RNA design rules from a massive open laboratory

Jeehyung Lee[a], Wipapat Kladwang[b], Minjae Lee[a], Daniel Cantu[b], Martin Azizyan[a], Hanjoo Kim[c], Alex Limpaecher[a], Sungroh Yoon[c], Adrien Treuille[a,1], Rhiju Das[b,d,1], and EteRNA Participants[2]

[a]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15206; Departments of [b]Biochemistry and [d]Physics, Stanford University, Stanford, CA 94305; and [c]Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-741, Korea

Edited by David Baker, University of Washington, Seattle, WA, and approved December 12, 2013 (received for review July 12, 2013)

Self-assembling RNA molecules present compelling substrates for the rational interrogation and control of living systems. However, imperfect in silico models—even at the secondary structure level—hinder the design of new RNAs that function properly when synthesized. Here, we present a unique and potentially general approach to such empirical problems: the Massive Open Laboratory. The EteRNA project connects 37,000 enthusiasts to RNA design puzzles through an online interface. Uniquely, EteRNA participants not only manipulate simulated molecules but also control a remote experimental pipeline for high-throughput RNA synthesis and structure mapping. We show herein that the EteRNA community leveraged dozens of cycles of continuous wet laboratory feedback to learn strategies for solving in vitro RNA design problems on which automated methods fail. The top strategies—including several previously unrecognized negative design rules—were distilled by machine learning into an algorithm, EteRNABot. Over a rigorous 1-y testing phase, both the EteRNA community and EteRNABot significantly outperformed prior algorithms in a dozen RNA secondary structure design tests, including the creation of dendrimer-like structures and scaffolds for small molecule sensors. These results show that an online community can carry out large-scale experiments, hypothesis generation, and algorithm design to create practical advances in empirical science.

RNA folding | citizen science | high-throughput experiments | crowdsourcing

Structured RNA molecules play critical roles in biological processes from genetic regulation to viral replication; the characterization, detection, and reengineering of these RNAs are major goals of modern molecular biology and bioengineering (1–7). Recent years have witnessed the emergence of elegant RNA folding models that accurately capture secondary structure formation of loops and simple helices (8–12). However, more complex motifs, such as multiloops, remain challenging to model (1), and thus, algorithmically designed RNAs frequently misfold in vitro. Practitioners must often fall back on trial-and-error refinement or problem-specific selection methods (1–7).

High-throughput synthesis and biochemical interrogation offer the prospect of developing better folding models. Nevertheless, a small group of professional scientists must interpret this torrent of empirical data, a challenging task even with modern machine learning and visualization tools. The results of such big data science often lack the parsimony, elegance, and predictive power of handcrafted models. This paper presents an alternative approach, a Massive Open Laboratory, that combines the parallelism of high-throughput experimental biochemistry with the advantages of detailed human-guided experimental design and analysis.

The 37,000-member EteRNA project has now generated many hundreds of designs probed at single nucleotide resolution, resulting in a database of nearly 100,000 data points. Instead of outpacing human curation, this unprecedented dataset of designs has been created concomitantly with detailed handcrafted hypotheses advanced by the community, most of which were previously unexplored in the RNA modeling literature. Sifting and automating these hypotheses by machine learning has resulted in an automated algorithm, EteRNABot, which parsimoniously describes a unique optimization function for RNA design. A gauntlet of additional design targets tested this algorithm, including previously unseen RNA secondary structures as well as complex scaffolds for small molecule sensors, with binding that provided independent readouts of folding accuracy. These tests confirmed that both EteRNABot-designed RNAs and handcrafted RNAs by the community outperform existing state of the art algorithms. Although previous internet-scale communities have solved difficult problems in silico (13–16), the results herein are unique in showing that such a community can collectively generate and test hypotheses through actual experiments, which are required for advancing empirical science.

Results

EteRNA combines an interactive interface for modeling biomolecules with a remote wet laboratory experimental pipeline (*Materials and Methods* and Fig. 1). A web-based interface challenges participants to design and rank sequences that will fold into a target structure when synthesized in vitro (*SI Appendix*, Fig. S1 and Table S1 give all design targets) and develop design rules that explain the community's experimental results. High-throughput synthesis and structure mapping measurements [selective 2′-hydroxyl acylation with primer extension (SHAPE)] (17) (*Materials and Methods* and Fig. 1*C*) assess nucleotide pairing of eight community-selected designs per week. EteRNA returns these experimental results to participants through visualization of the data at single nucleotide resolution (Fig. 1*D*) as well as an overall structure mapping score on a scale of 0–100 (*Materials and Methods*), indicating the percentage of nucleotides giving reactivities consistent with the target structure (experimental

Significance

Self-assembling RNA molecules play critical roles throughout biology and bioengineering. To accelerate progress in RNA design, we present EteRNA, the first internet-scale citizen science "game" scored by high-throughput experiments. A community of 37,000 nonexperts leveraged continuous remote laboratory feedback to learn new design rules that substantially improve the experimental accuracy of RNA structure designs. These rules, distilled by machine learning into a new automated algorithm EteRNABot, also significantly outperform prior algorithms in a gauntlet of independent tests. These results show that an online community can carry out large-scale experiments, hypothesis generation, and algorithm design to create practical advances in empirical science.

Author contributions: J.L., M.A., S.Y., A.T., R.D., and EteRNA Participants designed research; J.L., W.K., M.L., D.C., M.A., H.K., A.L., S.Y., A.T., R.D., and EteRNA Participants performed research; J.L., W.K., M.L., D.C., M.A., H.K., A.L., S.Y., A.T., and EteRNA Participants analyzed data; and J.L., A.T., and R.D. wrote the paper.

Conflict of interest statement: The editor, David Baker, is a recent coauthor with R.D. having published a paper with him in 2013.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. E-mail: treuille@cs.cmu.edu or rhiju@stanford.edu.

[2]A complete list of the EteRNA Group can be found in Dataset S1.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1313039111/-/DCSupplemental.

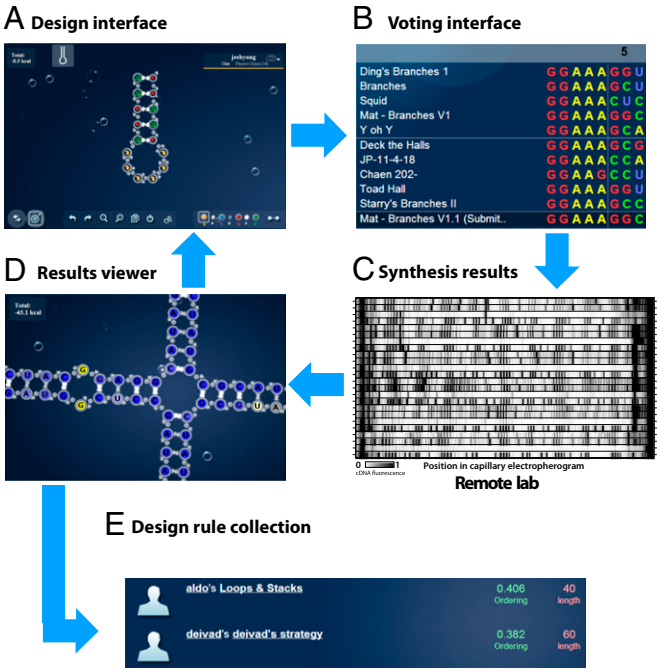


Fig. 1. EteRNA workflow. Each week, participants (*A*) design sequences that can fold into a target RNA structure in the sequence design interface and (*B*) review and vote for the best designs with the voting interface. (*C*) At the end of the round, the eight top-voted sequences are synthesized and verified by single nucleotide resolution chemical reactivity measurements. (*D*) The experimental results are published online and available for review in the results viewer. Participants then create new hypotheses and (*A*) start the next experimental cycle or (*E*) submit design rules learned from the results (text) that are codified and automatically ranked based on scores obtained to date (numbers).

error ±5) (*SI Appendix*, Fig. S2). As participants intuit features of experimentally successful designs, they can submit heuristics to a design rule collection (Fig. 1*E*). Additional descriptions of the platform design (*SI Appendix*, Fig. S3), training of participants, reward structure, visualization, participant distributions (*SI Appendix*, Fig. S4), and experimental reproducibility (*SI Appendix*, Fig. S2) are given in *Materials and Methods* and *SI Appendix*.

The initial 6-mo training period, called phase I, saw the EteRNA community engaging in six RNA design problems containing increasing numbers of nonhelical elements (bulges and multihelix junctions) and more complex topologies (Fig. 2), mimicking components of known functional RNAs (1, 5, 6, 18); 189 community-chosen sequences were synthesized along with 65 sequences from RNAInverse (11) and NUPACK (12) algorithms for comparison. Initially, the community was inexperienced, and their designs depended solely on computational folding models (11). These designs fared poorly: during the first laboratory competition (the three-helix finger) (Fig. 2*A*), many participants' designs gave structure mapping scores lower than 70 (compared with greater than 90 for all NUPACK designs) (*SI Appendix*, Table S2). However, as the community gained experience with empirical RNA design cycles, performance improved, and community submissions converged to successful designs (above 90) in two to three rounds for all targets (Fig. 2).

Beyond this target-specific learning, structure mapping scores from the first round of each new target increased over time, suggesting that the participants were developing generalizable design rules (blue symbols in Fig. 2). Over all six targets, these first round scores increased continuously. By the third target (Fig. 2*C*), participants outperformed both RNAInverse and NUPACK in their first round maximum score. By the fifth and sixth targets (Fig. 2 *E* and *F*), first round median participant scores exceeded the algorithms' maximum scores, with top participant designs achieving scores indistinguishable from perfect designs given experimental error (>95). In contrast, the increasing structural complexity (measured in stems and junctions) led to declining performances for RNAInverse and NUPACK (Fig. 2). First round designs from EteRNA participants were significantly better than designs from RNAInverse and NUPACK in the last three puzzles, with P values of 2.9×10^{-4} against both algorithms (Fig. 2, structure mapping data and *SI Appendix*, Table S3*B*). We independently confirmed these results from the EteRNA training period by additional tests based on several additional design challenges (*SI Appendix*, Fig. S5), automated SHAPE-directed secondary structure inference (*SI Appendix*, Fig. S6), a different chemical mapping method based on dimethyl sulfate alkylation (19) (*SI Appendix*, Fig. S7), 2D chemical mapping with the more information-rich mutate-and-map technique (20) (which suggests structural heterogeneity in failed designs; *SI Appendix*, *Supporting Results* and Fig. S8), and replicates by separate experimenters and with alternative techniques (next generation sequencing) (*SI Appendix*, Fig. S2). *SI Appendix*, *Supporting Results* gives a complete description of these results and structure models.

During the training challenges, the community-submitted collection of design rules grew to 40 contributions (Fig. 3 and *SI Appendix*, Table S4), most of which encoded unique insights into successful RNA design. On one hand, some of these rules involved features previously discussed in the RNA design literature [e.g., G-C content (*SI Appendix*, Table S4, A Basic Test), the ensemble defect (8, 12) (*SI Appendix*, *Supporting Results* and Table S4, Clean Dot Plot), and sequence symmetry minimization (21) (*SI Appendix*, Table S4, Repetition)]. Some features were similar to patterns highlighted in bioinformatic analyses of natural structured RNAs, such as the prevalence of G-C closing pairs at multiloop junctions (*SI Appendix*, Table S4, GC Pairs in Junctions) or the general prevalence of adenosines outside stems (*SI Appendix*, Table S4, Only As in the Loops) (18). On the other hand, to our knowledge, most of the EteRNA design rules were unique in the RNA folding and design field, including prescriptions for the identities of unpaired nucleotides adjacent to stems (*SI Appendix*, Table S4, No Blue Nucleotides in Hook Area), C-G vs. G-C edge base pairs in different contexts (*SI Appendix*, Table S4, Direction of GC Pairs in Multiloops + Neck Area), and placement of Gs within loops (*SI Appendix*, Table S4, Gs in Place of the Last As on the Righthand Side of Any End Loop).

Few of these rules have been previously encoded into energetic models or automated RNA design methods, much less confirmed experimentally, and it remained unclear if the participants' proposed rules accounted for their outperformance of prior design methods. We, therefore, sought to evaluate the rules independently from EteRNA participants through their integration into a single score function. Sparse machine learning regression (22) with cross-validation selected five rules (Fig. 3), which we tested by incorporation into a unique automated Monte Carlo algorithm called EteRNABot and rigorous experimental tests. *Materials and Methods* and *SI Appendix*, *Supporting Results* provide additional discussion on this algorithm, a less parsimonious algorithm EteRNABot-alt reweighting all 40 rules, and a variant algorithm using only features preexisting in the project interface.

In the subsequent testing period, called phase II, nine unique targets challenged EteRNA participants, the EteRNABot method, and prior algorithms (Fig. 4 and *SI Appendix*, Fig. S9). The first five targets (Fig. 4 *A*–*E*) were multijunction structures distinct from each other and the phase I structures in topology. We evaluated only one round of participant designs per target, thus testing whether community knowledge was generalizable across target structures. We again observed superior performance of the participant designs compared with RNAInverse and NUPACK ($P = 1.5 \times 10^{-4}$ and 2.9×10^{-4}, respectively) (*SI Appendix*, Table S3*C*). Furthermore, in three of five cases (Fig. 4 *B*, *D*, and *E*), automated designs from the unique EteRNABot algorithm achieved maximum scores within ±1.5 of the participant designs and median scores within ±5.5. In the two remaining cases

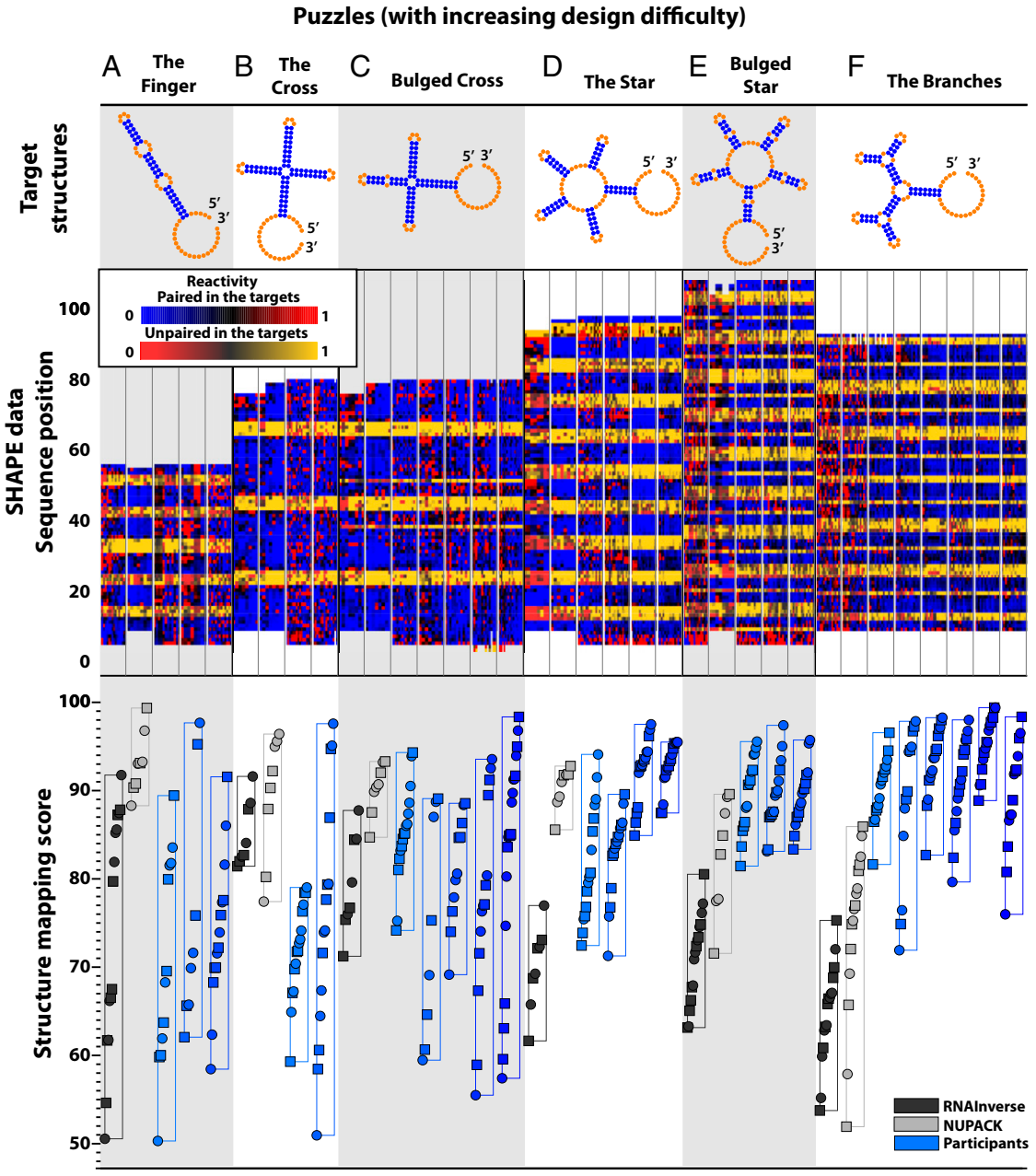


Fig. 2. Phase I puzzles and results in order of puzzle posting date. *Top* shows a target structure and title for each puzzle (*A–F*). Nucleotide coloring in target structures indicates the ideal SHAPE reactivity (gold for high reactivity and blue for low reactivity). *Middle* gives the single nucleotide resolution reactivity data measured for all designs. Yellow stripes indicate bases that should show high reactivity if the target secondary structure (*Top*) is formed. *Bottom* shows a summary of structure mapping scores for designs from the RNAInverse (black) and NUPACK (gray) algorithms compared with EteRNA participants (colored symbols; ordered by score within each design round). Each design was subjected to SHAPE chemical reactivity mapping in two solution conditions, 1 M NaCl (circles) and 10 mM $MgCl_2$ (squares), with 50 mM Na-Hepes (pH 8.0) at 24 °C. The colored border lines connect designs within the same round.

(Fig. 4 *A* and *C*), EteRNABot modestly underperformed participants, a gap that may close as more experimental data and design rules are collected. Importantly, EteRNABot outperformed RNAInverse and NUPACK ($P = 3.0 \times 10^{-4}$ and 1.2×10^{-3}, respectively) (*SI Appendix*, Table S3*C*), with higher maximum scores in all cases (Fig. 4 and *SI Appendix*, Table S2*A*) as well as better ability to rank top designs (*SI Appendix*, Fig. S10).

The last four puzzles (Fig. 4 *F–I*) of phase II presented challenges that arise in the engineering of RNA-based switches: the inclusion of sensor domains [in this case, a 13-nt internal loop that binds the small molecule flavin mononucleotide (FMN) whose sequence was held fixed] (23, 24). Consistent with previous results, EteRNA participants and the EteRNABot algorithm outperformed NUPACK and RNAInverse in terms of their structure mapping scores in FMN-free conditions ($P < 0.06$ in all comparisons) (Fig. 4 *F–I* and *SI Appendix*, Table S3*D*). Furthermore, these designs' association constants for FMN binding offered stringent tests of folding accuracy that were fully independent of the structure mapping scores (Fig. 4 *K–N* and *SI Appendix*, *Supporting Results* and Fig. S11). Here again, both EteRNA participants and EteRNABot outperformed RNAInverse and NUPACK in both best and median association constants ($P < 0.05$ in all pairwise comparisons) (Fig. 4*N* and *SI Appendix*, Table S3*E*). These small molecule binding measurements

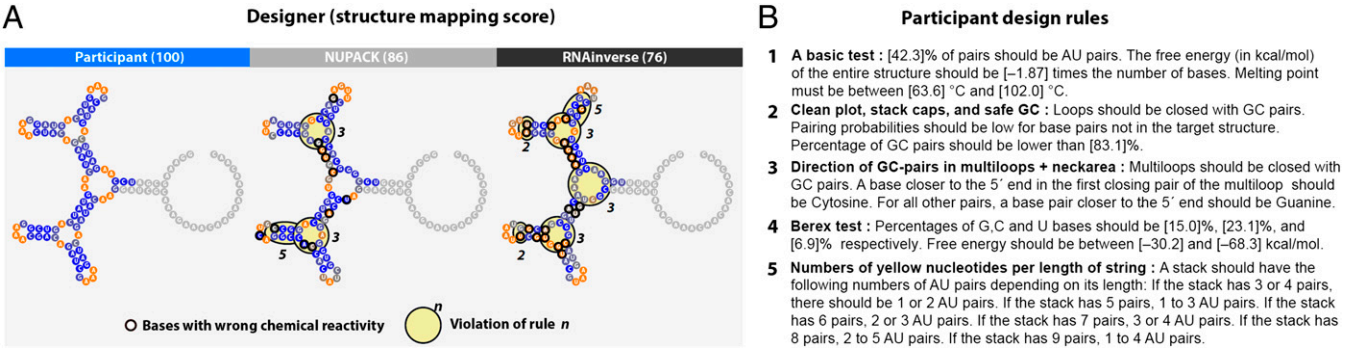


Fig. 3. RNA design rules proposed by EteRNA participants. (*A*) The best designs from each design agent (EteRNA participants, NUPACK, and RNAInverse) for the last target shape of phase I (Fig. 2*F*); the nucleotide coloring gives experimental chemical reactivity and is identical to the coloring used in Fig. 2. The designs are annotated with violations of the top 5 rules of 40 rules proposed by participants, which were assessed by sparse linear regression. (*B*) The five rule statements used for EteRNABot. The numerical parameters in brackets were optimized to best explain the results from a training set based on starting values proposed by participants (*Materials and Methods*).

independently confirmed the SHAPE results above: rules developed by the EteRNA community permit more accurate automated design of RNA secondary structures than has been previously possible. The resulting EteRNABot algorithm should be of immediate practical use.

Discussion

The EteRNA project has discovered unique RNA design rules by giving an internet-scale community of citizen scientists access to high-throughput wet laboratory experimentation, totaling nearly 100,000 single nucleotide resolution data points. The community's design rules have been empirically and rigorously validated through design tests involving nine target structures distinct from six structures of the training period and independent flavin mononucleotide binding titrations on four scaffold structures. Outperformance of prior in silico metrics in these tests (*SI Appendix*, Figs. S10 and S12) confirms the importance of experiments in inspiring the rules. Mechanistic work will be required to give atomic-level explanations for the rules' predictive power. In this sense, the EteRNA rules are analogous to energetic models, such as the nearest neighbor rules (25), which are also empirically derived but not yet derivable from first principles (26, 27) or other design heuristics (28). From a mechanistic perspective, one interesting feature shared across many of the EteRNA design rules collected so far is the use of negative design rules. For example, penalties on repeated *n*-mers (repetition), the disallowance of mixtures of strong and weak tetraloops (tetraloop similarity), and penalties for similarity between neighboring base pairs (twisted base pairs) are potentially strategies that would prevent misfolding in any reasonable energetic model. Such features may not be captured in prior RNA design algorithms, which may uncover designs that stabilize the target structure compared with misfolds by overoptimizing idiosyncrasies of a particular energetic model. The emergence of energy function-independent negative design rules in the EteRNA project underscores the importance of actual experimental falsification/validation in developing RNA design methods.

Beyond its implications for RNA engineering, our method represents a successful attempt to generate and experimentally test hypotheses through crowdsourcing. As the data throughput of experimental approaches continues to grow, this approach offers several benefits. Currently, small sets of professional scientists attempt to resolve the complexity of designing and analyzing high-throughput experiments and enumerate a space of folding hypotheses for computational analysis of these data. Instead, the approach herein enables a vastly larger number of participants to design and execute remote experiments in parallel, while machine learning algorithms sift through the community's catalog of hypotheses. This Massive Open Laboratory template could be generalized to a broad class of biomolecule design problems, including mechanistic dissection of current design rules (26), modeling of pseudoknots, engineering of RNA switches for cellular control (5, 6), and 3D modeling and design, all assessed by high-throughput mapping (1, 7, 16, 20, 29, 30). Other fields, such as taxonomy (31), astronomy (13), and neural mapping (32), are making pioneering efforts in internet-scale scientific discovery games. Our Massive Open Laboratory results suggest that integrating timely player-proposed experiments as part of the standard game play will be worthwhile challenges for such projects.

Materials and Methods

Online Interface. EteRNA is an online Flash (Adobe Systems Inc.) application that can be accessed within any web browser. EteRNA presents the RNA design problem as a set of puzzles; participants use an interactive sequence design interface to design RNA sequences that fold into target secondary structures. The interface visualizes each nucleotide with yellow, blue, red, and green circular symbols representing adenine, uracil, guanine, and cytosine, respectively. Symbols are laid out using the NAView drawing algorithm (33). The secondary structure display updates in real time with the minimum free energy pseudoknot-free solution predicted by the ViennaRNA package (11) (compiled into Flash). The interface also gives access to predicted melting curves and dot plots of alternative base pairings (11). Additional descriptions of the individual components of the EteRNA online interface are presented in *SI Appendix, Supporting Methods* and Fig. S3.

Design Rule Selection Method and EteRNABot. EteRNABot is a unique algorithm for design of pseudoknot-free secondary structures that optimize a function to predict structure mapping scores (see below) generated from participant-submitted design rules. To create the score predictor, each participant-submitted rule was coded into a scoring function. When a rule contained nondiscrete numeric parameters (given as numbers in brackets in Fig. 3*B* and *SI Appendix*, Table S4), its scoring function was optimized over the parameters using the downhill simplex algorithm (34) to minimize the average squared error between the predicted and actual structure mapping scores for the training set (results from phase I and the four follow-up puzzles before phase II). *SI Appendix*, Table S4 lists all design rules and corresponding scoring functions; *SI Appendix*, Table S5 gives a glossary of frequently used terms in design rules. The EteRNABot score is a linear combination of five scoring functions selected from 40 submitted rules using least angle regression (22) and cross-validation from analyses leaving out data for each target shape (Fig. 3). *SI Appendix, Supporting Results* and Fig. S10 report the predictive power of EteRNABot scores for structure mapping scores. To design a unique secondary structure, EteRNABot runs a loop of randomized nucleotide mutations to find a sequence that accepts a mutation if it increases the sequence's predicted score or decreases a base pair distance between the predicted minimum free energy structure [calculated with ViennaRNA (11)] and the target structure; its speed is nearly the same as RNAInverse. The search ends when the sequence's predicted score is over 90 and the base pair distance is less than 0.1 times the sequence length. The EteRNABot algorithm and its training data are freely available as a server at http://eternabot.org. We also report experimental results of an alternate EteRNABot that uses all 40 submitted rules in *SI Appendix, Supporting Results*, Fig. S10, and Tables S2 and S3.

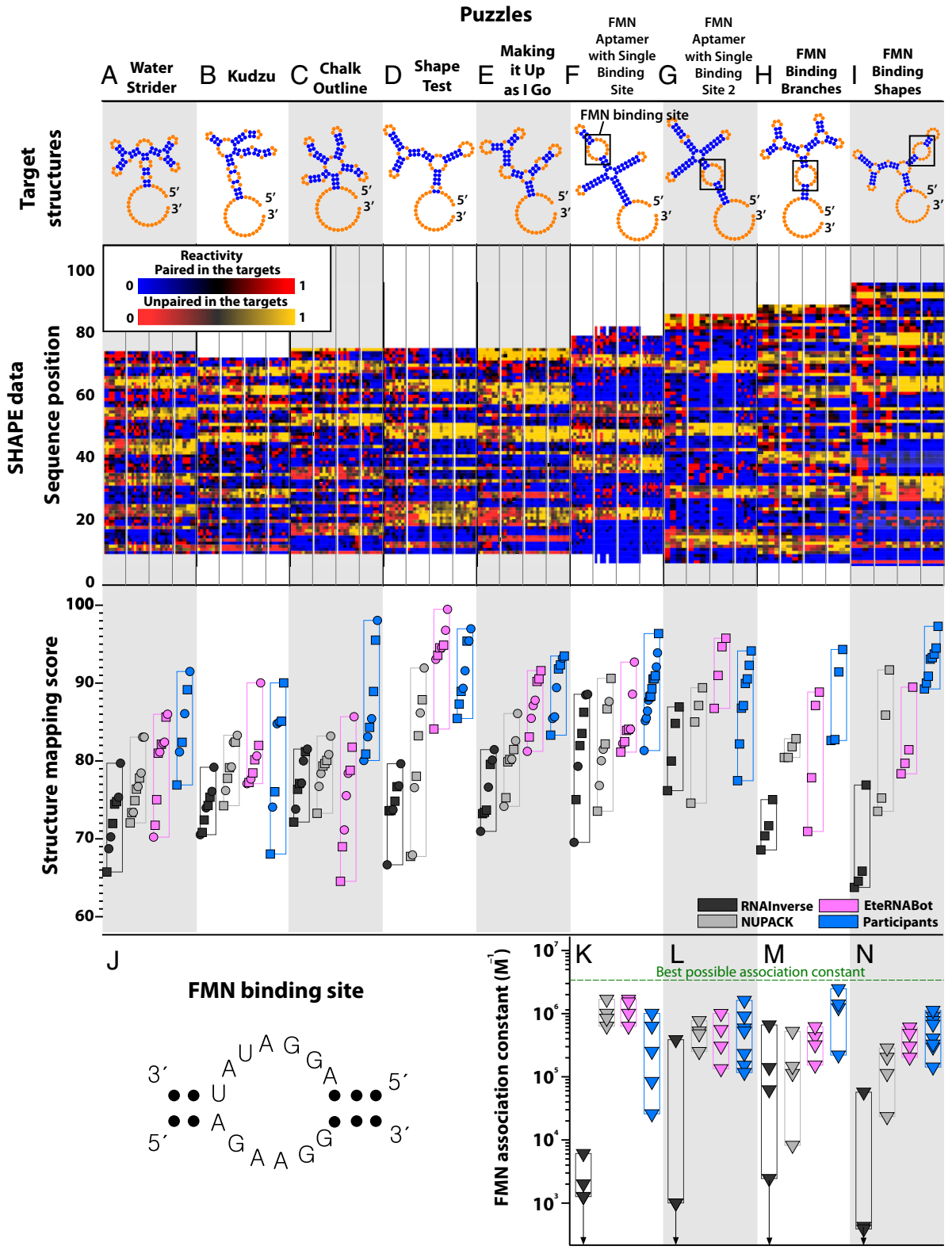


Fig. 4. Phase II puzzles and results, including the EteRNABot algorithm, in order of puzzle posting date. (*A–E*) Multijunction target structures distinct from each other and the phase I structures in topology. (*F–I*) Puzzles giving the binding site for FMN as a fixed sequence that must be displayed as an internal loop (boxed in the target structure; *J* shows the sequence) within a complex structure. The coloring scheme is identical to the coloring scheme used in Fig. 2, with EteRNABot shown in magenta. (*K–N*) FMN association constants for all designs synthesized for the last four puzzles measured by dimethyl sulfate chemical mapping as a function of FMN concentration. Measurements on a simple construct displaying FMN binding (green dashed line) give the best possible association constant, which is achievable only with correct secondary structure folding. The down arrows in *K–N* mark RNAInverse designs that gave no observable FMN binding, setting the upper bounds on the association constants.

RNA Synthesis and Structure Mapping. RNA sequences were prepared by in vitro transcription with T7 RNA polymerase from DNA templates encoding the sequence designs and probed with structure mapping based on *N*-methylisatoic anhydride [SHAPE chemistry (17)] using 96-well protocols described previously (35). All RNAs contained a shared primer binding site (AAAGAAACAACAACAACAAC) at their 3′ end, which was included as a fixed sequence in EteRNA puzzles. Measurements included SHAPE reactions [final concentration of *N*-methyl isatoic anhydride of 6 mg/mL with 20% DMSO (vol/vol)] or dimethyl sulfate reactions (final concentration of 0.2%) at 24 °C with 60 nM RNA in two solution conditions (10 mM $MgCl_2$ and 1 M NaCl) with 50 mM Na-Hepes (pH 8.0), control measurements without SHAPE reagent, and control measurements using 2′-3′-dideoxythymidine triphosphate in primer extension to generate reference ladders at adenosine residues. All data were aligned and quantified with the HiTRACE software (36), corrected for attenuation of long reverse transcription products, and background-subtracted as described in ref. 35. SHAPE-directed secondary structure models and confidence estimates were obtained with data-derived pseudoenergy terms and nonparametric bootstrapping (35). Binding titrations to FMN were monitored with dimethyl sulfate alkylation (19), which gave a strong protection signal on FMN binding to the aptamer (*SI Appendix*, Fig. S11 *A–C*). Titrations were analyzed with likelihood-based fits and error estimation (*SI Appendix*, Fig. S11 *C* and *D*) (37). Finally, for 30 sequences, we also carried out the SHAPE-seq method read out by Illumina sequencing, which is described in ref. 38, although this protocol's systematic errors (in PCR and ligation bias) precluded its general use for EteRNA scoring. Detailed protocols are in refs. 39 and 40.

Structure Mapping Scores. In addition to returning nucleotide by nucleotide SHAPE data, the quality of each synthesized design was summarized and reported to participants as a structure mapping score, which was analogous to the L1 norm scores used in prior work (41). A nucleotide was assigned a point if its reactivity exceeded 0.25 (if designed to be unpaired) or was less than 0.50 (if designed to be paired). The threshold for unpaired nucleotides was less stringent to allow for the possibility that the nucleotide could have reduced reactivity from non-Watson–Crick or other interactions and set based on calibration data on natural structured RNAs (35). The baseline and normalization of each dataset were determined using linear programming to optimize the total score. Scores were given as the percentage of nucleotides with points (0–100). An additional scoring system based on the ratio of likelihoods for the data given the target secondary structure and the best possible unpaired/paired status at each nucleotide was also tested using likelihood distributions derived from a benchmark of natural RNAs (35). The likelihood-based scheme gave rankings consistent with the point-based scheme and took into account experimental error; we chose to use the point-based scheme to calculate the EteRNA structure mapping score because of its simplicity.

Availability

EteRNA platform and all synthesis data used in this paper are available at http://eternagame.org. EteRNABot and its training data are available at http://eternabot.org. Please see Dataset S1 for a complete listing of the EteRNA participants.

ACKNOWLEDGMENTS. We thank the following colleagues for discussions and expert assistance with platform development: M. Baumgartner, B. Bethrum, J. P. Bida, J. Ciscon, S. Federwisch, N. Fishel, J. Fu, S. Gaikwad, T. Hu, P. Kinney, K. Kiser, D. Klionsky, S. Koo, S. Mohan, I. Riedel-Kruse, A. Singh, C. Spitale, C. C. VanLang, and M. Yao. This work was funded by seed grants from the National Science Foundation [EAGER (EArly-concept Grants for Exploratory Research) IIS-1043251], the Stanford Bio-X Interdisciplinary Initiative Program, Stanford Media-X, National Research Foundation of Korea Grant 2011-0009963 (to S.Y.), a Burroughs-Wellcome Foundation Career Award (to R.D.), the Keck Medical Research Foundation, and National Institutes of Health Grant R01 R01GM100953.

1. Bida JP, Das R (2012) Squaring theory with practice in RNA design. *Curr Opin Struct Biol* 22(4):457–466.
2. Mattick JS (2004) RNA regulation: A new genetics? *Nat Rev Genet* 5(4):316–323.
3. Goldberg MS, et al. (2011) Nanoparticle-mediated delivery of siRNA targeting Parp1 extends survival of mice bearing tumors derived from Brca1-deficient ovarian cancer cells. *Proc Natl Acad Sci USA* 108(2):745–750.
4. Stelzer AC, et al. (2011) Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. *Nat Chem Biol* 7(8):553–559.
5. Win MN, Smolke CD (2008) Higher-order cellular information processing with synthetic RNA devices. *Science* 322(5900):456–460.
6. Sinha J, Reyes SJ, Gallivan JP (2010) Reprogramming bacteria to seek and destroy an herbicide. *Nat Chem Biol* 6(6):464–470.
7. Delebecque CJ, Lindner AB, Silver PA, Aldaye FA (2011) Organization of intracellular reactions with rationally designed RNA assemblies. *Science* 333(6041):470–474.
8. Jacobson AB, Zuker M (1993) Structural analysis by energy dot plot of a large mRNA. *J Mol Biol* 233(2):261–269.
9. McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29(6–7):1105–1119.
10. Mathews DH, et al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA* 101(19):7287–7292.
11. Hofacker IL (2004) RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc BioInform*122004p 12.2.
12. Zadeh JN, et al. (2011) NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem* 32(1):170–173.
13. Land K, et al. (2008) Galaxy Zoo: The large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey. *Mon Not R Astron Soc* 388(4):1686–1692.
14. Kanefsky B, Barlow NG, Gulick VC (2001) Can distributed volunteers accomplish massive data analysis tasks? *Proceedings of the 32nd Annual Lunar and Planetary Science Conference*. Available at http://ti.arc.nasa.gov/m/pub-archive/212h/0212%20(Kanefsky).pdf. Accessed January 7, 2014.
15. Khatib F, et al. (2011) Algorithm discovery by protein folding game players. *Proc Natl Acad Sci USA* 108(47):18949–18953.
16. Eiben CB, et al. (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotechnol* 30(2):190–192.
17. Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* 127(12):4223–4231.
18. Laing C, Jung S, Iqbal A, Schlick T (2009) Tertiary motifs revealed in analyses of higher-order RNA junctions. *J Mol Biol* 393(1):67–82.
19. Cordero P, Kladwang W, VanLang CC, Das R (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* 51(36):7037–7039.
20. Kladwang W, VanLang CC, Cordero P, Das R (2011) A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem* 3(12):954–962.
21. Seeman NC (1982) Nucleic acid junctions and lattices. *J Theor Biol* 99(2):237–247.
22. Efron B, Hastie T, Johnstone T, Tibshirani R (2004) Least angle regression. *Ann Stat* 32(2):407–499.
23. Burgstaller P, Famulok M (1994) Isolation of RNA aptamers for biological cofactors by in vitro selection. *Angew Chem* 33(10):1084–1087.
24. Jose AM, Soukup GA, Breaker RR (2001) Cooperative binding of effectors by an allosteric ribozyme. *Nucleic Acids Res* 29(7):1631–1637.
25. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288(5):911–940.
26. Yildirim I, Turner DH (2005) RNA challenges for computational chemists. *Biochemistry* 44(40):13225–13234.
27. Ducongé F, Di Primo C, Toulme JJ (2000) Is a closing "GA pair" a rule for stable loop-loop RNA complexes? *J Biol Chem* 275(28):21287–21294.
28. Dirks RM, Lin M, Winfree E, Pierce NA (2004) Paradigms for computational nucleic acid design. *Nucleic Acids Res* 32(4):1392–1403.
29. Das R, et al. (2008) Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci USA* 105(11):4144–4149.
30. Jaeger L, Chworos A (2006) The architectonics of programmable RNA and DNA nanostructures. *Curr Opin Struct Biol* 16(4):531–543.
31. Schindel DE, Miller SE (2005) DNA barcoding a useful tool for taxonomists. *Nature* 435(7038):17.
32. Helmstaedter M (2013) Cellular-resolution connectomics: Challenges of dense neural circuit reconstruction. *Nat Methods* 10(6):501–507.
33. Bruccoleri RE, Heinrich G (1988) An improved algorithm for nucleic acid secondary structure display. *Comput Appl Biosci* 4(1):167–173.
34. Nelder JA, Mead R (1965) A simplex-method for function minimization. *Comput J* 7(4):308–313.
35. Kladwang W, VanLang CC, Cordero P, Das R (2011) Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* 50(37):8049–8056.
36. Yoon S, et al. (2011) HiTRACE: High-throughput robust analysis for capillary electrophoresis. *Bioinformatics* 27(13):1798–1805.
37. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods* 7(4):291–294.
38. Lucks JB, et al. (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci USA* 108(27):11063–11068.
39. Cordero P, Kladwang W, VanLang CC, Das R (2014) The mutate-and-map protocol for inferring base pairs in structured RNA. *Methods Mol Biol* 1086:53–77.
40. Seetin MG, Kladwang W, Bida JP, Das R (2014) Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *Methods Mol Biol* 1086:95–117.
41. Quarrier S, Martin JS, Davis-Neulander L, Beauregard A, Laederach A (2010) Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA* 16(6):1108–1117.

Supporting Information for 'RNA Design Rules from a Massive Open Laboratory'

Authors: Jeehyung Lee[1], Wipapat Kladwang[2], Minjae Lee[1], Daniel Cantu[2], Martin Azizyan[1], Hanjoo Kim[3], Alex Limpaecher[1], Sungroh Yoon[3], Adrien Treuille[1,*], Rhiju Das[2,4,*], and EteRNA participants

[1]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15206, USA
[2]Biochemistry Department, Stanford University, Stanford, CA 94305, USA
[3]Department of Electrical and Computer Engineering, Seoul National University, 151-741, Korea
[4]Physics Department, Stanford University, Stanford, CA 94305, USA
*Correspondence to: Adrien Treuille (treuille@cs.cmu.edu, 206-409-8246) and Rhiju Das (rhiju@stanford.edu, 650-723-5976)

This Supporting Information Document contains the following sections:

Supporting Methods

Components of the online interface.
EteRNA offers three kinds of puzzles. First, six tutorials (Figure S3D) introduce RNA design without assuming prior scientific knowledge. This entire series can be completed in 5-10 minutes. Second, challenges (Figure S3E) teach advanced folding concepts, such as nearest-neighbor base-stack energies (1); each challenge can be solved in approximately 1 to 10 minutes, and provides users with a continuous source of reward and competition. Both tutorials and challenges are *in silico* puzzles without experimental validation, designed to familiarize participants with RNA design. Participants are rewarded with lab points for clearing each tutorial and challenge.
Once a participant achieves 10,000 lab points, he/she can participate in lab competitions (Figure S3F) which integrate high-throughput experimental feedback into the problem solving. Unlike tutorials and challenges, participants win lab puzzles by designing RNA sequences that fold accurately *in vitro*, as assessed by experiments.
Every week, participants use the *voting interface* to vote for 8 sequences amongst those submitted which will be synthesized and experimentally scored (Figure S3H). During voting, participants can comment on any design and vote for up to 8 designs per round. The 8 selected sequences are synthesized and chemically mapped via the high-throughput synthesis pipeline (see below) and scored on a 0 to 100 scale. Once 8 sequences are experimentally scored, residue-by-residue reactivity data are presented to participants in the *results viewer* (Figure S3J). The *results viewer* renders sequences in the same representation used in the *sequence design interface* but with a different color scheme (nucleotides are colored in a blue-to-white-to-yellow spectrum representing low-to-intermediate-to-high reactivity), melding EteRNA's casual visualization with a standard colorings used in the expert literature on RNA folding and design (2, 3). This *results viewer* can also supply an estimate of the most likely single secondary structure given the SHAPE data, using pseudoenergy-directed modeling (4). EteRNA rewards participants who created or voted for successful designs, with lab points allotted in proportion to the structure mapping score. All participants can access the entire history of EteRNA designs through the *results browser* (Figure S3I) which supports global sorting by score, author, and properties including calculated melting points and G–C content.
Finally, participants are encouraged to submit their own lab design rules to the *design rule collection* (Figure S3K). Participants post to the *design rule collection* their individual methods of designing "good" sequences that are likely to work *in vitro;* submitted design rules are coded into scoring functions by EteRNA developers. Any participant can access the original posts and see how the corresponding scoring function scores past synthesized designs. 40 submitted design rules from the *design rule collection* were used to create *EteRNABot* (see Materials and Methods).
A history of the EteRNA project, from its release to present, is available as an RSS feed at: http://eterna.cmu.edu/rss/news.

Designs from prior automated algorithms. We carried out comparisons to designs from the NUPACK and RNAInverse algorithms. Two other available automated methods, RNA-SSD (5) and INFO-RNA (6) generally gave designs with strings of four or more G's, which pose difficulties for nucleic acid synthesis.

We used the online version of NUPACK (7) design available at NUPACK design server (8), which minimizes the ensemble defect predicted for a sequence. The following settings were used:

```
s.stop = 1.0
material = rna1999
prevent = GGGG, CCCC
```

To generate *N* sequences we always generated *10N* sequences with NUPACK and chose *N* sequences with the lowest normalized ensemble defects.

We compiled the official ViennaRNA (9) package (version 1.8.5) available on its official website (10) and used the default setting to run RNAInverse, which seeks a sequence whose minimum free energy structure is the target structure. In both NUPACK and RNAInverse design packages, we used an energy function consistent with ViennaRNA 1.8.5 package, as was also used in the EteRNA interface. We also tested later updates to these energy functions, but did not give significantly improved predictive power for the EteRNA experimental results (see Supporting Results and Figure S10).

EteRNABot-alt and EteRNABot-control. Aside from EteRNABot, which used 5 participant-proposed design rules, we tested EteRNABot-alt which utilized all 40 design rules. EteRNABot-alt involved linear regression with L2 regularization (17) to generate the score predictor. As a further control for our algorithm comparisons, we trained and tested a simplified 'EteRNABot-control' that only used RNA design features that expert developers had made available as part of the sequence browser interface: percentage of GC pairs, pairing probabilities, predicted minimum free energy, and melting point (Table S6).

Statistical cross-comparison of design methods. We carried out statistical comparisons of each pair of design methods (participants, RNAInverse, NUPACK, and EteRNABot algorithms) and report associated *p*-values in Table S3. The comparison is based on the non-parametric bootstrap. For each pair of methods, we randomly sampled 4 scores from each method and measured the difference between the 75% quantile scores to avoid outliers. After 20,000 rounds of sampling, we report the *p*-values given by the Wald hypothesis test that the difference is positive (Table S3A-E). While we used the 75% quantile comparison through the main text, we also calculated *p* values comparing mean (Table S3F-J) and median (Table S3K-O) values. Both gave consistent results with the 75% quantile comparison.

Supporting Results

Five independent lines of evidence confirm structure mapping results. Structure mapping data from the EteRNA training period (Phase I) indicated that the EteRNA community learned from experimental data, eventually outperforming existing design algorithms by a substantial margin (main text, Figure 2). By later challenges, at least one of 8 first-round designs from participants gave data that were indistinguishable from data expected for fully accurate folds. While these data provided single-nucleotide resolution evaluations of the RNA designs, we sought additional evidence for our conclusions from independent challenges, modeling tests, experimental data. We carried out five additional sets of experiments:

1. *Four additional 'simple' design challenges*. In principle, the improvement of EteRNA participant designs compared to automated algorithm designs could have arisen not from learning by the EteRNA community but instead through some special features of the target structures that were especially unfavorable for participants in early targets and then unfavorable for algorithms in later targets. To rule out such biases for or against EteRNA participants or prior algorithms, we carried out four additional design challenges on distinct 'simple' targets that contained a similar number of stems as the first three design targets (Figure S1). In two of these cases, NUPACK designs were able to achieve highly accurate designs (structure mapping scores indistinguishable from 100, given the experimental error of ±5; Figure S5 and Table S2), disfavoring bias in target selection against automated methods. Furthermore, in all four cases, first-round designs from EteRNA participants continued to outperform NUPACK and RNAinverse algorithms (Figure S5, Tables S2 and S3), confirming the results of the training period and disfavoring bias of simple challenges against EteRNA participants.

2. *Automated structure modeling.* EteRNA's continuous experimental assessment of RNA designs leverages the empirical correlation of a nucleotide's structure mapping data to its structural accessibility or flexibility. These data and the overall structure mapping scores are therefore independent of algorithms or energy functions used to model or design secondary structure, as is necessary for a rigorous evaluation of each design's *in vitro* folding accuracy. Nevertheless, we separately carried out tests based on methods for harnessing structure mapping data to automatic modeling algorithms (4, 11, 12). Because these data-directed modeling methods are biased by their underlying energy function, they can recover the target structure for a given design even if the data conflict with the assumed fold. However, if the data drive the modeling away from the target structure for a particular design, this can be taken as evidence against the folding accuracy of the design. Figure S6 shows the results of SHAPE-directed modeling for the top designs produced by each method through the six challenges of the EteRNA Phase I training period. For five designs created by the ViennaRNA algorithm and for two designs created by the NUPACK method, the SHAPE-directed model does not match the target

structure (red symbols). For all participant designs, however, the target structure is recovered by the data-directed modeling. These results lend additional support to the conclusion that the participant designs outperformed sequences from prior algorithms. We obtained similar rankings of methods from SHAPE-directed modeling of designs by EteRNA participants, the EteRNAbot algorithm, NUPACK, and RNAInverse in the later testing period (Figure S9 and Table S3).

3. *DMS modification* As a cross-check on the SHAPE methodology, we applied an alternative structure mapping method [DMS, dimethyl sulfate alkylation (11, 13)] to several designs. For example, Figure S7 displays DMS data as well as replicate SHAPE data for the highest scoring NUPACK design and highest scoring participant design for the 'Bulged Star' target structure. For the NUPACK design, nucleotides that should be sequestered in helices in the target structure were exposed to DMS and/or SHAPE modification, while the participant design gave both DMS and SHAPE data consistent with the target structure at all probed nucleotides.

4. *Information rich validation from the mutate-and-map technique.* The two-dimensional mutate-and-map technique gives an information-rich readout of RNA pairing interactions that is more accurate in *de novo* modeling of secondary structure than standard structure mapping data in benchmarks and blind tests (14). The premise of this technique is that mutation of each nucleotide involved in a base pair can potentially expose not only that nucleotide but also its pairing partners to chemical modification. Therefore, coupling systematic mutagenesis to chemical mapping can enable the detection of helices in a model-free manner, and coupling to secondary structure algorithms gives highly accurate models (14). While the current cost of the method precluded its application to all EteRNA designs, we acquired mutate-and-map data for the highest scoring NUPACK design and highest scoring participant design for the 'Bulged Star' target structure (Figure S8; see also above and Figure S7). On the highest scoring participant design, this independent method confirmed the accurate folding of each helical stem with $\geq$ 98% support values, as assessed by non-parametric bootstrapping (15) (Figure S8B, D). In contrast, the top-scoring NUPACK design for the same target structure gave data consistent with an ensemble of different structures (Figure S8A, C). Thus, while NUPACK explicitly minimizes the predicted heterogeneity of RNA structure ensembles (16) through an elegant search procedure (7), errors in available computational models appear to hinder its consistent use, especially for complex RNA shapes.

5. *Replicates with alternative techniques and by separate experimenters*. As a test of the reproducibility of the data, replicate measurements for the more difficult target shapes (the five-helix junction “Bulged Star” and the dendrimer shape “The Branches”) were carried out by different experimentalists (D.C. and W.K.). We observed results indistinguishable from those of the initial experiments (Figures S2A-C). Furthermore, after the EteRNA training period, Lucks and colleagues pioneered an alternative readout for SHAPE-based structure mapping using Illumina next-generation-sequencing (3). The SHAPE-seq method unfortunately carried biases

from PCR amplification and adapter ligation steps that precluded exact agreement with capillary electrophoresis, so we could not immediately adopt this technique for EteRNA experimental tests. However, in initial tests on 30 EteRNA designs, SHAPE-seq gave structure mapping scores that were consistent with capillary electrophoresis data within the estimated error of the methods (Figure S2D).

EteRNABot variants. EteRNABot-alt involved reweighting all 40 submitted rules, using L2 regularization rather than L1 regularization (which yielded a sparse rule set for EteRNABot). For all design targets in Phase II, we evaluated an equal number of designs from EteRNABot-alt as from EteRNABot. The alternative algorithm did not show a consistent difference from EteRNABot in its predictive power for experimental measurements (Figure S13, Tables S2 and S3). In particular, while EteRNAbot nominally out-ranked EteRNAbot-alt throughout Phase II, *p* values did not give statistical confidence to this ranking (e.g., $p = 0.31$ and 0.34 for 75^{th} percentile comparisons of puzzles 11-15 and 16-19, respectively; Table S3). Due to its parsimonious representation as five rules, we highlighted EteRNABot in the main text, but EteRNABot-alt is a viable alternative that can be tested in further applications and is available on the on-line server.

EteRNABot-control involved reweighting scores that were made available as columns in the sequence browser and therefore were not novel features proposed by participants. While designs from EteRNABot-control scored better than RNAInverse, its performance was significantly poorer than both EteRNABot and EteRNABot-alt (Figure S13 and Table S3).

Tables S6, S7, and S8 list strategies used by each EteRNABot variant and their weights in the score predictor. Table S3 gives pair-wise statistical comparisons of all design methods tested**.**

Predictive power of more recently developed energy functions. After the release of EteRNA, several RNA secondary structure modeling packages released modified energy functions with potentially greater accuracy at predicting the *in vitro* folding accuracy of designs. While future versions of EteRNA will include these alternative energy functions *in silico*, we chose to keep the computational energy function in the interface fixed for this study, and instead assessed the predictive power of the newer energy functions *post facto*.

We compared the predictions of the newer energy functions to the experimental structure mapping data for EteRNA designs. We calculated the base-pairing probabilities for 190 synthesized designs from Phase II with NUPACK (7), RNAfold from a recently updated version of ViennaRNA (9) (ver 2.0), and the most recent version of RNAstructure (18) (ver 5.3). The ensemble defect, normalized to the number of residues, summarizes the partition fraction of conformations in which each base is in its target base pair (or in an unpaired state, if that is the target). This metric provides an intuitive and robust summary score of the match of the base

pairing probabilities to the desired structure (7), and could be calculated based on base pairing probability matrices outputted by all tested software packages. We assessed whether choosing top-scoring designs by these metrics enriched for designs with structure mapping scores above 90, which constituted 27% of the designs (51 of 190). Focusing on the top 9 designs (top 5%) with lowest normalized ensemble defects (Figure S10A), all 3 algorithms selected 3 to 4 designs (33-44%) with structure mapping scores lower than 90, suggesting a modest enrichment. In comparison, EteRNABot and EteRNABot-alt (see section above, *EteRNABot variants*) outperformed these previous algorithms (Figure S10A), giving 7 of 9 (EteRNABot, 77%) and 8 of 9 (EteRNABot-alt, 88%) designs with structure mapping scores over 90. EteRNABot-control performed comparably to NUPACK (44%). We also applied the test to more samples, the top 19 (top 10%; Figure S10B), top 38 (top 20%; Figure S10C), and top 57 (top 30%; Figure S10D) designs selected by each scoring method. The relative predictive power of the algorithms remained the same, with EteRNABot and EteRNABot-alt giving clearly better enrichment of top-scoring designs compared to the most recent versions of NUPACK, RNAfold, and RNAstructure.

Relationship of small-molecule binding affinity to fraction of conformations presenting aptamer in correct secondary structure. Measuring the equilibrium affinity of a small molecule (here, flavin mononucleotide, FMN) to its RNA aptamer sequence provides a thermodynamic readout of folding accuracy – specifically, the partition fraction of RNA conformations that present the aptamer in the correct secondary structure. The partition function for the system is as follows:

$$Z = 1 + K_{\text{incorrect}} + K_a^0[\text{FMN}],$$

where the terms represent all conformational states presenting correct aptamer structure, all states with secondary structure incompatible with the aptamer, and all states with FMN bound to a correct aptamer structure, respectively.

The 'intrinsic' association constant K_a^0 is that determined for a minimal aptamer system, and has units of inverse concentration. In our experimental conditions (50 mM Na-HEPES, pH 8.0; 10 mM $MgCl_2$), we measured $K_a^0 = 3 \pm 1 \times 10^6$ M^{-1}. The parameter $K_{\text{incorrect}}$ is the (unit-less) equilibrium ratio between incorrect and correct conformations in the absence of FMN. We have assumed throughout that the concentration of RNA is negligible compared to FMN, as is the case in our experiments ([RNA] < 60 nM, compared to typical FMN binding midpoints of ~1 μM or higher). We have further assumed that in RNA conformational states with the correct secondary structure, the nature of base pairs that are outside the aptamer sequence do not influence the aptamer's intrinsic binding affinity. Tests changing the base pairs neighboring the aptamer sequence in a simple hairpin construct gave smaller than 2-fold effects, less than the error of the

titrations below. We wish to estimate the fraction of correct secondary structures in the absence of FMN, $f_{\text{correct}} = 1/\left(1+K_{\text{incorrect}}\right)$.

Consider an FMN titration experiment; the fraction of RNA conformations with FMN bound will be

$$\text{Fraction bound} = \frac{K_a^0[\text{FMN}]}{1+K_{\text{incorrect}}+K_a^0[\text{FMN}]} = \frac{K_a[\text{FMN}]}{1+K_a[\text{FMN}]}$$

where the apparent dissociation constant is:

$$K_a = K_{\text{a}}^0 \frac{1}{1+K_{\text{incorrect}}} = K_a^0 f_{\text{correct}}$$

Thus, the ratio of the measured K_a to the intrinsic K_{a}^0 of the isolated aptamer reports on the partition fraction f_{correct} of RNAs with correct aptamer structure. For example, if K_a is measured to be 100-fold lower than K_{a}^0, the fraction of the RNA's conformations that fold into secondary structures that correctly present the aptamer is 1%. Further, K_a should always be within experimental error or less than K_{a}^0. This bound is experimentally satisfied for all FMN-aptamer designs synthesized herein.

Supporting References

1. Mathews DH, Sabina J, Zuker M, & Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288(5):911-940.
2. Russell R, *et al.* (2006) The paradoxical behavior of a highly structured misfolded intermediate in RNA folding. *J Mol Biol* 363(2):531-544.
3. Lucks JB, *et al.* (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci USA* 108(27):11063-11068.
4. Deigan KE, Li TW, Mathews DH, & Weeks KM (2009) Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci USA* 106(1):97-102.
5. Andronescu M, Fejes AP, Hutter F, Hoos HH, & Condon A (2004) A new algorithm for RNA secondary structure design. *J Mol Biol* 336(3):607-624.
6. Busch A & Backofen R (2006) INFO-RNA--a fast approach to inverse RNA folding. *Bioinformatics* 22(15):1823-1831.
7. Zadeh JN, *et al.* (2011) NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem* 32(1):170-173.
8. Bois J, *et al.* (2012) NUPACK: Nucleic Acid Package.
9. Hofacker IL (2004) RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics* Chapter 12:Unit 12 12.
10. Hofacker I (2007) Vienna RNA Secondary Structure Package.
11. Cordero P, Kladwang W, VanLang CC, & Das R (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* 51(36):7037-7039.
12. Washietl S, Hofacker IL, Stadler PF, & Kellis M (2012) RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res* 40(10):4261-4272.
13. Tijerina P, Mohr S, & Russell R (2007) DMS footprinting of structured RNAs and RNA-protein complexes. *Nat Protoc* 2(10):2608-2623.
14. Kladwang W, VanLang CC, Cordero P, & Das R (2011) A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem* 3(12):954-962.
15. Kladwang W, VanLang CC, Cordero P, & Das R (2011) Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* 50(37):8049-8056.
16. Zadeh JN, Wolfe BR, & Pierce NA (2011) Nucleic acid sequence design via efficient ensemble defect optimization. *J Comput Chem* 32(3):439-452.
17. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* 58(1):267-288.
18. Reuter JS & Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11:129.
19. Das R, Karanicolas J, & Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods* 7(4):291-294.
20. Frederiksen JK, Li NS, Das R, Herschlag D, & Piccirilli JA (2012) Metal-ion rescue revisited: Biochemical detection of site-bound metal ions important for RNA folding. *RNA* 18(6):1123-1141.

Figure S1. EteRNA lab competition design targets. All of the secondary structure targets for lab competitions. The gray bases indicate fixed sequences for synthesis and structure mapping (see Materials and Methods). Puzzles 1-6 belong to Phase I (initial training period). Puzzles 7-10 were carried out to rule out bias for or against EteRNA participants. Puzzles 11-19 belong to Phase II (testing period).

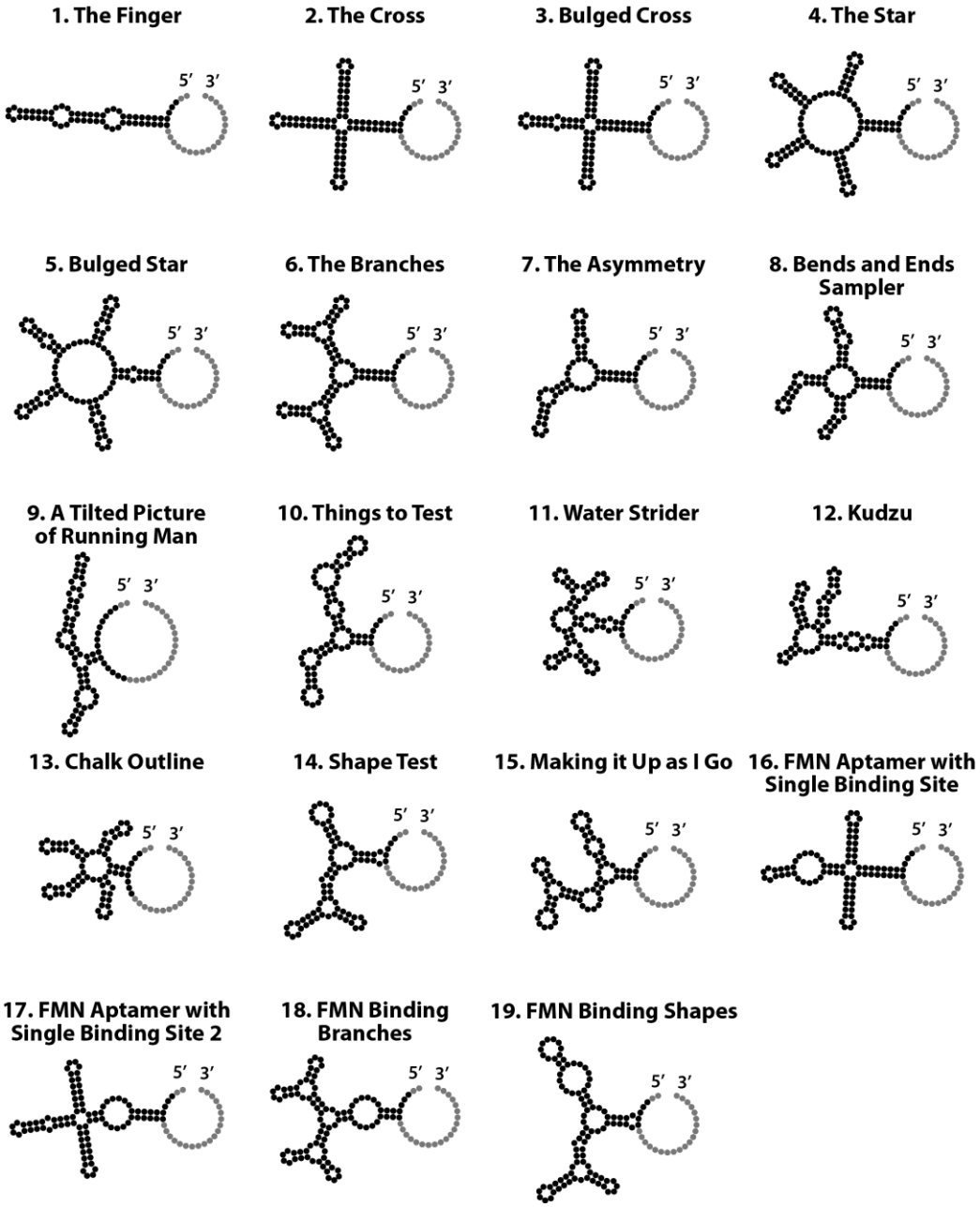


Figure S2. Reproducibility of the experimental EteRNA score. (A) Replicates of SHAPE structure mapping data by different experimenters agree within estimated error of capillary electrophoresis (CE) quantification for the top EteRNA design for the 'Bulged Star' target structure (target 5 in Figure S1). Pink lines show high/low reactivity scheme expected for the target structure (see also Figure S7). (B) Replicates of SHAPE measurements using two different readouts, capillary electrophoresis and an Illumina-based SHAPE-seq protocol, agree within estimated error for the top EteRNA design for the 'Finger' target structure (target 1 in Figure S1). (C) Comparison of CE-based structure mapping scores for independent replicates for 48 designs give a mean unsigned error of 5 score units. (D) Comparison of structure mapping scores from CE data to scores from Illumina-based SHAPE-seq data. Note that the error bars on the SHAPE-seq data, which reflect statistical error, are larger than those for the CE data.

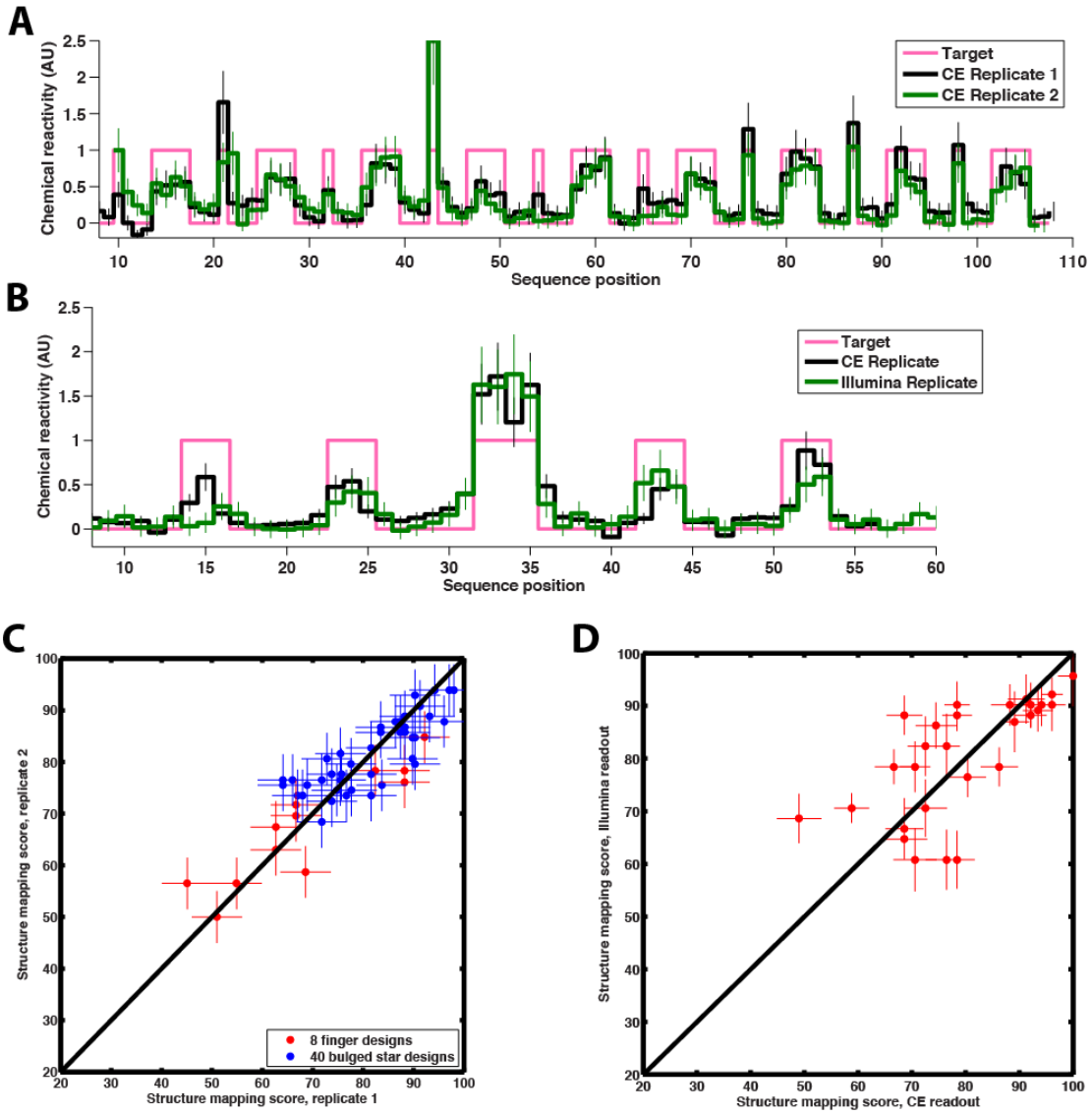


Figure S3. (following page) Overall platform design. A typical participant's progression through the EteRNA interface. When a participant logs in to EteRNA from the *splash page* (A), he/she lands on the *action page* (B), which summarizes the participant's activity and to-do list. The participant can immediately access community features (C) such as forum or participant rankings. The participant clears tutorials (D) to learn the basics of RNA secondary structure design. Then he/she engages in various *in silico* challenges (E) designing various RNA structures. Finally, when the participant earns 10,000 lab points he/she can enter the *RNA lab* (F) to participate in *in vitro* design challenges. The participant submits a sequence to the lab challenge using the *sequence design interface* (G). Each week the participant uses the *voting interface* (H) to review other participant designs and votes for the 8 best designs to be synthesized. Synthesized results are published back to participants through the *results browser* (I) and the *results viewer* (J). Participants study the experimental results and improve their hypotheses and strategies. Expert participants share and test their own lab design rules in the *design rule collection* (K).

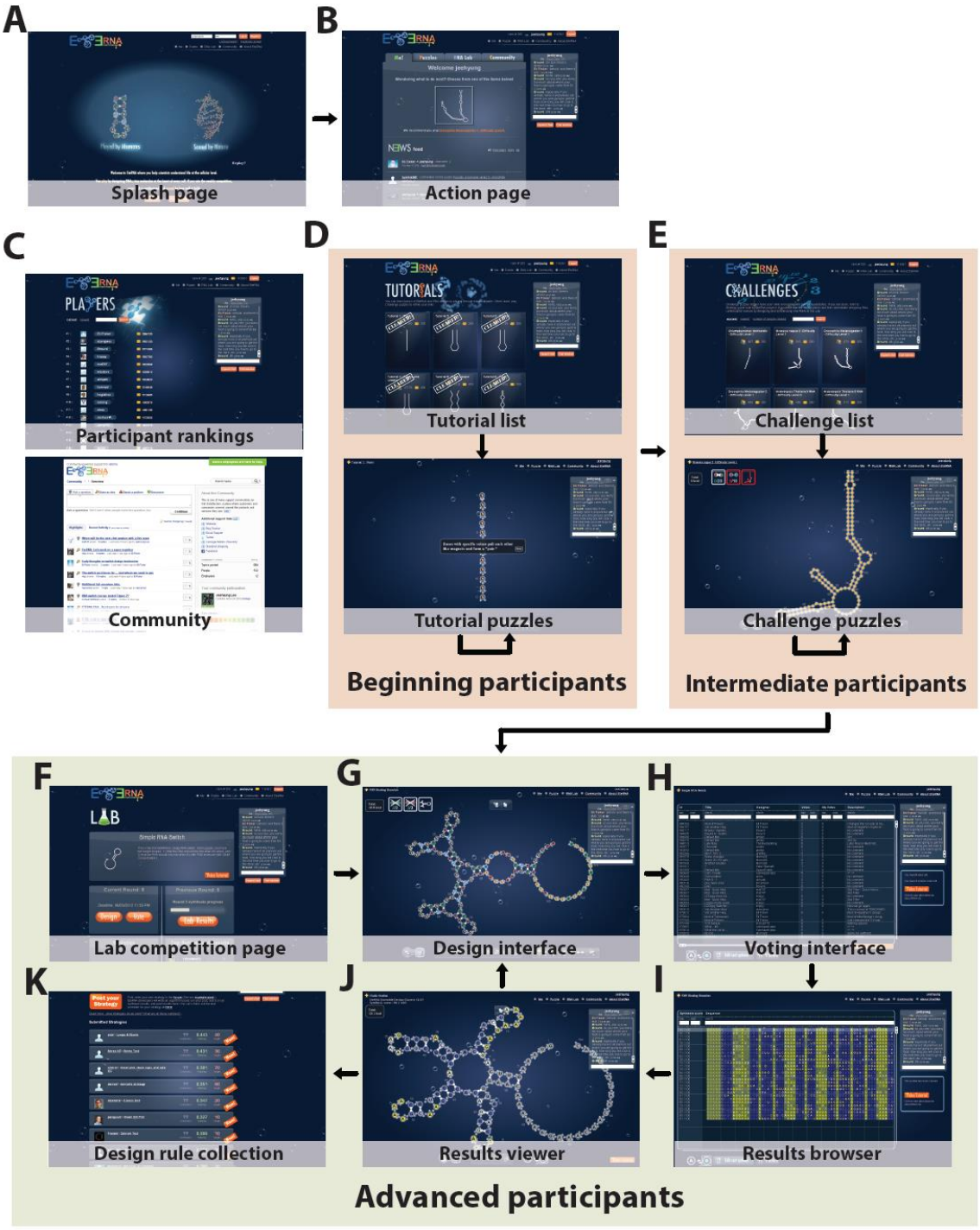

A
Splash page
B
Action page
C
Participant rankings
Community
D
Tutorial list
Tutorial puzzles
Beginning participants
E
Challenge list
Challenge puzzles
Intermediate participants
F
Lab competition page
G
Design interface
H
Voting interface
K
Design rule collection
J
Results viewer
I
Results browser
Advanced participants

Figure S4. Number of participants passing milestones. Number of EteRNA participants at different milestones (at end of Phase II, the testing phase; Figure 4 in main text).

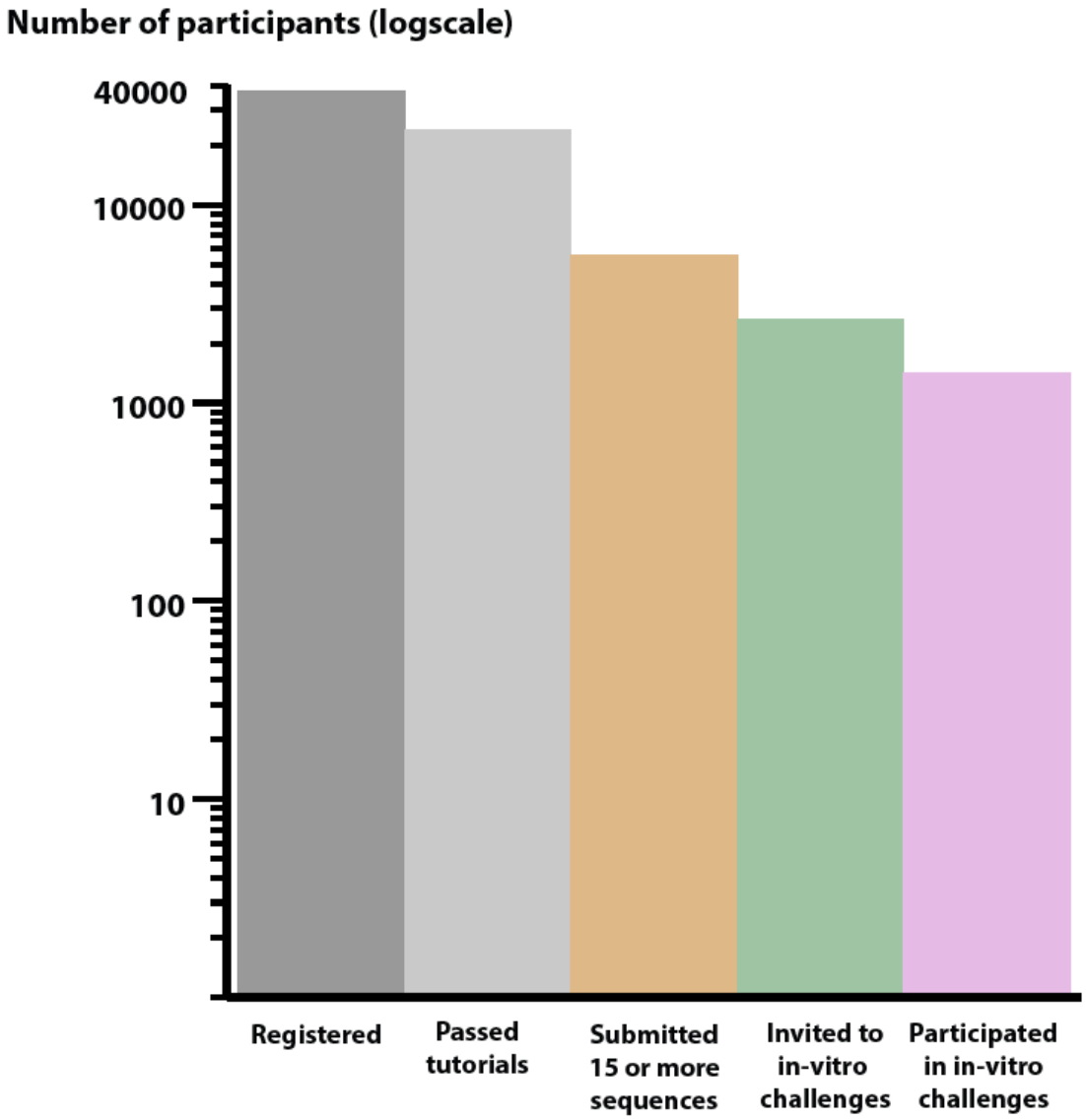


Figure S5. (following page) Structure mapping scores and nucleotide-by-nucleotide chemical accessibility data for four simple design challenges carried out between Phase I and Phase II. Figure details are the same as in Figure 2 of the main text.

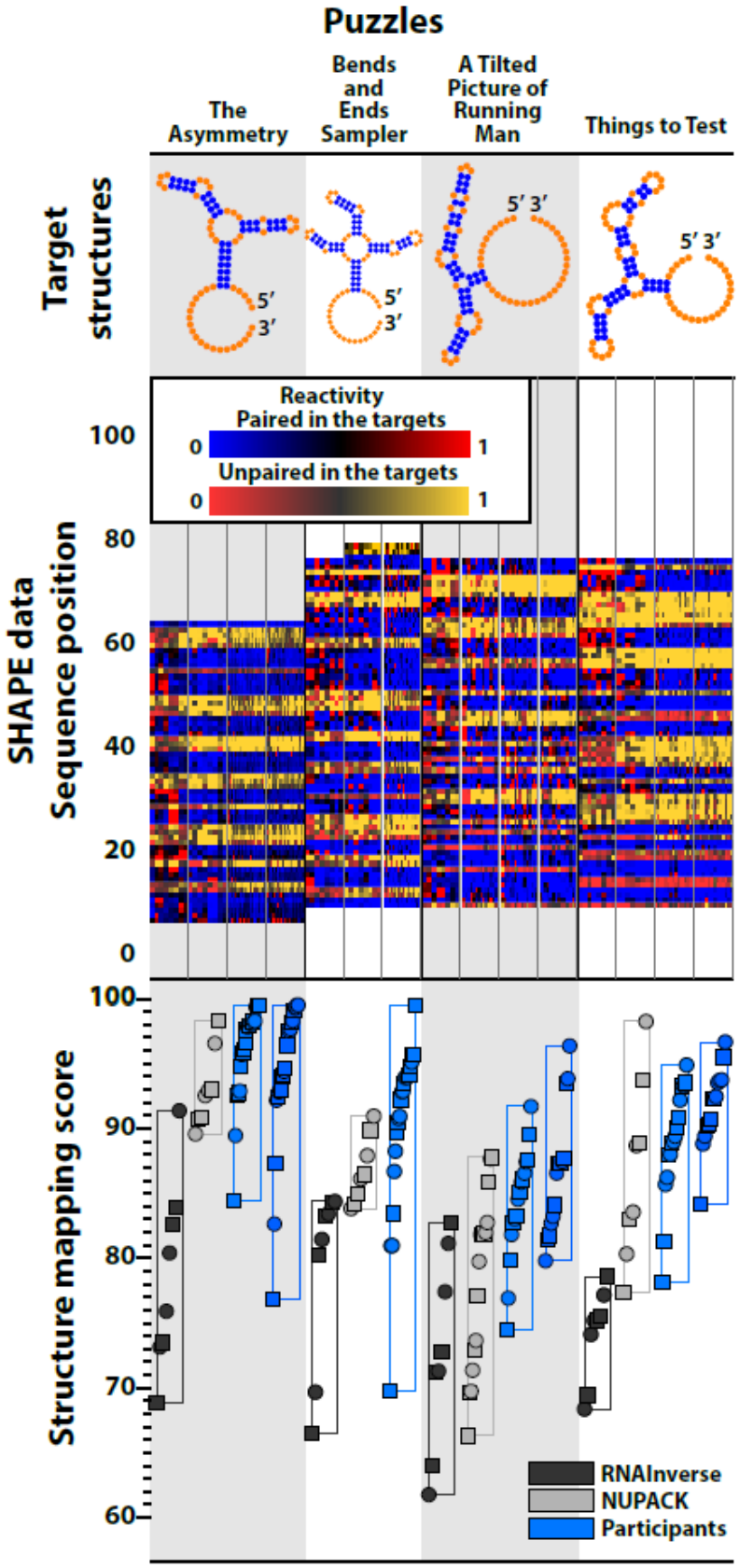

Puzzles
The Asymmetry
Bends and Ends Sampler
A Tilted Picture of Running Man
Things to Test
Target structures
5′
3′
5′ 3′
Reactivity
Paired in the targets
0
1
Unpaired in the targets
0
1
SHAPE data
Sequence position
100
80
60
40
20
0
Structure mapping score
100
90
80
70
60
RNAInverse
NUPACK
Participants

Figure S6. (following page) SHAPE-directed secondary structure models for the best scoring RNA designs from each design agent (phase I and 4 test puzzles before phase II). Blue, gold, and gray nucleotide coloring indicate low, high, and unmeasured SHAPE reactivity, respectively. The gray lines connect nucleotides that are paired in the target secondary structure and the red outlines indicate bases that do not have the same pairing with the target structure in the predicted fold.

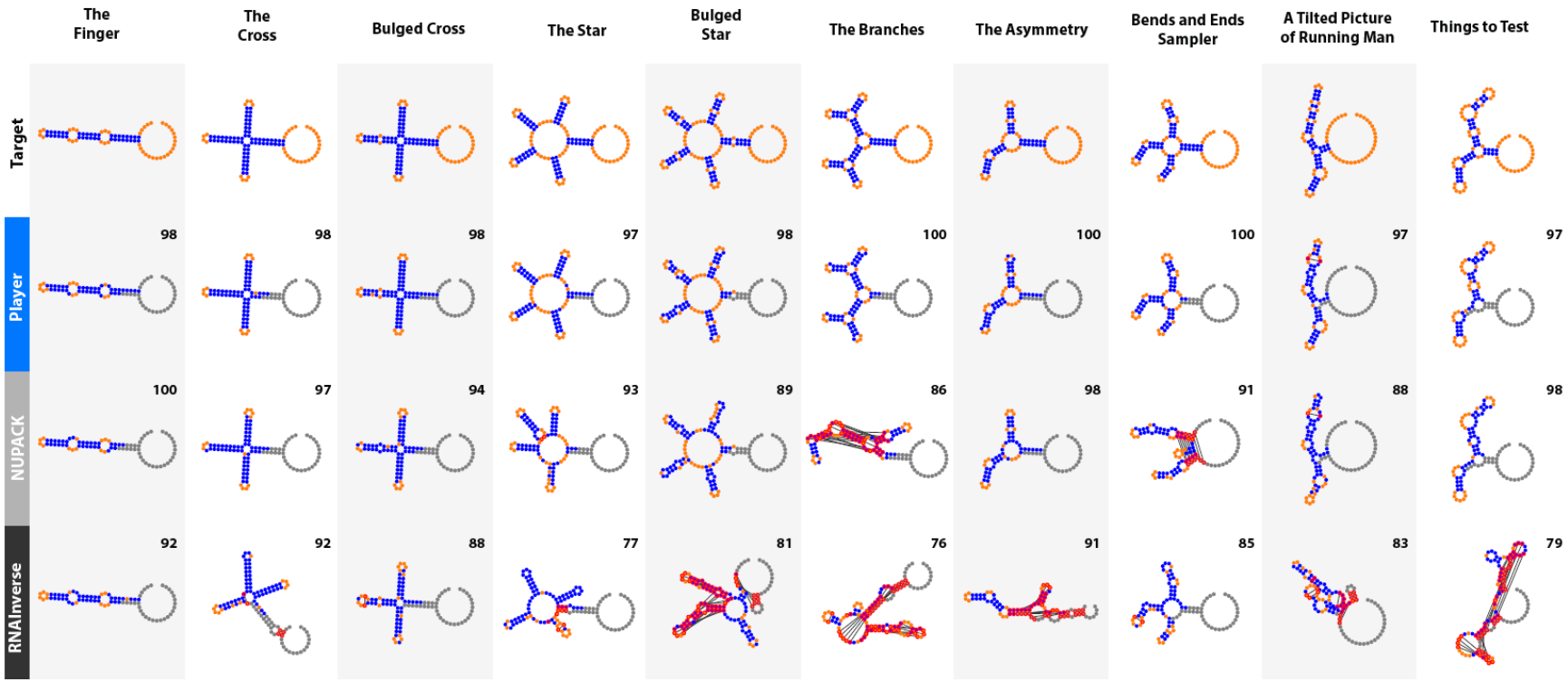
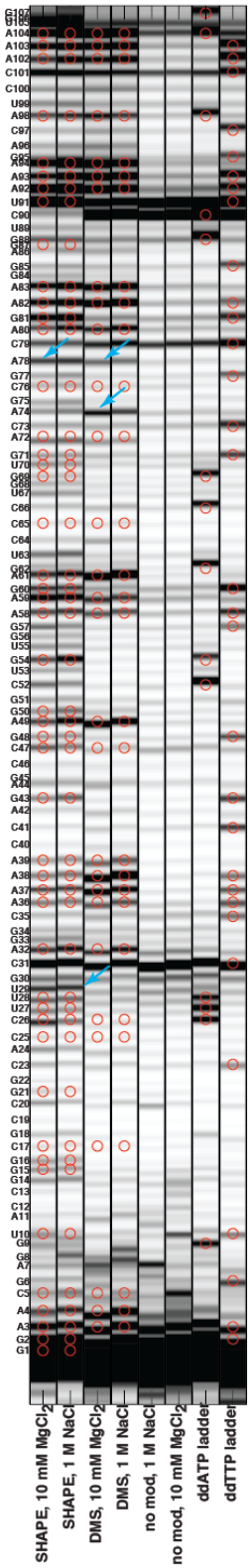

The Finger
The Cross
Bulged Cross
The Star
Bulged Star
The Branches
The Asymmetry
Bends and Ends Sampler
A Tilted Picture of Running Man
Things to Test
Target
Player
NUPACK
RNAInverse
98 98 98 97 98 100 100 100 97 97
100 97 94 93 89 86 98 91 88 98
92 92 88 77 81 76 91 85 83 79

Figure S7. (following page) Alternative chemical interrogation of top NUPACK design and top EteRNA community-created design for "Bulged Star" target shape. Capillary electropherograms are shown, with fluorescence in arbitrary units. The direction of electrophoresis is such that longer products (corresponding to reverse transcription stops closer to the 5´ end) are at the bottom of the figure. The red circles mark nucleotides that should give strong signals if the secondary structure is correct; nucleotides without circles should be protected if the target secondary structure is formed. For the NUPACK sequence, both SHAPE and DMS chemical mapping experiments give reactive nucleotides that should be protected (cyan arrows), in two different folding conditions, 10 mM $MgCl_2$ and 1 M NaCl. Data for the EteRNA community design are consistent with the target secondary structure.

A NUPACK design



B EteRNA community design

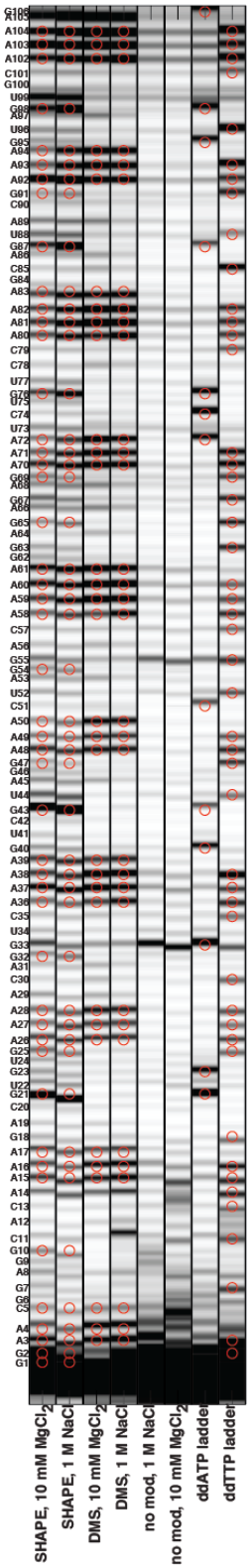


Figure S8. Two-dimensional chemical mapping of top NUPACK design (A,C) and top EteRNA-created design (B,D) for "Bulged Star" target shape. (A,B) Mutate-and-map measurements (14). Each possible single mutation of the sequences (to the complementary nucleotide) was synthesized and subjected to SHAPE chemical mapping (capillary electropherograms of each mutant are shown as horizontal traces). Diagonal features extending from the top-left to the bottom-right (near black line) indicate perturbations near the site-of-mutation. Off-diagonal features signal release of base-pairing partners upon each mutation. Data are displayed as Z-scores (difference to mean reactivity at each probed nucleotide, divided by standard deviation of reactivity across all mutants). Red squares mark base pairs of the target structure; gray squares in (A) mark off-target pairs derived from structure inference using these data. Arrows mark features consistent (red) or inconsistent (cyan) with target design. (C,D) Most probable secondary structure derived from automated modeling using the mutate-and-map Z scores of A and B. Low bootstrap support values (red percentages) for the NUPACK design are consistent with interconversion between several structures. In contrast, high bootstrap values for each helix of the EteRNA human-created design strongly support its accuracy of folding *in vitro*. The nucleotides are colored according to SHAPE reactivity.

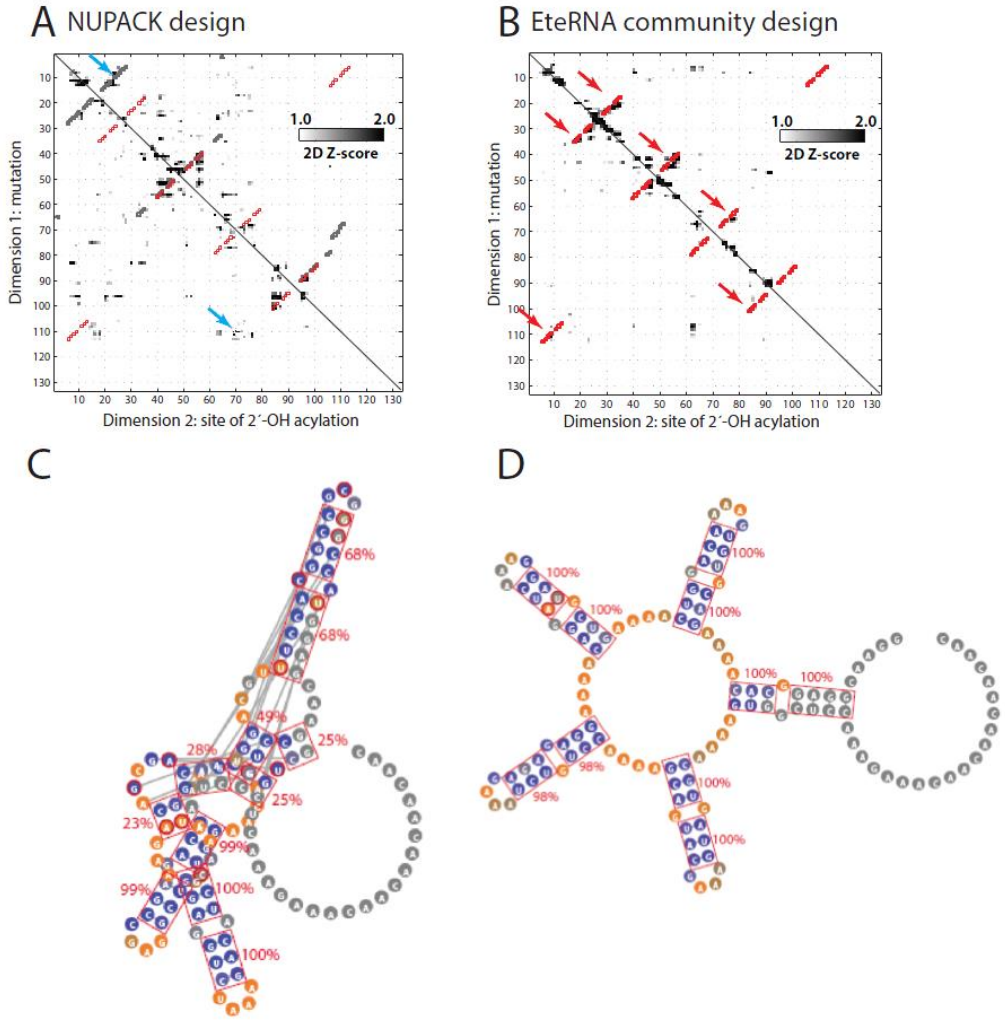


Figure S9. **SHAPE-directed secondary structure models for top designs from each design agent (phase II).** Blue, gold, and gray nucleotide coloring indicate low, high, and unmeasured SHAPE reactivity, respectively. The gray lines connect nucleotides that are paired in the target secondary structure and the red outlines indicate bases that do not have the same pairing with the target structure in the predicted fold.

Target
Participants
EteRNABot (EB)
EB-alt
EB-control
RNAInverse
NUPACK
Water Strider
Kudzu
Chalk Outline
Shape Test
Making it Up as I Go
FMN Aptamer with Single Binding Site
FMN Aptamer with Single Binding Site 2
FMN Binding Branches
FMN Binding Shapes

Figure S10. Structure mapping score distribution of top designs as ranked by RNAfold, RNAstructure, NUPACK, EteRNABot-control, EteRNABot-alt, and EteRNABot. Each algorithm picked the top (A) 9, (B) 19, (C) 38, and (D) 57 designs from 190 synthesized designs. RNAfold, RNAstructure, and NUPACK picked top designs based on normalized ensemble defects. EteRNABot variants picked top designs based on their score predictors.

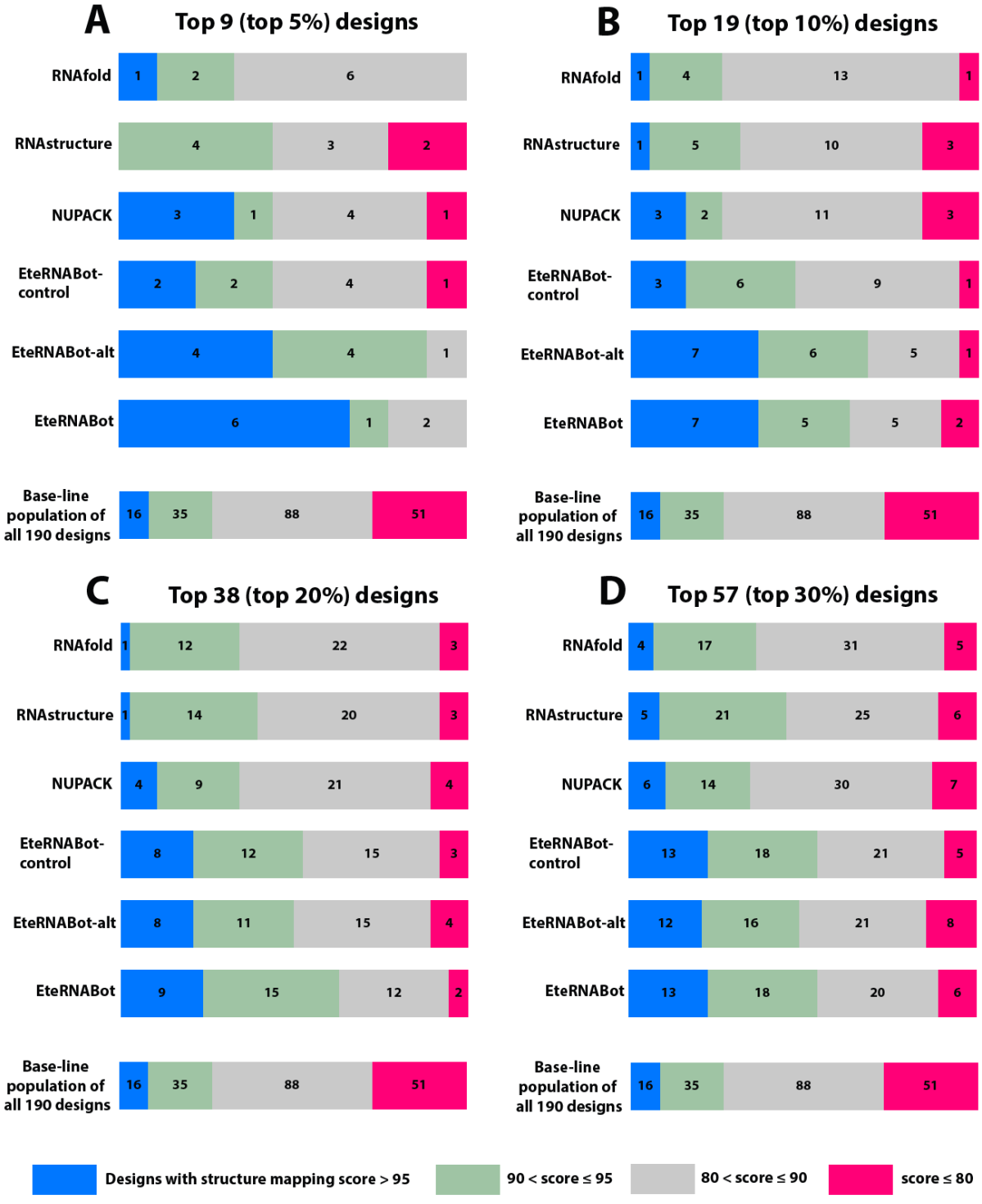


Figure S11. Determination of association constants for binding flavin mononucleotide (FMN). (A,B) DMS reactivity (read out at A and C bases by primer extension) for an EteRNABot FMN-binding-branches design in 10 mM $MgCl_2$, 50 mM Na-HEPES, pH 8.0 without FMN (A) and with 200 μM FMN. The magenta square marks nucleotide A15. This base's reactivity could be experimentally measured for all designs and showed strong protections upon FMN binding (yellow to blue coloring). (C) Capillary electropherograms used to read out DMS reactivities as a function of the FMN concentration; decrease in reactivity between 2 to 15 μM FMN is apparent at A15. (D) Likelihood-based fit (19, 20) (magenta curve) gives an FMN binding midpoint of 9 μM, with approximate errors of 2-fold (black curves). A fit was carried out to the entire data set at all nucleotides; the magenta symbols show reactivity at A15, rescaled to transition from 0 to 1 to reflect the fraction of RNA in the FMN bound state.

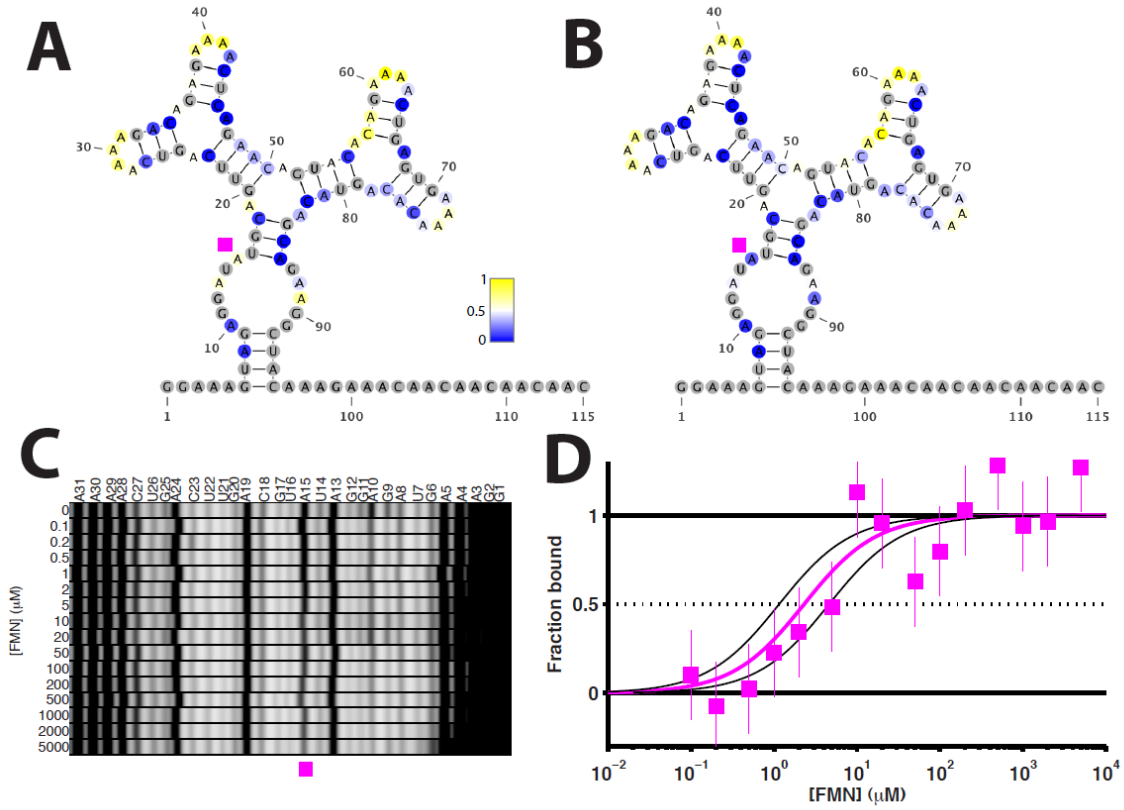


Figure S12. Ensemble defects calculated with the RNAfold energy function for all sequences submitted in phases I and II.

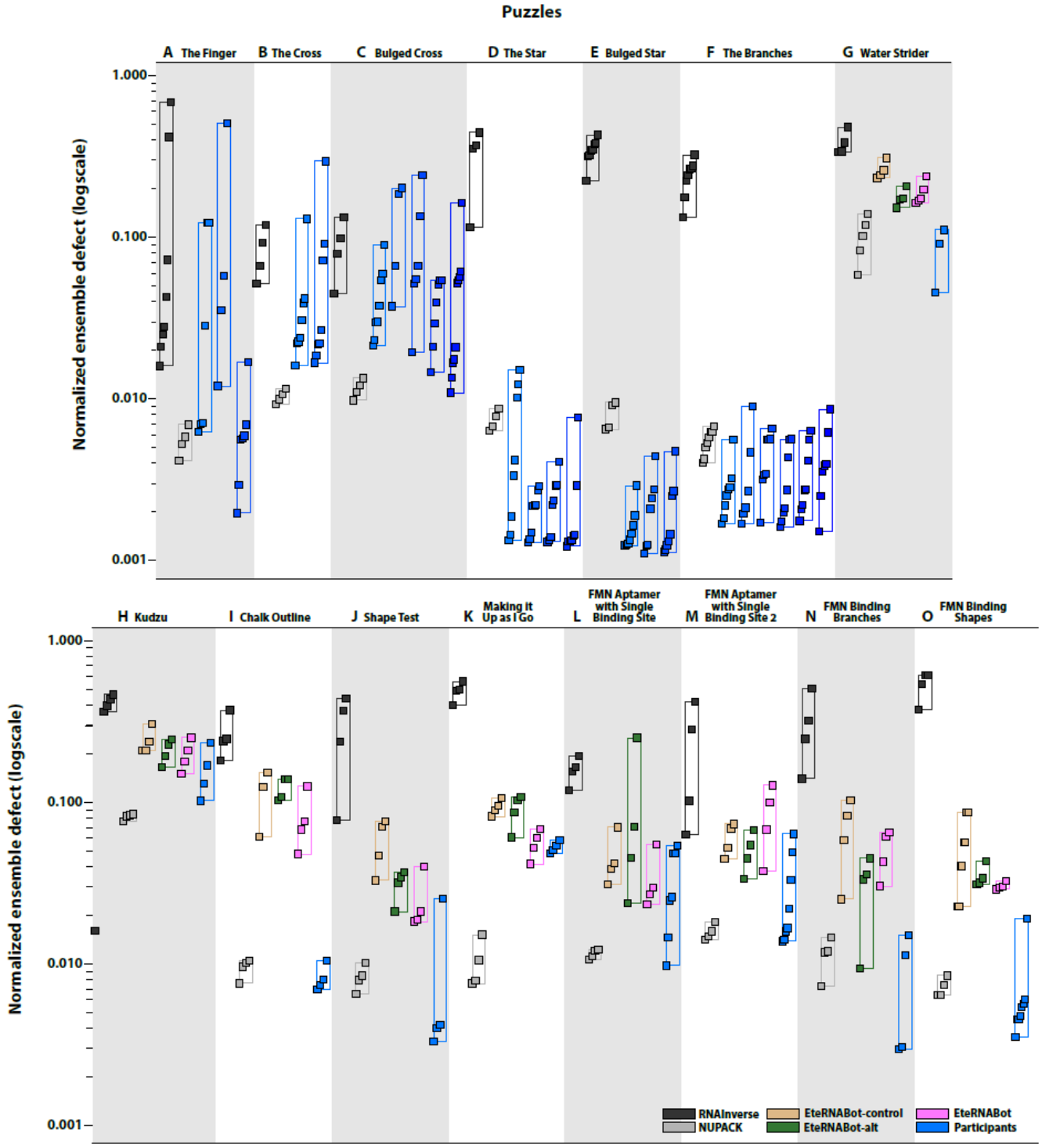
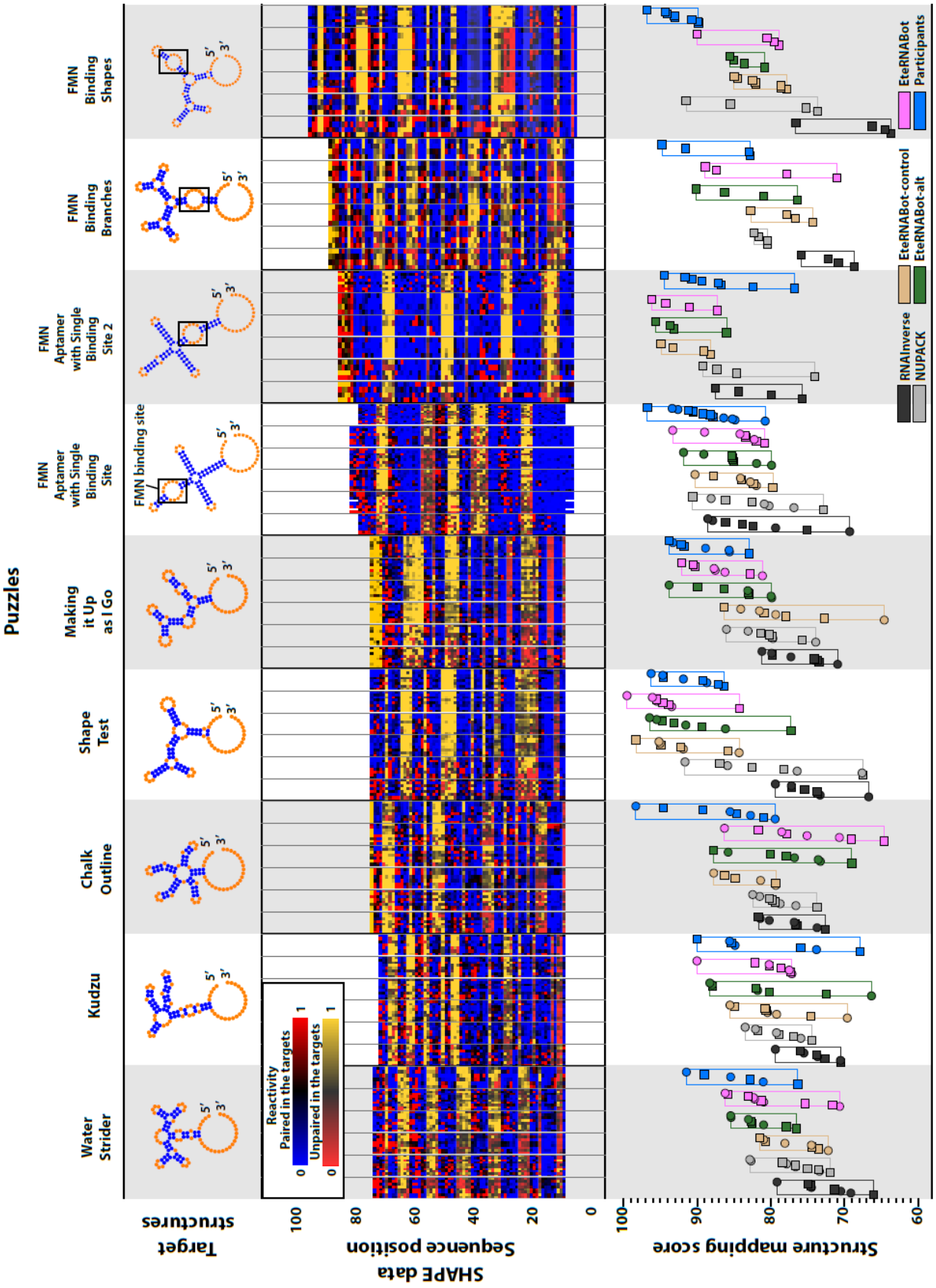


Figure S13. (following page) Structure mapping scores and nucleotide-by-nucleotide chemical accessibility data for Phase II puzzles including EteRNABot variants.


Puzzles
Water Strider
Kudzu
Chalk Outline
Shape Test
Making it Up as I Go
FMN Aptamer with Single Binding Site
FMN Aptamer with Single Binding Site 2
FMN Binding Branches
FMN Binding Shapes
FMN binding site
5′
3′
Target structures
Reactivity
Paired in the targets
Unpaired in the targets
0
1
SHAPE data
Sequence position
100
80
60
40
20
0
Structure mapping score
100
90
80
70
60
RNAinverse
NUPACK
EteRNABot-control
EteRNABot-alt
EteRNABot
Participants

Tables

Table S1. EteRNA lab competition design target secondary structures and design constraints. N denotes nucleotide that needs to be designed.

Title	Secondary structure
The Finger	.....((((((((...((((((...((((((....))))))...))))))...)))))))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NAAAGAAACAACAACAACAAC
The Cross	.....(((((((((((((((((....))))))))(((((((((....)))))))))((((((((....))))))))))))))))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
Bulged Cross	.....(((((((((((((((((....))))))))((((.((((....)))).))))((((((((....))))))))))))))))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
The Star	.....(((((((....((((((....))))))....((((((....))))))....((((((....))))))....((((((....))))))....)))))))................. ... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAAC AAC
Bulged Star	.....((((.(((....(((.(((....))).)))....(((.(((....))).)))... .(((.(((....))).)))....(((.(((....))).)))....))).))))....... .... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAA CAACAACAACAAC
The Branches	.....(((((((..(((((.((((....)))).((((....)))).)))))..(((((.((((....)))).((((....)))).)))))..))))))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
The Asymmetry	.....(((((((..(((.(((....))).)))...((((...((((....)))).)))). ...))))))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNAAAGAAACAACAACAACAAC
Bends and Ends Sampler	.....((((((..((((..(((.....))).))))..((((..((((....)))))))). .((((((...))).)))..)))))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
A tilted picture of running man	..........((((((.(((.((.(.(((...))).).)).)))...))).((((..((((....))))....)))))))......................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

	NNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
Things to Test	.....((((.(((..(((..((.((......)).))......)))..))).((((....((((.......)))).))))..)))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
Water Strider	.....((.((..(((((.(((....)))(((....)))))).....((((((....)))(((....))).)))))..)).))..................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
Kudzu	.....(((.(..(.(((((.((((.((...)))).))))((((((...)).))))..(((...)))...))).)..).)))..................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
Chalk Outline	.....((((.(((((....)).))).(((.(((....)))))).((((((....)).)))).(((((....))).)))))))....................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
Shape Test	.....(.((((..(((((........)))))..(((.(((.(((((.....))))).((((....)))).))))))..)))).).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
Making it up as I go	.....((((.(((..(((......)))..))).(((.((((.(((.....))).(((... ....))))))).....))).)))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
FMN Aptamer with single binding site	.....(((((((((((((((....)))))))((((......((((....)))).....))))(((((((....))))))))))))))).................... GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGGAUAUNNNNNNNNNNAGAAGGNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC
FMN Aptamer with Single Binding Site II	.....((((((......((((((((((....)))))))((((.((((....)))).)))) (((((((....)))))))))).....)))))).................... GGNNNNNNNNAGGAUAUNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNAGAAGGNNNNNNNAAAGAAACAACAACAACAAC
FMN Binding Branches	.....((((......(((.((((.(((....))).(((....))).)))).((((.(((. ...))).(((....))).)))).))).....)))).................... GGNNNNNNNAGGAUAUNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNAGAAGGNNNNAAAGAAACAACAACAACAAC
FMN Binding Shapes	.....(.((((..((((......(((........)))......))))..(((.(((.(((((.....))))).((((....)))).))))))..)))).).................... GGNNNNNNNNNNNNNNAGGAUAUNNNNNNNNNNNNNAGAAGGNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAAGAAACAACAACAACAAC

Table S2. Performance of all tested design methods. The max, median, and mean scores of all design methods. Each puzzle index corresponds to the numbers in Figure S1.

A

	Puzzle (maximum score)																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Participants	98.0	**98.6**	**98.7**	**97.8**	**98.0**	**100**	**100**	**100**	**97.0**	97.0	**92.3**	**90.5**	**98.5**	97.0	**93.9**	**97.1**	95.0	**95.2**	**97.8**
RNAInverse	92.2	92.5	88.1	77.6	81.6	76.2	91.4	85.1	83.6	79.1	80.0	79.4	81.8	80.3	81.8	88.6	87.5	75.9	77.2
NUPACK	**100**	97.1	94.3	93.2	89.8	86.9	98.3	91.4	88.1	**98.5**	83.1	84.1	83.3	92.4	86.4	90.8	90.0	83.1	92.4
EteRNABot (EB)	-	-	-	-	-	-	-	-	-	-	86.2	**90.5**	86.4	**100**	92.4	93.4	**96.3**	89.2	90.2
EB-alt	-	-	-	-	-	-	-	-	-	-	86.2	88.9	87.9	97.0	**93.9**	92.1	**96.3**	90.4	85.9
EB-control	-	-	-	-	-	-	-	-	-	-	81.5	85.7	87.9	98.5	86.4	90.8	95.0	83.1	85.9

B

	Puzzle (median score)																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Participants	72.5	73.5	84.6	88.7	**90.3**	**92.9**	**96.6**	**92.9**	**85.9**	**91.0**	**84.7**	**85.7**	**85.6**	90.9	**90.9**	**89.3**	90.7	**87.4**	**93.5**
RNAInverse	74.5	84.4	79.1	71.2	72.3	66.7	78.5	82.9	72.4	76.1	73.9	74.6	77.3	75.0	75.8	83.6	82.5	71.7	65.8
NUPACK	**93.5**	**92.2**	**91.4**	**92.0**	84.5	77.4	93.1	87.1	81.4	86.6	76.9	79.4	80.3	81.1	80.3	82.3	86.3	81.3	81.0
EteRNABot (EB)	-	-	-	-	-	-	-	-	-	-	82.3	80.2	77.3	**95.5**	87.9	84.2	93.2	83.2	81.0
EB-alt	-	-	-	-	-	-	-	-	-	-	83.1	82.5	78.1	93.2	83.3	85.5	**93.8**	84.3	84.8
EB-control	-	-	-	-	-	-	-	-	-	-	77.0	81.0	83.3	94.0	81.1	83.6	91.9	77.7	82.6

C

	Puzzle (mean score)																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Participants	73.9	73.4	81.1	88.1	**90.0**	**91.9**	**95.6**	**91.0**	**85.8**	**90.7**	**84.9**	81.0	**87.5**	91.7	**89.8**	**89.6**	88.3	**88.3**	**90.2**
RNAInverse	73.2	85.8	80.0	70.6	71.8	65.6	79.1	79.7	73.3	74.8	73.1	74.6	77.8	75.0	76.7	82.1	82.2	72.0	64.1
NUPACK	**93.8**	**90.0**	**91.1**	**90.9**	83.1	75.3	93.3	87.3	78.7	87.1	77.2	79.6	79.5	80.3	80.3	82.7	84.4	81.6	73.9
EteRNABot (EB)	-	-	-	-	-	-	-	-	-	-	80.3	81.0	76.0	**94.5**	87.7	85.4	92.5	81.7	79.3
EB-alt	-	-	-	-	-	-	-	-	-	-	82.3	**81.6**	78.6	90.9	85.2	85.9	**92.6**	84.0	81.5
EB-control	-	-	-	-	-	-	-	-	-	-	77.5	79.8	83.6	92.6	79.0	84.6	91.9	78.3	78.3

Table S3. Pair-wise performance comparison of all tested design methods. Each table cell reports p value that the row method performs better than the column method.

A. First half of Phase I (puzzle 1, 2 and 3) – 75% Quantile comparison

	Participants	RNAInverse
NUPACK	0.00031	0.00028
Participants	-	0.046

B. Second half of Phase I (puzzle 4, 5 and 6) – 75% Quantile comparison

	NUPACK	RNAInverse
Participants	0.00029	0.00029
NUPACK	-	0.00022

C. Phase II (puzzle 11-15) – 75% Quantile comparison

	EteRNABot	EteRNABot-alt	EteRNABot-control	NUPACK	RNAInverse
Participants	0.15	0.025	0.021	0.00029	0.00015
EteRNABot	-	0.31	0.11	0.0012	0.00030
EteRNABot-alt	-	-	0.095	0.00030	0.00028
EteRNABot-control	-	-	-	0.024	0.00026
NUPACK	-	-	-	-	0.00034

D. Phase II (puzzle 16-19) in FMN free condition – 75% Quantile comparison

	EteRNABot	EteRNABot-alt	EteRNABot-control	NUPACK	RNAInverse
Participants	0.035	0.073	0.0026	0.00032	0.00026
EteRNABot	-	0.34	0.088	0.057	0.010
EteRNABot-alt	-	-	0.11	0.16	0.048
EteRNABot-control	-	-	-	0.35	0.16
NUPACK	-	-	-	-	0.11

E. Phase II (puzzle 16-19) FMN association constants – 75% Quantile comparison

	EteRNABot	EteRNABot-control	NUPACK	EteRNABot-alt	RNAInverse
Participants	0.20	0.127	0.015	0.0022	0.00095
EteRNABot	-	0.024	0.045	0.0063	0.0028
EteRNABot-control	-	-	0.34	0.16	0.011
NUPACK	-	-	-	0.067	0.011
EteRNABot-alt	-	-	-	-	0.049

F. First half of Phase I (puzzle 1, 2 and 3) – Mean comparison

	RNAiInverse	Participants
NUPACK	1.2×10^{-5}	1.0×10^{-7}
RNAInverse	-	0.40

G. Second half of Phase I (puzzle 4, 5 and 6) – Mean comparison

	NUPACK	RNAInverse
Participants	3.0×10^{-11}	3.6×10^{-35}
NUPACK	-	1.2×10^{-7}

H. Phase II (puzzle 11-15) – Mean comparison

	EteRNABot	EteRNABot-alt	EteRNABot-control	NUPACK	RNAInverse
Participants	0.014	0.010	0.0016	1.1×10^{-7}	2.7×10^{-11}
EteRNABot	-	0.50	0.23	0.0018	3.6×10^{-7}
EteRNABot-alt	-	-	0.022	0.00088	7.6×10^{-8}
EteRNABot-control	-	-	-	0.013	3.2×10^{-6}
NUPACK	-	-	-	-	0.00012

I. Phase II (puzzle 16-19) in FMN free condition – Mean comparison

	EteRNABot-alt	EteRNABot	EteRNABot-control	NUPACK	RNAInverse
Participants	0.0084	0.0033	3.6×10^{-5}	1.5×10^{-5}	1.3×10^{-7}
EteRNABot-alt	-	0.30	0.045	0.016	0.00014
EteRNABot	-	-	0.16	0.068	0.00083
EteRNABot-control	-	-	-	0.23	0.0016
NUPACK	-	-	-	-	0.011

J. Phase II (puzzle 16-19) FMN association constants – Mean comparison

	Participants	NUPACK	EteRNABot-control	EteRNABot-alt	RNAInverse
EteRNABot	0.47	0.056	0.067	0.0081	5.0×10^{-5}
Participants	-	0.078	0.071	0.010	5.0×10^{-5}
NUPACK	-	-	0.45	0.15	0.0001
EteRNABot-control	-	-	-	0.18	0.00015
EteRNABot-alt	-	-	-	-	0.0033

K. First half of Phase I (puzzle 1, 2 and 3) – Median comparison

	RNAiInverse	Participants
NUPACK	0.00022	0.00027
RNAInverse	-	0.037

L. Second half of Phase I (puzzle 4, 5 and 6) – Median comparison

	NUPACK	RNAInverse
Participants	0.00025	0.00030
NUPACK	-	0.00025

M. Phase II (puzzle 11-15) – Median comparison

	EteRNABot	EteRNABot-alt	EteRNABot-control	NUPACK	RNAInverse
Participants	0.0060	0.0065	0.00034	0.00030	0.00025
EteRNABot	-	0.47	0.083	0.0020	0.00027
EteRNABot-alt	-	-	0.072	0.00060	0.00032
EteRNABot-control	-	-	-	0.0048	0.00024
NUPACK	-	-	-	-	0.00037

N. Phase II (puzzle 16-19) in FMN free condition – Median comparison

	EteRNABot-alt	EteRNABot	EteRNABot-control	NUPACK	RNAInverse
Participants	0.0033	0.0040	0.00030	0.00032	0.00028
EteRNABot-alt	-	0.25	0.0087	0.036	0.00094
EteRNABot	-	-	0.059	0.097	0.011
EteRNABot-control	-	-	-	0.22	0.022
NUPACK	-	-	-	-	0.032

O. Phase II (puzzle 16-19) FMN association constants – Median comparison

	EteRNABot	EteRNABot-control	NUPACK	EteRNABot-alt	RNAInverse
Participants	0.22	0.060	0.043	0.021	0
EteRNABot	-	0.070	0.081	0.012	0
EteRNABot-control	-	-	0.63	0.23	0.0004
NUPACK	-	-	-	0.11	0.00025
EteRNABot-alt	-	-	-	-	0.0011

Table S4 (following pages). Participant descriptions of design rules. Description shows the original design rule statement by participants. The scoring functions show the pseudocode of the scoring function coded from the corresponding design rule. In the scoring functions, the "Number *[Number]*" notation represents an originally proposed parameter and an optimized parameter (Materials and Methods) respectively.

Title	Repetition
Author	aldo
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_repetition
Description	Measure how many times each subsequence of length n is repeated in the whole sequence for each n in some reasonable range, e.g. from 2 to 5. For our purposes let "the number of repetitions" be defined as "one less than the number of occurrences", i.e. if GA occurs 3 times we say it is repeated twice. For example, for the sequence AGUCUGACGUCCGUA we get 1 repetition each of UC, CG, GUC and CGU and 2 repetitions of GU. The design is penalized for each repeating subsequence based on the length of the subsequence and the number of repetitions.
Scoring function	score = 100 – (number of repeated sequences of length 2) * 5.00 *[4.89]* – (number of repeated sequences of length 3) * 5.00 *[4.93]* – (number of repeated sequences of length 4) * 5.00 *[4.99]* – (number of repeated sequences of length 5) * 5.00 *[4.99]*

Title	A Basic Test
Author	Dejerpha
Original post	http://getsatisfaction.com/EteRNAgame/topics/a_basic_test
Description	Let's try out the Strategy Market feature with some simple criteria... 50% of pairs are UA Free energy = –1.5 * number of pairs [e.g. 48 kcal if there are 32 pairs] Melting point between 77 and 97°C
Scoring function	**score** = 100 – abs((number of UA pairs) / (total number of pairs) – 0.50 *[0.42]*) * 100.00 *[93.04]* – abs(–1.50 *[–1.87]* * (total number of pairs) – (free energy)) * 1.00 *[1.15]* – max (77.00 *[63.60]* – (melting point), (melting point) – 97.00 *[102.00]*, 0) * 1.00 *[0.94]*

Title	Eli Blue Line
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_blue_line
Description	I like to ad a strategy that says, no strands with 3, 4 or more blue nucleotides in line. Penalize for each extra with –1.
Scoring function	**score** = 100 for each loop **score** = **score** – (max((number consecutive Us) – 4, 0) * 1.00 *[13.15]*)

Title	G's in place of the last A's on the right hand side of any end loop
Author	Clollin
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_ensemble_strategy_or_beta_test_experiment_title_gs_in_place_of_the_last_as_on_the_right
Description	If any end loops (MUST be 4 or more nucleotides) (only one stack coming off the loop) exist, give a point for each loop when the designer has mutated the last available A on the right hand side into a G.
Scoring function	**score** = 80 + (number of hairpins with 4 or more bases with G at the end)

Title	Tests by Region – boundaries
Author	Quasispecies
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_tests_by_region
Description	This strategy breaks the RNA into regions and applies penalties to each region. Each sequence starts with 100 points. Each time you penalize, you deduct one point. You could probably add a weighing factor to each penalty, but I would only be guessing at what they should be. For all boundary regions: Give a penalty if the boundary is not a G/C pair Give a penalty If there is an adjacent base pair and it is identical to the boundary pair
Scoring function	**score** = 100 – ((number of pairs adjacent to loops that are not GC) + (number of consecutive identical pairs adjacent to loops)) * 1.00 *[1.01]*

Title	GC Pairs in Junctions
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_gc_pairs_in_junctions
Description	I would like to ad a strategy that says that there have to be a GC-pair in each junction (That is between a tetreloop and an arm, a multiloop and an arm, the neck and the multiloop, the neck and the hook.) For each place there is no GC-pair in a junction, subtract one point.
Scoring function	**score** = 100 – (number of pairs adjacent to loops that are not GC) * 1.00 *[3.46]*

Title	No Blue Nucleotides in Hook Area
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_no_blue_nucleotides_in_hook_area
Description	I would like to ad a strategy, that says: No blue nucleotides are allowed in the hook area (the nucleotides beside the junctions) Give –2 for each blue nucleotide in this position
Scoring function	**score** = 100 – (number of Us in dangling loops at 3' or 5' ends) * 2.00 *[19.48]*

Title	Closing GC pairs in 3-1 internal loops
Author	mat747
Original post	http://EteRNA.cmu.edu//sites/default/files/chat_screens/scr267_1311088579805.png
Description	3-1 internal loop must be closed with GC pairs
Scoring function	**score** = 100 – (number pairs other than GC closing 3-1 internal loops)

Title	The 1-1 Loop V1
Author	merryskies
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_beta_test_experiment_title_the_1_1_loop_v1
Description	If there is a 1-1 loop (only one un-bonded nt opposite only one un-bonded nt), then: If both nts are Gs, then add 5 points. If one nt is a G and the other nt is an A, then add 3 points. If both nts are A, then add 1 point.
Scoring function	**score =** 80 for each 1-1 internal loop if both bases are Gs **score = score** + 5.00 *[2.50]* else if bases are G and A or A and G **score = score +** 3.00 *[–0.89]* else if both bases are As **score = score +** 1.00 *[2.83]*

Title	Clean plot, stack caps, and safe GC
Author	xmbrst
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_clean_plot_stack_caps_and_safe_gc
Description	plot_score = (number of white cells in the upper triangle of the pairwise probabilities plot) / (total number of cells in the upper triangle of the pairwise probabilities plot) cap_score = ((number of GC pairs that are at the end of a stack) + 0.5 * (number of GC pairs that are 1 away from the end of a stack)) / (3 * total number of stacks) gc_penalty = 2 if 80% or more of the design's pairs are GC pairs, 0 otherwise. A design's total score is: (2 + plot_score + cap_score–gc_penalty) * 25 The +2 and *25 are just to make it come out to between 0 and 100.
Scoring function	**score** = (2 + (plot_score * 1.00 *[0.88]* + cap_score * 1.00 *[1.05]* – (gc_penalty with GC pair threshold on 0.80 *[0.83]*) * 2.00 *[2.10]*)) * 25

Title	JP Stratmark
Author	JerryP70
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_jp_stratmark
Description	45%-65% of the pairs should be GC 40%-50% of the pairs should be AU 5% or less of the pairs should be GU Tetraloops and pentaloops should be closed with a GC with any loops containing more unbonded pairs being closed with 2 consecutive GC's Stacks need to have a GC pair every third or fourth bonded pair in stacks with six or more bonded pairs
Scoring function	**score** = 10.00 *[11.12]* * ((1 if 0.45 *[0.47]*<= (number of GC pairs) / (total number of pairs) <= 0.65 *[0.65]*) + (1 if 0.40 *[0.46]*<= (number of AU pairs) / (total number of pairs) <= 0.50 *[0.46]*) + (1 if (number of GU pairs) / (total number of pairs) <= 0.05 *[0.04]*)) + 10.00 *[8.47]* * ((number of tetraloops and pentaloops closed with a GC pair) + (number of stacks with more than 6 pairs that does not have 4 consecutive non-GC pairs))

Title	Energy Limit in Tetraloops
Author	Eli Fisker
Original post	https://getsatisfaction.com/EteRNAgame/topics/_strategy_market_energy_limit_in_tetraloops
Description	I would like to ad a strategy that says: Max energy level in tetraloops allowed is 4,5. Anything above that should be penalized.
Scoring function	**score** = 100 – (number of tetraloops with free energy bigger than 4.50 *[4.14]*) * 10.00 *[8.51]*

Title	Double AU pairs strategy
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_double_au_pair_strategy
Description	I would like to ad a strategy that says, there can be 0 or 1 double AU-pair in the design (two AU-pairs turning the same way.) For more than one double AU-pair, ad –1 pr. extra. Neckarea is the exception to this strategy. Here up to 2 X 2 double AU-pair are allowed (if these quads are switched opposite to each other.) Au-pairs in the neck are not counting as AU pairs in the overall design. If there is two double AU-pair in the neck, then an additional AU pair is allowed elsewhere in the design, without penalty.
Scoring function	**score** = 100 – (number of consecutive AU pairs except in the first stack from 5' end) * 1.00 *[7.78]*

Title	Green and blue nucleotides a strand + Strong middle half
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_green_and_blue_nucleotides_a_strand_strong_middle_half
Description	This is an addition to my strategy Green and blue nucleotides on a strand, where I take in to account, that a strengthening appears to be taking place right after the middle of the design. (For the theory behind this see my getsat post Slightly skewed energy tendency for the whole lab) So this strategy should be excactly as my Green and blue nucleotides on a strand, with the only exception, that I allow a blue/green line on a strand that is one nucleotide longer than stated in my strategy - if this longer blue/green line is found up to 11 nucleotides after the exact middle of the design (this 11 nt limit may have to be higher, if we are starting to making bigger lab puzzles) So here my strategy is again with the addition. I would like to ad a strategy about how many blue and green nucleotides I will allow in line right after each other on a strand of a certain length. In a string 3 nt long, allow max 3 blue/green nt in line on a strand In a string 4 nt long, allow max 3 blue/green nt in line on a strand In a string 5 nt long, max 3 In a string 6 nt long, max 4 In a string 7 nt long, max 4 In a string 8 nt long, max 4 In a string 9 nt long, max 4 In neckarea up to 6 is allowed. (including this number) Penalize with –1 for each extra blue/green nucleotide above the number allowed, Except if one extra nucleotide is found in the area (number of nucleotides divided with 2 and found between this area and + 11 nucleotides)
Scoring function	**score** = 100 – (number of extra C or U bases) * 1.00 *[11.50]*

Title	Loop pattern for small multiloops
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_loop_pattern_for_small_loops
Description	Allowed these patterns in small multiloops (3 arms, one neck, and no nucleotides between the (4) gc-pairs) a) Rightturning GC-pairs all around allowed (red nucleotide to the right) b) Axis I: The GC-pair in the neck turn opposite, so does the one in top, the last two turn right (red nucleotide to the right), the rest turn the usual way. c) Axis II: Neck GC and top GC-pair (with red nucleotide to the right), left GC pair and bottom GC-pair opposite. For any other combinaiton give –1 for each GC-pair in the multiloop that is differing from these three patterns. See the reason for this in my getsat post loop pattern for small multiloops This is sort of meant as an adittion to more general strategies for GC-pairs in multiloops. Could be seperated into 3 different additions.
Scoring function	**score** = 100 – (number of GC pairs against the patterns) * 1.00 *[3.21]* – (number of non-GC pairs against the patterns) * 2.00 *[5.73]*

Title	Tetraloop Similarity
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_tetraloop_similarity
Description	I would like a strategy that ad points for having similar energy content in tetraloops (like quardruplet) as best, double twin tetraloops as next best. For designs with 4 tetraloops: a) If all tetraloops are similar in energy content, add 5 points b) If there are two pair of twin tetraloops, add 5 points (2 x 2 identical tetraloops) c) If three tetraloops are similar in energy content add 3 d) If two tetraloops are similar and the rest un even, add 2 points e) If all tetraloops are different ad 0 point For designs with only 3 tetraloop: If all 3 tetraloops are similar, add 5 points If only 2 tetraloops are similar, add 3 points If none are similar, add 0 points
Scoring function	if there are 4 tetraloops if all 4 tetraloops free energy differences are less than 0.30 *[0.32]*, **score** = 100 else if there are 2 groups of tetraloops with free energy differences less than 0.30 *[0.32]*, **score** = 100 else if there are 3 tetraloops with free energy differences less than 0.30 *[0.32]*, **score** = 60 else if there are 2 tetraloops with free energy differences less than 0.30 *[0.32]*, **score** = 40 else**score** = 0 else if there are 3 tetraloops if all 3 tetraloops free energy differences are less than 0.30 *[0.32]*, **score** = 100 else if there are 2 tetraloops with free energy differences less than 0.30 *[0.32]*, **score** = 60 else**score** = 0

Title	[Example] 60% of pairs should be GC pairs
Author	jeehyung
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_example_60_of_pairs_must_be_gc_pairs-1erc6
Description	60% of pairs should be GCs
Scoring function	**score** = 100 – abs((number of GC pairs) / (total number of pairs) – 0.60 *[0.59]*) * 100

Title	Clean dot plot
Author	Penguian
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_clean_dot_plot
Description	Penalize designs by one point for each dot in the pairing probabilities plot.
Scoring function	**score** = 100 – (sum of probabilities of incorrect pairs) / (total number of pairs)

Title	Twisted base pairs
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_twisted_basepairs
Description	I would like to add a strategy that penalizes base pair similarity. Give –1 for each base pair that is same colors and turn the same way as it's neighbor. In this strategy I will just aim at penalizing same turning base pairs, though I'm aware that they are allowed under certain conditions. I have made other strategies (not programmed yet) that opens up for use of double AU and GC pairs at certain places. This is just the broad strategy to bomb out designs with too much repetitive, same turning base pairs, no matter what their color are and length of numbers of them in line. This is because very few of the really good designs have them. Usually two repetitive base pairs are allowed in the neck, only one in the main design and sometimes two in the main design - if it's an asymmetric design. I'm aware that the really long armed designs like the cross, have a greater tolerance for repetitive base pairs. In strings 8 and more nucleotides long, 2 similar same turning base pairs are allowed pr. string, exactly like in the neck area. I just wanted to keep this strategy simple.
Scoring function	**score** = 100 – (number of consecutive identical pairs)

Title	Loops & Stacks
Author	aldo
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_loops_stacks
Description	Loops/unpaired bases (excluding locked bases): 1 point for every unpaired A 0.5 points for every unpaired G Stacks: 2 points for every GC pair at the end of a stack 1 point for every GC pair not at the end of a stack 1 point for every AU pair at the end of a stack 2 points for every AU pair not at the end of a stack 1 point for every GU pair Modifier = 1 –abs((number of GC pairs)/(total number of pairs) – 0.55) Score = 100 * (modifier) * (total number of points)/(total number of bases excluding locked unpaired bases)
Scoring function	**score** = 100 * (1 – abs((number of GC pairs) / (total number of pairs) – 0.55 *[0.59]*)) * ((number of GC pairs at the end of a stack) * 2.00 *[2.02]* + (number of GC pairs not at the end of a stack) * 1.00 *[1.02]* + (number of AU pairs at the end of a stack) * 1.00 *[0.99]* + (number of AU pairs not at the end of a stack) * 2.00 *[1.90]* + (number of GU pairs) * 1.00 *[1.03]*) / (sequence length)

Title	Direction of GC-pairs in multiloops + neckarea
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_direction_of_gc_pairs_in_multiloops_neckarea
Description	I make a wish for a strategy that says: All GC-pairs in the in multiloop junctions, have to turn in same direction. (Red nucleotide to the right and green nucleotide to the left.) Exception: the GC-pair connecting multiloop and neck, are allowed to turn in both directions, without being penalized. I would like to give –2 point for each wrong turning GC-pair.
Scoring function	**score** = 100 – (number of GC pairs in wrong directions adjacent to multiloops except those in the first stack from 5'end) * 1.00 *[5.71]* – (number of non-GC pairs adjacent to multiloops) * 2.00 *[6.63]*

Title	Multiloop similarity
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_market_strategy_multiloop_similarity
Description	Energy inside two similar and same sized multiloop should be the same. (Multiloops with same numbers of arms and same numbers of nucleotides between arms) I'm thinking about designs with multiple multiloops, like the branches designs. Penalize with –1 for each 0.5 energy difference between them.
Scoring function	**score** = 100 for each pair of identical multiloops if (free energy difference between the loops) > 0.50 *[0.44]* **score = score –** 1.00 *[1.11]*

Title	Eli Green Line
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_green_line
Description	I like to ad a strategy that says, no strands with 3 green nucleotides in line. Penalize this with –1. Exception for allowing 3 gren nucleotides in line, if string is 9 nucleotides long or more and with no internal loop inside.
Scoring function	**score** = 100 for each loop less than 9 bases long **score** = **score** – (max((number consecutive Cs) – 3, 0) * 1.00 *[13.98]*)

Title	Ding quad energy
Author	Ding
Original post	http://getsatisfaction.com/EteRNAgame/topics/tell_us_about_your_EteRNA_lab_algorithms
Description	Finally I look at color patterns as Eli Fisker has described, and quad energies. a) I don't like to see any quad energies over –0.9 kcal (a UA UA or AU AU quad). If there are GU bonds used, b) I like to see them stabilized on one side with a GC in a configuration that give –2.1 or –2.5 kcal.
Scoring function	**score** = 100 – (number of a 2 pair stacks that violates *a* with free energy threshold –0.90 *[–0.85]*) * 1.00 *[1.00]* – (number of 2 pair stacks that violates *b*) * 1.00 *[0.99]*

Title	Tests by Region – loops
Author	Quasispecies
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_tests_by_region
Description	For all loop and overhang regions: Penalize if the free energy contribution per nucleotide in the region is >0.5 kcal Penalize for each stretch of 4 consecutive bases complementary to 4 bases elsewhere in the molecule. Example: Round 1: 5' ...[GAGU]AACGGAC... 3' Penalized if more than one sequence in the molecule is complementary to "GAGU" Round 2: 5' ...G[AGUA]ACGGAC... 3' Penalize if more than one sequence in the molecule is complementary to "AGUA" And so on...
Scoring function	**score** = 100 – (number of loops whose free energy is bigger than 0.5 times the length of the loop) * 1.00 *[1.00]* – (number of length 4 loop pairs that are complementary to each other) * 1.00 *[1.00]*

Title	Berex Loop Basic
Author	Berex NZ
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_berex_loop_basic
Description	a) If length 1 bulge has a GC partner, +1 b) If 1-1 loop got 2 Gs, +3 c) If 1-2 loop got G in the bottom nucleotide, +2. If it got G in bottom and right and nucleotide, +2 d) If 2-2 loop got 2 Gs, + 3 e) If 3-1 loop got 2Gs on top, +3. If it got 1 G on bottom, +3 f) If G is on the right hand side of a tetraloop, + 3 g) For a loop with length 5 or more, if the loop got 1 G, +1. h) For a loop with length 5 or more, if the loop got 2 Gs, +2
Scoring function	**score** = 80 for each loop if *a* **score** = **score** + 1.00 *[1.04]* else if *b* **score** = **score** + 3.00 *[2.95]* else if *c* **score** = **score** + 2.00 *[2.11]* else if *d* **score** = **score** + 3.00 *[3.06]* else if *e* **score** = **score** + 3.00 *[3.04]* else if *f* **score** = **score** + 3.00 *[3.10]* else if *g* **score** = **score** + 1.00 *[0.89]* else if *h* **score** = **score** + 2.00 *[1.91]*

Title	Legal Placement of GU pairs
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_legal_placements_of_gu_pairs
Description	I would like to ad a strategy for placements of GU-pairs If there is no Gu pair (in the whole design) +2 a) Gu-pair between two opposite twisted GC-pairs +1 (GC-pairs should be twisted compared to each other) b) Gu-pair between 2 GC-pairs where the GC-pairs turn same way compared to each other +0,5 c) Gu-pair beside one GC-pair +0,5 d) Gu-pair besides no GC-pair –2 e) Two GU pairs beside each other, –4, if more –2 pr extra f) GU-pair in junction –1 Neckarea behaves different than the rest of the design, also when it comes to tolerance for GU-pairs, so here I wish for almost double price and half penalty: If GU-basepair is found in neck: Gu-pair between two opposite twisted GC-pairs +2 Gu-pair between 2 GC-pairs where the GC-pairs turn same way compared to each other +1 Gu-pair beside one GC-pair +1 Gu-pair besides no GC-pair –1 g) Two GU pairs beside each other, (turning same way) - one of those are allowed, if more, or the double GU pair are opposite turning, penalize with –2 GU-pair in junction –0.5

Scoring function	**score** = 80
	if (number of GU pairs) == 0
	score = score + 2.00 *[1.99]*
	else
	for each GU pair not in the first stack from the 5' end
	if *a*
	score = score + 1.00 *[0.99]*
	else if *b*
	score = score + 0.50 *[0.49]*
	else if *c*
	score = score + 0.50 *[0.49]*
	else if *d*
	score = score – 2.00 *[1.99]*
	else if *f*
	score = score – 1.00 *[1.03]*
	
	for each group of consecutive GU pairs not in the first stack from the 5' end
	if *e*
	score = score – 4.00 *[4.04]*
	score = score – max(0, (number of GU pairs in the group) – 2) * 2.00 *[2.02]*
	
	for each GU pair in the first stack from the 5' end
	if *a*
	score = score + 2.00 *[2.01]*
	else if *b*
	score = score + 1.00 *[1.00]*

	else if *c* **score = score +** 1.00 *[1.00]* else if *d* **score = score** – 1.00 *[0.99]* else if *f* **score = score –** 0.50 *[0.49]* for each group of consecutive GU pairs in the first stack from the 5' end if *g* **score = score –** 2.00 *[2.01]*

Title	1-1 Loop Energy V1
Author	Merryskies
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_beta_test_experiment_title_1_1_loop_energy_v1
Description	If there is a 1-1 loop (only one un-bonded nt opposite only one un-bonded nt), then: If the Energy of the loop is less than E1-1, then add 5 pts. For the first run, E1-1 = –0.39.
Scoring function	**score** = 80 for each 1-1 internal loop if (free energy of the loop) < –0.39 *[–0.46]* **score** = **score** +5.00 *[3.55]*

Title	Tetraloop Pattern
Author	Ding
Original post	http://getsatisfaction.com/EteRNAgame/topics/tell_us_about_your_EteRNA_lab_algorithms
Description	If the tetraloops aren't either AAAA or one of the known patterns that gets an energetic bonus in EteRNA I check to see if there are mispairing possibilities with a complementary sequence nearby (if for instance the four nucleotides in the tetraloop are AGUA and part of a nearby stack is UAC).
Scoring function	**score =** 100 **–** (number of tetraloops whose sequences are not AAAA) * 5.00 *[4.78]*

Title	Mismatch
Author	aldo
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_mismatch
Description	Slide the sequence past itself in reverse one base at a time and determine how many pairs form each time. Penalize the design for each pair. Optionally give different penalties for different types of pairs (AU, GU or GC). It would probably also be wise to subtract the number of intended pairs from the total.
Scoring function	**score =** 100 – (number of valid pairs after shifting the sequence by 1 in the target structure) * 40.00 *[19.01]* – (number of valid pairs after shifting the sequence by 2 in the target structure) * 40.00 *[6.26]* – (number of valid pairs after shifting the sequence by 3 in the target structure) * 40.00 *[29.83]*

Title	Tetraloop Blues
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_tetraloop_blues
Description	I would like to ad a strategy that says – no more than 2 blue nucleotides in tetraloops - that includes blue nucleotides in the closing base pair in the tetraloop too. I would like to penalize with –1 for each extra blue nucleotide here.
Scoring function	**score =** 100 for each tetraloop **score = score** – max((number of Us in the tetraloop and its closing pair) – 2, 0) * 4.00 *[10.16]*

Title	Eli Red line
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_red_line
Description	I like to ad a strategy that says, no strands with 3 red nucleotides in line. Penalize this with –1. Exception for allowing 3 red nucleotides in line, if string is 9 nucleotides long or more and with no internal loop inside.
Scoring function	**score** = 100 for each loop less than 9 bases long **score** = **score** – (max((number consecutive Gs) – 3, 0) * 1.00 *[13.68]*)

Title	Wrong Direction of GC-pairs in multiloops
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_wrong_direction_of_gc_pairs_in_multiloops
Description	I make a wish for a strategy that is the opposite of my strategy Direction of GC-pairs in multiloops. As last time all GC-pairs in the multiloop junctions, should turn in same direction. This time Green nucleotide to the right and red nucleotide to the left. I would like to give –1 point for each GC-pair in the multiloop, where red nucleotide is to the right and the green on is to the left. This strategy should preferable do worse than my Direction of GC-pairs in multiloops, as the tendency for best placement of GC-pairs in multiloops is red nucleotide to the right.
Scoring function	**score** = 100 – (number of GC pairs in correct directions adjacent to multiloops) * 1.00 *[2.29]*

Title	Deivad's strategy
Author	deivad
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_deivads_strategy
Description	- I count how many loops there are with its energy below 1. I add a point for each one. (+1 each) - I do the same for the loops with its energy higher than 4. I subtract a point for each loop like this. (–1 each). - I add a point for each short chain (less than 5 bonds) that has two or more GC's (+1 each chain). - I subtract a point for each short chain (less than 5 bond) that has one or more GU (–1 each chain). - If a GU bond is just near a loop with its energy higher than 2, I subtract 2 points for each one (–2 each). - I subtract a point for each C in a loop (–1 each), as they are potential undesirable strong bonds. I don't take into account the blocked, default elements. - I subtract 0.2 points for each G in a loop (–0.2 each), as they can try to make a strong bond with any C. I don't take into account the blocked, default elements. - I divide the energy between 10, and put it positive (I mean, if the total energy is–50 kcal, it would have 5 points. - If there are more than a 50% of GC's, I subtract 2 points. It can be a way of avoiding too many GC's which is an easy way for obtaining the desired design. - If there are more than a 15% of GU's, I subtract 5 points. Probably it will be too weak, and it won't fold correctly. - I divide the melting point between 20. As far as I've seen, good designs (with 95% or more, obtain more than 10 points (more than 20, sometimes), 80-90% designs obtain between 0-10 points, while designs with less than a 70% are often negative.

Scoring function	**score** = 80
	for each loop
	  if (free energy of the loop) < 1.00 *[0.91]*
	  score = score + 1.00 *[0.91]*
	  else if (free energy of the loop) < 4 *[4.67]*
	  score = score – 1.00 *[0.93]*
	  else if (free energy of the loop > 2.00 *[2.07]*
	  if there is a GU pair adjacent to the loop
	  score = score – 2.00 *[2.04]*
	for each stack
	  if (number of pairs in the stack) < 5
	  if (number of GC pairs in the stack) > 1
	  score = score + 1.00 *[1.00]*
	  if (number of GU pairs in the stack) > 0
	  score = score – 1.00 *[1.00]*
	score = score – (number of unpaired Cs) * 1.00 *[1.12]*
	score = score + (total free energy) / –10.00 *[–9.29]*
	if (number of GC pairs) / (total number of pairs) > 0.5
	score = score – 2.00 *[2.00]*
	if (number of GU pairs) / (total number of pairs) > 0.15
	score = score – 5.00 *[5.24]*

	score = score + (melting point) / 20.00 *[20.10]*

Title	Direction of GC-pairs in multiloops
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_direction_of_gc_pairs_in_multiloops
Description	Now that it thanks to the EteRNA crew were possible to run my strategy of GC-pairs in the junctions, I make a wish for a strategy that says: All GC-pairs in the in multiloop junctions have to turn in same direction. (Red nucleotide to the right and green nucleotide to the left.) I would like to give –1 point for each wrong turning GC-pair.
Scoring function	**score** = 100 – (number of GC pairs in wrong directions adjacent to multiloops) * 1.00 *[5.71]* – (number of non-GC pairs adjacent to multiloops) * 2.00 *[6.63]*

Title	No blue nucleotides in multiloop ring
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_no_blue_nucleotides_in_multiloop_ring
Description	I would like to add a strategy, that says: No blue nucleotides are allowed in multiloop ring. (the nucleotides beside the junctions) Give –1 for each blue nucleotide in this position
Scoring function	**score** = 100 – (number of Us in multiloops) * 1.00 *[3.66]*

Title	Berex Test
Author	Berex NZ
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_berex_test
Description	Melting Point between 97 and 107 Free Energy between –30 and –60 G bases of 22% U bases of 13% C bases of 20%
Scoring function	**score =** 100 – max (–60.00 *[–68.34]* – (free energy), (free energy) – (–30.00 *[–30.20]*), 0) * 1.00 *[1.12]* – abs((number of Gs) / (sequence length) – 0.22 *[0.15]*) * 100.00 *[116.23]* – abs((number of Us) / (sequence length) – 0.13 *[0.07]*) * 100.00 *[129.23]* – abs((number of Cs) / (sequence length) – 0.20 *[0.19]*) * 100.00 *[117.62]* – *max ((melting point) – 107.00 *[133.66]*, 97.00 *[35.47]* – (melting point), 0) * 1.00 *[1.36]* (* The game only gives melting points between 37 and 107. This term therefore is always 0 and was omitted from the description in the main text Figure 3.)

Title	Simplified Berex Test
Author	Jeehyung
Original post	https://getsatisfaction.com/EteRNAgame/topics/_strategy_market_simplified_berex_test
Description	Free Energy should be between –30 and –60 Melting Point should be between 97 and 107 Penalize RNAs if they don't satisfy these conditions
Scoring function	**score =** 100 – max ((melting point) – 107.00 *[102.87]*, 97.00 *[71.76]* – (melting point), 0) * 1.00 *[1.14]* – max (–60.00 *[–48.28]* – (free energy), (free energy) – (–30.00 *[–37.07]*), 0) * 1.00 *[0.67]*

Title	Numbers of yellow nucleotides pr length of string
Author	Eli Fisker
Original post	http://getsatisfaction.com/EteRNAgame/topics/numbers_of_yellow_nucleotides_pr_length_of_string
Description	I would like to ad a strategy for numbers of yellow nucleotides allowed pr. lengt of string (neckarea excluded): If a string/arm is this number of nucleotides long, then allow this number of yellow adenine. For each yellow nucleotide below the minimum or above the maximum, penalize with –2. String length (yellow nucleotides) 3 (1-2) String eg. the bulged cross and the asymmetry 4 (1-2) 5 (1-3) 6 (2-3) 7 (3-4) 8 (2-5) 9 (1-4) This could be used to rule out some of the cub scouts and a few christmas threes.
Scoring function	**score** = 100 for each stack **score**= **score** –(max((number of AU pairs) – (upper bound on number of AU pairs), (lower bound on number of AU pairs) – (number of AU pairs), 0) * 2.00 *[10.50]*)

Title	Test by kkohli
Author	Kkohli
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_test_by_kkohli
Description	Free energy in loops of 1.5 or less. Melting point between 97 and 107. G-C bond at the opening of at least 60% of all loops. Amount of G between 25 and 35%. Free energy of no more than –30.
Scoring function	**score** = 100 – (number of loops with free energy bigger than 1.50 *[0.54]*) * 10.00 *[14.68]* – (1 if melting point is not between 97 and 107) * 10.00 *[11.76]* – abs((number of loops whose adjacent pairs are all GCs) / (total number of loops) – 0.6)) * 10.00 *[12.58]* – (1 if (number of Gs) / (sequence length) is not between 0.25 and 0.35) * 10.00 *[6.28]* – max((free energy) + 30, 0) * 1.00 *[0.94]*

Title	Only As in the loops
Author	merryskies
Original post	http://getsatisfaction.com/EteRNAgame/topics/_strategy_market_beta_test_experiment_title_only_as_in_the_loops
Description	If any nucleotide is *not* supposed to be bonded, then let it be assigned A (Adenine).
Scoring function	**score** = 100 – (number of unpaired bases that are not Adenine)

Table S5. EteRNA glossary. Frequently used terms created by participants to describe the various features and rules for secondary structure design

Term	Definition
Arm	A stem
Asymmetric Multiloop	A multiloop with different numbers of nucleotides between the stems
Blocking Point	A blocking point is an all Adenine bulge closed by an AU pair
Boosting	Putting Guanines in the terminal mismatches to lower the free energy of loops
Boost Point	Base in terminal mismatches.
C-stifling	Putting Cytosines in loops to minimize their chance of pairing.
Catalyst Point	A base that is not a boosting point in the target structure, but is a boost point in a close alternative structure.
Christmas Tree	RNA sequences whose pairs are all GC (green-red in EteRNA color scheme) pairs.
Corner Loop	A bulge loop.
Cub Scout Project	RNA sequences whose pairs are all AU (yellow-blue in EteRNA color scheme) pairs.
End Loop	A hairpin loop.
GC-heat	Using GC-pairs in multiloops to minimize the free energy.
GC-line	Consecutive 3 GC pairs in a stem
Hidden instabilities	Features that indirectly stabilize the target structure by destabilizing close alternative structures.
Hook	Loops at the 3´end and the 5´end.
Junction	Loop closing pairs.
Neck	The first stem from the 5´end.
Optical illusion	RNA sequences whose pairs are all GU pairs.
Pentagon Bulge	A bulge loop with 5 bases.
Quad	4 bases in 2 consecutive pairs.
Roadkill plot	A pairing probability plot suggesting many chances for mispairings.
Zigzag	Consecutive alternating bulge loops.

Table S6. Weights of 3 design rules in the score predictor of EteRNABot-control. Only rules utilizing conventional features in the RNA literature were selected. Weights were determined by linear regression.

Title	Weight
[Example] 60% of pairs should be GC pairs	0.41
Clean Dot Plot	0.16
Simplified Berex Test	0.14

Table S7. Weights of 5 design rules in the score predictor of EteRNABot. Rules were selected with Least Angle Regression. Weights were determined by linear regression.

Title	Weight
Berex Test	0.36
Numbers of yellow nucleotides pr length of string	0.22
Direction of GC-pairs in multiloops + neckarea	0.21
Clean plot, stack caps, and safe GC	0.12
A basic test	0.092

Table S8. Weights of 40 design rules in the score predictor of EteRNABot-alt. Weights were determined by L2 regularized linear regression.

Title	Weight
Repetition	0.18
Berex Test	0.14
A basic test	0.13
Numbers of yellow nucleotides pr length of string	0.12
Direction of GC-pairs in multiloops + neckarea	0.12
Wrong Direction of GC-pairs in multiloops	0.11
Clean plot, stack caps, and safe GC	0.10
Loops & Stacks	0.081
Eli Green line	0.080
G's in place of the last A's on the right hand side of any end loop	–0.075
Ding quad energy	–0.059
Only As in the loops	0.058
Tests by Region - loops	0.057
Mismatch	0.046
Direction of GC-pairs in multiloops	0.046
deivad's strategy	0.044
Closing GC pairs in 3-1 internal loops	0.038
Test by kkohli	–0.035
GC-pairs in junctions	0.034
Clean Dot Plot	0.032
The 1-1 Loop V1	0.032
[Example] 60% of pairs should be GC pairs	0.030
Twisted Basepairs	–0.026
Eli Blue line	0.024
No blue nucleotides in multiloop ring	–0.023
Legal placement of GU-pairs	–0.022
Eli Red line	0.022
Multiloop similarity	–0.020
Loop pattern for small multiloops	–0.019
Tests by Region - boundaries	0.016
Tetraloop Pattern	0.016
JP-Stratmark	–0.013
1-1 Loop Energy V1	–0.012
Green and blue nucleotides a strand + Strong middle half	0.011
No Blue Nucleotides in Multiloop Rings	–0.0079
Double AU-pair strategy	–0.0071
Tetraloop similarity	0.0049
Energy Limit in Tetraloops	–0.0014
Tetraloop blues	–0.00085
Berex Loop Basic	–0.00038