Diverse Database and Machine Learning Model to narrow the generalization gap in RNA structure prediction

Albéric A. de Lajarte*[1], Yves J. Martin des Taillades*[2], Colin Kalicki[1], Federico Fuchs Wightman[1], Justin Aruda[1], Dragui Salazar[1], Matthew F. Allan[1], Casper L'Esperance-Kerckhoff[1], Alex Kashi[1], Fabrice Jossinet[3], Silvi Rouskin[1,†]

[1] Department of Microbiology, Harvard Medical School, Boston, Massachusetts, USA
[2] Department of Biochemistry, Stanford University, Stanford, California, USA
[3] Faculty of Life Sciences, University of Strasbourg, Strasbourg, France

* these authors contributed equally
† To whom correspondence should be addressed: E-mail: silvi@hms.harvard.edu

Abstract

Understanding macromolecular structures, such as proteins and nucleic acids, is critical for discerning their functions and biological roles. Advanced techniques like crystallography, NMR, and CryoEM have facilitated the determination of over 180,000 protein structures, all cataloged in the Protein Data Bank (PDB). This comprehensive repository has been pivotal in developing deep learning algorithms for predicting protein structures directly from sequences. In contrast, RNA structure prediction has lagged, primarily due to a scarcity of RNA structural data. Here, we present the secondary structures of 1098 pre-miRNAs and 1456 human mRNA regions determined through chemical probing. We develop a novel deep learning architecture, inspired from the Evoformer model of Alphafold and traditional architectures for secondary structure prediction. This new model, eFold, was trained on our newly generated database and over 100,000 secondary structures across multiple sources. We benchmark eFold on a set of challenging RNA structures and show that both our new architecture and dataset contributes to increasing the prediction performance compared to state of the art end-to-end methods. All together, these results reveal that merely expanding the database size is inadequate; rather, incorporating a greater diversity and complexity of structures is crucial for enhancing performance.

Introduction

The structural integrity of RNA is critical to its function, playing a vital role in how RNA molecules interact within the cell. Beyond its primary sequence, RNA harbors a second layer of information in its ability to form secondary structures that can conceal or expose crucial regulatory binding sites for proteins, microRNAs, and other RNAs. These secondary structures, dictated by Watson-Crick base pairings, are pivotal in RNA's lifecycle, affecting localization, translation, degradation, and more [1., 2., 3., 4.]. While X-ray crystallography, Nuclear Magnetic Resonance (NMR), and CryoEM are adept at elucidating RNA's tertiary structure, their applicability is limited for long RNA sequences. Moreover, it is plausible that many long RNAs, apart from specialized ones such as ribosomal RNAs, may not adopt stable tertiary structures, further complicating their studies with conventional methods. As a result, the focus shifts to RNA's secondary structure, which is an intrinsic characteristic of all RNA molecules. To predict this secondary structure, dynamic programming algorithms are frequently used, identifying the conformation with the lowest energy based on the Turner nearest neighbor rules [5., 6.], a set of experimental parameters for the energy of different base pairs in the context of diverse structure motifs.

Although widely used, these algorithms have their limitations, being confined to canonical base pairs and structures and often unable to predict more complex interactions such as non-canonical base pairs or pseudoknots. A contrasting approach is to learn RNA structure models directly from data. In this vein, recent advancements have seen the emergence of numerous deep learning methods [7., 8., 9.]. These methods aim to predict the secondary structure as a base pairing matrix from the sequence alone. This methodology makes minimal assumptions about the rules governing RNA structure and, with adequate data, could capture more complex motif types. Nonetheless, a significant challenge in deep learning is ensuring a sufficiently diverse and high-quality dataset. A training set lacking in diversity results in poor performance in out-of-domain contexts. This problem is exacerbated in RNA studies due to the unique sequences and structures of different RNA families, leading to many distinct domains. For instance, Flamm et al. [10.] demonstrated how performance significantly declines when algorithms are trained on a specific set of families or sequence lengths and then tested on different ones. Similarly, Bugnon et al. [11.] examined this effect, observing diminished performance when applying published algorithms to very long ncRNA and mRNA. Szikszai et al. [12.] also highlighted the risks of using test sets composed of similar families and advocated for the curation of inter-family test sets.

There are a few databases commonly used by all deep learning methods. The bpRNA database [13.] consists of over 100,000 sequences and structures aggregated from multiple sources including PDB and the RNA families database (Rfam). The PDB database [14.] includes 1,790 RNA tertiary structures as of December 2023. ArchiveII [15.] and RNAStralign [16.] contain 3,975 and 30,451 sequences, respectively. Additionally, a substantial dataset of 800,000 sequences, together with chemical probing data for DMS and SHAPE and data-derived structure models, was recently released for the Ribonanza competition [17.]. However, many of these datasets require extensive filtering to eliminate redundant and low-quality sequences, such as bpRNA90, a subset of bpRNA featuring 28,000 non-redundant sequences. Additionally, these databases tend to be heavily biased towards a limited number of short ncRNA families, which adversely affects the generalization capabilities of commonly used algorithms.

We examine this issue on four widely-used secondary structure algorithms, each employing different methodologies. RNAStructure Fold [18.] is a classic model that utilizes the Nearest Neighbors parameters and dynamic programming to determine the minimum free energy of a structure. EternaFold [19.], adapted from ContraFold [20.], employs context-free grammar to learn the probability of different motifs and structures. MXFold2 [8.], a deep learning model, predicts the probability of all potential base pairs and conformations as multiple base pairing matrices. Both EternaFold and MXFold2 are hybrid approaches, leveraging dynamic programming to determine the minimum free energy structure from the predicted energy motifs. UFold [7.], on the other hand, is an end-to-end model that directly predicts a base pairing matrix from the sequence, requiring only a minor post-processing step for conversion into a valid structure. We benchmark these algorithms on a set of challenging sequences, and show poor generalization capabilities. This leads us to the development of a new database of relevant and diverse sequences to mitigate this generalization gap.

The state-of-the-art method for high throughput RNA structure prediction involves probing the RNA molecule with reagents that measure the probability of each nucleotide being paired or unpaired. The two most frequently used methods are DMS-MaPseq and selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE). In DMS-MaPseq dimethyl sulfate (DMS) reacts at the Watson-Crick face of adenine and cytosine, while SHAPE identifies the most flexible parts of the RNA backbone, typically where unpaired bases are found. Both methods induce mutations in the sequence during reverse transcription, which can be detected through high-throughput next-generation sequencing. Analyzing a multitude of sequencing reads provides a mutation fraction per nucleotide which, after normalization, is interpreted as a pairing probability. This data, when added as a pseudo-energy term to classical algorithms, significantly enhances the accuracy of the predicted structure [21., 22.]. Here, we employed the DMS-MaPseq

method to probe and model the structure of a variety of RNAs from new and underrepresented families. Our efforts aim to improve generalization across different RNA domains. The data collected, covering biologically significant sequences such as the 3' ends of mRNAs and pri-miRNA of the human genome, is now accessible in our online database (rnandria.org/app/rnandria/page1-659f36682f84e02b0ff732d7). Using this new data along with existing RNA structure databases, we trained a new architecture, eFold, based on the Evoformer from Alphafold [23.] and Convolutional Neural Networks. We curate a set of challenging test structures and demonstrate enhanced generalization capabilities compared to state-of-the-art deep learning models.

Results

Secondary structure algorithms do not generalize to new domains

Current work on secondary structure prediction relies heavily on the same databases for their training sets. A notable issue is the lack of diversity in these datasets, which are predominantly biased towards short ncRNAs. Rfam [24.] is a database of all the known RNA families and clans, counting 146 clans to date, and consistently adding new ones. However, only a few clans such as tRNA, rRNA and sRNA are significantly represented in commonly used databases, and most sequences are typically less than 200 nucleotides long (Figure 1a, Supplementary Figure 2). This bias allows deep learning models to excel within these family domains, but their performance substantially diminishes outside of them.

In order to evaluate performance outside the common testing sets, we curated three datasets from published RNA structures. The first one is a set of 356 unique structures from the Protein Data Bank (PDB), that will serve as a baseline on which all algorithms excel. These sequences are obtained by filtering out from PDB any entry with multiple chains such as RNA-RNA, RNA-DNA or RNA-protein, which results in mainly short non-coding RNAs that fold well in the absence of proteins. The two other sets are more challenging, and aim at testing the algorithms for new families and different sequence lengths. The long non coding RNA dataset (Long ncRNA) is a collection of 15 long RNA curated by Bugnon et al. [11.] from various publications, with length in the range of 1000 to 2000 nucleotides. The other challenging set is composed of 58 sequences from multiple published viral genomes. As the viral RNA are initially greater than 10000 nucleotides, we segmented them into multiple modular structures of around 200 nucleotides. The full curation process for these three datasets is detailed in the Methods.

We evaluate four widely used algorithms on these curated test sets: RNAstructure Fold, EternaFold, MXFold2 and UFold by calculating an F1 score distribution between the predicted and published structure models (Figure 1b). The F1 score is a commonly used metric for classification, and averages between the precision and sensitivity of detecting correct base pairs. An F1 score of 1.0 indicates identical structures. All algorithms consistently demonstrate remarkable efficacy on PDB, with an average F1 score of around 0.9. Yet their performances significantly drop on the challenging test sets, with an average F1 score for the viral structures typically around 0.7, and even lower for the long ncRNAs with an average F1 score around 0.45 for RNAstructure Fold, EternaFold, MXFold2 .Among the four algorithms evaluated, UFold notably underperforms on new test sets, particularly in predicting base pairs within long non-coding RNAs. This suggests that most of these algorithms are primarily designed or trained for RNA types typically represented in the PDB test set, lacking exposure to newer RNA domains. The lack of generalization is exacerbated for UFold as it is fully an end-to-end algorithm with almost no prior on the structure of RNA. To address these limitations, we collected a new dataset for training these algorithms to enhance their adaptability and performance in diverse RNA domains.

A new database of biologically relevant RNAs to bridge the generalization gap

To address the generalization gap in RNA secondary structure prediction, we probed 4,550 sequences from the 3'end of human mRNAs (Methods) and 1292 primary microRNA (pri-miRNA) sequences of length 200 to 1000 nucleotides long using DMS-MapSeq (Figure 2a). Previous studies have shown that adding chemical probing constraints to RNAStructure yields models with a high degree of accuracy, between 90-100% identical to models obtained through crystallography, cryoEM, or phylogenetic covariation.[25., 26.]. In order to ensure high quality DMS signal we used a stringent coverage cut off of at least 3000 reads per base (Methods). The median signal for adenine and cytosine (AC) over uracil and guanine (GU) was 5.3 fold, with a two phase distribution between DMS reactive and non reactive bases (Figure 2b). DMS reactivity for each RNA was then used as constraints in *RNAstructure*'s Fold algorithm to determine the secondary structure.

We assess the quality of the final dataset in several ways. Over 70% of sequences underwent dual independent probing. The correlation between the DMS signal replicates, as well as the F1 score of the resulting structures, is depicted in Figure 2c. The reproducibility of the signals is highlighted by a high mean Pearson correlation between DMS replicates of 0.95 for the mRNA and 0.99 for the pri-miRNA, which results in a mean F1 score between structure replicates of 0.86 for the mRNA and 0.97 for the miRNA, indicating a strong consistency in the results. To assess the quality of

our data-derived structure models, we used the area under the receiver operating curve. An AUROC close to 1.0 indicates that DMS reactive positions are located within loops and DMS unreactive positions are located within stems, therefore the model is consistent with the data. Although a high AUROC does not validate the model, a low AUROC indicates the model is incorrect either due the presence of alternative structures or poor data quality. Indeed gold standard structures have high AUROCs [27.]. We report the AUROC value for each structure in the database and filter out any structure with AUROC less than 0.8 (Methods).

We validate that the DMS signal brings valuable information to the structure model by running RNAStructure Fold without the DMS constraint. Indeed, the mRNA and pri-miRNA datasets cannot be trivially predicted from sequence alone as seen by a range of F1 scores of 0.3 to 0.7 (Figure 2d, Supplementary Figure1). These lower F1 scores are comparable to the performance of RNAstructure Fold on the rRNA, which increases in accuracy from ~50 to ~90% upon incorporating chemical probing data [25]. In contrast, other large datasets such as Ribonanza[17.] can be predicted with high accuracy directly from the sequence. The low F1 scores of RNAstructure on our datasets, mirroring its performance on other challenging sets, implies the presence of difficult to predict, complex structures. This highlights the value of our dataset in enhancing the training of models to better generalize across a diverse array of RNA domains.

To further explore the level of diversity, we analyzed the inter-dataset sequence similarity for several datasets (Supplementary Figure 3). We find that our datasets have a similarity between sequences of 0.59 and 0.64 for the mRNA and pri-miRNA dataset, much lower than the average similarity of 0.74 of Ribonanza, and close to the average similarity between random RNA sequences of 0.55. More importantly, our datasets don't show clusters of highly similar sequences like we see for Archive II, and can be thus used entirely without filtering for redundant sequences.

Our final database post filtering for high read depth includes 1,456 mRNA and 1,098 pri-miRNA structures. This dataset is available in a new online database (rnandria.org/app/rnandria/page1-659f36682f84e02b0ff732d7), which allows users to search and filter by data features including species, condition, dataset, and more. Researchers can then visualize and download their data of interest. We included visualization of both the secondary structure and DMS signal (Figure 3). The secondary structures, along with their DMS reactivity per base, was performed using RNArtist (Methods), a tool suited for representing long and complex RNA structures.

A novel architecture for secondary structure prediction

Our approach to predicting RNA secondary structure directly from sequence data utilizes a new neural network architecture, prominently featuring four eFold modules. These modules draw inspiration from the Evoformer block used in AlphaFold, as depicted in Figure 4. The core design of each eFold block consists of dual channels: one channel processes the sequence representation through self-attention layers, and the other handles pairwise representation via ResNet convolutional layers. Specialized connections between layers in each channel facilitate effective information exchange. Convolutional neural networks (CNNs) have been heavily used in many deep learning structure prediction models, and are thus integral to our architecture. Their suitability for handling image-like representations of pairing matrices, along with their computational efficiency, makes them preferable over the graph neural network blocks found in the original Evoformer. However, CNNs typically have limitations in capturing long-range interactions, an issue we address by incorporating self-attention layers. These layers process the sequence representation without range limitations, effectively complementing the CNNs.

As shown in Figure 4b, the sequence representation is initialized using a simple table embedding, while the pair representation adopts a method similar to that in the UFold architecture. This setup encodes the sequence as a binary map representing all 16 possible base pair interactions and includes an additional channel that factors in the energy of different pairs. Following the eFold blocks, only the pair representation is advanced further. It is then processed through a ResNet block, resulting in the generation of the pairing matrix. This novel architecture is designed to balance the strengths of both CNNs and self-attention layers, aiming to enhance the accuracy and efficiency of RNA secondary structure prediction from sequence data.

Improving the performance of end-to-end model prediction

Our model's training was executed in two distinct phases.The first phase involved using existing databases like bpRNA, RNAstralign, and Ribonanza to compile a substantial pre-training dataset of over 120,000 unique sequences and structures. Redundant sequences were filtered out using BLAST [28.], and low-quality structures were removed(Supplementary Figure 4). This pre-training phase alone enabled our model to achieve commendable performance, surpassing UFold in the long non-coding RNA test set with a mean F1 score of 0.32, compared to UFold's 0.12 (Figure 5a). Performance on other test sets was similarly effective. In the second phase, we fine-tuned our model using our curated mRNA and pri-miRNA datasets. This led to an increase in the mean

F1 score across all test sets, underscoring the value of new and diverse data for enhancing generalization (Figure 5a).

The contributions of different training components were analyzed in the ablation test performance (Figure 5b). Comparing UFold's and eFold's architecture, each trained on the same pre-training dataset, we observe the most substantial improvement on the long non-coding RNA dataset and a modest improvement on the viral structures. This notable enhancement, particularly for long RNA, may be linked to the eFold blocks architecture's ability to facilitate both long and short-range interactions throughout the network, combined with the use of relative positional embeddings aiding in generalizing to new lengths. The fine-tuning stage on Rnandria marked the second most substantial improvement for both the viral structures and long non-coding RNA. We experimented with fine-tuning exclusively on the mRNA or pri-miRNA datasets. The results indicated that the mRNA dataset was more critical for generalization, as performance closely mirrored the baseline when trained solely on mRNA (Figure 5b). This outcome could be attributed to the mRNA dataset's larger size and its diversity in terms of sequence lengths, which are essential for generalizing to the long non-coding RNA test set.

Discussion

In this study we tackled two major challenges in RNA secondary structure prediction: the limited generalizability of existing algorithms and the need for a robust, end-to-end predictive model. We demonstrate that a large class of biologically relevant RNA molecules , such as mRNAs, viral RNAs, long ncRNAs, and pri-miRNAs have challenging RNA structures that are difficult to predict by conventional algorithms. We curate a list of published structures and show that machine-learning based approaches do not generalize well in new domains outside their training set. The focus of our work was twofold: first, we collected a new and diverse dataset of RNA structures, and second, we developed a state-of-the-art model, eFold, for predicting RNA secondary structures.

Our newly generated dataset includes secondary structures of 1,456 human mRNA regions and 1,098 pre-miRNAs, determined through chemical probing. This dataset represents a significant expansion in the diversity of RNA structures available for study, addressing a critical gap in existing data. The lack of variety in current datasets, predominantly featuring short ncRNAs, has limited the effectiveness of RNA structure prediction algorithms. By incorporating a broad spectrum of RNA families and lengths, we aimed to improve the generalization capability of predictive models.

The eFold model, inspired by the Evoformer block of Alphafold and traditional architectures, marks a considerable advancement in RNA secondary structure prediction. This end-to-end deep learning architecture was trained on our newly generated database as well as over 100,000 secondary structures from multiple sources. Our benchmarks demonstrate that eFold outperforms existing state-of-the-art end to end methods, especially in predicting complex RNA structures such as long-ncRNAs. This improvement highlights the importance of not just expanding the size of the training database but also enhancing its diversity and complexity.

Our research emphasizes the critical need for more comprehensive and broader datasets, with a particular emphasis on long RNAs featuring complex structures, to improve the accuracy of RNA structure prediction across various biologically significant RNAs. Currently, the prediction accuracy for the curated list of published structures of long ncRNAs is around 40%-50%, a stark contrast to the approximately 90-100% accuracy achieved for short nc-RNAs in the PDB. The development of eFold, alongside our novel database, signifies one step forward in this field. By integrating a substantially larger amount of data, including data on alternative structures, it should be possible to attain a level of precision in RNA secondary structure prediction for long RNAs that parallels the high accuracy we now see with short nc-RNAs. Such high accuracy is needed for comprehending the diverse roles and functions of RNA structures in biology.

Acknowledgment
This work was supported by the Burroughs Wellcome fund G-1018246.01

Author Contributions

A.L,Y.M., and S.R. conceived and designed the project with input from C.K. A.L. and Y.M. build and trained eFold with input from C.K. and A.K. F.W, J.A. carried out experiments and design of the sequencing and structure probing approach. M.A. designed the human mRNA dataset. D.S performed DMS-MaPseq for the mRNA sequences. C.LK. built the web interface. F.J. drew all the structure models in the RNAndria database.

Declaration of Interests
The authors declare no competing interests

<u>Figure legends</u>

Figure 1: Performance of state-of-the art algorithms on diverse RNA domains

(a) A violin plot showing the F1 score distribution between the structure model predicted by each algorithm and a published structure model for three different test sets - PDB, Viral structures, and Long ncRNAs. F1 score is the harmonic mean of precision and sensitivity and ranges between 0 and 1. Please be aware that the violin plots are truncated within the 0 to 1 range, resulting in a loss of some data at the extremes.
(b) RNA clans and sequence length distribution of bpRNA, RNAstralign and archiveII datasets.

Figure 2: Data quality, complexity , and diversity of RNAndria database
(a) Sequence length distribution for the pri-miRNA dataset (left scale,red) and mRNA dataset (right scale, teal)**(b).** Distribution of non-normalized DMS reactivity for the pri-miRNA and mRNA datasets. The non-reactive G/U bases are highlighted in gray with a mean reactivity of 0.006, while the reactive A/C bases are in orange with a mean reactivity of 0.03.**(c)** Distribution of Pearson correlation scores between DMS signal replicates (top) and F1 score between structure models derived from replicate data (bottom)**(d)** Distribution of F1 score between structures predicted by RNAstructure Fold from sequence alone compared to RNAstructure Fold with chemical probing constraints, for Ribonanza, Pri-miRNA, and human mRNA regions (mRNA) datatest. Please be aware that the violin plots are truncated within the 0 to 1 range, resulting in a loss of some data at the extremes.

Figure 3: RNAndria database architecture
(a) Main page of the online database Rnandria, showing the browsing, filtering, searching and download features.**(b)** Sub page for one sequence of the pri-miRNA hsa-mir-1205, showing its sequence, secondary structure, and DMS reactivity

Figure 4: eFold architecture
(a) High level schematic of the eFold architecture, with the two input representations and the output pairing matrix **b.** Low level view of one eFold block, showing the sequence and pair representation channels, as well as the inter channel connections

Figure 5: eFold performance
(a) F1 score distribution between each model prediction and the ground truth, for the three test sets. The PT model is eFold only pretrained and PT+FT is eFold pretrained then finetuned on the mRNA and pri-miRNA datasets. Please be aware that the violin plots are truncated within the 0 to 1 range, resulting in a loss of some data at the extremes.**(b)** Ablation study for different architecture and training strategies. The baseline model is eFold pretrained then finetuned. We try fine tuning on only the mRNA, only the pri-miRNA, no finetuning, and no finetuning on the UFold architecture. The delta F1 score is with respect to the baseline mean F1 score per dataset

Methods

Library design

To generate the human mRNA library, all human protein-coding transcripts were downloaded from GENCODE Release 43 [29.]. Every transcript with a non-AUG start codon or without an annotated 5' UTR, CDS, and 3' UTR was discarded. To focus on 3' UTRs and maximize the likelihood of successful RT-PCRs, for each gene, the longest common suffix (i.e. subsequence beginning at the 3' end) among all isoforms was selected. Each gene was discarded if it had no annotated name, if its longest common suffix was shorter than 253 nt, or (to minimize mispriming and misalignment) if its longest common suffix contained any 41-mers that were present in the longest common suffix of any other gene. To verify that alignment to the wrong gene would be minimal, a simulated FASTQ file containing every 128 nt segment in 64 nt increments from every gene's longest common suffix was aligned to all longest common suffixes with Bowtie2 v2.5.1 [30.].

Primers for human mRNAs were designed against the longest common suffixes of all remaining genes using Primer3 v2.6.1 [31.] with the parameters PRIMER_TASK=generic, PRIMER_PICK_LEFT_PRIMER=1, PRIMER_PICK_INTERNAL_OLIGO=0, PRIMER_PICK_RIGHT_PRIMER=1, PRIMER_PRODUCT_SIZE_RANGE=253-1021, PRIMER_OPT_SIZE=28, PRIMER_MIN_SIZE=20, PRIMER_MAX_SIZE=36, PRIMER_WT_SIZE_LT=0.0, PRIMER_WT_SIZE_GT=0.0, PRIMER_MIN_TM=65.0, PRIMER_OPT_TM=75.0, PRIMER_MAX_TM=80.0, PRIMER_PAIR_MAX_DIFF_TM=10.0, PRIMER_GC_CLAMP=0, PRIMER_MAX_NS_ACCEPTED=0, P3_FILE_FLAG=0, PRIMER_EXPLAIN_FLAG=1, SEQUENCE_OVERHANG_LEFT=TAATACGACTCACTATAGGG. A T7 promoter ending in three G bases was prepended to every forward primer. Genes that received no primer pairs were discarded; for those that received more than one pair, the pair that would produce the longest amplicon was selected. The sequences of the amplicons (including only the last 3 Gs of the T7 promoter) were used as the reference sequences in the subsequent analysis.

To select amplicons for verification in HEK 293T cells, four microarray datasets of gene expression in HEK 293T cells from two studies [32.] were downloaded from the National Center for Biotechnology Information's Gene Expression Omnibus [33.] (accession numbers GSM389756, GSM389757, GSM609202, and GSM609205). Microarray probe IDs were converted to gene names using DAVID [34.]. The amplicons of at most 600 nt (the maximum paired-end read length on Illumina sequencers) were ranked by the minimum expression level of their genes among the four datasets (to be conservative); the 96 highest-ranked amplicons were chosen for mutational profiling in HEK 293T cells.

To generate the pri-miRNA library, human genome annotations were downloaded from miRBase (mirbase.org, Kozomara A, Birgaoanu M, Griffiths-Jones S
Nucleic Acids Res (2019) 47:D155-D162; Kozomara A, Griffiths-Jones S

Nucleic Acids Res (2014) 42:D68-D73; Kozomara A, Griffiths-Jones S
Nucleic Acids Res (2011) 39:D152-D157; Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ Nucleic Acids Res (2008) 36:D154-D158; Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ Nucleic Acids Res (2006) 34:D140-D144, Griffiths-Jones S Nucleic Acids Res (2004) 32:D109-D111). After removing redundant sequences, 1292 miRBase hairpins were selected for downstream library creation. The miRBase hairpins were then padded with their flanking genomic sequence to a total of 192 nt using Biopython's Entrez interface. These 1292 sequences were then separated into 3 sub-libraries to reduce jackpotting during library processing. To each construct, a sub-library specific pair of 19 nt universal primer binding regions were added to minimize primer associated PCR and reverse transcription bias, bringing the full length library size to 230 nt. The final sequences were then synthesized as an oligo pool (Agilent Technologies).

DMS data generation

For the in vitro transcription of the human mRNA library, all the designed amplicons were amplified via PCR, adding a T7 promoter sequence at the forward primer (5′ TAATACGACTCACTATAG 3′) using a 2X PCR PreMix (Syd Labs, Cat. MB067-EQ2N), human cDNA (ZYAGEN, Cat. HD-UR-40) and an INTEGRA VIAFLO384 channel pipette. For each 384 plate, 5 µl of each individual reaction were pooled together and purified using a Zymo DNA clean up kit (Cat. D4003). The PCR product pool was used as template for the in vitro transcription, using the T7 MEGAscript kit (Thermo Fisher, Cat. AMB13345) according to the manufacturer's instructions. After DNase treatment, the transcribed RNAs were purified using a Zymo RNA clean up kit (Cat. R1017). 10 µg of the purified RNAs were diluted in 10 µl of nuclease free water, denatured for 1 minute at 95 degrees and placed on ice for 2 minutes. 88.5 µl of the refolding buffer (sodium cacodylate 397.5 mM and $MgCl_2$ 6 mM final) were added and the RNAs were left refolding at 20 minutes at 37ºC. After, 1.5 µl of DMS (dimethyl sulfate) was added to obtain a 1.5% final concentration, for 5 minutes at 37ºC, while shaking at 600 RPM. The reaction was stopped by adding 60 µl of β-mercaptoethanol, followed by Zymo RNA clean up.

Libraries were generated as previously described [35.], and the indexed samples were sequenced in a Illumina NextSeq 2000.

For the pri-miRNA library, *in vitro* transcription templates were prepared via 8 cycles of PCR using 2x Q5 Master Mix (New England Biolabs, Cat. M0492S) with a T7 promoter containing forward primer as previously described. Following column clean up (Zymo Research, Cat. D4013), 100 ng of template DNA was used in a 2 hour *in vitro* transcription reaction using the HiScribe T7 High Yield RNA (New England Biolabs, Cat. E2040S) kit according to the manufacturer's instructions. Following DNase digestion (Invitrogen, Cat. AM2238) and column clean up (Zymo Research, Cat. R1017), 1 ug of RNA from each library was DMS treated separately as described above, substituting a 1.5% final DMS concentration for 2.5%. Following DMS treatment, the samples were reverse transcribed with Induro Reverse Transcriptase (New England Biolabs, Cat. M0681L) for 30 minutes at 55°C using the appropriate reverse primer for each library according to the manufacturer protocol. After alkaline hydrolysis with 1 ul of 4M

NaOH, cDNA was cleaned and concentrated using Zymo Oligo Clean and Concentrator (Zymo Research, D4060). Sub-library specific primers were then used to amplify 1 ul of cDNA template for 24 cycles using Advantage 2 PCR Mix (Takara Bio, Cat. 639206). Finally, library indexing was carried out with the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Cat. E7645S) and sequenced on a NextSeq 1000 P2 cartridge (Illumina, Cat. 20046813) in accordance with the manufacturer's instructions.

DMS data processing and filtering

For the processing of raw reads, the fastq files were handled using SEISMIC [36.], which aligned the reads to the reference sequence and quantified mutations per base. Sequences with less than 3000 aligned reads or where less than 50% of bases had coverage of at least 3000 reads were excluded. In some cases, an untreated RNA sample was sequenced to provide a baseline, aiding in the identification of abnormally high mutation rates that might indicate inaccuracies in the reference sequence. If mutations consistently occurred towards a specific base, the reference sequence was updated; otherwise, the sequence was discarded. Biological replicates were analyzed to check experimental consistency. Replicates with a pearson correlation below 0.8 were removed, while those with higher correlation were averaged for further use. Bases were masked with a value of -1000 if their coverage was below 3000, if they were G or U bases, or if they were part of the primer region in the pri-miRNA dataset. Signal normalization was then performed per reference, using the 95th percentile as the maximum value, with values above this threshold clipped to 1. The RNAstructure Fold algorithm, incorporating the DMS signal as a constraint, was then applied. Structures with an AUROC of less than 0.8 in relation to the DMS signal were excluded. Through these measures, 15% of the pri-miRNA dataset and 70% of the human mRNA dataset were removed, ensuring the high quality of the remaining data.

Data quality assessment:

Pri-miRNA has replicates for 99% of the references and human mRNA has replicates for 50% of the references. This is because we have very stringent filtering steps, so some references were valid in one experiment and not in another. We estimated the reproducibility of our data by comparing our replicates. The DMS signals replicates were compared using the Pearson correlation score. We used each pair of replicates as DMS constraints for RNAstructure, and we compared the two structures using the F1 score. The median Pearson score was 0.997 for pri-miRNA and 0.95 for mRNA. The median F1 score for pri-miRNA was 1 and 0.86 for mRNA. Note that in the final dataset, both replicates were aggregated into a single signal.

We also estimated the robustness of the data as the number of reads decreases. The DMS-MaPseq method consists in counting the mutated bases at each position across a large number of reads. To estimate the error due to the number of reads, we model the

mutation fraction as a binomial distribution divided by the number of trials. We bootstrap 10 new dms signals for each reference from the original signal using this model. We predict the structure using RNAstructure and these DMS signals. The structures are systematically the same when predicted with the original signal and the bootstrapped signals, suggesting that subsampling error doesn't have an influence on the structure prediction.

Test sets

For our test sets, we sourced data from a variety of published papers by different research groups. In selecting entries from the Protein Data Bank (PDB), we focused on those classified under "Polymer Composition" as RNA, and with a "Number of Assemblies", "Number of Distinct Molecular Entities", and "Total Number of Polymer Instances" all set at 1. This approach yielded 356 entries, which were then converted from tertiary to secondary structures using the RNApdbee webserver*, applying the default settings: 3DNA/DSSR as the conversion software and the hybrid algorithm method.

The viral structures dataset was created by segmenting long viral RNA sequences into smaller, independent modules. We utilized the HIV structure from Watts et al. [37.], the SARS-CoV structure from Lan et al. [38.], the Hepatitis virus structure from Mauger et al.[39.], and the Alphavirus structure from Kutchko et al. [40.]. For each of these structures, we identified modular structures characterized by fully closed loops and high agreement between the structure and chemical probing data (AUROC > 0.8). After segmenting the chemical probing signal (DMS or SHAPE), RNAstructure was rerun with each fragment. Fragments were retained if the new structure aligned with the corresponding segment of the chemical probing signal (AUROC > 0.8). The long non-coding RNA (lncRNA) dataset was sourced from Bugnon et al. [11.], with the only modification being the removal of sequences exceeding 2,000 nucleotides in length.

External databases curation

We aggregated public databases such as bpRNA [13.], RNAstralign [15.], Zuber [41.], Ribonanza [17.] and ArchiveII [16.] into a pre-training dataset. We applied the following filtering steps: first, we removed duplicate sequences within the databases. We only kept sequences with the canonical bases ACGU. The T bases were converted to U. We filtered out the sequences below 10 nucleotides. We removed sequences for which we have a sequence but no structure. We removed sequences that are common between datasets. For ribonanza specifically, we applied the following filtering: we filtered low-reads and low S/N ratio data. The cutoff was set to more than 500 reads and a S/N ratio - a quality indicator provided by the Ribonanza dataset - greater than 1. The data includes structures predicted with EternaFold. To ensure that the EternaFold-predicted structure was matching the signal, we computed the AUROC between the structure and

the one of the chemical probing signals for each sequence. We used DMS by default, and SHAPE if DMS was filtered out. If the AUROC was below 0.8, the structure was filtered out. We removed redundant sequences within the dataset using BLAST. The primers were masked. If two sequences had over 80% matches on a sequence of over 112 nucleotides, we kept only the best covered sequence. Note that 112 nucleotides correspond to 80% of the most represented length minus the primers.

Structure visualization

All the secondary structures available in RNAndria, along with their DMS reactivity per base, have been plotted from the command line using the tool RNArtistCore (https://github.com/fjossinet/RNArtistCore). Built with Kotlin (https://kotlinlang.org/), RNArtistCore follows an heuristic approach to identify a non-overlapping 2D layout for an RNA secondary structure.

At first, RNArtistCore computes the structural domains (helices and junctions) from paired positions. In 3D, junctions connect and orient helices in order to produce the biologically relevant 3D architectures. Each junction is associated to a class depending on the number of helices connected (apical loops, inner loops, 3-way junctions, 4-way junctions….).Starting from the 5'-end, RNArtistCore is processing each junction in order to find orientations for its helices in order to avoid overlapping between upstream and downstream structural domains in a 2D sketch.

Except for apical loops, the drawing engine implemented in RNArtistCore splits the connected helices in a junction between a single ingoing helix and outgoing ones. The modeling of RNA concepts of RNArtistCore links each junction class to a default positioning for its outgoing helices, relatively to the orientation of its ingoing one. If the default position for an outgoing helix produces an overlap between the next junction and at least one upstream structural domains, RNArtistCore searches for a new position among those still available. When the orientation of an outgoing helix has been defined, this helix becomes the ingoing one for the next junction to be processed.

Once a non-overlapping layout found, it is used to produce an SVG drawing according to some user-defined instructions. RNArtistCore provides a simple and expressive syntax (a.k.a « Domain Specific Language ») to allow the user to parameter the final result. The instructions are described in a script whose name and location has to be specified on the command line. Among all the options available (details levels, colors, line widths,..), one can easily link a dataset (like DMS reactivities) to the structure and map its values to a color gradient.

Some 2D RNA structures can contain nested junctions for which RNArtistCore could not be able to find any non-overlapping organization. In such cases, RNArtistCore can be configured to go backwards in order to test different layouts for upstream junctions that

could unlock the nested ones. For this study, we allowed RNArtistCore to go back until 10 previous junctions if necessary.

Models architecture

Model inputs are RNA sequences only. The output is a 2D RNA structure prediction in the form of a pairing matrix. The sequence is one-hot encoded and then fed through a learned embedding layer (dimension = 64). This learned representation is the input to the eFold module. The first step in this module is constructing the pair representation. The one-hot encoded sequence is transformed using a Kronecker outer product (specified here [3.]) to create an LxLx17 pair representation. The sequence representation goes through a relative positional multihead self-attention module with bias from the paired representation. This module consists of a multiheaded attention mechanism with learned relative position encodings [42.] and pair representation bias added to the logits prior to the softmax function. The sequence representation goes through a feed-forward layer followed by a 1-D convolution module. This includes pointwise convolutions followed by depthwise convolution (number of output channels = 128) and finally another pointwise convolution (number of output channels = 64). This is followed by another feedforward layer and a 2-layer deep sequence representation transition MLP. There are skip connections between each module. The output of the transition MLP is transformed using an outer mean product and added to the pair representation for the second trunk of the eFold module. The pair representation is sent through a dilated 2D resenet with 2 residual-blocks of depth. The output is followed by a 2-layer deep pair representation transition MLP. The eFold module is repeated four times prior to the output layers. Pairing matrix predictions are sent through a final 2D residual block head (output dimensions = LxLx1). Loss is calculated with binary cross entropy. The pairing matrix is post-processed by enforcing constraints such as no sharp loops and only one nucleotide can be paired with another. We reuse the same optimization method as UFold to find the pairing matrix respecting the constraints with minimum modification of the model output.

Training and evaluation

The model is developed using the Pytorch and Ligthning libraries. We use the Binary Cross Entropy (BCE) loss for structure prediction. We use a cluster of 8 Nvidia RTX4090 and train using the DDP strategy for 40 epochs in 2 days on the pretraining set. Due to large memory constraints, we run the model with a batch size of 1, and thus don't require any padding. Using gradient accumulation, the effective batch size is 256. We train with Adam optimizer with a learning rate of 0.0007.

The test sets are run separately on one GPU to ensure maximum precision in the metric computations. All the code is freely available on Github at https://github.com/rouskinlab/DMSensei and the datasets on Huggingface at https://huggingface.co/rouskinlab.

Supplementary Figure legends

Supplementary figure 1: Comparison of RNAstructure prediction with and without chemical probing signal for each dataset
(a) For each sequence of each dataset, we predicted the structure without and with chemical probing signal using RNAstructure [14.]. Then we compared these predictions using the F1 score, using paired and unpaired as classes. The distributions of the F1 score are displayed for each dataset. Please be aware that the violin plots are truncated within the 0 to 1 range, resulting in a loss of some data at the extremes.

Supplementary figure 2: Representation of the external databases and of the families and lengths of the pre-train set
(a) The redundant sequences between bpRNA, RNAstralign and ArchiveII are represented using a Venn diagram.

(b) The distribution in families of the compiled train set, displayed as a pie chart. Ribonanza counts as a family since we don't have labels for it.

(c) The distribution of the length of the compiled train set, displayed as a histogram.

Supplementary Figure 3: Dataset diversity analysis
(a) For each dataset, we compared the sequence to each other sequence using. As a reference frame, we also added a dataset of random sequences. The results are displayed as heatmaps, where one pixel is a similarity indicator between two sequences, from 0 to 1.

Supplementary Figure 4: External databases filtering plot
(a) This plot displays the steps that we used to curate the external databases. The x axis is a series of filtering steps applied to the datasets. The y axis is the number of sequences left at that stage for each dataset.

Supplementary Figure 5: Median of the performance of the state-of-the-art algorithm on multiple datasets
(a) The plot is a table that shows the median values for each violin plot shown on figure 1.b.

References

1. David M. Mauger, B. Joseph Cabral, Vladimir Presnyak, Stephen V. Su, David W. Reid, Brooke Goodman, Kristian Link, Nikhil Khatwani, John Reynders, Melissa J. Moore, and Iain J. McFadyen. mRNA structure regulates protein expression through changes in functional half-life. Biophysics and computational biology, 2019 https://doi.org/10.1073/pnas.1908052116
2. Das S, Vera M, Gandin V, Singer RH, Tutucci E. Intracellular mRNA transport and localized translation. Nat Rev Mol Cell Biol. 2021 Jul;22(7):483-504. doi: 10.1038/s41580-021-00356-8. Epub 2021 Apr 9. Erratum in: Nat Rev Mol Cell Biol. 2021 Apr 28;: PMID: 33837370; PMCID: PMC9346928.
3. Wong JLC, David S, Sanchez-Garrido J, Woo JZ, Low WW, Morecchiato F, Giani T, Rossolini GM, Beis K, Brett SJ, Clements A, Aanensen DM, Rouskin S, Frankel G. Recurrent emergence of Klebsiella pneumoniae carbapenem resistance mediated by an inhibitory ompK36 mRNA secondary structure. Proc Natl Acad Sci U S A. 2022 Sep 20;119(38):e2203593119. doi: 10.1073/pnas.2203593119. Epub 2022 Sep 12. PMID: 36095213; PMCID: PMC9499542.
4. Zhang Y, Burkhardt DH, Rouskin S, Li GW, Weissman JS, Gross CA. A Stress Response that Monitors and Regulates mRNA Structure Is Central to Cold Shock Adaptation. Mol Cell. 2018 Apr 19;70(2):274-286.e7. doi: 10.1016/j.molcel.2018.02.035. Epub 2018 Apr 5. PMID: 29628307; PMCID: PMC5910227.
5. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res. 2010 Jan;38(Database issue):D280-2. doi: 10.1093/nar/gkp892. Epub 2009 Oct 30. PMID: 19880381; PMCID: PMC2808915.
6. Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998 Oct 20;37(42):14719-35. doi: 10.1021/bi9809425. PMID: 9778347.
7. Laiyi Fu, Yingxin Cao, Jie Wu, Qinke Peng, Qing Nie, Xiaohui Xie, UFold: fast and accurate RNA secondary structure prediction with deep learning, Nucleic Acids Research, Volume 50, Issue 3, 22 February 2022, Page e14, https://doi.org/10.1093/nar/gkab1074

8. Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun 12, 941 (2021). https://doi.org/10.1038/s41467-021-21194-4
9. Xinshi Chen, Yu Li, Ramzan Umarov, Xin Gao, Le Song. RNA Secondary Structure Prediction By Learning Unrolled Algorithms. https://doi.org/10.48550/arXiv.2002.05810 (2020)
10. Flamm Christoph, Wielach Julia, Wolfinger Michael T., Badelt Stefan, Lorenz Ronny, Hofacker Ivo L. Caveats to Deep Learning Approaches to RNA Secondary Structure Prediction. Frontiers in Bioinformatics, Volume 2, 2022, 10.3389/fbinf.2022.835422
11. L A Bugnon, A A Edera, S Prochetto, M Gerard, J Raad, E Fenoy, M Rubiolo, U Chorostecki, T Gabaldón, F Ariel, L E Di Persia, D H Milone, G Stegmayer, Secondary structure prediction of long noncoding RNA: review and experimental comparison of existing approaches, Briefings in Bioinformatics, Volume 23, Issue 4, July 2022, bbac205, https://doi.org/10.1093/bib/bbac205
12. Szikszai M, Wise M, Datta A, Ward M, Mathews DH. Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. Bioinformatics. 2022 Aug 10;38(16):3892-3899. doi: 10.1093/bioinformatics/btac415. PMID: 35748706; PMCID: PMC9364374.
13. Danaee P, Rouches M, Wiley M, Deng D, Huang L, Hendrix D. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. Nucleic Acids Res. 2018 Jun 20;46(11):5381-5394. doi: 10.1093/nar/gky285. PMID: 29746666; PMCID: PMC6009582.
14. Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, Green RK, Goodsell DS, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose AS, Shao C, Tao YP, Valasatava Y, Voigt M, Westbrook JD, Woo J, Yang H, Young JY, Zardecki C, Berman HM, Burley SK. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. 2017 Jan 4;45(D1):D271-D281. doi: 10.1093/nar/gkw1000. Epub 2016 Oct 27. PMID: 27794042; PMCID: PMC5210513.
15. Zhen Tan, Yinghan Fu, Gaurav Sharma, David H. Mathews, TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs, Nucleic Acids Research, Volume 45, Issue 20, 16 November 2017, Pages 11570–11581, https://doi.org/10.1093/nar/gkx815
16. Zuber J, Cabral BJ, McFadyen I, Mauger DM, Mathews DH. Analysis of RNA nearest neighbor parameters reveals interdependencies and quantifies the uncertainty in RNA secondary structure prediction. RNA. 2018

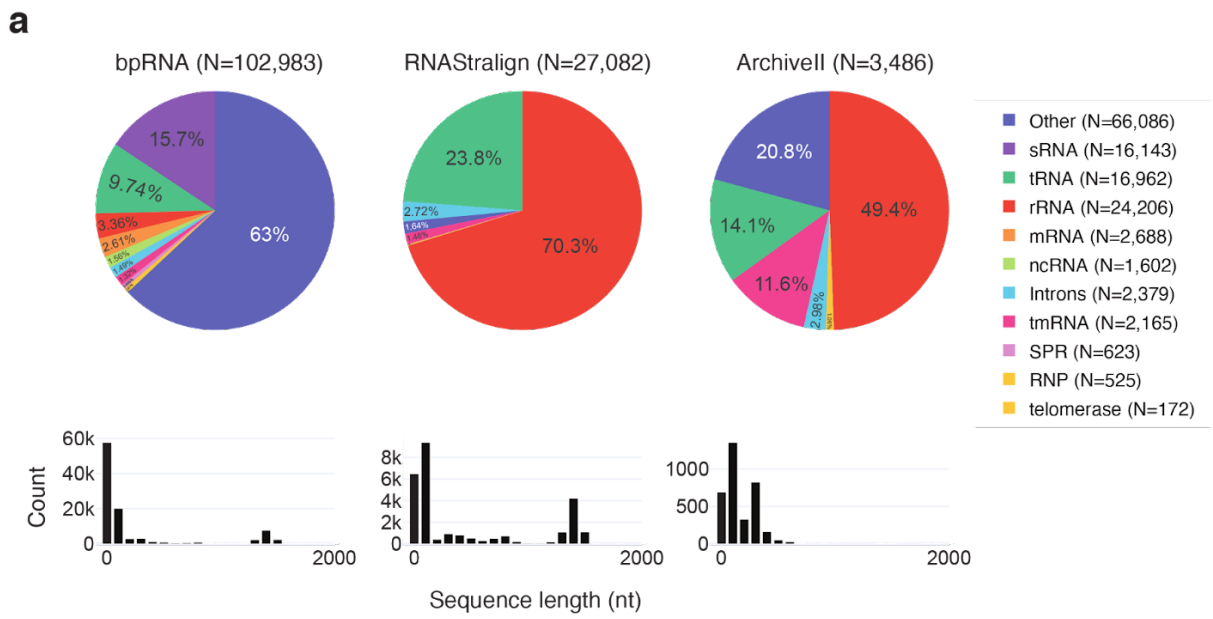
Nov;24(11):1568-1582. doi: 10.1261/rna.065102.117. Epub 2018 Aug 13. PMID: 30104207; PMCID: PMC6191722.

17. Rhiju Das, Shujun He, Rui Huang, Jill Townley, Rachael Kretsch, Thomas Karagianes, John Nicol, Grace Nye, Christian Choe, Jonathan Romano, Maggie Demkin, Walter Reade, and Eterna players . (2023). Stanford Ribonanza RNA Folding. Kaggle. https://kaggle.com/competitions/stanford-ribonanza-rna-folding
18. Reuter, J.S., Mathews, D.H. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11, 129 (2010). https://doi.org/10.1186/1471-2105-11-129
19. Wayment-Steele, H.K., Kladwang, W., Strom, A.I. *et al.* RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nat Methods* 19, 1234–1242 (2022). https://doi.org/10.1038/s41592-022-01605-0
20. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics (Oxford, England). 2006 Jul;22(14):e90-8. DOI: 10.1093/bioinformatics/btl246. PMID: 16873527.
21. Katherine E. Deigan, Tian W. Li, David H. Mathews, and Kevin M. Weeks. Accurate SHAPE-directed RNA structure determination, Proceedings of the National Academy of Sciences, Volume 106, page 97-102, https://doi.org/10.1073/pnas.08069291
22. Zubradt, M., Gupta, P., Persad, S. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat Methods 14, 75–82 (2017). https://doi.org/10.1038/nmeth.4057
23. Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
24. Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, Elena Rivas, Sean R Eddy, Robert D Finn, Alex Bateman, Anton I Petrov, Rfam 14: expanded coverage of metagenomic, viral and microRNA families, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D192–D200, https://doi.org/10.1093/nar/gkaa1047
25. Katherine E. Deigan, Tian W. Li, David H. Mathews, and Kevin M. Weeks. Accurate SHAPE-directed RNA structure determination. Biophysics and computational biology. 2009. https://doi.org/10.1073/pnas.0806929106
26. Cordero P, Kladwang W, VanLang CC, Das R. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. Biochemistry. 2012 Sep 11;51(36):7037-9. doi: 10.1021/bi3008802. Epub 2012 Aug 29. PMID: 22913637; PMCID: PMC3448840.

27. Lan, T.C.T., Allan, M.F., Malsick, L.E. et al. Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. Nat Commun 13, 1128 (2022). https://doi.org/10.1038/s41467-022-28603-2
28. Camacho, C., Coulouris, G., Avagyan, V. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009). https://doi.org/10.1186/1471-2105-10-421
29. Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Carles Boix, Silvia Carbonell Sala, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Kevin L Howe, Toby Hunt, Osagie G Izuogu, Rory Johnson, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Ferriol Calvet Riera, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Maxim Y Wolf, Jinuri Xu, Yucheng T Yang, Andrew Yates, Daniel Zerbino, Yan Zhang, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Michael L Tress, Paul Flicek, GENCODE 2021, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D916–D923, https://doi.org/10.1093/nar/gkaa1087
30. Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012). https://doi.org/10.1038/nmeth.1923
31. Andreas Untergasser, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, Steven G. Rozen, Primer3—new capabilities and interfaces, Nucleic Acids Research, Volume 40, Issue 15, 1 August 2012, Page e115, https://doi.org/10.1093/nar/gks596
32. Joshua Abd Alla, Kristin Reeck, Andreas Langer, Thomas Streichert, Ursula Quitterer, Calreticulin enhances B2 bradykinin receptor maturation and heterodimerization, Biochemical and Biophysical Research Communications, Volume 387, Issue 1,2009,Pages 186-190, ISSN 0006-291X, https://doi.org/10.1016/j.bbrc.2009.07.011.
33. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5. doi: 10.1093/nar/gks1193. Epub 2012 Nov 27. PMID: 23193258; PMCID: PMC3531084.

34. Brad T Sherman, Ming Hao, Ju Qiu, Xiaoli Jiao, Michael W Baseler, H Clifford Lane, Tomozumi Imamichi, Weizhong Chang, DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update), Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W216–W221, https://doi.org/10.1093/nar/gkac194
35. Gabriel Romero-Agosto, Ethan Cox, Silvi Rouskin. DMS-MapSeq Analysis of Antisense Oligonucleotide Binding to lncRNA PANDA. bioRxiv 2023.10.22.563486; doi: https://doi.org/10.1101/2023.10.22.563486
36. Allan, Matthew F. Investigating and Reprogramming RNA Folding with Molecular Probes. DSpace. 2023. https://hdl.handle.net/1721.1/152722
37. Watts, J., Dang, K., Gorelick, R. et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460, 711–716 (2009). https://doi.org/10.1038/nature08237
38. Lan, T.C.T., Allan, M.F., Malsick, L.E. et al. Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. Nat Commun 13, 1128 (2022). https://doi.org/10.1038/s41467-022-28603-2
39. Mauger DM, Golden M, Yamane D, Williford S, Lemon SM, Martin DP, Weeks KM. Functionally conserved architecture of hepatitis C virus RNA genomes. Proc Natl Acad Sci U S A. 2015 Mar 24;112(12):3692-7. doi: 10.1073/pnas.1416266112. Epub 2015 Mar 9. PMID: 25775547; PMCID: PMC4378395.
40. Kutchko KM, Madden EA, Morrison C, Plante KS, Sanders W, Vincent HA, Cruz Cisneros MC, Long KM, Moorman NJ, Heise MT, Laederach A. Structural divergence creates new functional features in alphavirus genomes. Nucleic Acids Res. 2018 Apr 20;46(7):3657-3670. doi: 10.1093/nar/gky012. PMID: 29361131; PMCID: PMC6283419.
41. Sloma MF, Mathews DH. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. RNA. 2016 Dec;22(12):1808-1818. doi: 10.1261/rna.053694.115. Epub 2016 Oct 19. PMID: 27852924; PMCID: PMC5113201.
42. Shaw, Peter, et al. Self-Attention with Relative Position Representations. arXiv:1803.02155, arXiv, 12 Apr. 2018. arXiv.org, https://doi.org/10.48550/arXiv.1803.02155.

Figure 1


a
bpRNA (N=102,983)
RNAStralign (N=27,082)
ArchiveII (N=3,486)
15.7%
9.74%
3.36%
2.61%
63%
23.8%
2.72%
70.3%
20.8%
49.4%
14.1%
11.6%
2.98%
Other (N=66,086)
sRNA (N=16,143)
tRNA (N=16,962)
rRNA (N=24,206)
mRNA (N=2,688)
ncRNA (N=1,602)
Introns (N=2,379)
tmRNA (N=2,165)
SPR (N=623)
RNP (N=525)
telomerase (N=172)
Count
60k
40k
20k
0
8k
6k
4k
2k
1000
500
2000
Sequence length (nt)

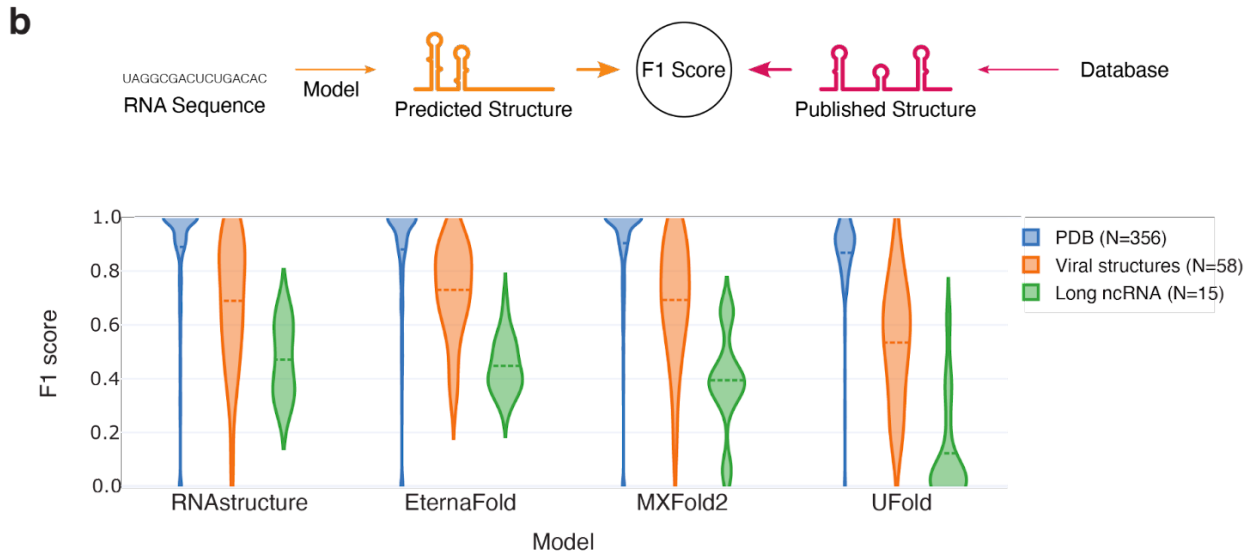

b
UAGGCGACUCUGACAC
RNA Sequence
Model
Predicted Structure
F1 Score
Published Structure
Database
F1 score
1.0
0.8
0.6
0.4
0.2
0.0
RNAstructure
EternaFold
MXFold2
UFold
Model
PDB (N=356)
Viral structures (N=58)
Long ncRNA (N=15)

Figure 2

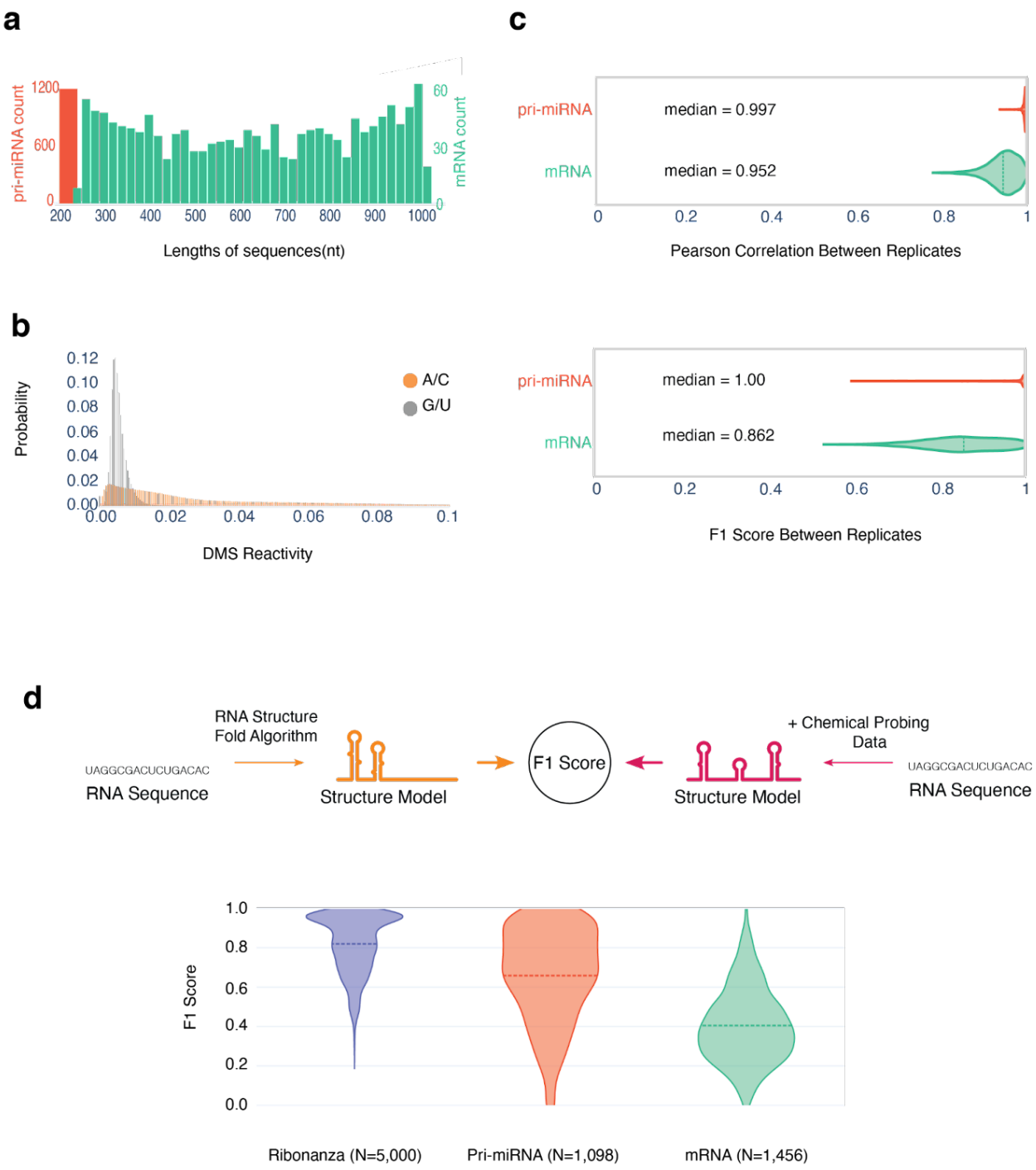
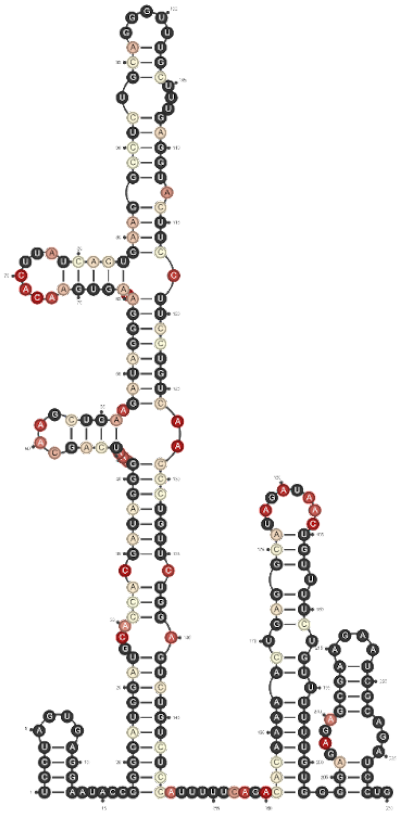

a
pri-miRNA count
mRNA count
Lengths of sequences(nt)
b
Probability
A/C
G/U
DMS Reactivity
c
pri-miRNA
mRNA
median = 0.997
median = 0.952
Pearson Correlation Between Replicates
pri-miRNA
mRNA
median = 1.00
median = 0.862
F1 Score Between Replicates
d
RNA Structure Fold Algorithm
UAGGCGACUCUGACAC
RNA Sequence
Structure Model
F1 Score
+ Chemical Probing Data
Structure Model
UAGGCGACUCUGACAC
RNA Sequence
F1 Score
Ribonanza (N=5,000)
Pri-miRNA (N=1,098)
mRNA (N=1,456)

Figure 3

a

Image	Reference	Dataset	Species	Family	Sequen...	Date	Experimental ...	Cond...
	ENSG00000120696.9	human_mRNA	human	mRNA	GGGGCUA...	2024-01-09	DMS-MaPseq	in vitro
	ENSG00000120688.9	human_mRNA	human	mRNA	GGGUAAG...	2024-01-09	DMS-MaPseq	in vitro
	ENSG00000151208.17	human_mRNA	human	mRNA	GGGAAGC...	2024-01-09	DMS-MaPseq	in vitro
	ENSG00000189120.5	human_mRNA	human	mRNA	GGGACCA...	2024-01-09	DMS-MaPseq	in vitro
	hsa-mir-1204	pri_miRNA	human	pri_miRNA	UCCUAGU...	2024-01-10	DMS-MaPseq	in vitro
	hsa-mir-1203	pri_miRNA	human	pri_miRNA	CAGAGAU...	2024-01-10	DMS-MaPseq	in vitro
	hsa-mir-1205	pri_miRNA	human	pri_miRNA	UCCUAGU...	2024-01-10	DMS-MaPseq	in vitro
	hsa-mir-1200	pri_miRNA	human	pri_miRNA	UCCUAGU...	2024-01-10	DMS-MaPseq	in vitro
	hsa-mir-1202	pri_miRNA	human	pri_miRNA	UCCUAGU...	2024-01-10	DMS-MaPseq	in vitro
	hsa-mir-1207	pri_miRNA	human	pri_miRNA	UCCUAGU...	2024-01-10	DMS-MaPseq	in vitro
	hsa-mir-1206	pri_miRNA	human	pri_miRNA	UCCUAGU...	2024-01-10	DMS-MaPseq	in vitro

b



Dataset

pri_miRNA

Family

pri_miRNA

Species

human

Sequence

UCCUAGUGAGGAAUACCGGGGAUGGAUGCAC
CACGAUAGGGAAAUCAGCAAGCUGAAGAUAGG
GAAAGUGAACACUUAUCACUGAAGGCCUCUGC
AGGGUUUGCUUUGAGGUACUUCCUUCCUGUC
AACCCUGUUCUGGAGUCUGUCUCCAUUUUUC
AGACACAAAAACUGAGGCAUAGAUAACUGUUU
UCUGUUUUUUGUGGGGAGAGAGCGAAGAAUC
GCAGAUCCUG

Dot Bracket

((((....))))....((((((((((...(((.(((((((...((((....)))).
((((((((.(((((.....)))))(((((((((.
(((.....)))...)))))).))))).))))))))..))))))).))).))))))))).....
.((((((((((.((((((.....))).))).)).)))))))).(((....
((((....))))...)))..

Figure 4

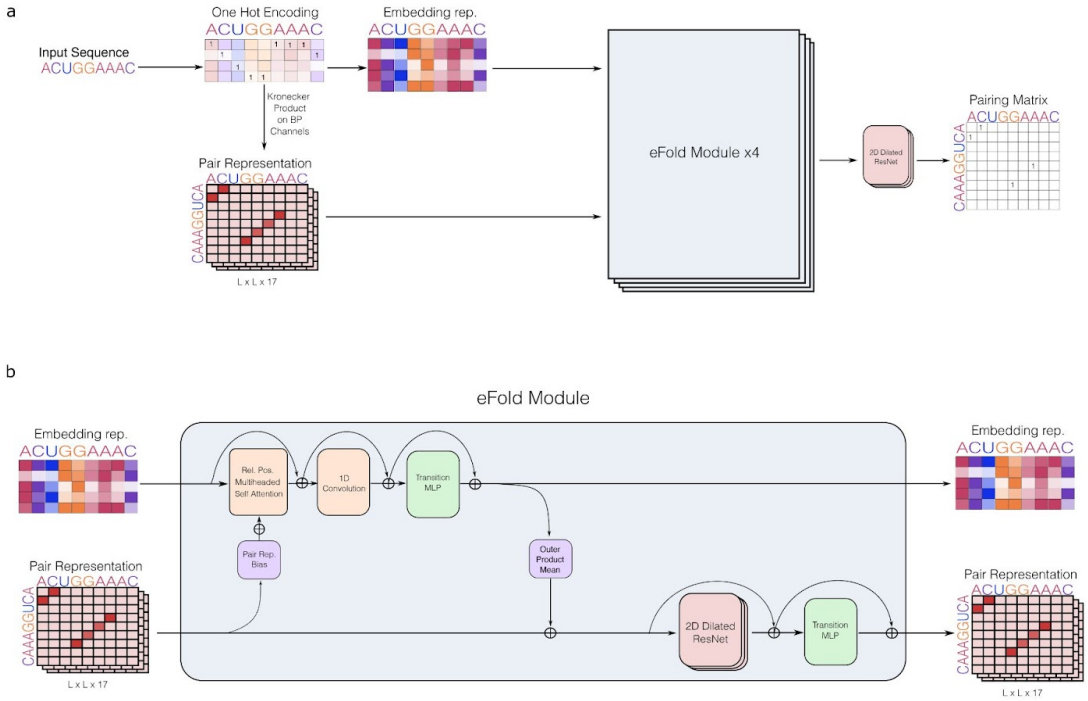
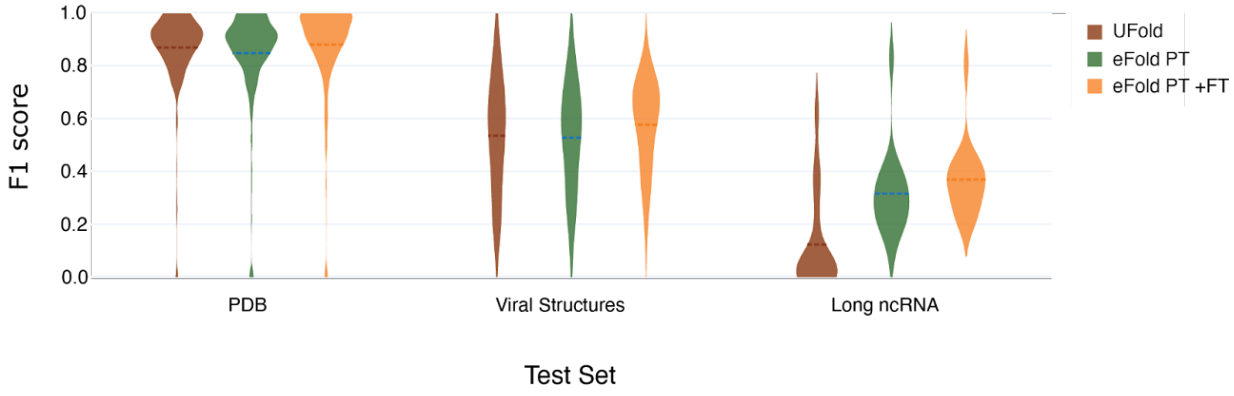
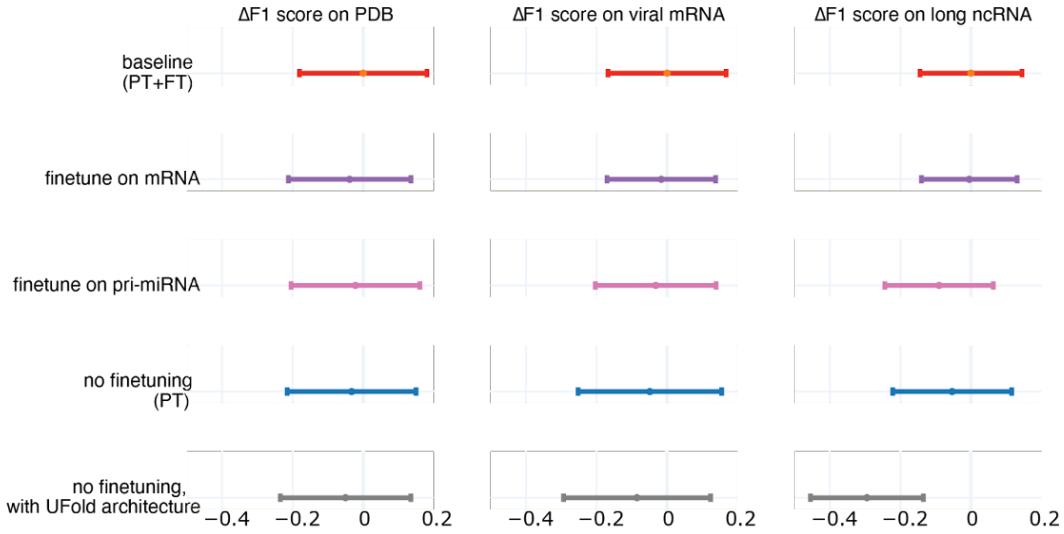

a
Input Sequence
ACUGGAAAC
One Hot Encoding
ACUGGAAAC
Kronecker Product on BP Channels
Pair Representation
ACUGGAAAC
CAAAGGUCA
L x L x 17
Embedding rep.
ACUGGAAAC
eFold Module x4
2D Dilated ResNet
Pairing Matrix
ACUGGAAAC
CAAAGGUCA
b
eFold Module
Embedding rep.
ACUGGAAAC
Pair Representation
ACUGGAAAC
CAAAGGUCA
L x L x 17
Rel. Pos. Multiheaded Self Attention
1D Convolution
Transition MLP
Pair Rep. Bias
Outer Product Mean
2D Dilated ResNet
Transition MLP
Embedding rep.
ACUGGAAAC
Pair Representation
ACUGGAAAC
CAAAGGUCA
L x L x 17

Figure 5

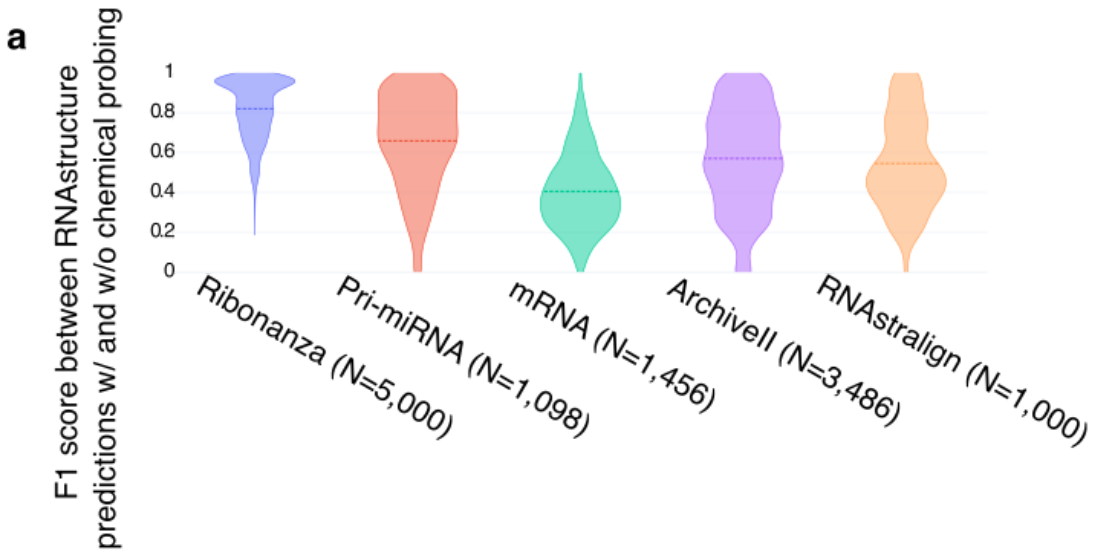
a


1.0
0.8
0.6
0.4
0.2
0.0
F1 score
PDB
Viral Structures
Long ncRNA
Test Set
UFold
eFold PT
eFold PT +FT

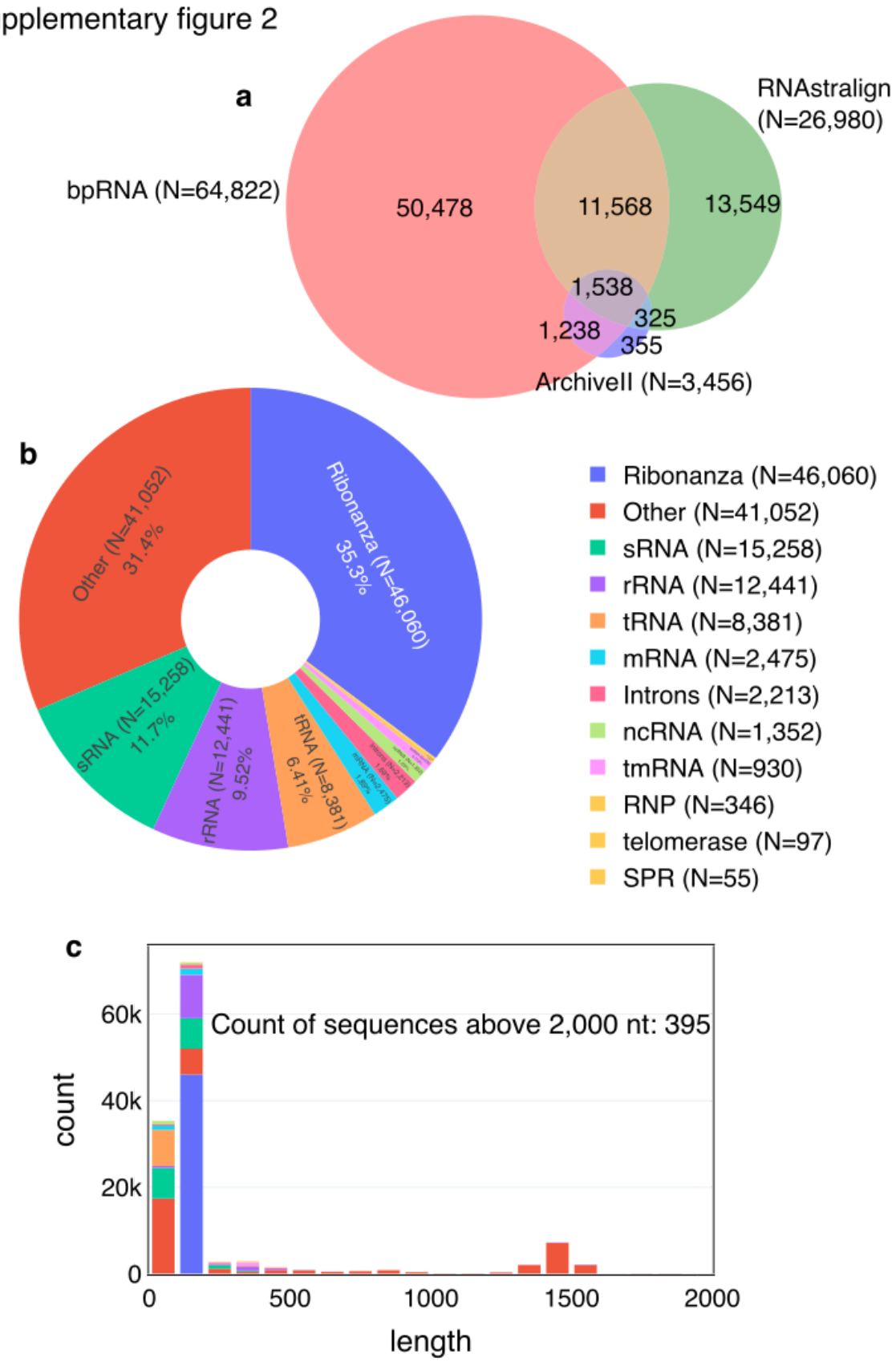
b


ΔF1 score on PDB
ΔF1 score on viral mRNA
ΔF1 score on long ncRNA
baseline (PT+FT)
finetune on mRNA
finetune on pri-miRNA
no finetuning (PT)
no finetuning, with UFold architecture
−0.4
−0.2
0
0.2

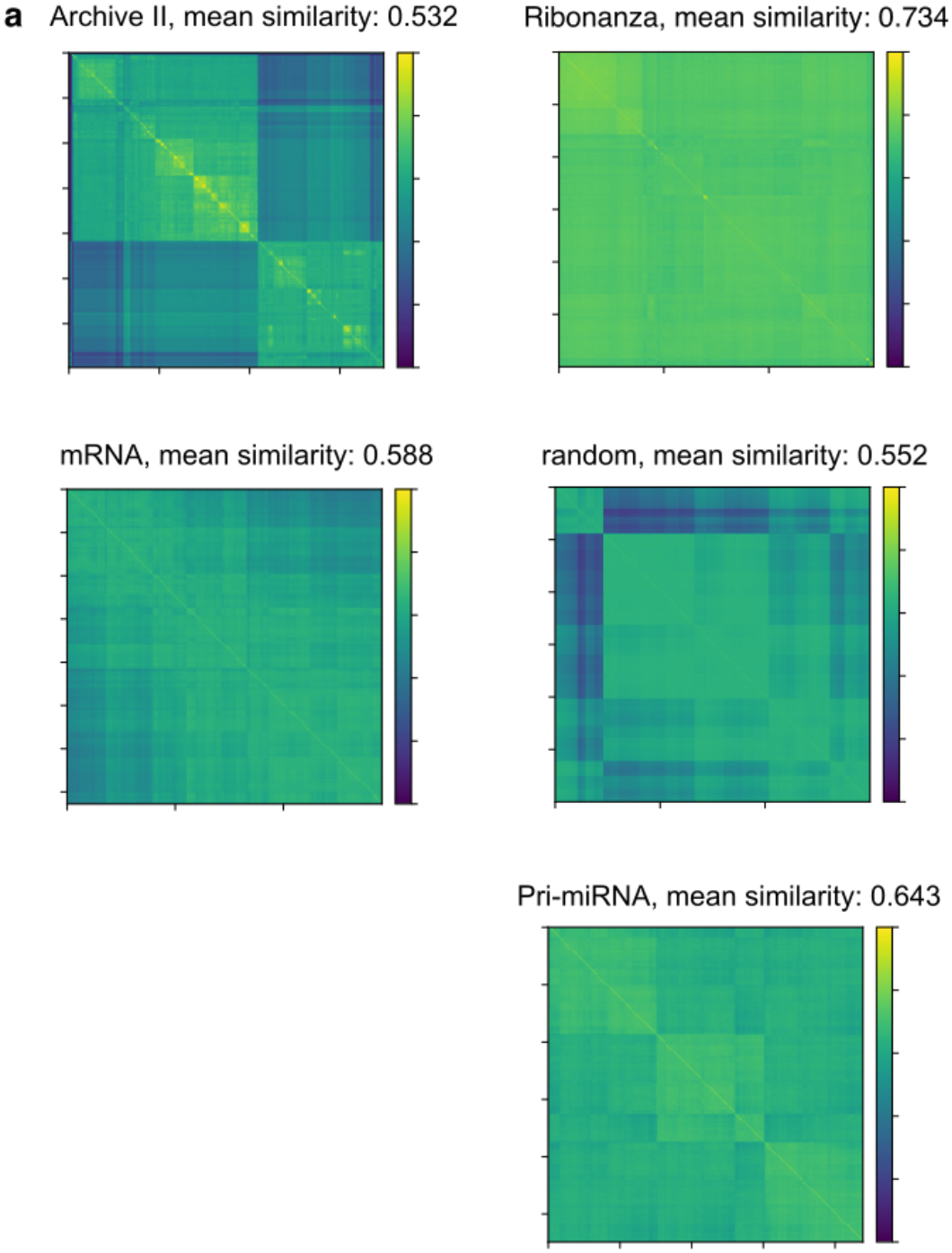
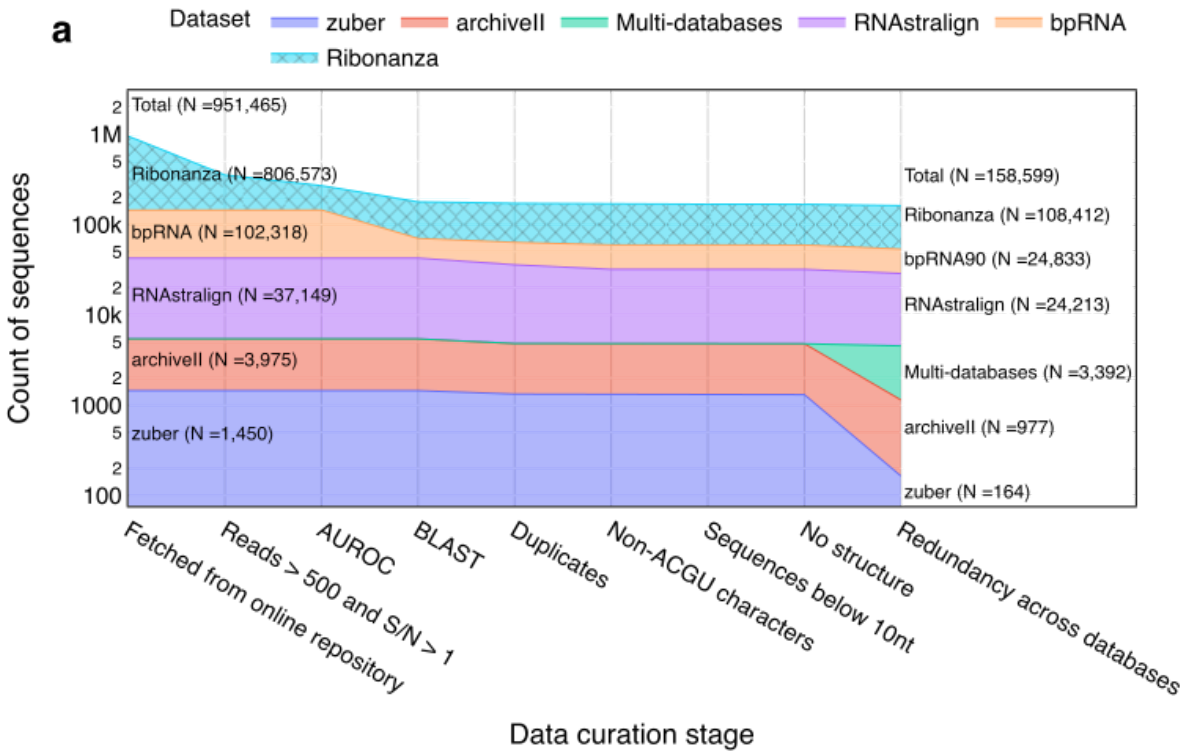
Supplementary figure 1



Supplementary figure 2



Supplementary figure 3



Supplementary figure 4