ON INPUTS TO DEEP LEARNING FOR RNA 3D STRUCTURE PREDICTION

Marcell Szikszai,[1,2] **Marcin Magnus,**[2] **Sachin Kadyan,**[3] **& Elena Rivas**[2,*]
[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 Systems Biology, Columbia University, New York 10027, NY, USA
[*] corresponding author.

ABSTRACT

Today, there are several effective deep learning models for predicting the 3D structure of proteins. Building on their success, models have been developed for predicting the 3D structure of non-coding RNAs. Unfortunately, these models are much less accurate than their protein counterparts. In this paper, we highlight differences between protein and RNA structure, and demonstrate methods for deep learning targeted at addressing those differences, with the aim of prompting discussion on these topics. We present an RNA-specific pipeline for generating structural Multiple Sequence Alignments (MSAs). Derived from the structural alignments, we introduce engineered evolutionary features that strongly inform RNA structure. Further, from the crystal structure, we derive structural features describing RNA base pairing. These evolutionary and structural features can be used in loss functions at different stages of training. Finally, we discuss different cropping strategies informed by RNA structure.

1 INTRODUCTION

The prediction of 3D protein structure was revolutionized via deep learning by AlphaFold in 2018 (Senior et al., 2020). Since then, the number of tools that apply deep learning to broader problems in structural biology has skyrocketed. It is no surprise that researchers immediately began adapting the lessons from AlphaFold toward non-coding ribonucleic acid (RNA) 3D structure prediction (Chen et al., 2020; Wang et al., 2021; Sato et al., 2021; Fu et al., 2021; Pearce et al., 2022; Shen et al., 2022; Baek et al., 2022; Feng et al., 2022; Li et al., 2022; Abramson et al., 2024). The core problems are essentially analogous: take a 1D polymer sequence as input, and predict the 3D conformation of the molecule. For both proteins and structured RNAs, the 3D structure is a consequence of the 1D polymer sequence, and the 3D structure has strong ties to the molecule's function.

With increasing interest in 3D prediction of the RNA structure, there was a need for more robust tools to benchmark performance. The blind-assesment competition CASP15 (Kryshtafovych et al., 2023) joined RNA-Puzzles (Cruz et al., 2012; Miao et al., 2015; 2017; 2020) in 2022 to include RNA-only targets in the competition. These assessments of novel RNA structures indicates that to this day, deep-learning methods have yet to catch up to other existing traditional methods for RNA structure prediction, as reported by the latest CASP16 and RNA-Puzzle (Miao et al., 2020) competitions.

The amount of RNA structural data available to perform deep-learning RNA 3D structure prediction pales in comparison to that available for protein structure prediction (Szikszai et al., 2024). Methods like RNA3DB (Szikszai et al., 2024) have been created recently to exhaustively characterize structural homologies in existing RNA PDB chains, and to provide flexible tools to avoid structural homology overlap when designing training and testing sets for robust benchmarking.

As we learn from the successes in protein structure prediction, several additional considerations have to be taken into account when creating deep-learning methods for RNA structure prediction, owing

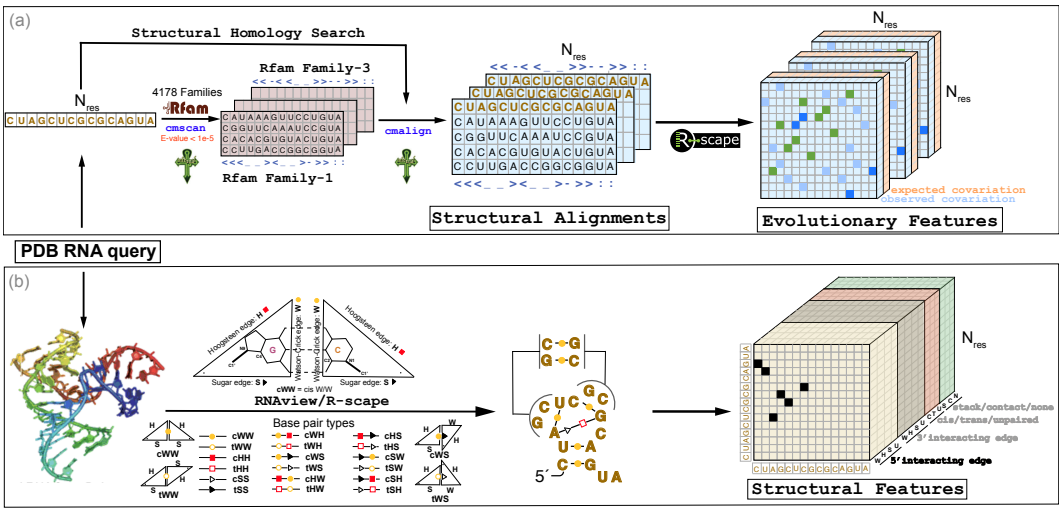


Figure 1: **Overview of RNA structural inputs for training deep-learning methods.** **(a)** The structural alignments are generated by structural homology search to the Rfam database (Ontiveros-Palacios et al., 2024), with the Infernal (Nawrocki & Eddy, 2013) method. We build one structural alignment for each Rfam family with significant homology to the query (E-value < 1e-5). See Appendix section B for details. From each structural alignment, we extract evolutionary features (both observed and expected covariation) using R-scape (Rivas et al., 2017; 2020). **(b)** From the query PDB file, we extract structural properties related to the base pairing geometry (including all possible base-pair types) using RNAview (Yang et al., 2003).

to the distinct properties of RNA compared to proteins. We investigate some of these considerations in detail in this manuscript. Our contribution is a method to generate RNA structural alignments, and evolutionary and structural features that can be used to inform the training of RNA 3D structure prediction methods. See Figure 1.

2 RFAM-BASED STRUCTURAL MULTIPLE SEQUENCE ALIGNMENTS

The idea behind using multiple sequence alignments (MSAs) as the input to AlphaFold-like deep learning models is the expectation that evolutionary information informs structure. This is true for both proteins and RNAs. However, with RNAs, the degree to which alignments inform structure varies highly by type. Structured RNAs, such as tRNA and rRNA, rely on specific conformations to perform their function. As a result, their structure is highly conserved. This conservation is easily detectable in their MSAs. Base pairing in particular can be inferred from alignments by looking at positive and negative evolutionary information (Rivas, 2020). On the other hand, some RNAs, like mRNAs, rely primarily on their conserved codon organization to determine their function. These mRNAs still form base pairs and fold into some 3D conformation, but they are generally not conserved (even though certain folded configurations may be more stable).

Since most interest in RNA 3D structure is focused on structured RNAs, alignments should be made with evolutionary conservation of structure in mind, when possible. However, the methods used by current deep learning models either assume all RNAs are structural, or do not incorporate structural conservation into their alignments. There are, broadly speaking, two pipelines representative of the approaches used by all existing methods: the rMSA (Zhang et al., 2023) pipeline that fits alignments to proposed structures, or all structure-agnostic HMMER-based (Eddy, 2008; 2009; 2011) pipelines as used by AlphaFold 3.

The pipeline used by AlphaFold 3 relies on a large database of clustered representative RNA sequences from Rfam (Ontiveros-Palacios et al., 2024), RNAcentral (RNAcentral Consortium, 2021), and Nucleotide Collection (Sayers et al., 2023). HMMER is then used to find homologous sequences in this database. HMMER uses profile hidden Markov models to search sequence databases

for homologues using sequence only. It does not consider the secondary structure of RNAs in the homology search.

It is well established that alignments for structural RNAs can be improved by using both sequence and secondary structure (Freyhult et al., 2007; Nawrocki, 2009), using structural homology methods such as Infernal (Nawrocki & Eddy, 2013). Infernal works by constructing profile stochastic context-free models (*covariance models* or CMs) of RNA families, which are trained from a family-specific MSA along with a consensus secondary structure.

The rMSA pipeline, used by RoseTTAFoldNA (Baek et al., 2023) and others (Wang et al., 2023), starts out with an initial HMMER alignment to first identify homologous sequences. Then it creates an Infernal covariance model using a predicted RNA secondary structure. This Infernal covariance model is then used for homology searches to arrive at a final alignment. This pipeline does consider secondary structure, but a relatively unreliable one, since the consensus is found through thermodynamic folding (Lorenz et al., 2016), and not tested for evolutionary conservation of the structure. Importantly, this approach assumes that the query sequence conserves its secondary structure, which may not be the case (e.g. mRNAs or synthetic constructs). Additionally, Infernal is used with an E-value cut-off of 10, which is prone to favor the inclusion of false positive homologues. The artifacts created by the high E-value cut-off and the assumption of conserved secondary structure was previously documented by Gao et al. (2022).

For known structural RNAs used in training, we propose to take advantage of the Rfam database (Ontiveros-Palacios et al., 2024). Rfam compiles a database of Infernal covariance models which classifies structural RNAs into families. Each family has a *seed alignment*, the MSA used to build the covariance model, along with a carefully created consensus secondary structure. Our method described in Figure 1a starts by finding Rfam families that show statistical significant homology to the query RNA sequence. For each homologous RNA family, a structural alignment including the query and sequences that belong to the Rfam family can be constructed. For the PDB database, as reported by RNA3DB (2024-12-04-full-release)(Szikszai et al., 2024), of the 1,869 RNA representative chains (clustered at 99% identity), 67% (1,198) have homology to at least one Rfam family with an E-value cutoff of $1e^{-5}$. RNA chains without homology to Rfam usually fall in the category of synthetic sequences, mRNAs or small fragments lacking structure.

From our structural alignments, we can directly feed covariation evolutionary information into the model in order to provide a signal about secondary and tertiary structure. See Figure 1a and Appendix Section A for a discussion.

Figure 2 presents a comparison of the performance of the alignment methods on a 5S rRNA structural RNA chain and a Purine riboswitch aptamer. Even though the AlphaFold-like alignment includes many more sequences, the alignment is less accurate identifying the positions that are base paired. Similar results are presented in the supplement for two other structural RNAs. Other examples are given in Appendix Section B.

These examples show the power of using Rfam's covariance models, and building MSAs using Infernal with appropriate E-value cutoffs. In cases where no Rfam CM produces a significant hit, it may be difficult to determine whether the RNA is structural or not. As a result, alignments made with HMMER (as done by AlphaFold 3) will be the most adequate and informative without making assumptions about a secondary structure that could introduce circular analysis artifacts (Gao et al., 2022).

3 Loss functions associated to RNA base pairing

RNA folding is hierarchical (Tinoco & Bustamante, 1999) meaning that the secondary structure (that is, the collection of base pairs) is more stable and forms faster than the 3D structure (Banerjee et al., 1993; Mathews et al., 1997; Onoa et al., 2003). The RNA secondary structure heavily informs the 3D (Shapiro et al., 2007; Miao et al., 2020). As a result, it is often argued that for a model to predict 3D correctly, it must correctly predict the 2D structure first (Kerpedjiev et al., 2015). However, there is little discussion from the RNA 3D structure prediction community about loss functions that target RNA base pairing specifically.

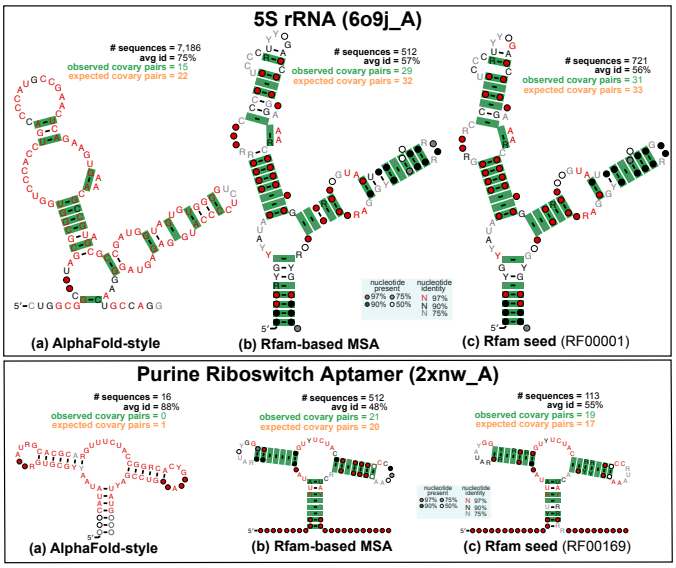


Figure 2: **Comparison of alignment methods.** For a 5S rRNA PDB chain and a Purine riboswitch aptamer, we show: **(a)** A HMMER alignment made against Rfam, RNAcentral and the nucleotide databases. **(b)** A structural alignment created with our Rfam-based method. **(c)** A curated structural alignment for the RNA family from Rfam. We show the evolutionary information present in each alignment as the significantly covarying base pairs (depicted in green) found using R-scape. Covarying base pairs are given in the context of a CaCoFold (Rivas, 2020) consensus secondary structure that incorporates all significantly covarying pairs found in the alignment.

AlphaFold-like models distinguish two different kinds of losses. Structural losses rely on the entire 3D structure. These are usually end-to-end, and are evaluated at the end of the structure module or the diffusion module in the case of AlphaFold 3, e.g. Frame Aligned Point Error (FAPE) (Jumper et al., 2021). Also, there are auxiliary losses that apply to linear projections from the internal pair representation, usually just before the structure module and after the Pairformer or Evoformer, e.g. distogram loss (Jumper et al., 2021). Here we propose two loss functions, one structural and one auxiliary to inform specifically about RNA base pairing.

We propose a loss termed *pairtogram* loss–a play on AlphaFold's distograms[1]. A pairtogram describes the base pairing geometry and can be used as an auxiliary loss. To construct a pairtogram, we extract an augmented Leontis-Westhof base pair geometry classification matrix (Leontis & Westhof, 2001) for all $N_{\text{res}} \times N_{\text{res}}$ pairs. This classification provides 12 basic geometric types, distinguishing between Watson-Crick, Hoogsteen, or Sugar-edge interacting edges, and cis or trans bond orientations. For instance, canonical RNA base pairs `A:U`, `G:C` are cis interactions between the Watson-Crick edges of both residues. The standard annotation is augmented with whether the pair is stacked (but not any of the defined base pairs), or a contact (defined as residues at a distance smaller than 8 Å), or neither. See Appendix section C for details on how the pairtogram loss is calculated.

Our RNA structural loss considers the dihedral angle between the planes of the two nucleotide bases. Pyrimidine bases are completely planar, while the base in purines is nearly planar (but can be functionally treated as such) (Callahan, 2011). As a result, we can assign a plane corresponding to a base via three atoms in all residues. In principle, any three atoms can be used since the bases are largely planar, but we consider two planes for both pyrimidines and purines for redundancy. The planes for purines are defined by C1′-N9-C4 and C8-N9-C4, while the planes for pyrimidines are defined by C1′-N1-C2 and C6-N1-C2 (see Figure 3). These planes are then used to calculate the dihedral angles between bases. Residues forming Watson-Crick base pairs will have angles close to $180°$. Moreover, residues in one side of a canonical helix will also have angles close to $180°$ with residues on the other side of the helix. On the other hand, residues in one side of the helix will have very small base angles amongst each other.

[1] *Distance histograms*, an output of discretised pairwise distances between all atoms.

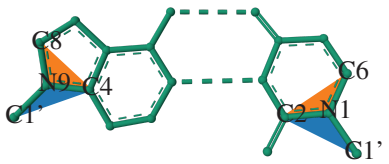


Figure 3: **Between-Base Angle Planes.** An example of how the planes are constructed for a purine (left, adenine) and a pyrimidine (right, uracil). The first planes are shown in blue, while the second planes are shown in orange. The example base pair is from a tRNA (PDB: `1ehz_A`) (Shi & Moore, 2000) and drawn with Mol* (Sehnal et al., 2021).

Let the plane for a residue i be defined by three atom coordinates $(\mathbf{A}_{\text{C1'/C8/C6}}, \mathbf{A}_{\text{N9/N1}}, \mathbf{A}_{\text{C4/C2}})$ where $\mathbf{A} \in \mathbb{R}^3$. Then, we calculate the normal of the plane,

$$\mathbf{N}_i = (\mathbf{A}_{\text{N9/N1}} - \mathbf{A}_{\text{C1'/C8/C6}}) \times (\mathbf{A}_{\text{C4/C2}} - \mathbf{A}_{\text{C1'/C8/C6}}). \tag{1}$$

For each pair of planes i and j, we calculate the sine and cosine of the dihedral angle,

$$\sin\theta_{ij} = \frac{\|\mathbf{N}_i \times \mathbf{N}_j\|}{\|\mathbf{N}_i\|\|\mathbf{N}_j\|}, \quad \cos\theta_{ij} = \frac{\mathbf{N}_i \cdot \mathbf{N}_j}{\|\mathbf{N}_i\|\|\mathbf{N}_j\|}. \tag{2}$$

Then we define our loss, Between Base Angle Error (BBAE) as,

$$\mathcal{L}_{\text{BBAE}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{TRUE}}) = \sum_{i<j}^{N_{\text{res}}} [(\sin\theta_{ij} - \sin\theta_{ij}^{\text{TRUE}})^2 + (\cos\theta_{ij} - \cos\theta_{ij}^{\text{TRUE}})^2]^{\dagger}, \tag{3}$$

where $\boldsymbol{\theta}$ are the dihedral angles for the predicted structure, and $\boldsymbol{\theta}^{\text{TRUE}}$ are the dihedral angles for the ground-truth structure.

4 Residue Cropping

A key consideration for deep learning models is memory. Recently, this issue has received a lot of attention, since large-scale state-of-the-art models (such as large language models) may have hundreds of billions of parameters (Kaplan et al., 2020; Grattafiori et al., 2024), which must be loaded into memory during both inference and training. In the case of 3D structure prediction models, the number of parameters is much more manageable (on the order of around 100M parameters for AlphaFold 2 (Jumper et al., 2021), for example), but the memory consumption is still relatively high, and can scale cubically or quadratically with sequence length (Vaswani et al., 2017; Senior et al., 2020; Jumper et al., 2021). As a result, deep learning models for 3D structure prediction often have to crop the sequences to shorter sub-sequences during training and process these crops one at a time. This is commonly referred to as *residue cropping*.

Initially, the AlphaFold 1 convolutional neural network (CNN) produced crops of length 64 in order to generate 64×64 distograms (Senior et al., 2020). This cropping strategy is easy to motivate for proteins: existing literature has shown that protein contact prediction only needs a limited context window (Jones & Kandathil, 2018; Senior et al., 2020). Unfortunately, this is not the case for RNAs. Some structural RNA families, such as SSU and LSU rRNA subunits, contain long-range base pairs that are more than 500 residues apart, and are difficult to predict for traditional dynamic programming algorithms (Huang et al., 2019).

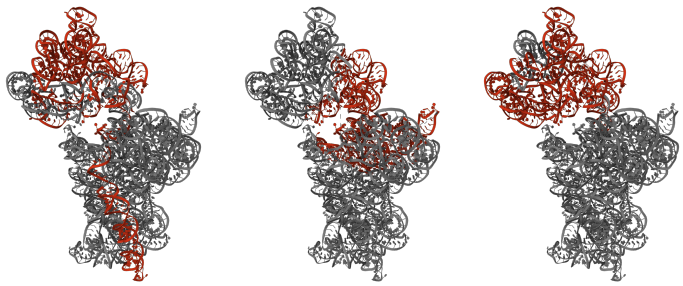
In 2021, AlphaFold 2 moved from a CNN-based architecture to a transformer model, which requires $O(n^3)$ memory for a sequence of length n. This memory requirement comes from *triangular* self-attention. As a result, AlphaFold 2 takes highly restrictive crops of 256 residues during the initial training phase and crops of 384 residues during fine-tuning. The starting position of these crops is randomly sampled from $\mathcal{U}\{1, N_{\text{res}} - \texttt{crop\_size} + 1\}$ where N_{res} is the length of the sequence.

[†]Note that $(\sin\theta_{ij} - \sin\theta_{ij}^{\text{TRUE}})^2 + (\cos\theta_{ij} - \cos\theta_{ij}^{\text{TRUE}})^2 = 2 - 2\cos(\theta_{ij} - \theta_{ij}^{\text{TRUE}})$.

These crops are *contiguous* in sequence, so that given a starting position i, the window contains residues $[i, i + \texttt{crop\_size} - 1]$.

While this strategy works well for proteins, since they only need small context windows, it is sub-optimal at best for RNAs. AlphaFold 2 contiguous crops break RNA base pairs, and include only one half of the RNA canonical helices that are the foundation of any RNA 3D structure (Figure 4a). Despite this, some methods for RNA 3D structure prediction, such as DRfold (Li et al., 2022) use continuous cropping for training. Consequently, DRfold also restricts their test set to RNAs with lengths between 14 and 392 nucleotides (Li et al., 2022), likely masking the performance degradation of their cropping strategy.

When DeepMind debuted AlphaFold-Multimer for protein complex prediction in 2022, they introduced *spatial* crops where residues are selected by their spatial distance in the 3D structure. For these spatial crops, a starting position is sampled from $\mathcal{U}\{1, N_{\text{res}}\}$. A reference atom is chosen (in the case of AlphaFold-Multimer C_α atoms), from which all distances are measured. Then, the `crop_size` nearest residues, measured by Euclidean distance to the reference atoms, are taken as the crop (Figure 4b). In AlphaFold-Multimer, spatial crops are chosen in a 50:50 ratio along with contiguous crops with a `crop_size` of 384. Although only used by AlphaFold-Multimer for multimer interface residues, the concept can be easily adapted to monomers.



(a) Contiguous (b) Spatial (c) Graph-based

Figure 4: **Different cropping strategies.** Example for a 30S rRNA (PDB: `5no2_A`) (López-Alonso et al., 2017). The red regions indicate the cropped sequence. The starting position is 1,100 with $\texttt{crop\_size} = 384$, with Watson-Crick base pairs extracted using RNApdbee 2.0 (Zok et al., 2018) and RNAview (Leontis & Westhof, 2001). The visualisations were created with Mol* (Sehnal et al., 2021).

Beyond *spatial* crops, RoseTTAFoldNA (Baek et al., 2023) also developed an alternate strategy for cropping nucleic acid–protein complexes and RNA monomers to explicitly avoid breaking base pairs and to pick context windows better suited to preserve RNA canonical helices composed of stacked Watson-Crick base pairs (Figure 4c). For RNA monomers, RoseTTAFoldNA builds a weighted undirected graph of the sequence where sequential residues have edges with a weight of one, and Watson–Crick base pairs have a weight of zero. As before, a random starting position is sampled from $\mathcal{U}\{1, N_{\text{res}}\}$, and minimum-weight graph traversal is used to find the nearest $\texttt{crop\_size} = 256$ residues based on the graph-distance.

We suggest a combined method using *contiguous* cropping together with RoseTTAFoldNA's *interaction graph-based* and AlphaFold 3 *spatial* crops. The ratios of the different cropping categories can be customized for the different training sets and different environments, with the goal of balancing time and performance with structure prediction accuracy.

5 Summary

The prediction of RNA 3D structure from sequence by deep learning methods is challenged by the small amount of structural data existing to train the models in comparison to proteins. This manuscript aims at lowering the impact of such a fundamental problem by making sure that the in-

formation obtained from the existing inputs extracts the maximal amount of structural RNA specific properties, both structural and at the level of RNA base pairing.

We have presented structural alignments for PDB RNA chains that capture significantly more pairing information than other agnostic homology methods. We have introduced losses capturing RNA base pairing information, including non-canonical base pairs, and also structural losses that capture the stacked nature of the 3D helices formed by the RNA canonical base pairs. Finally, we have discussed base pairing-aware cropping strategies. By introducing these topics in this manuscript, we hope to encourage further research and discussion on RNA-specific models in the field of RNA 3D structure prediction.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w. Publisher: Nature Publishing Group.

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, David Baker, and Frank DiMaio et al. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA, September 2022. URL https://www.biorxiv.org/content/10.1101/2022.09.09.507333v1. Pages: 2022.09.09.507333 Section: New Results.

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, pp. 1–5, November 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-02086-5. URL https://www.nature.com/articles/s41592-023-02086-5. Publisher: Nature Publishing Group.

A. R. Banerjee, J. A. Jaeger, and D. H. Turner. Thermal unfolding of a group I ribozyme: the low-temperature transition is primarily disruption of tertiary structure. *Biochemistry*, 32(1):153–163, January 1993. ISSN 0006-2960. doi: 10.1021/bi00052a021.

Michael P. Callahan. Nucleic Acid Base. In Muriel Gargaud, Ricardo Amils, José Cernicharo Quintanilla, Henderson James (Jim) Cleaves, William M. Irvine, Daniele L. Pinti, and Michel Viso (eds.), *Encyclopedia of Astrobiology*, pp. 1138–1140. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-11274-4. doi: 10.1007/978-3-642-11274-4_1080. URL https://doi.org/10.1007/978-3-642-11274-4_1080.

Xinshi Chen, Yu Li, Ramzan Umarov, Xin Gao, and Le Song et al. RNA Secondary Structure Prediction By Learning Unrolled Algorithms. In *International Conference on Learning Representations*, 2020. doi: 10.48550/arXiv.2002.05810. URL https://openreview.net/forum?id=S1eALyrYDH.

Young-In Chi, Monika Martick, Monica Lares, Rosalind Kim, William G. Scott, and Sung-Hou Kim. Capturing Hammerhead Ribozyme Structures in Action by Modulating General Base Catalysis. *PLOS Biology*, 6(9):e234, September 2008. ISSN 1545-7885. doi: 10.1371/journal.pbio.0060234.

José Almeida Cruz, Marc-Frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das. RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, 18(4):610–625, April 2012. ISSN 1355-8382. doi: 10.1261/rna.031054.111. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3312550/.

Sean R. Eddy. A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. *PLOS Computational Biology*, 4(5):e1000069, May 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000069. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000069. Publisher: Public Library of Science.

Sean R. Eddy. A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, 23(1):205–211, October 2009. ISSN 0919-9454.

Sean R. Eddy. Accelerated Profile HMM Searches. *PLOS Computational Biology*, 7(10):e1002195, October 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002195. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002195. Publisher: Public Library of Science.

Chenjie Feng, Wenkai Wang, Renmin Han, Ziyi Wang, Lisa Ye, Zongyang Du, Hong Wei. Accurate de novo prediction of RNA 3D structure with transformer network, October 2022. URL https://www.biorxiv.org/content/10.1101/2022.10.24.513506v1. Pages: 2022.10.24.513506 Section: New Results.

Eva K. Freyhult, Jonathan P. Bollback, and Paul P. Gardner. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Research*, 17(1):117–125, January 2007. ISSN 1088-9051. doi: 10.1101/gr.5890907. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716261/.

Laiyi Fu, Yingxin Cao, Jie Wu, Qinke Peng, Qing Nie, and Xiaohui Xie. UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Research*, pp. gkab1074, November 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1074. URL https://doi.org/10.1093/nar/gkab1074.

W. Gao, A. Yang, and E. Rivas. Thirteen dubious ways to detect conserved structural RNAs. *IUBMB Life*, 75:471–492, 2022. doi: https://doi.org/10.1002/iub.2694. URL https://iubmb.onlinelibrary.wiley.com/doi/10.1002/iub.2694.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Liang Huang, He Zhang, Dezhong Deng, Kai Zhao, Kaibo Liu, David A Hendrix, and David H Mathews. LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics*, 35(14):i295–i304, July 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz375. URL https://doi.org/10.1093/bioinformatics/btz375.

David T Jones and Shaun M Kandathil. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, 34(19):3308–3315, October 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty341. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6157083/.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs].

Aayush Karan and Elena Rivas. All-at-once RNA folding with 3D motif prediction framed by evolutionary information, December 2024. URL https://www.biorxiv.org/content/10.1101/2024.12.17.628809v1. Pages: 2024.12.17.628809 Section: New Results.

Peter Kerpedjiev, Christian Höner zu Siederdissen, and Ivo L Hofacker. Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, 21:1110–1121, 2015.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult et al. Critical assessment of methods of protein structure prediction (CASP)-Round XV. *Proteins*, 91(12):1539–1549, December 2023. ISSN 1097-0134. doi: 10.1002/prot.26617.

N B Leontis and E Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, April 2001. ISSN 1355-8382. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1370104/.

Yang Li, Chengxin Zhang, Chenjie Feng, Peter L. Freddolino, and Yang Zhang et al. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction, December 2022. URL https://www.biorxiv.org/content/10.1101/2022.12.30.522296v1. Pages: 2022.12.30.522296 Section: New Results.

Ronny Lorenz, Ivo L. Hofacker, and Peter F. Stadler. RNA folding with hard and soft constraints. *Algorithms for Molecular Biology*, 11(1):8, April 2016. ISSN 1748-7188. doi: 10.1186/s13015-016-0070-z. URL https://doi.org/10.1186/s13015-016-0070-z.

Jorge Pedro López-Alonso, Tatsuya Kaminishi, Takeshi Kikuchi, Yuya Hirata, Idoia Iturrioz, Neha Dhimole, Andreas Schedlbauer. RsgA couples the maturation state of the 30S ribosomal decoding center to activation of its GTPase pocket. *Nucleic Acids Research*, 45(11):6945–6959, June 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx324. URL https://doi.org/10.1093/nar/gkx324.

D. H. Mathews, A. R. Banerjee, D. D. Luan, T. H. Eickbush, and D. H. Turner et al. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA (New York, N.Y.)*, 3(1):1–16, January 1997. ISSN 1355-8382.

Zhichao Miao, Ryszard W. Adamiak, Marc-Frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Clarence Cheng. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*, 21(6):1066–1084, June 2015. ISSN 1355-8382, 1469-9001. doi: 10.1261/rna.049502.114. URL http://rnajournal.cshlp.org/content/21/6/1066. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Zhichao Miao, Ryszard W. Adamiak, Maciej Antczak, Robert T. Batey, Alexander J. Becka, Marcin Biesiada, Michał J. Boniecki. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*, 23(5):655–672, May 2017. ISSN 1355-8382, 1469-9001. doi: 10.1261/rna.060368.116. URL http://rnajournal.cshlp.org/content/23/5/655. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Zhichao Miao, Ryszard W. Adamiak, Maciej Antczak, Michał J. Boniecki, Janusz Bujnicki, Shi-Jie Chen, Clarence Yu Cheng. RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, 26(8):982–995, August 2020. ISSN 1355-8382. doi: 10.1261/rna.075341.120. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7373991/.

Eric P. Nawrocki and Sean R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, November 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt509. URL https://doi.org/10.1093/bioinformatics/btt509.

Eric Paul Nawrocki. *Structural RNA homology search and alignment using covariance models*. Washington University in St. Louis, 2009.

Bibiana Onoa, Sophie Dumont, Jan Liphardt, Steven B. Smith, Ignacio Tinoco, and Carlos Bustamante. Identifying kinetic barriers to mechanical unfolding of the T. thermophila ribozyme. *Science (New York, N.Y.)*, 299(5614):1892–1895, March 2003. ISSN 1095-9203. doi: 10.1126/science.1081338.

N. Ontiveros-Palacios, E. Cooke, E. P. Nawrocki, S. Triebel, M. Marz, E. Rivas, S. Griffiths-Jones. Rfam 15: RNA families database in 2025. *NAR*, gkae1023, 2024. doi: https://doi.org/10.1093/nar/gkae1023.

Robin Pearce, Gilbert S. Omenn, and Yang Zhang. De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning, May 2022. URL https://www.biorxiv.org/content/10.1101/2022.05.15.491755v1. Pages: 2022.05.15.491755 Section: New Results.

Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić et al. RiNALMo: General-Purpose RNA Language Models Can Generalize Well on Structure Prediction Tasks, February 2024. URL http://arxiv.org/abs/2403.00043. arXiv:2403.00043 [cs, q-bio].

Aiming Ren, Yi Xue, Alla Peselis, Alexander Serganov, Hashim M. Al-Hashimi, and Dinshaw J. Patel. Structural and Dynamic Basis for Low-Affinity, High-Selectivity Binding of L-Glutamine by the Glutamine Riboswitch. *Cell Reports*, 13(9):1800–1813, December 2015. ISSN 2211-1247. doi: 10.1016/j.celrep.2015.10.062.

Elena Rivas. RNA structure prediction using positive and negative evolutionary information. *PLOS Computational Biology*, 16(10):e1008387, October 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008387. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008387. Publisher: Public Library of Science.

Elena Rivas. RNA covariation at helix-level resolution for the identification of evolutionarily conserved RNA structure. *PLOS Computational Biology*, 19(7):e1011262, July 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011262. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011262. Publisher: Public Library of Science.

Elena Rivas, Jody Clements, and Sean R. Eddy. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods*, 14(1):45–48, January 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4066.

Elena Rivas, Jody Clements, and Sean R Eddy. Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics*, 36(10):3072–3076, May 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa080. URL https://doi.org/10.1093/bioinformatics/btaa080.

RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49(D1):D212–D220, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa921. URL https://doi.org/10.1093/nar/gkaa921.

Kengo Sato, Manato Akiyama, and Yasubumi Sakakibara. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications*, 12(1):941, February 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21194-4. URL https://www.nature.com/articles/s41467-021-21194-4. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Machine learning;Non-coding RNAs;RNA;Structure determination Subject_term_id: machine-learning;non-coding-rnas;rna;structure-determination.

Eric W. Sayers, Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jessica Chan, Donald C. Comeau, Catherine M. Farrell. Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Research*, 51(D1):D29–D38, January 2023. ISSN 1362-4962. doi: 10.1093/nar/gkac1032.

David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1):W431–W437, July 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab314. URL https://doi.org/10.1093/nar/gkab314.

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, January 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1923-7. URL https://www.nature.com/articles/s41586-019-1923-7. Number: 7792 Publisher: Nature Publishing Group.

Bruce A. Shapiro, Yaroslava G. Yingling, Wojciech Kasprzak, and Eckart Bindewald. Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology*, 17(2):157–165, April 2007. ISSN 0959-440X. doi: 10.1016/j.sbi.2007.03.001.

Tao Shen, Zhihang Hu, Zhangzhi Peng, Jiayang Chen, Peng Xiong, Liang Hong, Liangzhen Zheng. E2Efold-3D: End-to-End Deep Learning Method for accurate de novo RNA 3D Structure Prediction, July 2022. URL http://arxiv.org/abs/2207.01586. arXiv:2207.01586 [cs, q-bio].

Huijing Shi and Peter B. Moore. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA*, 6(8):1091–1105, August 2000. ISSN 1355-8382. doi: 10.1017/S1355838200000364. URL

https://www.cambridge.org/core/journals/rna/article/abs/crystal-structure-of-yeast-phenylalanine-trna-at-193-a-resolution-a-classic-struct AC4EBBDBBABEEC91D6B0D48E511B707C. Publisher: Cambridge University Press.

Marcell Szikszai, Marcin Magnus, Siddhant Sanghi, Sachin Kadyan, Nazim Bouatta, and Elena Rivas. RNA3DB: A structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction. *Journal of Molecular Biology*, pp. 168552, March 2024. ISSN 0022-2836. doi: 10.1016/j.jmb.2024.168552. URL https://www.sciencedirect.com/science/article/pii/S0022283624001475.

I. Tinoco and C. Bustamante. How RNA folds. *Journal of Molecular Biology*, 293(2):271–281, October 1999. ISSN 0022-2836. doi: 10.1006/jmbi.1999.3001.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Linyu Wang, Xiaodan Zhong, Shuo Wang, Hao Zhang, and Yuanning Liu et al. A novel end-to-end method to predict RNA secondary structure profile based on bidirectional LSTM and residual neural network. *BMC bioinformatics*, 22(1):169, March 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04102-x.

Wenkai Wang, Chenjie Feng, Renmin Han, Ziyi Wang, Lisha Ye, Zongyang Du, Hong Wei. trRosettaRNA: Automated prediction of RNA 3D structure with transformer network. *Nature Communications*, 14:7266, November 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-42528-4.

H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H. M. Berman, and E. Westhof. Tools for the automatic identification and classification of RNA base pairs. *NAR*, 31.13:3450–3460, 2003.

Chengxin Zhang, Yang Zhang, and Anna Marie Pyle. rMSA: A Sequence Search and Alignment Algorithm to Improve RNA Structure Modeling. *Journal of Molecular Biology*, 435 (14):167904, July 2023. ISSN 0022-2836. doi: 10.1016/j.jmb.2022.167904. URL https://www.sciencedirect.com/science/article/pii/S0022283622005241.

Tomasz Zok, Maciej Antczak, Michal Zurkowski, Mariusz Popenda, Jacek Blazewicz, Ryszard W Adamiak, and Marta Szachniuk. RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Research*, 46(W1):W30–W35, July 2018. ISSN 0305-1048. doi: 10.1093/nar/gky314. URL https://doi.org/10.1093/nar/gky314.

A STRUCTURAL EVOLUTIONARY FEATURES

As discussed in Section 2, the purpose of feeding MSAs into deep learning structure prediction pipelines is to provide evolutionary context about the residues. In traditional RNA structure prediction pipelines, MSAs can allow the model to identify covariation resulting from the presence of conserved RNA structure. For structured RNAs, the covariation derived from their alignments has been shown to be highly informative towards both the secondary and tertiary structures (Rivas et al., 2017; Rivas, 2020; Rivas et al., 2020; Rivas, 2023; Karan & Rivas, 2024).

Here we describe a method for how these features can be used by an AlphaFold-like architecture. Using the software R-scape (Rivas et al., 2017; 2020; Rivas, 2020), for any pair of positions, we can compute both the statistically significant covariation above phylogenetic expectation or observed covariation (in the form of an expected E-value), as well as the expected covariation (or power) given the number of total substitutions in the pair. Both E-value and power are binned into 8 bins, and E-value is clamped to the range $[0, 10.0]$. The feature tensor produced is of shape $[N_{\text{res}}, N_{\text{res}}, 16]$. This is then linearly projected to the pair representation channel dimension c_z, and added to the input of, for example, the main transformer block. This method also allows for calculating the E-value and power from multiple MSAs by computing the minimum, mean, and maximum across the MSAs, and producing a tensor of shape $[N_{\text{res}}, N_{\text{res}}, 16 \times 3]$.

Under ideal circumstances, it may seem unnecessary to explicitly feed in these features. We may expect that since deep learning is highly effective at representation learning, i.e. the ability to learn the useful representations from the raw data, we can just directly input the raw alignments and the network can learn these features implicitly as part of its internal representation. Currently we have no evidence to conclusively show whether inputting these features directly improves performance, however, there are two other possible motivations for explicitly calculating these.

First, these evolutionary maps provide an efficient way of embedding covariation information from a high-dimensional MSA in a way that is independent of the MSA dimensionality. AlphaFold 2's Evoformer memory cost is $O(N_{\text{seq}}^2 \times N_{\text{res}})$, where N_{seq} is the number of sequences in the alignment[3]. To address this, the model reduced the depth of the alignments using ***MSA clustering***, where a relatively small number of sequences ($N_{\text{clust}} = 128$ during the initial training phase, $N_{\text{clust}} = 512$ during fine-tuning[4]) cluster centres are picked. Then the remaining sequences in the MSA are assigned to their closest cluster by Hamming distance, and a number of statistics (e.g. distribution of amino acids) are computed for the cluster. Our proposed evolutionary features avoid the effect of any dimensionality reduction.

Second, while we currently suggest using these features as inputs, they may also be useful as auxiliary losses. It would be easy to create an auxiliary head that linearly projects the internal representation into the desired dimension (either $[N_{\text{res}}, N_{\text{res}}, 16]$ or $[N_{\text{res}}, N_{\text{res}}, 16 \times 3]$), and calculates the averaged cross-entropy loss as done for distograms by AlphaFold 2. We further suggest that these features may also be a useful auxiliary head for other models such as RNA language models like RiNALMo (Penić et al., 2024).

B STRUCTURAL ALIGNMENTS

For our Rfam-based MSAs, each family-specific alignment includes the query sequence, and up to 256 seed sequences from the Rfam family, or up to 512 full sequences if there are fewer than 256 seed sequences in the family.

In Figure 5, we describe two other examples of structural RNAs and the comparison of our Rfam/Infernal based structural alignments and the structure-agnostic alignments used by AlphaFold 3. For the two selected PDB chains: `5ddp_A` (a glutamine riboswitch aptamer) (Ren et al., 2015) and `2qus_A` (a Hammerhead_3 ribozyme) (Chi et al., 2008), we observe that Infernal is able to find multiple homologs in the Rfam database, and produces a structural alignment of quality comparable to that of the Rfam seed. On the other hand, a structure-agnostic search in the same database renders

[3]This is a simplification from the original AlphaFold 2 paper. We omit templates from our explanations.

[4]This is a simplification. $N_{\text{seq}} = N_{\text{clust}} + N_{\text{templ}}$.

few homologs and very sequence-conserved alignments that offer no evolutionary information about the secondary structures.

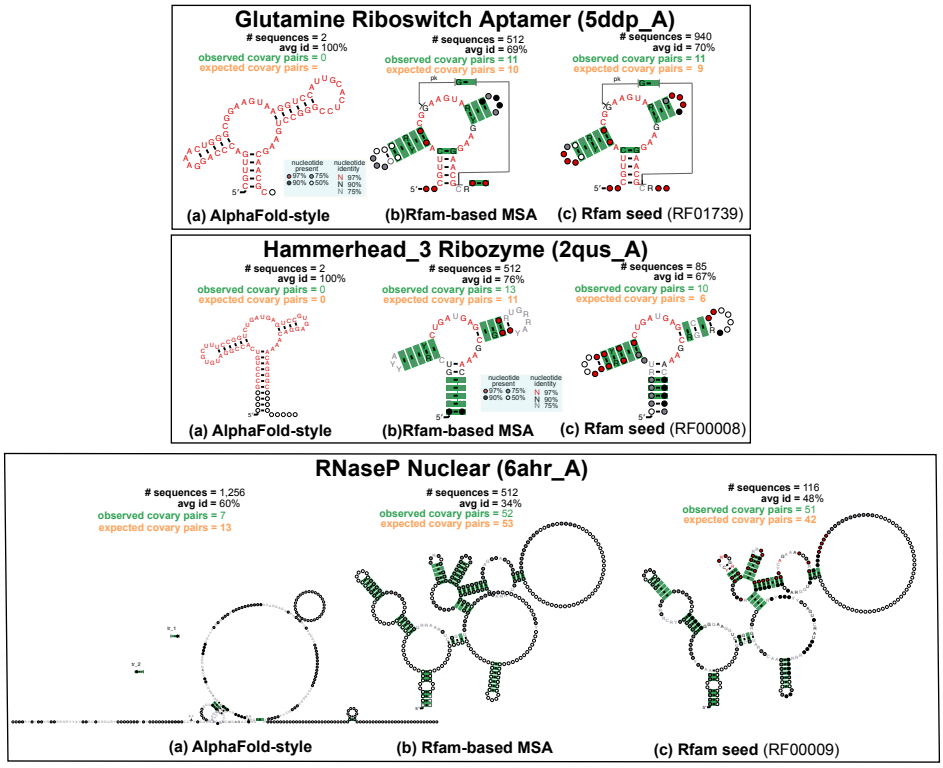


Figure 5: **Comparison of alignment methods.** **(a)** The AlphaFold 3-like alignments were created in-house using the following database: Nucleotide collection (nt) 112,177,963 sequences; 2,688,129,930,104 total bases (Feb 2, 2025 4:42 AM); BLASTDB Version: 5; RNAcentral Release 24, 07/03/2024. **(b)** Our structural alignments were created using Rfam v15.0, and Infernal v 1.1.4. **(c)** We compare to the Rfam seed alignment for the RNA family to which the queries belong. Evolutionary conserved base pairs are depicted in green. For other details, see caption of Figure 2.

C PAIRTOGRAM LOSS DETAILS

Pairtogram data is extracted from the 3D structures using R-scape (Rivas et al., 2017; 2020; Rivas, 2020), which includes a modified version of the software RNAView (Leontis & Westhof, 2001). The final pairtogram matrix has 14 total dimensions for each pair (see Table 1 and Figure 1b for a breakdown of the features). We treat these 14 dimensions as four separate one-hot vectors, and calculate the average cross-entropy loss across the four features between the ground-truth and a linear projection from the pair representation.

Description	Values	Dimension
$5'$ interacting edge	Watson-Crick, Hoogsteen, Sugar-Edge, unpaired	4
$3'$ interacting edge		4
Bond orientation	cis, trans, unpaired	3
Stacked/contact	stacked, contact, neither	3

Table 1: Structural features used in the *pairtogram* loss. The final tensor has size $[N_{\text{res}}, N_{\text{res}}, 14]$.