Improving RNA Secondary Structure Prediction Through Expanded Training Data

Conner J. Langeberg[*‡] Taehan Kim[§] Roma Nagle[§] Charlotte Meredith[‡]

Dimple A. Garuadapuri[¶] Jennifer A. Doudna[*‡†] Jamie H. D. Cate[*‡‡‡]

Abstract

In recent years, deep learning has revolutionized protein structure prediction, achieving remarkable speed and accuracy. RNA structure prediction, however, has lagged behind. Although several methods have shown some success in predicting RNA secondary and tertiary structures, none have reached the accuracy observed with contemporary protein models. The lack of success of these RNA structure prediction models has been proposed to be due to limited high-quality structural information that can be used as training data. To probe this proposed limitation, we developed a large and diverse dataset comprising paired RNA sequences and their corresponding secondary structures. We assess the utility of this enhanced dataset by retraining on a deep learning model, SincFold. We find that SincFold exhibited improved generalization to some previously unseen RNA families, enhancing its capability to predict accurate de novo RNA secondary structures. The RNASSTR dataset provides a substantial advance for RNA structure modeling, laying a strong foundation for the development of future RNA secondary structure prediction algorithms.

1 Introduction

Structured RNAs play essential regulatory roles across all domains of life and in viruses (1, 2), and participate in diverse regulatory processes, including transcription and translation (2, 3), catalysis (4–6), epigenetic modulation (7), and ribonucleoprotein complex function (8, 9). Similarly, many viruses make use of structured RNA motifs during infection to enhance virulence and replication efficiency (10, 11). Recent advances in mRNA vaccines have specifically leveraged RNA structural stability to enhance half-life and protein expression (12), highlighting the role of RNA structure in molecular therapeutics. The utility of RNA structure prediction represents a promising frontier for antiviral drug design (13, 14), RNA-targeting small molecules (13), CRISPR guide RNA design (15, 16), and RNA-based synthetic biology applications (17–19). However, the effective use of RNA in these areas necessitates a robust and comprehensive understanding of RNA folding and structure.

[*] Innovative Genomics Institute, University of California, Berkeley, CA, USA
[†] California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, CA, USA
[‡] Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA
[§] Department of Computer Science, University of California, Berkeley, CA, USA
[¶] Department of Bioengineering, University of California, Berkeley, CA, USA
[‖] Department of Electrical Engineering & Computer Science, University of California, Berkeley, CA, USA
[**] Howard Hughes Medical Institute, University of California, Berkeley, CA, USA
[††] Department of Chemistry, University of California, Berkeley, CA, USA
[‡‡] MBIB Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Submitted to the AI for Science workshop (NeurIPS 2025).

Structure determination methods including X-ray crystallography, cryo-electron microscopy, and nuclear magnetic resonance spectroscopy (20–25) are the only methods which allow direct experimental confirmation of an RNA's structure, including the three dimensional topology and interaction network within a fold. These methods are essential in defining the complex, non-Watson Crick Franklin (WCF) type base pairs, long-range interactions, and pseudoknots which define the higher order topology of structured RNA. However, these methods are resource-intensive and frequently fail due to technical limitations or inadequate resolution for directly modeling the RNA structure. High-throughput methods such as DMS-MaP Seq (26, 27) and SHAPE-MaP (28–32) allow for efficient scaling of experiments but sacrifice atomic-level resolution, relying heavily on expectation-maximization algorithms that may obscure non-standard structural features.

Inspired by the recent success of deep learning methods for protein structure prediction like AlphaFold2 (33) and ESMFold (34), efforts have been directed at applying similar approaches to RNA structure prediction (35). However, the limited availability of experimental RNA structures significantly hampers these data-intensive approaches, making it challenging to achieve comparable accuracy to protein predictions (36, 37). Beyond the challenges posed by the limited training data for RNA 3D structure prediction, the multiple sequence alignment algorithms used for protein structure prediction do not work well for RNA (38, 39). Protein sequence alignments leverage evolutionary conservation that can be detected in the primary structure directly, further informing structural conservation and relationship with other homologs.

RNA alignments however, are dominated by secondary structure due to the coevolution of paired bases (39, 40) which appear as compensatory mutations. While compensatory mutations conserve structure, they may result in highly degenerate sequences that are therefore difficult to align. RNA relies on base pairing to form the initial topology of the fold, necessitating secondary structure informed sequence alignments rather than relying on the primary structure. Accurate RNA secondary structure prediction is therefore an essential prerequisite to accurate 3D structure prediction as the WCF base pairs define the stems, junctions, and pseudoknots in the RNA structure around which the tertiary contacts form (41, 42).

Historically, a combination of experimental and bioinformatic methods have been used to infer RNA secondary structure (39, 43–47). However, these approaches largely require specific expertise, making them hard to disseminate and to scale. Presently-available computational algorithms that aim to minimize the requirement for user expertise while providing accurate predictions frequently fail to accurately predict large, multistem RNA folds due to their reliance on reductionist thermodynamic models which do not recapitulate the often many and non-nested stems present in highly structured RNA. For example, computational methods such as ViennaRNA (RNAfold) and MFold (48, 49), rely heavily on the Turner Rules (46, 50–54), utilizing empirically determined nearest-neighbor thermodynamic parameters to minimize the folding free energy. Although these programs proved effective for motifs with fewer stems and shorter lengths, they often perform poorly on long RNAs containing many stems and pseudoknots due to the oversimplified energetic approximations and folding assumptions. Recent machine learning methods, including UFold, SincFold, and DMfold (55–57), integrate deep neural networks to enhance the accuracy of RNA secondary structure prediction. However, these deep learning models frequently exhibit poor generalization (42), performing well only on RNA folds the models are directly trained on. This can be observed in careful validation where existing models demonstrate a significantly decreased prediction accuracy with RNA folds not used in model training (42). More recently, hybrid methods such as the MXFold suite and CDPfold (58, 59), which integrate both machine learning and thermodynamic approaches have shown promise in improving model performance. However these methods have not yet achieved accurate and general RNA secondary structure prediction (38, 42). Integrating supplementary data has been suggested as a method to improve model performance, such as the use of chemical probing or enzymatic mapping data (14, 60). The scalability of these methods enabled by high throughput sequencing and ease of data generation make integrating these orthogonal approaches a promising future direction in RNA structure prediction.

One possible solution to improve present machine learning models for RNA secondary structure prediction is to increase the size and diversity of training data. The existing training data, such as bpRNA (61) and ArchiveII (41) contain a limited set of distinct RNA folds, few sequences, and highly biased compositions which may encourage machine learning models to memorize the most abundant classes, limiting their usability with novel RNA folds and motifs. ArchiveII, for example, consists of 10 distinct RNA folds and 3847 sequences whereas the version 14.10 of the Rfam (62, 63) database recognises 4170 distinct RNA families. The bpRNA dataset provides a more structurally diverse set of sequence structure pairs than ArchiveII, with 2,588 structural families from Rfam, though only consisting of 102,318 sequences. Additionally, the data lack a standardized grammar for model training, which makes comparisons between users challenging to achieve. Finally, there is a need for robust methods to prepare training data that accounts for both sequence and structural diversity in splitting RNAs into the training and test data.

To overcome these challenges, we have developed the RNA Secondary Structure Repository (RNASSTR), a rigorously curated dataset comprising 4170 annotated RNA families described in Rfam and nearly 5 million unique sequences. These sequences span all domains of life and viruses, and were assembled leveraging robust bioinformatic workflows to identify novel RNA structural homologs. Using RNASSTR, we retrained two existing RNA secondary structure prediction models, demonstrating that increased depth and diversity of RNAs may improve model generalization, but at the cost of performance for some structural families. Our analysis also identifies additional limitations suffered by current models including slow training speed and sequence-structure memorization. The RNASSTR dataset and associated benchmarking model parameters thus provide a powerful foundational resource that can be used for further model development, and represents an important step in developing training data comparable to those now available for proteins.

2 Results

2.1 Dataset curation

In order to construct a deeper and more diverse RNA secondary structure dataset compared to those presently available, we leveraged the existing bioinformatic tool Infernal to gather and curate a set of structurally homologous RNA sequences. We first retrieved all covariation models from Rfam version 14.10 (63), which is composed of 4170 covariation models of structured RNAs (39). These models describe the sequence and secondary structure space of each unique RNA family using a stochastic context free grammar and can be used by the bioinformatic tool Infernal (47) to search for sequences which can adopt a similar secondary structure topology. We then used these covariation models to search all reference genomes available from the GTDB (version 214) (65) and NCBI RefSeq databases (resease 229) (64), from which we identified 8,910,328 putative homologous sequences. Because of the underlying false positive rate inherent in these search strategies, we chose to refine the dataset using a number of metrics to minimize the inclusion of false positive structures in the final dataset. We curated this set of data using a number of statistical metrics as described in the methods section, based on a similar analysis used for a prior dataset (66). Briefly, we removed those sequences which we determined to be outliers in terms of sequence length and structure conservation, as well as those which fell outside of the expected phylogeny as defined by Rfam. Following this curation process, we recovered 4,779,435 high confidence RNA sequence-structure pairs. At present, the resulting dataset, RNASSTR, does not include pseudoknots due to the limitation that the stochastic context free grammars used by Infernal can only handle nested stems and as such cannot be used to identify pseudoknots (47).

Further analysis of the resulting dataset revealed that the bulk of the sequences were centered around 80 nucleotides in length, composed largely of tRNAs, pre-miRNAs, and bacterial sRNAs (Fig. S4). In addition to this large population, several other notable populations exist, such as the bacterial small subunit ribosomal RNA at approximately 1600 nucleotides in length. Notably, the composition of RNASSTR is not equally distributed between the distinct RNA families. A subset of these families constitutes the bulk of the sequences, with the tRNA family accounting for 39.5% of sequences within the dataset (Fig. 1A,B). However, when compared to other frequently used RNA sequence structure training datasets RNASSTR provides a greatly increased depth across other RNA families for model training (Fig. 1C). In addition to the significant increase in sequence-structure pairs over other datasets, RNASSTR also demonstrates superior sequence diversity at a variety of fractional sequence identities, as can be seen in the top 6 most abundant RNA families from RNASSTR (Fig. 1D). Because the goal of these RNA secondary structure prediction models is to accurately predict the native fold of a given RNA, ideally generalizing to unseen structural classes, we partitioned the dataset into three parts, a training set containing 90% of sequences belonging to one-third of the Rfam families, a validation set containing 5% of the sequences belonging to one-third of the Rfam families, and a test set containing the final 5% of sequences and one-third of Rfam families. By ensuring that the families were mutually exclusive at the structural level we ensured no data leakage

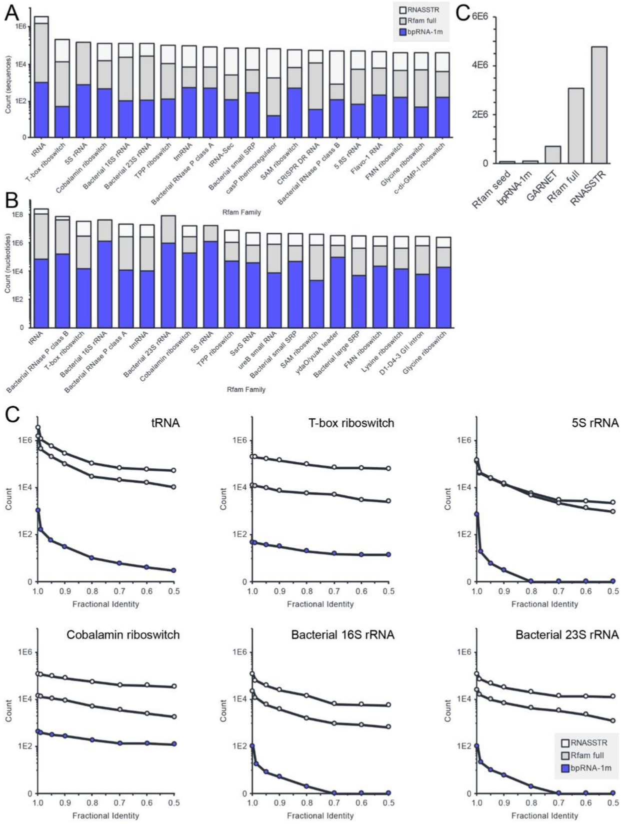


Figure 1: Depth and diversity of RNASSTR RNA sequence-structure pairs. A) Stacked histogram of the top 20 most abundant classes in RNASSTR by sequences compared to Rfam full and bpRNA-1m. B) Stacked histogram of the top 20 most abundant classes in RNASSTR by nucleotides compared to Rfam full and bpRNA-1m. C) Graphical representation of the total abundance of sequence-structure pairs in multiple RNA secondary structure datasets. D) Fractional identity of six abundant classes of RNA families comparing the sequence depth at multiple thresholds comparing RNASSTR, Rfam full, and bpRNA-1m.

between the training, validation, and testing sets, which should aid model generalization and mitigate against memorization.

2.2 Model retraining

We next tested whether using RNASSTR to retrain existing machine learning algorithms would improve their performance and generalization. To do this we identified a previously published 2D RNA prediction model, SincFold (56), which we could retrain from scratch. We used the structure-stratified dataset split within RNASSTR to prevent data leakage between training and testing (Methods). We trained SincFold until the validation F1 score converged, to a training F1 of 0.983 and a validation F1 of 0.420, which took 15 epochs. Notably, one full training and validation cycle for SincFold required 64 GPU hours per training epoch and 2 hours per validation iteration.

2.3 Model performance

Following model training, we used the retrained SincFold model as well as its default published parameter sets to predict the secondary structures of the RNASSTR held-out testing data. We measured the performance using two standard metrics for the field, F1 and MCC scores, both measures of confusion matrix categories (Methods). We computed these metrics for the published model parameters and for our retrained parameter sets to assess how data scaling during training affected the model's ability to generalize to unseen RNA families and folds. To compare against

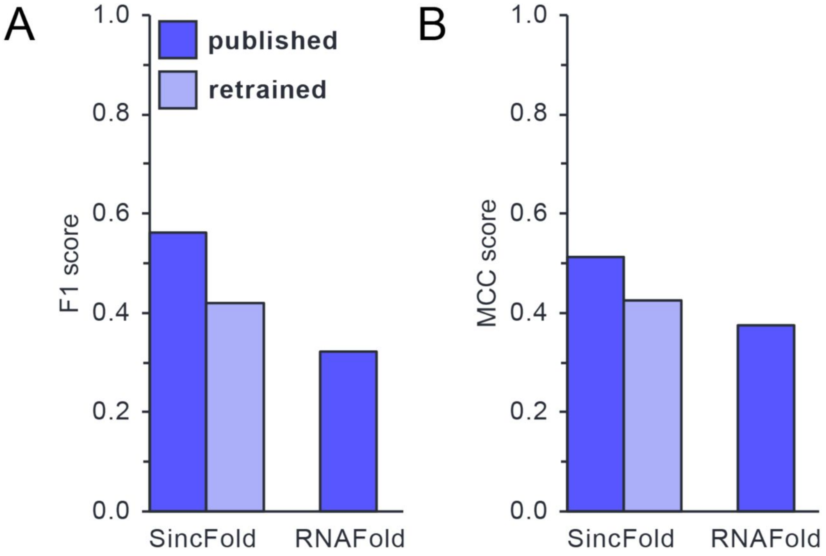


Figure 2: Model performance pre and post retraining. A-B) Model performance on RNASSTR test partition using the published model and RNASSTR retrained model for SincFold. RNAfold is included as a minimum free energy comparison. Scores are calculated for F1 (A) and MCC (B).

Table 1: Model performance. Shown are the RNASSTR test partition for published and RNASSTR models as well as minimum free energy model RNAFold.

	SincFold		MXFold2		MFold
Metric	Published	RNASSTR retrained	Published	RNASSTR retrained	
Test F1	0.561	0.420	0.571	0.171	0.321
Test MCC	0.513	0.424	0.554	0.169	0.691

non-ML based methods, we included predictions from a popular minimum free energy program, RNAfold (49). For RNAfold, we subsampled 100 sequences per family in the testing set due to the computational complexity of computing all structures.

For SincFold, the retrained model performed worse than the published model on the RNASSTR testing set. This is striking, given that we made no changes to the underlying model architectures or hyperparameters. The retrained SincFold model did, globally, outperform RNAfold, the minimum free energy model (Fig. 2A,B, Table 1). To better understand if specific features of the dataset impacted the retrained SincFold model performance, we assessed a range of sequence features: absolute number of ground truth paired bases, GC content, ground truth fraction of paired bases, and sequence length. While no clear trend is visible in the data, we noted many sequences with a F1 score of 0 indicating the retrained SincFold model failed to predict any true positives (Fig. S6), defined as correctly predicted base pairs. In these cases, the retrained model appears to catastrophically fail during the inference stage.

We then assessed per-family F1 scores to determine if specific RNA structures were more prone to poor prediction accuracy. We observe that the distribution of F1 scores for the top 20 families did not change significantly across epochs (Fig. S7). However, in later epochs the retrained SincFold model showed improvement in performance for some Rfam groups, RF01359 and RF00730 as examples, suggesting that the retrained SincFold model learned to better predict some Rfam groups beyond early epochs. While this was true for a subset of classes, most Rfam groups are already learned at early epochs.

When comparing the published SincFold model to the RNASSTR retrained model, we observed a large variance in per family performance (Fig. 3A,B), suggesting these two models learned different features resulting in varying results. While the published model parameters yielded a marginally higher average F1 score across the full test set, this improvement was spread thinly across many RNA

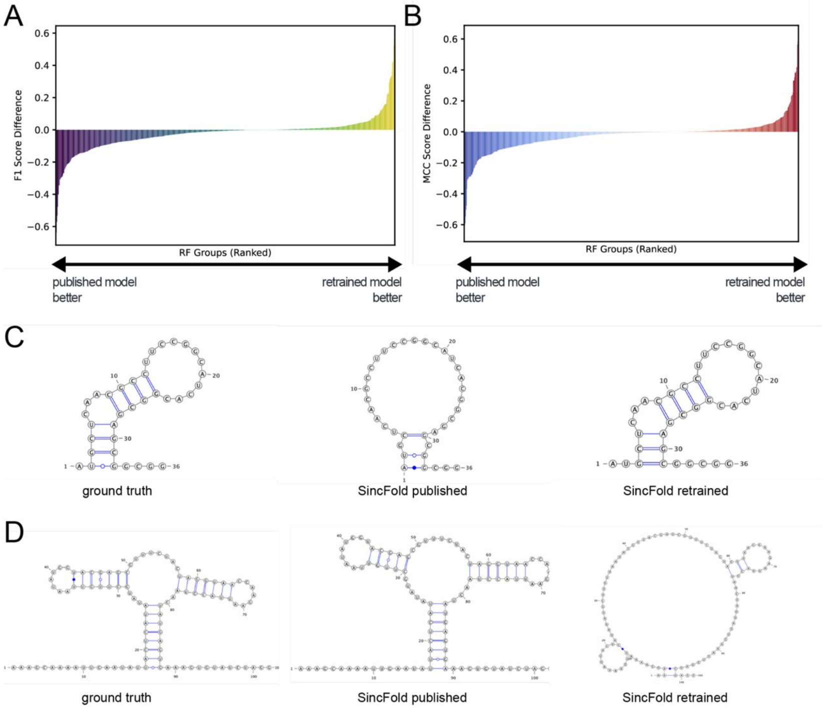


Figure 3: Differential performance across Rfam families by retrained SincFold. A) Rank order plot of per family average F1 score difference between the published SincFold model and the RNASSTR retrained SincFold model. Left-shifted families perform better with the published model and right-shifted families perform better with the RNASSTR retrieved model. B) Rank order plot of per family average MCC scores differences between published SincFold model and the RNASSTR retrained SincFold model. C) Representative RNA secondary structure of a family, RF01336 CRISPR RNA direct repeat, where the RNASSTR retrained model performed better than the published model. D) Representative RNA secondary structure of a family, RF00167 purine riboswitch, where the RNASSTR retrained model performed worse than the published model.

families. In contrast, the RNASSTR retrained model showed more substantial gains within a smaller subset of families, suggesting that although it performs less consistently overall, it captures specific structural features more effectively. This implies that while the retrained model does not perform as well broadly, within specific classes the retrained model performs better. To exemplify this we show two representative sequences, showing the ground truth structure, the published SincFold prediction, and the RNASSTR retrained SincFold prediction (Fig. 3C,D). In one case where the retrained model performed better, it is able to recover all canonical WCF base pairs in a CRISPR RNA direct repeat (Fig. 3C), only missing a single U•G pair. However, the published SincFold model does not recover any pairs in the ground truth structure and proposes a non-canonical A•G pair. In a counterexample of a purine riboswitch, the retrained model fails to correctly predict any ground truth base pairs while the published model recovers all pairs except a non-canonical U•U pair (Fig. 3D). Taken together, these results highlight how the size of the training data and method for partitioning the sequences into rigorously separated training, validation, and test sets dramatically change how a given model architecture performs.

3 Discussion

3.1 Current Benchmark Datasets

Within the field of RNA secondary structure prediction, there currently exist three widely-used RNA secondary structure benchmark datasets: RNAStrAlign, ArchiveII, and bpRNA (41, 61, 71). Each of these has distinct characteristics in terms of size, sequence length distribution, and RNA family composition, and each has been utilized to various levels of success in training RNA ML models to predict RNA secondary structures.

RNAStrAlign (71) represents an alignment-based dataset aggregating known RNA secondary structures from 8 diverse RNA families (5S rRNA, tRNA, group I introns, 16S rRNA, tmRNA, SRP RNA, RNase P, and telomerase RNA) containing 37,149 sequence-structure pairs with lengths ranging from approximately 30 nucleotides to 1,851 nucleotides. While this represents a robust grouping of structurally diverse RNAs, this dataset only contains representative sequences from 8 structural families, limiting its utility in training general RNA structure models. Similarly, approximately 50% of the sequences in this dataset belong to 5S rRNA family, which limits the structural diversity despite the size of the dataset.

ArchiveII is a highly curated collection of 2,975 RNA sequences from 10 distinct RNA families (41) including several rRNAs as well as other common classes such as tRNA, RNase P, tmRNA, and self-splicing introns. While this dataset contains many fewer sequences than either RNAStrAlign (71) or bpRNA (61), it represents one of the older datasets for training and was originally compiled to provide a high-quality test set for RNA folding algorithms. Additionally, its curation ensures that each example is biologically relevant and non-redundant. For instance, ArchiveII contains representative rRNA sequences from different organisms rather than many near-duplicates. Because of its inclusion of large structured RNAs, ArchiveII is used as a stringent benchmark and has often been used as a hold-out test set in prior studies (i.e., models are sometimes trained on RNAStrAlign and evaluated on ArchiveII) (72, 73).

bpRNA represents the largest of the three datasets and the de facto standard for the field (61). This larger meta-database includes 102,318 RNA secondary structures, drawing from multiple sources, the largest being Rfam which contributes approximately half of the sequences and over 2,000 unique structural families. The scale of bpRNA makes it a popular choice for training deep learning models, but it carries an uneven family distribution. A few RNA families (tRNAs, 5S rRNAs) make up more than half of the sequences, while many other families are sparsely represented. To specifically test generalization to novel folds, an updated bpRNA-new dataset was introduced in Sato et al. (58) based on Rfam 14.2 (63). The bpRNA-new set contains sequences from approximately 1,500 new RNA families that were not present in the original bpRNA-1m compilation. By design, none of these families overlap with prior training sets, making bpRNA-new a benchmark for testing the cross-family generalization of current RNA secondary structure prediction models.

3.2 The RNASSTR dataset

We mined three published RNA sequence databases to comprehensively guarantee diverse RNA fold representation in a new curated RNA secondary structure dataset we term RNASSTR. These databases, GTDB, NCBI RefSeq, and Rfam (63–65), represent an encompassing representation of our current understanding of biological sequence space. Using only reference genomes present in these data to limit the overrepresentation of model organisms, we identified 4,779,435 RNA sequences homologous to known RNA families defined in the Rfam database (Fig. 1). These sequences span all domains of life as well as viruses and represent a significantly larger sequence space than that queried in the three published datasets. Because previous datasets were limited in the folds they contained, we ensured that RNASSTR provides coverage of all RNA families present in Rfam, consisting of 4,028 unique folds representing all known RNA structural families. A notable feature of RNASSTR, much like bpRNA, is the overrepresentation of specific RNA families. As mentioned above, tRNAs account for 39.5% of the total sequences. However because of their short length this only accounts for 13.7% of the total nucleotides. Given training is performed on single nucleotide tokens we expect this overrepresentation to be mitigated, with the top 10 families by number of sequences making up a relatively more equal distribution of training families. In the future, it will be interesting to test alternative training schemes in which overrepresented families by number of sequences are subsampled, so family bias is less prevalent.

With RNASSTR, we are able to provide significantly more sequence depth and diversity compared to other existing RNA secondary structure datasets. A general trend in machine learning is improved model performance with data scaling (74). Because previous RNA secondary structure prediction models had been trained primarily with smaller and less diverse datasets, we hypothesized the lack of training data reduced their predictive power. However, upon retraining SincFoldusing RNASSTR we were not able to increase the average accuracy above that seen with the published model, suggesting there may be alternative aspects beyond the size of the training data driving lower model performance. However, we did observe a subset of families performed better with RNASSTR retrained models as compared to published models (Fig. 2,4). This suggests that the two versions of SincFold, the published version and the retrained version presented here, have learned different features in the data resulting in differential model performance. At this point, we have not been able to determine any specific features which may drive this difference (Fig. S6). Similarly, the performance differences between the SincFold models and RNAfold appear distinct, particularly for sequences where predictions failed to recover any correct base pairs, though further analysis of specific instances or features which allow thermodynamics-based models to outperform ML models, or vice versa represent an open question in the field.

One unsolved problem for both the published model parameters and the retrained model is data memorization. In deep learning, models can overfit the training data resulting in a pathology termed "memorization" (75), where instead of learning dataset features which allow the model to generalize, they instead memorize the sequence and secondary structures present in the training data. While we cannot directly comment on memorization in the published models, their large difference in performance on families on which they have been trained compared to novel families suggests model overfitting (42). Similarly, when retraining SincFold, the large discrepancy between the training and testing F1 and MCC scores (Table 1) also suggests this same pathology. One possible explanation for the difference in performance of the same model architecture trained on previous data versus RNASSTR could be the approach to splitting the data into training, test, and validation sets. In the case of secondary structure prediction, the approach used for bpRNA-1m, which performed data splitting using sequence identity rather than structural overlap, may artificially inflate F1 values in test sets. In other words, the RNASSTR test set used here to assess the published SincFold model may have some representation in the bpRNA-1m training data used to originally train SincFold.

To prevent a similar issue with RNASSTR, we performed a rigorous structure-based partitioning of the data into three sets, a training, a validation, and a testing set. By using the Rfam defined structural grammars, we ensured that these splits retained families which were mutually identifiable and remained in the same group. While the RNASSTR-retrained SincFold model did not outperform the published model parameters, the discrepancy between the training and testing F1 and MCC scores may represent a more accurate estimation of model generality compared to previous benchmarks. While the current RNA secondary structure prediction architectures we tested were unable to provide general RNA secondary structure prediction, we anticipate future models that make use of RNASSTR will be able to overcome this limitation and provide true generalization for RNA secondary structure prediction.

3.3 Scaling Issues in current RNA secondary structure prediction models

Increased dataset size requires more efficient and faster training to make training models viable. Because RNASSTR is nearly 50x times larger than the most commonly used dataset, bpRNA-1m, the computational burden of training a single epoch using RNASSTR becomes much larger than with other datasets. The computational cost was highly limiting in the case of another model, MXFold2 (58) where each epoch required 268 GPU hours using a NVIDIA RTX A4500 GPU, and inference required 110 GPU hours on the same device. Training times of this length are infeasible, especially for academic groups with limited computational resources. As such we only completed 3 rounds of MXFold2 training before it became too resource-intensive to continue and we chose to devote more resources to the other model, SincFold. While SincFold was significantly quicker to train than MXFold2, each training epoch still required 64 GPU hours on a NVIDIA RTX A4500 GPU and 2 GPU hours for inference. A key aspect of improving future model architectures for RNA secondary structure prediction tasks would be to prioritize efficiency to allow for more throughput with the increasing sizes of training datasets. Models which enable parallel training using multiple GPUs should alleviate some of the burden of these large training sets where compute is not limiting.

RNASSTR provides a new step towards a standardized community wide benchmark for RNA secondary structure prediction, including rigorous methods for structure-based data splitting. However, it remains unclear whether sequence and secondary structure pairs will be sufficient to overcome the challenge of model generalization for RNA secondary structure prediction. For example, annotation of the full set of noncanonical base pairs in the secondary structure representations may be required. Furthermore, RNASSTR presently does not include information on pseudoknots, a key feature of many RNA structures that has been difficult to capture in training data and secondary structure prediction models. To date, since most secondary structures are determined computationally, it is not clear how to accomplish this extension. Alternatively, the incorporation of experimental data may improve the prediction accuracy of RNA secondary structure prediction models. RNA chemical probing data is scalable and high-throughout, with initial attempts to integrate it into RNA structure prediction pipelines already showing promise (51, 52).

Acknowledgments and Disclosure of Funding

Funding

This work was supported by 5R35GM148352 to J.H.C. J.A.D. is an investigator of the Howard Hughes Medical Institute, and research in the Doudna lab is supported by the Howard Hughes Medical Institute (HHMI), NIH/NIAID (U19AI171110, U54AI170792, U19AI135990, UH3AI150552, U01AI142817), NIH/NINDS (U19NS132303), NSF (2334028), DOE (DE-AC02-05CH11231, 2553571, B656358), Lawrence Livermore National Laboratory, Apple Tree Partners (24180), UCB-Hampton University Summer Program, Mr. Li Ka Shing, Emerson Collective and the Innovative Genomics Institute (IGI).

Conflict of Interest Disclosure

J.H.C. is founder, board and SAB member of Initial Therapeutics. The Regents of the University of California have patents issued and pending for CRISPR technologies on which J.A.D. is an inventor. J.A.D. is a cofounder of Azalea Theratupics, Caribou Biosciences, Editas Medicine, Evercrisp, Scribe Therapeutics, Intellia Therapeutics, and Mammoth Biosciences. J.A.D. is a scientific advisory board member at Evercrisp, Caribou Biosciences, Intellia Therapeutics, Scribe Therapeutics, Mammoth Biosciences, The Column Group and Inari. J.A.D. is Chief Science Advisor to Sixth Street, a Director at Johnson & Johnson, Altos and Tempus, and has a research project sponsored by Apple Tree Partners. The remaining authors declare no competing interests.

Acknowledgments

The authors would like to thank members of the Cate and Doudna Labs for their helpful discussions and insightful comments. The dataset filtering was inspired by similar analyses from Yekaterina Shulgina and would not have been possible without her suggestions.

References

[1] Cech, T.R. (2012) The RNA worlds in context. Cold Spring Harb. Perspect. Biol., 4, a006742.

[2] Serganov, A. and Nudler, E. (2013) A decade of riboswitches. Cell, 152, 17–24.

[3] Breaker, R.R. (2012) Riboswitches and the RNA world. Cold Spring Harb. Perspect. Biol., 4, a003566.

[4] Breaker, R.R. (1997) In vitro selection of catalytic polynucleotides. Chem. Rev., 97, 371–390.

[5] Doudna, J.A. and Cech, T.R. (2002) The chemical repertoire of natural ribozymes. Nature, 418, 222–228.

[6] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell, 35, 849–857.

[7] Rinn, J.L. and Chang, H.Y. (2012) Genome regulation by long noncoding RNAs. Annu. Rev. Biochem., 81, 145–166.

[8] Bothe, J.R., Nikolova, E.N., Eichhorn, C.D., Chugh, J., Hansen, A.L. and Al-Hashimi, H.M. (2011) Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. Nat. Methods, 8, 919–931.

[9] Nissen, P., Hansen, J., Ban, N., Moore, P.B. and Steitz, T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. Science, 289, 920–930.

[10] Jaafar, Z.A. and Kieft, J.S. (2019) Viral RNA structure-based strategies to manipulate translation. Nat. Rev. Microbiol., 17, 110–123.

[11] Boerneke, M.A., Ehrhardt, J.E. and Weeks, K.M. (2019) Physical and functional analysis of viral RNA genomes by SHAPE. Annu. Rev. Virol., 6, 93–117.

[12] Pardi, N., Hogan, M.J., Porter, F.W. and Weissman, D. (2018) mRNA vaccines - a new era in vaccinology. Nat. Rev. Drug Discov., 17, 261–279.

[13] Childs-Disney, J.L., Yang, X., Gibaut, Q.M.R., Tong, Y., Batey, R.T. and Disney, M.D. (2022) Targeting RNA structures with small molecules. Nat. Rev. Drug Discov., 21, 736–762.

[14] Boyd, N., Anderson, B.M., Townshend, B., Chow, R., Stephens, C.J., Rangan, R., Kaplan, M., Corley, M., Tambe, A., Ido, Y., et al. (2023) ATOM-1: A foundation model for RNA structure and function built on chemical mapping data. bioRxiv, doi:10.1101/2023.12.13.571579.

[15] Lee, K., Mackley, V.A., Rao, A., Chong, A.T., Dewitt, M.A., Corn, J.E. and Murthy, N. (2017) Synthetically modified guide RNA and donor DNA are a versatile platform for CRISPR-Cas9 engineering. Elife, 6.

[16] Liu, Y., Liu, W. and Wang, B. (2023) Engineering CRISPR guide RNAs for programmable RNA sensors. Biochem. Soc. Trans., 51, 2061–2070.

[17] Isaacs, F.J., Dwyer, D.J. and Collins, J.J. (2006) RNA synthetic biology. Nat. Biotechnol., 24, 545–554.

[18] Pfeifer, B.A., Beitelshees, M., Hill, A., Bassett, J. and Jones, C.H. (2023) Harnessing synthetic biology for advancing RNA therapeutics and vaccine design. NPJ Syst. Biol. Appl., 9, 60.

[19] Suess, B. (2024) Synthetic RNA biology. RNA Biol., 21, 1–2.

[20] Robertus, J.D., Ladner, J.E., Finch, J.T., Rhodes, D., Brown, R.S., Clark, B.F. and Klug, A. (1974) Structure of yeast phenylalanine tRNA at 3 Å resolution. Nature, 250, 546–551.

[21] Yan, C., Hang, J., Wan, R., Huang, M., Wong, C.C.L. and Shi, Y. (2015) Structure of a yeast spliceosome at 3.6-angstrom resolution. Science, 349, 1182–1191.

[22] Nozinovic, S., Fürtig, B., Jonker, H.R.A., Richter, C. and Schwalbe, H. (2010) High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. Nucleic Acids Res., 38, 683–694.

[23] Cate, J.H., Gooding, A.R., Podell, E., Zhou, K., Golden, B.L., Kundrot, C.E., Cech, T.R. and Doudna, J.A. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. Science, 273, 1678–1685.

[24] Kappel, K., Zhang, K., Su, Z., Watkins, A.M., Kladwang, W., Li, S., Pintilie, G., Topkar, V.V., Rangan, R., Zheludev, I.N., et al. (2020) Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures. Nat. Methods, 17, 699–707.

[25] Langeberg, C.J. and Kieft, J.S. (2023) A generalizable scaffold-based approach for structure determination of RNAs by cryo-EM. Nucleic Acids Res., 51, e100.

[26] Tomezsko, P., Swaminathan, H. and Rouskin, S. (2021) DMS-MaPseq for genome-wide or targeted RNA structure probing in vitro and in vivo. Methods Mol. Biol., 2254, 219–238.

[27] Zubradt, M., Gupta, P., Persad, S., Lambowitz, A.M., Weissman, J.S. and Rouskin, S. (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat. Methods, 14, 75–82.

[28] Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E. and Weeks, K.M. (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods, 11, 959–965.

[29] Mustoe, A.M., Lama, N.N., Irving, P.S., Olson, S.W. and Weeks, K.M. (2019) RNA base-pairing complexity in living cells visualized by correlated chemical probing. Proc. Natl. Acad. Sci. U. S. A., 116, 24574–24582.

[30] Mortimer, S.A. and Weeks, K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. J. Am. Chem. Soc., 129, 4144– 4145.

[31] Deigan, K.E., Li, T.W., Mathews, D.H. and Weeks, K.M. (2009) Accurate SHAPE-directed RNA structure determination. Proc. Natl. Acad. Sci. U. S. A., 106, 97–102.

[32] Wilkinson, K.A., Merino, E.J. and Weeks, K.M. (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat. Protoc., 1, 1610–1616.

[33] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

[34] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379, 1123–1130.

[35] Shen, T., Hu, Z., Sun, S., Liu, D., Wong, F., Wang, J., Chen, J., Wang, Y., Hong, L., Xiao, J., et al. (2024) Accurate RNA 3D structure prediction using a language model-based deep learning approach. Nat. Methods, 21, 2287–2298.

[36] Kretsch, R.C., Andersen, E.S., Bujnicki, J.M., Chiu, W., Das, R., Luo, B., Masquida, B., McRae, E.K.S., Schroeder, G.M., Su, Z., et al. (2023) RNA target highlights in CASP15: Evaluation of predicted models by structure providers. Proteins, 91, 1600–1615.

[37] Kretsch, R.C., Albrecht, R., Andersen, E.S., Chen, H.-A., Chiu, W., Das, R., Gezelle, J.G., Hartmann, M.D., Höbartner, C., Hu, Y., et al. (2025) Functional relevance of CASP16 nucleic acid predictions as evaluated by structure providers. bioRxiv, https://doi.org/10.1101/2025.04.15.649049.

[38] Schneider, B., Sweeney, B.A., Bateman, A., Cerny, J., Zok, T. and Szachniuk, M. (2023) When will RNA get its AlphaFold moment? Nucleic Acids Res., 51, 9522–9532.

[39] Eddy, S.R. (2014) Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annu. Rev. Biophys., 43, 433–456.

[40] Petrov, A.I., Zirbel, C.L. and Leontis, N.B. (2013) Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA, 19, 1327–1340.

[41] Sloma, M.F. and Mathews, D.H. (2016) Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. RNA, 22, 1808–1818.

[42] Szikszai, M., Wise, M., Datta, A., Ward, M. and Mathews, D.H. (2022) Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. Bioinformatics, 38, 3892–3899.

[43] Lorenz, R., Wolfinger, M.T., Tanzer, A. and Hofacker, I.L. (2016) Predicting RNA secondary structures from sequence and probing data. Methods, 103, 86–98.

[44] Tieng, F.Y.F., Abdullah-Zawawi, M.-R., Md Shahri, N.A.A., Mohamed-Hussein, Z.-A., Lee, L.-H. and Mutalib, N.-S.A. (2023) A Hitchhiker's guide to RNA-RNA structure and interaction prediction tools. Brief. Bioinform., 25.

[45] Zhao, Q., Zhao, Z., Fan, X., Yuan, Z., Mao, Q. and Yao, Y. (2021) Review of machine learning methods for RNA secondary structure prediction. PLoS Comput. Biol., 17, e1009291.

[46] Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M. and Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. U. S. A., 101, 7287–7292.

[47] Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 29, 2933–2935.

[48] Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. Algorithms Mol. Biol., 6, 26.

[49] Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31, 3406–3415.

[50] Mathews, D.H., Burkard, M.E., Freier, S.M., Wyatt, J.R. and Turner, D.H. (1999) Predicting oligonucleotide affinity to nucleic acid targets. RNA, 5, 1458–1469.

[51] Brion, P. and Westhof, E. (1997) Hierarchy and dynamics of RNA folding. Annu. Rev. Biophys. Biomol. Struct., 26, 113–137.

[52] Turner, D.H. and Mathews, D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res., 38, D280–2.

[53] Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol., 288, 911–940.

[54] Tinoco, I.,Jr and Bustamante, C. (1999) How RNA folds. J. Mol. Biol., 293, 271–281.

[55] Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q. and Xie, X. (2022) UFold: fast and accurate RNA secondary structure prediction with deep learning. Nucleic Acids Res., 50, e14.

[56] Bugnon, L.A., Di Persia, L., Gerard, M., Raad, J., Prochetto, S., Fenoy, E., Chorostecki, U., Ariel, F., Stegmayer, G. and Milone, D.H. (2024) sincFold: end-to-end learning of short- and long-range interactions in RNA secondary structure. Brief. Bioinform., 25.

[57] Wang, L., Liu, Y., Zhong, X., Liu, H., Lu, C., Li, C. and Zhang, H. (2019) DMfold: A novel method to predict RNA secondary structure with pseudoknots based on Deep Learning and Improved Base Pair Maximization Principle. Front. Genet., 10, 143.

[58] Sato, K., Akiyama, M. and Sakakibara, Y. (2021) RNA secondary structure prediction using deep learning with thermodynamic integration. Nat. Commun., 12, 941.

[59] Zhang, H., Zhang, C., Li, Z., Li, C., Wei, X., Zhang, B. and Liu, Y. (2019) A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. Front. Genet., 10, 467.

[60] Rouskin, S., de Lajart, A., des Taillades, Y.M., Kalicki, C., Wightman, F.F., Aruda, J., Salazar, D., Allan, M., L'Esperance-Kerckhoff, C., Kashi, A., et al. (2024) Diverse database and machine learning model to narrow the generalization gap in RNA structure prediction. Research Square, doi:10.21203/rs.3.rs-4159627/v1.

[61] Danaee, P., Rouches, M., Wiley, M., Deng, D., Huang, L. and Hendrix, D. (2018) bpRNA: large-scale automated annotation and analysis of RNA secondary structure. Nucleic Acids Res., 46, 5381–5394.

[62] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Rfam: an RNA family database. Nucleic Acids Res., 31, 439–441.

[63] Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., et al. (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res., 49, D192–D200.

[64] Goldfarb, T., Kodali, V.K., Pujar, S., Brover, V., Robbertse, B., Farrell, C.M., Oh, D.-H., Astashyn, A., Ermolaeva, O., Haddad, D., et al. (2025) NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Res., 53, D243– D257.

[65] Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A. and Hugenholtz, P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res., 50, D785–D794.

[66] Shulgina, Y., Trinidad, M.I., Langeberg, C.J., Nisonoff, H., Chithrananda, S., Skopintsev, P., Nissley, A.J., Patel, J., Boger, R.S., Shi, H., et al. (2024) RNA language models predict mutations that improve RNA function. Nat. Commun., 15, 10627.

[67] Lipman, D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity searches. Science, 227, 1435–1441.

[68] Chen, C.-C. and Chan, Y.-M. (2023) REDfold: accurate RNA secondary structure prediction using residual encoder-decoder network. BMC Bioinformatics, 24, 122.

[69] Singh, J., Hanson, J., Paliwal, K. and Zhou, Y. (2019) RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat. Commun., 10, 5407.

[70] Chen, X., Li, Y., Umarov, R., Gao, X. and Song, L. (2020) RNA secondary structure prediction by learning unrolled algorithms. arXiv [cs.LG].

[71] Tan, Z., Fu, Y., Sharma, G. and Mathews, D.H. (2017) TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. Nucleic Acids Res., 45, 11570–11581.

[72] Zhao, Q., Mao, Q., Zhao, Z., Yuan, W., He, Q., Sun, Q., Yao, Y. and Fan, X. (2023) RNA independent fragment partition method based on deep learning for RNA secondary structure prediction. Sci. Rep., 13, 2861.

[73] Wang, Z., Feng, Y., Tian, Q., Liu, Z., Yan, P. and Li, X. (2024) RNADiffFold: generative RNA secondary structure prediction using discrete diffusion models. Brief. Bioinform., 26.

[74] Ahsan, M., Mahmud, M., Saha, P., Gupta, K. and Siddique, Z. (2021) Effect of data scaling methods on machine learning algorithms and model performance. Technologies (Basel), 9, 52.

[75] Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. (2017) A closer look at memorization in deep networks. arXiv [stat.ML].

A Technical Appendices and Supplementary Material

A.1 Materials and Methods

A.1.1 Dataset Generation

To generate sequence-structure pairs we leveraged the bioinformatic tool Infernal v1.1.5 and the RNA Family database (Rfam) v14.10 (47, 63). Using the 4170 covariance models of RNA structure families deposited on Rfam, we searched against a database containing all eukaryotic and viral reference genomes from NCBI, release 229 (64), as well as all bacterial and archaeal reference genomes from the Genome Taxonomy Database, release 214 (65). The resulting sequences were realigned to their respective Rfam families using Infernal, thus inferring secondary structure from the consensus model. The resulting hits were then filtered by several features. First, only sequences with a reported E-value of 0.01 or less were considered, an approximate false positive rate of 1 out of 100 or better. Sequences more than 2 standard deviations from the Rfam defined length were removed, in line with previously reported filtering thresholds (66). Similarly, sequences with more than 2 standard deviations below Rfam defined consensus base pairs were removed as well as those sequences with more than 2 standard deviations fewer WCF base pairs than defined in the Rfam seed alignment were removed. Sequences which were identified outside of their Rfam defined phylogeny were removed. Finally, overlapping hits were assessed and only the better E-value sequence was retained.

In order to facilitate rigorous training and prevent data leakage we performed data splitting accounting for secondary structure. To do this we defined mutually exclusive RNA structural groups using Infernal cross-validation, identifying RNA families incapable of cross-identification, thus preventing structural data leakage. This was accomplished by searching all sequences of one Rfam family against all other Rfam models. If any sequences were identified as having statistically significant similarity, an E-value less than or equal to 0.01, they were considered not mutually exclusive and were placed into one of the data splits. Those families which did not have identifiable structural homology were placed in different data splits. This resulted in a partitioning scheme containing approximately 90% training, 5% validation, and 5% test sequences, with exact counts provided in the supplementary materials. For model training purposes, sequence-structure pairs were converted into several formats using a custom python script: standard FASTA (67) and dot-bracket notation, BPSEQ format (61), and an expanded BPSEQ format specific to the SincFold (56) we here call SincFold format.

A.1.2 SincFold retraining

Model retraining was performed on a single NVIDIA RTX A4500 GPU. Default parameters were used for model training from a random seed initialization, as specified in the initial publication (56). This ensured models were trained from scratch without any prior knowledge. Training was monitored using the calculated F1 score of the train and validation split at the end of each epoch of training to externally monitor model progress. Training was allowed to progress until the validation F1 score plateaued for multiple epochs, in this case training required 15 epochs to stabilize as we observed early convergence. The final trained model was assessed using the calculated F1 score of the train and test split.

A.1.3 MFold secondary structure calculation

MFold v3.6 (49) was used to calculate the minimum free energy structures of a subset of each RNA family to compare against the ML models. A subset of 100 sequences was randomly selected from each RNA family and subjected to folding using the default parameters in mFold. For those RNA families with less than 100 members all sequences were used.

A.1.4 F1 and MCC score calculation

Both F1 and Matthews Correlation Coefficient (MCC) scores were calculated using a custom script which allowed us to analyze all predicted secondary structures independently. In order to ensure a rigorous calculation, base pair partners were enumerated using the same strategy the BPSEQ format uses. From these enumerated pairing schemes, both the F1 and MCC scores were calculated. A true positive is defined as a nucleotide predicted to be involved in a base pair with the correct pairing partner, a false positive is defined as a nucleotide predicted to be involved in a base pair but with the incorrect pairing partner, and a false negative is the number of base pairs in the ground truth structure not predicted. True negatives are ignored for this adaptation of MCC as is standard practice in RNA 2D structure quantification.

True Positive (TP): Predicted base pair is in the true structure.

False Positive (FP): Predicted base pair is not in the true structure.

False Negative (FN): True base pair was not predicted.

$$F_1 = \frac{TP}{TP + {}^{1}\!/_{2}(FP + FN)} \tag{1}$$

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

A.1.5 Models which could not be retrained

In addition to retraining both SincFold, we attempted to retrain a number of other ML models. However, these were unable to be retrained for various reasons ranging from missing training scripts to bugs in the deposited code. The following is an overview of those models which we attempted to retrain and why we were unable to do so.

UFold (55): During model training, an error arose stating the input dimension did not match the shape of the data preventing the model from being retrained. This error could not be resolved. GitHub link: `https://github.com/uci-cbcl/UFold`

REDfold (68): Script typos prevented the retraining script from functioning. GitHub link: `https://github.com/aky3100/REDfold`

SPOT-RNA (69): Model lacked a retraining script. GitHub link: `https://github.com/jaswindersingh2/SPOT-RNA`

E2Efold (70): Unresolvable munch incompatibility prevented model retraining. GitHub link: `https://github.com/ml4bio/e2efold`

MXFold2 (58): Training times were prohibitively long, approaching 270 hours per epoch. GitHub link: `https://github.com/mxfold/mxfold2`

A.2 Supplementary Figures

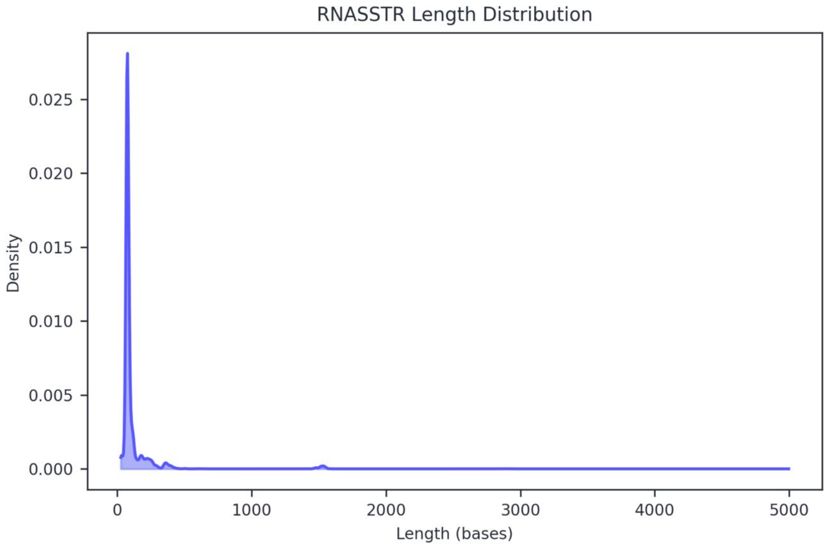


Figure 4: RNASSTR dataset sequence length distribution density plot. The peak at approximately 1550 nucleotides corresponds to bacterial small subunit rRNA. Longer sequences correspond to primarily bacterial and archeal large subunit rRNA.

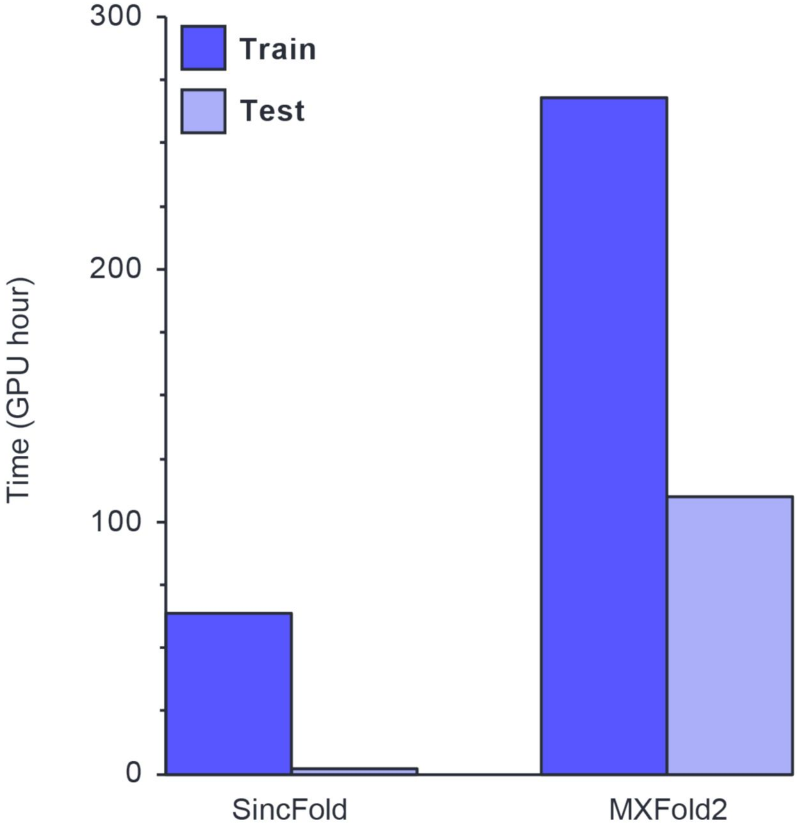


Figure 5: Model retraining times. SincFold (left) and MXFold (right) training and testing times per epoch in GPU hours using a single NVIDIA RTX A4500 GPU.

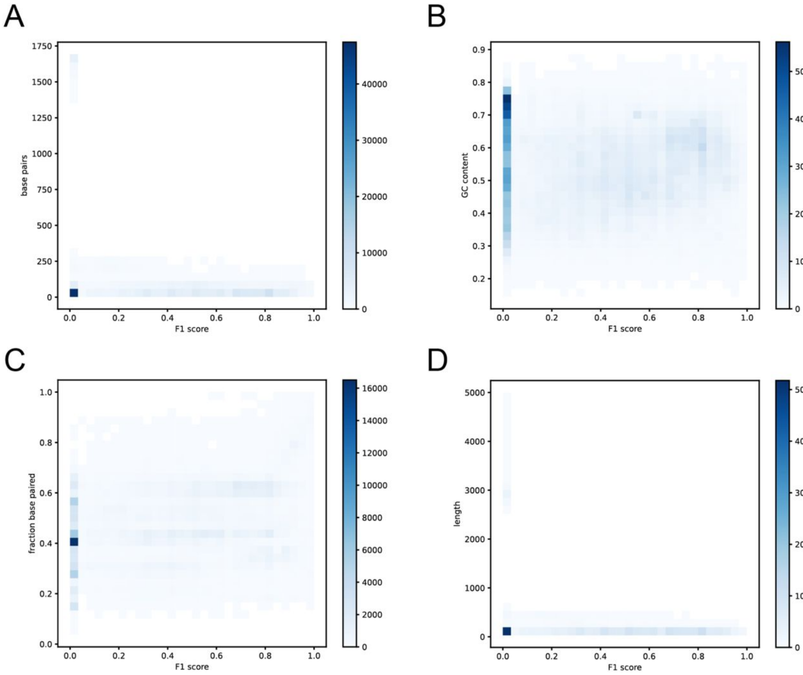


Figure 6: RNASSTR features versus F1 score. A-D) Density of RNASSTR test partition features versus the resulting F1 score from the retrained SincFold model. Features analyzed are as follows: absolute number of base pairs in the ground truth structure (A), sequence GC content (B), fraction of bases paired in the ground truth structure (C), and sequence length (D).

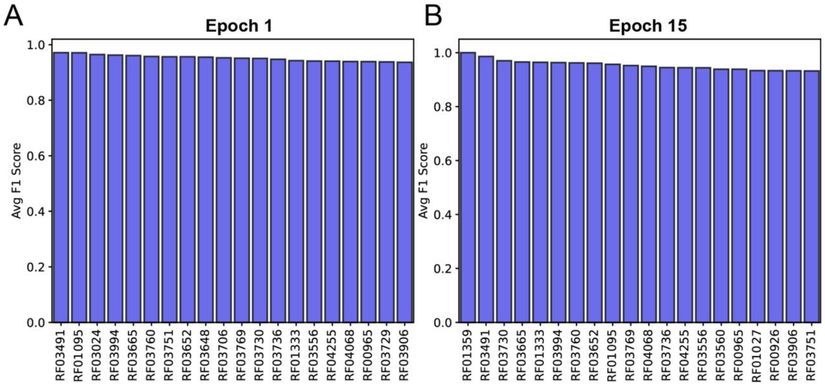
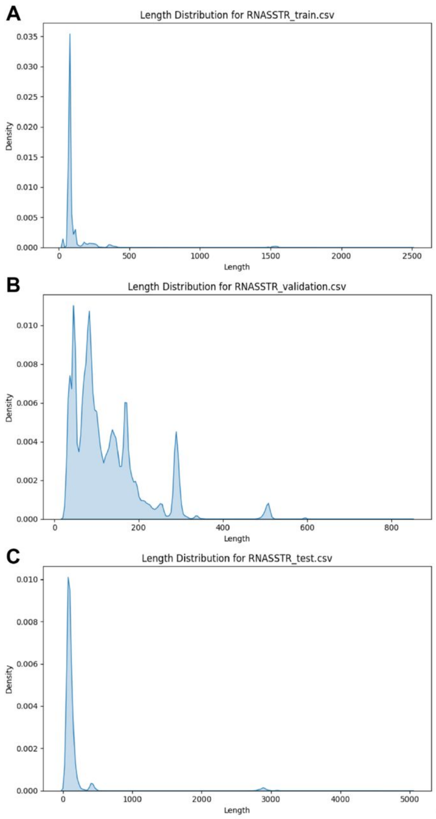


Figure 7: Average F1 score of top 20 performing classes. Shown are the F1 scores predicted by RNASSTR retrained SincFold at the end of epoch 1 (A) and epoch 15 (B).



\

Figure 8: RNASSTR features versus F1 score. A-C) Length distributions of the sequences contained in the three data partitions of RNASSTR: train (A), validation (B), and test (C). Note that the length scales on the horizontal axis differ for each partition, based on the longest sequence present.