

Deep Learning for RNA Secondary Structure Determination: Gauging Generalizability and Broadening the Scope of Traditional Methods

SUPPLEMENTAL METHODS

Marcell Szikszai^{1,2}, Ting-Yuan Wang³, Ryan Krueger⁴, David H. Mathews^{5,6}, Max Ward¹, Sharon Aviran^{3,7}

Author Affiliations

1 Department of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA 6009, Australia

2 Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

3 Department of Biomedical Engineering, University of California Davis, Davis, CA 95616, USA

4 School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

5 Department of Biochemistry and Biophysics, University of Rochester Medical Center, Rochester, NY 14642, USA

6 Center for RNA Biology, University of Rochester Medical Center, Rochester, NY 14642, USA

7 Genome Center, University of California Davis, Davis, CA 95616, USA

Data Collection

Details of the datasets analyzed in this study are summarized below. In total, four datasets were utilized, encompassing bacterial, viral, and human RNAs. Structural motifs shorter than 25 nucleotides were excluded from analysis. Below, we describe each dataset in detail.

E. coli

All structured regions and reactivity data were collected from prior work (Mustoe et al., 2018). Mustoe et al. identified 58 structural motifs using low SHAPE/low-entropy methods, 9 of which have been independently verified in earlier studies (Chiaruttini et al., 1996; Desnoyers et al., 2009; Fu et al., 2013). They split *rpmI* 5' UTR into two well-defined sub-motifs for ease of analysis, yielding a total of 10 motifs that comprise our validated dataset. For the remaining 49 motifs, we excluded the 5' UTR *rpoS* region (-) 2865600-2865661 due to its high proportion of missing values. The final putative dataset, therefore, consists of 48 motifs. Reactivity profiles were obtained from RASP v2.0 (Mu et al., 2025).

SARS-CoV-2

We used reactivity profiles from a prior study (Manfredonia et al., 2020), which we downloaded from the Incarnato laboratory website (https://www.incarnatolab.com/datasets/SARS_Manfredonia_2020.php). The conserved motifs reported in this study were included and designated as validated. In addition, we incorporated 5' UTR stem-loops 1–7 as validated elements, as they are consistently reported across multiple studies, while 5' UTR stem-loops 1–8 was included as putative since stem-loop 8 was not supported by their data. Additional structures were collected from other works. Hagey et al. identified two conserved structures, both of which were included and classified as validated (Hagey et al., 2022). Lan et al. investigated the frameshifting element (FSE), which we incorporated and designated as putative (Lan et al., 2022). Furthermore, Wacker et al. applied NMR spectroscopy and verified 15 structures (Wacker et al., 2020). Of these, one was excluded because it contained only a single stem within a stem-loop, another was excluded due to a high proportion of missing values, and three were excluded due to redundancy, as two overlapped with stem-loop 5 in the 5' UTR identified by Manfredonia et al. and one overlapped with the FSE reported by Lan et al. The remaining Wacker et al. structures were included and marked as validated, except for two that were noted in their study as requiring further investigation. Yang et al. applied evolutionary covariance analysis and reported 20 structural elements (Yang et al., 2025). Four of these overlapped with motifs identified by Manfredonia et al. or with the FSE described by Lan et al. and were therefore excluded. The remaining structures were retained and designated as putative. Because different viral strains were analyzed across studies, we performed manual sequence alignment to reconcile coordinate differences, particularly for motifs reported by Hagey et al. and Yang et al.

HIV-1

Structural motifs and reactivity profiles were obtained directly from (Siegfried et al., 2014). In that study, SHAPE-MaP was applied to identify structured elements based on low SHAPE reactivity and low Shannon entropy. We collected the *de novo* structures shown in Figure 4, which highlights regions with low SHAPE reactivity and low Shannon entropy. As these motifs were newly proposed and not well-validated, we designated them as putative. In addition, their Supp. Table 2 lists previously verified functional structures. From this list, we excluded splice donor and acceptor sites as well as all pseudoknots. The remaining motifs were incorporated into our analysis and designated as validated. Reactivity profiles were downloaded from RASP v2.0 (Mu et al., 2025).

H. sapiens

The RNAndria dataset comprises 1,456 human mRNA regions and 1,098 pri-miRNA sequences, each probed *in vitro* using DMS-MaPseq (Lajarte et al., 2024). It provides high-resolution structural information for long and diverse RNAs, extending our bacterial and viral datasets. There was no need to delineate independently folding regions or label them as validated or putative since these regions were folded *in vitro* in their entirety and then probed by DMS. The data were downloaded from <https://huggingface.co/rousseinlab/datasets>.

Experimental padding

Some of the sequences in the Ribonanza data are padded prior to probing, such that most probed sequences are of a fixed length. The overwhelming majority of training sequences are 177 nt long. By subtracting 30, which is the barcode length in most training sequences, and further subtracting 47, which is the total length of the constant 5' and 3' flanking regions, we are left with 100 nt as the region of interest. Based on this and on He et al.'s description of their work, we concluded that they predominantly padded training sequences shorter than 100 (He et al., 2024). We therefore padded only sequences shorter than 100 using He et al.'s code (downloaded from https://github.com/DasLab/big_library_design on 10/13/25), which aims to ensure that the added nucleotides do not interfere with the region of interest during probing. A detailed description of padding and barcoding procedures and other design considerations is provided in this site.

Computational padding

When training deep learning models, it is common practice to use batch processing to improve efficiency. However, RNA sequences vary in length, requiring padding to achieve uniform input dimensions. During inference, sequences are sorted by length before batching, so sequences of similar size are grouped together, thereby minimizing padding. Nevertheless, padding remains necessary in some cases. In RibonanzaNet, sequences are padded with the token "N" (in addition to A, U, C, and G). In our experiments, sequences were padded to lengths of 177, 207, 307, and 457 nucleotides. The lengths 177 and 207 nt are the most prevalent in the Ribonanza dataset, while 307 and 457 nt were chosen to evaluate the effect of longer padding, as these lengths are also

present in the testing set. Padding was applied at inference by modifying the parameter `max_len` in `dataset.py` of the RibonanzaNet repository, set to `max([max(length), length_to_pad])`.

Barcodes and constant regions

In the Ribonanza data, barcodes are designed as stem-loop structures containing a conserved UUCG tetraloop and variable stem lengths, with the most common stem length being 13 base-pairs, resulting in a total barcode length of 30 nucleotides. For our analyses, we used barcodes of 30 nucleotides taken from Ribonanza test sequences. All sequences were also flanked by constant sequences: a 5' sequence (5'-GGGAACGACUCGAGUAGAGUCGAAAA-3') and a 3' sequence (5'-AAAAGAAACAACAACAACAAC-3'). For the comparison experiments, sequences were processed in the same manner, with both barcodes and constant regions added to our dataset.

REFERENCES

- Chiaruttini, C., Milet, M., & Springer, M. (1996). A long-range RNA-RNA interaction forms a pseudoknot required for translational control of the IF3-L35-L20 ribosomal protein operon in *Escherichia coli*. *The EMBO Journal*, 15(16), 4402–4413. <https://doi.org/10.1002/j.1460-2075.1996.tb00813.x>
- Desnoyers, G., Morissette, A., Prévost, K., & Massé, E. (2009). Small RNA-induced differential degradation of the polycistronic mRNA *iscRSUA*. *The EMBO Journal*, 28(11), 1551–1561. <https://doi.org/10.1038/emboj.2009.116>
- Fu, Y., Deiorio-Hagggar, K., Anthony, J., & Meyer, M. M. (2013). Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria. *Nucleic Acids Research*, 41(6), 3491–3503. <https://doi.org/10.1093/nar/gkt055>
- Hagey, R. J., Elazar, M., Pham, E. A., Tian, S., Ben-Avi, L., Bernardin-Souibgui, C., Yee, M. F., Moreira, F. R., Rabinovitch, M. V., Meganck, R. M., Fram, B., Beck, A., Gibson, S. A., Lam, G., Devera, J., Kladwang, W., Nguyen, K., Xiong, A., Schaffert, S., ... Glenn, J. S. (2022).

Programmable antivirals targeting critical conserved viral RNA secondary structures from influenza A virus and SARS-CoV-2. *Nature Medicine*, 28(9), 1944–1955.

<https://doi.org/10.1038/s41591-022-01908-x>

He, S., Huang, R., Townley, J., Kretsch, R. C., Karagianes, T. G., Cox, D. B. T., Blair, H., Penzar, D., Vyaltsev, V., Aristova, E., Zinkevich, A., Bakulin, A., Sohn, H., Krstevski, D., Fukui, T., Tatematsu, F., Uchida, Y., Jang, D., Lee, J. S., ... Das, R. (2024). Ribonanza: Deep learning of RNA structure through dual crowdsourcing. *bioRxiv*, 2024.02.24.581671.

<https://doi.org/10.1101/2024.02.24.581671>

Lajarte, A. A. de, Taillades, Y. J. M. des, Kalicki, C., Wightman, F. F., Aruda, J., Salazar, D., Allan, M. F., L'Esperance-Kerckhoff, C., Kashi, A., Jossinet, F., & Rouskin, S. (2024). Diverse database and machine learning model to narrow the generalization gap in RNA structure prediction (p. 2024.01.24.577093). *bioRxiv*. <https://doi.org/10.1101/2024.01.24.577093>

Lan, T. C. T., Allan, M. F., Malsick, L. E., Woo, J. Z., Zhu, C., Zhang, F., Khandwala, S., Nyeo, S. S. Y., Sun, Y., Guo, J. U., Bathe, M., Nääär, A., Griffiths, A., & Rouskin, S. (2022). Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. *Nature Communications*, 13(1), 1128. <https://doi.org/10.1038/s41467-022-28603-2>

Manfredonia, I., Nithin, C., Ponce-Salvatierra, A., Ghosh, P., Wirecki, T. K., Marinus, T., Ogando, N. S., Snijder, E. J., van Hemert, M. J., Bujnicki, J. M., & Incarnato, D. (2020). Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant

elements. *Nucleic Acids Research*, 48(22), 12436–12452. <https://doi.org/10.1093/nar/gkaa1053>

Mu, K., Fei, Y., Xu, Y., & Zhang, Q. C. (2025). RASP v2.0: An updated atlas for RNA structure probing data. *Nucleic Acids Research*, 53(D1), D211–D219.

<https://doi.org/10.1093/nar/gkae1117>

Mustoe, A. M., Busan, S., Rice, G. M., Hajdin, C. E., Peterson, B. K., Ruda, V. M., Kubica, N., Nutiu, R., Baryza, J. L., & Weeks, K. M. (2018). Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing. *Cell*, *173*(1), 181-195.e18.

<https://doi.org/10.1016/j.cell.2018.02.034>

Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., & Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods*, *11*(9), 959–965.

<https://doi.org/10.1038/nmeth.3029>

Wacker, A., Weigand, J. E., Akabayov, S. R., Altincekic, N., Bains, J. K., Banijamali, E., Binas, O., Castillo-Martinez, J., Cetiner, E., Ceylan, B., Chiu, L.-Y., Davila-Calderon, J., Dhamotharan, K., Duchardt-Ferner, E., Ferner, J., Frydman, L., Fürtig, B., Gallego, J., Grün, J. T., ... Zetzsche, H. (2020). Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. *Nucleic Acids Research*, *48*(22), 12415–12435.

<https://doi.org/10.1093/nar/gkaa1013>

Yang, S. L., DeFalco, L., Wang, S., Wong, Y. H., Han, J., Mok, C. K., Tan, K. Y., Lim, S. Y., Zhao, Z., Zhang, Y., Lim, J. J. A., Xiang, J. S., Sobota, R., Wang, L.-F., Chu, J. J. H., Merits, A., Huber, R. G., & Wan, Y. (2025). A conserved long-range RNA interaction in SARS-CoV-2 recruits ADAR1 to enhance virus proliferation. *Nature Communications*, *16*(1), 8460.

<https://doi.org/10.1038/s41467-025-63297-2>