})</th>
<th>FP(\mathrm{b})</th>
<th>FNR (%)</th>
<th>FDR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no SHAPE data (control)</td>
<td>26 21</td>
<td>38.1 44.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAPE-directed, default parameters</td>
<td>35 9</td>
<td>16.7 20.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>remove residues with large errors(\mathrm{c})</td>
<td>36 8</td>
<td>14.3 18.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>use only data collected with dITP during primer extension</td>
<td>31 12</td>
<td>26.2 27.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>use only data collected with dGTP during primer extension</td>
<td>37 6</td>
<td>11.9 14.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>use 1M7 instead of NMIA reagent</td>
<td>35 9</td>
<td>16.7 20.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cap outliers(\mathrm{d}) at cutoff value</td>
<td>35 9</td>
<td>16.7 20.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cap outliers(\mathrm{d}) at 2.0</td>
<td>35 9</td>
<td>16.7 20.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>remove five additional residues from 5' and 3' ends</td>
<td>35 9</td>
<td>16.7 20.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>remove residues with SHAPE &lt; 0.5</td>
<td>32 14</td>
<td>23.8 30.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>optimized m and b in pseudoenergy relation(\mathrm{e})</td>
<td>35 8</td>
<td>16.7 18.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adjust normalization 2-fold</td>
<td>35 8</td>
<td>16.7 18.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adjust normalization 1.5-fold</td>
<td>35 9</td>
<td>16.7 20.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adjust normalization 0.75-fold</td>
<td>34 12</td>
<td>19.0 26.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adjust normalization 0.5-fold</td>
<td>31 14</td>
<td>26.2 31.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNAstructure (T = 37^\circ\mathrm{C}) (not 24 \time{}\mathrm{C})</td>
<td>35 9</td>
<td>16.7 20.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViennaRNA(\mathrm{f}) instead of RNAstructure</td>
<td>32 10</td>
<td>23.8 23.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\mathrm{a}\)All variations are described relative to “default conditions” (bold) using RNAstructure version 5.3. \(\mathrm{b}\)The total number of crystallographic helices is 42. TP, true positives; FP, false positives. \(\mathrm{c}\)Any residues whose estimated measurement error of SHAPE reactivity would give errors of more than ±0.4 kcal/mol if included in a base pair, using the SHAPE pseudoenergy relation. \(\mathrm{d}\)Outliers were defined as in the normalization procedure: those with values above a cutoff equal to 1.5 times the interquartile range. \(\mathrm{e}\)Pseudoenergy applied to base-paired nucleotides given by \(m \log(1.0 + \text{SHAPE}) + b\). Default parameters in RNAstructure are as follows: \(m = 2.6\), and \(b = -0.8\). The combinations of \(m\) and \(b\) that gave optimal accuracies for this benchmark were 3.0 and \(-0.6\), respectively. \(\mathrm{f}\)ViennaRNA version 1.8.4, using the default parameter set of ref 15.

We therefore estimated the helix-by-helix confidence of SHAPE-directed models through a nonparametric bootstrapping procedure, inspired by techniques developed to evaluate phylogenetic trees from multiple-sequence alignments.\(\mathrm{27,28,50}\) We generated 400 mock replicates of each data set by resampling with replacement the SHAPE data for individual residues, generating secondary structure models directed by these mock data sets, and evaluating the frequency with which each predicted helix appeared in these replicates (Figure S5b of the Supporting Information and percentage values in Figure 2). One-fourth of the modeled helices (11 of 44) appeared with bootstrap values of <55%, suggesting insufficient information for confident determination of their structure; 7 of these 11 helices were indeed incorrect. Encouragingly, the 33 helices with bootstrap values of >55% included only two errors, of which one was a single-nucleotide register shift. Further, these bootstrap values are robust to small changes in the SHAPE data (see the tRNA example in Figure S6d,e of the Supporting Information). Finally, the overall mean of the helix bootstrap values was 77%. This result predicts a false discovery rate of 23% (100% − 77%), in accord with the actual rate of 21%. Bootstrap analysis therefore appears to be well-suited for evaluating confidence in SHAPE-directed models.

**Bootstrap Analysis of an Independent Test Case: The HIV-1 Genome Model.** As a final demonstration of the utility of bootstrapping confidence estimation, we investigated the information content of an external data set. Recent application of the SHAPE method to the 9173-nucleotide RNA genome extracted from the NL4-3 HIV-1 virion gave a secondary structure hypothesis containing 429 helices,\(\mathrm{51}\) and the quantitated SHAPE reactivity data have been published. Employing these data and previously used modeling constraints (including division of the modeled genome into five separate domains), the current version of RNAstructure (version 5.3) largely recovers the prior working model (Table S4 of the Supporting Information). Furthermore, bootstrapping revealed additional useful information. The 57-nucleotide S’ TAR element, two helices with lengths greater than 10 bp in the gag-pol region, and the signal-peptide stem at the S’ end of gp120 give bootstrap values of >95%. These regions are thus high-confidence features of the SHAPE-directed model. Overall, however, 236 of 429 helices in the model have bootstrap confidence estimates of <50%. (If base pairs across the five assumed domains are permitted, more helices are found with such low bootstrap values.) The bootstrap value averaged over all predicted helices is 49%; excluding 59 stems in the prior model that are not recovered with the current version of RNAstructure gives a similar value of 55%. These results suggest that much of the HIV-1 secondary structure remains uncertain, even in regions that are strongly protected from SHAPE modification (Figure S7 of the Supporting Information). A tabulation of the helix-by-helix confidence estimates in Table S4 of the Supporting Information should help guide further dissection of these uncertain regions by other chemical and structural approaches.

**DISCUSSION**

With recent experimental and computational acceleration, nucleotide-resolution chemical mapping permits the characterization of noncoding RNAs at an unprecedented rate. Nevertheless, the resulting data are not always sufficient to determine the molecule’s secondary structure, especially if additional tertiary interactions are present. The helix-level error rates found in this study of six highly structured RNAs (false negative rate and false discovery rate of 17 and 20%, respectively) are significantly better than those of models generated without data (38 and 45%, respectively) but higher than those for prior SHAPE modeling test cases (FNR of 0–2%). The modeling inaccuracy found herein is similar to error rates (FNR of ~24%) found in benchmarks with other chemical modifiers, including dimethyl sulfate, kethoxal, and carboxamide,\(\mathrm{16}\) albeit on different RNAs and with different modeling protocols. Side-by-side tests on the same model RNAs will be necessary to rigorously compare conventional chemical approaches with SHAPE-based methods.

As with all structure characterization methods, SHAPE-directed models cannot be considered “determined structures” but instead are useful hypotheses, especially if accompanied by confidence estimates. This work proposes a bootstrapping analysis for SHAPE-directed modeling that provides such confidence values for novel RNAs. In addition to giving correct predictions for helix accuracy in six crystallized RNAs, bootstrapping analysis of the HIV-1 RNA genome finds numerous regions with high degrees of uncertainty in the RNA’s current SHAPE-directed working model. More information-rich multidimensional methods, such as NMR and the mutate-and-map chemical
approaches,\textsuperscript{13,14} should be able to test these predictions and, more generally, help attain accurate models of noncoding RNAs.

\section*{ASSOCIATED CONTENT}

\subsection*{Supporting Information}

Methods for likelihood-based data processing, four tables with detailed benchmark information and systematic error analyses, and eight supporting figures. This material is available free of charge via the Internet at http://pubs.acs.org.

\section*{AUTHOR INFORMATION}

\subsection*{Corresponding Author}

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

\subsection*{ACKNOWLEDGMENTS}

We thank authors of RNAstructure and ViennaRNA for making the source code freely available; M. Elazar and J. Glenn for the gift of 1M7; and J. Lucks, D. Mathews, K. Weeks, and the Das lab for comments on the manuscript.

\section*{REFERENCES}


Supporting Information for: Understanding the errors of SHAPE-directed RNA structure modeling

Wipapat Kladwang¹, Christopher C. VanLang², Pablo Cordero³, and Rhiju Das¹,³,⁴*

Departments of Biochemistry¹, Chemical Engineering², Biomedical Informatics³, and Physics⁴, Stanford University, Stanford CA 94305
* To whom correspondence should be addressed: rhiju@stanford.edu. Phone: (650) 723-5976. Fax: (650) 723-6783.

This document contains the following Supporting Information:

Supporting Methods.

Table S1. Benchmark for SHAPE-directed secondary structure modeling.

Table S2. Base-pair-level statistics of secondary structure recovery by RNAstructure with and without SHAPE data.

Table S3. Sources of poor discrimination of correct from incorrect secondary structures.

Table S4. Helix-by-helix bootstrap confidence estimates for the SHAPE-directed model of the HIV-1 RNA genome.

Figure S1. RNAstructure secondary structure models for a benchmark of six structured RNAs.

Figure S2. Recovery of SHAPE-directed model for a previously studied HCV RNA.

Figure S3. SHAPE data acquired with different dNTP mix for primer extension, refolding prior to chemical modification, and different DMSO backgrounds.

Figure S4. Demonstration that solution SHAPE data reflect folded or ligand-bound conformations.

Figure S5. Partition-function and bootstrap analysis of SHAPE-directed secondary structure models.

Figure S6. Sensitivity of minimum-energy model and robustness of bootstrap values to small changes in tRNA SHAPE data.

Figure S7. HIV-1 secondary structure helix confidence values compared to SHAPE reactivities.

Figure S8. Histogram and fit of SHAPE reactivities.

References for Supporting Information
Supporting Methods

Likelihood-based attenuation correction, background subtraction, and normalization of SHAPE data

Quantified SHAPE data were corrected for attenuation of longer reverse transcriptase products due to chemical modification, normalized, and background-subtracted. This procedure involved optimization carried out over three parameters, the overall modification rate $\gamma$, a normalization factor for the data $\alpha$, and a scaling for the background $\beta$, defined and estimated as follows. First, let $y_i$ be the observed fraction of products stopping at each nucleotide $i = 1, 2, \ldots N$ in the reverse transcription reaction, ordered so that longer (attenuated) products have larger indices $i$; and define

$$y_{N+1} = 1 - \sum_{i=1}^{N} y_i$$

be the fraction of products that are fully extended. These probabilities are equal to the underlying stopping probabilities $p_i$ times the product of probabilities that the reverse transcriptase has not stopped earlier:

$$y_i = p_i \prod_{j=1}^{i-1} (1 - p_j), \quad (1)$$

which can be inverted by recursively calculating $p_1 = y_1$, then $p_2 = y_2(1-p_1)^{-1}$, then $p_3 = y_3(1-p_1)^{-1}(1-p_2)^{-1}$, etc. For convenience, define $c_i = p_i / y_i$ as an attenuation correction factor.

[In the limit that the $y_i$ are approximately constant and much smaller than 1, the solution reduces to an exponential fall-off $c_i^{-1} = \exp[-<y_i>i]$, as was effectively assumed in (1, 2), but this approximation is unnecessary.]

Fluorescence measurements of reverse transcription products from capillary electrophoresis detectors are in arbitrary units; evaluating (1) requires that the $y_i$ be
properly normalized, and this is in principle achieved by the constraint that
\[ \sum_{i=1}^{N} y_i + y_{N+1} = 1. \]
However, in practice, the intensity of the longest products (e.g., \( y_{N+1} \))
cannot be measured precisely. The strong fluorescence of these fully extended products
typically saturates the experimental detector. Thus, the values \( y_i \) are defined only up to a
proportionality constant \( \gamma \), i.e., \( y_i = \gamma s_i / S \), where \( s_i \) are the observed fluorescence
intensities and \( S = \sum_{j=1}^{N_{\text{discarded}}} s_j \); we need to select amongst \( \gamma < 1 \).

We estimate \( \gamma \) at the same time as the two other unknown proportionality constants in
the data normalization and background subtraction. Let \( b_i \) be the quantified band
intensities from control measurements (no SHAPE reaction), which we assume require
negligible correction from attenuation. The background-subtracted SHAPE reactivities
are given in terms of unknown constants \( \alpha \) and \( \beta \) by:
\[
 s_i^{\text{correct}} = \alpha |p_i - \beta b_i| = \alpha |c_i(\gamma) s_i - \beta b_i| 
\]
(2)

We then optimized the log-likelihood function:
\[
 L = \prod_{i=1}^{n} \alpha c_i(\gamma) \exp[F(\alpha (c_i(\gamma) s_i - \beta b_i))] 
\]
(3)

Here, \( F \) was chosen as a piecewise linear function that (i) gives an exponential
distribution of positive SHAPE reactivities [similar to what is empirically observed (SI
Fig. S8)], (ii) penalizes negative SHAPE reactivities, and (iii) results in a convex
optimization problem for maximum likelihood estimation. The functional form was
\[ F(x) = F_+ (x - x_0) \] if \( x > x_0 \), and \( F(x) = -F_-(x - x_0) \) if \( x \leq x_0 \). The parameters \( F_+ = 5.0 \), \( F_- = 25.0 \), and \( x_0 = 0.06 \) were used, and this formulation appears robust: varying \( F_+ \) and \( F_- \) by
two-fold, changing \( x_0 \) to zero, or using a double exponential fit for \( x > x_0 \), did not change the resulting corrected data beyond experimental errors (as estimated in the next section).

We optimized (3) by performing a grid search of \( \zeta = \gamma/(1-\gamma) \) from 0.0 to 2.0 in 0.05 increments and iteratively solving for \( \alpha \) and then \( \beta \) (through relations obtained by \( dL/d\alpha = 0 \) and \( dL/d\beta = 0 \)) until convergence. Applying the maximum likelihood values of \( \alpha \), \( \beta \), and \( \gamma \) in (2) gave the corrected SHAPE reactivities. The algorithm is available in the function `overmod_and_background_correct_logL.m` within the freely available HiTRACE software package (3).

**Averaging across replicates, estimation of errors, and normalization**

The acquisition of multiple replicates across several experiments permitted high-quality final averaged data \( \overline{S_i} \) with error estimates \( \overline{\sigma_i} \). To carry out the averaging, we noted that individual experiments might have different levels of measurement precision, and the variance of measurements within each experiment provide an estimate of that precision. These estimates, however, do not include systematic errors that differ between experiments, e.g., differing fluorescent backgrounds in different capillary electrophoresis instruments. We therefore carried out a two-part averaging. First, the data within each experiment \( j = 1, 2, \ldots M \) were averaged to give \( S_i^j \) and \( \sigma_i^j \). As an example, suppose we have available 20 replicate measurements of each background/overmodification-corrected SHAPE profile \( s_i^k \), where \( k = 1 \) to 20. Suppose these data were measured for 4 independently prepared RNAs across \( M = 5 \) different days/experiments. Then each \( k \) is the member of one and only one of 5 subsets \( E_j \). Let \( N_j \) be the number of RNAs in set \( E_j \) (here \( N_j = 4 \)). For \( j = 1, 2, \ldots 5 \),
\[ S_i^j = \frac{1}{N_j} \sum_{k \in S_i^j} s_i^k \] (4)

The estimated errors for these SHAPE data are:

\[ \sigma_i^j = \left( \frac{1}{N_j} \sum_{k \in S_i^j} [s_i^k - S_i^j]^2 \right)^{1/2} \] (5)

To combine measurements across multiple experiments, these merged data were averaged, with the inclusion of a position-dependent scale-factor \( \alpha_i \) that accounts for additional sources of experiment-to-experiment error. Explicitly, the assumed likelihood model was:

\[ L(S_i) \propto \prod_j \frac{1}{2\pi \alpha_i \sigma_i^j} e^{-\frac{[S_i^j - S_i]^2}{2[\alpha_i \sigma_i^j]^2}} \] (6)

This gives maximum-likelihood combined signal values \( \bar{S}_i \) and final Gaussian errors \( \bar{\sigma}_i \) of:

\[ \bar{S}_i = \frac{\sum_j \left[ s_i^j \left( \sigma_i^j \right)^2 \right]}{\sum_j \left[ 1 \left( \sigma_i^j \right)^2 \right]} \] (7)

\[ \bar{\sigma}_i = \alpha_i \left( \sum_j \left[ 1 \left( \sigma_i^j \right)^2 \right] \right)^{-1/2} \]

Here, \( \alpha_i \) is a scale factor and is again determined by optimizing the likelihood:
\[ \alpha_i = \frac{1}{M} \sum_{j=1}^{M} \left[ \left( s_j - s_i^j \right)^2 \right. \left/ \left( \sigma_j^i \right)^2 \right] \] (8)

In practice, to obtain a robust estimate of this error scale factor, the average in (8) is taken across a 5-nucleotide window of bands around each nucleotide \( i \). An example of this averaging is given in SI Fig. S3. These data, averaged across multiple replicates, were then normalized following a previously described procedure that was found to be optimal for \( E. coli \) ribosomal RNA (1). Briefly, the data sets were divided by a normalization factor, determined as the average of the top tenth percentile of band intensities. ‘Outliers’, identified as band intensities that exceeded 1.5 times the interquartile range, were removed before determining this factor. The resulting values lie mostly between 0 and 2 (see e.g., main text Fig. 1). The overall algorithm is available in the function `get_average_standard_state.m` within the freely available HiTRACE software package (3).
MPD data
P5abc region; buffer f is not nested in the given secondary structure evidence. Domain. Structure is given in dot cases, designs were checked in RNAstructure and ViennaRNA to give negligible base pairing between added sequences and target variants, different complexes, or different crystallographic space groups within each experiment, two or more independently prepared and purified RNA stocks were assayed.

Table S1. Benchmark for SHAPE-directed secondary structure modeling.

<table>
<thead>
<tr>
<th>RNA, source</th>
<th>Solution conditions¹</th>
<th>Replicates²</th>
<th>Exp ts³</th>
<th>Offset⁴</th>
<th>PDB⁵</th>
<th>Sequence &amp; Secondary Structure⁶</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRNA⁷, E. coli</td>
<td>Standard</td>
<td>14</td>
<td>7</td>
<td>-15</td>
<td>1L0U</td>
<td>1E8H</td>
</tr>
<tr>
<td>P4-P6 domain, Tetrahymena ribozyme</td>
<td>Standard⁷</td>
<td>28</td>
<td>11</td>
<td>89</td>
<td>10G1D</td>
<td>1LBYV 1HR2 2R8S</td>
</tr>
<tr>
<td>5S RNA, E. coli</td>
<td>Standard</td>
<td>12</td>
<td>6</td>
<td>-20</td>
<td>30PF</td>
<td>30AS 30B 2WQ</td>
</tr>
<tr>
<td>Adenine riboswitch, V. vulnificus (add)</td>
<td>Standard + 5 mM adenine</td>
<td>19</td>
<td>6</td>
<td>-8</td>
<td>1X26 1X27 2G9C 3GO2</td>
<td></td>
</tr>
<tr>
<td>c-di-GMP riboswitch, V. cholerae (VC1722)</td>
<td>Standard + 10 µM cyclic di-guanosine mono-phosphate</td>
<td>15</td>
<td>6</td>
<td>0</td>
<td>3MXX 3TWN 3MUV 3MUT</td>
<td></td>
</tr>
<tr>
<td>Glycine riboswitch, F. nucleatum</td>
<td>Standard + 10 mM glycine</td>
<td>22</td>
<td>8</td>
<td>-10</td>
<td>3P49</td>
<td></td>
</tr>
</tbody>
</table>

¹ Standard conditions are: 10 mM MgCl₂, 50 mM Na-HEPES, pH 8.0 at 24 °C.
² All data average over experiments carried out on at least four different days to minimize systematic errors in sample preparations; within each experiment, two or more independently prepared and purified RNA stocks were assayed.
³ Number added to sequence index to yield numbering used in previous biophysical studies, and in Figs. 1 and 2 of the main text.
⁴ The first listed PDB ID was the source of the assumed crystallographic secondary structure; other listed IDs contain sequence variants, different complexes, or different crystallographic space groups and confirm this structure.
⁵ In the sequence, lowercase symbols denote 5’ and 3’ buffer sequences, including primer binding site (last 20 nucleotides). In all cases, designs were checked in RNAstructure and ViennaRNA to give negligible base pairing between added sequences and target domain. Structure is given in dot-bracket notation, and here denotes Watson/Crick base pairs for which there is crystallographic evidence. Only helices with two or more base pairs are included. For the adenine riboswitch, a two-base-pair helix [25-50, 26-49] that is not nested in the given secondary structure and involved in an extensive non-canonical loop-loop interaction is not included.

Additional measurements were carried out with 30% methylpentanediol (MPD) due to reports that its presence in crystallization buffer can change SHAPE reactivity of the P4-P6 RNA (d). Measurements with MPD (10 replicates) gave different reactivities in the P5abc region; final SHAPE-directed secondary structure models mispredicted an additional helix compared to models guided by no-MPD data.
Table S2. Base-pair-level statistics of secondary structure recovery by \textit{RNAsstructure} with and without SHAPE data. TP=true positives; FP=false positives. TP’ and FP’ are the same, but allowing matches of base pair (i,j) with (i±1, j±1).

<table>
<thead>
<tr>
<th>RNA</th>
<th>Len.</th>
<th>Number of base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cryst</td>
<td>RNAstructure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TP</td>
</tr>
<tr>
<td>tRNA\textsubscript{phe}</td>
<td>76</td>
<td>20</td>
</tr>
<tr>
<td>P4-P6 RNA</td>
<td>158</td>
<td>48</td>
</tr>
<tr>
<td>5S rRNA</td>
<td>118</td>
<td>34</td>
</tr>
<tr>
<td>Adenine ribosw.</td>
<td>71</td>
<td>21</td>
</tr>
<tr>
<td>c-di-GMP ribosw.</td>
<td>80</td>
<td>25</td>
</tr>
<tr>
<td>Glycine riboswitch</td>
<td>158</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>661</td>
<td>188</td>
</tr>
</tbody>
</table>

| False negative rate\textsuperscript{b} | 34.0% | 31.9% | 9.6%  | 8.5%  |
| False discovery rate\textsuperscript{c} | 40.7% | 36.6% | 17.1% | 12.2% |
| Sensitivity\textsuperscript{d} | 66.0% | 68.1% | 90.4% | 91.5% |
| Positive predictive value\textsuperscript{e} | 59.3% | 63.4% | 82.9% | 87.8% |

\textsuperscript{a}Cryst = number of helices in crystallographic model.
\textsuperscript{b}False negative rate = 1 - TP/Cryst.
\textsuperscript{c}False discovery rate = FP/(TP+FP).
\textsuperscript{d}Sensitivity = (1 - false negative rate) = TP/Cryst.
\textsuperscript{e}Positive predictive value = (1 - false discovery rate) = TP/(TP+FP).

Table S3. Sources of poor discrimination of correct from incorrect secondary structures.
Thermodynamic energies of base pairs and SHAPE pseudoenergies in kcal/mol, calculated in \textit{RNAsstructure}.

<table>
<thead>
<tr>
<th>RNA</th>
<th>SHAPE-directed model</th>
<th>Crystallographic model\textsuperscript{d}</th>
<th>Difference\textsuperscript{b}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E_{\text{total}}$</td>
<td>$E_{\text{thermo}}$</td>
<td>$E_{\text{SHAPE}}$</td>
</tr>
<tr>
<td>tRNA\textsubscript{phe}</td>
<td>-40.1</td>
<td>-20.3</td>
<td>-19.8</td>
</tr>
<tr>
<td>P4-P6 RNA</td>
<td>-125.6</td>
<td>-54.8</td>
<td>-70.8</td>
</tr>
<tr>
<td>5S rRNA</td>
<td>-95.5</td>
<td>-47.5</td>
<td>-48.0</td>
</tr>
<tr>
<td>Ade ribosw.</td>
<td>-48.2</td>
<td>-16.6</td>
<td>-31.6</td>
</tr>
<tr>
<td>c-di-GMP ribosw.</td>
<td>-63.6</td>
<td>-26.3</td>
<td>-37.3</td>
</tr>
<tr>
<td>Gly. riboswitch</td>
<td>-98.5</td>
<td>-24.8</td>
<td>-73.7</td>
</tr>
<tr>
<td>Average</td>
<td>-78.5</td>
<td>-31.7</td>
<td>-46.9</td>
</tr>
</tbody>
</table>

\textsuperscript{a}For a fair comparison to the SHAPE-directed model, this is the lowest energy secondary structure produced by \textit{RNAsstructure} with the same SHAPE data, but forced to contain the crystallographically observed base pairs. For the adenine riboswitch, an ‘extra’ two-base-pair helix appears in this structure.
\textsuperscript{b}Negative values indicate inaccuracy in structure discrimination.
\textsuperscript{c}$E_{\text{total}}$ and $E_{\text{thermo}}$ are derived from from \textit{efn2} (the \textit{RNAsstructure} package) run with and without SHAPE data, respectively. $E_{\text{SHAPE}}$ is the difference of the two values.
Table S4. Helix-by-helix bootstrap confidence estimates for the SHAPE-directed model of the HIV-1 RNA genome. Models were generated by applying RNAstructure 5.3 to SHAPE data from ref. (5). Following prior work, the temperature was set to the default (37 °C); slope m and intercept b of SHAPE pseudoenergy relation were set to 3.0 kcal/mol and -0.6 kcal/mol, respectively; maximum sequence distance between base pairs was set to 600; modeling was carried out for separate subsegments 1-2844, 2836-5722, 5676-6832, 6807-7791, and 7779-9173; and positions at termini of these subsegments (2836-2845, 5676-5724, 6799-6838, 7779-7791, 9171-9173) and in pseudoknotted regions (179-216, bound to tRNAlys primer; 255-263 in the dimerization loop DIS; and 74-86 and 408-375, forming the 5' polyA signal) were forced to remain unpaired. “BP1”, “BP2”, and “len” give two residues marking the starting base pair of each stem and stem length; “P(boot)” and “BPP” are bootstrap confidence value and maximum Boltzmann probability in the stem as percentages; and “Modeled” gives whether the stem was in the working model of (5) and recovered with RNAstructure 5.3 (Y) or not (N), or whether the stem is newly predicted herein (X). Continued on next two pages.
<table>
<thead>
<tr>
<th>BP1</th>
<th>BP2</th>
<th>seq</th>
<th>P(boot)</th>
<th>BPP</th>
<th>Modeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP3</td>
<td>BP4</td>
<td>seq</td>
<td>P(boot)</td>
<td>BPP</td>
<td>Modeled</td>
</tr>
</tbody>
</table>

`signal-peptide stem at 5' end of gp120.`
**Figure S1.** RNAstructure secondary structure models for a benchmark of six structured RNAs. Cyan lines mark incorrect base pairs; orange lines mark crystallographic base pairs missing in each model. (Figure is in two parts.)

(a) tRNA(phe), E. coli

(b) P4-P6 domain, Tetrahymena group I ribozyme

(c) 5S ribosomal RNA, E. coli
(d) adenosine riboswitch, V. vulnificus

(e) cyclic diGMP riboswitch, V. cholerae

(f) glycine riboswitch, F. nucleatum
Figure S2. Recovery of SHAPE-directed model for a previously studied HCV RNA. As a control, SHAPE data were collected herein for the hepatitis C virus internal ribosomal entry site (HCV IRES) domain II; the resulting SHAPE-directed model agrees with prior work (1) and phylogenetic and NMR analyses [see, e.g., (6)]. Model is colored by SHAPE reactivity (see color scalebar). Helix confidence estimates from bootstrap analyses (see main text) are given as red percentage values. Flanking sequences (similar to those added to benchmark RNAs) are shown in gray.
Figure S3. SHAPE data acquired with different dNTP mix for primer extension, refolding prior to chemical modification, and different DMSO backgrounds. Colored error bars and lines give background-subtracted data for tRNA^phe (E. coli) from six experiments: two experiments in which the dATP, dCTP, dITP, and dTTP were used for reverse transcription of modified RNA; and four experiments in which standard dATP, dCTP, dGTP, and dTTP were used. Each experiment involved at least two replicate measurements; error bars represent standard deviations within each experiment. Arrows mark high-variance bands at C nucleotides in dITP experiments (red) due to poor incorporation of dITP, and near G nucleotides in dGTP experiments (blue) due to band compression. ‘Refold’ experiment 5 (green) involved incubation of RNA at 10 mM MgCl₂, 10 mM Na-MES, pH 6.0 at 50 °C for thirty minutes and gave reactivities indistinguishable from conditions without incubation. Low DMSO experiment 6 (dark green) contained 10% DMSO during chemical modification and gave reactivities indistinguishable from conditions used for other experiments (25% DMSO). Black error bars and lines gives the final averaged SHAPE reactivity averaged over all experiments, taking into account the estimated errors (see SI Methods).
Figure S4. Demonstration that solution SHAPE data reflect folded or ligand-bound conformations. Significant differences were observed upon addition of 10 mM MgCl₂ with a background of 50 mM Na-HEPES (for the P4-P6 domain & 5S rRNA) and upon addition of ligand with a background of 10 mM MgCl₂, 50 mM Na-HEPES, pH 8.0 (for the ligand-binding domains of riboswitches for adenine, c-di-GMP, and glycine). Regions that become protected upon Mg²⁺-induced tertiary folding or ligand binding are annotated on the data, and compare well to expectations from previous biophysical and crystallographic studies (7-13).
**Figure S5. Partition-function and bootstrap analysis of SHAPE-directed secondary structure models.** A confidence estimate for each helix in each of the six benchmark SHAPE-directed models was determined by (1) partition-function-based Boltzmann probabilities and (2) a nonparametric bootstrap analysis (repeating the modeling with ‘replicate’ data sets generated by randomly resampling the data with replacement). The confidence estimates for the two analyses correlate approximately, but partition function probabilities are skewed to higher values than bootstrap probabilities. Helices that agree (blue) or disagree (red) with crystallographic secondary structures are plotted separately. Dashed lines mark 80% and 55% separatrix values, above which two incorrect helices are observed, and 29 and 31 correct helices are observed for Boltzmann probability and bootstrap analyses, respectively.
Figure S6. Sensitivity of minimum-energy model and robustness of bootstrap values to small changes in tRNA SHAPE data. (a) Comparison of SHAPE data sets obtained by averaging over all collected data (14 replicates; blue) and by averaging over just those data collected with primer extension with standard deoxynucleotide triphosphates (no dITP; i.e., dATP, dCTP, dGTP, dTTP) (6 replicates; red). (b,c) Minimum-energy SHAPE-directed secondary structures are different for the two data sets in the pairings of the third helix; bootstrap values given as red percentage values. (d,e) Helix probabilities from bootstrap analysis shown in grayscale, with 0 to 100% shown as white to black. Bootstrap values at alternate locations of third helix are shown as red percentage values; they are similar for the two data sets. Red squares mark crystallographic base pairs.
Figure S7. HIV-1 secondary structure helix confidence values compared to SHAPE reactivities. For each helix in the HIV-1 secondary structure model (5), the median SHAPE reactivity for nucleotides in the helix was computed, and plotted against bootstrap values. Blue line marks median reactivity over all nucleotides. High-bootstrap-value helices (e.g., four helices in TAR, three helices in gag-pol, and the gp120 signal peptide stem; shown as red x’s) typically have low median SHAPE reactivities. However, the converse is not true. Low-reactivity helices frequently have poor bootstrap values, indicating the existence of multiple secondary structures consistent with the data while still protecting the associated regions.
**Figure S8. Histogram and fit of SHAPE reactivities.** SHAPE reactivities of all residues for the six test RNAs (black; see SI Table S1), compared to least-squares fit (red) to a simple probability distribution \( P(x) \). The distribution was assumed to take the form \( P(x) = N \exp(F|x-x_0|) \) for \( x > x_0 \); and \( P(x) = N \exp(F|x-x_0|) \) for \( x \leq x_0 \). The presented fit is for \( x_0 = 0.06; F_+ = 5.0; F_- = 25.0 \).
References for Supporting Information