RNA secondary structure packages ranked and improved by high-throughput experiments

Hannah K. Wayment-Steele[1,2], Wipapat Kladwang[2,3], Eterna Participants[2], Rhiju Das[2,3,4]

[1] Department of Chemistry, Stanford University, Stanford, CA, 94305
[2] Eterna Massive Open Laboratory. List of consortium members given in Table S1.
[3] Department of Biochemistry, Stanford University, Stanford, CA, 94305
[4] Department of Physics, Stanford University, Stanford, CA, 94305

Abstract

The computer-aided study and design of RNA molecules is increasingly prevalent across a range of disciplines, yet little is known about the accuracy of commonly used structure prediction packages in real-world tasks. Here, we evaluate the performance of current packages using EternaBench, a dataset comprising 23 *in vitro* structure mapping and 11 riboswitch activity datasets involving 18,509 synthetic sequences from the crowdsourced RNA design project Eterna. We find that CONTRAfold and RNAsoft, packages with parameters derived through statistical learning, achieve consistently higher accuracy than more widely used packages like the ViennaRNA software, which derive parameters primarily from thermodynamic experiments. Motivated by these results, we develop a multitask-learning-based model, EternaFold, which demonstrates improved performance that generalizes to diverse external datasets, including complete viral genomes probed in vivo and synthetic designs modeling mRNA vaccines.

Introduction

RNA molecules perform essential roles in cells, including regulating transcription, translation, and molecular interactions, and performing catalysis.[1] Synthetic RNA molecules are gaining increasing interest and development for a variety of applications, including genome editing,[2] biosensing,[3] and vaccination.[4] Characterizing RNA secondary structure, the collection of base

pairs present in the molecule, is typically necessary for understanding the function of natural RNA molecules and is of crucial importance for designing better synthetic molecules. Some of the most widely-used packages use a physics-based approach[5] that assigns thermodynamic values to a set of structural features (Vienna RNAfold,[6] NUPACK,[7] and RNAstructure[8]), with parameters traditionally characterized via optical melting experiments and then generalized by expert intuition.[9] However, a number of other approaches have also been developed that utilize statistical learning methods to derive parameters for structural features (RNAsoft,[10] CONTRAfold,[11] CycleFold,[12] LearnToFold,[13] MXfold[14], SPOT-RNA[15]).

Secondary structure modeling packages are typically evaluated by comparing single predicted structures to secondary structures of natural RNAs[16]. While important, this practice has limitations for accurately assessing packages, including bias toward structures more abundant in the most well-studied RNAs (tRNAs, ribosomal RNA, etc.) and neglect of energetic effects from these natural RNAs' tertiary contacts or binding partners. Furthermore, scoring on single structures fails to assess the accuracy of ensemble-averaged RNA structural observables, such as base-pairing probabilities, affinities for proteins, and ligand-dependent structural rearrangements, which are particularly relevant for the study and design of riboswitches[17, 18], ribozymes, pre-mRNA transcripts, and therapeutics[19] that occupy more than one structure as part of their functional cycles. Existing packages are, in theory, capable of predicting ensemble properties through so-called partition function calculations, and, in practice, are used to guide RNA ensemble-based design, despite not being validated for these applications.

New sources of structural data now offer opportunities to evaluate secondary structure packages incisively and with less bias. In particular, experimental data accumulating in the RNA design crowdsourcing platform Eterna has potential to mitigate effects of bias incurred in natural RNA datasets, as the probed constructs are designed by citizen scientists to fold to synthetic

secondary structures of their choosing.[20] In this work, we evaluate the performance of commonly-used packages capable of making thermodynamic predictions in two tasks that have been crowdsourced on Eterna and are emerging as central to RNA characterization and design: 1) predicting chemical reactivity data through calculating probabilities that nucleotides are unpaired, and 2) predicting relative stabilities of multiple structural states that underlie the functions of riboswitch molecules, a task that involves predicting affinities of both small molecules and proteins of interest. Both of these evaluations are made possible through EternaBench, a collection of high-throughput datasets acquired in Eterna. We find striking, consistent differences in package performance across these quantitative tasks, with the packages CONTRAfold and RNASoft performing better than packages that are in much wider use. Furthermore, we develop a multitask-learning-based framework to train a thermodynamic model on these tasks concurrently with the task of single-structure prediction. The resulting multitask-trained model, called EternaFold, demonstrates increased accuracy both on held-out data from Eterna as well as completely independent datasets encompassing viral genomes and mRNAs probed with distinct methods and under distinct solution and cellular conditions. We conclude with suggestions of how RNA thermodynamic secondary structure models might continue to be improved and with discussion of design applications.

Results

Evaluated packages

We evaluated commonly used secondary structure modeling packages in their ability to make thermodynamic predictions on a compilation of large datasets of diverse synthetic molecules from Eterna, which we termed EternaBench (Figure 1A). The packages Vienna RNAfold, NUPACK, RNAstructure, RNAsoft, and CONTRAfold, were analyzed across different

package versions, parameter sets, and modelling options, where available (Supplementary Table 2). Packages discussed in the main text reflect their standard settings; the performance of other package options are included in Supplementary Information. We also evaluated packages trained more recently through a varied set of statistical or deep learning methods (LearnToFold, SPOT-RNA, MXfold, and CycleFold), but these packages demonstrated poor performance (Figure S1, S9) and have been omitted from further discussion.

Ranking of packages based on predictions for RNA chemical mapping

Our first ensemble-based structure prediction task investigates the capability of these packages to predict chemical mapping reactivities. Chemical mapping is a widely-used readout of RNA secondary structure[21-23] and has served as a high-throughput structural readout for experiments performed in the Eterna massive open online laboratory[20]. A nucleotide's reactivity in a chemical mapping experiment depends on the availability of the nucleotide to be chemically modified, and hence provides an ensemble-averaged readout of the nucleotide's deprotection from base pairing or other binding partners.[24] We wished to investigate if current secondary structure packages differed in their ability to recapitulate information about the ensembles of misfolded states that are captured in chemical mapping experiments. We used the Eterna "Cloud Labs" for this purpose: 23 datasets of player-designed constructs, ranging from 80-130 nucleotides in length, all of which were chemically modified by selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) and read out using the MAP-seq chemical mapping protocol.[25] After filtering for sequences with > 80% redundancy, these 23 datasets comprise 11,903 individual constructs (Figure 1A, "EternaBench-CM").

Figure 2A shows an example heatmap of SHAPE data for Eterna-player-designed synthetic RNA molecules from the Eterna Cloud Lab Round 69, the round with the most

constructs and a relatively high signal-to-noise ratio (Figure S2). Figure 2B shows two of these player-designed constructs in the minimum free energy (MFE) structures predicted in the Eterna gameplay with Vienna RNAfold 1.8.5 ("Vienna 1"), colored by experimental SHAPE reactivity data. Vienna 1 predicted that both sequences would form the same structure, and low reactivity in stem regions in construct (i) indicates that the molecule predominantly folded to the target structure. However, high reactivity values in the stems of construct (ii) indicate that the molecule did not fold to the correct structure, and instead remained largely unfolded, in conflict with the predictions of Vienna 1.

Figure 2C depicts package-predicted unpaired probabilities per nucleotide, p_{unp}, plotted in the same heatmap arrangement as the experimental data in Figure 2A (see Figure S3 for example heatmaps for all package options tested). In this subset of constructs, all packages are largely able to identify which regions should be completely paired (p_{unp} = 0, white) or unpaired (p_{unp} = 1, black), but some packages are able to predict p_{unp} values between 0 and 1 that more accurately reflect intermediate reactivity levels. Arrows (blue, green, magenta) indicate intermediate reactivity values that are captured by CONTRAfold and RNAsoft but not Vienna RNA models. Figure 2D depicts the reactivity of all nucleotides in Cloud Lab Round 69 (1,088 constructs) plotted vs. package-predicted p_{unp} values. The comparisons demonstrate improved correlation coefficients of experimental data to CONTRAfold 2 and RNAsoft BLstar predictions (0.718(2), 0.724(2), respectively) over Vienna 2, NUPACK, and RNAstructure (0.673(2), 0.667(1), 0.671(2), respectively). Correlation coefficient values for all package options are in Figure S1 and listed in Supplementary Table 3, and pairwise significance tests are in Figure S4.

We computed the correlation coefficients between chemical mapping data and p_{unp} values across the 23 Cloud Lab experiments (Figure 2E, Supplementary Table 3). These large-scale comparisons confirmed the ranking based on a single Cloud Lab experiment, with CONTRAfold 2 and RNAsoft Blstar outperforming Vienna 2, NUPACK, and RNAstructure in the

majority of experiments. Bootstrapping pairwise comparisons of package correlations to the data (see Online Methods) allowed us to create an overall package ranking, listed lowest to highest, of NUPACK, Vienna 2, RNAstructure, RNAsoft Blstar, and CONTRAfold 2 (Figure 2F, pairwise comparisons separated by dataset in Figure S5). For all packages, correlation coefficients were weakly positively associated with number of constructs in the dataset and strongly negatively correlated with the experimental noise in the data (Figure S2), confirming the importance of having large, high-quality data sets in evaluating these secondary structure packages.

We observed that CONTRAfold and RNAsoft generally predict that the constructs studied are more melted than the other packages predict at their default temperatures of 37 °C, even though the actual chemical mapping experiments were carried out at lower temperature (24 °C; see Methods). Motivated by this observation, we wished to ascertain if a simple change in temperature might account for differences in performance between packages. To address this, Vienna 2, NUPACK, and RNAstructure packages include parameters for both enthalpy and entropy, allowing for altering the temperature used in prediction. We found that increasing the temperature from the default value of 37 ˚C used in these packages to 60–70 °C improved their correlation to experimental data, but the maximum correlation values did not surpass the correlation observed from CONTRAfold or RNAsoft with default settings (Figure S6).

Ranking of packages based on predictions for riboswitch energetics

Our second ensemble-based structure prediction task involved predicting the relative populations of states occupied by riboswitch molecules. Riboswitches are RNA molecules that alter their structure upon binding of an input ligand, which effects an output action such as regulating transcription, translation, splicing, or the binding of a reporter molecule.[18, 26, 27] We

compared these packages in their ability to predict the relative binding affinity of synthetic riboswitches to their output reporter, fluorescently-tagged MS2 viral coat protein. These riboswitches were designed to use a small molecule input (flavin mononucleotide (FMN), tryptophan, or theophylline) to regulate formation of the MS2 hairpin aptamer (Figure 1A, "EternaBench-Switch"). These riboswitches came from two sources: the first consisted of 4,404 riboswitches designed by citizen scientists on Eterna,[28] filtered from the original dataset to exclude sequences with over 80% similarity. The second consisted of 2,202 riboswitches designed fully computationally using the RiboLogic package,[29] probed concomitantly with Eterna riboswitches.

Three important metrics that secondary structure packages can predict[17] for these datasets (depicted in Figure 3A, overview of experimental measurements in Figure S7) include MS2 coat protein binding affinity in the absence of input ligand, $\mathrm{K}_{\mathrm{MS2}}^{-lig}$, MS2 coat protein binding affinity in the presence of input ligand, $\mathrm{K}_{\mathrm{MS2}}^{+lig}$, and the ratio of these two values, termed the Activation Ratio (AR), which serves as a measure of the efficacy of the riboswitch,

$$\mathrm{AR} = \mathrm{K}_{\mathrm{MS2}}^{+lig}/\mathrm{K}_{\mathrm{MS2}}^{-lig}.$$

Figure 3B depicts predicted and experimental values for $\log K_{MS2}^{-lig}$ for FMN riboswitches from the RiboLogic dataset, the largest dataset included, with 2,202 constructs. Example scatterplots for all package options tested are in Figure S8, and correlations and pairwise significance tests for all package options are in Figure S9 and Supplementary Table 4. When performance across 11 independent experimental datasets was evaluated (Figure 3C, Figure S10) and pairwise comparisons bootstrapped over all datasets, we obtain a ranking, from lowest to highest, of NUPACK, Vienna RNAfold, RNAstructure, RNAsoft Blstar, and CONTRAfold, identical to the ranking obtained from chemical mapping data (Figure 3C, 3D, pairwise significance tests per dataset in Figure S11). Another metric to evaluate predictive accuracy is the root mean-squared

error (RMSE) between predicted and experimental values; this metric revealed a similar ranking led by CONTRAfold (Figure S12, Supplementary Table 4).

Predicting MS2 binding affinity in the presence of the riboswitch input ligand, $\mathrm{K}_{\mathrm{MS2}}^{+lig}$, as well as the activation ratio, AR, requires computing constrained partition functions, a capability limited to Vienna RNAfold and CONTRAfold (see Online Methods). CONTRAfold predictions of $\log K_{MS2}^{+lig}$ exhibited higher correlation to experimental values than Vienna 2 across the majority of data sets (>99% bootstraps, Figure 3D, Figure S14). For log AR, CONTRAfold predictions also exhibited higher correlation than Vienna 2 (87% of bootstraps; Figure 3E, Figure S15). Overall, these comparisons consistently ranked CONTRAfold 2 as the most accurate package for modeling riboswitches; this ranking matches the entirely independent ranking based on chemical mapping measurements of distinct RNA sequences described in the previous section.

EternaFold gives best-of-class performance in multiple structure prediction tasks

Given that the two packages that performed best in both structure prediction tasks were developed with statistical methods, we hypothesized that performance in these two tasks might be improved by incorporating these tasks in the process of training a secondary structure package. Multi-task learning, the process of learning tasks in parallel using a shared model, has proven useful in image classification and natural language processing.[30] For learning to be transferred across tasks, the tasks must share significant commonalities and structure. However, multi-task learning might fail if modelling assumptions made for one task do not hold across other tasks, or if the model's representational capacity is not large enough to correctly model all the data types. We used the CONTRAfold code as a framework to explore multi-task learning on RNA structural data, since it has previously been extended to train on chemical mapping data to maximize the expected likelihood of chemical mapping data.[31] We further extended the CONTRAfold loss function to include a term to minimize the mean squared error

of riboswitch affinities for MS2 protein (Online Methods). We trained models with a variety of combinations of data types to explore interactions in multitask training (Table 1) and evaluated performance on held-out test sets for single-structure prediction accuracy, chemical mapping prediction accuracy, and riboswitch fold change prediction. For single-structure data, we used the S-Processed dataset[32] used previously in training CONTRAfold 2 and RNAsoft[33].

Comparing performance across models trained with different types of input data indicates some tradeoffs in performance. Model "S", trained only on the S-Processed structure dataset, exhibited the highest accuracy for a held-out single-structure prediction test set (0.71(1)), outperforming models that included training on other data types. Likewise, Model "R", trained only on riboswitch K_{MS2}^{-lig} values, exhibited the highest performance in a held-out riboswitch K_{MS2}^{-lig} prediction test set. However, Model "SCRR", trained on four data types (structure, chemical mapping, riboswitch K_{MS2}^{-lig} and K_{MS2}^{+lig}) exhibited the highest performance on riboswitch K_{MS2}^{+lig} and chemical mapping, and its performance was within error of Models "S" and "R" on single structure prediction and K_{MS2}^{-lig} prediction test sets, respectively. We termed this SCRR model "EternaFold".

Independent tests confirm EternaFold performance

We wished to test if EternaFold's improvements in recovering Eterna measurements generalized to improvement in predictions for datasets from other groups, experimental protocols, and RNA molecules. We first tested the ability of EternaFold to predict the thermodynamics of protein binding not included in its training data, analogous to the MS2 binding comparisons above. We made use of a large data set of precisely measured $\Delta\Delta G$ values for binding of human Pumilio proteins 1 and 2 to a number of RNAs designed to include the consensus single-stranded Pumilio binding sequence.[34] EternaFold showed no significant increase or decrease in predictive ability when compared CONTRAfold or Vienna 2 (correlation

coefficients of 0.78(1), 0.78(1), and 0.80(1), respectively, Figure 4A; RMSE values of 0.70(2), 0.72(2), and 0.72(2), respectively, Figure S16, Supplementary Table 5).

We achieved stronger discrimination amongst packages by using a variety of *in vitro* and *in vivo* chemical mapping data collected with methods different from those used in the Eterna Cloud Labs (SHAPE-CE, SHAPE-MaP, DMS-MaP-seq vs. MAP-seq) and for distinct molecules, including viral genomes[21, 35-40] ribosomal RNAs[40, 41], and mRNAs designed to improve protein expression[19] (Table 2, Figure 4B). Most of these test molecules were much longer (thousands of nucleotides) than the 85-nucleotide RNAs used as the primary training data for EternaFold. Nevertheless, compared to all other packages tested, EternaFold exhibited the highest correlation in 6/19 datasets with $p < 0.01$ and an additional 3/19 datasets with $p < 0.05$ (Figure 4D, Figure S17, Supplementary Table 6, and demonstrated the highest correlation in a pairwise significance analysis bootstrapped over all datasets (Figure 4C). We were curious as to whether the differences in packages arise from consistent accuracy differences across all regions of these RNAs or from a net balance of increased and decreased accuracies at specific subregions of the RNAs, which might reflect particular motifs that are handled better or worse by the different packages. Figures 4D-F depict windowed correlation values along the length of representative molecules, the full-length HIV-1 and Venezeulan equine encephalitis virus genomes and mRNAs encoding luciferase. Generally, improvements between newer, statistically derived packages (RNAsoft, CONTRAfold, EternaFold) and previous packages are seen across the entire lengths of these RNAs. In molecules with noticeably improved predictions from EternaFold (see, e.g., Figure 4D), EternaFold gives better or similar predictions across all regions rather than showing regions with worse accuracy, supporting a consistent accuracy improvement by this package. Comparisons using different window sizes for structure prediction (Figures S18) and for other constructs over length 1000 (Figure S19) further support these observations.

Discussion

In this work, we have established EternaBench, benchmark datasets and analysis methods for evaluating package accuracy for two modeling tasks important in RNA structural characterization and design. These include 1) predicting unpaired probabilities, as measured through chemical mapping experiments, and 2) predicting relative stabilities of different conformational states, as exhibited in riboswitch systems. We discovered that RNAsoft and CONTRAfold, two packages that have had their thermodynamic parameters improved or estimated from scratch, respectively, by training on natural RNA secondary structures, showed significantly improved performance in these tasks over Vienna RNAfold, NUPACK, and RNAstructure, packages with parameters derived from thermodynamic experiments[9]. The results were particularly notable since the probed RNA molecules were designed for two distinct tasks EternaBench-CM and EternaBench-Switch, with no relationship between these two sets of sequences and no relationship between the synthetic sequences and natural sequences. We further investigated if combining these tasks in a multitask-learning framework could improve performance. We found that models trained on four types of data – single structures, chemical mapping data, and riboswitch affinities for both protein and small molecules – showed improved performance in predictions for held-out sub-sets of EternaBench datasets as well as improvements in datasets involving virus RNA genomes and mRNAs collected by five independent groups.

The improved performance of CONTRAfold and RNAsoft – two packages developed by maximum likelihood training approaches – was not obvious prospectively. Statistically-learned packages could incorporate bias towards common motifs in the RNA structures that they were trained on and might overstabilize motifs simply due to their increased frequency rather than actual thermodynamic stability.[42] Indeed, methods developed with a variety of more recent

methodological advances, including deep learning methods (SPOT-RNA), extended parameter sets (CONTRAfold-noncomplementary, CycleFold, MXfold), or accelerated folding packages (LearnToFold), demonstrated diminished performance in the EternaBench tasks (Figures S1,S9). It was surprising that well-developed and more widely used packages like ViennaRNA and RNAstructure gave worse performance than CONTRAfold and RNAsoft across all tasks. This observation might be rationalized by discrepancies in Mg^{2+}-free solution conditions used to measure these packages' thermodynamic parameters and the *in vitro* and *in vivo* conditions tested here. Our study underscores the need for independent measurements with distinct experimental methods and distinct molecules to test RNA secondary structure prediction packages.

Further advances in thermodynamic structure prediction might come from incorporating more data in a data-driven framework like EternaFold as well as advances in modelling. To analyze how much accuracy increases from incorporating more data, we compared differences in accuracy in models trained on the holdout dataset, roughly 40% the size of the full training dataset (Figure S20). The EternaFold model, when trained on this smaller data set, shows similar or slightly better accuracy in structure prediction and chemical mapping correlation coefficient to the full EternaFold model; there is no improvement with more data. However, including more riboswitch data does improve riboswitch performance on the test set. This observation suggests that improvements are likely to come from both including more data as well as through expansions in the EternaFold model. With regards to the latter, there is still significant room for improvement in RNA secondary structure prediction methodology, particularly in selecting which structural features to include in modelling. Feature selection involves a tradeoff: including too few features may result in a model unable to span the desired scope of structure space, but including too many features can lead to decreased performance through overfitting or antagonistic training on different tasks, especially if these features are not

well represented in the training data. By analyzing the parameter values of EternaFold models trained on different types of data, we found signs of degeneracy in parameters for dangling nucleotides, terminal mismatches, and 1x1 nucleotide internal loops, parameters that all reflect energetic contributions of nucleotides neighboring helices (Figure S21). Further data-driven investigations will be necessary to improve performance and aspects of the model that need to expanded, which may include noncanonical pairs[43], more sophisticated treatment of junctions[44], next-nearest-neighbor effects[14], and chemically modified nucleotides[45, 46].

This work demonstrates that RNA secondary structure prediction can be methodically evaluated and improved by combining thermodynamic predictions with statistical learning on diverse, high-throughput experimental datasets. As more sources of structural data continue to be made available in higher quantity, resolution, and scope, the EternaBench and EternaFold frameworks will allow for the usage of these datasets in continually assessing and improving RNA structure prediction. An important use case for EternaFold will be computationally guided design of RNA medicines, including structured mRNAs[19] for viral pandemics like COVID-19. EternaFold already affords increases in predictive power for structure mapping data acquired for synthetic mRNAs that model mRNA vaccines as well as for alphaviruses, which form the basis of self-amplifying mRNA vaccines[47, 48]. These observations suggest immediate applications for the packages herein and potentially rapid further improvements in EternaFold as further structural data are acquired on these new RNA molecules.

Acknowledgements

We thank members of the Das and Barna labs (Stanford University), C. Pop, and C.-S. Foo for useful discussions. We thank I. Jarmoskaite, V. V. Topkar, R. Wellington-Oguri, and J. Townley for helpful comments on the manuscript. Calculations and model training were performed on the

Stanford Sherlock cluster. We acknowledge funding from the National Science Foundation (GRFP to H.K.W.S.), the National Institute of Health (R35 GM122579 to R.D.), and gifts to the Eterna OpenVaccine project from Dave Baszucki and Jan Ellison, Bisconti Family Foundation, FastGrants, Baidu Research U.S.A., Camilla Kao, Patrick Collison, Ganesh Pusapati, Stephanie Federwisch, Yili Zhu, Roger Wellington-Oguri, Paul Berg, Gerry Smith, Feriel Melaine, Kabita Das, Rohan Koodli,Trent Smith, Phymirus LLC, Good Earth Bags, Beverly Morris Armstrong, Sagnik Ghosh, Gordon Stallings, Gordon Hug, Sachidulal Dhar, Naresh Saha, Laura Gable, Ro Ruan, Steven Clark, Tapan Ganguly, Anjali Paul, Sovan Banerjee, Cat Trivigno, Nagaraj Nagaraj, Upendra Bhadauria, A.K. Fazlur Rahman, Sumit Nag, Sushawn Nag, Samir Nayak, Anne Freedman, Leonard Oppenheimer, G N R Tripathi, Joseph McLaughlin, Gouri Datta, Amanda O'Mara, Subhash Gupta, Loretta Franklin, Amine Rehioui, and 10 anonymous donors.

Contributions

H.K.W.S. and R.D. designed the EternaBench benchmark approach and EternaFold multitask training method. H.K.W.S. prepared the EternaBench datasets, performed analyses, and implemented and trained the EternaFold model. H.K.W.S. and R.D. wrote the manuscript. W.K. designed methods, acquired data for high-throughput chemical mapping experiments, and reviewed the manuscript. Eterna participants created online design projects, provided RNA solutions, analyzed data, and reviewed the manuscript (see Supplemental Table 1).

Figure and table captions

Figure 1. A) EternaBench is a database of high-throughput data on diverse synthetic RNAs that can be used to evaluate secondary structure prediction packages in two different multi-state structure prediction tasks. The first dataset, EternaBench-CM, represents 11,903 constructs across 23 independent experimental rounds. Chemical mapping data can be compared to ensemble-wide probability of each nucleotide being paired or unpaired. The second dataset, EternaBench-Switch, contains 6,606 riboswitch constructs across 11 which can be used to evaluate predictions of base pair probabilities of protein-binding aptamers. B) Incorporating these data types into a multitask-trained package (EternaFold) results in a model that shows improved predictive power on completely independent datasets.

Figure 2. A) Heatmap representation of 60 example constructs from Eterna Cloud Lab Round 69. Representative partially reactive features are indicated with blue, green, and magenta arrows. B) Structures predicted in the Eterna gameplay with Vienna 1 of two example player-designed constructs, colored by experimental SHAPE reactivity data. Low reactivity in stems in (i) indicates predominantly correct structure prediction, whereas high reactivity in stems (ii) indicates incorrectly predicted structure. C) Example unpaired probabilities for the 60 example constructs, whose reactivity data are shown in (A), across 5 representative packages. Blue, green, magenta arrows indicate package predictions that recapitulate experimental partially reactive features. D) Plotting SHAPE reactivity values vs. predictions of p(unpaired) indicates that CONTRAfold and RNAsoft predictions for p(unpaired) have greater similarity to experimental reactivity data. E) Correlation values for Vienna 2, NUPACK, RNAstructure, CONTRAfold 2, and RNAsoft Blstar over 23 independent datasets. F) A pairwise significance analysis, bootstrapped over all datasets, reveals a ranking for predicting chemical reactivity

data, listed lowest to highest: NUPACK, Vienna 2, RNAstructure, RNAsoft Blstar, and CONTRAfold 2.

Figure 3. (A) Example set of states for a riboswitch that toggles binding of the fluorescent MS2 protein as an output, controlled by binding the small molecule FMN. The equilibrium constant for forming the MS2 aptamer in the absence of ligand, K_{MS2}^{-lig}, is estimated using the probability of forming the closing base pair for all packages (magenta). (B) Representative scatter plots for the largest dataset, RiboLogic-FMN riboswitches, of experimental vs. predicted values of K_{MS2}^{-lig}. (C) Correlation values between predicted and experimental measurements of K_{MS2}^{-lig} for 11 datasets of riboswitches. D) A pairwise significance analysis, bootstrapped over all datasets, reveals a ranking for predicting MS2 affinity, listed lowest to highest: NUPACK, Vienna 2, RNAstructure, RNAsoft Blstar, and CONTRAfold 2. Vienna and CONTRAfold are algorithmically capable of predicting K_{MS2}^{+lig} (correlations in E) as well as the activation ratio, $AR = K_{MS2}^{+lig}/K_{MS2}^{-lig}$ (correlations in F). In bootstrapping over all datasets, CONTRAfold 2 had higher correlation coefficient over Vienna 2 for K_{MS2}^{+lig} in > 99% of bootstraps, and higher correlation coefficient for AR in 87% of bootstraps.

Figure 4. A) In the one external RNA-protein binding set tested, EternaFold showed same correlation as Vienna (within error) at 37 °C. B) EternaFold shows increased correlation coefficient in 6 out of 19 independent chemical mapping datasets over all other evaluated packages with $p < 0.01$ (denoted with **), and 3 out of 19 datasets with $p < 0.05$ (denoted with *). C) Bootstrapped significance of pairwise package comparisons across all external datasets. D) Windowed correlation values (window size = 500) along reactivity data from the HIV-1 genome as measured in ref. [35], indicating that in particular regions, EternaFold shows pronounced improvement in correlation to experimental structure probing data. EternaFold had

the highest correlation coefficients with $p < 0.01$. E) A representative profile of correlation values for the Venezuelan Equine Encephalitis Virus (VEEV) genome[36], for which EternaFold gave the highest correlation coefficient with $p < 0.05$. F) Averaged windowed correlation profiles (window size = 100) for the 39 Luciferase constructs probed in ref. [19]. EternaFold had the highest correlation coefficient with $p < 0.01$. References in (A): Jarmoskaite, 2019[49]. References in (B): Watts, 2009[35]; Siegfried, 2014[38]; Rice, 2014[41]; Lavender, 2015[37]; Watters, 2018[21]; Kutchko, 2018[36]; Dadonaite, 2019[39]; Simon, 2019[40]; Mauger, 2019[19].

Table 1. Results of EternaFold multitask training with varying data type inputs and details of dataset types.

Table 2. Details of chemical mapping datasets used for external validation and correlation coefficient between predicted p_{unp} values and reactivity data across all packages tested.

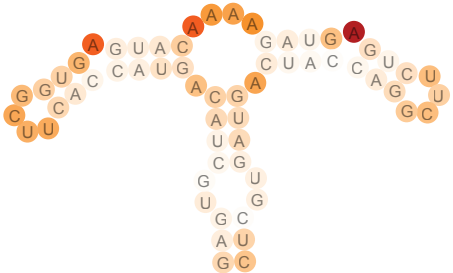
1. Amaral, P.P., Dinger, M.E., Mercer, T.R. & Mattick, J.S. The eukaryotic genome as an RNA machine. *Science* **319**, 1787-1789 (2008).
2. Singh, V., Braddick, D. & Dhar, P.K. Exploring the potential of genome editing CRISPR-Cas9 technology. *Gene* **599**, 1-18 (2017).
3. Jaffrey, S.R. RNA-Based Fluorescent Biosensors for Detecting Metabolites in vitro and in Living Cells. *Adv Pharmacol* **82**, 187-203 (2018).
4. Kramps, T. & Elbers, K. Introduction to RNA Vaccines. *Methods Mol Biol* **1499**, 1-11 (2017).
5. Zuker, M. & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* **9**, 133-148 (1981).
6. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
7. Zadeh, J.N. et al. NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem* **32**, 170-173 (2011).
8. Reuter, J.S. & Mathews, D.H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
9. Xia, T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719-14735 (1998).
10. Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H. & Murphy, K.P. in RNA, Vol. 16 2304-2318 (2010).
11. Do, C.B., Woods, D.A. & Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90-98 (2006).

12. Sloma, M.F. & Mathews, D.H. Base pair probability estimates improve the prediction accuracy of RNA non-canonical base pairs. *PLoS Comput Biol* **13**, e1005827 (2017).
13. Rezaur Rahman Chowdhury, F.A., Zhang, H. & Huang, L. Learning to Fold RNAs in Linear Time. *bioRxiv*, 852871 (2019).
14. Akiyama, M., Sato, K. & Sakakibara, Y. A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model. *J Bioinform Comput Biol* **16**, 1840025 (2018).
15. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications* **10**, 1-13 (2019).
16. Puton, T., Kozlowski, L.P., Rother, K.M. & Bujnicki, J.M. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res* **41**, 4307-4323 (2013).
17. Wayment-Steele, H., Wu, M., Gotrik, M. & Das, R. Evaluating riboswitch optimality. *Methods Enzymol* **623**, 417-450 (2019).
18. Berens, C. & Suess, B. Riboswitch engineering---making the all-important second and third steps. *Curr. Opin. Biotechnol.* **31**, 10-15 (2015).
19. Mauger, D.M. et al. mRNA structure regulates protein expression through changes in functional half-life. *Proc Natl Acad Sci U S A* **116**, 24075-24083 (2019).
20. Lee, J. et al. RNA design rules from a massive open laboratory. *Proc Natl Acad Sci U S A* **111**, 2122-2127 (2014).
21. Watters, K.E. & Lucks, J.B. Mapping RNA Structure In Vitro with SHAPE Chemistry and Next-Generation Sequencing (SHAPE-Seq). *Methods Mol Biol* **1490**, 135-162 (2016).
22. Wilkinson, K.A., Merino, E.J. & Weeks, K.M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc* **1**, 1610-1616 (2006).
23. Tian, S. & Das, R. RNA structure through multidimensional chemical mapping. *Q Rev Biophys* **49**, e7 (2016).
24. Eddy, S.R. Computational Analysis of Conserved RNA Secondary Structure in Transcriptomes and Genomes. *Annual Review of Biophysics* (2014).
25. Cordero, P., Lucks, J.B. & Das, R. An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics* **28**, 3006-3008 (2012).
26. Beisel, C.L. & Smolke, C.D. Design principles for riboswitch function. *PLoS Comput. Biol.* **5**, e1000363 (2009).
27. Breaker, R.R. Prospects for riboswitch discovery and analysis. *Mol. Cell* **43**, 867-879 (2011).
28. Andreasson, J.O.L. et al. Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular sensors. *bioRxiv*, 2019.2012.2016.877183 (2019).
29. Wu, M.J., Andreasson, J.O.L., Kladwang, W., Greenleaf, W. & Das, R. Automated Design of Diverse Stand-Alone Riboswitches. *ACS Synth Biol* **8**, 1838-1846 (2019).
30. Caruana, R. Multitask Learning. *Mach. Learn.* **28**, 41-75 (1997).
31. Foo, C.-S. & Pop, C. Learning RNA secondary structure (only) from structure probing data. *bioRxiv*, 152629 (2017).
32. Andronescu, M., Bereg, V., Hoos, H.H. & Condon, A. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics* **9**, 340 (2008).
33. Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H. & Murphy, K.P. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* **23**, i19-28 (2007).
34. Becker, W.R. et al. Quantitative high-throughput tests of ubiquitous RNA secondary structure prediction algorithms via RNA/protein binding. *bioRxiv*, 571588 (2019).

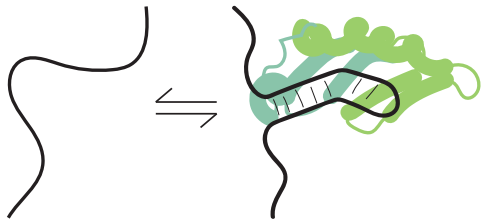
35. Watts, J.M. et al. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-716 (2009).
36. Kutchko, K.M. et al. Structural divergence creates new functional features in alphavirus genomes. *Nucleic Acids Res* **46**, 3657-3670 (2018).
37. Lavender, C.A. et al. Model-Free RNA Sequence and Structure Alignment Informed by SHAPE Probing Reveals a Conserved Alternate Secondary Structure for 16S rRNA. *PLoS Comput Biol* **11**, e1004126 (2015).
38. Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A. & Weeks, K.M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* **11**, 959-965 (2014).
39. Dadonaite, B. et al. The structure of the influenza A virus genome. *Nat Microbiol* **4**, 1781-1789 (2019).
40. Simon, L.M. et al. In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Res* **47**, 7003-7017 (2019).
41. GM, R., CW, L. & KM, W. RNA Secondary Structure Modeling at Consistent High Accuracy Using Differential SHAPE. *RNA (New York, N.Y.)* **20** (2014).
42. Rivas, E., Lang, R. & Eddy, S.R. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* **18**, 193-212 (2012).
43. Base pair probability estimates improve the prediction accuracy of RNA non-canonical base pairs. (2020).
44. M, W., A, D., M, W. & DH, M. Advanced Multi-Loop Algorithms for RNA Secondary Structure Prediction Reveal That the Simplest Model Is Best. *Nucleic acids research* **45** (2017).
45. BS, Z., IA, R. & C, H. Post-transcriptional Gene Regulation by mRNA Modifications. *Nature reviews. Molecular cell biology* **18** (2017).
46. Burnett, J.C. & Rossi, J.J. RNA-based Therapeutics- Current Progress and Future Prospects. *Chem Biol* **19**, 60-71 (2012).
47. Geall, A.J. et al. Nonviral delivery of self-amplifying RNA vaccines. (2012).
48. McKay, P.F. et al. Self-amplifying RNA SARS-CoV-2 lipid nanoparticle vaccine induces equivalent preclinical antibody titers and viral neutralization to recovered COVID-19 patients. (2020).
49. Jarmoskaite, I. et al. A Quantitative and Predictive Model for RNA Binding by Human Pumilio Proteins. *Mol Cell* **74**, 966-981.e918 (2019).

A. EternaBench: Evaluate multistate structure prediction

Chemical mapping reactivity



Riboswitch-protein affinity



Evaluate algorithms by predicting

prob(nucleotide unpaired)

prob(protein bound)

Eterna "Cloud Lab" Challenges
40,858 constructs

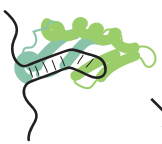
Eterna Riboswitch Challenges
28,706 constructs

Filtered for sequence redundancy, data quality

EternaBench-CM
11,903 constructs
23 datasets

EternaBench-Switch
6,606 constructs
11 datasets

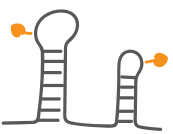
B. EternaFold: Multitask training improves external predictions

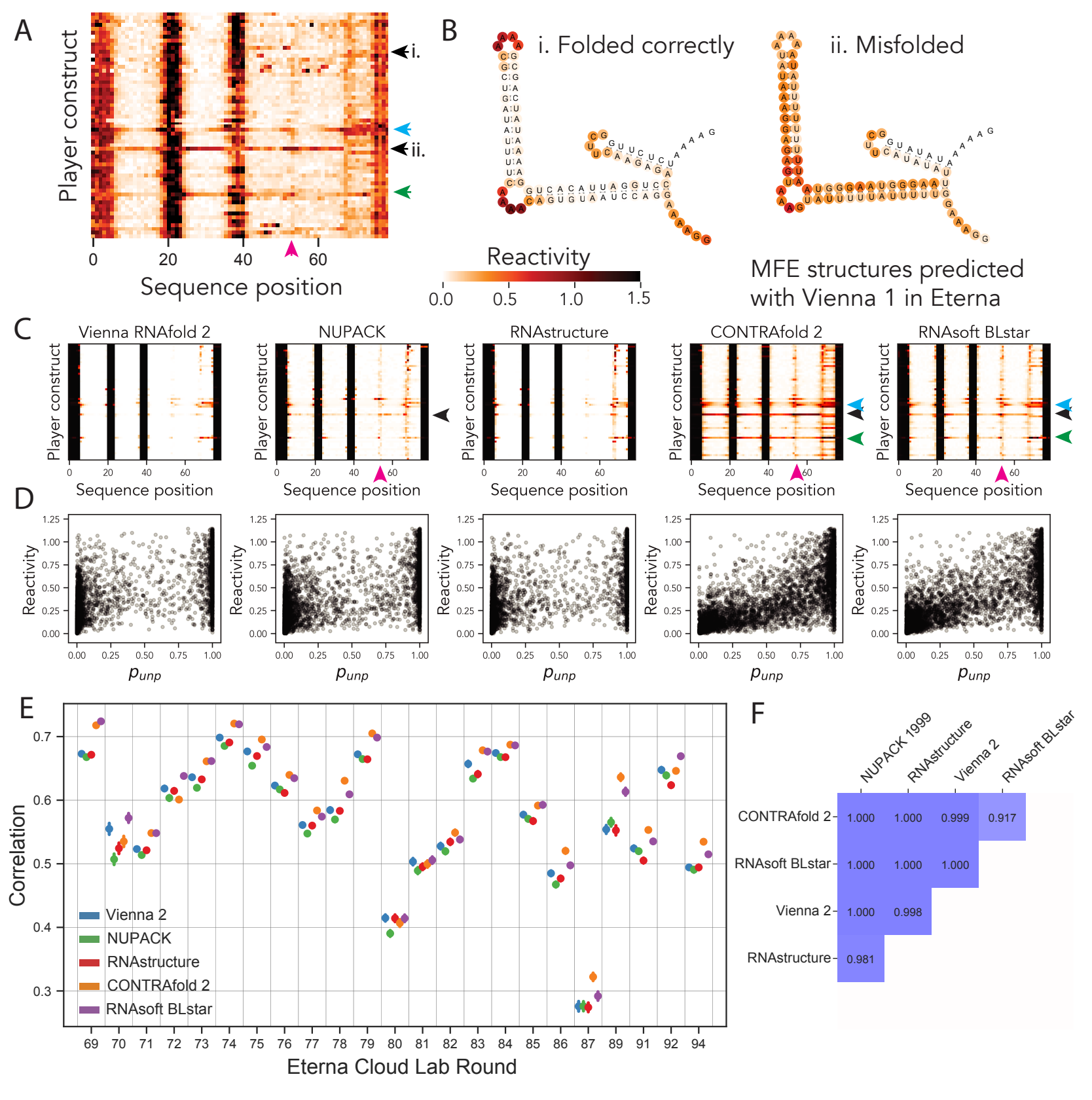


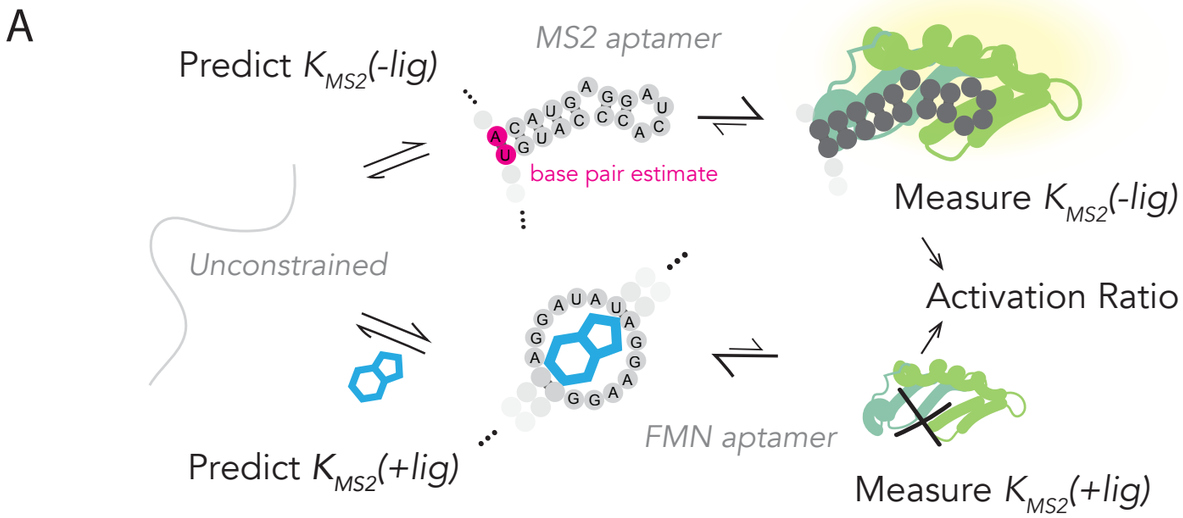
Structure data

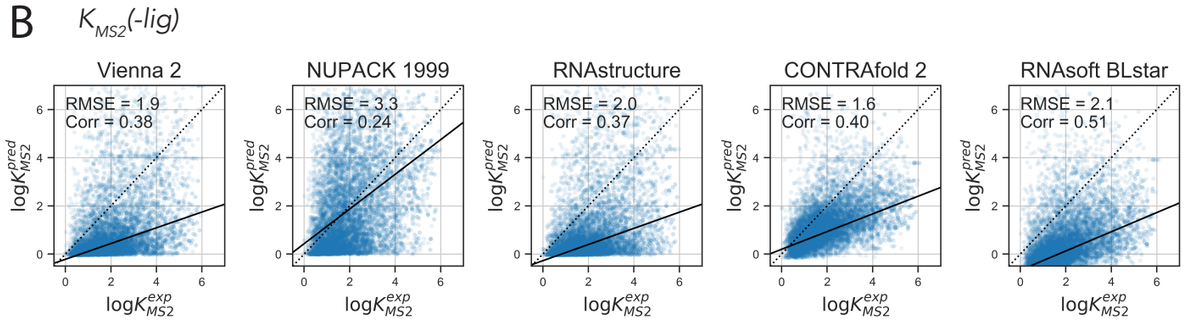
Conditional Log-linear Model

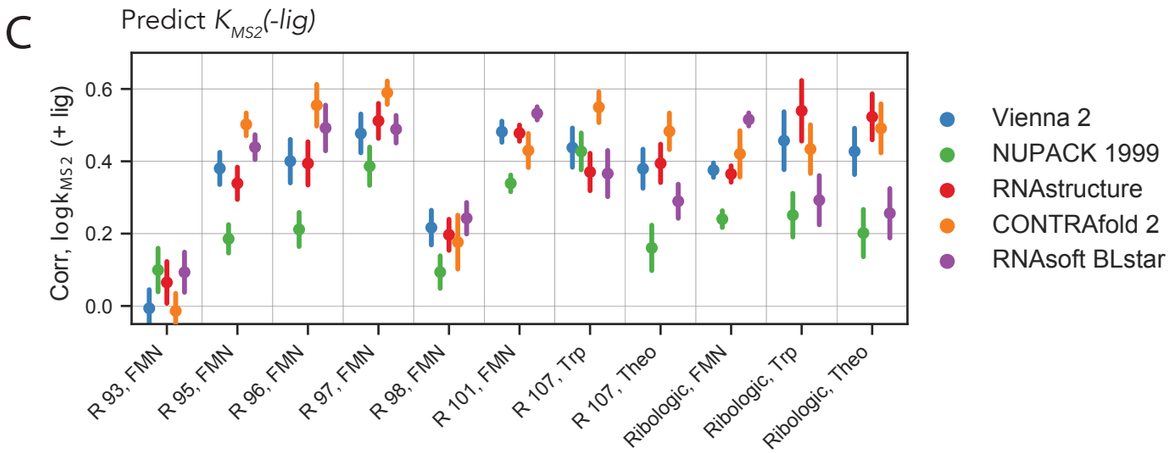
In Vitro/ In Vivo chemical mapping

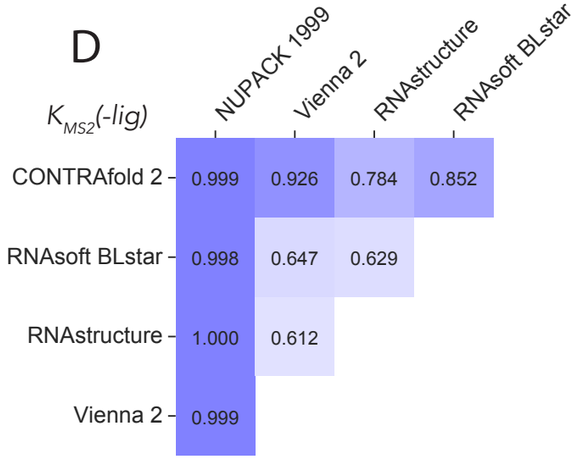


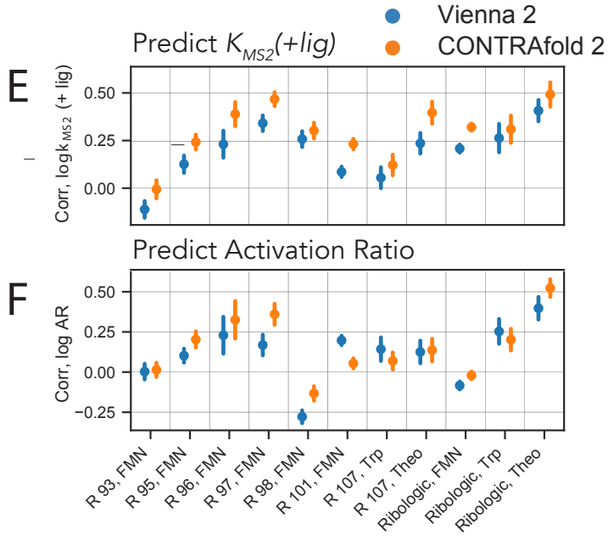

A
Player construct
i.
ii.
0
20
40
60
Sequence position
B
i. Folded correctly
ii. Misfolded
Reactivity
0.0
0.5
1.0
1.5
MFE structures predicted with Vienna 1 in Eterna
C
Vienna RNAfold 2
NUPACK
RNAstructure
CONTRAfold 2
RNAsoft BLstar
Player construct
Sequence position
D
Reactivity
p_{unp}
E
Correlation
Vienna 2
NUPACK
RNAstructure
CONTRAfold 2
RNAsoft BLstar
Eterna Cloud Lab Round
F
NUPACK 1999
RNAstructure
Vienna 2
RNAsoft BLstar
CONTRAfold 2
RNAsoft BLstar
RNAstructure
1.000
1.000
0.999
0.917
1.000
1.000
1.000
1.000
0.998
0.981

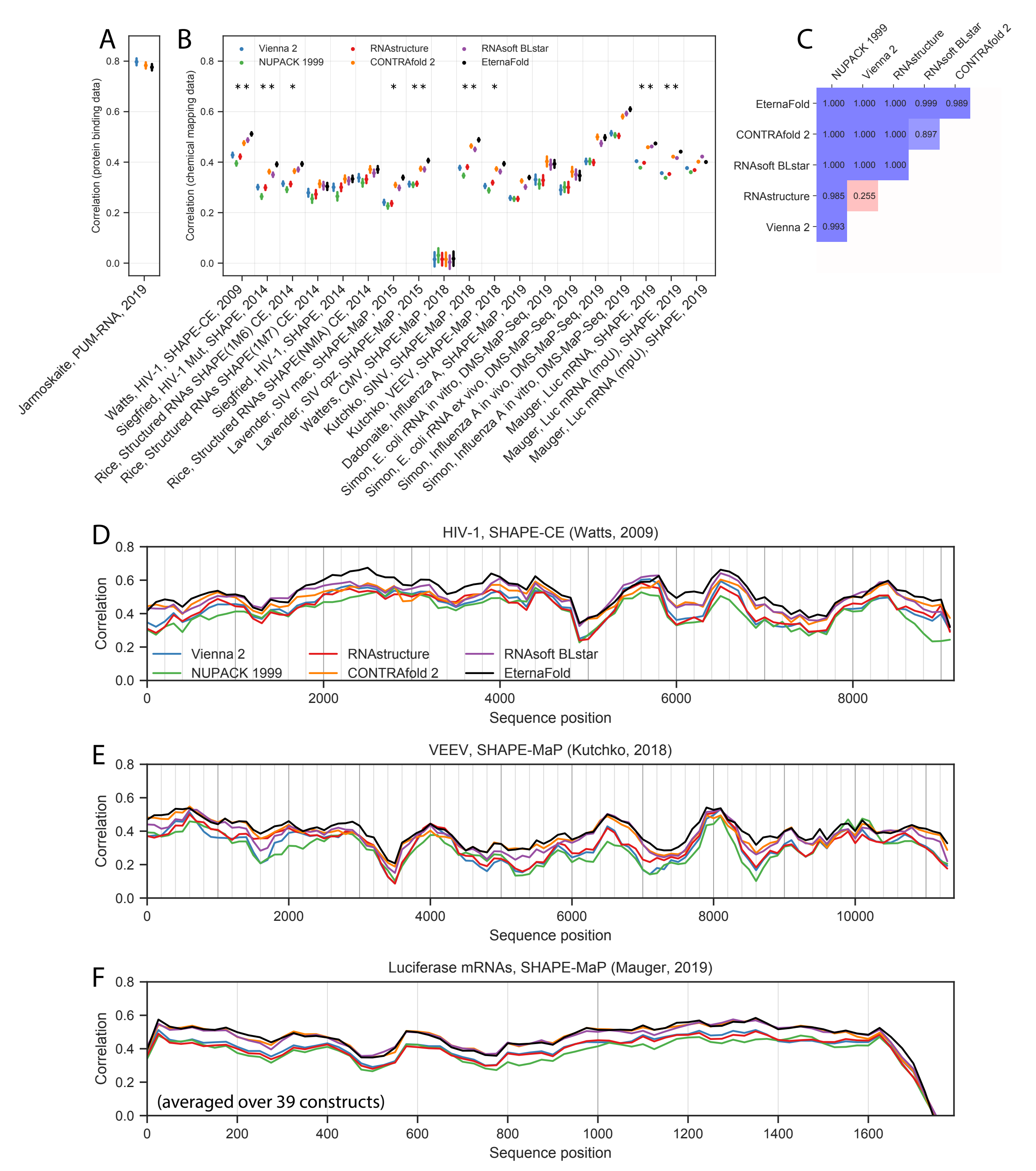

A
Predict K_{MS2}(-lig)
MS2 aptamer
base pair estimate
Measure K_{MS2}(-lig)
Unconstrained
Activation Ratio
FMN aptamer
Predict K_{MS2}(+lig)
Measure K_{MS2}(+lig)


B
K_{MS2}(-lig)
Vienna 2
RMSE = 1.9
Corr = 0.38
NUPACK 1999
RMSE = 3.3
Corr = 0.24
RNAstructure
RMSE = 2.0
Corr = 0.37
CONTRAfold 2
RMSE = 1.6
Corr = 0.40
RNAsoft BLstar
RMSE = 2.1
Corr = 0.51
$\log K_{MS2}^{pred}$
$\log K_{MS2}^{exp}$


C
Predict K_{MS2}(-lig)
Corr, $\log k_{MS2}$ (+ lig)
0.0
0.2
0.4
0.6
R 93, FMN
R 95, FMN
R 96, FMN
R 97, FMN
R 98, FMN
R 101, FMN
R 107, Trp
R 107, Theo
Ribologic, FMN
Ribologic, Trp
Ribologic, Theo
Vienna 2
NUPACK 1999
RNAstructure
CONTRAfold 2
RNAsoft BLstar


D
K_{MS2}(-lig)
NUPACK 1999
Vienna 2
RNAstructure
RNAsoft BLstar
CONTRAfold 2
0.999
0.926
0.784
0.852
RNAsoft BLstar
0.998
0.647
0.629
RNAstructure
1.000
0.612
Vienna 2
0.999


Vienna 2
CONTRAfold 2
E
Predict K_{MS2}(+lig)
Corr, $\log k_{MS2}$ (+ lig)
0.00
0.25
0.50
F
Predict Activation Ratio
Corr, log AR
−0.25
0.00
0.25
0.50
R 93, FMN
R 95, FMN
R 96, FMN
R 97, FMN
R 98, FMN
R 101, FMN
R 107, Trp
R 107, Theo
Ribologic, FMN
Ribologic, Trp
Ribologic, Theo


A
Correlation (protein binding data)
Jarmoskaite, PUM-RNA, 2019
B
Correlation (chemical mapping data)
Vienna 2
NUPACK 1999
RNAstructure
CONTRAfold 2
RNAsoft BLstar
EternaFold
Watts, HIV-1, SHAPE-CE, 2009
Siegfried, HIV-1 Mut, SHAPE, 2014
Rice, Structured RNAs SHAPE(1M6) CE, 2014
Rice, Structured RNAs SHAPE(1M7) CE, 2014
Siegfried, HIV-1, SHAPE, 2014
Rice, Structured RNAs SHAPE(NMIA) CE, 2014
Lavender, SIV mac, SHAPE-MaP, 2015
Lavender, SIV cpz, SHAPE-MaP, 2015
Watters, CMV, SHAPE-MaP, 2018
Kutchko, SINV, SHAPE-MaP, 2018
Kutchko, VEEV, SHAPE-MaP, 2018
Dadonaite, Influenza A, SHAPE-MaP, 2018
Simon, E. coli rRNA in vitro, DMS-MaP, 2019
Simon, E. coli rRNA ex vivo, DMS-MaP, 2019
Simon, Influenza A in vivo, DMS-MaP-Seq, 2019
Simon, Influenza A in vitro, DMS-MaP-Seq, 2019
Mauger, Luc mRNA, SHAPE, 2019
Mauger, Luc mRNA (moU), SHAPE, 2019
Mauger, Luc mRNA (mpU), SHAPE, 2019
C
NUPACK 1999
Vienna 2
RNAstructure
RNAsoft BLstar
CONTRAfold 2
EternaFold 1.000 1.000 1.000 0.999 0.989
CONTRAfold 2 1.000 1.000 1.000 0.897
RNAsoft BLstar 1.000 1.000 1.000
RNAstructure 0.985 0.255
Vienna 2 0.993
D
HIV-1, SHAPE-CE (Watts, 2009)
Correlation
Sequence position
E
VEEV, SHAPE-MaP (Kutchko, 2018)
Correlation
Sequence position
F
Luciferase mRNAs, SHAPE-MaP (Mauger, 2019)
Correlation
(averaged over 39 constructs)
Sequence position

Online Methods

The algorithms evaluated in this work model secondary structure in the following manner. Given a model Θ, which is comprised of a set of structural features $\{\theta\}$, the partition function of an RNA sequence *x* is computed as

$$Z(x|\Theta) = \sum_{s\in\{S\}}\sum_{k\in s}\exp(-\frac{\Delta G(\theta_k)}{k_B T}), \quad (1)$$

where $\Delta G(\theta_k)$ is the free energy contribution of structural feature *k*, k_B is Boltzmann's constant, and *T* is temperature. Z represents a sum over the set of all possible structures {*S*}. From this expression, the probability of any particular structure *s* is defined as

$$p(s|x,\Theta) = Z^{-1}\sum_{k\in s}\exp(-\frac{\Delta G(\theta_k)}{k_B T}). \quad (2)$$

Chemical mapping prediction

Structure prediction algorithms are able to estimate the ensemble-averaged probability that a nucleotide is paired or unpaired. Let $p(i{:}j|x,\Theta)$ be the probability of bases *i* and *j* being paired, given sequence *x* and model Θ. This may be computed as

$$p(i{:}j|x,\Theta) = \Sigma_{s_{i:j}\in\{S\}}\, p(s_{i:j}), \quad (3)$$

where $s_{i:j}$ denotes a structure containing the base pair i:j, and {S} is the full set of possible structures. These posterior probabilities can be analytically calculated by all the algorithms tested here. The probability of any single base being unpaired can be computed as

$$p(i\ unpaired) = 1 - \sum_j p(i{:}j). \quad (4)$$

The relationship between the probability of a nucleotide being unpaired and its experimentally-measured reactivity has served as a locus for many efforts for improving structure prediction of RNA constructs incorporating chemical mapping data from those constructs, and several

functional forms have been used to describe the relationship between unpaired probability and chemical mapping reactivity[1-3]. In this work, we use the correlation coefficient between unpaired probability and experimentally-measured reactivity as a nonparametric measure of model quality.

Riboswitch activity prediction

A thermodynamic framework discussed in greater detail in ref. [4] allows us to relate the observed binding affinity of an output molecule to the relative populations of a riboswitch molecule in different states. In the absence of input ligand, we may relate the probability that a riboswitch adopts a structural feature that can bind its output, $p(out)$, to an experimentally-measured binding affinity, $\mathrm{K}_{\mathrm{obs}}^{-lig}$, via the relative ratios of both values to those of a reference state:

$$\frac{K_{obs}^{-lig}}{K_{obs}^{ref}} = \frac{p^{ref}(out)}{p(out)} \equiv K_{MS2}^{-lig}. \qquad (5)$$

We selected the MS2 hairpin aptamer as a reference state whose probability of forming, $p^{ref}(out)$, can be estimated by the secondary structure algorithm. For each separate independent experimental dataset, K_{obs}^{ref} is estimated as the highest affinity measured (Figure S7). We refer to the estimated ratio $\frac{p^{ref}(out)}{p(out)}$ as K_{MS2}^{-lig} in the main text, as the equilibrium constant of forming the MS2 hairpin as normalized to the reference state (labeled in Figure 3A).

Although there may be error introduced in which experimental point is selected to be K_{obs}^{ref}, relative error should be constant when comparing packages on the same dataset. To compare packages, we report the correlation between $\log(K_{obs}^{\pm lig}/K_{obs}^{ref})$ and $\log(K_{MS2}^{\pm lig})$, which excludes the effect of selection for $\mathrm{K}_{\mathrm{obs}}^{ref}$. However, the Root Mean Squared Error (RMSE) between predicted and experimental values is also of interest to be able to consider error in terms of

energetic units (kcal/mol) and is reported in Supplementary tables S4, S5, S6, S7, and plotted in Figure S11 for the Ribologic-FMN dataset.

In general, the probability of any an RNA molecule forming any structure motif is computed as

$$p(motif|x,\theta) = \Sigma_{s_{motif}\in\{S\}} p(s_{motif}), \quad (6)$$

where s_{motif} denotes a structure containing that motif. Computing this probability requires a dynamic programming routine that is able to constrain the sampled structure space to only structures containing that motif to estimate a so-called "constrained partition function". However, not all secondary structure algorithms have implemented constrained partition function estimation. Because the MS2 aptamer is a hairpin, we can approximate its probability of forming as the probability of forming the final base pair of the MS2 hairpin aptamer (colored pink in Figure 3A), an experimental observable that can be estimated by all the packages tested here. Thus, our prediction of interest is

$$\frac{K_{pred}^{-lig}}{K_{pred}^{ref}} = \frac{\mathrm{p}^{\mathrm{ref}}(i{:}j)}{p(i{:}j)}, \qquad (7)$$

where *i* and *j* are the nucleotides forming the terminal base pair in the MS2 aptamer stem. The value $p^{ref}(i{:}j)$ is accordingly computed as the probability of closing the base pair in the reference sequence. We confirmed that calculations using eqn. 5 and eqn. 7 agree for Vienna and CONTRAfold packages. RNAstructure is capable of computing constrained partition functions, but the constrained partition function calculations did not match those from base pair probabilities (Figure S13). Hence, RNAstructure constrained partition function predictions were excluded from comparisons.

The estimation of $\mathrm{K}_{\mathrm{fold}}^{+lig}$ follows similarly to above but must take into account increased thermodynamic weights for states that correctly display the aptamer of the input small molecule

ligand. Therefore, it cannot be estimated via the simplified single base pair calculation and must make use of constrained partition functions (eq. 6).

For all constructs as well as the reference MS2 hairpin construct, we performed $K_{MS2}^{\pm lig}$ estimations including a flanking hairpin included in the Illumina array experiments (described in ref. [5]). In example, the full reference MS2 hairpin construct, as well as the constraint used for estimating K_{pred}^{ref} with constrained-partition-function-based estimation, is reproduced below. The MS2 hairpin construct is underlined and the nucleotides in the base used for base-pair-based prediction are bolded.

GGGUAUGUCGCAGAA**A**CAUGAGGAUCACCCAUG**U**AACUGCGACAUACCC

.....((((x((xxxx))))))).....

The riboswitches in EternaBench-Switch are controlled by the small molecules FMN, Tryptophan, or Theophylline. The motifs, concentrations, and intrinsic K_d values used for K_{MS2}^{+lig} prediction, taken from refs. [5] and [6], are repeated below:

Ligand	Aptamer sequence	Aptamer constraint	Concentration	K_d (μM)
FMN	nAGGAUAU&AGAAGGn	(xxxxxx(&)xxxxx)	200 μM	2.2
Tryptophan	AGGACCGG&CCGCCACU	((xxx(((&)))xxx))	2.4 mM	1.3
Theophylline	GAUACCAG&CCCUUGGCAGC	(xxx((((&)xxx)))xxx)	2.0 mM	20

Multi-task learning

In brief, the CONTRAfold[7] loss function optimizes the conditional log-likelihood of ground-truth structure $s^{(i)}$ given sequence $x^{(i)}$ over dataset D:

$$L_{CONTRAfold} = L_{Struct}(\theta) = \sum_{i \in D} log\ P(s^{(i)} |\ x^{(i)}, \{\theta\}). \qquad (8)$$

In CONTRAfold-SE[8], the authors include a term to also use chemical mapping data to optimize structure prediction by maximizing the likelihood of observing the included chemical mapping dataset. The loss function then becomes

$$L_{CONTRAfold-SE} = L_{Struct} + w_{CM} L_{CM}, L_{CM}(\theta, \phi) = \sum_{i \in D} \log \sum_{s}\ P(s, d\ | x, \{\theta\}, \phi), \qquad (9)$$

where *d* are the chemical mapping datapoints from construct *x*. CONTRAfold-SE fits reactivity signals to gamma distributions for each nucleotide type (A,C,G,U) and whether the base is paired or unpaired, parameters for which are represented by ϕ.

We further included a term to minimize the mean squared error of predicted $\log K_{fold}^{-lig}$ and $\log K_{fold}^{+lig}$:

$$L_{MS2} = w_{-lig}[\log K_{MS2}^{exp}(-lig) - \log K_{MS2}^{pred}(-lig)]^2 + w_{+lig}[\log K_{MS2}^{exp}(-lig) - \log K_{MS2}^{pred}(-lig)]^2\ . \quad (10)$$

The full loss function for EternaFold is thus written as

$$L_{EternaFold} = L_{Struct} + L_{CM} + L_{MS2}. \quad (11)$$

The hyperparameters $w_{CM}, w_{-lig}, w_{+lig}$, corresponding to the relative weights placed on different data types, were selected through a grid search on a holdout set. The final values used for training were $w_{CM} = 0.5,\ w_{-lig} = 30,\ w_{+lig} = 30.$

Code availability

The datasets used here for evaluation, as well as scripts and Python notebooks for reproducing the filtered datasets and the chemical mapping and riboswitch fold change calculations described here, are available at www.software.eternagame.org in the package "EternaBench". The code for training EternaFold, as well as the training and test sets used, are available at www.software.eternagame.org as the package "EternaFold". The EternaFold code is derived from the CONTRAfold-SE[8] codebase, which is derived from the CONTRAfold[7] codebase. A server to run EternaFold is being made available for noncommercial use.

Package predictions

All base-pairing probability calculations and constrained partition function calculations were performed using standardized system calls through Python wrappers developed in Arnie (www.github.com/DasLab/arnie). Example command-line calls for each package option evaluated are provided in Supplementary Table 2. Datasets were handled with Pandas (https://github.com/pandas-dev/pandas) and visualized with Seaborn (https://seaborn.pydata.org/).

Error and Significance Estimation

The following significance test was performed to make pairwise package comparisons for each type of data. For each bootstrapping round, the datasets within the data type (for instance, the 23 Cloud Lab chemical mapping rounds) were sampled with replacement, and within each of the sampled datasets, the datapoints were sampled with replacement. The correlations for both packages on the sampled datasets were calculated, and the number of datasets for which package A had a higher correlation than package B was recorded. If package A had a higher correlation in more than 50% of the datasets, this was counted as a "win" for package A. The

fraction across bootstrapping rounds that package A outperformed package B are reported in the final reported pairwise comparison matrices. Reported correlation and RMSE values reflect averages obtained over 1000 rounds of bootstrapping. Reported standard error values are estimated by calculating the standard deviation across bootstrapping rounds.

Chemical mapping data processing

Chemical mapping data for the Eterna Cloud Lab experiments were downloaded from RMDB[9] and processed with RDATKit (https://ribokit.github.io/RDATKit/). The RNA was probed with the MAP-seq protocol with a co-loaded standard molecule (P4-P6-2HP RNA) to enable normalization, as described in ref. [10]; measurements were carried out at ambient temperatures (24 ˚C) with 10 mM $MgCl_2$, 50 mM Na-HEPES, pH 8.0. Within each chemical mapping dataset, CD-HIT-EST[11] was used to filter sequences with greater than 80% redundancy (excluding a shared 3' primer binding site). From each sequence cluster identified, the sequence with the highest signal-to-noise ratio from chemical mapping experiments was selected as the representative sequence. Nucleotides with reactivities less than zero or greater than the 95th percentile of the dataset were removed from analysis. Cloud Lab Round 70 was filtered to exclude certain experiments that had FMN present, pertaining to Eterna Cloud Lab challenges to design riboswitches. Adenosine nucleotides preceded by 6 or more As were also removed due to evidence of anomalous transcription effects in such stretches [12], though this removal was not shown to alter package correlations to data (data not included).

External chemical mapping datasets were obtained from the supplementary information from the papers and processed similarly (outliers, nucleotides in poly-A stretches removed). For molecules longer than 600 nucleotides, *p(unp)* predictions were performed using a sliding window size of 600, with overlapping regions of length 25. Changing the window size was not shown to affect correlation values (Figure S18).

Riboswitch data processing

Riboswitch data were downloaded from supplementary materials from refs. [5] and [6]. Measurements were carried out at 37 °C in 100 mM Tris-HCl, pH 7.5, 80 mM KCl, 4 mM $MgCl_2$, 0.1 mg/mL BSA, 1 mM DTT, 0.01 mg/mL yeast tRNA, 0.01% Tween-20, and varying concentrations of small molecule ligand (FMN, tryptophan, theophylline) and MS2 coat protein. Datasets were filtered to only include constructs with sequences that included the canonical MS2 and small molecule aptamers, and filtered using CD-HIT-EST[11] to remove sequence redundancy over 80%.

1. Washietl, S., Hofacker, I.L., Stadler, P.F. & Kellis, M. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res.* **40**, 4261-4272 (2012).
2. Deng, F., Ledda, M., Vaziri, S. & Aviran, S. Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA* **22**, 1109-1119 (2016).
3. Eddy, S.R. Computational Analysis of Conserved RNA Secondary Structure in Transcriptomes and Genomes. *Annual Review of Biophysics* (2014).
4. Wayment-Steele, H., Wu, M., Gotrik, M. & Das, R. Evaluating riboswitch optimality. *Methods Enzymol* **623**, 417-450 (2019).
5. Andreasson, J.O.L. et al. Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular sensors. *bioRxiv*, 2019.2012.2016.877183 (2019).
6. Wu, M.J., Andreasson, J.O.L., Kladwang, W., Greenleaf, W. & Das, R. Automated Design of Diverse Stand-Alone Riboswitches. *ACS Synth Biol* **8**, 1838-1846 (2019).
7. Do, C.B., Woods, D.A. & Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90-98 (2006).
8. Foo, C.-S. & Pop, C. Learning RNA secondary structure (only) from structure probing data. *bioRxiv*, 152629 (2017).
9. Cordero, P., Lucks, J.B. & Das, R. An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics* **28**, 3006-3008 (2012).
10. Kladwang, W. et al. Standardization of RNA Chemical Mapping Experiments. *Biochemistry* **53**, 3063-3065 (2014).
11. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
12. Kladwang, W. et al. Anomalous reverse transcription through chemical modifications in polyadenosine stretches. *bioRxiv*, 2020.2001.2007.897843 (2020).