nature methods

Article

https://doi.org/10.1038/s41592-024-02487-0

Accurate RNA 3D structure prediction using a language model-based deep learning approach

Received: 31 January 2024

Accepted: 25 September 2024

Published online: 21 November 2024

Check for updates

Tao Shen[1,2,3,17], Zhihang Hu[1,17], Siqi Sun[4,5,17] ✉, Di Liu[6,7,8,9,17] ✉, Felix Wong[10,11,12,13], Jiuming Wang[1,14], Jiayang Chen[1], Yixuan Wang[1], Liang Hong[1], Jin Xiao[1], Liangzhen Zheng[2,3], Tejas Krishnamoorthi[15], Irwin King[1], Sheng Wang[2,3] ✉, Peng Yin[6,7] ✉, James J. Collins[6,10,11,12] ✉ & Yu Li[1,6,10,11,16,17] ✉

Accurate prediction of RNA three-dimensional (3D) structures remains an unsolved challenge. Determining RNA 3D structures is crucial for understanding their functions and informing RNA-targeting drug development and synthetic biology design. The structural flexibility of RNA, which leads to the scarcity of experimentally determined data, complicates computational prediction efforts. Here we present RhoFold+, an RNA language model-based deep learning method that accurately predicts 3D structures of single-chain RNAs from sequences. By integrating an RNA language model pretrained on ~23.7 million RNA sequences and leveraging techniques to address data scarcity, RhoFold+ offers a fully automated end-to-end pipeline for RNA 3D structure prediction. Retrospective evaluations on RNA-Puzzles and CASP15 natural RNA targets demonstrate the superiority of RhoFold+ over existing methods, including human expert groups. Its efficacy and generalizability are further validated through cross-family and cross-type assessments, as well as time-censored benchmarks. Additionally, RhoFold+ predicts RNA secondary structures and interhelical angles, providing empirically verifiable features that broaden its applicability to RNA structure and function studies.

RNA molecules occupy a key role in the central dogma of molecular biology. How RNA structures impinge on gene regulation and function has been a subject of intense study[1]. Studies focusing on RNA targeting have demonstrated that it can be an important, druggable target for drug development[2–4] and a useful synthetic biology design element[5]. Over 85% of the human genome is transcribed, but a mere 3% encodes proteins, underscoring the substantial portion of transcribed RNAs with unknown functions and structures. In many cases, obtaining high-resolution structural information can enable a more predictive understanding of the RNA molecules of interest[4,6].

The conformational flexibility of RNA molecules has made the experimental determination of their three-dimensional (3D) structures challenging. As of December 2023, RNA-only structures comprise less than 1.0% of the ~214,000 structures in the Protein Data Bank (PDB), and RNA-containing complexes account for only 2.1% (refs. 6,7). Despite advances in X-ray crystallography, NMR spectroscopy and cryogenic electron microscopy, these low-throughput techniques are limited by specialized requirements. Computational methods have emerged as a complementary approach for RNA 3D structure prediction, leveraging RNA sequence data. These methods fall into two main categories:

A full list of affiliations appears at the end of the paper. ✉e-mail: siqisun@fudan.edu.cn; di.liu@asu.edu; wangsheng@zelixir.com; peng_yin@hms.harvard.edu; jimjc@mit.edu; liyu@cse.cuhk.edu.hk

https://doi.org/10.1038/s41592-024-02487-0

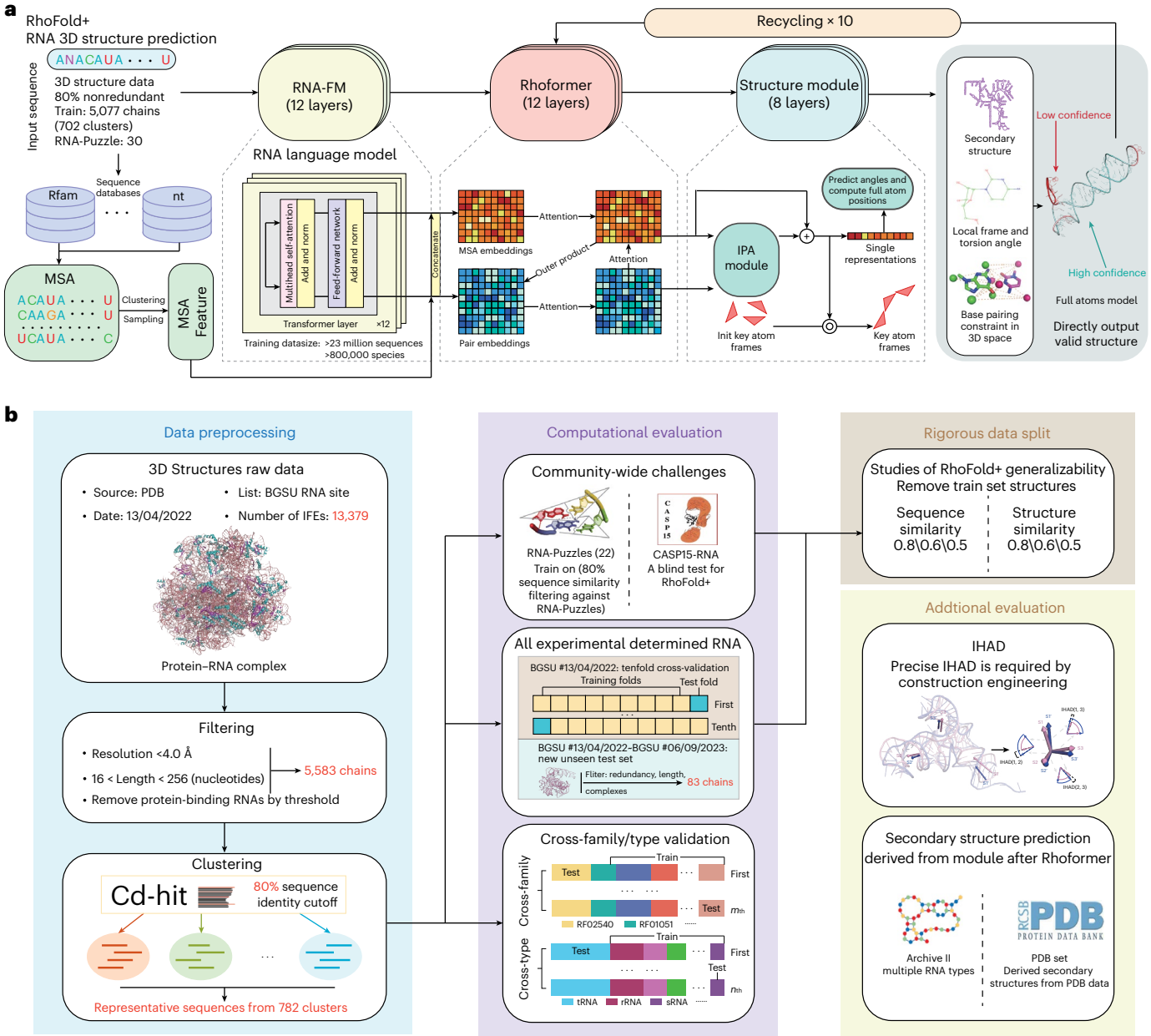


Fig. 1 | The architecture of RhoFold+ and the tasks used for performance evaluation. a, The architecture of RhoFold+, a fully automated and differentiable end-to-end approach to de novo RNA 3D structure prediction from the sequence. Using an RNA language model (RNA-FM) pretrained on 23,735,169 unannotated RNA sequences and several deep learning modules—including an IPA module that models 3D positions—RhoFold+ can generate valid and largely accurate RNA 3D structures of interest typically within ~0.14 s (without MSA searching). init, initialized; norm, normalize. **b**, The preprocessing step of RhoFold+ to extract all available nonredundant single-stranded RNA 3D structures from the PDB database. IFE, integrated functional element. RhoFold+ is comprehensively benchmarked on community-wide challenges including RNA-Puzzles targets and CASP15 natural RNA targets, and on all available experimentally determined RNA 3D structures. RhoFold+ also demonstrates high accuracy in cross-validation experiments, as well as generalizability to unseen, newly determined RNA structures and unseen RNA families and types in cross-family and cross-type validation experiments. Data split evaluations reveal that RhoFold+ does not overfit its training set. RhoFold+ is also capable of predicting secondary structures and parameters that are useful for construct engineering.

template-based modeling, such as ModeRNA[8] and RNAbuilder[9], which are constrained by limited template libraries, and de novo prediction approaches, including FARFAR2 (ref. 10), 3dRNA[11] and SimRNA[12], which are more predictive but computationally intensive due to large-scale sampling requirements.

An orthogonal de novo prediction approach is to leverage deep learning, which has been successfully applied to various biological problems. These applications include predicting protein 3D structures[13], RNA secondary structures[14,15] and scoring RNA structures generated by other methods[16]. Previous methods for RNA 3D structure prediction focused on template-based or energy-based sampling techniques, which were informed by the scarcity of available RNA 3D structural data. Despite the scarcity of data, the success of AlphaFold2 (ref. 13) for protein structure prediction has catalyzed the development of de novo deep learning methods for RNA 3D structure prediction. These de novo methods often begin with a single input sequence and then construct multiple sequence alignments (MSAs) from it, which are subsequently used to build the 3D structures.

MSAs have been shown to provide additional information helpful for protein modeling and this may be similarly true for RNAs.

For instance, DeepFoldRNA[17] and trRosettaRNA[18] utilize transformer networks (for example, RNAformer) to convert built MSAs and predicted secondary structures into various one-dimensional (1D) and two-dimensional (2D) distances, orientations and torsion angles. These predicted geometries are then leveraged as constraints to predict RNA 3D structures using energy minimization, integrating sampling and scoring processes into their frameworks. Several models, including E2Efold-3D[19] and RoseTTAFoldNA[20], employ fully differentiable end-to-end pipelines that directly predict all-atom 3D models using built MSAs and secondary structure constraints. AlphaFold3 (ref. 21), the successor to AlphaFold2 (ref. 22), is also capable of predicting RNA 3D structures directly from input sequences, while still relying on its constructed MSAs during the prediction process. In contrast to other methods, AlphaFold3 (ref. 21) employs a diffusion-based process to predict raw atom coordinates, replacing the AlphaFold2 structure module operating on amino acid-specific frames and side-chain torsion angles. While these MSA-based methods are capable of accurately predicting RNA 3D structures, they require extensive searches across large sequence databases, which can be time consuming. In contrast, models based on single sequences, including DRFold[23], do not utilize MSAs and thus do not require extensive searches in large sequence databases. Instead, DRFold[23] relies solely on predicted secondary structures to inform 3D structure predictions. This approach is faster, but typically has a lower accuracy compared with MSA-based methods. Next-generation deep learning methods might better leverage MSA-based approaches in a way that improves both speed and accuracy.

Here we present a language model-based deep learning method, RhoFold+, for accurate and fast de novo RNA 3D structure prediction. RhoFold+ represents a fully automated and differentiable improvement over its predecessor, RhoFold[19], leveraging improved integration of MSAs and other features to enhance performance. Our primary focus is on determining the structures of single-chain RNAs, which have limited interactions with other molecules. Addressing this challenge can help us better understand RNA biology and provide a starting point for solving more complex structural problems.

Results

Automated end-to-end platform for RNA 3D structure prediction

The development of RhoFold+ was guided by RNA-specific knowledge and the limitations of existing RNA 3D structure data. To build our training dataset, we curated all available RNA 3D structures from the PDB, using the BGSU representative sets of RNA structures (version 2022-04-13)[24]. We focused on single-chain RNAs and reduced redundancy by clustering sequences with Cd-hit[25] at an 80% sequence similarity threshold, resulting in 782 unique sequence clusters from 5,583 RNA chains. These RNA sequences were then processed through our pipeline, RhoFold+. First, the sequences were transformed using RNA-FM, our large RNA language model, to extract evolutionarily and structurally informed embeddings. Concurrently, MSAs were generated by searching through extensive sequence databases. The embeddings and MSA features were then fed into our transformer network, Rhoformer, and iteratively refined for ten cycles. Following this, our structure module employed a geometry-aware attention mechanism and an invariant point attention (IPA) module to optimize local frame coordinates and torsion angles for key atoms in the RNA backbone. Structural constraints, such as secondary structure and base pairing, were applied after reconstructing the full-atom coordinates (Fig. 1a and detailed discussion in Supplementary information). After developing RhoFold+, we rigorously benchmarked and evaluated its performance across a broad range of tests (Fig. 1b).

Benchmarking RhoFold+ on RNA-Puzzles

We performed a comprehensive retrospective comparison between RhoFold+ and other existing computational methods on two previously held community-wide challenges: RNA-Puzzles and CASP15. We first used the results from the RNA-Puzzles[26–30] competition, where the submissions were produced and optimized by human knowledge or computational methods. Importantly, here RhoFold+ was trained using nonoverlapping training data with respect to the RNA-Puzzles targets tested (Methods). We conducted preprocessing to obtain 24 single-chain RNA targets and excluded RNA complexes. This set of RNA targets contained two puzzles (PZs), PZ34 and PZ38, that were introduced after our development of RhoFold+ (Fig. 2a and Supplementary Fig. 3) and thus served as a blind test. After collecting the predictions of other methods from the official server (http://www.rnapuzzles.org/), we found that the performance of RhoFold+ surpassed that of all other methods, including FARFAR2/ARES, on nearly all targets, except for PZ24. Notably, RhoFold+ outperformed the second-best method on more than half of the targets by ~4 Å r.m.s.d. On 17 targets, RhoFold+ achieved r.m.s.d. values of <5 Å, and only one target exhibited an r.m.s.d. of >10 Å (Fig. 2a and Supplementary Table 5). As a whole, RhoFold+ produced an average r.m.s.d. of 4.02 Å, 2.30 Å better than that of the second-best model (FARFAR2: top 1%, 6.32 Å). Assessed using the template modeling (TM) score[31], RhoFold+ achieved an average of 0.57 (Supplementary Table 5), higher than the scores of other top performers (0.41 and 0.44).

To show that the promising results on RNA-Puzzles did not arise from overfitting, we studied whether the sequence similarity between the test set and our training data was substantially positively correlated with the performance of RhoFold+, as measured by the TM score and the local distance difference test (LDDT), a superposition-free score that evaluates local distance differences for all atoms in a model[32,33]. Such a correlation was previously found in protein structure prediction[13], yet here we found that R^2 values, which represent whether the slope is significantly nonzero, were 0.23 for the TM score and 0.11 for the LDDT (Fig. 2b,c), indicating no significant correlation between model performance and the similarity of our training and testing sets. These results suggest that RhoFold+ can generalize in predicting accurate RNA structures. A case study of a representative RNA-Puzzles target, PZ7 (a 186-nucleotide-long Varkud satellite ribozyme RNA), exemplifies this finding. Here, the structure of the most similar RNA in the training set differed substantially from the structure of PZ7 (Fig. 2b): the r.m.s.d. between these structures was 34.48 Å. As another example, PZ38 exhibited the highest sequence similarity of 53% with respect to all RNAs in our training set, and the r.m.s.d. between the structure of the most sequence-similar RNA and PZ38 was 16.46 Å (Fig. 2b). This was larger than the r.m.s.d. of 8.92 Å between PZ38 and the RhoFold+ prediction.

To test the ability of RhoFold+ to generalize for structure-dissimilar (in addition to mainly sequence-dissimilar) targets, we sought to determine whether the predictions of RhoFold+ could surpass the best single template (the most structurally similar model) in the training set for a given query. To investigate this, we compared the TM scores between our predictions and experimentally determined structures against the TM scores between the best single templates and experimentally determined structures across all RNA-Puzzles. For the majority of puzzles, RhoFold+ produced predictions with a higher global similarity and an average TM score of 0.574, surpassing the best single template by 0.05 (Fig. 2e and Supplementary Table 13). It is important to highlight that for proteins, surpassing the best single template required substantial progress. Indeed, it was only during CASP14 that computational methods outperformed the best single template. Although RhoFold+ generated considerably more accurate predictions than other methods under the conventional sequence similarity data splitting paradigm, we further tested the adaptability of RhoFold+ by eliminating 3D structures from the training set whose TM score, with respect to any target, surpassed a specified threshold (Supplementary Fig. 6 and Supplementary Tables 6 and 10). Even under this more demanding condition, RhoFold+ continued to exhibit a promising performance (Supplementary Table 10).

https://doi.org/10.1038/s41592-024-02487-0

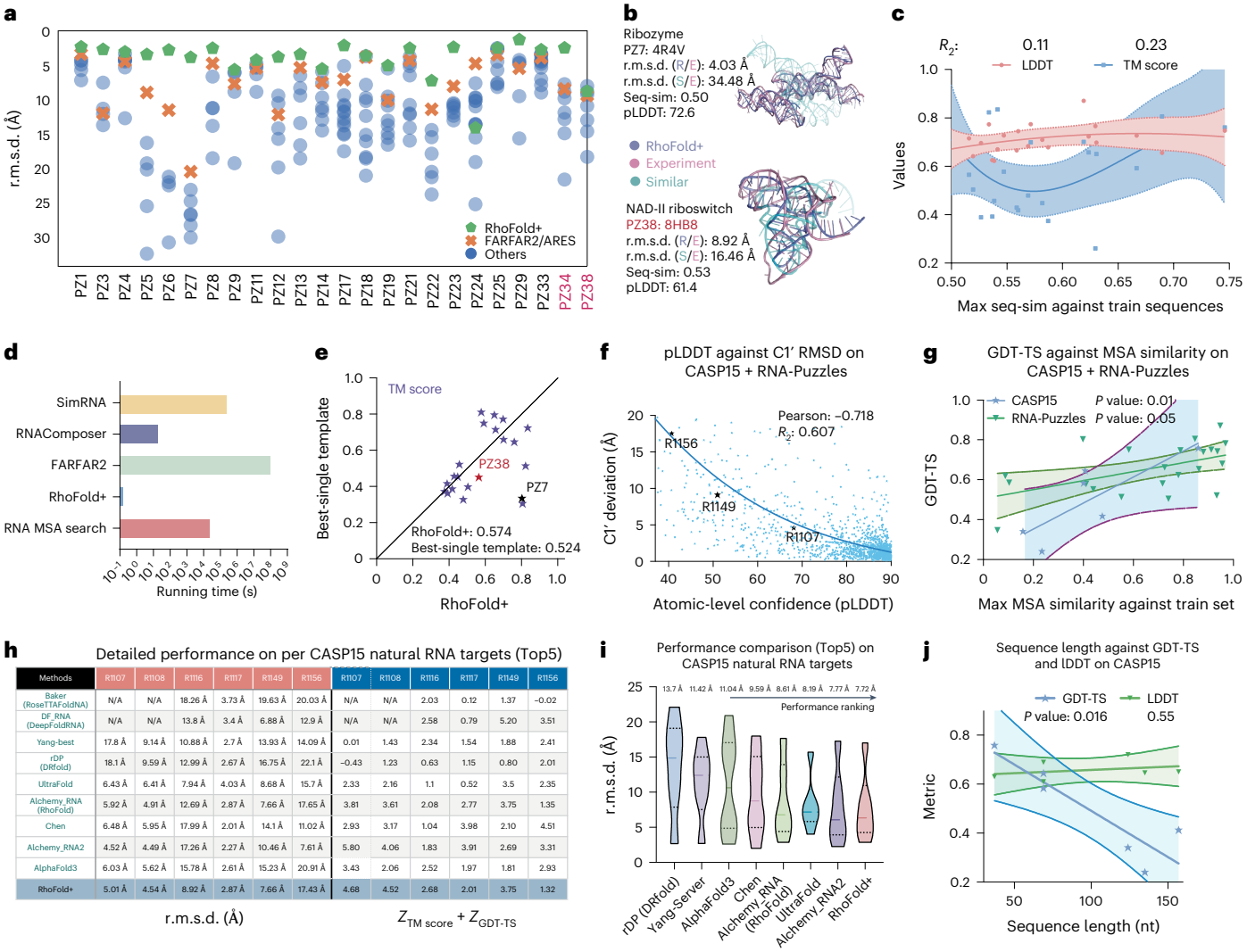


Fig. 2 | Benchmarking RhoFold+ on previously held community-wide challenges. a, The r.m.s.d. performance scatter plot of RhoFold+ and other methods across 24 nonoverlapping, nonredundant RNA-Puzzles targets. Each point represents a predicted model from the specific method. **b**, Visualization of RNA-Puzzles 7 and 38. In addition to the aligned RhoFold+ prediction, we show the most similar training structure with respect to each target, suggesting that RhoFold+ neither overfits the training set nor simply reproduces the most similar structure to the target. Seq-sim, sequence similarity. **c**, Regression plot of the TM score and LDDT of RhoFold+ predictions against the maximum sequence similarity among all the training sequences, across all RNA-Puzzles targets. Each point represents an RNA-Puzzles target. **d**, The running time comparison for different methods. **e**, Comparison of RhoFold+ predictions against the respective best single templates from our training set across all RNA-Puzzles targets. **f**, A regression plot for the r.m.s.d. of against atom-level pLDDT across all RNA-Puzzles and CASP15 targets. **g**, A regression plot for structure GDT-TS against MSA similarity across all RNA-Puzzles and CASP15 targets. **h**, A detailed performance comparison for CASP15 natural RNA targets. The pink columns record detailed r.m.s.d. values and the blue columns record the sum of *Z*-scores for the GDT-TS and TM score. Entries missing officially reported CASP15 data are marked as N/A; Yang-Sever and Chen are CASP15 registered groups. **i**, A comparison of RhoFold+'s average performance against the average reported performance of CASP15 groups and published works on CASP15 natural RNA targets. **j**, A regression plot for the structure GDT-TS and LDDT against sequence length across all CASP15 targets. The central curve in **c**, **g** and **j** represents the fit regression model, while the two surrounding curves indicate the 95% percentile intervals. **k**, A comparison of RhoFold+ predictions against Alchemy_RNA2 and UltraFold on the R1116 target from CASP15. MSA-sim, MSA profile similarity. **l**, For the R1156 target, showing a RhoFold+ potential failure case, involving incorrect stacking patterns and orientations.

In applying computational models to large-scale, real-world settings, speed is often a top priority. In addition to generating largely accurate folding results, we found that RhoFold+ is fast, with typical RNA-Puzzles predictions completed within ~0.14 s (Fig. 2d). In contrast, other approaches, including SimRNA[12], FARFAR2[10] and RNAComposer[34], exhibited significantly longer running times, probably due to the large-scale sampling processes employed by these methods (Fig. 2d).

Benchmarking RhoFold+ on CASP15 targets

As RNA-Puzzles was first released over a decade ago[26], we next used RhoFold+ to predict RNA targets from the more recent CASP15 (refs. 35,36). We focused on CASP15's six natural RNA targets (Fig. 2h and Supplementary Fig. 4). Artificially designed targets, which fell outside the expected domain of application for RhoFold+, were not included: in particular, the excluded targets were characterized by their lack of homology and divergence from our training set or their being RNA–protein complexes. We followed the CASP15 guidelines, which specified that participating teams were permitted to submit up to five models. Utilizing different, randomly sampled MSAs (Methods), we modeled five candidate structures for each target using RhoFold+ and considered only the highest-performing prediction (Supplementary Table 6).

Several top-ranking CASP15 groups and recent published works on RNA 3D structure prediction[17,18,20,21,23] were included in our benchmarking. Particularly, CASP15 groups were divided into two categories, 'server' and 'expert', depending on whether or not human expert knowledge and fine tuning were used. Regardless of the category, many CASP15 groups employed computational pipelines that were based on comparative or statistical learning for natural targets, thus allowing us to assess the learning capability of RhoFold+. Our preliminary model, AIchemy_RNA (RhoFold), was a participant in the 'expert' category. Building on RhoFold, RhoFold+ represents a fully automated and end-to-end pipeline that is more similar to participants in the 'server' category. Here, we found that RhoFold+ outperformed RhoFold on CASP15's natural RNA targets by an average r.m.s.d. of ~1 Å. Furthermore, RhoFold+ outperformed other methods whose predictions were available for all six natural RNA targets, including the first-ranked AIchemy_RNA2, the second-ranked Chen method and other computational methods, including DRfold[23], DeepFoldRNA[17], AlphaFold3[21] and trRosettaRNA[18] (Fig. 2h,i). Although RhoFold+ outperformed AIchemy_RNA2 marginally by 0.06 Å (average r.m.s.d.; Fig. 2i), AIchemy_RNA2 required expert knowledge. Additionally, RhoFold+ demonstrated accuracy comparable to each top-performing method on almost every natural RNA target, with the exception of R1156 (Fig. 2h).

Following CASP15's assessment approach[36], we also computed *Z*-scores for the predictions from all participating groups. CASP15 prioritized the TM score and the global distance test-total score (GDT-TS), which evaluates both overall structure similarity and local alignment, leading us to assess these models based on the cumulative *Z*-scores of these metrics (Fig. 2h). On the six natural RNA targets and among the subset of all CASP15 participants ranked on these specific targets, RhoFold (AIchemy_RNA) was fourth, while the performance of RhoFold+ was on par with that of AIchemy_RNA2 (with a difference of 0.4 in the *Z*-score) and surpassed that of other methods. In a detailed analysis of performance on specific targets, we found that, for target R1108, RhoFold+ achieved the best *Z*-score and r.m.s.d. Interestingly, RhoFold+ also attained the best *Z*-score for R1116, although the r.m.s.d. was ~1 Å higher than that of UltraFold (other methods produced predictions with significantly lower accuracy, all with r.m.s.d. >10 Å). Upon further investigation, we found that, while UltraFold outperformed RhoFold+ on this metric by producing accurate local predictions, the predicted global structure was less accurate, as evidenced by a TM score of 0.497 and a GDT-TS score of <0.4. In contrast, RhoFold+ inaccurately predicted a helix angle, resulting in an r.m.s.d. of 8.92 Å, but its correctly predicted topology resulted in a higher TM score of >0.55. For this target, AIchemy_RNA2 incorrectly predicted the stem stackings and RNA topology, resulting in a high r.m.s.d. of 17.26 Å and a TM score of ~0.49. Notably, the RhoFold+ prediction for R1116 did not arise from overfitting, as indicated by the low maximum structural similarity (TM score) and maximum sequence similarity of R1116 with respect to the training set (Fig. 2k and Supplementary Table 6).

We also looked into targets where RhoFold+ may achieve reduced performance and found that higher MSA quality correlated with better performance. While RhoFold+ accurately predicted local structural topologies, it struggled with aligning helices, particularly at junctions. This discrepancy may be due to the dynamic and flexible nature of RNA junctions, which often adopt multiple conformations[37–39], making them challenging for fully automated models to represent accurately (Fig. 2k,l and detailed discussion in Supplementary information).

Factors influencing prediction accuracy

Building on the findings above, we performed a more comprehensive study involving all CASP15 natural RNAs and RNA-Puzzles targets. We observed that the prediction accuracy of RhoFold+ is sensitive toward the query's MSA profile similarity (Supplementary information) against the training set (Fig. 2g) and the complexity of RNA structures (query length; Fig. 2j). Additionally, predicted LDDT (pLDDT) scores were found to correlate with the confidence of RhoFold+, providing a useful metric for identifying regions with lower prediction accuracy, especially in more complex or less homologous queries (Fig. 2f and detailed discussion and analysis in Supplementary information).

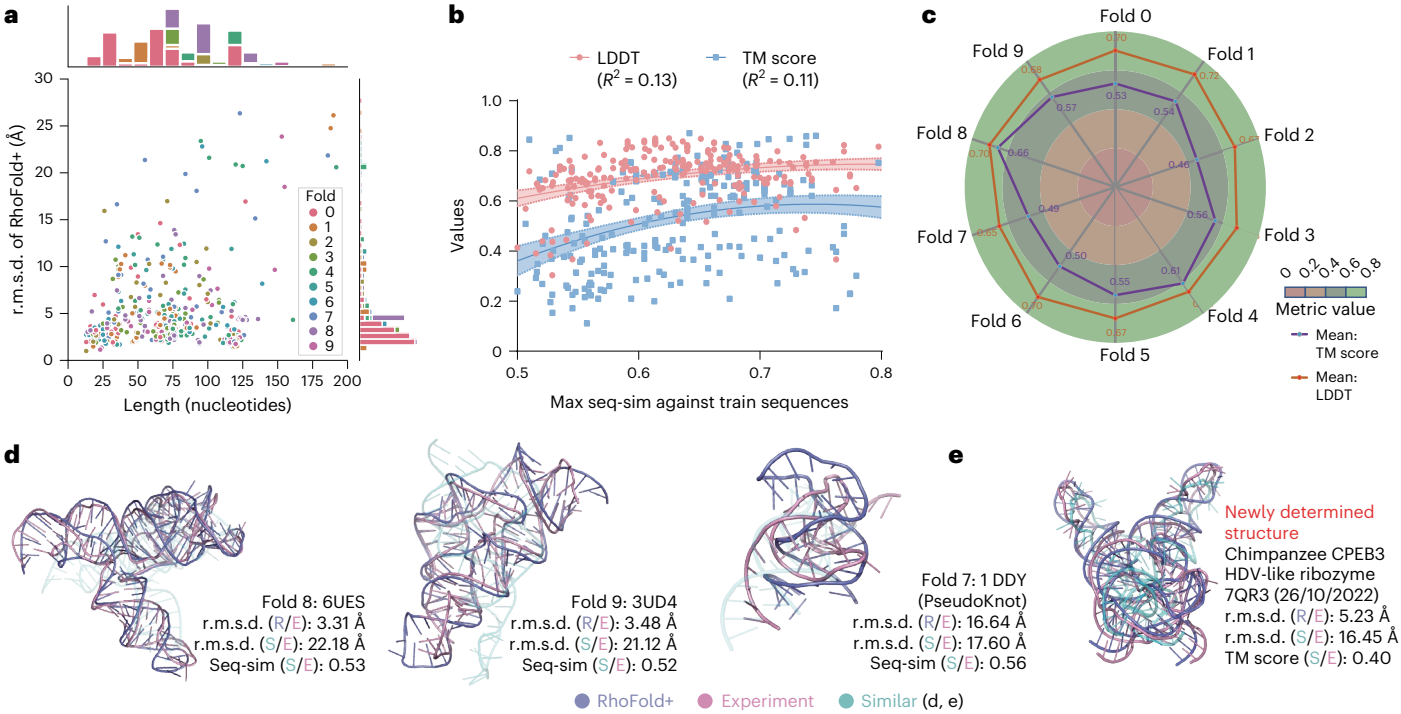
Benchmarking RhoFold+ on all determined RNA 3D structures

After benchmarking RhoFold+ with RNA-Puzzles and CASP15, we next evaluated RhoFold+ in greater detail using all experimentally determined RNA structures, as defined by the BGSU representative sets of RNA structures (preprocessed to remove redundancy). To further study the performance of RhoFold+, we performed tenfold cross-validation

Fig. 3 | Benchmarking RhoFold+ on all experimentally determined RNA structures supports the accuracy and ability of RhoFold+ to generalize to unseen structures. a, A plot of r.m.s.d. values against sequence length for all cross-validation experiments. Each point represents an RNA structure and is colored according to the cross-validation fold. **b**, A regression analysis for each prediction's TM score (blue) and LDDT (pink) against the maximum sequence similarity with respect to all training data. Each point represents an RNA structure. **c**, The average TM score and LDDT for each fold. **d**, Visualization of two representative riboswitch structures, 6UES and 3UD4, and a pseudoknot 1DDY (pink), along with the corresponding RhoFold+ predictions (slate) and the training RNA structures with the highest sequence similarity (cyan). In **a**–**d**, the tenfold cross-validation of RhoFold+ using all experimentally determined RNA structures is shown. **e**, Visualization of a newly determined RNA structure, 7QR3, an hepatitis delta virus (HDV)-like ribozyme, which has a low structural similarity with respect to the training set, but whose structure (pink) is accurately predicted by RhoFold+ (slate). The most similar structure, 7DLZ, is shown in cyan. **f**, A comparison of average r.m.s.d. values generated by RhoFold+ and other methods on the new PDB set, a set of 76 newly determined solo RNA structures. **g**, A regression plot of the prediction r.m.s.d. values against maximum sequence similarity to the training set for RhoFold+ and other baseline methods. **h**, A regression plot of the correlation between the RhoFold+ predictions TM score/LDDT and the maximum MSA profile similarity against the training set. The central curve in **b** and **h** represents the fit regression model, while the two surrounding curves indicate the 95% percentile intervals. **i**, An overview of cross-type validation performance of RhoFold+ measured by LDDT and TM score. All structures in the type used for validation were masked during model training. sRNA, small RNA. **j**, A violin plot of RhoFold+ r.m.s.d. values in the cross-family validation. Here, all the structures in a family to be tested were masked during model training and RhoFold+ accurately predicted RNA structures from most unseen families. The numbers of sequences in each family are shown in parentheses.

https://doi.org/10.1038/s41592-024-02487-0

by iteratively masking 80 sequence clusters for validation and leaving 702 sequence clusters for training. We found that the performance of RhoFold+ across all RNA structures was robust regardless of the train–test data split and fairly consistent across all folds (Fig. 3a–c). Slight variations in TM score might be caused by challenging targets such pseudoknot cases in Fold2 and Fold7 similar to PZ24 (Fig. 3c,e), and we expect that the predictions of RhoFold+ on such targets could be improved if secondary structure constraints were provided. Also, during our cross-validation test, the accurate predictions of RhoFold+ were not due to merely mimicking the most sequence-similar training data (Fig. 3b,d,e). A plot of the r.m.s.d. against the sequence length shows that r.m.s.d. values were largely distributed below 10 Å, independent of the sequence length (Fig. 3a). Outliers with r.m.s.d. >20 Å were more likely to occur for sequences longer than 200 nt, where we



RF00059: TPP riboswitch (THI element) (12,375)
RF00162: SAM riboswitch (S box leader) (7,052)
RF00504: glycine riboswitch (4,600)
RF01051: cyclic di-GMP-I riboswitch (4,213)
RF00379: YdaO/yuaA leader (3,917)
RF00167: purine riboswitch (2,661)
RF00028: group I catalytic intron (2,611)
RF01852: selenocysteine transfer RNA (2,359)
RF00168: lysine riboswitch (2,261)
RF01750: ZMP/ZTP riboswitch (1,719)
RF02540: Archaeal large subunit ribosomal RNA (1,539)
RF00234: GlmS glucosamine-6-phosphate activated ribozyme (943)
RF01763: guanidine-III riboswitch (706)
RF01831: THF riboswitch (700)
RF01344: CRISPR RNA direct repeat element (382)
RF03013: NadA RNA (195)

expect further improvement by more tuning on long RNAs (detailed discussion in Supplementary information).

As a further evaluation of the capabilities of RhoFold+, we considered the model's performance on newly determined RNA single-stranded structures released subsequent to the compilation of our training dataset. This approach acted as an additional blind test, similar to the CASP15 competition. We included comparisons against FARFAR2 and recent deep learning methods[17,18,21,23], all of which have inference code and/or servers available and some of which also participated in CASP15 (Methods). RhoFold+ outperformed all benchmarked models, achieving the highest average accuracy as measured by r.m.s.d. RhoFold+ produced an average r.m.s.d. of 7.74 Å, which was approximately 0.8 Å and 10.5 Å better than the second-ranked DeepRNAFold and the lowest-ranked FARFAR2, respectively. Notably, on average, RhoFold+ also outperformed AlphaFold3 and RoseTTAFold2NA by approximately 2.2 Å and 1.8 Å, respectively (Fig. 3f and detailed discussion in Supplementary information). These results were consistent with the performance observed in our previous benchmark on CASP15, suggesting that RhoFold+ accurately generalizes to newly determined structures not seen in our training set. Furthermore, these results support that AlphaFold3 and RoseTTAFold2NA, which are designed to predict biomolecular complexes, do not perform as well as RhoFold+ when applied to single RNA molecules. Further examining sequence and structural similarities to our training set reveals that RhoFold+ maintained strong performance even with sequence similarities below 0.5 (Fig. 3g), and the TM score was greatly influenced by MSA profile similarity while local accuracy (LDDT) remained high and robust (Fig. 3h). Additionally, RhoFold+ demonstrated strong generalizability, accurately folding structures such as 7QR3 despite its low similarity to the closest training template, 7DLZ (TM score of 0.40, r.m.s.d. of 16.45 Å; Fig. 3e).

RhoFold+ generalizes to unseen RNA types and families

Having demonstrated that RhoFold+ can generalize to predicting RNA structures with divergent sequence similarities, structural similarities and dates of release, we next investigated the ability of RhoFold+ to handle different RNA types and families defined by expert knowledge. In particular, RNA types and families—such as those curated in Rfam[40]—are often classified manually based on factors including function, structure and co-evolutionary information. Addressing the challenge of generalizing to different RNA types and families may be considerably more demanding for deep learning methods such as RhoFold+ as such a task requires larger domain shifts.

We benchmarked the cross-type performance of RhoFold+ by training the model on a subset of all RNA types while testing on the others. RhoFold+ showed robustness across RNA types. Though struggling with introns and riboswitches, it performed well on transfer RNA (tRNA) and micro RNA (miRNA) types, achieving TM scores up to 0.73 (Fig. 3i). When compared with FARFAR2, RhoFold+ outperformed it across all RNA types, particularly in tRNAs and ribosomal RNAs (rRNAs), with smaller margins for riboswitches (detailed discussion in Supplementary information). For cross-family tests, RhoFold+ achieved an average r.m.s.d. of 6.69 Å (Fig. 3j), but struggled with complex families such as group I introns (RF00028). This difficulty is consistent with challenges observed in cross-type tests, such as for complex RNA types such as introns and CRISPR RNA elements (RF01344). These elements interact with various proteins and enzymes, and focusing solely on RNA structure without considering these interactions may limit the prediction accuracy (detailed discussion in Supplementary information). Overall, these tests demonstrate the ability of RhoFold+ to generalize across unseen RNA types and families, though challenges remain for complex structures and datasets with limited available data.

RhoFold+ predicts secondary structures and substructures

RhoFold+ can accurately predict RNA 3D structures, but the limited number of experimentally determined RNA structures and types makes it difficult to understand the space of all possible RNA folds. This is particularly true for complicated and large RNA types, including internal ribosomal entry sites, introns, synthetic RNAs and long noncoding RNAs. RNA secondary structures, however, can be more easily determined in experiments and accurate secondary structure predictions can supplement the predictions of 3D structures, offering valuable insights into RNA folding and function. Therefore, we adapted RhoFold+ to predict secondary structures as well. As RhoFold+ was designed to predict RNA 3D structures, we incorporated a postprocessing module that utilizes the features retrieved from RhoFold+'s Rhoformer to predict secondary structures (since Rhoformer's features show attention maps highly aligned with the contact maps; Supplementary Fig. 8 and Supplementary Table 14). This module takes into account the same structural information as the module performing 3D reconstruction but operates under distinct geometric and biological constraints imposed to predict secondary structure.

We benchmarked the performance of RhoFold+ on newly determined PDB structures (the 'new PDB set') and the ArchiveII dataset[41], which includes secondary structure information for diverse RNAs. On the new PDB set, RhoFold+ outperformed UFold[41] by 0.035 in the average F1 score (Fig. 4a), even when UFold was trained on all available data (PDB and bpRNA-1M, a database with over 100,000 annotated RNA secondary structures). On the ArchiveII dataset comprising 2,975 RNA samples, RhoFold+ also outperformed other secondary structure prediction methods (Fig. 4b), particularly on larger RNA types (Fig. 4c). For instance, it achieved an F1 score of 0.60 on structured domains in the dengue virus transcriptome (Supplementary Table 19), aligning with results from mutational profiling (RING-MaP)[42,43]. Similarly, the strong performance of RhoFold+ did not stem from mimicking training data, as it maintained an F1 score of ~0.7 even when sequence similarity dropped below 50% (Fig. 4e), and achieved a perfect F1 score of 1.0 on the CASP15 target R1117 (Fig. 4f). These results suggest that RhoFold+ not only excels in predicting 3D structures, but also generates rich, meaningful representations that enable state-of-the-art secondary structure prediction.

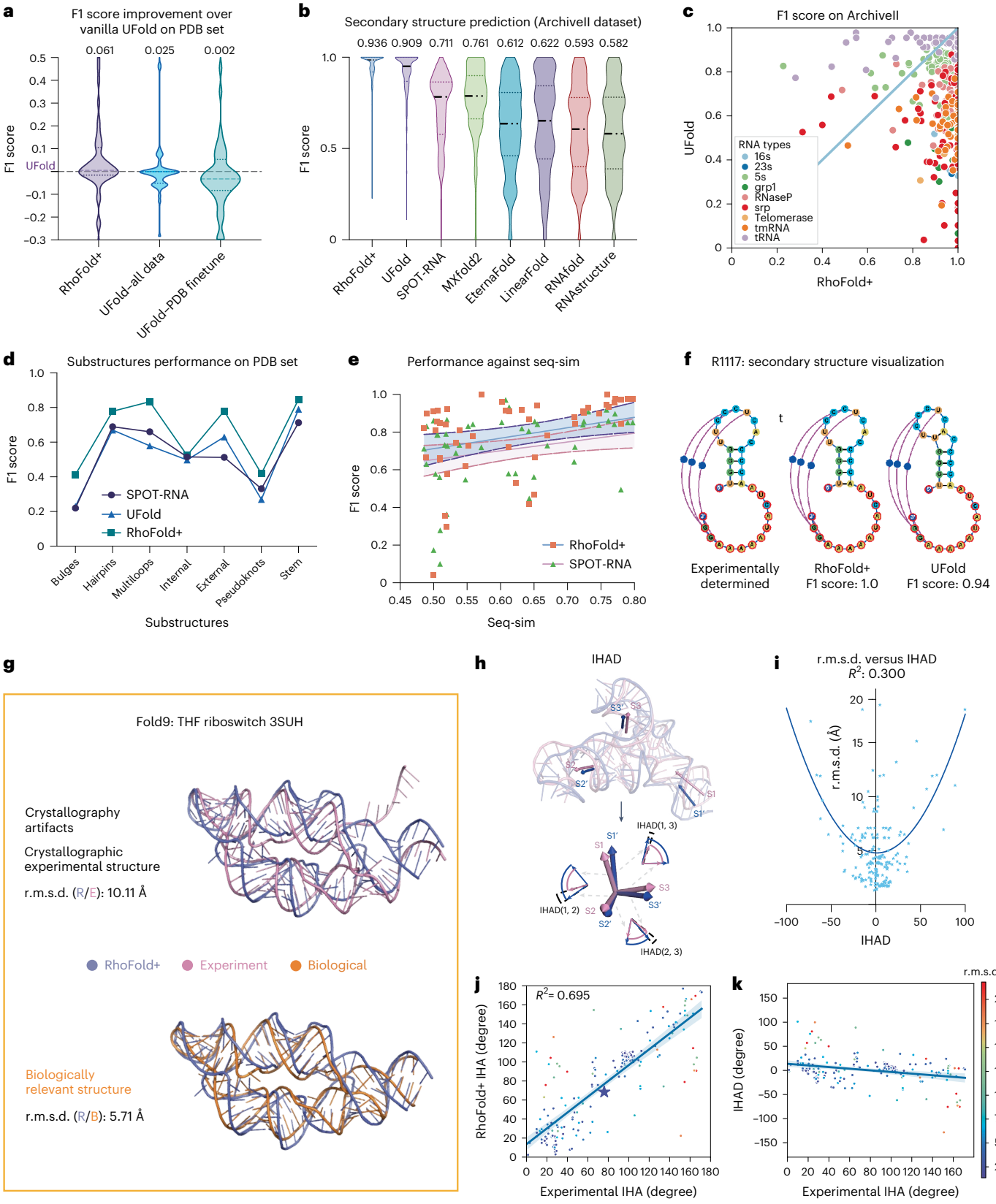
Fig. 4 | RhoFold+ accurately predicts secondary structures and IHAs from experimental data. a, F1 score comparison against multiple configurations of UFold on the PDB set. Here, a version of UFold trained on bpRNA is also presented as a baseline, to evaluate the improvement in terms of F1 score. **b**, The F1 score distribution of various methods on the ArchiveII dataset. Average scores are indicated at the top of the plot. **c**, F1 score comparison between RhoFold+ and UFold on the ArchiveII dataset. Each point represents an RNA structure and is colored according to its RNA type. srp, signal recognition particle RNA; tmRNA, transfer-messenger RNA. **d**, F1 score comparison of RhoFold+ versus UFold and SPOT-RNA on RNA substructures in the new PDB set. **e**, F1 score comparison of RhoFold+ versus UFold and SPOT-RNA against sequence similarity of RNA structures in the new PDB set. **f**, Visualization of a CASP15 RNA target where RhoFold+ predicted the correct secondary structures including pseudoknots. **g**, Visualization of a swapped dimer, tetrahydrofolate (THF) ribozyme, 3SUH, for which the RhoFold+ prediction (purple) resembles the biologically meaningful structure (orange) instead of the crystallographic artifact found in the PDB (pink). **h**, Visualization showing the definition of the IHAD, which is the difference between the IHAs derived from the RhoFold+ prediction and the experimentally determined structure. **i**, Regression analysis between the IHAD and r.m.s.d. of the RhoFold+ predictions. Each point represents an RNA. **j**, Comparison between the IHAs derived from the RhoFold+ predictions against those from experimental structures. Each point represents an angle instance and is colored according to the r.m.s.d. between the experimental structure containing the angle and the structure predicted by RhoFold+. **k**, A plot of the IHAD against experimentally determined IHA values. The coloring is the same as in **j**. The central curve in **e**, **j** and **k** represents the fit regression model, while the two surrounding curves indicate the 95% percentile intervals.

https://doi.org/10.1038/s41592-024-02487-0

We further evaluated substructures within RNA secondary structures, finding that RhoFold+ consistently outperformed SPOT-RNA[44] and UFold[41] across all substructures, with the most significant improvements in multiloops and external loops, while internal loops and pseudoknots showed similar performance across methods (Fig. 4d). These results underscore the potential capability of RhoFold+ in predicting RNA secondary structures and enhancing our understanding of RNA function.

Correcting artifacts and IHA prediction

As RhoFold+ accurately predicts RNA structures at both the secondary and tertiary levels, we asked whether we could leverage RhoFold+ for



https://doi.org/10.1038/s41592-024-02487-0

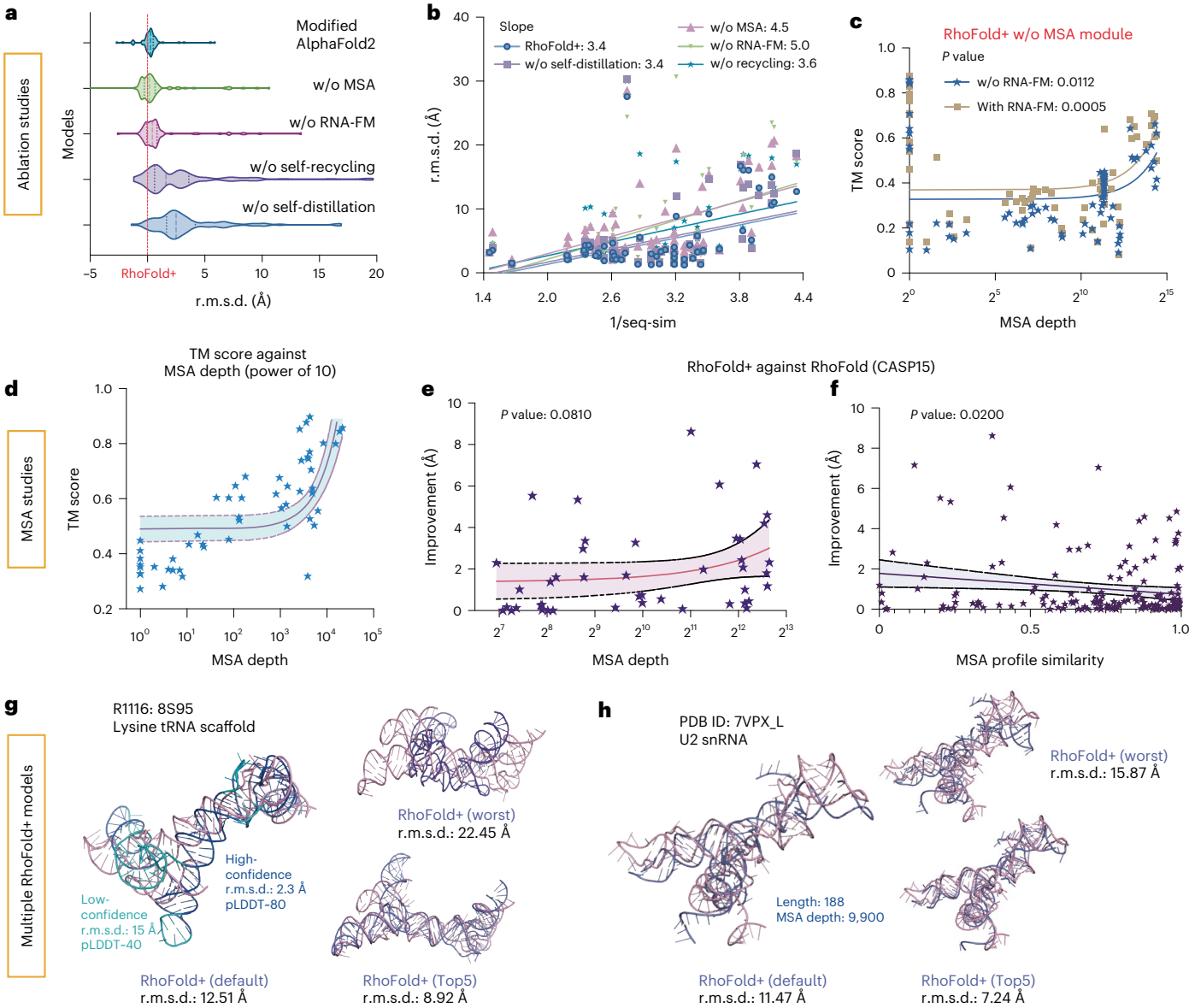


Fig. 5 | Ablation studies of RhoFold+ and sampling of multiple models. **a**, Ablation studies of RhoFold+ without (w/o) corresponding modules in RhoFold+ with performance measured by r.m.s.d. **b**, A regression analysis for prediction accuracy (measured by r.m.s.d.) against the reciprocal of sequence similarity. **c**, A regression analysis of the TM score against MSA depth for the ablation study of the RNA-FM module. Note that the *x* axis is log scaled. **d**, A plot of prediction accuracy (measured by the TM score) against MSA depth. **e**, A plot of the improvement of RhoFold+ against RhoFold (measured by r.m.s.d.) across different MSA depths. **f**, A plot of the improvement of RhoFold+ against RhoFold (measured by r.m.s.d.) across different MSA profile similarities. The central curve in **e** and **f** represents the fit regression model, while the two surrounding curves indicate the 95% percentile intervals. **g**, Visualization of a CASP15 target where RhoFold+ produces an r.m.s.d. of 12.51 Å, but improves by 8.92 Å using the Top5 prediction from MSA sampling. **h**, Visualization of a newly determined RNA structure where the r.m.s.d. of RhoFold+ improves by 7.92 Å using Top5 prediction from MSA sampling.

experimental efforts. Toward this, we investigated two use cases of RhoFold+: (1) for correcting experimental structural artifacts and (2) for guiding RNA construct engineering.

X-ray crystallography is widely used to resolve RNA 3D structures, but it can introduce artifacts such as domain-swapped dimers[45], potentially misleading machine learning models that do not generalize well. In one case, the RhoFold+ prediction for 3SUH initially yielded a high r.m.s.d. of 10.11 Å compared with the PDB structure. However, further analysis revealed that the crystal structure involved a domain-swapped dimer. When comparing the RhoFold+ prediction with the inferred monomeric structure, the r.m.s.d. improved to 5.71 Å, indicating RhoFold+ accurately predicted the biologically relevant structure (Fig. 4g). Similar findings were also observed for the ZTP riboswitch[46] (Supplementary Fig. 9), suggesting that RhoFold+ can effectively correct for such experimental artifacts.

When comparing experimental data with RNA 3D models, additional geometric metrics, such as interhelical angles (IHAs), can provide insights beyond standard global alignment measures such as r.m.s.d., LDDT and TM score. IHAs, which can be estimated using experimental methods, are useful for validating predicted models and guiding RNA nanostructure design. We introduced the IHA difference (IHAD) as a metric to benchmark the predictions of RhoFold+ (Fig. 4h and Supplementary information), finding that IHAD can reveal discrepancies in stem orientations that are not captured by r.m.s.d. alone (Fig. 4i). Our analysis shows that RhoFold+ generally predicted stem directions accurately (Fig. 4j,k), though performance decreased for IHAs near 0° or 180°, probably due to underfitting of parallel stems in large and complex structures (Fig. 4k and detailed discussion in Supplementary information). We further demonstrated the practical application of

IHAs by predicting values for RNA constructs such as the FMN riboswitch and the P4–P6 domain from the Tetrahymena group I intron (Supplementary Fig. 9).

Ablation studies and generation of multiple predictions

Given the high accuracy and speed of RhoFold+, we finally conducted ablation studies to understand which components and information are important to the RhoFold+ predictions. The architectural components we investigated included four different modules (Fig. 5a and Methods). Ablation studies were performed on 138 PDB targets (collected between April 2022 and December 2023) with sequence similarities below 80% to our training set and lengths ranging from 16 to 300 nt (the 'Ablation set'). By removing each RhoFold+ component, we observed that all contributed to improving the performance, with the MSA module being the most critical, followed by the RNA-FM language model (Fig. 5a). The RNA-modified version of AlphaFold2, without the MSA module, performed worse than RhoFold+ (Fig. 5a). Notably, removing RNA-FM led to a sharper performance decline for dissimilar sequences (Fig. 5b), and the RNA-FM module seemed to compensate for the loss of the MSA module, maintaining higher TM scores (Fig. 5c). Additionally, removing the recycling module most significantly affected predictions for longer sequences, probably due to its role in effectively deepening the model (Supplementary Fig. 7 and detailed discussion in Supplementary information).

These findings are consistent with our results for CASP15's natural RNA targets and RNA-Puzzles, where MSA quality significantly impacts predictions. We also explored how the number of sequences in the extracted MSA influences accuracy. While RhoFold+ is limited to 256 MSAs due to training constraints, this limit did not compromise its effectiveness. A key enhancement in RhoFold+ is its ability to generate multiple predictions by sampling or clustering from a fixed number of MSAs, allowing for broader prediction selection and improved outcomes. Performance on RNA-Puzzles showed an inverse correlation with reduced MSA counts, with a marked improvement when the MSA number exceeded 100 (Fig. 5d), indicating that a larger MSA pool enhances model optimization (detailed discussion in Supplementary information). With this expanded MSA sampling, the lowest r.m.s.d. of the RhoFold+ Top5 predictions significantly decreased compared with RhoFold, correlating positively with increased MSA depth and yielding an up to 10 Å improvement (Fig. 5e). This improvement was more pronounced when the MSA profile similarity between the query and training sequences was high, resulting in smaller gains when similarity was already strong (Fig. 5f). Overall, additional MSA sampling is crucial for high performance, as demonstrated for CASP15 target R1116 and PDB 7VPX_L (Fig. 5g,h).

Discussion

In this study, we have developed an end-to-end language model-based deep learning method, RhoFold+, to predict RNA 3D structure from sequence. RhoFold+ is a fully automated and differentiable model that integrates an RNA language model pretrained on ~23.7 million RNA sequences without structural information leakage and multiple strategies to augment the scarce training data. RhoFold+ outperforms other RNA structure prediction approaches based on deep learning on CASP15 natural RNA targets and achieves a sub-4 Å mean r.m.s.d. for the nonoverlapping and nonredundant RNA-Puzzles structures. As RhoFold+ does not require any time-consuming and computationally intensive sampling processes, RhoFold+ is also fast and efficient, and neither does it rely on expert knowledge, which has been used in the most high-performing approaches to RNA structure prediction so far. RhoFold+ is able to generalize from different sets of training data and accurately predict both available RNA 3D structures and newly determined ones, an observation underscored by the strong robustness of RhoFold+ during cross-fold validation. Additionally, RhoFold+ can accurately predict unseen RNA structures during cross-family and cross-type validations. Although RhoFold+ was designed to predict 3D structures, it can also accurately predict RNA secondary structures. Applying RhoFold+ for the prediction of IHAs—a task inspired by cryogenic electron microscopy-based and NMR-based construct engineering design—suggests its potential to accelerate the process of experimentally determining more RNA structures.

Although RhoFold+ shows promising performance, it shares limitations with other deep learning methods for RNA structure prediction. First, our knowledge of RNA structural diversity is limited, making it challenging to predict different conformations of the same RNA molecule due to their dynamic nature and interactions with other molecules. RNA junctions, for example, can adopt multiple conformations and are better represented as dynamic ensembles[37–39]. Second, due to insufficient data, predicting large and complex RNA structures, particularly those with multiple helices or pseudoknots, remains difficult, especially for sequences longer than 500 nucleotides. Third, RNA complexes involving ligands or proteins present additional challenges as current methods often fail to account adequately for these interactions, reducing accuracy. While methods such as AlphaFold3 (ref. 21) and RoseTTAFoldNA[20] can predict RNA complexes, their accuracy is still limited and they perform less well than RhoFold+ on single-strand RNAs. Fourth, RhoFold+ and similar models are trained on datasets derived from specific environmental conditions, which may not generalize well to the diverse and dynamic solution conditions that RNA molecules encounter in vivo. These conditions include varying concentrations of ions, such as magnesium and potassium, and the presence of ligands, which are known to play critical roles in RNA folding and stability.

Methods that rely on MSAs are limited by the availability of these alignments, making accurate predictions difficult for artificially designed or orphan RNAs lacking corresponding MSAs. Although RNA-FM has helped mitigate this dependency, challenges remain. RhoFold+ and similar deep learning models, while accurate, are hindered by limited knowledge of RNA structural diversity, difficulties in predicting large and complex structures and the reliance on MSAs. To mitigate these obstacles, integrating probing methods to define secondary structures, incorporating molecular dynamics and energy function techniques, and improving the MSA extraction process could potentially enhance the accuracy of RhoFold+. Additionally, addressing RNA–protein and RNA–ligand interactions remains crucial, and integrating RhoFold+ with protein structure prediction tools such as RoseTTAFoldNA or AlphaFold3 could improve its capabilities in these areas.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02487-0.

References

1. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **15**, 469–479 (2014).
2. Warner, K. D., Hajdin, C. E. & Weeks, K. M. Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discov.* **17**, 547–558 (2018).
3. Kulkarni, J. A. et al. The current landscape of nucleic acid therapeutics. *Nat. Nanotechnol.* **16**, 630–643 (2021).
4. Sheridan, C. First small-molecule drug targeting RNA gains momentum. *Nat. Biotechnol.* **39**, 6–9 (2021).
5. Zhao, E. M. et al. RNA-responsive elements for eukaryotic translational control. *Nat. Biotechnol.* **40**, 539–545 (2022).
6. Liu, D., Thélot, F. A., Piccirilli, J. A., Liao, M. & Yin, P. Sub-3-Å cryo-em structure of RNA enabled by engineered homomeric self-assembly. *Nat. Methods* **19**, 576–585 (2022).

7. Xu, B. et al. Recent advances in RNA structurome. *Sci. China Life Sci.* **65**, 1285–1324 (2022).
8. Rother, M., Rother, K., Puton, T. & Bujnicki, J. M. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.* **39**, 4007–4022 (2011).
9. Flores, S. C., Wan, Y., Russell, R. & Altman, R. B. Predicting RNA structure by multiple template homology modeling. In *Proc. Pacific Symposium on Biocomputing 2010* (ed. Altman, R. B. et al.) 216–227 (World Scientific, 2010).
10. Watkins, A. M., Rangan, R. & Das, R. Farfar2: improved de novo rosetta prediction of complex global RNA folds. *Structure* **28**, 963–976 (2020).
11. Wang, J., Wang, J., Huang, Y. & Xiao, Y. 3DRNA v2.0: an updated web server for RNA 3D structure prediction. *Int. J. Mol. Sci.* **20**, 4116 (2019).
12. Boniecki, M. J. et al. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **44**, e63 (2016).
13. Jumper, J. M. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
14. Chen, X., Li, Y., Umarov, R., Gao, X. & Song, L. RNA secondary structure prediction by learning unrolled algorithms. In *Proc. International Conference on Learning Representations* (OpenReview, 2020); https://openreview.net/forum?id=S1eALyrYDH
15. Chen, J. et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. Preprint at https://arxiv.org/abs/2204.00300 (2022).
16. Townshend, R. J. et al. Geometric deep learning of RNA structure. *Science* **373**, 1047–1051 (2021).
17. Pearce, R., Omenn, G. S. & Zhang, Y. De novo RNA tertiary structure prediction at atomic resolution using geometric potentials from deep learning. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.15.491755 (2022).
18. Wang, W. et al. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nat. Commun.* **14**, 7266 (2023).
19. Shen, T. et al. E2Efold-3D: end-to-end deep learning method for accurate de novo RNA 3D structure prediction. Preprint at https://arxiv.org/abs/2207.01586 (2022).
20. Baek, M. et al. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **21**, 117–121 (2023).
21. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
22. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
23. Li, Y. et al. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nat. Commun.* **14**, 5745 (2023).
24. Danaee, P. et al. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res.* **46**, 5381–5394 (2018).
25. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
26. Cruz, J. A. et al. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* **18**, 610–625 (2012).
27. Miao, Z. et al. RNA-Puzzles round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **21**, 1066–1084 (2015).
28. Miao, Z. et al. RNA-Puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* **23**, 655–672 (2017).
29. Miao, Z. et al. RNA-Puzzles round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA* **26**, 982–995 (2020).
30. Magnus, M. et al. RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Res.* **48**, 576–588 (2020).
31. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
32. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
33. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
34. Popenda, M. et al. Automated 3D structure composition for large RNAs. *Nucleic Acids Res.* **40**, e112 (2012).
35. Critical assessment of techniques for protein structure prediction. *Protein Structure Prediction Center* https://predictioncenter.org/casp15/index.cgi (2022).
36. Das, R. et al. Assessment of three-dimensional RNA structure prediction in CASP15. *Proteins* **91**, 1747–1770 (2023).
37. Gupta, P., Khadake, R. M., Panja, S., Shinde, K. & Rode, A. B. Alternative RNA conformations: companion or combatant. *Genes* **13**, 1930 (2022).
38. Zhang, Q., Stelzer, A. C., Fisher, C. K. & Al-Hashimi, H. M. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature* **450**, 1263–1267 (2007).
39. Ding, J. et al. Visualizing RNA conformational and architectural heterogeneity in solution. *Nat. Commun.* **14**, 714 (2023).
40. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).
41. Fu, L. et al. Ufold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **50**, e14 (2022).
42. Dethoff, E. A. et al. Pervasive tertiary structure in the dengue virus RNA genome. *Proc. Natl Acad. Sci. USA* **115**, 11513–11518 (2018).
43. Rice, G. M., Leonard, C. W. & Weeks, K. M. RNA secondary structure modeling at consistent high accuracy using differential shape. *RNA* **20**, 846–854 (2014).
44. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **10**, 5407 (2019).
45. Bou-Nader, C. & Zhang, J. Structural insights into RNA dimerization: motifs, interfaces and functions. *Molecules* **25**, 2881 (2020).
46. Trausch, J. J., Marcano-Velázquez, J. G., Matyjasik, M. M. & Batey, R. T. Metal ion-mediated nucleobase recognition by the ZTP riboswitch. *Chem. Biol.* **22**, 829–837 (2015).

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China. [2]Shanghai Zelixir Biotech Company Ltd, Shanghai, China. [3]Shenzhen Institute of Advanced Technology, Shenzhen, China. [4]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai, China. [5]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [6]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA. [7]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [8]Center for Molecular Design and Biomimetics at the Biodesign Institute, Arizona State University, Tempe, AZ, USA. [9]School of Molecular Sciences, Arizona State University, Tempe, AZ, USA. [10]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [11]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [12]Synthetic Biology Center, Massachusetts Institute of Technology, Cambridge, MA, USA. [13]Integrated Biosciences, Redwood City, CA, USA. [14]OneAIM Ltd, Hong Kong SAR, China. [15]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA. [16]The CUHK Shenzhen Research Institute, Shenzhen, China. [17]These authors contributed equally: Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Yu Li. ✉e-mail: siqisun@fudan.edu.cn; di.liu@asu.edu; wangsheng@zelixir.com; peng_yin@hms.harvard.edu; jimjc@mit.edu; liyu@cse.cuhk.edu.hk

Methods

The RhoFold+ platform

MSA feature generation. We used the MSAs constructed by Infernal[47] and rMSA (https://github.com/pylelab/rMSA) to capture co-evolutionary information of the sequence as an additional input. Using Infernal, it is possible to locate homologous sequences with conserved secondary structures, while on the other hand, rMSA employs an iterative search strategy based on RNA sequence databases. We utilized the nucleic acid sequence databases Rfam and RNAcentral[48]. In AlphaFold2, a similar approach was used with different alignment tools and sequence databases. Given the need to produce several models and the constraints imposed by hardware memory, we reduced our fully extracted MSAs to a maximum of 256 sequences during the training phase. Subsequently, during the inference phase, 256 MSAs were either randomly selected or chosen through clustering and then fed into RhoFold+. We implemented clustering with conserved secondary structure or sequence embeddings from our pretrained RNA language model. Different sampled and clustered results can be thus used for multiple predictions, as marked by Top5, Top10 and so on. By default, the top 256 MSAs are chosen as input features for predicting the standard structure, which we refer to as standard RhoFold+. RhoFold+ (Top$K$) refers to the optimal model selected from K different models generated using distinct sampled MSAs.

RNA-FM language model. *Overview of RNA-FM*. Our foundation model provides meaningful representations that are inferred from standalone sequence information. These representations may improve performances in various downstream tasks, especially for those with insufficient annotated data. Inspired by recent studies[49,50], we leverage a general transformer architecture. In particular, our framework was built on the bidirectional transformer language model proposed in BERT (Bidirectional Encoder Representations from Transformers)[51], followed by the unsupervised training scheme. We named our framework 'RNA-FM' as it represents a foundational model for future RNA-related studies (Supplementary Fig. 2). Below, we detail how we constructed the large-scale noncoding RNA (ncRNA) dataset, followed by model and training details.

Large-scale pretraining dataset. The large-scale dataset used in the pretraining phase was collected from RNAcentral[48], the largest ncRNA dataset available to date. This dataset is a comprehensive collection of ncRNA sequences, representing all ncRNA types from a broad range of organisms. It combines ncRNA sequences across 47 different databases, resulting in a total of ~27 million RNA sequences (Supplementary Tables 2–4).

We preprocessed all ncRNA sequences by replacing all instances of 'T' with 'U' since they are both complementary to adenine and similar in structure ('T' representing thymine in DNA, while 'U' is for uracil in RNA). This resulted in a dataset involving four main types of bases (16 counted types of combinations in total: 'A', 'C', 'G', 'U', 'R', 'Y', 'K', 'M', 'S', 'W', 'B', 'D', 'H', 'V', 'N' and '–'). Moreover, to minimize redundancy without compromising the size of our dataset (that is, to preserve as many sequences as possible), we removed duplicate sequences using Cd-hit-est, which was set to a 100% similarity threshold. After the above preprocessing steps, a final, large-scale dataset consisting of over 23.7 million ncRNA sequences was obtained. We named this final dataset 'RNAcentral100', and we used this dataset to train our RNA foundation model in a self-supervised manner (see Supplementary Information for more details).

RNA-FM training details. Our RNA-FM framework comprises 12 transformer–encoder blocks, inspired by BERT[49,51]. Each block includes a 640 hidden size feed-forward layer and a multihead self-attention layer with 20 heads, along with layer normalization and residual connections applied pre- and postblock, respectively. For an RNA sequence of length L, RNA-FM takes raw sequential tokens as input, mapping each nucleotide into a 640 dimensional vector via an embedding layer, forming an $L \times 640$ embedding matrix. This matrix passes through each encoder block, retaining its size throughout and is followed by a softmax layer to predict corresponding tokens, including 16 nucleotides and four specific functional identifiers. Additional model details are in Supplementary Information.

During pretraining, we employed self-supervised training akin to BERT[51], randomly replacing 15% of nucleotide tokens with a special mask token. If the ith token was chosen, it was replaced with (1) the (MASK) token 80% of the time, (2) a random token 10% of the time and (3) left unchanged 10% of the time. We trained the model using masked language modeling (MLM)[51], predicting the original masked token via cross-entropy loss. This training strategy is formulated as an objective function as follows:

$$\mathcal{L}_{\mathrm{MLM}} = \mathbb{E}_{x\sim\mathcal{X}}\mathbb{E}_{x_{\mathcal{M}}\sim x}\sum_{i\in\mathcal{M}} -\log p(x_i|x_{/\mathcal{M}}). \quad (1)$$

A set of indices $\mathcal{M}$ is randomly sampled from each input sequence x, covering 15% of the sequence, and the corresponding tokens are replaced with mask tokens. For each masked token, given the masked sequence ($x_{/\mathcal{M}}$) as context, the objective function minimizes the negative log-likelihood of the true nucleotide x_i. This approach captures dependencies between the masked and unmasked parts of the sequence, leading to accurate predictions for masked positions. Training with the objective function in equation (1) allows RNA-FM to effectively model representations of each sequential token. We trained RNA-FM on eight 80 GB A100 graphics processing units (GPUs) for 1 month, using an inverse square root learning rate schedule with a 0.0001 base rate, 0.01 weight decay and 10,000 warm-up steps. To optimize memory usage and batch size, we set the maximum input sequence length to 1,024, accelerating the training process.

Efficient development of a self-distillation dataset. Although our RNA-FM can alleviate the problem of data scarcity, there is still less structural data available for RNAs than for proteins. As a result, we collected a nonredundant, self-distillation dataset with ground truth secondary structure from the RNAStralign and bpRNA-1M databases. We filtered this dataset by removing sequences with more than 256 or fewer than 16 nucleotides, resulting in a dataset of 27,732 sequences. RhoFold+ was initially trained using only PDB data, which was then used to generate a self-distillation dataset by inferring pseudo-structural labels. We retrained the model by sampling 25% of the PDB data and 75% of the distillation data for further improvement. During training, we masked out pseudo-label residues with pLDDT scores <0.7 and uniformly subsampled the MSAs to augment the distillation dataset.

A structure prediction module. The structure module of RhoFold+ aims to predict the 3D structure of an RNA based on the sequence and pair representation extracted by Rhoformer. The structure module of AlphaFold2 directly predicts the rotation and translation matrices of the backbone frames, as these are the most influential factors in protein folding. However, RNA folding is primarily driven by nucleotide base pairing. Due to their irregular structural patterns, directly predicting the base frame (C1′, N1/N9, C2/C4) defined over the nucleotides may pose a convergence problem in our experiments (Supplementary Table 1). To efficiently reconstruct the RNA full-atom coordinates, we used frame (C4′, C1′, N1/N9) and four torsion angles α, β, γ and ω to resolve this issue. Supplementary Table 1 provides the definitions of torsion angles and corresponding rigid groups. The 3D positions are modeled using IPA, a geometry-aware attention operation. On the basis of Rhoformer's output features and pair presentation, the IPA operation predicts the rotation and translation matrices for each frame. In addition, the predicted structure is refined iteratively using a recycling strategy, in which the Rhoformer receives the prediction from the

previous iteration. The recycling process ends when the pLDDT, which is one of the outputs generated by IPA that measures the quality of the predicted 3D structure, converges. With the reconstructed full-atom coordinates, biological constraints, such as base pairing, can be enforced directly in 3D space to optimize the structure module and generate biologically valid structural predictions.

Feature processing with Rhoformer. As with the Evoformer introduced in AlphaFold2, our main module, Rhoformer, is composed of a series of transformer modules with gated self-attention layers, which are employed to learn evolutionary information and simultaneously update the pairwise sequence embeddings and MSA representations. A transition block comprising two linear layers is added to the resulting pair and MSA representations to increase the embedding dimension by a factor of four, thereby increasing the model's capacity. Lastly, four self-attention blocks are stacked on the Rhoformer to refine the pair and MSA representations. These representations are then fed into the structure module to obtain the predicted full atom coordinates in three-dimensional space, as described in the following sections.

The structure prediction loss. The loss function is defined at 1D, 2D and 3D levels. Each of these levels is discussed in detail below. We first employed a MLM loss L_{mlm} to improve the extraction of co-evolutionary information from the MSAs at the 1D level without adding curated correlation features. In our experiments, 5% of the nucleotides were randomly masked and a linear projection layer was utilized to reconstruct them.

At the 2D level, a distance loss L_{dis} and a secondary structure loss L_{ss} were applied to supervise RhoFold+ to learn the pairwise positional correlations between each residue. In particular, three feed-forward layers were used for distance prediction to predict the pairwise distance between the P, C4 and N atoms. The distance was divided into 40 bins, where the first and last bins indicate <2 Å and >38 Å, respectively, and the distances between 2 Å and 38 Å were evenly divided into 36 bins. Additionally, the cross-entropy loss was used to determine whether the distance predictions belong to the correct bin. For the secondary structure prediction loss L_{ss}, a feed-forward layer was leveraged on top of pairwise features to predict the secondary structure. The secondary structure C is a $L \times L$ binary matrix, where L denotes the sequence length, and $C_{i,j} = 0$ or 1 indicates if the ith and jth residue from a base pair.

At the 3D level, gradients were derived from the main frame aligned point error (FAPE) loss, denoted as L_{FAPE}, the secondary structure constraint loss and the clash violation loss L_{clash}. AlphaFold2's FAPE loss compares a set of predicted atom coordinates under a set of predicted local frames with the corresponding ground truth atom coordinates and ground truth local frames. Loss is independent of rigid motions. The loss remains constant when the predicted structure differs from the actual structure by arbitrary rotation and translation.

The secondary structure constraint loss, L_{ss3d}, encodes secondary structural information directly into 3D prediction. To unify the calculation of different types of base pairing constraints in 3D space, we introduced four fixed pseudo atoms (T1, T2, T3 and T4) in the local coordinate system of a base[52] (Supplementary Fig. 1). L_{ss3d} aims to constrain the pseudo atoms in two base-paired nucleobases to satisfy the base-pairing property (base–base interactions). For two residue m and n, we computed the pairwise distance of the fixed points: $\mathbb{D}^{m,n} = \{d_{i,j}^{m,n} | i,j \in \{1,2,3,4\}\}$, where m and n denote two RNA residues and i and j are the indexes of the four atoms. We defined L_{ss3d} as follows:

$$L_{\text{ss3d}} = \sum_{\substack{m=1 \\ n=1}}^{N_{\text{nbpairs}}} \max\left(\hat{d}_{i,j}^{m,n} - \tau - d_{i,j}^{m,n}, 0\right), \tag{2}$$

where m and n are the indices of two residues that form a base pair; $i,j \in \{1,2,3,4\}$ denote the index of four pseudo atoms; $\hat{d}_{i,j}^{m,n}$ is the distance between two pseudo atoms i and j in the predicted structure; $d_{i,j}^{m,n}$ is the corresponding standard pairwise distance; N_{nbpairs} is the number of all base pair residues in this structure and τ is a tolerance distance threshold. The L_{ss3d} penalizes pairwise atom distances in the nucleotides when two residues form a base pair. The calculation of the standard pairwise distance $d_{i,j}^{m,n}$ is divided into two scenarios: (1) when the training sample comes from PDB data with 3D native structures, $d_{i,j}^{m,n}$ comes directly from the structure and (2) when the training sample comes from self-distilled data, $d_{i,j}^{m,n}$ are the statistical values generated from all PDB structures of the corresponding type of base pair. This can prevent RhoFold+ from overfitting the pseudo-labels and make full use of secondary structure information.

L_{clash} expects the model to learn to avoid atom clashes by penalizing distances that are too short between atoms according to their van der Waals radii. Additionally, we employ a loss, L_{pLDDT}, to train an LDDT evaluator that scores the predicted 3D RNA models as an indicator for global recycling (as introduced above). The purpose of the L_{pLDDT} loss is to train an LDDT evaluator that predicts the LDDT of the predicted 3D model based on the ground truth structure. The LDDT value is discretized with a 0.02 bin interval into 50 bins. Once a predicted 3D model has been generated, its LDDT is computed against the ground truth structure as the ground truth pLDDT label, and the LDDT evaluator generates the predicted pLDDT bin. Cross-entropy loss is used as L_{pLDDT} to determine whether the predicted LDDT falls within the ground truth bin.

The overall loss function is

$$\begin{aligned} L = {} & L_{\text{mlm}} + 0.3 \times L_{\text{dis}} + 0.1 \times L_{\text{ss}} + 0.03 \times L_{\text{clash}} \\ & + 2 \times L_{\text{FAPE}} + 0.1 \times L_{\text{ss3d}} + 0.01 \times L_{\text{pLDDT}}. \end{aligned} \tag{3}$$

Structure relaxation by force fields. As a preventive measure to resolve any remaining structural clashes and violations, we may relax our model predictions using a restrained energy minimization procedure, such as AMBER[53] and BRiQ[52]. Specifically, we minimized the AMBER force field using harmonic restraints, allowing the system to maintain a close relationship with its input structure. This postprediction relaxation also enforces the geometric features of phosphodiester bonds. Our empirical evidence indicates, as measured by r.m.s.d. and TM score, that while this final relaxation does not improve the model's accuracy, it eliminates distracting stereochemical violations without compromising accuracy.

Implementation details and running time. We used the Adam optimizer with a 0.0003 learning rate for 300,000 iterations, alongside a polynomial decay scheduler with 10,000 warm-up steps and a batch size of 16. A dropout ratio of 0.1 was applied to the Rhoformer and structure modules during training. The hardware setup included a GPU cluster with 768 GB memory and eight NVIDIA A100 GPUs (80 GB each), supported by an Intel Xeon Gold 6230 central processing unit (CPU) @ 2.10 GHz with 64 cores. RhoFold+ was trained for 1,600 epochs over 300,000 iterations, taking approximately 1 week. Posttraining, the inference is rapid, with RhoFold+ predicting a structure in about 0.14 s on a single A100 GPU. For FARFAR2 benchmarking, which demands significant computational resources, a Slurm job was run on the cluster using a single central processing unit core and 8 GB memory, with execution times detailed in Supplementary Table 9.

Running other baselines

In our benchmarking experiments, we obtained DeepFoldRNA, DRfold, RoseTTAFold2NA, FARFAR2 and trRosettaRNA (v1.0) from their official code repositories (either their homepages or their GitHub repositories). The authors of AlphaFold3 did not publish the code, so we used their server to perform predictions. For FARFAR2, we followed the default settings specified in the corresponding code documentation or on the server and we trained 100 models. Our benchmarking and evaluation used CASP15 natural RNA targets and newly determined

single-stranded RNA structures. Following CASP15 guidelines, we collected five candidate models for each competing method to compute r.m.s.d. and *Z*-score. For AlphaFold3, which generates five models per input sequence per run, we conducted one run and collected the five models. For RhoFold+, we ran it five times with different sampled MSAs, and for the other methods, we collected their five models from CASP15 website. For the newly determined single-stranded RNA structures, we ran the default configurations of RhoFold+ and other methods to produce default predictions for each sequence. For AlphaFold3, only the top model ('model_0') of a single run was evaluated. The input of all methods consisted solely of RNA sequences, without ions or other molecules.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in our work were obtained from related public datasets. We obtained all the RNA 3D structures using the data list arranged by BGSU RNA representative sets (version 2022-04-13) (http://rna.bgsu.edu/rna3dhub/nrlist/release/3.226) and downloaded them from the PDB (https://www.rcsb.org). For pretraining our language model (RNA-FM), we downloaded the unannotated RNA sequences from RNAcentral (https://rnacentral.org/). For RNA MSA construction, we built the database using a nucleotide database (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz), Rfam (https://rfam.org) and RNAcentral (https://rnacentral.org) and use rMSA (https://github.com/pylelab/rMSA) for searching and construction tools. We used secondary structural information for self-distillation. For this data, we downloaded the bpRNA dataset from SPOT-RNA at https://sparks-lab.org/server/spot-rna/, bprna-1m data from https://bprna.cgrb.oregonstate.edu/ and used RNAStralign, based on E2Efold available via GitHub at https://github.com/ml4bio/e2efold. The family/type information in Rfam (https://rfam.xfam.org) was used for cross-family/type validation. For RNA-Puzzles, we downloaded native structures and submissions of other methods from GitHub at https://github.com/RNA-Puzzles/standardized_dataset and http://www.rnapuzzles.org/results/, respectively. Similarly, CASP15 data were obtained via https://predictioncenter.org/casp15/index.cgi. Source data are provided with this paper.

Code availability

For the RhoFold+ model, trained weights and inference scripts are available under an open-source license via GitHub at https://github.com/ml4bio/RhoFold. RhoFold+ is also freely available as a server for academic purposes at https://proj.cse.cuhk.edu.hk/aihlab/RhoFold/#/. Our pretrained language model (RNA-FM) and its inference pipeline can be found via GitHub at https://github.com/ml4bio/RNA-FM. The RNA MSA search was performed by combining Infernal (http://eddylab.org/infernal/), Blastn (https://blast.ncbi.nlm.nih.gov/Blast.cgi), HMMER (http://hmmer.org) and rMSA (https://github.com/pylelab/rMSA), we also used openmm 7.7 for AMBER force field relaxation. Source codes are written under Python 3.7. We also utilized the following software for data collection, data analysis and visualization: Infernal 1.1.3 (cmbuild, cmcalibrate, cmscan, cmsearch), Cd-hit 4.8.1 (cd-hit-est), HMMER 3.3 (nhmmer), HH-suite 2.0.15, numpy 1.21.2, PyTorch 1.10.2, pandas 1.3.1, matplotlib 3.4, scikit-learn 0.24, scipy 1.7.1, biopython 1.79, PyTorch-Ignite 0.4.6 and TensorBoard 2.6.0.

References

47. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
48. Sweeney, B. A. et al. Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **49**, D212–D220 (2021).
49. Vaswani, A. et al. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems 30 (NIPS 2017)* (eds Guyon, I. et al.) 5998–6008 (Curran Associates, 2017).
50. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
51. Kenton, J.D.M.-W.C. & Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT 2019* Vol. 1 (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019).
52. Xiong, P., Wu, R., Zhan, J. & Zhou, Y. Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement. *Nat. Commun.* **12**, 2777 (2021).
53. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 198–210 (2013).

Acknowledgements

This work was supported by the Chinese University of Hong Kong (CUHK; award numbers 4937025, 4937026, 5501517, 5501329, 8601603, 8601663 and SHIAE BME-p1-24 to Y.L.) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Hong Kong SAR; project numbers CUHK 24204023 to Y.L., CUHK 14222922 and RGC GRF 2151185 to I.K.). Additional support was provided by the Innovation and Technology Commission of the Hong Kong SAR (project no. GHP/065/21SZ to Y.L.). J.W. was supported by a Hong Kong PhD Fellowship (award no. PF22-73180) from the Research Grants Council of the Hong Kong SAR, China and an IdeaBooster Fund (project no. IDBF24ENG06) from CUHK. The work is part of the Antibiotics-AI Project, directed by J.J.C., and supported by the Audacious Project, Flu Lab, LLC, the Sea Grape Foundation, R. Zander, H. Wyss for the Wyss Foundation, and an anonymous donor. F.W. was supported by the National Institute of Allergy and Infectious Diseases of the NIH (award no. K25AI168451). D.L. acknowledges the startup funding from ASU. We thank X.-J. Lu for guidance on interhelical twist experiments. This work was partly supported by the Shenzhen-HongKong Joint Funding Project (Category A) under grant no. SGDX20211123112401002 (to S.W.) and no. SGDX20230116092056010 (to S.W.). This project was partially supported by funds from the Focus Project of AI for Science of Comprehensive Prosperity Plan for Disciplines of Fudan University (to S.S.) and Shanghai Artificial Intelligence Laboratory (to S.S.).

Author contributions

Y.L. conceived the study. Z.H., T.S., S.S., J.C., L.H., S.W. and Y.L. conducted the study. Z.H., T.S. and J.C. processed the data. Z.H., D.L., J.W., Y.W., J.C., F.W. and Y.L. wrote the manuscript. Z.H., D.L., F.W. and Y.L. designed the experiments. L.Z. assisted with the relaxation section. J.X. assisted with setting up the server. T.K. and D.L. assisted with implementing the helical experiments. I.K., J.J.C. and P.Y. co-supervised the study.

Competing interests

S.W. and L.Z. are the co-founders of Zelixir Biotech Co. Ltd. P.Y. is a co-founder, equity holder, board member and consultant of Ultivue, Inc., Spear Bio, Inc. and Digital Biology, Inc. J.J.C. is the founding scientific advisory board chair of Integrated Biosciences. F.W. is a co-founder of Integrated Biosciences. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02487-0.

Correspondence and requests for materials should be addressed to Siqi Sun, Di Liu, Sheng Wang, Peng Yin, James J. Collins or Yu Li.

Peer review information *Nature Methods* thanks Hashim Al-Hashimi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Arunima Singh, in collaboration with the *Nature Methods* team.

nature research

Corresponding author(s): Yu Li

Last updated by author(s): Sep 10, 2024

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	cmbuild, cmcalibrate, cmscan and cmsearch from INFERNAL 1.1.3, cd-hit-est from CD-HIT 4.8.1, Infernal 1.1.4, nhmmer from HMMER 3.3, HH-suite 2.0.15
Data analysis	numpy=1.21.2, an open-source python package for numerical calculations. pytorch=1.10.2, an open-source deep learning framework in python. pandas>=1.3.1, an open-source python package for data analysis. matplotlib>=3.4, an open-source python package for visualization. scikit-learn>=0.24, an open-source python package for machine learning. scipy=1.7.1, an open-source python package for mathematics, science, and engineering. biopython=1.79, an open-source python package for biological computation. pytorch-ignite=0.4.6, an open-source python package for training and evaluating neural networks in PyTorch flexibly and transparently. openmm=7.7, a high-performance toolkit for molecular simulation including AMBER relaxation. tensorboard=2.6.0, an open-source python package for visualizing training process. RNA-FM, our custom code (open source) for (https://github.com/ml4bio/RNA-FM) pretraining. rMSA (https://github.com/pylelab/rMSA), an open source code for RNA MSA construction. RhoFold, our primary model code (open source) for (https://github.com/RFOLD/RhoFold) structure prediction.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used in our work were obtained from related public datasets. We obtained all the training RNA 3D structures using the data list arranged by BGSU RNA Representative Sets (Ver.04-13.2022) (http://rna.bgsu.edu/rna3dhub/nrlist/release/3.226), all train and test data by BGSU RNA Representative Sets (Ver.02-15.2023) (http://rna.bgsu.edu/rna3dhub/nrlist/release/3.270)
and downloaded them from Protein Data Bank (https://www.rcsb.org).
For RNA-FM pre-training, we downloaded the unannotated RNA sequences from RNAcentral (https://rnacentral.org).
For RNA MSA, we built the database upon Nucleotide database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz), Rfam (\https://rfam.xfam.org) and RNAcentral (\https://rnacentral.org/) and use rMSA (https://github.com/pylelab/rMSA) for searching and construction tools.
We used secondary structural information for self-distillation training and testing. For such information, we downloaded the bpRNA dataset from SPOT-RNA at https://sparks-lab.org/server/spot-rna/, bpRNA-1m data from https://bprna.cgrb.oregonstate.edu/ and RNAStralign based on E2Efold from https://github.com/ml4bio/e2efold. The family/type information in Rfam (https://rfam.xfam.org) is used for cross-family/type validation. Secondary structures for PDB data from PDB website and ArchiveII from https://rna.urmc.rochester.edu/publications.html.
For RNA-Puzzles, we downloaded native structures and submissions of other methods from http://www.rnapuzzles.org/results/ and https://github.com/RNA-Puzzles/standardized_dataset.
Similarly, CASP15 data is downloaded via https://predictioncenter.org/casp15/index.cgi

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We use 16,349 single chain RNA 3D structures extracted from 4213 RNA 3D structures in the PDB database for training. RNA-FM is trained on 23,735,169 sequences extracted from the RNACentral database in an unsupervised manner to generate rich information to benefit structural prediction. We used 2,4183 RNA secondary structure data in bprna90-1m for self-distillation training.
Data exclusions	All data were used following previous researches , no exclusion was done prior to analysis.
Replication	The performance of our model could be reproduced, and we also offer codes, on-line service, and tutorials for the key downstream tasks.
Randomization	There are 3 cross-validation experiments in our tests. 1. 10-fold: After obtain 782 sequence clusters from CD-HIT, we randomly masked 80 sequence clusters for validation, leaving 702 sequence clusters for training. 2. cross-type/cross-fam: RNA types and families are obtained via Rfam. We then randomly selected an RNA type/fam and masked all structures (structure number depending on the RNA type) associated that type/fam, trained the model using the remaining types/families, and evaluated the model on the masked type/fam.
Blinding	Not applicable. All experiments, data collection, and analysis are performed in an unbiased way.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging