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], Snehal Gaikwad[d], Sungroh Yoon[c], Adrien Treuille[a,1], Rhiju Das[b,e,1], and EteRNA Participants[2]

[a]Department of Computer Science and [d]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15206; Departments of [b]Biochemistry and [e]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.G., 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.G., 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.

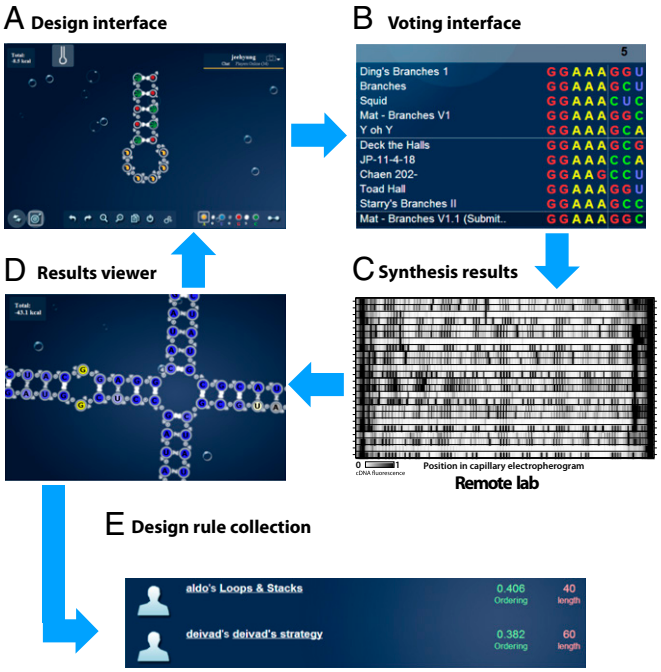

Fig. 1. EteRNA workflow. Each week, participants (*A*) design sequences that can fold into a target RNA structure in the sequence design interface and (*B*) review and vote for the best designs with the voting interface. (*C*) At the end of the round, the eight top-voted sequences are synthesized and 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

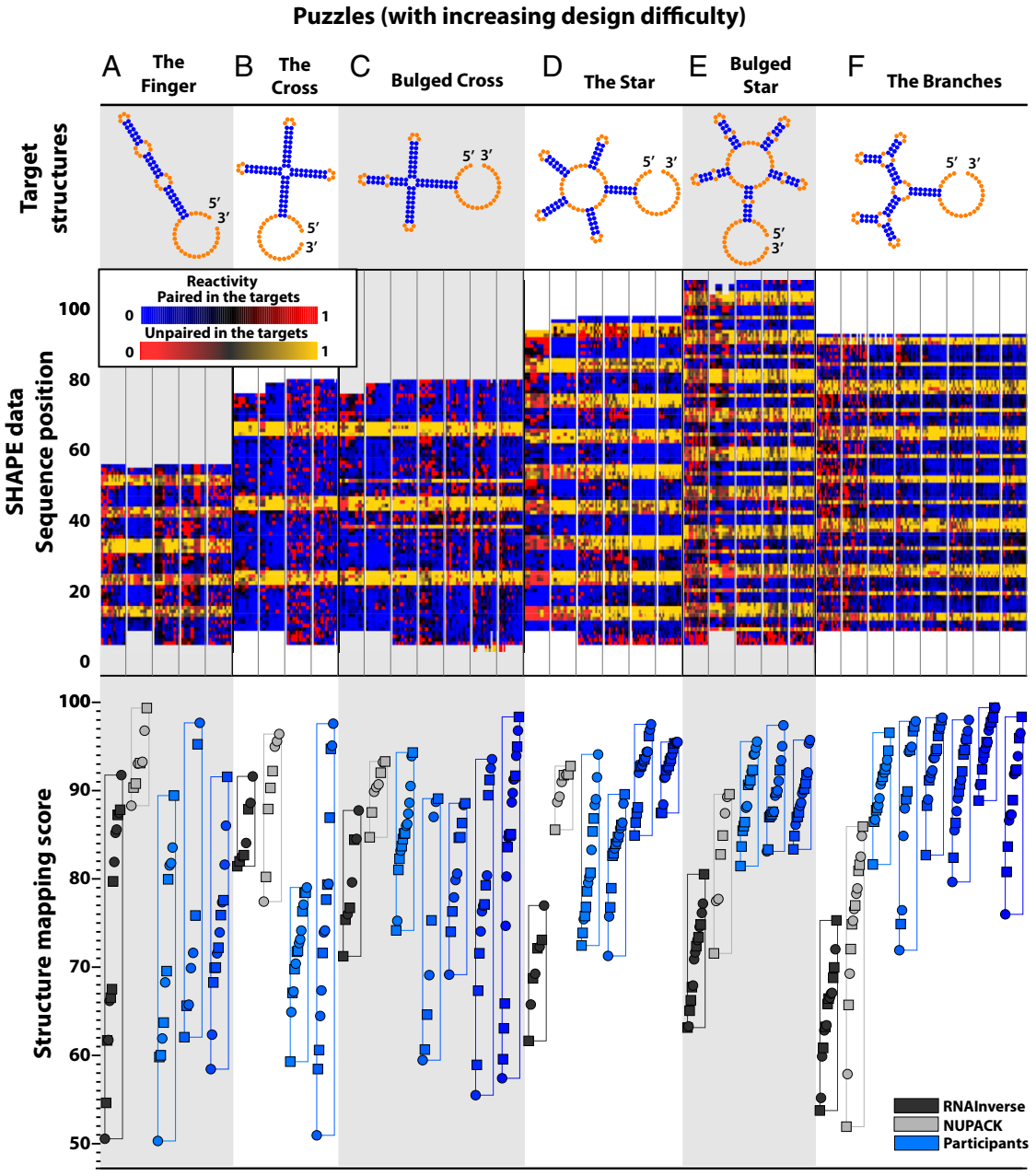

Fig. 2. Phase I puzzles and results in order of puzzle posting date. *Top* shows a target structure and title for each puzzle (*A–F*). Nucleotide coloring in target structures indicates the ideal SHAPE reactivity (gold for high reactivity and blue for low reactivity). *Middle* gives the single nucleotide 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

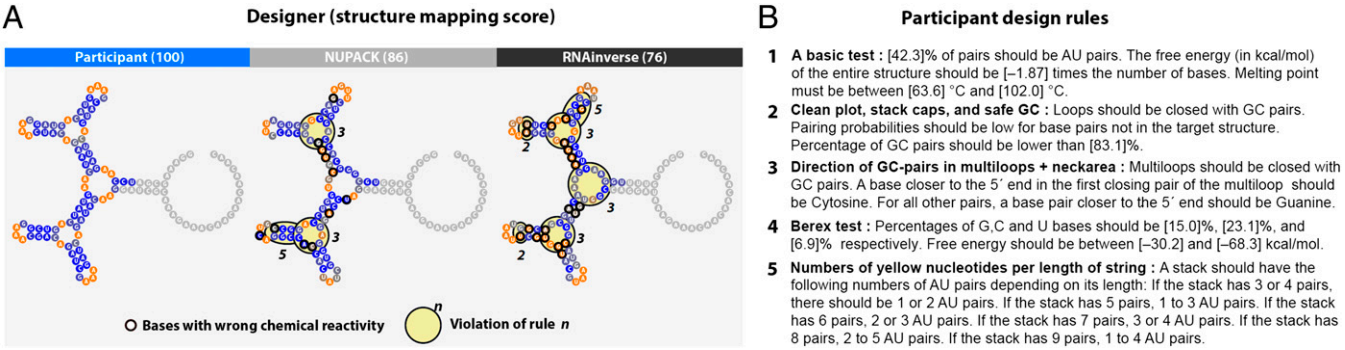

Fig. 3. RNA design rules proposed by EteRNA participants. (*A*) The best designs from each design agent (EteRNA participants, NUPACK, and RNAInverse) for the last target shape of phase I (Fig. 2*F*); the nucleotide coloring gives experimental chemical reactivity and is identical to the coloring used 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.

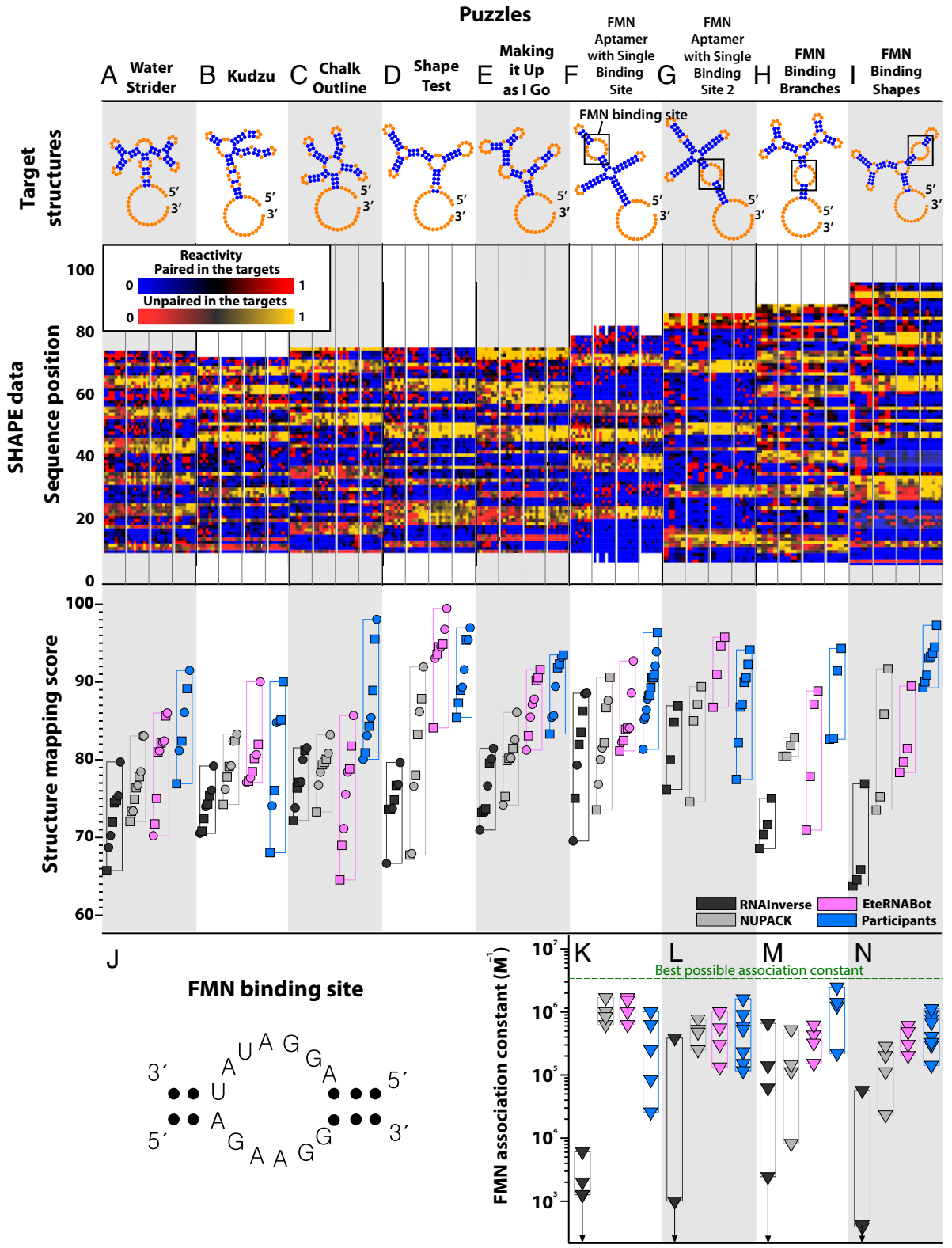

Fig. 4. Phase II puzzles and results, including the EteRNABot algorithm, in order of puzzle posting date. (*A–E*) Multijunction target structures distinct from each other and the phase I structures in topology. (*F–I*) Puzzles giving the binding site for FMN as a fixed sequence that must be displayed 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.