Critical Reviews and Perspectives

When will RNA get its AlphaFold moment?

Bohdan Schneider [1], **Blake Alexander Sweeney** [2], **Alex Bateman** [2,*], **Jiri Cerny** [1], **Tomasz Zok** [3] **and Marta Szachniuk** [3,4]

[1]Institute of Biotechnology of the Czech Academy of Sciences, Prumyslova 595, CZ-252 50 Vestec, Czech Republic, [2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, CB10 1SD, UK, [3]Institute of Computing Science and European Centre for Bioinformatics and Genomics, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland and [4]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland

ABSTRACT

The protein structure prediction problem has been solved for many types of proteins by AlphaFold. Recently, there has been considerable excitement to build off the success of AlphaFold and predict the 3D structures of RNAs. RNA prediction methods use a variety of techniques, from physics-based to machine learning approaches. We believe that there are challenges preventing the successful development of deep learning-based methods like AlphaFold for RNA in the short term. Broadly speaking, the challenges are the limited number of structures and alignments making data-hungry deep learning methods unlikely to succeed. Additionally, there are several issues with the existing structure and sequence data, as they are often of insufficient quality, highly biased and missing key information. Here, we discuss these challenges in detail and suggest some steps to remedy the situation. We believe that it is possible to create an accurate RNA structure prediction method, but it will require solving several data quality and volume issues, usage of data beyond simple sequence alignments, or the development of new less data-hungry machine learning methods.

GRAPHICAL ABSTRACT
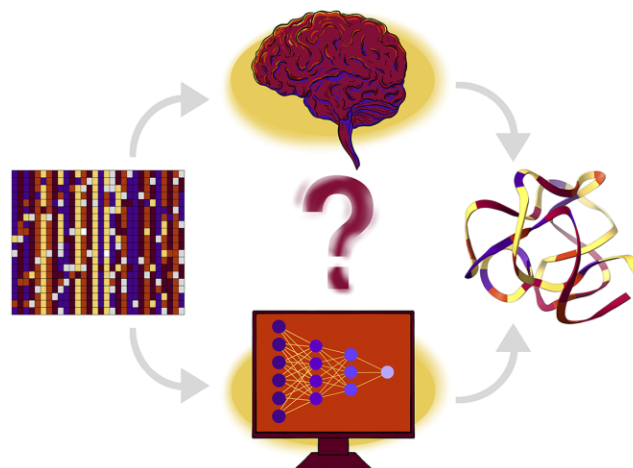
INTRODUCTION

RNA molecules play many key functions within cells. Perhaps the most striking example is in translation, where it has been shown that the ability to build proteins is orchestrated by ribosomal particles, with the crucial catalytic step being performed by the ribosomal RNA itself, with amino acid residues delivered specifically by transfer RNAs. Untranslated regions of mRNAs and viruses harbor numerous regulatory elements. There are also a large number of noncoding RNAs (ncRNA) for which, despite decades of research, we have only a scant understanding of their functions. An example is the large class of long noncoding RNAs in animal genomes. These RNA genes are numerous, perhaps exceeding the number of protein-coding genes and seem to play a range of subtle regulatory roles (1). Many ncRNA functions depend on the stable (ribosome, tRNA) or

*To whom correspondence should be addressed. Tel: +48 616652999; Fax: +44 1223494100; Email: agb@ebi.ac.uk
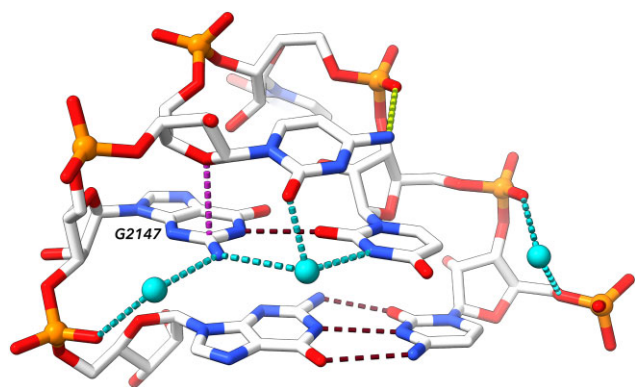
Figure 1. Examples of interactions in an RNA molecule. Some of the most important interactions are highlighted in dashed lines: base pairing hydrogen bonds in dark red, sugar-base stacking in dark violet, phosphate-base hydrogen bond in yellow, water-formed hydrogen bonds in cyan (waters are depicted as cyan balls). The bottom pair is canonical Watson–Crick, the pair above is a G–U pair 'locked' by interaction with bridging water molecule. G2147 is in *syn* orientation and dinucleotide C2146–G2147 is in the left-handed Z-form conformation (note the inverted direction of the ribose of C2146 further stabilized by stacking its O4' to the guanine aromatic ring). Displayed is a six nucleotide loop from 80 nucleotide long fragment of 23S RNA from *Thermus thermophilus* complexed with ribosomal protein L1 (PDB ID: 4qvi) (5).

transient (spliceosome) structure of RNA. Knowledge of RNA structures can answer basic scientific questions and can be of great help in design of new types of drugs and therapies. Structures can help answering the fundamental question of evolution whether life started with RNA as 'RNA World' (2) or other, perhaps peptide-type molecules. Rational drug design would without a doubt benefit from reliable predictions of RNA structures. Increasingly, the growing issue of bacterial drug resistance is approached from different perspectives but specific inhibition of ribosome particles offers a promising route to effective treatment (3). RNA therapies are attracting more attention from large pharmaceutical companies (4).

RNA building blocks, nucleotides, are chemically complex with aromatic nitrogenous bases, chiral ribose sugar rings and phosphate groups. The bases are able to stack on each other by van der Waals interactions, but they also carry large electrical moments and can form strong hydrogen bonds. Ribose rings strongly constrain backbone geometries by their puckers; the C3'-endo pucker prevails in RNA, but a ribose can also locally adopt the C2'-endo pucker, thus radically changing the backbone geometry. The phosphate groups are perhaps structurally the most complex parts of the RNA molecules due to d-orbitals in phosphorous atoms. Both torsion angles describing the conformations around the phosphodiester bonds O3'-P and P-O5' called ζ and α prefer -gauche orientations, but the torsions can adopt any other combinations of gauche, trans and -gauche (+60°, 180° and –60°) conformations. Phosphates in nucleic acids under normal conditions are charged and render whole RNA or DNA molecules strongly negative, which needs to be neutralized by interacting positive ions. The single negative charge of each phosphate is distributed between its unbound oxygen atoms that are highly polarizable and capable of forming hydrogen bonds to other RNA atoms,

proteins and water, but also of forming charge-charge interactions to amino acids, other cellular components such as amines and prominently also to metals. All intra- and inter-molecular interactions in which RNA molecules are involved determine their structures. Figure 1 illustrates at least some of these physically complex interactions as they were observed in a small six-nucleotide loop from an 80-nt fragment of rRNA from a crystal structure 4qvi (5).

RNA 3D STRUCTURE PREDICTION: STATE OF THE ART

In the 1960s, first attempts began to reconstruct *in silico* the 3D structures of RNA molecules based on sequence homology (6). These efforts became more frequent with a growing number of experimentally determined 3D RNA structures. Building *in silico* models relied largely on manual manipulation of structure templates in a computational environment. The first interactive tool targeting RNA tertiary structure modeling was published in 1998 (7). Several years later, systems that could fully or semi-automatically process from RNA sequence to a 3D model began to appear, using *ab initio* folding such as FARFAR (8), iFoldRNA (9), NAST (10), SimRNA (11) and Vfold (12); or homology modeling such as RNABuilder (13) and ModeRNA (14), or a fragment-based assembly approach used in MC-Fold/MC-Sym (15), Assemble (16), RNAComposer (17) and 3dRNA (18). In the past two years, deep learning (DL)-based predictive models have begun to emerge. The paper by Townshend *et al.* (19) presented a DL model that predicted the quality (RMSD) of a new computer-generated 3D RNA structure. Meanwhile, other works (20–22) described methods that used deep learning for the end-to-end 3D prediction of the RNA structure.

With the increasing availability of computer-based methods for predicting 3D RNA structures, the question of the reliability and quality of the generated models became more important. In response, RNA-Puzzles, a collective blind experiment to critically evaluate the prediction of 3D RNA structures, was started in 2010 (23). During the past 12 years, RNA-Puzzles organized 38 competitive challenges (24) and two dedicated projects—modeling structures from unknown Rfam families and untranslated region of SARS-CoV-2 (25). Within each, participants predicted the tertiary structure of a single RNA target. The predictions were evaluated mainly by comparing them with a reference structure, once the latter was published in the Protein Data Bank and the assessments for 34 challenges are currently known (data as of February 2023). Several similarity and distance measures were used for evaluation, some of which were specifically developed for RNA (26–30). For example, Interaction Network Fidelity (INF), a similarity measure, scores the prediction of base pairs, Watson–Crick (INF-WC), non-Watson–Crick (INF-NWC) and stacking (INF-stacking). As shown in Figure 2, during the 12 years of challenges in RNA-Puzzles, INF-WC generally ranged between 0.75 and 1.0, demonstrating that most models had accurately predicted double helical stem motifs (INF = 1 means ideal prediction and 0 is failure). However, INF-NWC scored close to 0 for most predictions, which is of concern since non-Watson–Crick base pairs play a crucial role in
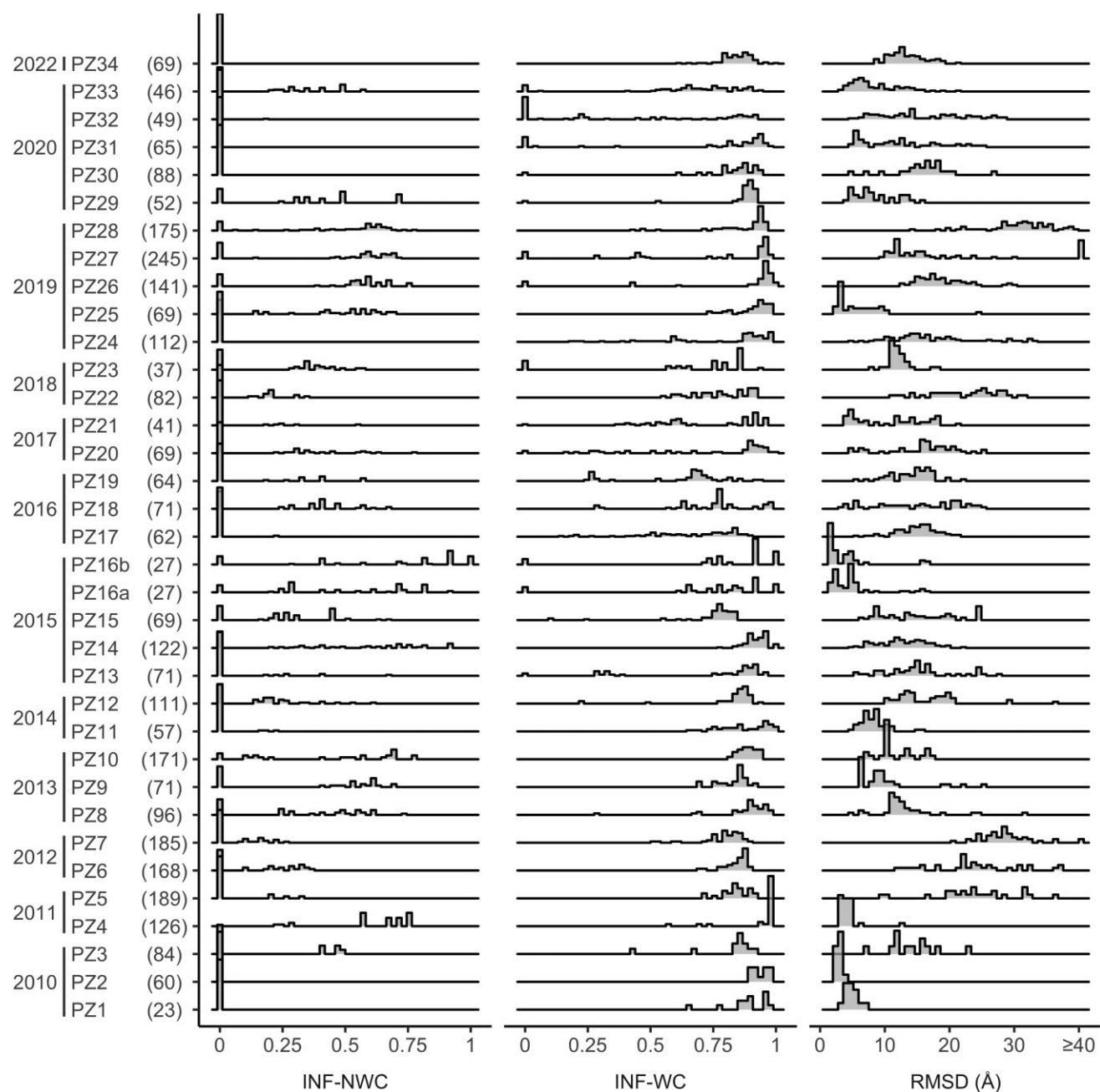
Figure 2. Distribution of values of selected evaluation measures for the predictions submitted to RNA-Puzzles from inception to 2022. Numbers in parentheses next to each puzzle indicate the total number of nucleotides for all structures in each puzzle.

determining the overall fold of the RNA, influencing stem packing and junction topologies. RMSD indicates how the predicted 3D coordinates diverge from those of the reference structure and shows only a few models with RMSD<5Å. For most RNA-Puzzles, the distribution of RMSD values is multimodal and spreads over a wide range. Therefore, despite significant advances in modeling approaches, predicting RNA coordinates with native-like features remains challenging and requires improvements in both accuracy and quality (31).

The RNA-Puzzles initiative has adopted many mechanisms that were developed in CASP, the biennial experiment

for the critical assessment of protein structure prediction. The first CASP competition was launched in 1994 (32), a quarter of a century after pioneering research into 3D computer modeling of protein structure began (33). Twenty-seven participating groups were challenged to predict the atomic coordinates of 33 amino acid sequences. In subsequent editions of CASP, the number of targets and participants increased (Figure 3), and new competition categories emerged. This included a fully automatic prediction by web servers, a category that started in 2000 (CASP4). Eighteen years later, AlphaFold (34) entered the game in CASP13 (35) to make a breakthrough in protein structure prediction
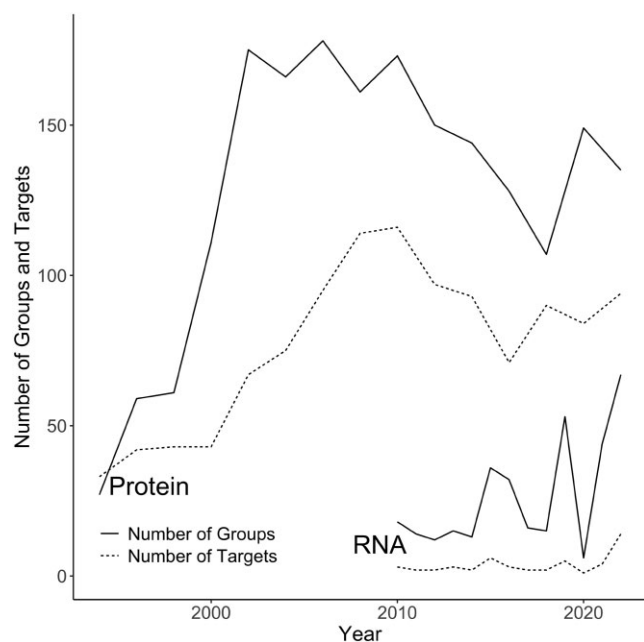
Figure 3. Numbers of RNA and protein structure predictions made in RNA-Puzzles and CASP competitions. The solid lines represent the numbers of groups competing in CASP and RNA-Puzzles; the dashed lines are for the number of protein/RNA targets. From 2010 to 2021, RNAs were predicted only in RNA-Puzzles and in 2022, CASP included also RNA targets, which is responsible for the recent spike in targets and groups involved in 3D RNA structure prediction.

in 2020 (CASP14) (36). RNA-Puzzles opened its own web server category in 2015. In 2022, this competition saw the first teams using deep learning models to predict 3D RNA structures. In the same year, CASP-RNA was launched, a contest co-organized by CASP and RNA-Puzzles (37). It coincided with an explosion of interest in the prediction of the 3D RNA structure (38) resulting, among other things, from the success of AlphaFold and the Covid-19 pandemic caused by an RNA virus. 42 groups participating in CASP-RNA tried their hand at modeling three-dimensional structures for 12 RNA sequences. Eighteen contributing teams used deep learning models (including Deep-FoldRNA, RhoFold, trRosettaRNA and OpenComplex-RNA) at various stages of prediction (20–22). The final CASP-RNA ranking gave the top 4 places to teams that combined expert modeling with non-machine learning algorithms.

THE CHALLENGES

AlphaFold and other highly accurate methods (34,39–48) applied deep learning to predict the protein structure based on the sequence. Training these tools required huge amounts of data. For example, AlphaFold implemented a bootstrap technique in which its final version used both experimentally determined and predicted structures of high accuracy. A fundamental question is whether we have enough RNA structure data for training and whether they are of sufficiently high quality and diversity.

RNA content in the Protein Data Bank

Since the first tRNA structures were solved in the mid-1970s (49) and published about ten years later (50,51) it was known that RNA molecules could adopt complex 3D architectures. However, it was not until the late 1990s that structures of functionally new types of RNA emerged: first several types of ribozymes (52–54), and then impressive ribosome particles (55–57). These revealed the structural richness of the RNA architectures, which was later confirmed by more structures solved mostly by X-ray crystallography and recently by cryo-electron microscopy (cryo-EM). Despite all the discoveries about RNA structures, the sheer volume of experimental structural data available for RNA and proteins remains strongly in favor of the latter (Table 1). There are about 25 times more protein depositions than RNA. The ratio is slightly more favorable for DNA, but even so, both nucleic acids account for <10% of the PDB archive, and this ratio has remained fairly stable over time. The situation is even more dramatic when restricted to high-resolution data: among X-ray and cryo-EM structures with a resolution better than 2.0 Å, proteins are about 100 times more abundant than RNA (Table 1). Considering all structures with resolution <3.0 Å, RNA nucleotides constitute only 2% of all residues (nucleotides and amino acids) (58,59). Unfortunately, these proportions cannot be expected to change quickly. Newly solved crystal and cryo-em structures tend to have a limited resolution. The reason is the inherent flexibility of RNA molecules that can be estimated, for instance, by factors B and R in the crystal phase; they are higher for RNAs than for proteins with comparable resolution. A limited number of high-resolution RNA structures is a severe constraint, as these structures are the source of the most reliable experimental information about the 3D structures, and some believe the only.

RNA architectures crucial for the global fold

The main architectural element of RNA is an antiparallel double helix of form A that constitutes approximately 60% of RNA in ribosome particles. The structure of this element is the easiest to identify and predict. The overall three-dimensional arrangement of a molecule results from the assembly of these helical regions. It is orchestrated by various types of 3D motifs such as sharp turns, loops, n-way junctions, coaxial stacking of duplexes and triple and quadruple helical regions (56,60). A junction consists of at least three helical regions arranged in a way that significantly influences the overall fold. There are three families of three-way junctions, which differ by the coaxial stacking pattern (60). For junctions with higher multiplicity, it becomes more complicated (61). The correct prediction of the junction topology and the resulting stem orientation is of utmost importance, but poses a significant challenge, as there are usually only single or no homologous junctions in experimental structures of RNA (62). All of the aforementioned regions often form between sequentially distant parts of the RNA molecule and are stabilized by non-Watson–Crick base pairs (NWC). Reliable information on structurally critical NWCs is necessary for the correct 2D/3D structural predictions. However, the collection of NWCs

Table 1. Numbers of all PDB-released structures (*) and residues in X-ray and cryo-EM structures (**) with high resolution (≤2.0 Å) over decades. In the first column, amino acids are abbreviated as AAs, and nucleotides as nts

	≤1980	1981–1990	1991–2000	2001–2010	2011–2022	Total	% of the total
Proteins (*)	78	634	12 121	43 205	108 677	164 715	91.57
AAs ≤2.0 Å (**)	5050	45 236	1 609 401	11 390 238	28 513 777	41 563 702	99.78
RNA (*)	2	23	306	1392	4488	6211	3.45
RNA nts ≤2.0 Å (**)	0	0	1270	5974	26 921	34 165	0.08
DNA (*)	1	91	1061	2009	5800	8962	4.98
DNA nts ≤2.0 Å	0	238	5430	15 730	38 107	59 505	0.14

in high-resolution PDB structures is not sufficient to infer their sequence and structural features (63). There are ∼34 thousand RNA nucleotides in high resolution (≤2.0 Å) crystal and cryo-EM structures, compared to ∼42 million amino acids; it is <0.1% of all PDB-deposited residues (Table 1).

3D modules are another group of crucial yet hard to predict motifs (64) (Figure 4). They are primarily defined by NWCs that form an intricate network of interactions. These networks remain coherent even in RNAs from different phylogenetic groups. 3D modules serve as loops, turns and foundations for protein-RNA or RNA–RNA interactions. Their accurate modeling is essential to catch the global RNA fold, but it is hardly possible due to the low amount of data available.

RNA architectures are also stabilized by interactions such as base-ribose hydrogen bonding, intramolecular interactions with charged phosphates, and coordination with metal ions. The roles of these interactions are even less understood than those of non-Watson–Crick base pairs.

Quality of experimental RNA data

Not only does the shortage of high-resolution structures complicates the accurate annotation of RNAs. There are problems with the quality of deposited RNA (and DNA) data that arise from the lack of community-accepted quality standards. They are related to base pairing, valence geometry and backbone geometry; their combination can lead to a flood of imprecisely and unreliably refined structures.

A formal description of base pairing is essential to build reliable 3D models. However, base pairing in public archives is not described reliably; it is often incomplete or incorrect. The programs used to assign base pair topology to 3D structures, such as MC-Annotate (66), RNAview (67), FR3D (68), ClaRNA (69), CompAnnotate (69), RNApdbee (70), bpRNA (71), baRNAba (72), BP-NET (73) and DSSR (74), often provide incomplete or conflicting information (manuscript in preparation). Therefore, comprehensive benchmarking must be performed along with a consistent update of public archives with topology data from the consensus algorithm(s).

Perhaps of lesser but existing importance for the prediction of large RNA structures is the inconsistency of targets used in the refinement of bond distances and angles. These valence geometry targets differ in various refinement programs, validation packages and the PDB, leading to confusion in the community. Therefore, an ELIXIR-led effort was undertaken by the Nucleic Acid Valence Geometry
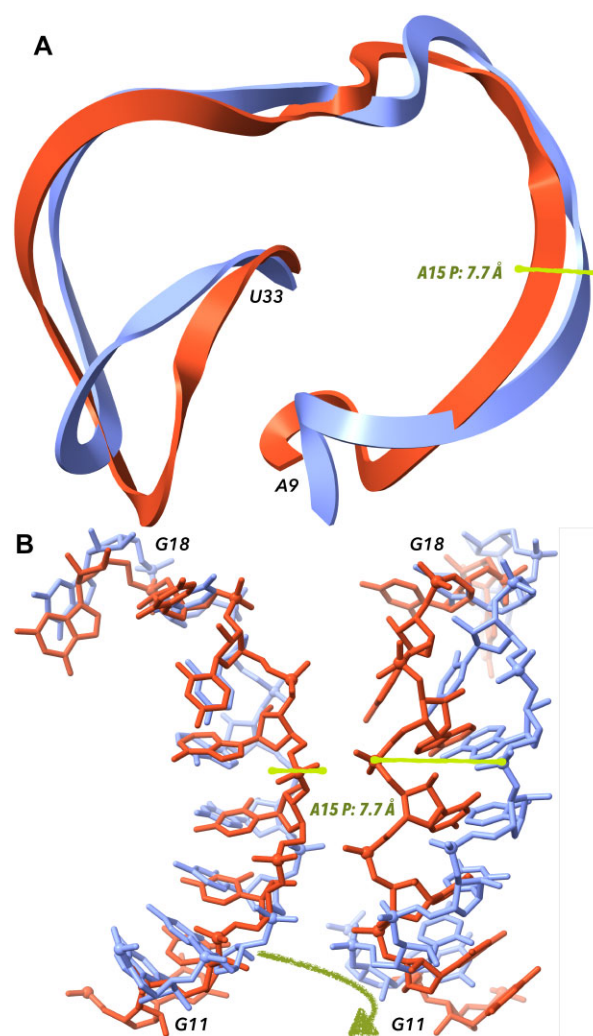
Figure 4. Comparison of predicted and experimentally determined structures. Displayed is hammerhead ribozyme RNA: the structure determined experimentally by X-ray diffraction at the 2.9 Å resolution (PDB ID 5di4) (65) is shown in light blue, the model PZ15_Adamiak_15 is in red. Cartoon representation of the residues A9-U33 in panel (**A**) suggests that the prediction follows the overall topology of the ribozyme correctly but with local deviations. Panel (**B**) shows segments between residues G11 and G18. The overall backbone direction is predicted correctly but local deviations are large. They include differences in base orientations and subsequently in base pairing and also the distances between the corresponding phosphorous atoms are quite large; one such distance between Ps of adenosines 15 of the target and model is highlighted by the green rod. Segments in panel B on the left and right show the same atoms, the view is rotated by ∼90°.

Working Group (75) to formulate community-agreed validation targets (76–78).

A significant source of errors in the structural description of RNA (and DNA) is the misconception about the geometry of the nucleic acid backbone. The structural complexity of the backbone was understood early on (79), but the topic attracted much less attention until the end of the 1990s. At that time, large RNA ribozyme and ribosome structures started to emerge and it became possible to analyze their structural variability based on experimental data. The smallest unit that makes sense to categorize structurally is a dinucleotide, which includes two riboses and captures the complexity of the phosphodiester linkage C3'–O3'–P–O5'–C5'. However, even this relatively small fragment has nine torsional degrees of freedom. The first conformer definitions of dinucleotide fragments were published at the beginning of 2000, first for RNA (80–82), later for DNA (83) and recently for both RNA and DNA as a structural alphabet CANA built from dinucleotide conformational classes NtC (84). Perhaps the relative novelty of the concept of conformational classes and technical difficulties with their implementation into routine refinement and validation protocols is the reason why the classes are not widely used. We see this fact as one of the reasons why the quality of newly determined structures does not improve.

Sequences and sequence alignments

The efficiency of 3D RNA structure prediction is likely to be improved using information from multiple sequence alignments (MSA). MSA has already been incorporated into several expert-based modeling methods in the human categories of RNA-Puzzles and CASP-RNA (24). Such a strategy is also applied in AlphaFold and other recent protein prediction methods. In these methods, correlated mutations are used to detect residues that are in close contact in 3D space, despite the distance in sequence. This principle has been understood for a long time in RNA (63). Unfortunately, creating high-quality RNA alignments is difficult and often requires the manual work of an expert. This difficulty has led to there being far fewer RNA vs. protein alignments.

To illustrate the difference in quantity, we can compare two resources, Pfam and Rfam. Pfam and Rfam are collections of protein/RNA alignments and models annotate them in genomes. Rfam is the oldest and largest source of alignments for ncRNAs. Although there are other resources that collect similar data, for example, miRBase (85) or MirGeneDB (86) for RNA, they are smaller and focus on one particular type of molecule. Pfam was founded in 1997 (87), while Rfam in 2003 (88). Each member of Rfam/Pfam is made up of a curated seed alignment which is used to build the model that allows finding more examples of the family and produces what is known as a full alignment. The models in Pfam are based on hidden Markov models, while in Rfam they are covariance models and also include a consensus secondary structure. Here, we will discuss some of the issues facing machine learning practitioners that want to use RNA alignments by comparing these resources.

First, while Rfam is similar to Pfam in spirit and goals, it contains far less data than Pfam. At the time of writing
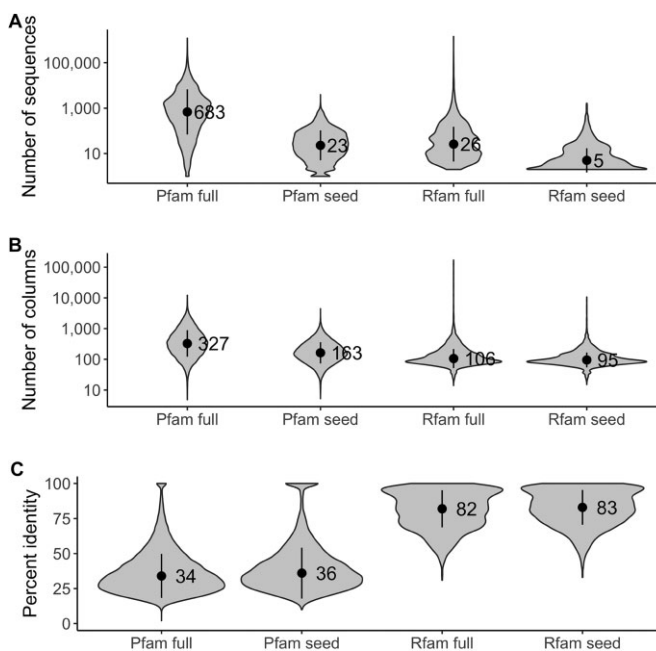
Figure 5. Rfam versus Pfam alignments compared based on (**A**) a number of sequences, (**B**) a number of columns and (**C**) the average pairwise percent identity for each family. The points on the plots indicate the mean, and the vertical bars indicate the standard deviation.

this paper, the current version of Rfam, 14.9, contains 4108 alignments, while the current release of Pfam, 35.0, contains 19 632. The difference in resource size is due to historical bias towards RNA gene discovery, the difficulty in identifying homology between related RNAs, and the difficulty in building new alignments for Rfam. Constructing Rfam alignments requires using covariance models, which are much more computationally expensive compared to the hidden Markov models applied to build Pfam alignments.

Second, RNA alignments are on average smaller than protein alignments. This relationship relates to the number of sequences, with seed alignments containing an average of 5 sequences in Rfam versus 23 in Pfam (Figure 5A), as well as the number of columns, 95 columns in Rfam versus 163 in Pfam (Figure 5B). There is also a significant difference in the degree of conservation, with the Rfam alignments 83% conserved versus 26% in Pfam (Figure 5C). Together, it means that there are few RNA alignments compared to proteins, and the existing alignments are smaller and lack variation. Therefore, it is likely that there is not enough RNA data yet to effectively train machine learning methods. This is also supported by the fact that the currently best-performing RNA-dedicated methods in CASP are not machine learning based.

Third, Rfam alignments have several global biases that make working with them difficult. One is that the most common alignments are for simple molecules. Taking into account the type of RNA, most alignments concern miRNA precursors (35%) followed by snoRNA (19%) (Figure 6). miRNA precursors are simple molecules, essentially a helix with a few small loops and mismatches; in proteins, this is most similar to a single alpha helix. Such simple structures do not represent the complexity of RNA folds; for
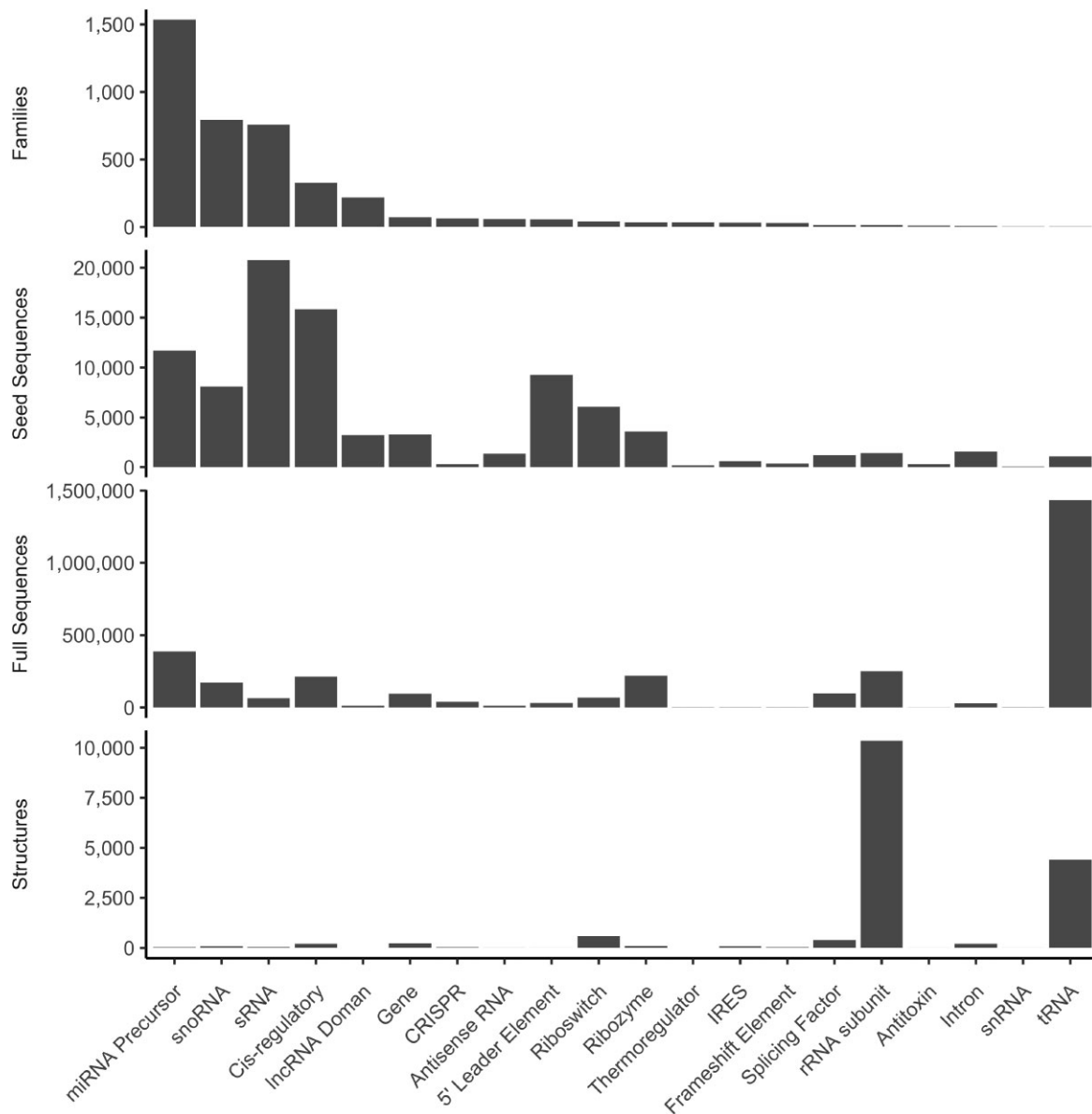
Figure 6. Counts of Rfam families, seed sequences, full sequences and structures for all Rfam families organized by Rfam RNA type.

example, they do not contain any junctions, while—as discussed above—the junction topology is essential to determine the overall structure of more complex RNAs.

Another global bias is observed in the number of seed or full sequences, Rfam has the most data for bacterial small RNA (sRNA) sequences. However, there are few structures of these molecules with <50 in PDB at the time of writing. In terms of full alignments, tRNAs constitute the largest group (45%), and rRNA subunits are the third largest, accounting for another 8% (Figure 6). These families are the most commonly solved structures, representing 26% and 61% of all known 3D structures of RNA, respectively (Figure 6). Although a large collection of these sequences and structures is valuable, we recommend caution. Creating ML models that generalize to other structures is unlikely if their training is based only on ribosomes. Several predic-

tion methods that train off currently existing datasets have not yet produced high-quality models.

In addition to the global bias in the RNA data, there are specific issues with Rfam alignments that must be considered in machine learning. For example, not all non-Watson–Crick base pairs are aligned in Rfam, and the aligned ones have not been handled in a consistent manner. Moreover, Rfam consensus secondary structures can represent parts of the structure as unfolded. However, looking at the 3D structure, when available, in that region often shows a clear secondary structure. These regions include places known to have species-specific structure or their unstructured form results from Rfam limitations. Rfam families are intended to cover a wide phylogenetic range. For example, the eukaryotic large subunit ribosomal RNA family (RF02543) represents all large rRNA subunits in all eukaryotes.

However, rRNA is well known to vary considerably within the kingdom, or even within a species, with important functional consequences (89). Since the 2D structures in Rfam must represent what is common to all members of the family, they are often underfolded in many regions. This should be dealt with when building a useful ML training set. Finally, pseudoknots—a key factor in 3D RNA structures—have been shown to help organize the global structure, but are not consistently annotated in Rfam alignments. Unfortunately, current 2D and 3D prediction methods struggle to predict them. Rfam is working to annotate more observed pseudoknots but many families lack them.

In summary, there are several issues with the RNA alignment dataset that will pose problems for deep learning. The data set is small compared to proteins, is highly biased in several ways, and the existing alignments have some shortcomings. While work is ongoing to fix all these issues, it will be challenging to use these data to successfully predict 3D structures. One key issue will be creating a test/train dataset that represents the observed complexity, while not being overly biased.

CONCLUSIONS

Given the history of protein fold prediction, can we anticipate when the RNA realm will see similar results? AlphaFold's success came 50 years after the first work on computer-based protein structure prediction. This period of time was necessary to accumulate a sufficient volume of high-quality, reliable data on protein sequences and structures. At the same time, information and computer technology were developed, enabling efficient applications of artificial intelligence models to solve problems that traditional computational methods could not deal with. Artificial neural networks as an idea are already 80 years old (90), but it was only in the second decade of the 21st century that they came into widespread use. In 2012, the power of deep learning was demonstrated (91,92). It has triggered a flood of projects that have applied DL models to various areas of life. Among other things, this wave has brought about new predictive methods dedicated to molecular structures. All of them are data-hungry; AlphaFold has been trained on structures of more than 170,000 proteins combined with very large sequence alignments. We expect to have similar requirements to successfully use neural networks for RNA 3D structure prediction.

A simple way to estimate when AlphaFold for RNA will be created is to consider when the number of RNA structures or sequence alignments are comparable to the currently available protein data. As mentioned above Pfam contains 19 632 protein sequence alignments. Historically, the growth of Rfam has been linear due to the requirement for manual work to build each alignment and we observe that on average Rfam adds approximately 205 alignments per year. Thus, we estimate Rfam will contain 19 000 alignments in approximately 70 years. This is undoubtedly a vast overestimate as we expect the RNA 3D structure prediction problem to be solved by then. One technique which may help is automatic family building. While this is still unsolved for RNA, there has been recent work on this issue which

may be promising (93). Automatically built families were used in training AlphaFold and may prove useful for RNA as well (34).

We believe that there are several viable approaches to enable the prediction of the 3D RNA structure in the near future. First, the RNA community can improve knowledge of RNA structure through more data, second, we can diversify the data used in prediction, and finally, we can improve the machine learning methods used.

What data is missing that would improve predictions? We do not seem to know enough about RNA motifs to predict their global structures. We may provide an educated guess, at least for the small structural motifs, of which the most important are base-pair topologies. Concerning the latter, it is very likely that they exist in known structures of reasonably high resolution and can provide reliable geometries. There are also strong reasons to believe that the CANA alphabet describes more than 90% of the existing dinucleotide conformers; only a few of them may be missing (84). In our opinion, more research is needed on intramolecular interactions other than base pairs, namely hydrogen bonding bridges of the O2' group to bases, ribose, phosphates and interactions between phosphate oxygens (mostly charged) and other RNA constituents. Benchmarking the quality of 3D structures, as well as streamlined and consistent principles of their validation, is required to ensure reliability in data repositories.

Another approach is to improve the size and scope of multiple sequence alignments of RNA. Alignments of four-letter RNA sequences are more challenging than those of 20-letter protein sequences. Some classes of RNA, such as ribosomes, have a large number of sequences and we know how to align them. However, more well-aligned sequences of underrepresented RNA classes are needed. Perhaps the Tree-of-Life projects (94,95) will provide a sufficiently large number of sequences. Currently, RNA gene prediction is inconsistent across known genomes, so we encourage the community to annotate ncRNA genes in newly sequenced genomes. Annotated ncRNAs from Tree of Life projects can show low sequence diversity, and we recommend that ncRNA gene annotation in metagenomes be used as a solution. We note that AlphaFold required metagenomic sequences in order to reach its maximum performance, and we suspect that RNA will show a similar trend. Solving these challenges involves finding all the ncRNA genes and making the data reusable.

Consistently annotating RNA families across all genomes will be useful and may increase the diversity of RNA sequences available; however, it seems that a prediction method would benefit from a wider range of RNA families. As discussed above, many Rfam families are structurally similar. We believe that providing a more diverse training set would be useful. While Rfam is the global repository of RNA families, not all known families can be found there. Correcting this and working to create new families that are different from existing ones should be a focus of the RNA community. Additionally, creating high-quality alignments remains a challenge (96).

If the current amount and growth rate of currently available sequence and structure data are not sufficient, can they be supplemented with other sources of data? We think so.

In particular, RNA biochemistry has a rich history and has developed many methods to rapidly probe 3D structures (97,98). A subset of these data, SHAPE probing, has proven useful to classical prediction methods, and we expect it to be helpful to DL-based approaches. Although many labs probe the structure of RNA, these data are not readily available to ML practitioners. Working as a community to standardize, collect and distribute such data seems valuable for predictions. Additionally, there are other low resolution methods, such as SAXS and AFM, which may prove useful in modelling structures (97).

Finally, the rapid and hard-to-predict development of ML methods may potentially change our pessimistic predictions about the ability to accurately predict 3D RNA structures. Development of methods that are less data hungry, e.g. transfer learning, may allow successful prediction sooner. We believe that RNA structure prediction is an excellent test case for researchers interested in machine learning in the face of limited data. At the moment, we do not believe that reliable 3D RNA prediction will be available in the 2020s, but we challenge the community to prove us wrong.

DATA AVAILABILITY

The data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.8167407.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **46**, D8–D13.
2. Cech,T.R., Steitz,J.A. and Atkins,J.F. (2019) In: *RNA worlds: New tools for deep exploration*. Cold Spring Harbor Laboratory Press, NY.
3. Matzov,D., Bashan,A. and Yonath,A. (2017) A bright future for antibiotics? *Ann. Rev. Biochem.*, **86**, 567–583.
4. n.a. (2023) Big pharma craves slice of AI-based RNA drug discovery. *Nat. Biotechnol.*, **41**, 305.
5. Tishchenko,S., Kostareva,O., Gabdulkhakov,A., Mikhaylina,A., Nikonova,E., Nevskaya,N., Sarskikh,A., Piendl,W., Garber,M. and Nikonov,S. (2015) Protein–RNA affinity of ribosomal protein L1 mutants does not correlate with the number of intermolecular interactions. *Acta Crystallogr. D*, **71**, 376–386.
6. Levitt,M. (1969) Detailed molecular model for transfer ribonucleic acid. *Nature*, **224**, 759–763.
7. Massire,C. and Westhof,E. (1998) MANIP: an interactive tool for modelling RNA. *J. Mol. Graph. Model*, **16**, 197–205.
8. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14664–14669.
9. Sharma,S., Ding,F. and Dokholyan,N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.
10. Jonikas,M.A., Radmer,R.J., Laederach,A., Das,R., Pearlman,S., Herschlag,D. and Altman,R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
11. Boniecki,M.J., Lach,G., Dawson,W.K., Tomala,K., Lukasz,P., Soltysinski,T., Rother,K.M. and Bujnicki,J.M. (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.*, **44**, e63.
12. Zhao,C., Xu,X. and Chen,S.-J. (2017) Predicting RNA structure with Vfold. *Methods Mol. Biol.*, **1654**, 3–15.
13. Flores,S.C., Wan,Y., Russell,R. and Altman,R.B. (2010) Predicting RNA structure by multiple template homology modeling. *Pac. Symp. Biocomput.*, 216–227.
14. Rother,M., Rother,K., Puton,T. and Bujnicki,J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.*, **39**, 4007–4022.
15. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
16. Jossinet,F., Ludwig,T.E. and Westhof,E. (2010) Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics*, **26**, 2057–2059.
17. Popenda,M., Szachniuk,M., Antczak,M., Purzycka,K.J., Lukasiak,P., Bartol,N., Blazewicz,J. and Adamiak,R.W. (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res.*, **40**, e112.
18. Zhao,Y., Huang,Y., Gong,Z., Wang,Y., Man,J. and Xiao,Y. (2012) Automated and fast building of three-dimensional RNA structures. *Sci. Rep.*, **2**, 734.
19. Townshend,R. J.L., Eismann,S., Watkins,A.M., Rangan,R., Karelina,M., Das,R. and Dror,R.O. (2021) Geometric deep learning of RNA structure. *Science*, **373**, 1047–1051.
20. Ramakers,J., Blum,C.F., König,S., Harmeling,S. and Kollmann,M. (2021) *De Novo* prediction of RNA 3D structures with Deep Learning. bioRxiv doi: https://doi.org/10.1101/2021.08.30.458226, 01 September 2021, preprint: not peer reviewed.
21. Pearce,R., Omenn,G.S. and Zhang,Y. (2022) De novo RNA tertiary structure prediction at atomic resolution using geometric potentials from Deep Learning. bioRxiv doi: https://doi.org/10.1101/2022.05.15.491755, 15 May 2022, preprint: not peer reviewed.
22. Shen,T., Hu,Z., Peng,Z., Chen,J., Xiong,P., Hong,L., Zheng,L., Wang,Y., King,I., Wang,S. *et al.* (2022) E2Efold-3D: end-to-end deep learning method for accurate de novo RNA 3D structure prediction. arXiv doi: https://arxiv.org/abs/2207.01586, 04 July 2022, preprint: not peer reviewed.
23. Cruz,J.A., Blanchet,M.-F., Boniecki,M., Bujnicki,J.M., Chen,S.-J., Cao,S., Das,R., Ding,F., Dokholyan,N.V., Flores,S.C. *et al.* (2012) RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **18**, 610–625.
24. Miao,Z., Adamiak,R.W., Antczak,M., Boniecki,M.J., Bujnicki,J., Chen,S.-J., Cheng,C.Y., Cheng,Y., Chou,F.-C., Das,R. *et al.* (2020) RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, **26**, 982–995.
25. Gumna,J., Antczak,M., Adamiak,R.W., Bujnicki,J.M., Chen,S.-J., Ding,F., Ghosh,P., Li,J., Mukherjee,S., Nithin,C. *et al.* (2022)

Computational pipeline for reference-free comparative analysis of RNA 3D structures applied to SARS-CoV-2 UTR models. *Int. J. Mol. Sci.*, **23**, 9630.

26. Parisien,M., Cruz,J.A., Westhof,E. and Major,F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.

27. Zok,T., Popenda,M. and Szachniuk,M. (2014) MCQ4Structures to compute similarity of molecule structures. *Cent. Eur. J. Oper Res.*, **22**, 457–473.

28. Wiedemann,J., Zok,T., Milostan,M. and Szachniuk,M. (2017) LCS-TA to identify similar fragments in RNA 3D structures. *BMC Bioinformatics*, **18**, 456.

29. Gong,S., Zhang,C. and Zhang,Y. (2019) RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics*, **35**, 4459–4461.

30. Magnus,M., Antczak,M., Zok,T., Wiedemann,J., Lukasiak,P., Cao,Y., Bujnicki,J.M., Westhof,E., Szachniuk,M. and Miao,Z. (2020) RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Res.*, **48**, 576–588.

31. Carrascoza,F., Antczak,M., Miao,Z., Westhof,E. and Szachniuk,M. (2022) Evaluation of the stereochemical quality of predicted RNA 3D models in the RNA-Puzzles submissions. *RNA*, **28**, 250–262.

32. Moult,J., Pedersen,J.T., Judson,R. and Fidelis,K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii–v.

33. Scheraga,H.A. (1969) Calculation of polypeptide conformation. *Harvey Lect.*, **63**, 99–138.

34. 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.

35. AlQuraishi,M. (2019) AlphaFold at CASP13. *Bioinformatics*, **35**, 4862–4865.

36. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Applying and improving AlphaFold at CASP14. *Proteins: Struct. Funct. Bioinformatics*, **89**, 1711–1721.

37. Kryshtafovych,A., Antczak,M., Szachniuk,M., Zok,T., Kretsch,R.C., Rangan,R., Pham,P., Das,R., Robin,X., Studer,G. *et al.* (2023) New prediction categories in CASP15. *Proteins: Struct. Funct. Bioinform.*, **91**, 1–8.

38. Zhang,J., Fei,Y., Sun,L. and Zhang,Q.C. (2022) Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat. Methods*, **19**, 1193–1207.

39. Wang,S., Sun,S., Li,Z., Zhang,R. and Xu,J. (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, **13**, e1005324.

40. Adhikari,B., Hou,J. and Cheng,J. (2018) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, **34**, 1466–1472.

41. Hou,J., Wu,T., Cao,R. and Cheng,J. (2019) Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins*, **87**, 1165–1178.

42. Du,Z., Su,H., Wang,W., Ye,L., Wei,H., Peng,Z., Anishchenko,I., Baker,D. and Yang,J. (2021) The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.*, **16**, 5634–5651.

43. Kandathil,S.M., Greener,J.G., Lau,A.M. and Jones,D.T. (2022) Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2113348119.

44. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S. and Steinegger,M. (2022) ColabFold: making protein folding accessible to all. *Nat. Methods*, **19**, 679–682.

45. Zhang,X., Zhang,B., Freddolino,P.L. and Zhang,Y. (2022) CR-I-TASSER: assemble protein structures from cryo-EM density maps using deep convolutional neural networks. *Nat. Methods*, **19**, 195–204.

46. Chowdhury,R., Bouatta,N., Biswas,S., Floristean,C., Kharkar,A., Roy,K., Rochereau,C., Ahdritz,G., Zhang,J., Church,G.M. *et al.* (2022) Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.*, **40**, 1617–1623.

47. Ferruz,N., Schmidt,S. and Höcker,B. (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.*, **13**, 4348.

48. Brandes,N., Ofer,D., Peleg,Y., Rappoport,N. and Linial,M. (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, **38**, 2102–2110.

49. Suddath,F.L., Quigley,G.J., McPherson,A., Sneden,D., Kim,J.J., Kim,S.H. and Rich,A. (1974) Three-dimensional structure of yeast phenylalanine transfer RNA at 3.0angstroms resolution. *Nature*, **248**, 20–24.

50. Brown,R.S., Dewan,J.C. and Klug,A. (1985) Crystallographic and biochemical investigation of the lead(II)-catalyzed hydrolysis of yeast phenylalanine tRNA. *Biochemistry*, **24**, 4785–4801.

51. Westhof,E., Dumas,P. and Moras,D. (1988) Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Crystallogr. A*, **44**, 112–123.

52. Tuschl,T., Gohlke,C., Jovin,T.M., Westhof,E. and Eckstein,F. (1994) A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science*, **266**, 785–789.

53. Pley,H.W., Flaherty,K.M. and McKay,D.B. (1994) Three-dimensional structure of a hammerhead ribozyme. *Nature*, **372**, 68–74.

54. 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.

55. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**, 905–920.

56. Tocilj,A., Schlünzen,F., Janell,D., Glühmann,M., Hansen,H.A., Harms,J., Bashan,A., Bartels,H., Agmon,I., Franceschi,F. *et al.* (1999) The small ribosomal subunit from Thermus thermophilus at 4.5 A resolution: pattern fittings and the identification of a functional site. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 14252–14257.

57. Wimberly,B.T., Brodersen,D.E., Clemons,W.M., Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T. and Ramakrishnan,V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.

58. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Costanzo,L.D., Christie,C., Dalenberg,K., Duarte,J.M., Dutta,S. *et al.* (2018) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.

59. Adamczyk,B., Antczak,M. and Szachniuk,M. (2022) RNAsolo: a repository of cleaned PDB-derived RNA 3D structures. *Bioinformatics*, **38**, 3668–3670.

60. Lescoute,A. and Westhof,E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83–93.

61. Laing,C. and Schlick,T. (2009) Analysis of four-way junctions in RNA structures. *J Mol. Biol.*, **390**, 547–559.

62. Wiedemann,J., Kaczor,J., Milostan,M., Zok,T., Blazewicz,J., Szachniuk,M. and Antczak,M. (2022) RNAloops: a database of RNA multiloops. *Bioinformatics*, **38**, 4200–4205.

63. Stombaugh,J., Zirbel,C.L., Westhof,E. and Leontis,N.B. (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.*, **37**, 2294–2312.

64. Leontis,N.B. and Westhof,E. (1998) A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *J. Mol. Biol.*, **283**, 571–583.

65. Mir,A., Chen,J., Robinson,K., Lendy,E., Goodman,J., Neau,D. and Golden,B.L. (2015) Two divalent metal ions and conformational changes play roles in the hammerhead ribozyme cleavage reaction. *Biochemistry*, **54**, 6369–6381.

66. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.

67. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.

68. Sarver,M., Zirbel,C.L., Stombaugh,J., Mokdad,A. and Leontis,N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.

69. Walen,T., Chojnowski,G., Gierski,P. and Bujnicki,J.M. (2014) ClaRNA: a classifier of contacts in RNA 3D structures based on a

comparative analysis of various classification schemes. *Nucleic Acids Res.*, **42**, e151.

70. Zok,T., Antczak,M., Zurkowski,M., Popenda,M., Blazewicz,J., Adamiak,R.W. and Szachniuk,M. (2018) RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res.*, **46**, W30–W35.

71. 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.

72. Bottaro,S., Bussi,G., Pinamonti,G., Reißer,S., Boomsma,W. and Lindorff-Larsen,K. (2019) Barnaba: software for analysis of nucleic acid structures and trajectories. *RNA*, **25**, 219–231.

73. Roy,P. and Bhattacharyya,D. (2022) Contact networks in RNA: a structural bioinformatics study with a new tool. *J. Comput. Aided Mol. Des.*, **36**, 131–140.

74. Lu,X.-J., Bussemaker,H.J. and Olson,W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142.

75. Schneider,B., Bruno,I., Burley,S.K., Case,D.A., Černý,J., Das,R., Egli,M., Emsley,P., Feng,Z., Jaskolski,M. *et al.* (2020) Nucleic acid valence geometry working group. *Int. Union Crystallogr. Newslett.*, **28**, https://www.iucr.org/news/newsletter/volume-28-number-4/nucleic-acid-valence-geometry-working-group.

76. Kowiel,M., Brzezinski,D. and Jaskolski,M. (2016) Conformation-dependent restraints for polynucleotides: I. Clustering of the geometry of the phosphodiester group. *Nucleic Acids Res.*, **44**, 8479–8489.

77. Gilski,M., Zhao,J., Kowiel,M., Brzezinski,D., Turner,D.H. and Jaskolski,M. (2019) Accurate geometrical restraints for Watson–Crick base pairs. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.*, **75**, 235–245.

78. Kowiel,M., Brzezinski,D., Gilski,M. and Jaskolski,M. (2020) Conformation-dependent restraints for polynucleotides: the sugar moiety. *Nucleic Acids Res.*, **48**, 962–973.

79. Kim,S.-H., Berman,H.M., Seeman,N.C. and Newton,M.D. (1973) Seven basic conformations of nucleic acid structural units. *Acta Crystallogr. B*, **29**, 703–710.

80. Murray,L. J.W., Arendall 3rd,W.B., Richardson,D.C. and Richardson,J.S. (2003) RNA backbone is rotameric. *Proc. Natl. Acad. Sci. U.S.A.*, 13904–13909.

81. Hershkovitz,E., Tannenbaum,E., Howerton,S.B., Sheth,A., Tannenbaum,A. and Williams,L.D. (2003) Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. *Nucleic Acids Res.*, **31**, 6249–6257.

82. Schneider,B., Morávek,Z. and Berman,H.M. (2004) RNA conformational classes. *Nucleic Acids Res.*, **32**, 1666–1677.

83. Svozil,D., Kalina,J., Omelka,M. and Schneider,B. (2008) DNA conformations and their sequence preferences. *Nucleic Acids Res.*, **36**, 3690–3706.

84. Černý,J., Božíková,P., Svoboda,J. and Schneider,B. (2020) A unified dinucleotide alphabet describing both RNA and DNA structures. *Nucleic Acids Res.*, **48**, 6367–6381.

85. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.

86. Fromm,B., Domanska,D., Høye,E., Ovchinnikov,V., Kang,W., Aparicio-Puerta,E., Johansen,M., Flatmark,K., Mathelier,A., Hovig,E. *et al.* (2020) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.*, **48**, D132–D141.

87. Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

88. 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.

89. Rothschild,D., Susanto,T.T., Spence,J.P., Genuth,N.R., Sinnott-Armstrong,N., Pritchard,J.K. and Barna,M. (2023) A comprehensive rRNA variation atlas in health and disease. bioRxiv doi: https://doi.org/10.1101/2023.01.30.526360 , 02 February 2023, preprint: not peer reviewed.

90. McCulloch,W.S. and Pitts,W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*, **5**, 115–133.

91. Krizhevsky,A., Sutskever,I. and Hinton,G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira,F., Burges,C.J., Bottou,L. and Weinberger,K.Q. (eds). *Advances in Neural Information Processing Systems*. Vol. **25**.

92. Dean,J., Corrado,G., Monga,R., Chen,K., Devin,M., Mao,M., Ranzato,M., Senior,A., Tucker,P., Yang,K. *et al.* (2012) Large scale distributed deep networks. *Adv. Neural. Inf. Process Syst.*, **25**, 1223–1231.

93. Zhang,C., Zhang,Y. and Pyle,A.M. (2023) rMSA: a sequence search and alignment algorithm to improve RNA structure modeling. *J. Mol. Biol.*, **435**, 167904.

94. Darwin Tree of Life Project Consortium (2022) Sequence locally, think globally: The Darwin Tree of Life Project. *Proc. Natl. Acad. Sci. USA*, **119**, e2115642118.

95. Gupta,P.K. (2022) Earth Biogenome Project: present status and future plans. *Trends Genet.*, **38**, 811–820.

96. Gao,W., Yang,A. and Rivas,E. (2022) Thirteen dubious ways to detect conserved structural RNAs. *IUBMB Life*, **75**, 471–492.

97. Ponce-Salvatierra,A., Merdas Astha,K., Nithin,C., Ghosh,P., Mukherjee,S. and Bujnicki,J.M. (2019) Computational modeling of RNA 3D structure based on experimental data. *Biosci. Rep.*, **39**, BSR20180430.

98. Spitale,R.C. and Incarnato,D. (2023) Probing the dynamic RNA structurome and its functions. *Nat. Rev. Genet.*, **24**, 178–196.