ARTICLES
https://doi.org/10.1038/s41592-022-01605-0

nature methods

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

Hannah K. Wayment-Steele[1,2], Wipapat Kladwang[2,3], Alexandra I. Strom[3,4], Jeehyung Lee[2,5], Adrien Treuille[2,5], Alex Becka[3], Eterna Participants* and Rhiju Das[2,3,6] ✉

Despite the popularity of computer-aided study and design of RNA molecules, little is known about the accuracy of commonly used structure modeling packages in tasks sensitive to ensemble properties of RNA. Here, we demonstrate that the EternaBench dataset, a set of more than 20,000 synthetic RNA constructs designed on the RNA design platform Eterna, provides incisive discriminative power in evaluating current packages in ensemble-oriented structure prediction tasks. We find that CONTRAfold and RNAsoft, packages with parameters derived through statistical learning, achieve consistently higher accuracy than more widely used packages in their standard settings, which derive parameters primarily from thermodynamic experiments. We hypothesized that training a multitask model with the varied data types in EternaBench might improve inference on ensemble-based prediction tasks. Indeed, the resulting model, named EternaFold, demonstrated improved performance that generalizes to diverse external datasets including complete messenger RNAs, viral genomes probed in human cells and synthetic designs modeling mRNA vaccines.

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 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 (ViennaRNA[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 use 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 (transfer RNAs, ribosomal RNA and so on) 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.

Data from high-throughput RNA structure experiments, such as high-throughput chemical mapping[20–22] and RNA-MaP experiments[23,24], offer the opportunity to make incisive tests of secondary structure models with orders-of-magnitude more constructs than previously. Unlike datasets of single secondary structures, both of these experiments provide ensemble-averaged structural properties, which allow for directly evaluating the full ensemble calculation of secondary structure algorithms, obviating the need to also evaluate the further nontrivial inference of a most-likely structure from the calculated ensemble. Furthermore, experimental data on human-designed synthetic RNA libraries have the potential to mitigate effects of bias incurred in natural RNA datasets.

In this work, we evaluate the performance of commonly used packages capable of making thermodynamic predictions in two tasks for which large datasets of synthetic RNAs have been collected via the RNA design crowdsourcing platform Eterna[25]: (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. 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 wider use.

We hypothesized that these data, although shorter than many natural RNAs of interest and not designed to bear similarity to natural RNAs, might still sufficiently represent RNA thermodynamics to allow for developing an improved algorithm. We tested this by developing 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

[1]Department of Chemistry, Stanford University, Stanford, CA, USA. [2]Eterna Massive Open Laboratory, Stanford, CA, USA. [3]Department of Biochemistry, Stanford University, Stanford, CA, USA. [4]Department of Chemistry and Biochemistry, San Diego State University, San Diego, CA, USA. [5]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. [6]Department of Physics, Stanford University, Stanford, CA, USA. *A list of members and their affiliations appears in the Supplementary Information. ✉e-mail: rhiju@stanford.edu

model, called EternaFold, did indeed demonstrate increased accuracy both on held-out data from Eterna as well as a collection of 31 independent datasets gathered from other literature sources, which encompass viral genomes, mRNAs and other small synthetic RNAs, probed with distinct methods and under distinct solution and cellular conditions. Compared to prior studies, this represents a very large collection of datasets used to evaluate RNA secondary structure algorithms.

Results

Evaluated packages. We initially evaluated commonly used secondary structure modeling packages in their ability to make thermodynamic predictions on two datasets of diverse synthetic molecules from Eterna: EternaBench-ChemMapping ($n=12{,}711$) and EternaBench-Switch ($n=7{,}228$). The packages ViennaRNA (v.1.8.5, 2.4.10), NUPACK (v.3.2.2), RNAstructure (v.6.2), RNAsoft (v.2.0) and CONTRAfold (v.1.0, 2.02) were analyzed across different package versions, parameter sets and modeling options where available (Supplementary Table 1). We also evaluated packages trained more recently through a varied set of statistical or deep learning methods (LearnToFold[13], SPOT-RNA[15], MXfold[14], CycleFold[12] and CROSS[26]), but these packages demonstrated poor performance on a subset of chemical mapping data (Extended Data Fig. 1a) and, due to their intensive runtimes, were omitted from further comparison.

Package ranking based on RNA chemical mapping predictions. 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[20–22] and has served as a high-throughput structural readout for experiments performed in the Eterna massive open online laboratory[25]. 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[27]. We wished to investigate whether current secondary structure packages differed in their ability to recapitulate information about the ensembles of misfolded states that are captured in chemical mapping experiments.

To make this comparison, we used the Eterna 'Cloud Labs' for this purpose: 24 datasets of 38,846 player-designed constructs, ranging from 80 to 130 nucleotides in length (dataset statistics in Supplementary Table 2, participant information in Supplementary Table 3). These constructs were designed in iterative cycles on the Eterna platform (Fig. 1a). Participants launched 'projects', each of which contained one 'target structure', and posed a design challenge or tested a hypothesis about RNA structure (project information in Supplementary Table 4). The constructs designed in these laboratories were periodically collected and mapped in vitro using selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) and the multiplexed accessibility probing read out through sequencing (MAP-seq) chemical mapping protocol[28]. These data were returned to participants, and the results guided future laboratory development and construct design[29].

The community of Eterna participants collectively developed highly diverse sequence libraries across target structures ranging from 0 to 12 loops (a proxy for design complexity[30]), as assessed by analyzing the positional sequence entropy of collected constructs as grouped by project (Fig. 1b). Example project target structures, colored by the mean reactivity of the probed solutions, are shown in Fig. 1b (inset). Some projects sought to design intricate structures, for example, 'The Nonesuch by rnjensen45' and 'Robot serial killer 1', while other participant projects focused more on better understanding experimental signals from particular structure motifs, for example, 'SHAPE Profile U-U mismatch', which consisted of a single stem and a U-U mismatch.

Figure 1a depicts an example heatmap of SHAPE data for Eterna-player-designed synthetic RNA molecules from the project 'Aires' by participant wateronthemoon. Figure 1c depicts calculated ensemble-averaged unpaired probabilities per nucleotide, P(unpaired), for five example package options, plotted in the same heatmap arrangement as the experimental data in Fig. 1a (see Extended Data Fig. 2 for heatmaps from all package options tested). In this subset of constructs, all packages are largely able to identify which regions are completely paired (P(unpaired) equal to 0, white) or unpaired (P(unpaired) equal to 1, black), but some packages predict P(unpaired) values between 0 and 1 that more accurately reflect intermediate reactivity levels. Arrows (blue, green and magenta) indicate intermediate reactivity values that are captured by predictions from CONTRAfold and RNAsoft but not ViennaRNA, NUPACK and RNAstructure. We quantified similarity between reactivity and P(unpaired) by calculating the Pearson correlation coefficient between the experimental reactivity values and P(unpaired) values (Methods). As an example, predictions from CONTRAfold 2 and RNAsoft BLstar for Cloud Lab round 1 (1,088 constructs) demonstrate improved correlation of $R=0.718(2)$ and 0.724(3) (respectively) over Vienna 2, RNAstructure and NUPACK (0.673(2), 0.671(2) and 0.667(2), respectively) (Supplementary Table 5). Noting that some projects had low sequence diversity, and to make the dataset a more manageable size for benchmarking while maintaining the same degree of sequence diversity, we filtered constructs to remove highly similar sequences (Methods and Extended Data Fig. 3). Clustering the resulting sequences per project (Fig. 1d) demonstrates that low-entropy projects were reduced in size. The final 24 EternaBench-CM datasets comprised 12,711 individual constructs.

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; Methods). Motivated by this observation, we wished to ascertain whether a simple change in temperature might account for differences in performance between packages. Because ViennaRNA, NUPACK and RNAstructure packages include parameters for both enthalpy and entropy, we calculated correlations across predictions from a range of temperatures (Extended Data Fig. 1b). We found that increasing the temperature from the default value of 37 °C used in these packages to 60 °C improved the correlation to experimental data for ViennaRNA ($R=0.708(2)$) and RNAstructure ($R=0.707(2)$), but not NUPACK ($R=0.639(2)$). We hence included each of these packages also at 60 °C as options to test.

We established a ranking of all package options for each dataset (Fig. 1e, Supplementary Table 5 and representative heatmaps for all datasets in Extended Data Fig. 4) by computing the z-score for each package correlation in comparison to all packages tested, and averaging over all datasets (Fig. 1f). The top three package options were CONTRAfold 2, ViennaRNA at 60 °C and RNAsoft with 'BLstar' parameters. Using a Pearson correlation assumes a linear relationship between P(unpaired) and reactivity and relies on a two-state model with inherent limitations (Methods). We therefore also ranked all packages with a Spearman rank correlation coefficient and found a similar global overall ranking (Extended Data Fig. 1c). Overall package performance and the resulting ranking was not strongly dependent on guanine-cytosine (GC) content, sequence length or total number of loops in the project target structure, which was investigated by calculating correlations and rankings when grouping constructs by project (Methods and Extended Data Fig. 1d).

Package ranking based on riboswitch affinity predictions. Our second ensemble-based structure prediction task involved

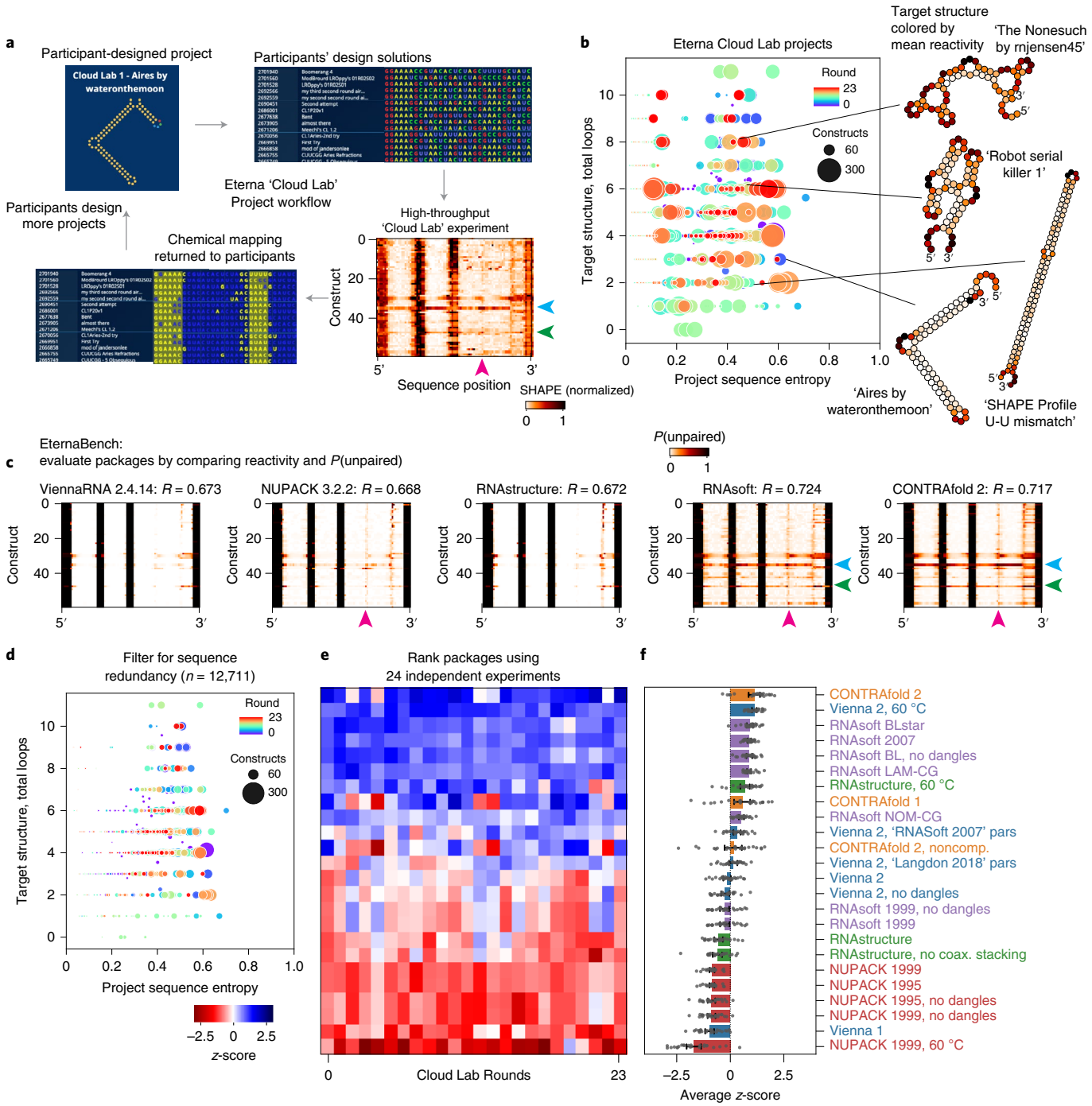


Fig. 1 | Community-science-designed RNA datasets from the Eterna 'Cloud Lab' experiments identify consistent discrepancies in ensemble calculations from secondary structure packages. a, Workflow of Cloud Lab rounds: Eterna participants design 'projects', typically intended as RNA design challenges. Players submit solutions, all of which are synthesized in high-throughput via MAP-seq experiments. Example reactivity data are depicted from the project 'Aires' by participant wateronthemoon. Data are returned to participants in the in-game browser, which served as the basis for more player-designed projects. norm., normalized. **b**, Calculating the average positional entropy for all solutions collected for each project reveals that participants were able to design a diverse set of solutions, independent of target structure complexity (monitored as number of loops in the target structure). Example target structures are colored by average reactivity. **c**, Example unpaired probabilities for 60 example constructs from the project 'Aires', for which reactivity data are shown in **a**, across five representative secondary structure packages. Blue, green and magenta arrows indicate package predictions that recapitulate experimental partially reactive features. CONTRAfold and RNAsoft predictions for *P*(unpaired) have higher correlation to experimental reactivity data. **d**, Analogous representation to **b** for the redundancy-filtered EternaBench dataset. **e**, We compared many commonly used packages and secondary structure prediction options over 24 Cloud Lab independent experiments. We calculated the Pearson correlation coefficient and calculated the *z*-score across all packages evaluated for each dataset. **f**, Final ranking is obtained by averaging the *z*-scores obtained across all datasets. Error bars represent 95% confidence interval of the mean obtained over 1,000 iterations of bootstrapping over $n = 24$ independent experiments, which comprised 12,711 independent constructs in total.

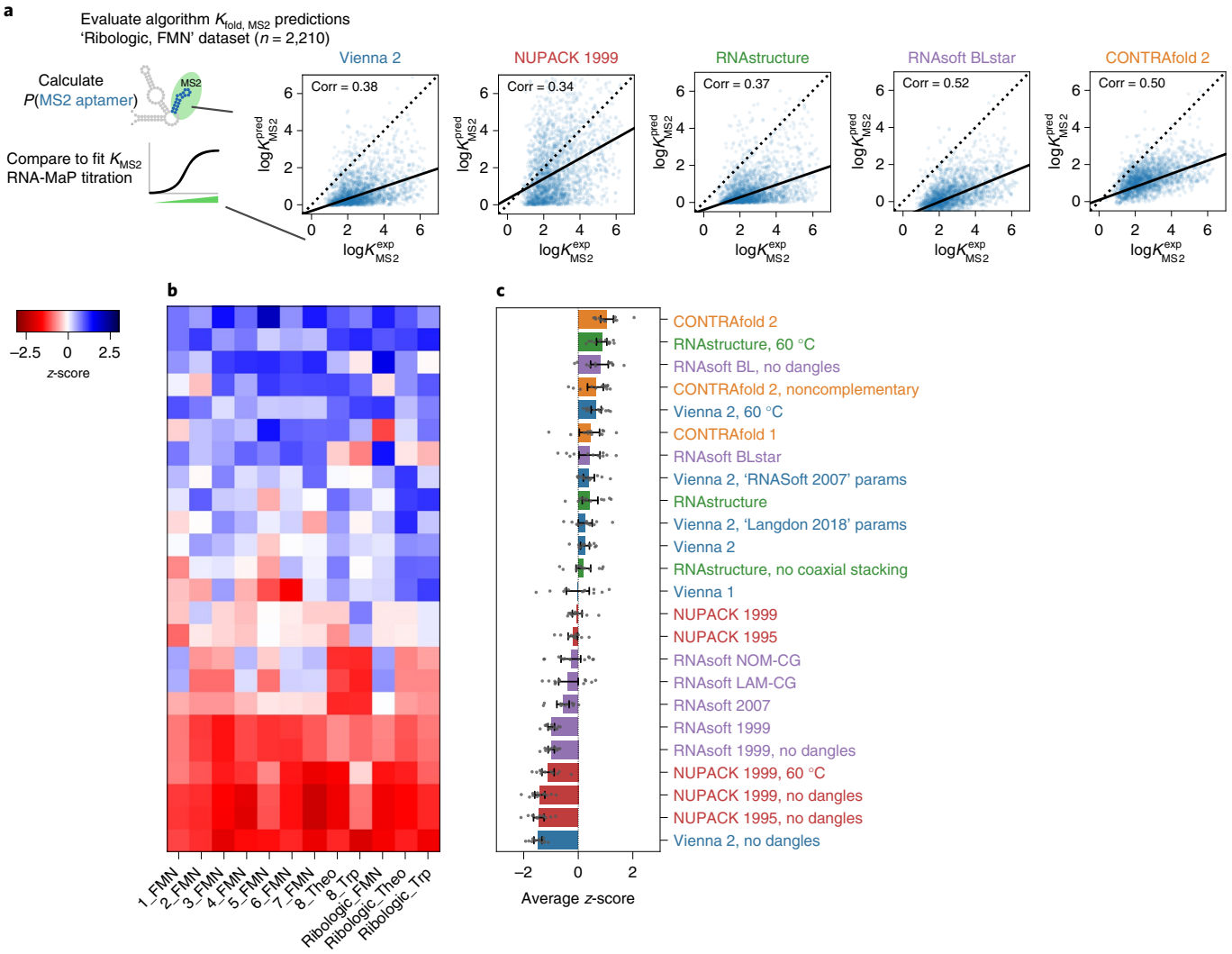


Fig. 2 | Riboswitch affinity predictions reveal similar package ranking. a, Representative scatterplots for the Ribologic-FMN dataset of experimental versus predicted values of K^{-lig}_{MS2}. Corr, Pearson correlation. **b**,**c**, Calculating the *z*-scores of Pearson correlation coefficients across 12 independent datasets of riboswitches (**b**) result in an overall ranking (**c**) consistent with the chemical mapping dataset. Error bars represent 95% confidence interval of the mean obtained over 1,000 iterations of bootstrapping over $n = 12$ independent experiments of 7,228 independent constructs in total.

predicting the relative populations of states occupied by riboswitch molecules. Riboswitches are RNA molecules that alter their structure on binding of an input ligand, which effects an output action such as regulating transcription, translation, splicing or the binding of a reporter molecule[18,31,32]. 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 in the absence of input ligand, K^{-lig}_{MS2} (Methods and Extended Data Fig. 5a). As with the chemical mapping datasets, each riboswitch dataset was filtered to exclude highly similar sequences (Extended Data Fig. 3 and Supplementary Table 6). These riboswitches came from two sources: the first consisted of 4,849 riboswitches (after filtering) designed by citizen scientists on Eterna[33]. The second consisted of 2,509 riboswitches (after filtering) designed fully computationally using the RiboLogic package[34], probed concomitantly with Eterna riboswitches. These riboswitches were designed using aptamers for three small molecules: flavin mononucleotide (FMN), theophylline and tryptophan.

Figure 2a depicts experimental values for $\log K^{-lig}_{MS2}$ for FMN riboswitches from the RiboLogic dataset versus predicted $\log K^{-lig}_{MS2}$ values. Again, CONTRAfold and RNAsoft BLstar packages exhibit higher correlations to the experimental data (Pearson $R = 0.50(2)$ and 0.51(2), respectively) than ViennaRNA, NUPACK and RNAstructure ($R = 0.37(2)$, 0.34(2), 0.36(2), respectively). Example predictions for all package options tested are in Extended Data Fig. 6. We evaluated performance across 12 independent experimental datasets (Fig. 2b, Supplementary Table 7 and representative predictions in Extended Data Fig. 7), and obtained a ranking (Fig. 2c) similar to the ranking obtained from chemical mapping data. CONTRAfold 2, RNAsoft (model 'BL, no dangles', equivalent to BLstar but without dangles) and RNAstructure 60 °C were ranked as the top three out of the package options tested. The top ranking of CONTRAfold 2 matches the entirely independent ranking based on chemical mapping measurements of distinct RNA sequences described in the previous section. These riboswitches were designed using aptamers for three small molecules: FMN, theophylline and tryptophan. Calculating *z*-scores over each individual subset resulted in slightly differing rankings but consistently favored Contrafold methods (Extended Data Fig. 5b). Predicting MS2 binding affinity in the presence of the riboswitch input ligand, K^{+lig}_{MS2}, as well as

Table 1 | Ranking by *z*-score over 24 chemical mapping datasets ($n = 12{,}711$ constructs), 12 riboswitch datasets ($n = 7{,}228$ constructs) and averaged over both dataset types

Package	ChemMapping *z*-score mean (s.d.)	Riboswitch *z*-score mean (s.d.)	Both dataset types mean (s.d.)
CONTRAfold 2	**1.14 (0.69)**	**1.03 (0.43)**	**1.09 (0.61)**
Vienna 2, 60 °C	1.12 (0.29)	0.65 (0.34)	0.89 (0.38)
RNAsoft BL, no dangles	0.88 (0.34)	0.79 (0.57)	0.84 (0.43)
RNAstructure, 60 °C	0.71 (0.57)	0.86 (0.36)	0.78 (0.51)
RNAsoft BLstar	0.93 (0.36)	0.42 (0.67)	0.67 (0.53)
CONTRAfold 1	0.57 (0.99)	0.45 (0.65)	0.51 (0.88)
CONTRAfold 2, noncomplementary	0.15 (1.01)	0.66 (0.53)	0.40 (0.91)
Vienna 2, 'RNASoft 2007' params	0.33 (0.45)	0.38 (0.37)	0.35 (0.42)
RNAsoft LAM-CG	0.87 (0.25)	−0.37 (0.69)	0.25 (0.74)
Vienna 2, 'Langdon 2018' params	0.13 (0.48)	0.26 (0.46)	0.20 (0.47)
RNAsoft 2007	0.89 (0.21)	−0.54 (0.40)	0.17 (0.74)
RNAsoft NOM-CG	0.52 (0.30)	−0.26 (0.66)	0.13 (0.58)
Vienna 2	−0.15 (0.47)	0.25 (0.30)	0.05 (0.46)
RNAstructure	−0.55 (0.50)	0.43 (0.51)	−0.06 (0.68)
RNAstructure, no coaxial stacking	−0.60 (0.55)	0.19 (0.51)	−0.21 (0.65)
NUPACK 1999	−0.86 (0.30)	−0.06 (0.33)	−0.46 (0.49)
Vienna 1	−0.96 (0.56)	−0.02 (0.79)	−0.49 (0.78)
NUPACK 1995	−0.86 (0.31)	−0.19 (0.33)	−0.52 (0.45)
RNAsoft 1999	−0.27 (0.55)	−0.98 (0.22)	−0.63 (0.58)
RNAsoft 1999, no dangles	−0.27 (0.55)	−0.99 (0.22)	−0.63 (0.58)
Vienna 2, no dangles	−0.24 (0.47)	−1.48 (0.26)	−0.86 (0.72)
NUPACK 1999, no dangles	−0.88 (0.47)	−1.42 (0.36)	−1.15 (0.50)
NUPACK 1995, no dangles	−0.88 (0.47)	−1.44 (0.35)	−1.16 (0.51)
NUPACK 1999, 60 °C	−1.72 (0.89)	−1.13 (0.41)	−1.42 (0.81)

Standard deviation of *z*-score over datasets in parentheses. The top-performing package for each is in bold.

the activation ratio requires computing constrained-partition functions, a capability limited to Vienna RNAfold, RNAstructure and CONTRAfold. Rankings for predicting K_{MS2}^{+lig} and activation ratio followed the same trends (Extended Data Fig. 5d,e and Methods).

EternaFold gives best-of-class performance in multiple tasks. We hypothesized that performance in both secondary structure prediction tasks above might be improved by incorporating these tasks in the process of training a secondary structure package. The RNAsoft[10,35] and CONTRAfold[11] packages, which performed well in both tasks (Table 1), both take advantage of the property that the gradient of the partition function with respect to any feature is related to the expected counts of that feature[10], which can be readily computed in dynamic programming scheme. We generalized this framework beyond maximizing the likelihood of one single structure to matching the experimentally determined probability of a particular structural motif in the ensemble through minimizing the root-mean-squared error to the logarithm of riboswitch affinities for MS2 protein (Methods). We used the CONTRAfold code as a framework to explore multitask 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[36].

We tested training from three data types: secondary structures, chemical mapping reactivity and riboswitch affinities. We used the STRAND S-Processed dataset for secondary structures ($n = 3{,}439$), which was the same data used to train RNAsoft and CONTRAfold[10]. The chemical mapping training data ($n = 2{,}603$) came from Cloud Lab datasets used in previous model development[36]. We used riboswitches designed by the automated RiboLogic[34] algorithm for riboswitch training data ($n = 1{,}295$). We trained models with a variety of combinations of data types to explore interactions in multitask training (Fig. 3a), used holdout sets to determine hyperparameter weights (Methods) and evaluated performance on separate test sets for single-structure prediction accuracy[37], chemical mapping prediction accuracy and riboswitch affinity prediction. To ensure a rigorous separation of training and test data, each test dataset was filtered for sequence similarity to all training data at 80% using a windowed Levenshtein metric (Methods). Marked sequence similarity overlap between the S-Processed train and test sets motivated us to develop an orthogonal dataset for secondary structure prediction testing based on the dataset ArchiveII[38]. Test sets for chemical mapping and riboswitch data came from completely different experimental rounds than those used in training to avoid learning experiment-specific biases.

Comparing performance across models trained with different types of input data indicates some tradeoffs in performance. CONTRAfold 2 exhibited the highest accuracy, followed by 'Model S', trained only on single-structure prediction training data, exhibited the highest accuracy on the separate single-structure prediction test set (Fig. 3b, *F*-score = 0.56(0.22)). Incorporating other data types in model training resulted in *F*-scores worse than Model S on the ArchiveII-NR single-structure prediction test set but within error of CONTRAfold 2 (Fig. 3b). Model 'SCRR', trained on four data types (single-structure data, chemical mapping, riboswitch K_{MS2}^{-lig} and K_{MS2}^{+lig}) exhibited the highest performance on separate test sets for chemical mapping (Fig. 3c) and riboswitch K_{MS2}^{-lig} prediction (Fig. 3d, data for all test sets in Supplementary Table 8). We termed this SCRR model 'EternaFold'.

Independent tests confirm EternaFold performance. We wished to test whether EternaFold's improvements in correlating *P*(unpaired) values to chemical mapping and protein-binding data generalized to improvement in predictions for datasets from other groups, experimental protocols and RNA molecules. We compiled 3 datasets of chemical mapping data for molecules including viral genomes[39–49] in cells and in virions, ribosomal RNAs[44,50,51] both in cells and extracted from cells, synthetic mRNAs and RNA fragments designed to improve protein expression and in vitro stability[19,52], and mRNAs probed in various subcellular compartments and extracted from human embryonic kidney 293 (HEK293) cells[53] (Fig. 4a and Supplementary Table 9). These datasets spanned structure probing methods different from those used in the Eterna Cloud Labs (SHAPE-CE, SHAPE-MaP, DMS-MaP-seq versus MAP-seq) as well as a variety of chemical modifications (DMS, icSHAPE, NAI). Most of these test molecules were much longer (thousands of nucleotides) than the 85-nucleotide RNAs used as the primary training data for EternaFold. Notably, six of these involved the SARS-CoV-2 genome[46–48], which came into prevalence after the development of the EternaFold model, and represented a test of new data. The following results are for *P*(unpaired) values calculated for overlapping windows of size 900, but other window sizes and Levenshtein distance metrics gave qualitatively similar results (Extended Data Fig. 8).

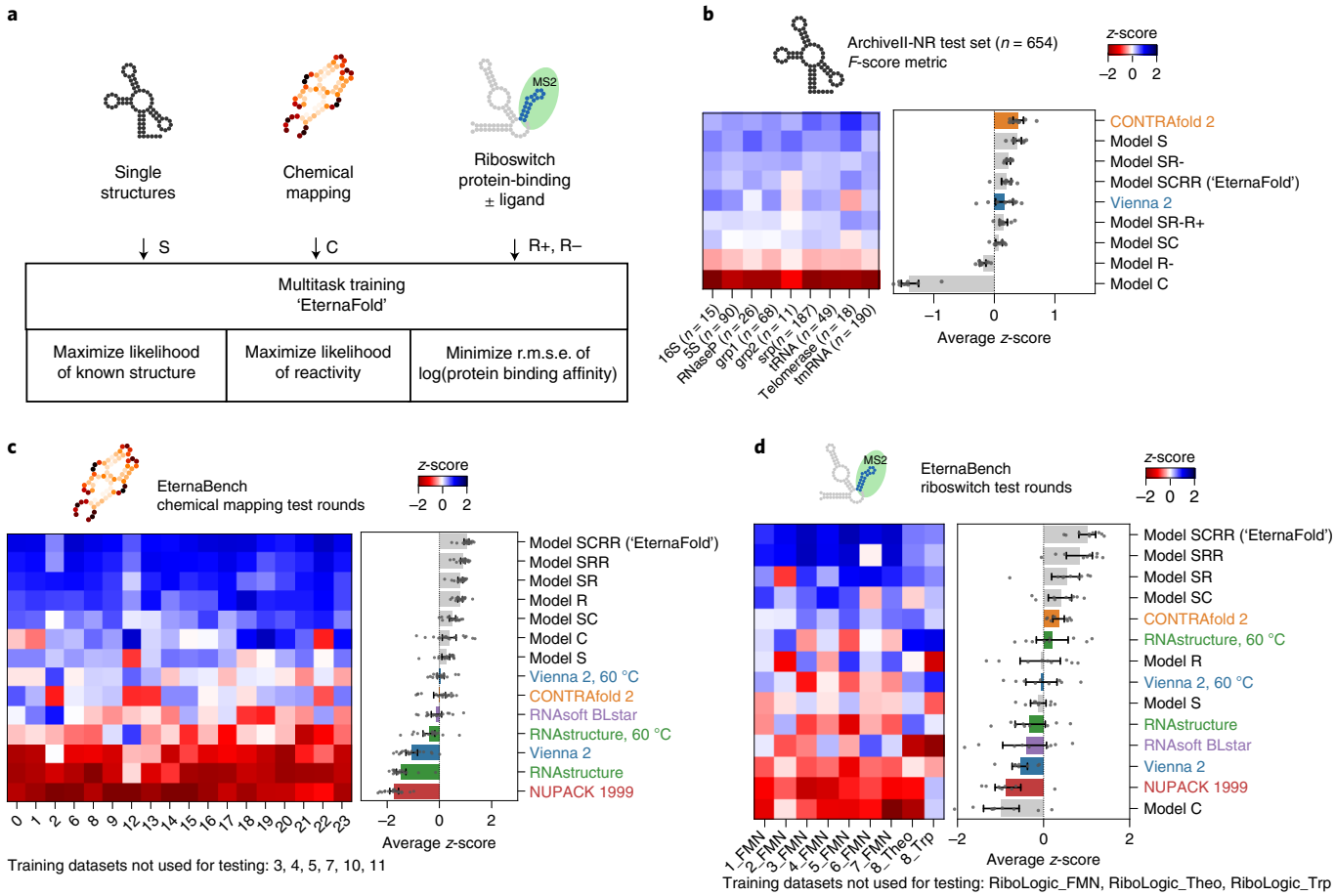


Fig. 3 | Multitask training using EternaBench datasets results in improved thermodynamic prediction. a, Scheme of data types used in multitask training and loss function used for each. R.M.S.E., root mean-squared error. **b**, Secondary structure prediction on ArchiveII-NR test set, prepared to contain <80% sequence similarity to secondary structure training data (Methods), $n = 9$ independent datasets with 654 constructs total. **c**, *z*-score ranking over 18 test datasets for EternaBench chemical mapping filtered to contain constructs with <60% sequence similarity to all training data, $n = 18$ datasets with 1,492 independent constructs total. **d**, *z*-score ranking over nine riboswitch test sets for riboswitch K_{MS2} prediction filtered to contain constructs with <80% sequence similarity to all training data, $n = 9$ datasets with 4,018 independent constructs total. In **b**–**d**, Error bars represent 95% confidence interval of the mean obtained over 1,000 iterations of bootstrapping.

We wished to ascertain that the sequences in these datasets did not overlap with sequences that EternaFold had been trained on, so we also filtered these data using a windowed Levenshtein distance metric at a cutoff of 60% sequence similarity. This removed 37% of the originally collected sequences for a dataset size of 8,734 sequences (Supplementary Table 10).

For 15 out of 31 datasets across all categories, EternaFold exhibited the highest correlation coefficient (with $P < 0.05$, determined by 95% overlapping confidence intervals; Methods), and had the highest average *z*-score (Fig. 4b and Table 2). For the other 16 datasets, EternaFold was tied with other packages for having the highest correlation. EternaFold showed significant improvement ($P < 0.05$) in datasets from varying sources including RNAs probed in cell (5 of 7 in cell datasets), extracted from cells (6 of 8), in virion (1 of 3), extracted from viral particles (1 of 2) and with other modifiers, including DMS (2 of 5) and icSHAPE (8 of 11). EternaFold was the top-scoring package ($P < 0.05$) in five of the six datasets of new SARS-CoV-2 data.

We were curious as to whether the differences in packages arose 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. We calculated correlations along the length of example constructs—the Zika ILM genome probed in virion[45] (Fig. 4c), HEK293 mRNA for gene *RPS27A*, extracted from chromatin and probed ex vivo[53] (Fig. 4d)—and observed that EternaFold correlations generally demonstrated a fixed improvement across compared packages across all regions, supporting a consistent accuracy improvement by this package.

We also tested the ability of EternaFold to predict the thermodynamics of binding of human Pumilio proteins 1 and 2 in a dataset of 1,405 constructs[54]. EternaFold showed no significant increase or decrease in predictive ability ($P > 0.05$) when compared to CONTRAfold or ViennaRNA 2 at 37 °C (Extended Data Fig. 9a and Supplementary Table 11).

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

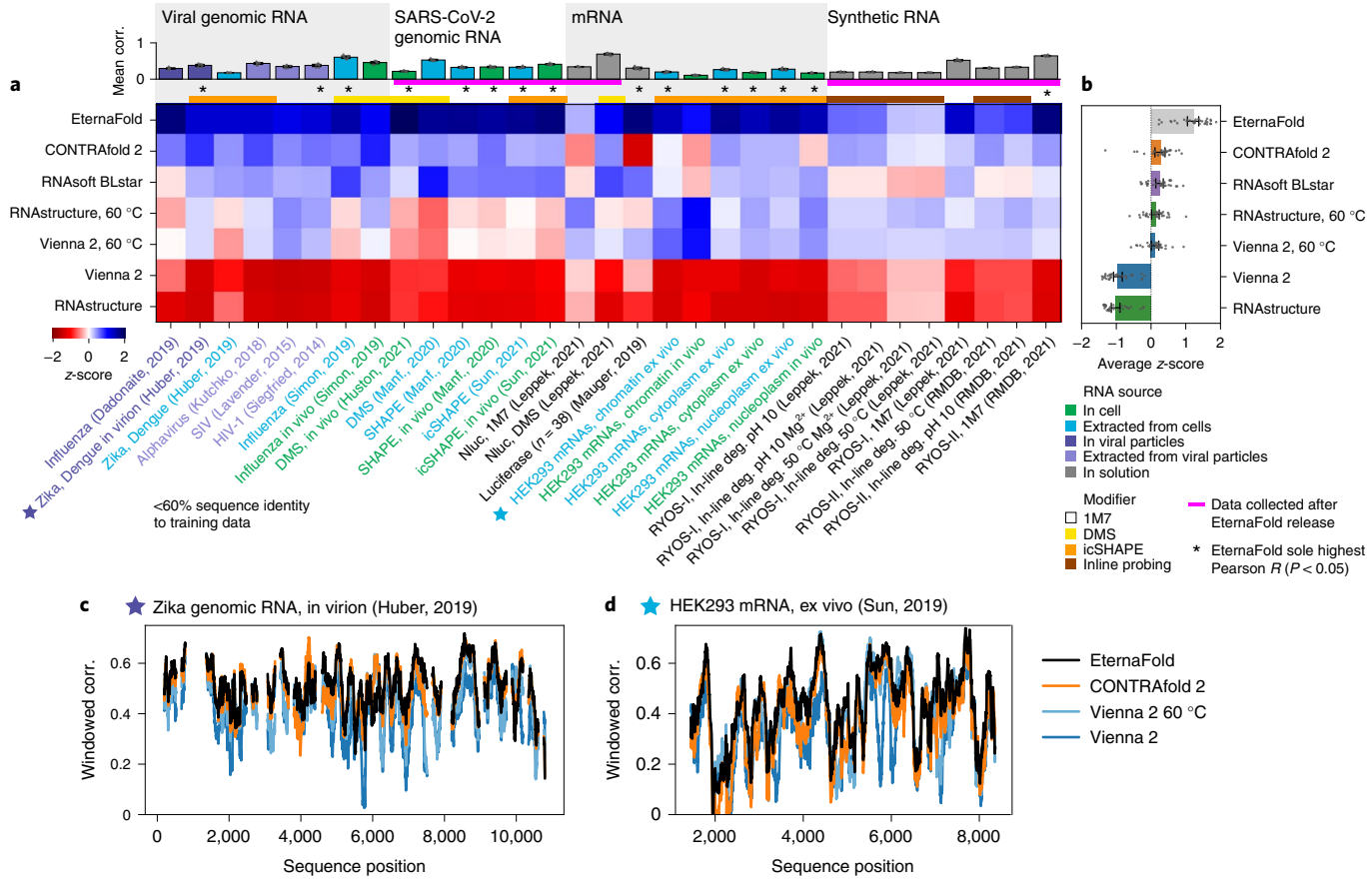


Fig. 4 | EternaFold improved prediction extends across diverse natural RNA contexts and experiments. a, Mean package correlation for top-performing representative packages selected to benchmark for 31 independent chemical mapping datasets from a variety of biological contexts and with other chemical modifiers. Each dataset was filtered to contain sequences with <60% sequence similarity to the EternaFold training set. Corr., correlation. **b**, EternaFold is ranked highest in average *z*-score. Error bars represent 95% confidence interval of the mean obtained over 1,000 iterations of bootstrapping over $n = 31$ datasets from literature. **c**,**d**, Calculating correlations over sequence windows indicates that EternaFold demonstrates uniformly higher correlation across sequence position for two representative datasets: Zika genome probed in virion (**c**) (ref. [45]) and Human mRNA for gene *RPS27A* probed ex vivo (**d**) (ref. [53]).

Table 2 | Average *z*-score for each external RNA class

	Viral genomic RNA	SARS-CoV-2 genomic RNA	mRNA	Synthetic RNA	Average across all datasets
Number of datasets	8	6	9	8	31
EternaFold	**1.29 (0.21)**	**1.65 (0.12)**	**1.26 (0.43)**	**0.75 (0.50)**	**1.21 (0.47)**
CONTRAfold	0.61 (0.17)	0.38 (0.09)	−0.10 (0.56)	0.27 (0.13)	0.27 (0.41)
RNAsoft BLstar	0.34 (0.23)	0.54 (0.24)	0.23 (0.31)	−0.07 (0.19)	0.24 (0.32)
RNAstructure, 60 °C	0.10 (0.27)	−0.25 (0.20)	0.32 (0.36)	0.21 (0.07)	0.12 (0.32)
ViennaRNA 2, 60 °C	0.04 (0.26)	−0.25 (0.22)	0.37 (0.25)	0.18 (0.03)	0.11 (0.30)
ViennaRNA 2	−1.19 (0.30)	−1.00 (0.16)	−0.97 (0.43)	−0.65 (0.32)	−0.95 (0.37)
RNAstructure	−1.20 (0.26)	−1.06 (0.16)	−1.11 (0.34)	−0.70 (0.38)	−1.02 (0.35)

Standard deviation of *z*-scores in parentheses. The top-performing package for each is in bold.

riboswitch systems. Unlike in single secondary structure prediction tasks, we demonstrate that both widely used and state-of-the-art machine-learning algorithms demonstrate a wide range in performance on these tasks. We averaged both rankings to acquire a final ranking of the tested external packages in Table 1.

We discovered that CONTRAfold 2, which inferred thermodynamic parameters by feature representation in datasets of natural RNA secondary structures, performed best in this ranking and performed significantly better than 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 (chemical mapping and riboswitch binding affinities), with no relationship between these two sets of sequences and no relationship between the synthetic sequences and natural sequences. We further investigated whether combining these tasks in a multitask-learning frame-

work could improve performance. We found that models trained on four types of data—single structures, chemical mapping data and riboswitch affinities for an output protein with and without an input ligand—showed improved performance in predictions for held-out subsets of EternaBench datasets as well as improvements in datasets involving virus RNA genomes and mRNAs collected by 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 toward 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. Indeed, methods developed with a variety of more recent methodological advances, including machine learning from chemical mapping datasets (CROSS), deep learning methods for secondary structure prediction (SPOT-RNA), extended parameter sets (CONTRAfold-noncomplementary, CycleFold, MXfold) or accelerated folding packages (LearnToFold), demonstrated diminished performance in the EternaBench tasks (Extended Data Fig. 1a). It was surprising that well-developed and more widely used packages such as ViennaRNA and RNAstructure gave worse performances than CONTRAfold and RNAsoft across all tasks, but that predictions from ViennaRNA and RNAstructure at 60 °C showed notable improvement over the default of 37 °C. This observation might be rationalized by discrepancies in ionic conditions used to measure these packages' thermodynamic parameters, and the in vitro and in vivo conditions tested here.

We used the EternaBench datasets to train a thermodynamic model via multitask learning on secondary structure prediction, chemical mapping signal likelihood maximization and minimizing error for riboswitch protein-binding prediction. The resulting model, termed EternaFold, performed best across 31 external datasets in four categories of natural and synthetic RNAs (Table 2) in a variety of cellular contexts, including RNAs probed in and extracted from cells and viral particles. It was not obvious that a model trained on datasets collected in vitro would demonstrate improvement on the variety of contexts for which we collected datasets. Although many factors influence RNA structure in cells beyond thermodynamic base pairing[55], this demonstrates that existing natural RNA datasets are indeed capable of discriminating between ensemble-averaged base-pairing predictions and that accurate prediction of chemical mapping signal presents an ensemble-aware target for RNA secondary structure algorithm improvement.

The improvements from multitask training in EternaFold indicated that the nearest-neighbor model encoded in CONTRAfold had sufficient representational capacity to gain improvement on the chemical mapping and riboswitch prediction tasks. A notable area of algorithm development and potential improvement is the systematic evaluation of structure prediction methods that incorporate structure mapping data[8,55,56]. We implemented data-driven folding in EternaFold and tested on a collection of 13 structured RNAs as well as three other independent datasets. We found that EternaFold-SHAPE resulted in the highest mean Mathews correlation coefficient (MCC) over all these datasets (0.842), but this improvement was not statistically significant over several other algorithms in use for SHAPE-directed folding, such as SHAPEknots[57] and the heuristic developed by Zarringhalam et al.[58]. implemented in ViennaRNA (mean MCCs of 0.820 and 0.830, respectively, Extended Data Fig. 9c and Supplementary Table 12), indicating potential for improvement. Another limitation of the resulting EternaFold algorithm is that it does not contain distinct terms for entropy, enthalpy and ionic concentrations. Future work creating temperature and salt-dependent models may benefit from analogous ensemble-aware fitting procedures collected at varying temperatures and ionic concentrations. Further improvements in modeling may arise from applying more sophisticated graph-[59] and language-based[60] architectures to predicting RNA thermodynamics. Further investigations will also be necessary to improve performance and aspects of the model that need to be expanded, which may include noncanonical pairs[12], more sophisticated treatment of junctions[61], next-nearest-neighbor effects[14] and chemically modified nucleotides[62]. Orthogonal 3D structure methods such as nuclear magnetic resonance spectroscopy[63] and cryogenic-electron microscopy[64] will likely be instrumental to these pursuits. Taken together, the datasets presented here serve as an important starting point for evaluating and improving future RNA structure prediction algorithms.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01605-0.

Received: 29 May 2020; Accepted: 10 August 2022;
Published online: 3 October 2022

References

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. Pharm.* **82**, 187–203 (2018).
4. Kramps, T. & Elbers, K. Introduction to RNA Vaccines. In: Kramps, T., Elbrs, K. (eds) RNA Vaccines. *Methods Mol. Biol.* Vol. 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 Bioinf.* **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. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* **23**, i19–i28 (2007).
11. Do, C. B., Woods, D. A. & Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90–e98 (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. Preprint at *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. *Nat. Commun.* **10**, 5407 (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. USA* **116**, 24075–24083 (2019).
20. 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).

21. 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).
22. Tian, S. & Das, R. RNA structure through multidimensional chemical mapping. *Q. Rev. Biophys.* **49**, e7 (2016).
23. Denny, S. K. et al. High-throughput investigation of diverse junction elements in RNA tertiary folding. *Cell* **174**, 377–390 e320 (2018).
24. Buenrostro, J. D. et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
25. Lee, J. et al. RNA design rules from a massive open laboratory. *Proc. Natl Acad. Sci. USA* **111**, 2122–2127 (2014).
26. Delli Ponti, R., Marti, S., Armaos, A. & Tartaglia, G. G. A high-throughput approach to profile RNA structure. *Nucleic Acids Res.* **45**, e35 (2017).
27. Eddy, S. R. Analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu. Rev. Biophys.* **43**, 433–456 (2014).
28. Cordero, P., Lucks, J. B. & Das, R. An RNA mapping database for curating RNA structure mapping experiments. *Bioinformatics* **28**, 3006–3008 (2012).
29. Wellington-Oguri, R. et al. Evidence of an unusual Poly(A) RNA signature detected by high-throughput chemical mapping. *Biochemistry* **59**, 2041–2046 (2020).
30. Anderson-Lee, J. et al. Principles for predicting RNA secondary structure design difficulty. *J. Mol. Biol.* **428**, 748–757 (2016).
31. Beisel, C. L. & Smolke, C. D. Design principles for riboswitch function. *PLoS Comput. Biol.* **5**, e1000363 (2009).
32. Breaker, R. R. Prospects for riboswitch discovery and analysis. *Mol. Cell* **43**, 867–879 (2011).
33. Andreasson, J. O. L. et al. Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular switches. *Proc. Natl Acad. Sci. USA* **119**, e2112979119 (2022).
34. 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).
35. Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H. & Murphy, K. P. Computational approaches for RNA energy parameter estimation. *RNA* **16**, 2304–2318 (2010).
36. Foo, C.-S. & Pop, C. Learning RNA secondary structure (only) from structure probing data. Preprint at *bioRxiv*, 152629 (2017).
37. Andronescu, M., Bereg, V., Hoos, H. H. & Condon, A. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinf.* **9**, 340 (2008).
38. Sloma, M. F. & Mathews, D. H. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA* **22**, 1808–1818 (2016).
39. Watters, K. E. et al. Probing of RNA structures in a positive sense RNA virus reveals selection pressures for structural elements. *Nucleic Acids Res.* **46**, 2573–2584 (2018).
40. Watts, J. M. et al. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711–716 (2009).
41. Kutchko, K. M. et al. Structural divergence creates new functional features in alphavirus genomes. *Nucleic Acids Res.* **46**, 3657–3670 (2018).
42. 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).
43. Dadonaite, B. et al. The structure of the influenza A virus genome. *Nat. Microbiol* **4**, 1781–1789 (2019).
44. 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).
45. Huber, R. G. et al. Structure mapping of dengue and Zika viruses reveals functional long-range interactions. *Nat. Commun.* **10**, 1408 (2019).
46. Huston, N. C. et al. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol. Cell* **81**, 584–598 e585 (2021).
47. Manfredonia, I. et al. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res.* **48**, 12436–12452 (2020).
48. Sun, L. et al. In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. *Cell* **184**, 1865–1883 e1820 (2021).
49. Lavender, C. A., Gorelick, R. J. & Weeks, K. M. Structure-based alignment and consensus secondary structures for three HIV-related RNA genomes. *PLoS Comput. Biol.* **11**, e1004230 (2015).
50. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl Acad. Sci. USA* **106**, 97–102 (2009).
51. McGinnis, J. L. & Weeks, K. M. Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry* **53**, 3237–3247 (2014).
52. Leppek, K. et al. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat. Commun.* **13**, 1536 (2022).
53. Sun, L. et al. RNA structure maps across mammalian cellular compartments. *Nat. Struct. Mol. Biol.* **26**, 322–330 (2019).
54. Becker, W. R. et al. Quantitative high-throughput tests of ubiquitous RNA secondary structure prediction algorithms via RNA/protein binding. Preprint at *bioRxiv*, 571588 (2019).
55. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).
56. Morandi, E. et al. Genome-scale deconvolution of RNA structure ensembles. *Nat. Methods* **18**, 249–252 (2021).
57. Hajdin, C. E. et al. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl Acad. Sci. USA* **110**, 5498–5503 (2013).
58. Zarringhalam, K., Meyer, M. M., Dotu, I., Chuang, J. H. & Clote, P. Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS ONE* **7**, e45160 (2012).
59. Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **12**, 941 (2021).
60. Chen, X., Li, Y., Umarov, R., Gao, X. &, Song, L. RNA secondary structure prediction by learning unrolled algorithms. In *Proceedings of the 8th International Conference on Learning Representations* (2020).
61. Ward, M., Datta, A., Wise, M. & Mathews, D. H. Advanced multi-loop algorithms for RNA secondary structure prediction reveal that the simplest model is best. *Nucleic Acids Res.* **45**, 8541–8550 (2017).
62. Zhao, B. S., Roundtree, I. A. & He, C. Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol.* **18**, 31–42 (2017).
63. Rinnenthal, J. et al. Mapping the landscape of RNA dynamics with NMR spectroscopy. *Acc. Chem. Res.* **44**, 1292–1301 (2011).
64. Kappel, K. et al. Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures. *Nat. Methods* **17**, 699–707 (2020).

Methods

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

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

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 $\{S\}$ (ref. [65]). From this expression, the probability of any particular structure s is defined as

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

Chemical mapping prediction theoretical basis. 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. For simplifying notation, we continue with implicit x and Θ, that is, $P(i:j|x,\Theta) = P(i:j)$. This is computed as

$$P(i:j) = \sum_{s_{i:j}\in\{S\}} P(s_{i:j}), \quad (3)$$

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 are analytically calculated by all the algorithms tested here. The probability of any single base being unpaired can be computed as

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

The relationship between the probability of a nucleotide being unpaired and its experimentally measured reactivity has served as a locus for efforts to improve 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[27,66,67]. In this work, we use the linear Pearson correlation coefficient between unpaired probability and experimentally measured reactivity as a measure of model quality. In the following, we describe the simple model under which this linear assumption holds. We write the probability that nucleotide i (nt_i) is modified at time t as

$$P(\text{nt}_i\,\text{modified, time}\,t) = 1 - e^{-k_{\text{mod}}(i)t}, \quad (5)$$

where $k_{\text{mod}}(i)$ is the rate of modification for nucleotide i. The measured chemical modification signal is an ensemble population average, where the time exposure of the ensemble to the modifier has been limited to aim to achieve 'single-hit kinetics' with single-hit frequency, so that the degree of modification in experiment is proportional to the rate of modification[68]. In other words, because $k_{\text{mod}}(i)t \ll 1$, we can approximate

$$P(\text{nt}_i\,\text{modified, time}\,t) \approx k_{\text{mod}}(i)\,t \propto k_{\text{mod}}(i). \quad (6)$$

This expression assumes that each RNA molecule is not heavily modified, such that $k_{\text{mod}}(i)$ for each nucleotide is independent of the modification state of other nucleotides. If we assume that the timescale of chemical modification is much slower than the timescale of fluctuation between structural ensemble states, then we may write the overall modification rate for each nucleotide i as averaged over the equilibrated structure ensemble of the RNA,

$$k_{\text{mod}}(i) = \sum_{s\in\{S\}} P(s)\,k_{\text{mod}}(i|s) \quad (7)$$

If we consider a simplest two-state model for each nucleotide, with modification rate k_{pr} if paired and a rate k_{unp} if unpaired, then this reduces to

$$\begin{aligned} k_{\text{mod}}(i) &= k_{\text{unp}}P(i\,\text{unpaired}) + k_{\text{pr}}P(i\,\text{paired}) \\ &= k_{\text{pr}} + (k_{\text{unp}} - k_{\text{pr}})P(i\,\text{unpaired}), \end{aligned} \quad (8)$$

which demonstrates that under this simple model, the modification rate is linear with respect to P(unpaired). The model above is limited in its assumption of two states and does not account for reactivity effects caused by sequence and local environment. For instance, Hoogsteen conformations in G–A and G–G mismatches expose the Watson–Crick faces of purine nucleobases, resulting in higher DMS reactivity[69]. A Spearman rank correlation (Extended Data Fig. 1c), which will be more dominated by relative rankings, results in a similar overall ranking.

Chemical mapping data. Chemical mapping data for the Eterna Cloud Lab experiments were downloaded from the RNA Mapping DataBase (RMDB)[28] and processed with RDATKit (https://ribokit.github.io/RDATKit/). The RNA was probed with the MAP-seq protocol with a coloaded standard molecule (P4-P6-2HP RNA) to enable normalization, as described in ref. [70] measurements were carried out at ambient temperatures (24 °C) with 10 mM $MgCl_2$ and 50 mM Na-HEPES, pH 8.0. Data were processed using MAPseeker[71] with standard settings.

Within each chemical mapping dataset, CD-HIT-EST[72] 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. These datasets ranged in size from 605 (round 15) to 3,378 constructs (round 23), with a median size of 1,577; after filtering, they ranged from 101 (round 12) to 1,088 (round 1), with a median size of 562 (Extended Data Fig. 3 and Supplementary Table 2). The filtered 24 datasets comprised 12,711 individual constructs, and distributions of GC content, average sequence length and number of loops in the target structures were not significantly affected (Extended Data Fig. 3).

Nucleotides with reactivities less than or equal to zero or greater than the 95th percentile of the dataset were removed from analysis. Cloud Lab round 2 was filtered to exclude experiments that had FMN present, which pertained to Eterna Cloud Lab challenges to design riboswitches. Adenosine nucleotides preceded by six or more As were also removed due to evidence of anomalous reverse transcription effects in such stretches[73]. External chemical mapping datasets were obtained from the supplementary information from the papers and processed similarly (outliers, nucleotides in poly-A stretches removed).

Analyzing package performance by the Cloud Lab project. We wished to understand whether factors such as target structure complexity, GC content and sequence length influenced package predictions. We performed the same package ranking analysis, grouping constructs by their projects instead of by the 24 datasets. Because grouping constructs into projects sometimes resulted in a small number of nucleotides over which to calculate correlations, we omitted package predictions where the standard error of the calculated Pearson correlation was greater than 0.05. This resulted in a total of 612 project groupings remaining, names and calculated metrics for which are contained in Supplementary Table 4.

We found weak correlation between the per-project z-score of the top-performing package, CONTRAfold 2 and GC content (Spearman $R = 0.15$), sequence length (0.07) and total loops in the target structure ($R = 0.16$). There were also weak correlations between the average Pearson correlation for all packages and GC content (Spearman $R = 0.10$), sequence length ($R = -0.24$) and total target structure loops ($R = -0.01$) (Extended Data Fig. 1d).

Riboswitch activity prediction theoretical basis. A thermodynamic framework discussed in greater detail in ref. [17] 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, $K_{\text{obs}}^{-\text{lig}}$, via the relative ratios of both values to those of a reference state:

$$\frac{K_{\text{obs}}^{-\text{lig}}}{K_{\text{obs}}^{\text{ref}}} = \frac{P^{\text{ref}}(\text{out})}{P(\text{out})} \equiv K_{\text{MS2}}^{-\text{lig}}. \quad (9)$$

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, $K_{\text{obs}}^{\text{ref}}$ is estimated as the strongest affinity measured (Extended Data Fig. 10a). We refer to the estimated ratio $\frac{P^{\text{ref}}(\text{out})}{P(\text{out})}$ as $K_{\text{MS2}}^{-\text{lig}}$ in the main text, as the equilibrium constant of forming the MS2 hairpin as normalized to the reference state.

Although there may be error introduced in which experimental point is selected to be $K_{\text{obs}}^{\text{ref}}$, relative error should be constant when comparing packages on the same dataset. To compare packages, we therefore report the correlation between $\log(K_{\text{obs}}^{\pm\text{lig}}/K_{\text{obs}}^{\text{ref}})$ and $\log(K_{\text{MS2}}^{\pm\text{lig}})$, which excludes the effect of selection for $K_{\text{obs}}^{\text{ref}}$.

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

$$P(\text{motif}|x,\theta) = \sum_{s_{\text{motif}}\in\{S\}} P(s_{\text{motif}}), \quad (10)$$

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, an

experimental observable that can be estimated by all the packages tested here. Thus, our prediction of interest is

$$\frac{K^{-\text{lig}}_{\text{pred}}}{K^{\text{ref}}_{\text{pred}}} = \frac{P^{\text{ref}}(i:j)}{P(i:j)}, \tag{11}$$

where i and j are the nucleotides forming the terminal base pair in the MS2 aptamer stem. The value $P^{\text{ref}}(i;j)$ is accordingly computed as the probability of closing the base pair in the reference sequence. We confirmed that calculations using equations (9) and (11) agree for Vienna, RNAstructure and CONTRAfold packages.

Predicting protein-binding affinities with input ligand bound. The estimation of $K^{+\text{lig}}_{\text{fold}}$ follows similarly to the above but accounts for 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 (equation (10)).

Analogously to equation (9), we define $K^{+\text{lig}}_{\text{MS2}}$ as

$$K^{+\text{lig}}_{\text{MS2}} = \frac{K^{+\text{lig}}_{\text{obs}}}{K^{\text{ref}}_{\text{obs}}} \tag{12}$$

which is calculated as

$$K^{+\text{lig}}_{\text{MS2}} = \frac{Z + bZ_{\text{lig}}}{Z_{\text{MS2}} + bZ_{\text{lig,MS2}}} \tag{13}$$

where Z_{lig} is the constrained-partition function of the state including the ligand aptamer (calculated in each algorithm as described in the next section), Z_{MS2} is the partition function for the state including the MS2 aptamer and $Z_{\text{lig,MS2}}$ is the partition function of the state including both ligand aptamer and MS2 aptamer. The constant $b = \frac{[\text{ligand}]}{K_{d,\text{ligand}}}$ is the Boltzmann weight of binding the ligand when the bulk concentration of the ligand is [ligand]. Values used for calculating b are in Supplementary Table 14. Representative predictions of $K^{+\text{lig}}_{\text{MS2}}$ versus experimental $K^{+\text{lig}}_{\text{MS2}}$ values are in Extended Data Fig. 10b.

Riboswitch data. Riboswitch data were downloaded from supplementary materials from refs. [33,34]. In brief, 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^{-1} BSA, 1 mM DTT, 0.01 mg ml^{-1} yeast tRNA, 0.01% Tween-20 and varying concentrations of small molecule ligand (FMN, theophylline, tryptophan) and MS2 coat protein. Datasets were filtered to include only constructs with more than 50 copies of the sequence represented in the RNA-MaP experiment, constructs that included the canonical MS2 and small molecule aptamers, and filtered using CD-HIT-EST[72] to remove sequence redundancy over 80%. As per the CD-HIT-EST algorithm default, the longest sequence per cluster was maintained. If all sequences were the same length, the first sequence was used. After filtering, the riboswitch datasets comprised 7,228 constructs in total. Scripts to replicate data processing from refs. [33,34]. are included in the EternaBench software repository.

For all constructs as well as the reference MS2 hairpin construct, we performed $K^{\text{lig}}_{\text{MS2}}$ estimations including a flanking hairpin included in the Illumina array experiments, described in ref. [33]. As an example, the full reference MS2 hairpin construct, as well as the constraint used for estimating $K^{\text{ref}}_{\text{pred}}$ 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 bold.

Sequence: GGGUAUGUCGCAGAA**A**CAUGAGGAUCACCCAUG**U**AACUG CGACAUACCC

Structure:.....((((((x((xxxx)))))))).....

The riboswitches in EternaBench-Switch are controlled by the small molecules FMN, tryptophan or theophylline. Motifs, concentrations and intrinsic K_d values used for $K^{+\text{lig}}_{\text{MS2}}$ prediction, taken from refs. [33,34], are provided in Supplementary Table 14.

EternaFold multitask learning. The CONTRAfold[11] loss function optimizes the conditional log-likelihood of ground-truth structure $s^{(i)}$ given sequence $x^{(i)}$ over dataset D:

$$L_{\text{CONTRAfold}} = L_{\text{Struct}}(\theta) = \sum_{i \in D} \log P(s^{(i)} | x^{(i)}, \{\theta\}). \tag{14}$$

In CONTRAfold-SE[36], 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

$$\begin{aligned} L_{\text{CONTRAfold-SE}} &= L_{\text{Struct}} + w_{\text{CM}} L_{\text{CM}}, \\ L_{\text{CM}}(\theta, \phi) &= \sum_{i \in D} \log \sum_{s} P(s, \mathbf{d} | x, \{\theta\}, \phi), \end{aligned} \tag{15}$$

where $\mathbf{d}$ 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 $\log K^{-\text{lig}}_{\text{fold}}$ and $\log K^{+\text{lig}}_{\text{fold}}$:

$$\begin{aligned} L_{\text{MS2}} = {} & w_{-\text{lig}} \left[\log K^{\text{exp}}_{\text{MS2}}(-\text{lig}) - \log K^{\text{pred}}_{\text{MS2}}(-\text{lig}) \right]^2 \\ & + w_{+\text{lig}} \left[\log K^{\text{exp}}_{\text{MS2}}(+\text{lig}) - \log K^{\text{pred}}_{\text{MS2}}(+\text{lig}) \right]^2. \end{aligned} \tag{16}$$

The full loss function for EternaFold is thus written as

$$L_{\text{EternaFold}} = L_{\text{Struct}} + L_{\text{CM}} + L_{\text{MS2}}. \tag{17}$$

The hyperparameters w_{CM}, $w_{-\text{lig}}$, $w_{+\text{lig}}$, corresponding to the relative weights placed on different data types, were selected through a grid search on the holdout sets STRAND-holdout, EternaBench-CM-holdout and EternaBench-Switch-holdout (data not shown). The final values used for training were $w_{\text{CM}} = 0.5$, $w_{-\text{lig}} = 30$, $w_{+\text{lig}} = 30$.

Dataset selection for training and testing EternaFold. *Single-structure data.* For training EternaFold, we used the S-Processed dataset[37] train and holdout sets used previously in training CONTRAfold 2 and RNAsoft[10], to keep the same datasets consistent with these algorithms. However, we found that the S-Processed test set had 68 and 52% redundancy to the S-Processed train and holdout sets, respectively, using CD-HIT-EST-2D. We therefore created a new secondary structure test set by filtering the more recent ArchiveII dataset[38] for constructs with <80% sequence similarity to any sequence across all three data types used in EternaFold training. We also evaluated EternaFold performance on structure prediction for the S-Processed test set, and found qualitatively similar results to the ArchiveII-NR test set (Extended Data Fig. 9b, compare to Fig. 3b).

Cloud Lab chemical mapping data. We used rounds 3, 4, 5, 7, 10 and 11 as training and holdout data. This was to be consistent with the training data used in CONTRAfold-SE[36], and to reserve rounds 0 and 1 as test rounds, given their large size and high signal-noise ratio. GC content, sequence length, total loops in the target structure and signal/noise ratio were equivalent across train, holdout and test rounds (Extended Data Fig. 3c).

Riboswitch data. We partitioned the RiboLogic dataset into our training, holdout and test sets due to the high signal-noise ratio and diversity of structures, subdividing the riboswitches so that each split contained identical fractions of FMN-, theophylline- and tryptophan-responsive riboswitches. This left the rest of the Eterna riboswitch rounds as test sets (Extended Data Fig. 3d).

Test dataset filtering. To filter test datasets based on sequence similarity to the EternaFold training data, we implemented a 'windowed Levenshtein distance'. We calculated Levenshtein distance across sliding windows of the longer sequence that are the length of the shorter sequence. A sequence was counted as redundant at X% cutoff if any window had a Levenshtein edit distance smaller than (100-X)% the window size. Supplementary Table 10 contains test dataset sizes before and after filtering at a windowed Levenshtein distance cutoff of 80, 60 and 40%. As a point of comparison, uniformly distributed, randomly generated 50-mers, 100-mers and 200-mers were calculated to have average Levenshtein distances of 42, 44 and 45%, respectively.

Evaluating base pair probabilities for external datasets. For comparing P(unpaired) calculations to natural RNAs, many of which are thousands of nucleotides long, we compared several practices for calculating, which includes predicting base pair probabilities from overlapping windows, constraining the nucleotides under consideration using a beam search algorithm implemented in LinearPartition[74], and conventional folding of the entire RNA. Windows of length 300, 600, 900 and 1,200 with 25-nt overlap. Results from length 900 are shown in the main text, although results are similar for other window sizes (Extended Data Fig. 8a).

SHAPE-directed folding evaluation. We implemented SHAPE-directed folding in EternaFold in the following way: for an RNA sequence x with length L, let d_j be the probing signal at nucleotide j in the sequence. The joint probability for structure s and the vector of reactivities $\mathbf{d}$ is given as

$$P(s, \mathbf{d} | x; \theta, \phi) = P(s | x; \theta) \prod_{j=1}^{L} P\left(d_j | x_j, s; \phi\right)^{\kappa} \tag{18}$$

where θ represents the learned set of thermodynamic parameters and ϕ represents the parameters learned for eight gamma distributions defining the reactivities of A,C,G,U being paired or unpaired (Extended Data Fig. 9d), and κ is a parameter

specifying the relative weight of the evidence. Predicting a maximum likelihood structure given an observed reactivity vector **d**, is calculated as

$$s_{\mathrm{MLE}} = \underset{\hat{s}\in\{S\}}{\operatorname{argmax}} P(s, \mathbf{d}|x;\theta,\phi). \quad (19)$$

The maximum expected accuracy structure is calculated using the same SHAPE-weighted partition function and the expression

$$s_{\mathrm{mea}} = \underset{\hat{s}}{\operatorname{argmax}} \mathbb{E}_s\left[\mathrm{Acc}\left(\hat{s}, s^*\right)\right] \quad (20)$$

where Acc $(\hat{s}, s^*)$ is the pseudo-accuracy measure described in detail in ref. [11] and s^* is the (unknown) true structure.

When the EternaFold parameters were initially trained, κ was set to 1. To fit κ in the context of SHAPE-directed folding, we used the SHAPEknots training dataset and calculated the MCC. This dataset consists of 16 RNAs with known 3D structure and was used similarly to tune parameters in SHAPEknots[57] and for the default settings of three formulas present in the ViennaRNA package. We refer to this model as EternaFold-SHAPE.

We compared EternaFold-SHAPE to SHAPEknots[57], RNAstructure with structure probing (but not pseudoknots as in SHAPEknots), three algorithms implemented in ViennaRNA from Washietl[66], Deigan[50] and Zaringhalam[58], as well as RNAstructure, ViennaRNA and EternaFold predictions without reactivity data. We also evaluated the algorithms on the SHAPEknots-TEST dataset, as well as datasets from Chen and Kappel that included DMS probing data for RNAs with secondary structures validated by other methods (Extended Data Fig. 9c, full dataset in Supplementary Table 12). In addition, 13 further RNA constructs were probed by SHAPE and DMS as described in the following section.

We calculated mean MCC across datasets and averaged these values. We found that EternaFold+SHAPE resulted in the highest mean MCC over test constructs of 0.842, but this was not statistically significant (evaluated as $P<0.05$) over SHAPEknots (MCC = 0.818), EternaFold without SHAPE data (MCC = 0.814), ViennaRNA with the heuristic developed by Zarringhalam (MCC = 0.828), RNAstructure with SHAPE data (MCC = 0.803) or Vienna RNAfold 2 (MCC = 0.801). Statistical significance was evaluated using a two-sided *t*-test for related values. Supplementary Table 12 contains predicted SHAPE- or DMS-directed MFE structures for the dataset in all evaluated algorithms.

SHAPE and DMS probing by capillary electrophoresis of 13 structured RNAs for SHAPE-directed folding evaluation. *DNA template preparation.* DNA templates were designed to include the 20-nt T7 RNA polymerase promoter sequence followed by a sequence encoding the desired RNA flanked by two hairpins used to normalize the resulting signal[70]. Double-stranded templates were prepared by the extension of 60-nt DNA oligomers (Integrated DNA Technologies) with Phusion polymerase, using the following thermocycler protocol: denaturation for 30 s at 98 °C, 35 cycles of denaturation for 10 s at 98 °C, annealing for 30 s at 60 to 64 °C, extension for 30 s at 72 °C, final extension for 10 min at 72 °C and cooling to 4 °C. DNA samples were purified with AMPure XP beads (Beckman Coulter), following the manufacturer's instructions. Sample concentrations were estimated based on ultraviolet absorbance at 260 nm measured on Nanodrop spectrophotometer. Verification of template length was accomplished by electrophoresis of all samples and 10- and 20-bp ladder length standards (Thermo Scientific O'RangeRuler SM1313 and SM1323) in 4% agarose gels (containing 0.5 mg ml^{-1} ethidium bromide) and 1× TBE (100 mM Tris, 83 mM boric acid, 1 mM disodium EDTA).

Preparation of RNA templates. In vitro transcription reactions were carried out in 40 µl volumes with 10 pmol of DNA template, using the TranscriptAid T7 High Yield Transcription Kit (Thermo Fisher). Reactions were incubated for 3 h at 37 °C, followed by degradation of DNA template with 2 µl of DNase I at 37 °C for 30 min. RNA samples were purified using the Zymo RNA Clean and Concentrator-25 kit (Zymo Research). Concentrations were measured by absorbance at 260 nm on Nanodrop spectrophotometers.

SHAPE mapping. 1.2 pmol of purified RNA was added to 2 µl of 500 mM Na-HEPES buffer (pH 8.0) and denatured at 90 °C for 3 min. The reaction was then cooled down to room temperature over 10 min. Then 2 µl of 100 mM $MgCl_2$ was added, followed by incubation at 50 °C for 30 min. The sample was cooled down to room temperature over 20 min before addition of 5 µl of nuclease-free water (negative control) or 1-methyl-7-nitroisatoic anhydride (8.48 mg ml^{-1} of dimethylsulfoxide) followed by incubation at room temperature for 15 min and brought to a final volume of 20 µl with nuclease-free water. The SHAPE-RNA sample was further purified by incubating the sample with 5.0 µl of Na-MES, pH 6.0, 3.0 µl of 5 M NaCl, 1.5 µl of Oligo dT bead, 0.25 µl of 10 µM FAM-A20-Tail2 and brought to a final volume of 10 µl with nuclease-free water. The reaction mixture was incubated at room temp for 15 min, pulled down by 96-post magnetic stand for 10 min, washed twice with 70% ethanol and allowed to dry, before adding 2.5 µl of nuclease-free water.

DMS mapping. 5 µl of RNA stock in H_2O containing 12.5 pmol of RNA was mixed with 5 µl of 1× TE (Ambion) and denatured by incubating at 95 °C for 2 min, and then cooling on ice for 1 min. Then 12.5 µl of 2× buffer (600 mM Na-cacodylate, pH 7.0 and 20 mM $MgCl_2$) was added, and the RNA was incubated at 37 °C for 30 min to fold. RNAs were modified by adding 2.5 µl of DMS (1.7 M in 100% ethanol); for no-modification controls, 2.5 µl of 100% ethanol was added instead. Reactions were incubated at 37 °C for 6 min, and then quenched with 25 µl of 2-mercaptoethanol.

Preparing samples for capillary electrophoresis. Compementary DNA (cDNA) was prepared from in-line probing and SHAPE-RNA samples as follows (note that above procedures leave RNA bound to FAM-A20-Tail2 reverse-transcription primers that are in turn bound to Oligo dT beads). Next, 2.5 µl of purified RNA was added to a reaction mixture containing 1× First Strand buffer (Thermo Fisher), 5 mM dithiothreitol (DTT), 0.8 mM dNTPs, 0.2 µl of SS-III RTase (Thermo Fisher) to a final volume of 5.0 µl. The reaction was incubated at 48 °C for 40 min, and stopped with 5 µl of 0.4 M sodium hydroxide. The reaction was then incubated at 90 °C for 3 min, cooled on ice for 3 min and neutralized with 2 µl of quench mix (2 m of 5 M sodium chloride, 3 ml of 3 M sodium acetate, 2 ml of 2 M hydrochloric acid). For four cDNA reference ladders, each of four ddNTPs (GE Healthcare 27-2045-01) with a ddNTP:dNTP ratio of 1.25 (0.1:0.08 mM) was used in the reverse-transcription reaction.

cDNA was pulled down on a 96-post magnetic stand and washed twice with 100 µl of 70% ethanol. To elute the bound cDNA, the magnetic beads were resuspended in 10.0625 µl of ROX350 (Thermo Fisher Scientific 401735)/Hi-Di (0.0625 µl of ROX350 ladder in 10 µl of Hi-Di formamide) and incubated at room temperature for 20 min. The cDNA was further diluted by 1/3 and 1/10 in ROX350/Hi-Di and samples loaded onto capillary electrophoresis sequencers (ABI-3730) on capillary electrophoresis services rendered by ELIM Biopharmaceuticals. Capillary electrophoresis data were analyzed using the HiTRACE v.2.0 package (https://github.com/ribokit/HiTRACE), following the recommended steps for sequence assignment, peak fitting, background subtraction of the no-modification control, correction for signal attenuation and reactivity profile normalization.

Error and significance estimation. We estimated confidence intervals on reported Pearson correlation values by bootstrapping the datapoints under consideration and reporting the 2.5th and 97.5th percentile over 1,000 rounds of bootstrapping. Reported standard error values are estimated by calculating the standard deviation across bootstrapping rounds. We inferred significance in differences between package correlations by analyzing overlap between 95% confidence interval estimates[75,76]. All code to reproduce significance analyses is included in the EternaBench repository.

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 1. Datasets were processed with Pandas (https://github.com/pandas-dev/pandas) and visualized with Seaborn (https://seaborn.pydata.org/).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All datasets used here for evaluation are available at https://www.github.com/eternagame/EternaBench. The original Cloud Lab datasets are available at the RNA Mapping Database[28] under accession IDs ETERNA_R00_0000 (round 00), ETERNA_R69_0000 (round 01), ETERNA_R70_0000 (round 02), ETERNA_R71_0000 (round 03), ETERNA_R72_0000 (round 04), ETERNA_R73_0000 (round 05), ETERNA_R74_0000 (round 06), ETERNA_R75_0000 (round 07), ETERNA_R76_0000 (round 08), ETERNA_R77_0002 (round 09), ETERNA_R78_0001 (round 10), ETERNA_R79_0001 (round 11), ETERNA_R80_0001 (round 12), ETERNA_R81_0001 (round 13), ETERNA_R82_0001 (round 14), ETERNA_R83_0003 (round 15), ETERNA_R84_0000 (round 16), ETERNA_R85_0000 (round 17), ETERNA_R86_0000 (round 18), ETERNA_R87_0001 (round 19), ETERNA_R89_0000 (round 20), ETERNA_R91_0000 (round 21), ETERNA_R92_0000 (round 22) and ETERNA_R94_0000 (round 23). A list of RMDB accession IDs or URLs corresponding to the data used for benchmarking SHAPE-guided folding is in Supplementary Table 12. Source data are provided with this paper.

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 affinity calculations described here, are available at https://www.github.com/eternagame/EternaBench. The code for training EternaFold is available

at https://www.github.com/eternagame/EternaFold. A server to run EternaFold is available at https://eternafold.eternagame.org/. The EternaFold code is derived from the CONTRAfold-SE[36] codebase, which is derived from the CONTRAfold[11] codebase.

References

65. McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**, 1105–1119 (1990).
66. 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).
67. Deng, F., Ledda, M., Vaziri, S. & Aviran, S. Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA* **22**, 1109–1119 (2016).
68. Cordero, P. & Das, R. Rich RNA structure landscapes revealed by mutate-and-map analysis. *PLoS Comput. Biol.* **11**, e1004473 (2015).
69. Xu, Y. et al. Hoogsteen base pairs increase the susceptibility of double-stranded DNA to cytotoxic damage. *J. Biol. Chem.* **295**, 15933–15947 (2020).
70. Kladwang, W. et al. Standardization of RNA chemical mapping experiments. *Biochemistry* **53**, 3063–3065 (2014).
71. Seetin, M. G., Kladwang, W., Bida, J. P. & Das, R. Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *Methods Mol. Biol.* **1086**, 95–117 (2014).
72. 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).
73. Kladwang, W. et al. Anomalous reverse transcription through chemical modifications in polyadenosine stretches. *Biochemistry* **59**, 2154–2170 (2020).
74. Zhang, H., Zhang, L., Mathews, D. H. & Huang, L. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics* **36**, i258–i267 (2020).
75. Zou, G. Y. Toward using confidence intervals to compare correlations. *Psychol. Methods* **12**, 399–413 (2007).
76. Diedenhofen, B. & Musch, J. cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS ONE* **10**, e0121945 (2015).

Acknowledgements

We thank members of the Das and Barna laboratories (Stanford University), C. Pop and C.-S. Foo for useful discussions. We thank I. Jarmoskaite, V.V. Topkar, R. Rangan 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 (grant no. R35 GM122579 to R.D.) and gifts to the Eterna OpenVaccine project from donors listed in Supplementary Table 13.

Author 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. A.I.S. performed data analyses and visualizations. W.K., J.L., A.T. and R.D. designed and implemented the Eterna Cloud Lab initiative. A.B. generated SHAPE and DMS data for RNAs of known structure used in SHAPE-directed folding benchmarking. Eterna participants created online design projects, provided RNA solutions and reviewed the manuscript (Supplementary Table 3).

Competing interests

The authors declare no competing interests.

Additional information

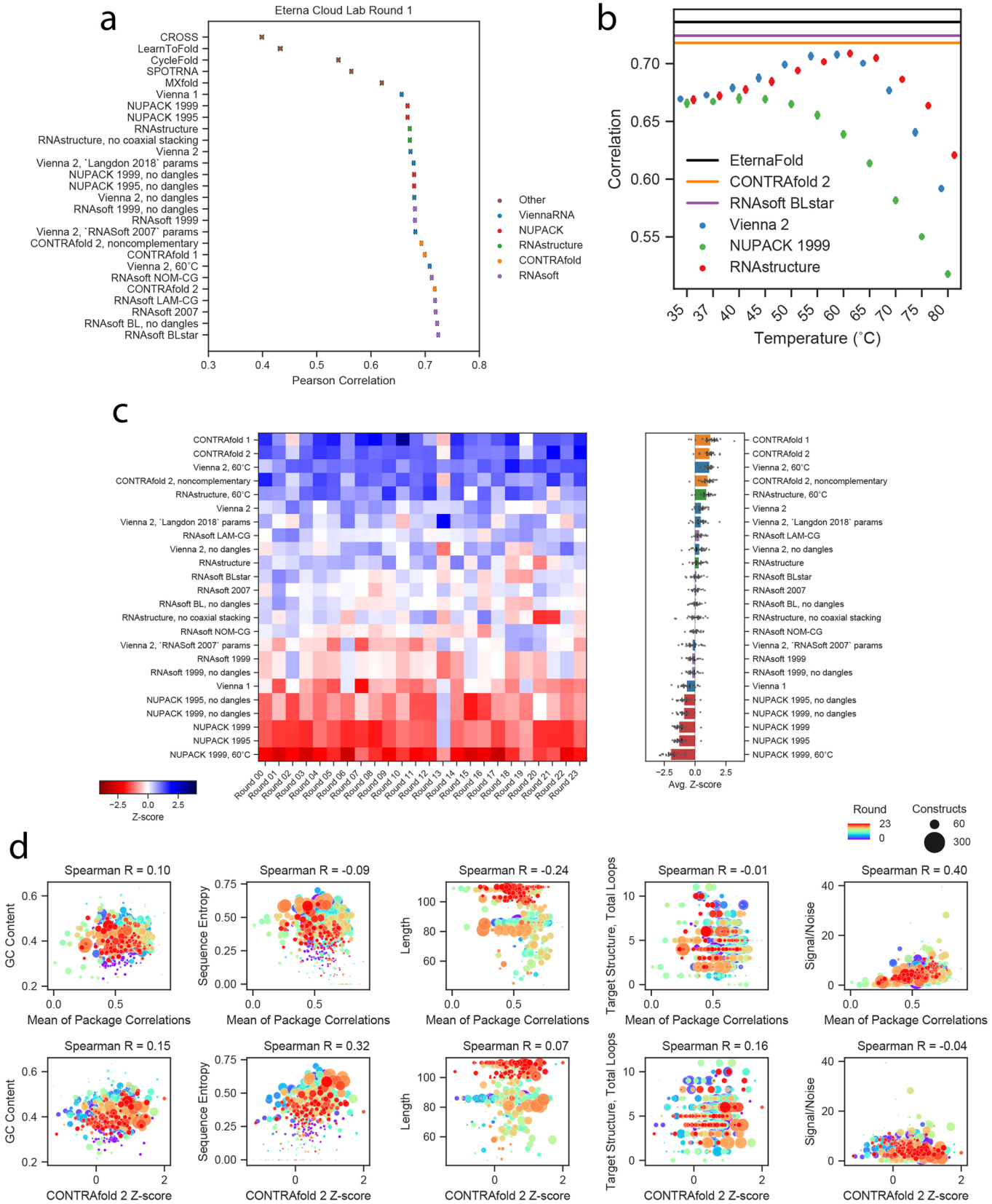
Extended data are available for this paper at https://doi.org/10.1038/s41592-022-01605-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01605-0.

Correspondence and requests for materials should be addressed to Rhiju Das.

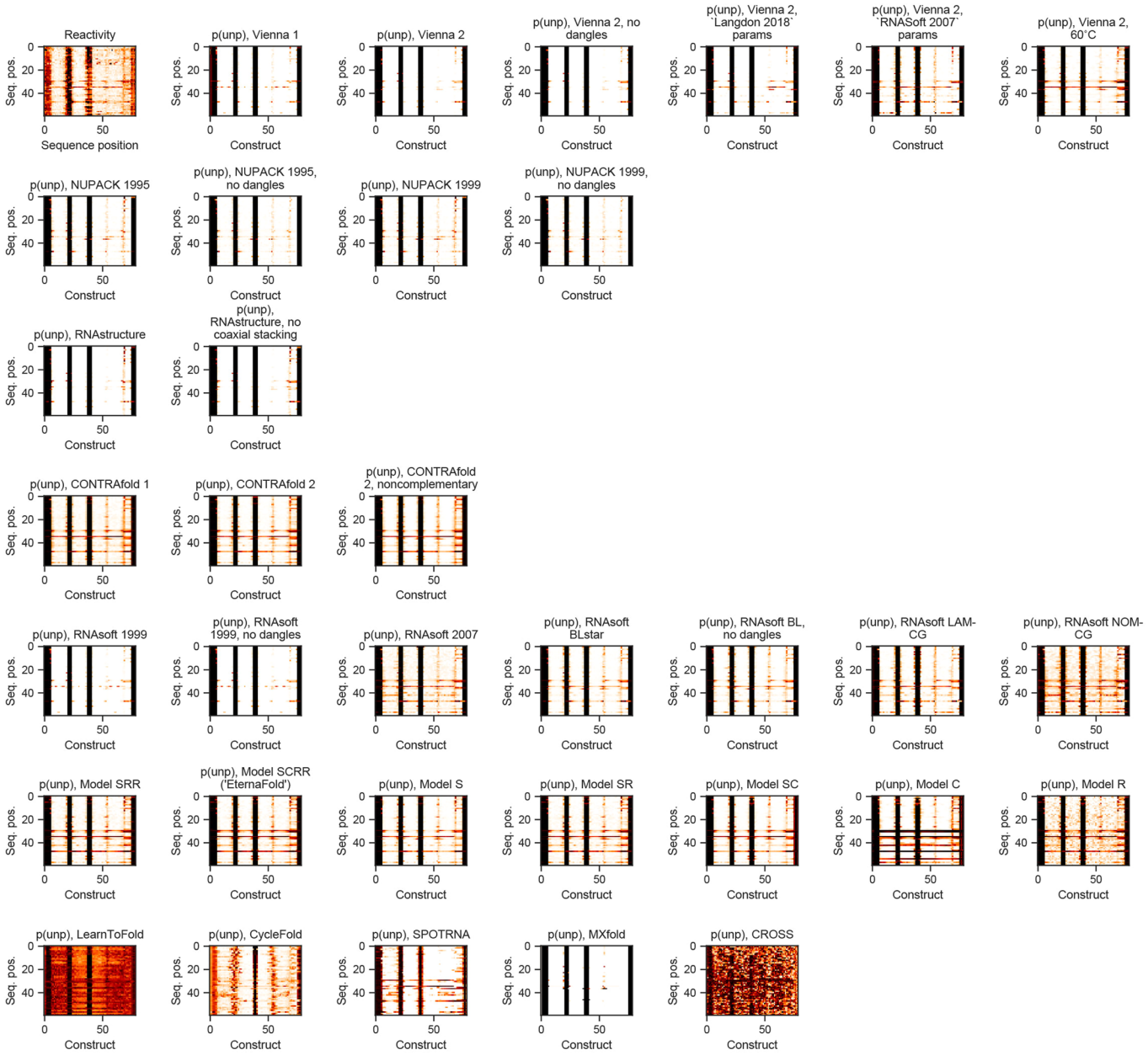
Peer review information *Nature Methods* thanks Hashim Al-Hashimi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

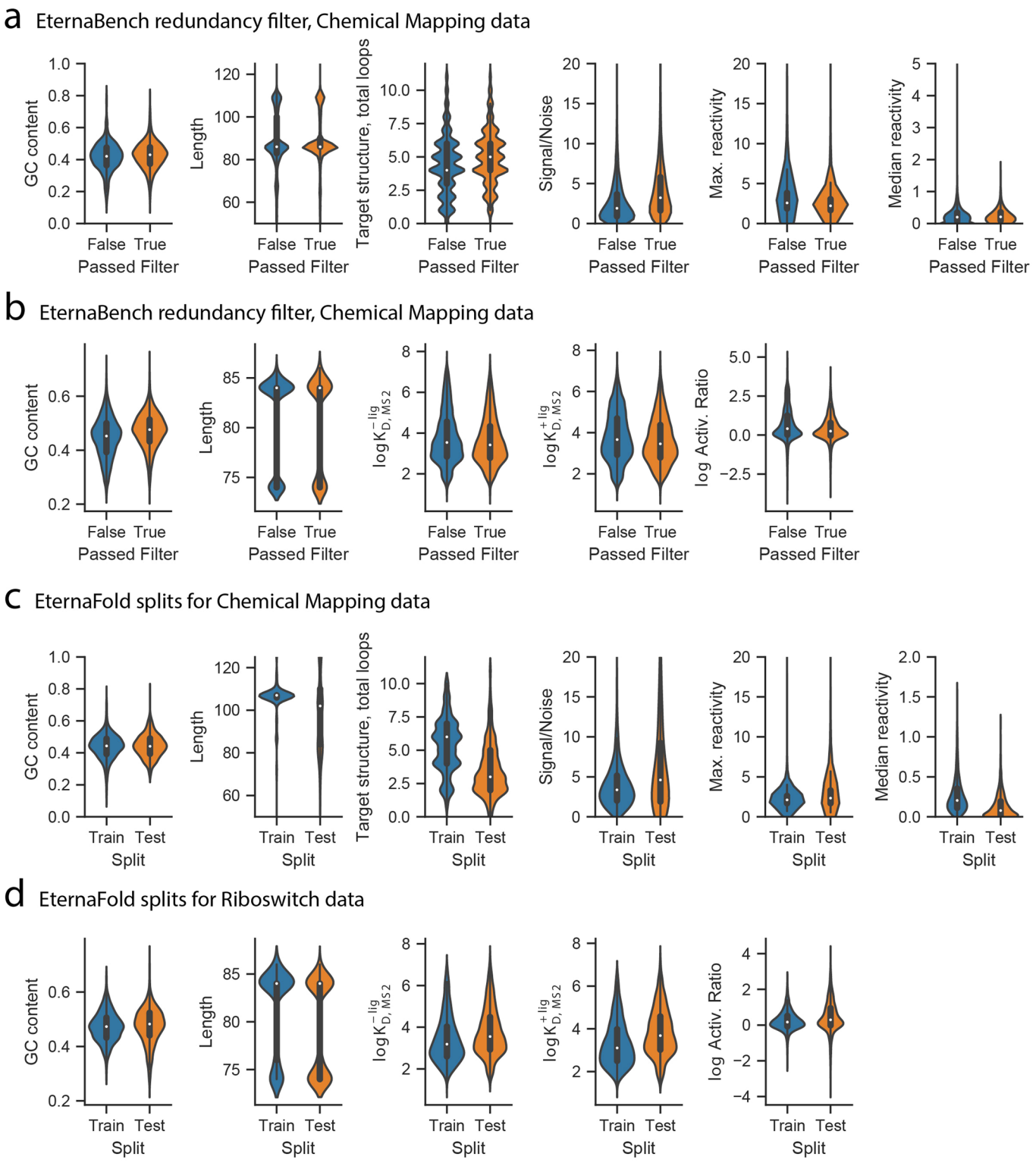


Extended Data Fig. 1 | See next page for caption.

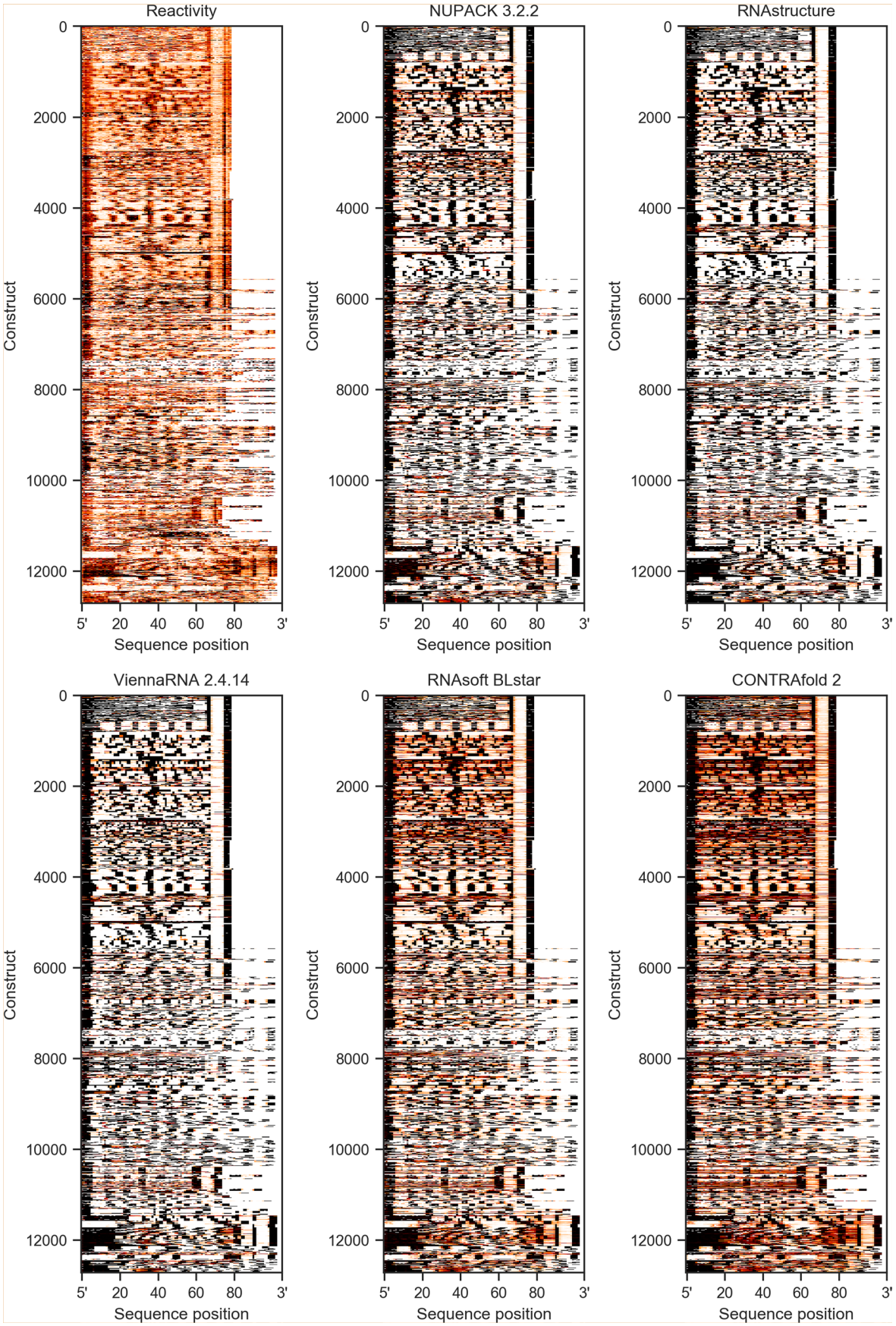
Extended Data Fig. 1 | Extended analysis of package rankings based on Eterna Cloud lab chemical mapping data. a) Pearson correlation of all package options tested on Cloud Lab Round 1, which was also a holdout test set for EternaFold training studies. Mean ± SEM of Pearson correlation calculated via bootstrapping, n = 1088 independent constructs. **b**) ViennaRNA 2, NUPACK 1999, and RNAstructure show maximum Pearson correlation to chemical mapping data at 60 °C, 40 °C, and 60 °C respectively for Eterna Cloud Lab Round 1. Mean ± SEM of Pearson correlation calculated via bootstrapping, n = 1088 independent constructs. **c**) Ranking across Cloud lab dataset rounds using Spearman rank correlation (compare to Fig. 1e, f). Error bars represent 95% confidence interval of the mean obtained over 1000 iterations of bootstrapping over 24 independent experiments, n = 12,711 independent constructs total. d) (Top) Mean Pearson correlations, calculated over each project (as opposed to each dataset), compared to sequence metrics of the Cloud Lab projects. The strongest correlation to mean correlation was Signal/Noise ratio. (Bottom) Z-score of CONTRAfold-2, calculated over each project, compared to sequence metrics of the Cloud Lab projects.



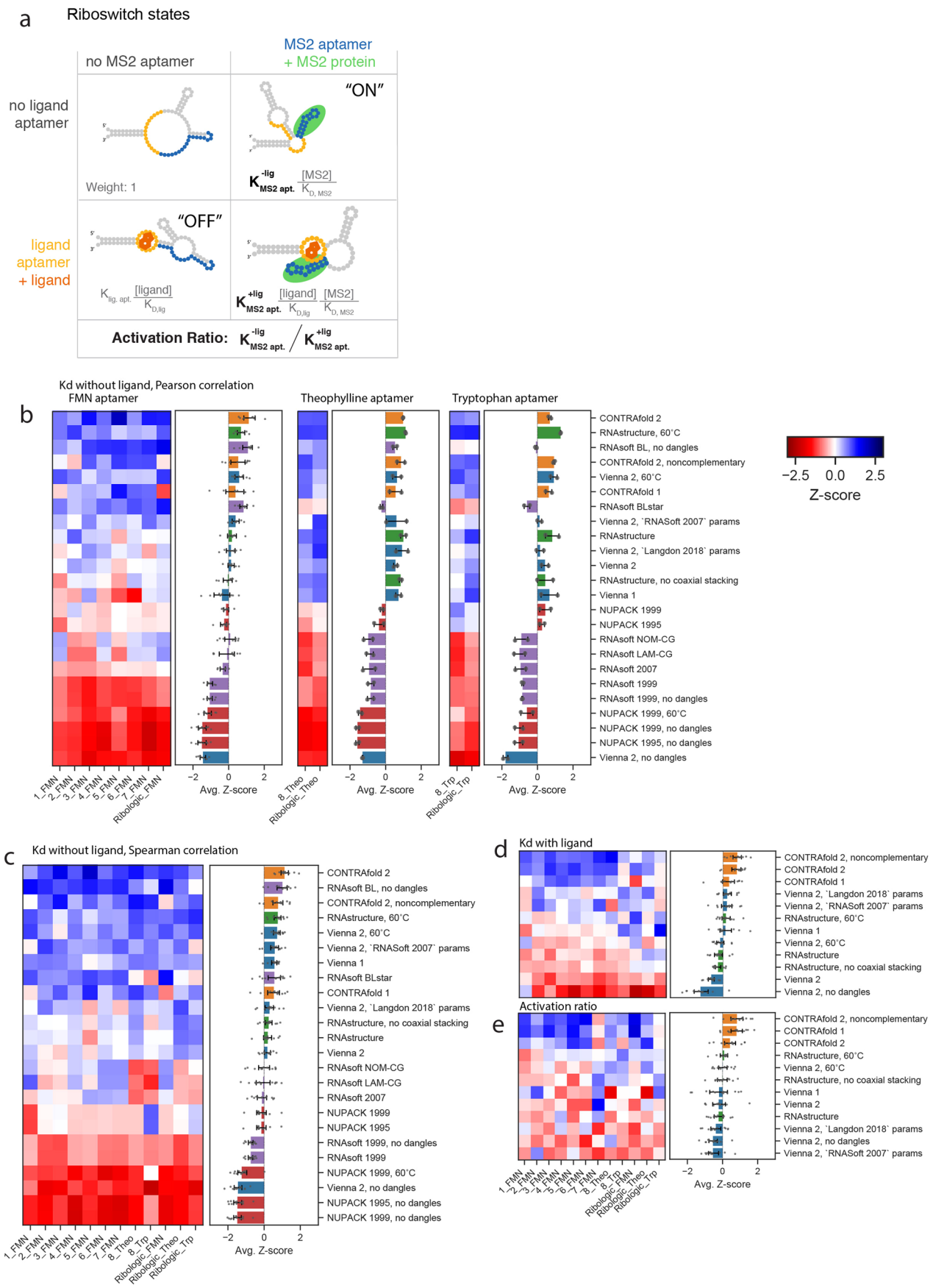
Extended Data Fig. 2 | Example chemical mapping predictions from all package options tested. Example heatmaps of all package options tested for the 'Aires' project (compare to Fig. 1c).



Extended Data Fig. 3 | Summary statistics for EternaBench datasets before and after performing CD-HIT filtering. a) Distributions of sequence properties for chemical mapping data (n = 38,846 before filtering and n = 12,711 independent constructs after filtering, collected across 24 experiments), and B) riboswitch constructs (n = 19,016 independent constructs and n = 7,228 independent constructs after filtering, collected in 12 experiments). Dataset statistics of EternaBench train and test experimental rounds for (**c**) Chemical Mapping (Train set: n = 3,476 independent constructs collected over 6 experiments. Test set: n = 1,492 independent constructs collected over 18 experiments) and (**d**) Riboswitch data (Train set: n = 2,508 independent constructs collected over 3 experiments. Test set: n = 4,018 independent constructs collected over 9 experiments). Center dot, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. For all subplots: center dot, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range.

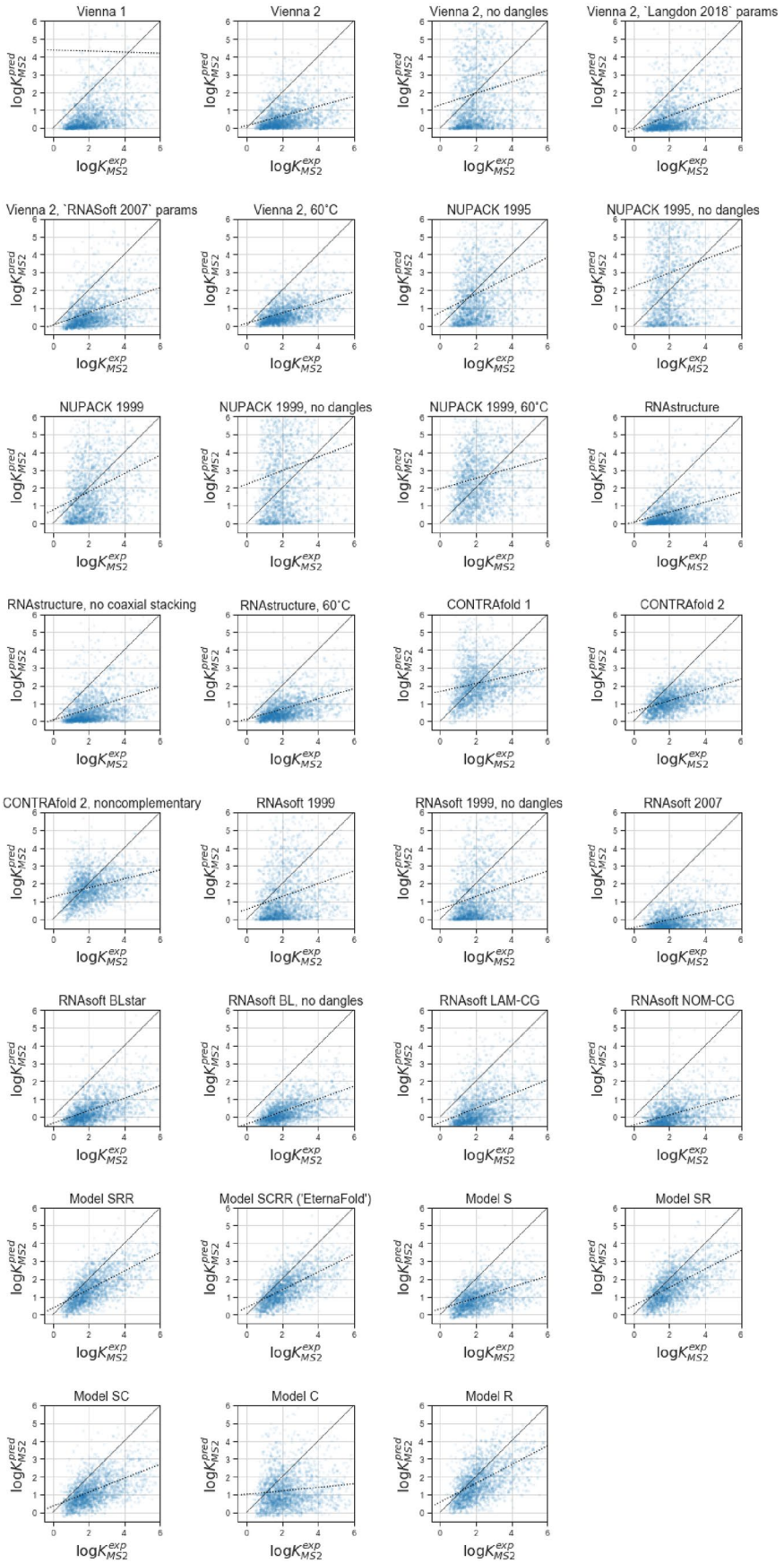


Extended Data Fig. 4 | Overview of all Cloud Labs data. Example reactivity and p(unpaired) heatmaps from example packages for all 24 Cloud Lab rounds. Data have been filtered to exclude nucleotides with reactivity equal to zero or less.

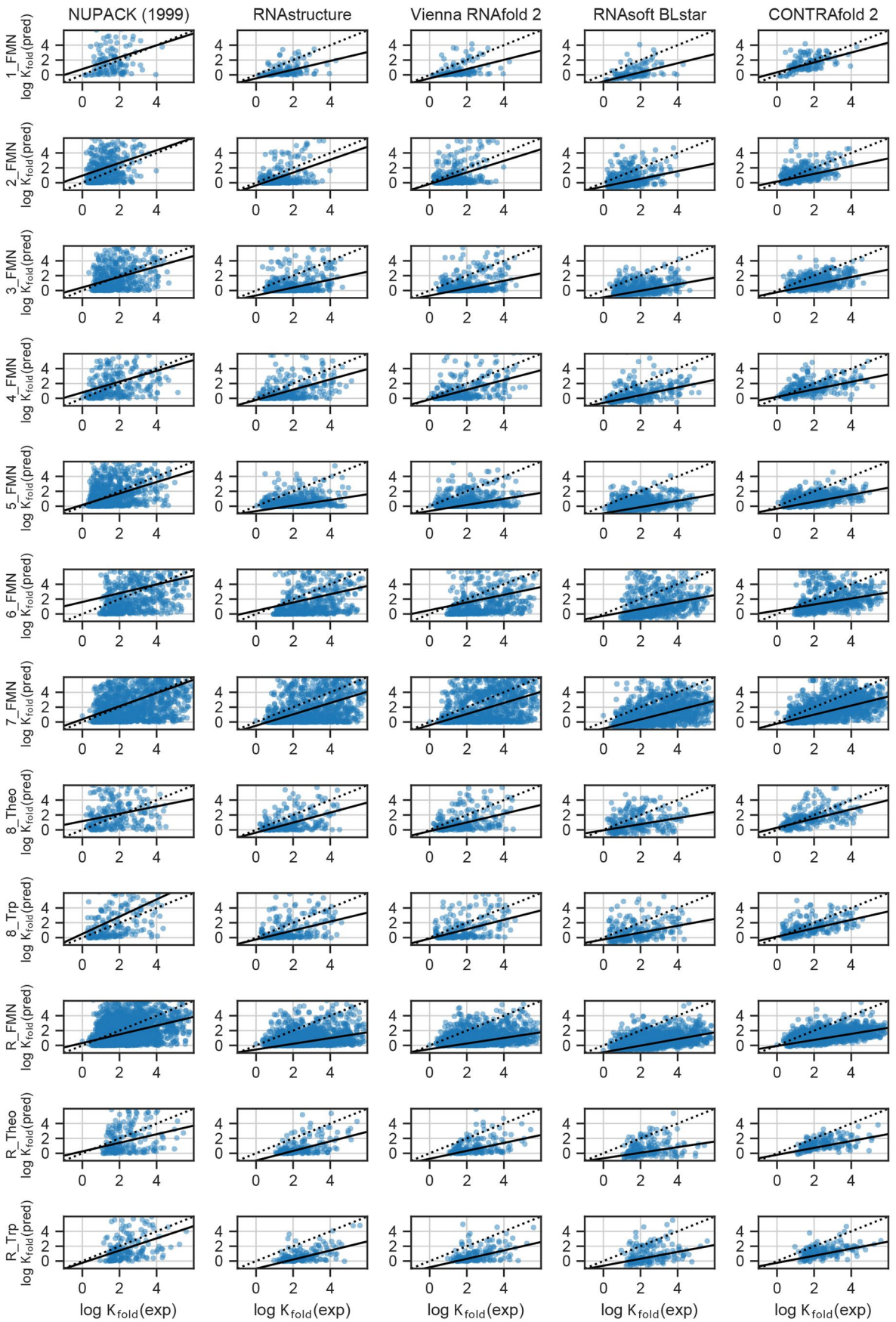


Extended Data Fig. 5 | See next page for caption.

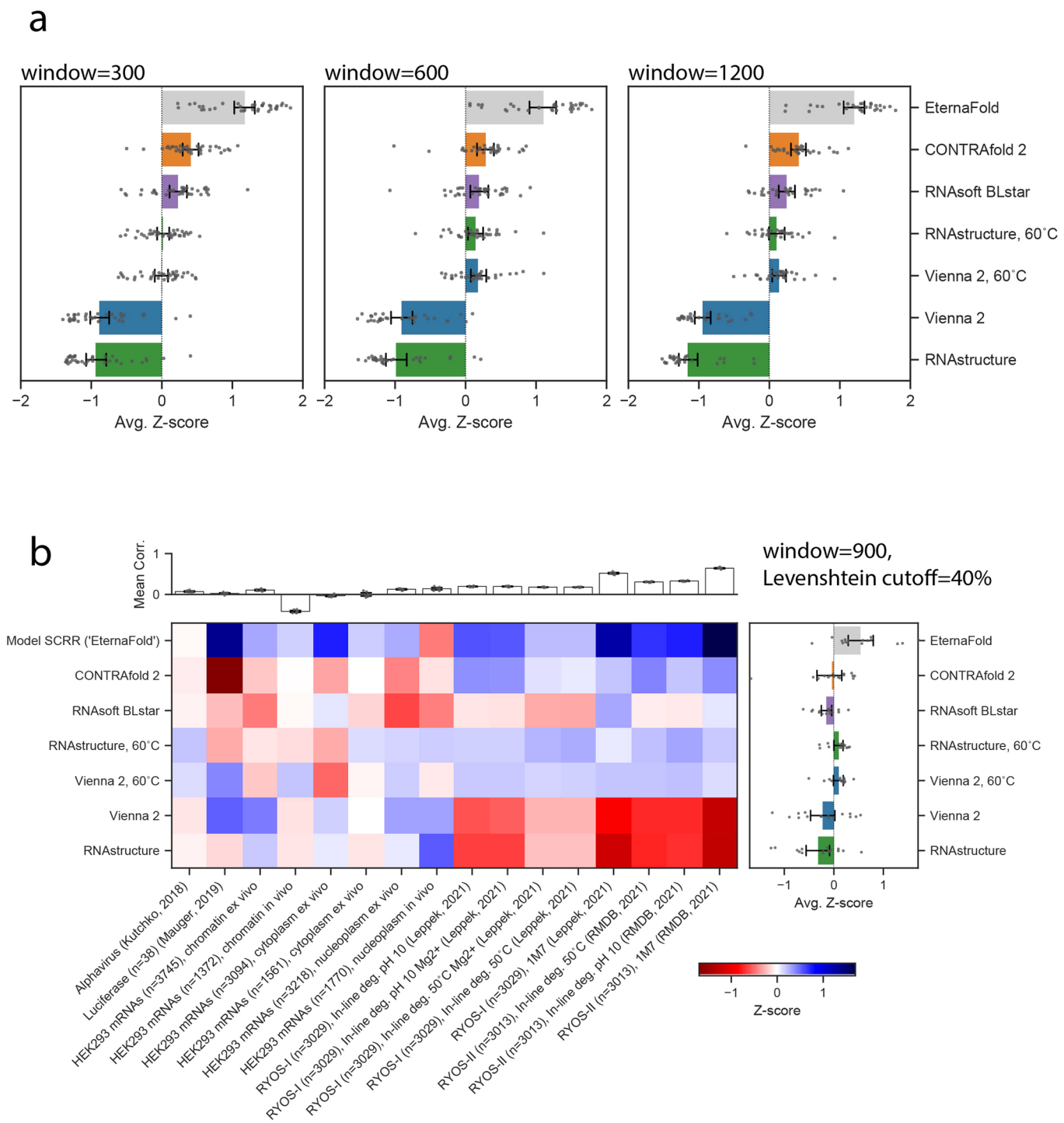
Extended Data Fig. 5 | Extended analysis of package rankings based on riboswitch activity predictions. a) Example set of states for a riboswitch that toggles binding of the fluorescent MS2 protein as an output, controlled by binding the small molecule FMN. The equilibrium constant for forming the MS2 aptamer in the absence of ligand, K_{MS2}^{-lig}, is estimated using the probability of forming the closing base pair for all packages. **b**) Riboswitch Z-scores stratified by input ligand type. Error bars represent standard error on Z-score as calculated by bootstrapping from 6402, 440, and 386 constructs collected over 8, 2, and 2 experiments, respectively. **c**) Overall ranking K_{MS2}^{-lig} calculations using the calculated Spearman correlation (no linear assumption, compare to Fig. 2b). Evaulating the Pearson Correlation of package calculations for (**d**) K_{MS2}^{+lig} as well as (**e**) riboswitch Activation Ratio results in a similar ranking. In C, D, E, error bars represent 95% confidence interval of the mean obtained over 1000 iterations of bootstrapping across datasets, n = 7,228 independent constructs collected over 12 experiments.



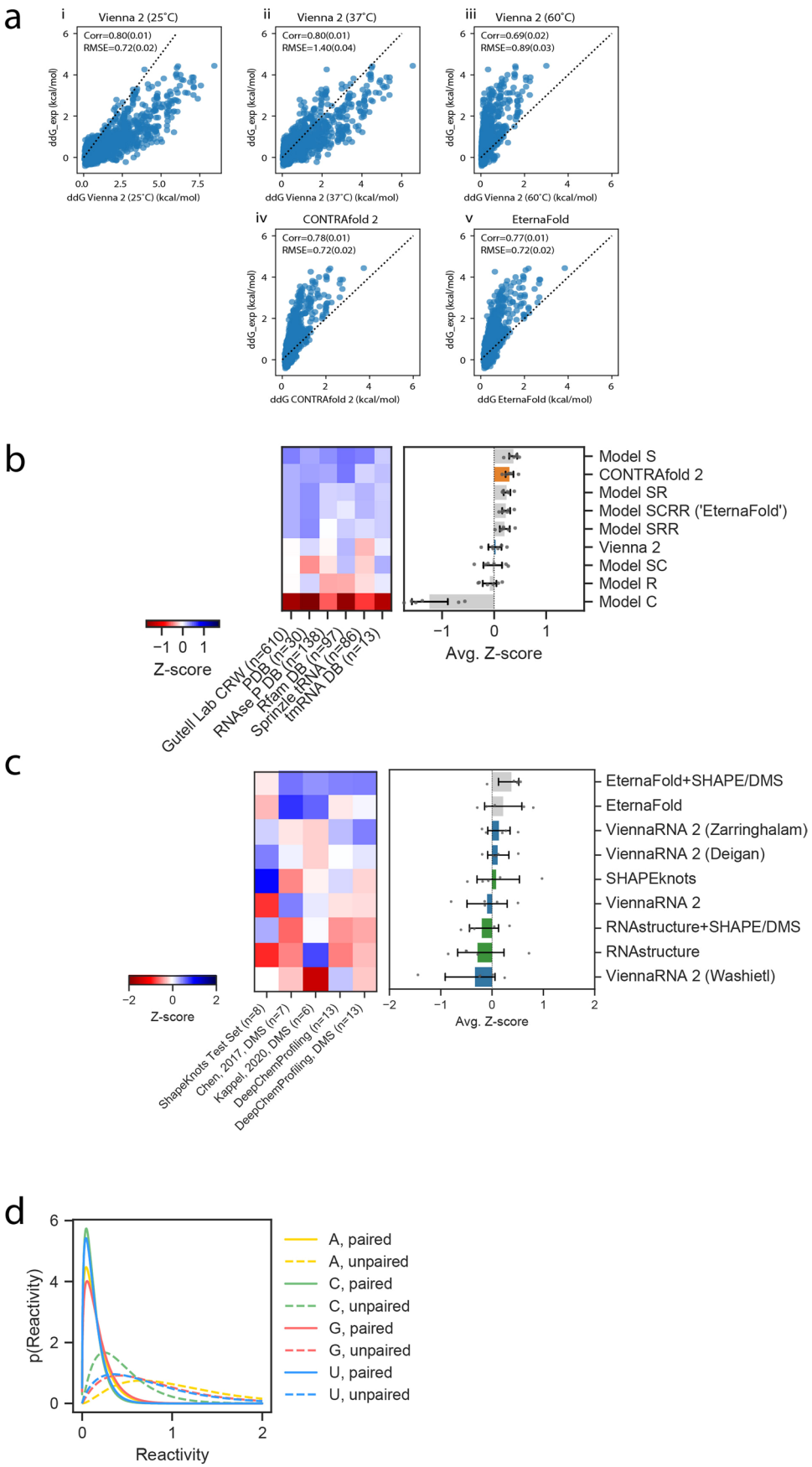
Extended Data Fig. 6 | Example riboswitch predictions from all package options tested. Scatterplots for all options tested for Ribologic dataset. Black solid line indicates line of best fit.



Extended Data Fig. 7 | Example riboswitch predictions across all datasets. Scatterplots for representative packages on all riboswitch datasets. Black solid line indicates line of best fit.

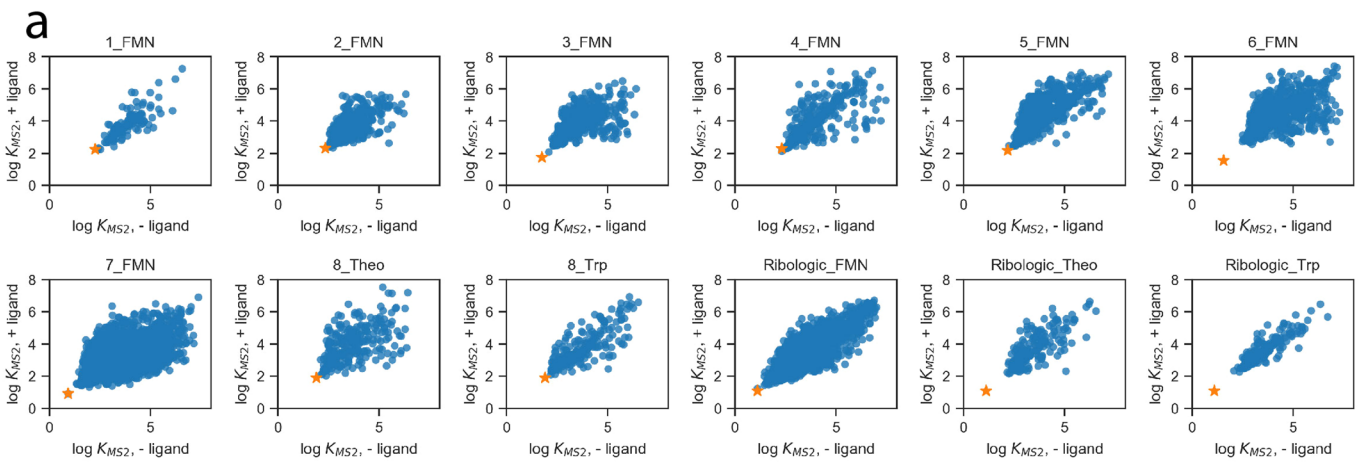


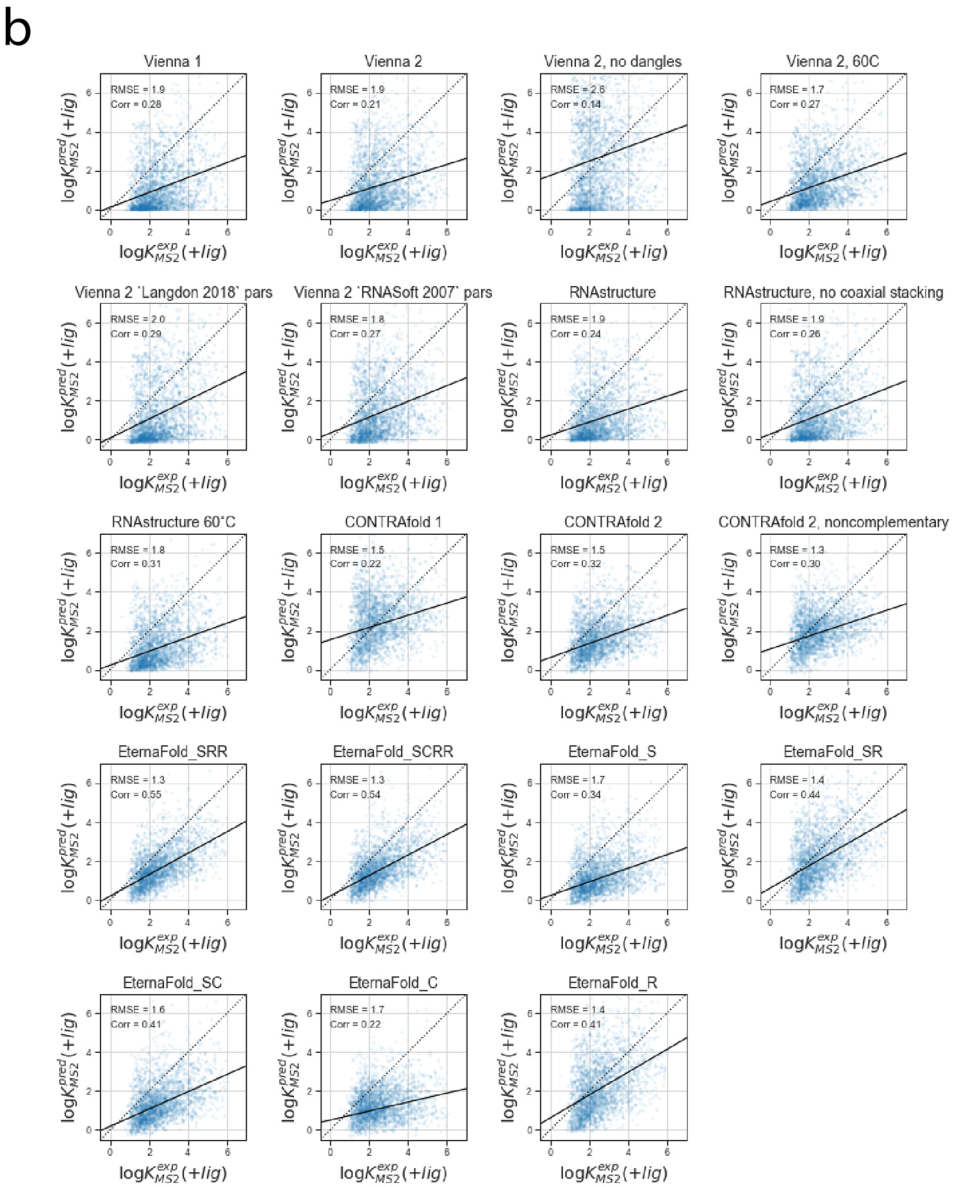
Extended Data Fig. 8 | Effect of window size and Levenshtein distance filtering for independent chemical mapping test set. a) Calculating p(unpaired) using varying sliding windows of size 300, 600, and 1200 does not change the overall ranking obtained across datasets, compare to Fig. 4b, which was calculated for window size 900 (n = 31 datasets for all). Package ranking is also consistent for a redundancy cutoff of 40% b) (n = 16 datasets included after filtering based on 40% cutoff by windowed Levenshtein distance). Error bars in A and B represent 95% confidence interval for the mean Z-score as calculated by bootstrapping across respective number of datasets for each.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Extended data corresponding to EternaFold development and test set evaluation. a) Comparing Vienna, CONTRAfold, and EternaFold predictions in predicting free energy of PUM binding. i) Replication of ddG_exp for both PUM WT and mutant binding from (Becker, 2019). The same calculation in Vienna 2 at 37 °C shows lower Root-mean-squared error (RMSE) (ii), but higher RMSE at 60 °C (iii). CONTRAfold 2 shows no improvement over Vienna at 37 °C (iv), but EternaFold shows modest improvement over both (v). **b**) Package performance for the S-Processed test set is qualitatively similar to results on the ArchiveII-NR test set (cf. Figure 3b). Error bars represent 95% confidence interval of the mean calculated with 1000 iterations of bootstrapping over n = 6 independent datasets, which contain 974 independent constructs total. **c**) Evaluating SHAPE- and DMS- directed folding. Error bars represent 95% confidence interval of the mean calculated with 1000 iterations of bootstrapping over n = 5 independent datasets of RNAs with known secondary structures,, which contain 47 constructs total. **d**) Potentials learned from EternaFold training and used in SHAPE-directed structure prediction.





Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Extended data corresponding to predicting riboswitch affinity in the presence of small molecule ligands. a) log K_{MS2}^{-lig} and log K_{MS2}^{+lig} values of riboswitches included in filtered datasets. Black starred datapoint indicates reference value used for K_{obs}^{ref}. **b**) Estimates for the RiboLogic FMN dataset for log K_{MS2}^{+lig} in all package options able to make estimates with constrained-partition functions.

nature research

Corresponding author(s): Rhiju Das

Last updated by author(s): May 26, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No code was used in data collection process.
Data analysis	Mapseeker v2.0 (https://github.com/eternagame/MAPseeker) was used to process RNA MAP-seq "Cloud Lab" datasets. HiTrace was used to process "DeepChemicalProfiling" datasets. Requirement: MATLAB 2011a. EternaBench code, used to run algorithms and perform statistical tests, is available at https://www.github.com/eternagame/EternaBench. EternaBench requirements: Python 3.7, numpy 1.19.5, seaborn 0.11.1, scipy 1.3.2, tqdm 4.60.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets used here for evaluation are available at https://www.github.com/eternagame/EternaBench. The original cloud lab datasets are available at the RNA Mapping Database28 under accession IDs ETERNA_R00_0000 (Round 00), ETERNA_R69_0000 (Round 01), ETERNA_R70_0000 (Round 02), ETERNA_R71_0000 (Round 03), ETERNA_R72_0000 (Round 04), ETERNA_R73_0000 (Round 05), ETERNA_R74_0000 (Round 06), ETERNA_R75_0000 (Round 07), ETERNA_R76_0000 (Round 08), ETERNA_R77_0002 (Round 09), ETERNA_R78_0001 (Round 10), ETERNA_R79_0001 (Round 11), ETERNA_R80_0001 (Round 12), ETERNA_R81_0001

(Round 13), ETERNA_R82_0001 (Round 14), ETERNA_R83_0003 (Round 15), ETERNA_R84_0000 (Round 16), ETERNA_R85_0000 (Round 17), ETERNA_R86_0000 (Round 18), ETERNA_R87_0001 (Round 19), ETERNA_R89_0000 (Round 20), ETERNA_R91_0000 (Round 21), ETERNA_R92_0000 (Round 22), ETERNA_R94_0000 (Round 23).
The riboswitch raw datasets are downloadable from the supporting information of Andreasson et al. PNAS (2022), at https://www.pnas.org/doi/10.1073/pnas.2112979119.
The "Deep Chemical Profiling" SHAPE and DMS test sets are available for download at https://github.com/DasLab/DeepChemicalProfiling.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed. Sample sizes for high-throughput RNA structure mapping "cloud lab" experiments were determined by the maximum allowable size to achieve sufficient read depth on each library construct. The "Deep Chemical Profiling" constructs were selected to represent 13 RNAs of known structure from a range of biological contexts. The remaining datasets, the riboswitch datasets and chemical mapping datasets of natural RNAs, were taken from literature and therefore the sample size not determined.
Data exclusions	Chemical mapping data was processed to remove nucleotides with reactivity over the 98th percentile of reactivity values, also called "Winsorization". Constructs probed in the context of small molecules were removed from the Cloud Lab datasets in order to only compare chemical mapping data in the context of standard buffer conditions. Nucleotides in stretches of polyA > 6 were also removed due to demonstrated reduced signal from polyA stretches.
Replication	The algorithm rankings obtained in this work were evaluated over 24 independent structure mapping and 12 riboswitch datasets, as well as over more than 20 datasets collected from independent groups. Algorithm rankings replicated over all these contexts. Experimental replications were performed for the first cloud lab dataset, and the Deep Chemical Profiling data. This work did not include experimental replicates other than initial checks that the chemical mapping protocol was replicable.
Randomization	Randomization is not relevant because conditions were constructed and there was not subjective allocation of samples to experimental groups.
Blinding	Investigators were not blinded to the study, but all constructs considered were either designed by citizen scientists not involved in experimental characterization or analysis, or data taken from other independent peer-reviewed works.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging