nature computational science

Brief Communication

https://doi.org/10.1038/s43588-024-00720-6

Deep generative design of RNA aptamers using structural predictions

Received: 22 February 2024

Accepted: 7 October 2024

Published online: 06 November 2024

Check for updates

Felix Wong[1,2,3,12], Dongchen He[4,5,12], Aarti Krishnan[1,2], Liang Hong[4], Alexander Z. Wang[1,2], Jiuming Wang[4], Zhihang Hu[4], Satotaka Omori[1,3], Alicia Li[3], Jiahua Rao[6], Qinze Yu[4], Wengong Jin[7,8], Tianqing Zhang[2], Katherine Ilia[2], Jack X. Chen[2], Shuangjia Zheng[9], Irwin King[4], Yu Li[1,2,4,10,11] & James J. Collins[1,2,11]

RNAs represent a class of programmable biomolecules capable of performing diverse biological functions. Recent studies have developed accurate RNA three-dimensional structure prediction methods, which may enable new RNAs to be designed in a structure-guided manner. Here, we develop a structure-to-sequence deep learning platform for the de novo generative design of RNA aptamers. We show that our approach can design RNA aptamers that are predicted to be structurally similar, yet sequence dissimilar, to known light-up aptamers that fluoresce in the presence of small molecules. We experimentally validate several generated RNA aptamers to have fluorescent activity, show that these aptamers can be optimized for activity in silico, and find that they exhibit a mechanism of fluorescence similar to that of known light-up aptamers. Our results demonstrate how structural predictions can guide the targeted and resource-efficient design of new RNA sequences.

In addition to their roles as carriers of transcriptional information, RNAs perform a myriad of biologically relevant functions including catalyzing biochemical reactions[1,2], regulating transcription[3], signaling[4] and binding to other molecules[5]. RNA aptamers that bind to target molecules can inhibit viral replication[6], facilitate biomolecular detection[7–11] and generate fluorescence[12]. In these contexts, the three-dimensional (3D) structures of RNA aptamers play key roles in their functions[13]. Deep learning methods for the accurate prediction of protein 3D structure—including AlphaFold and RoseTTAFold—are now available[14,15]. However, deep learning methods for the accurate prediction of RNA 3D structure have only recently emerged, due to various challenges including the relative scarcity of RNA 3D structures for training and their conformational flexibility[16–19].

Although the top-performing groups at the previous CASP15 competition for RNA 3D-structure prediction required expert knowledge and fine-tuning[20], several deep learning-based methods for predicting RNA 3D structure from sequence, including RhoFold[16], trRosettaRNA[17], DeepFoldRNA[18] and AlphaFold 3[19], are now capable of being fully automated. While diverse approaches to RNA design focusing on secondary structures, including LEARNA[21] and RiboLogic[22], have been developed, here we hypothesized that in silico platforms predicting RNA 3D structure can enable the structure-informed design of novel RNA aptamers. We tested this hypothesis by developing an approach for the generative design of RNAs using structural predictions (Fig. 1a). Given a 3D point cloud of a target structure, we aimed to solve the

[1]Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Institute for Medical Engineering & Science and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Integrated Biosciences, Redwood City, CA, USA. [4]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China. [5]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [6]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. [7]Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [9]Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, China. [10]The CUHK Shenzhen Research Institute, Shenzhen, China. [11]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA. [12]These authors contributed equally: Felix Wong, Dongchen He. e-mail: liyu@cse.cuhk.edu.hk; jimjc@mit.edu

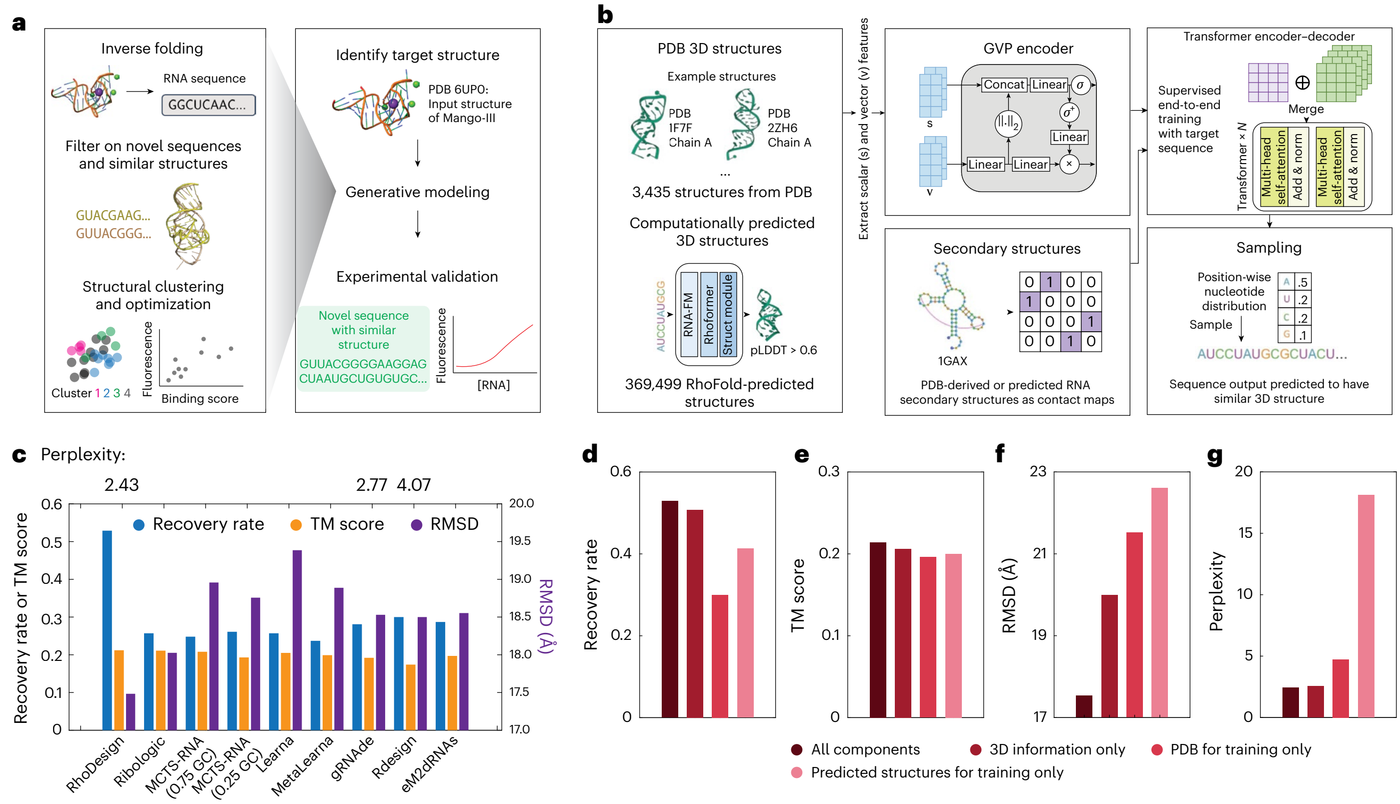


Fig. 1 | A deep learning approach for the 3D-structure-based generative design of RNAs. a, Schematic of the approach. RhoDesign is trained on 3D RNA structures predicted using a sequence-to-structure model (RhoFold), in addition to available 3D RNA structures from the PDB. RhoDesign solves the problem of producing RNA sequences whose 3D structures are predicted to approximate an input structure, and its designed sequences can be synthesized, experimentally validated and optimized for activity. **b**, Architecture of RhoDesign. 3D structural information is first encoded by a GVP encoder, and the output is combined with secondary-structure information and passed into a transformer. The output of the transformer is supervised using ground-truth sequences during training, and RhoDesign outputs designed RNA sequences through sampling. **c**, Benchmarking RhoDesign and other models for RNA design. The recovery rate (which measures the fraction of nucleotides in ground-truth sequences predicted by the model; higher is better), TM score (which measures the fidelity of the predicted 3D structures of the output sequences, as compared with the input structure; higher is better), RMSD (lower is better) and perplexity (which measures how well a model can predict a given sequence; lower is better) on a withheld test set were used. Optimal values of these metrics, corresponding to those calculated for perfect sequences, are 1, 0.247, 17.45 Å and 1, respectively (Benchmarking metrics). Models were used out of the box with default parameters where applicable; for MCTS-RNA, 0.75 and 0.25 GC indicate the target GC content of the RNA sequence. Perplexity values are provided for RhoDesign, gRNAde and RDesign, as only these models allow for its calculation. **d**–**g**, Ablation studies. RhoDesign was evaluated on a withheld test set when trained using a combination of 3D- and secondary-structure information or 3D-structure information only, as well as when trained using 3,435 PDB structures only or 369,499 RhoFold-predicted structures only.

inverse problem of producing sets of candidate RNA sequences whose structures are predicted to approximately match the input structure.

Results

A platform for the generative design of RNAs using structural predictions

Our platform, RhoDesign, builds on the heuristics and structural predictions developed as part of RhoFold[16], but is generalizable across different sequence-to-structure models including trRosettaRNA[17], DeepFoldRNA[18] and AlphaFold 3[19]. Here, we have focused on using RhoFold, building on its promising performance when benchmarked against other sequence-to-structure models[16]. We first curated a set of 369,499 RhoFold-predicted 3D RNA structures representative of different types of RNA (Supplementary Fig. 1), as well as 3,435 3D RNA structures from the Protein Data Bank (PDB)[23], and used these structures along with their corresponding sequences to train a deep learning model comprising a geometric vector perceptron (GVP) encoder[24] and a transformer module[25,26] (Fig. 1b). Predicted or ground-truth secondary structures corresponding to the 3D structures were provided as additional inputs to improve accuracy (‘Structure-to-sequence model’). After model training, we benchmarked the trained model using the recovery rate, the TM (template modeling) score, the root-mean-square deviation (RMSD) and the perplexity (Fig. 1c–g and ‘Benchmarking metrics’).

When benchmarked on training–testing subsets of our training set, we found that RhoDesign outperformed alternative models, including LEARNA[20], Meta-LEARNA[20], RiboLogic[21], Monte Carlo tree search (MCTS)-RNA[27], gRNAde[28], RDesign[29] and eM2dRNAs (enhanced M2dRNAs)[30] (Fig. 1c, Supplementary Table 1 and ‘Comparison with other models’). Because here the TM score and RMSD depend on RhoFold-predicted 3D structures, these metrics are bounded by imperfect values corresponding to fully recovered sequences, and we find that RhoDesign-generated sequences approach these bounds (Fig. 1c). Additional analyses demonstrate RhoDesign’s promising performance in cross-fold validation experiments and suggest that inclusion of both the RhoFold-predicted and PDB components of the training set, as well as secondary-structure inputs, are important for improving performance (Fig. 1d–g and Supplementary Table 1). Benchmarking RhoDesign models trained on PDB data only, with the GVP encoder, transformer encoder and transformer decoder modules ablated, indicated that each of these modules contributes to performance (Supplementary Table 2). These findings suggest the ability of the GVP and

https://doi.org/10.1038/s43588-024-00720-6

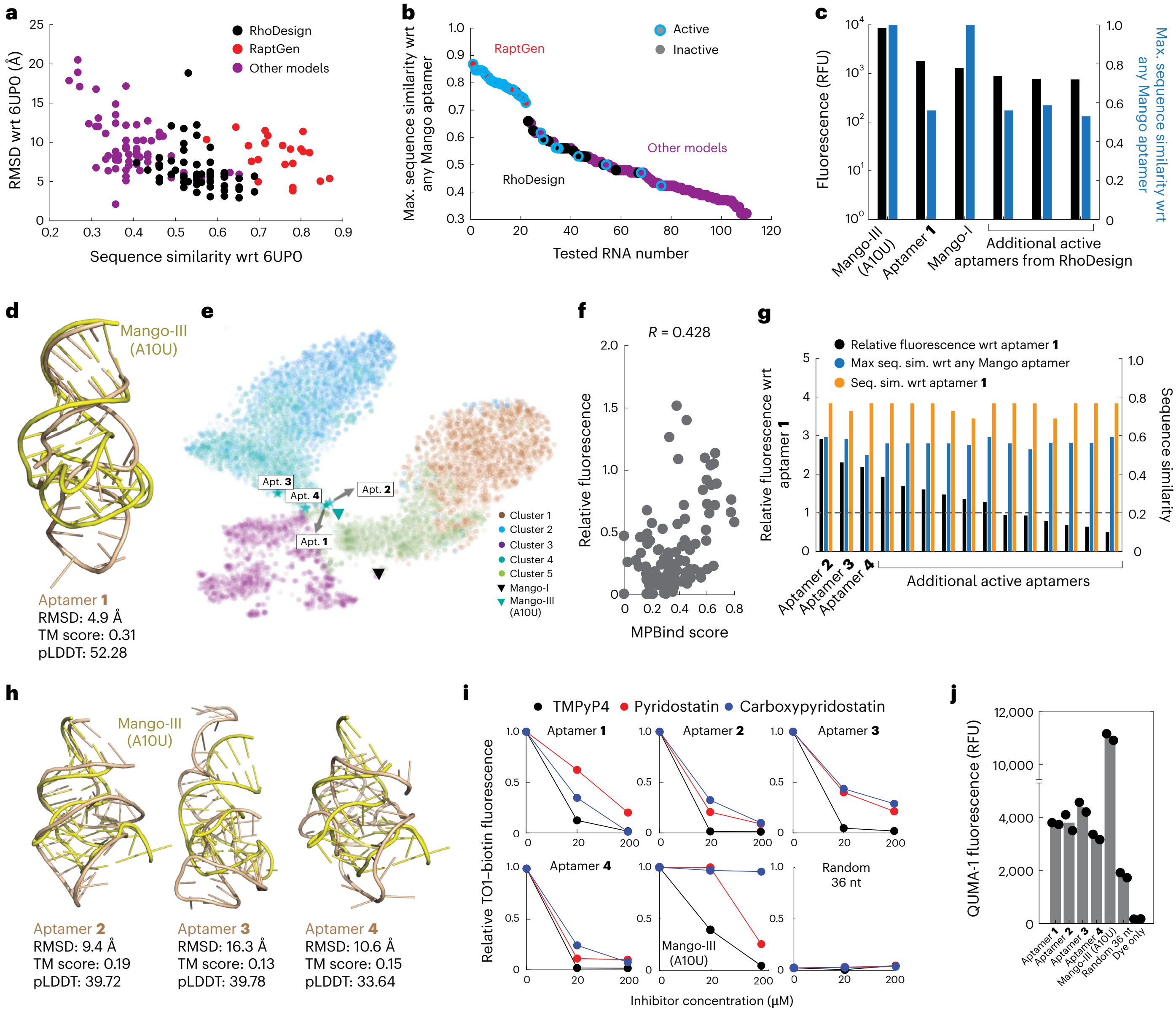


Fig. 2 | Experimental validation, optimization and mechanism of fluorescence of generated light-up RNA aptamers. a, Plot of sequence similarity against predicted RMSD—used here as a starting point for measuring structural fidelity that is correlated with TM scores—with respect to PDB 6UPO for 60 RhoDesign-generated aptamers, 22 RaptGen-generated aptamers and 70 aptamers generated using seven other structure-to-sequence models. RaptGen is a variational autoencoder developed for SELEX data, which was trained using sequence information from Mango aptamers alone. For each set of generated sequences, where applicable, sequences were downsampled to consider only those with the lowest RhoFold-predicted RMSD with respect to Mango-III in 6UPO. Overall, the 70 aptamers generated using other structure-to-sequence models were the least sequence similar and most structurally divergent from Mango-III. **b**, Rank-ordered maximal sequence similarity with respect to any Mango aptamer of 110 RNA aptamers synthesized and tested for TO1–biotin fluorescence induction. **c**, Fluorescence and sequence similarity of active RNA aptamers. Fluorescence was measured for each RNA by reacting 25 μM RNA with 10 μM TO1–biotin, and values represent one of two similar biological replicates. Mango-III, as expected from it being a highly optimized variant of Mango-I, exhibited about four times and about seven times brighter fluorescence than did aptamer **1** and Mango-I, respectively. RFU, relative fluorescence units. **d**, Predicted 3D structure of aptamer **1**, aligned to the ground-truth structure for Mango-III in PDB 6UPO. pLDDT, predicted local distance difference test. **e**, Spectral clustering using the predicted 3D structures of 10,000 RhoDesign-generated sequences, in addition to the predicted 3D structures of Mango aptamers. **f**, Correlation between MPBind score and relative fluorescence intensity for the 110 designed and tested aptamers shown in **b**. The Pearson correlation coefficient, R, is shown. **g**, Rank-ordered relative fluorescence of the 15 active aptamers from the additional 20 RNA aptamers that were synthesized and tested for TO1–biotin fluorescence induction, along with maximal sequence similarity with respect to any Mango aptamer and sequence similarity with respect to aptamer **1**. Fluorescence values represent one biological replicate. Relative fluorescence is with respect to aptamer **1** (dashed line). **h**, Predicted 3D structures of aptamers **2**–**4**, aligned to the ground-truth structure for Mango-III in 6UPO. **i**, Pharmacologic inhibition of RNA fluorescence using three small-molecule G-quadruplex binders. Values are normalized by the fluorescence intensities of each RNA with no inhibitor, except for values for a random 36-nucleotide RNA, which for comparison are normalized relative to corresponding Mango-III values. Each value represents one biological replicate. **j**, Detection of G-quadruplexes in active RNA aptamers using QUMA-1 dye. Bars represent the mean of two biological replicates (black circles).

https://doi.org/10.1038/s43588-024-00720-6

transformer architectures to learn better from datasets of size similar to those used to train other approaches.

Generation and validation of fluorescent RNA aptamers

Having benchmarked our model, we applied RhoDesign to generate RNA sequences from structures. We hypothesized that our approach could generate RNAs that are predicted to be structurally similar, but sequence dissimilar, to input RNAs. Given that the structures of RNAs impinge on their functions, we focused on generating fluorescent light-up RNA aptamers[12], whose binding to small molecules allows for rapid and quantitative testing of the designed sequences. Fluorogenic Mango aptamers binding to TO1–biotin, a small molecule, have been extensively characterized and structurally resolved in complex[31,32]. As a starting point, we considered the Mango-III (A10U) aptamer in PDB 6UP0. This aptamer was discovered using a large-scale microfluidic screen of variants of the original Mango aptamer, Mango-I[31]. We retrained RhoDesign on RNA structures to exclude structures with a sequence similarity greater than 0.5 with respect to the Mango-III aptamer in 6UP0 to test the models' ability to generalize. We then applied the trained models to generate 60 candidate sequences using the structure of 6UP0 as input, which, consistently, is well predicted using RhoFold (Fig. 2a and Extended Data Fig. 1). From the generated sequences, we downsampled candidates according to low predicted RMSDs and low sequence similarity with respect to the input aptamer. This resulted in 18 aptamers with predicted RMSDs between 2.9 and 7.5 Å, TM scores between 0.21 and 0.39 and sequence similarity between 0.43 and 0.65 that we selected for synthesis and empirical evaluation (Fig. 2b and Supplementary Data 1). For comparison, we also synthesized and evaluated 22 and 70 similarly downsampled aptamers generated using RaptGen[33] and seven other structure-to-sequence models, respectively.

Upon testing the synthesized aptamers for fluorescence in the presence of TO1–biotin (10 μM), we found that 4 of the 18 RhoDesign-generated aptamers exhibited activity, defined herein as exhibiting fluorescence brightness of at least half that of Mango-I—a lenient criterion, which nevertheless robustly filters out inactive aptamers (Fig. 2c and Supplementary Data 1). Indeed, random 36-nucleotide sequences that are similar in length to the active aptamers did not exhibit any activity. Twenty of the 22 RaptGen-generated aptamers also exhibited activity, yet all of these had maximal sequence similarities greater than 0.7 with respect to any Mango aptamer, indicating that these sequences are largely redundant (Fig. 2b). In contrast, the four active RhoDesign-generated aptamers exhibited a maximal sequence similarity of 0.59 with respect to any Mango aptamer; of these, aptamer **1** is notable in that it displayed higher fluorescence than did Mango-I (Fig. 2c). Only four of the 70 aptamers from the seven other structure-to-sequence models exhibited activity; of these, one was generated by gRNAde and three were generated by MCTS-RNA. The latter method yields a working hit rate of 30%, which is slightly higher than the working hit rate from RhoDesign (22%). However, none of the four active aptamers generated by these other methods displayed substantially stronger fluorescence than did aptamer **1**, and RhoDesign exhibited approximately twofold higher recovery rates than did MCTS-RNA in our benchmarks (Fig. 2b and Supplementary Data 1). Despite its predicted structural similarity to Mango-III, aptamer **1** did not contain conserved sequence motifs that are known to underlie the fluorescence activity of Mango aptamers, suggesting that it might be an unprecedented aptamer with Mango-like activity (Fig. 2d and Extended Data Fig. 1).

Optimization and mechanism of generated aptamers

To investigate aptamer **1** further, we aimed to model and optimize its fluorescence activity. To study whether the predicted 3D structure of aptamer **1** might provide insight into its function, we performed spectral clustering on the basis of 3D-aligned predicted structures (Fig. 2e). This revealed that, in contrast to other generated sequences based on Mango-I and Mango-III, aptamer **1** clustered more similarly to Mango-III than did Mango-I, suggesting that it was sampling a structural space associated with high fluorescence, and that 3D structural features conserved with Mango-III's may underlie its activity. Testing several approaches ('Binding predictions'), we found that using MPBind[34], a motif- and sequence-based statistical framework originally developed to process systematic evolution of ligands by exponential enrichment (SELEX) data, to score the 110 aptamers resulted in an encouraging correlation (Pearson's $R = 0.428$) with fluorescence activity (Fig. 2f).

Building on these results, we generated a set of aptamer **1** derivatives by providing the RhoFold-predicted structure of aptamer **1** as input to RhoDesign. We sampled 5,000 RNA sequences predicted to lie in a structural cluster surrounding aptamer **1**, filtered the sequences to ensure that they exhibited greater sequence similarity with respect to aptamer **1** than any Mango aptamer, and removed low (<0.4) MPBind-scoring sequences, resulting in 1,818 candidate aptamers. We then downsampled candidates on the basis of sequence by performing *t*-distributed stochastic neighbor embedding on one-hot encodings and selected the highest MPBind-scoring sequence in each *t*-distributed stochastic neighbor embedding cluster, resulting in 20 RNA aptamers. Synthesizing and testing these aptamers, we found that 15 of the 20 aptamers were active, and 9 of the 15 active aptamers exhibited higher activity than did aptamer **1** (Fig. 2g). All active aptamers exhibited low maximal sequence similarities (<0.6) with respect to any Mango aptamer and high sequence similarities (>0.6) with respect to aptamer **1** (Fig. 2g). We shortlisted the top three aptamers, aptamers **2**–**4**, for further study.

Aptamers **2**–**4** exhibited fluorescence intensities greater than that of aptamer **1**. Their predicted 3D structures, although more structurally dissimilar to Mango-III than aptamer **1**'s, differed only marginally (Fig. 2h and Extended Data Fig. 2), suggesting that they may fluoresce, in part, through a mechanism similar to that of Mango aptamers. Indeed, G-quadruplexes—tertiary structures formed through self-recognition of guanines—have been demonstrated to bind TO1 derivatives and other fluorescence-generating small molecules[35,36]. To study this potential mechanism further, we performed experiments relying on pharmacologic inhibition of aptamers **1**–**4** as well as direct detection of G-quadruplexes using a fluorescent dye (Fig. 2i,j). Administration of three different G-quadruplex-selective small-molecule binders, TMPyP4, pyridostatin and carboxypyridostatin[37–39], largely resulted in competitive inhibition of TO1–biotin fluorescence in aptamers **1**–**4** as well as Mango-III (Fig. 2i). Reacting QUMA-1, a fluorescent probe specific to G-quadruplexes[40], with aptamers **1**–**4** as well as Mango-III resulted in substantial increases in QUMA-1 fluorescence (Fig. 2j), consistent with the potential presence of G-quadruplexes in these aptamers. Together, these results indicate that aptamers **1**–**4**, and likely other generated aptamers, exhibit a mechanism of fluorescence similar to that of Mango-III despite their sequence dissimilarity, consistent with the notion that our approach leverages structural predictions to design new RNA sequences.

Discussion

Our work provides an in silico platform for the de novo design of RNA sequences that leverages structural predictions and is resource efficient (requiring the synthesis of a limited number of RNAs), which contrasts with traditional time- and resource-intensive approaches such as SELEX[41]. Other platforms, including gRNAde and MCTS-RNA, can contribute to the structure-guided design of active RNA aptamers, and we anticipate that integrating these methods with RhoDesign can further improve the accuracy of RNA design. Given the limitations of the current benchmarking, RhoDesign will also benefit from further validation using more reliable RNA 3D-structure prediction methods and additional experiment tests with diverse RNAs in the future.

 https://doi.org/10.1038/s43588-024-00720-6

Nevertheless, our study suggests a framework for the structure-informed design of additional aptamers, including therapeutic candidates[5,6] and diagnostic aptamers[7–11]. Further work, for instance in characterizing and making available more RNA 3D structures, will improve this type of approach and extend its use to designing other types of RNA with diverse functions.

Methods

Structure-to-sequence model

RhoDesign employs a GVP module to encode tertiary structures, learning vector-valued functions from backbone coordinates and scalar-valued functions from calculated dihedral angles, which capture geometric information[24]. As secondary-structure constraints may provide useful information, RhoDesign concatenates the output of the GVP encoder with the contact map derived from secondary-structure information, which for PDB structures was produced using DSSR[42] with default settings, and for RhoFold-predicted structures was produced using RhoFold (as detailed further below). Together, these components capture local and global structural features. RhoDesign then employs a transformer architecture[25], which may capture long-range dependences and relationships within encoded structures, to generate sequences. To address the challenge posed by limited RNA structure data, we employed RhoFold[16] to predict 3D structures for 369,499 RNAcentral sequences[43] and used these structures to augment the training of RhoDesign (as detailed further below).

After RhoDesign was trained, sequences were generated by sampling from the predicted distributions. Specifically, we utilized the torch.multinomial function for sampling. Additionally, we incorporated a temperature parameter to control the diversity of the samples. For all our benchmarking tests, we employed a relatively small temperature value of 1×10^{-5}, which resulted in the generation of nearly identical sequences for each sampling call. Hence, we sampled only one sequence per structure to maximize reproducibility. During the subsequent sequence design process, we were interested in sampling a diversity of sequences. To do so, we used a temperature of 1 and performed repeated sampling.

Training for RhoDesign was performed using eight NVIDIA RTX 3090 graphics processing units, and required approximately 16 h to complete. When generating new sequences using the trained model and a standard computer (for example, a current-generation MacBook Pro), RhoDesign exhibited an average generation time of 4.75 sequences per second.

RhoDesign architecture and training

RhoDesign is a deep learning model that enables RNA sequence design based on a fixed backbone structure. As described in the main text, the model includes the following components.

- *Geometric vector perceptron*. The GVP module encodes tertiary RNA structures. As implemented here, its architecture combines two key elements: the encoding of RNA backbone coordinates (C4′, C1′, N1) into vector-valued functions and the incorporation of scalar-valued functions derived from calculated dihedral angles.

    - First, the GVP module processes backbone coordinates by learning vector-valued functions. The vector values are derived from the directional vectors between specific atoms in the backbone. For instance, the vector from the C4′ of one nucleotide *i* to the C4′ of the next nucleotide ($C4'_i$ to $C4'_{i+1}$) provides a quantity that describes the orientation of one part of the backbone relative to the next. We associate each backbone coordinate (for example, C4′, C1′, N1) with a set of such vectors, each of which is mapped by the GVP to a high-dimensional vector representation capturing the spatial arrangement of the RNA backbone.
    - Additionally, the GVP module integrates information from calculated dihedral angles derived from all seven atoms (C4′, C1′, N1, C2, C5′, O5′, P). These dihedral angles are scalar quantities that capture information pertaining to the local geometry of the RNA molecule and are encoded as scalar-valued functions. These angles describe the rotational states around the backbone bonds and are invariant under rotations and translations of the backbone by construction.
    - Combining these encodings, the resulting representation captures both global and local features. As output, the GVP module produces a feature-rich representation of the input RNA structure: namely, the scalar and vector features are processed by a series of linear transformations, concatenations and nonlinear transformations based on the L_2 norm, preserving equivariance and invariance of the vector and scalar outputs, respectively, with respect to an arbitrary composition of rotations and reflections (as previously shown in ref. 24). This representation is then passed as input into a transformer module to enable sequence design.

- *Transformer*. The GVP-encoded structural features along with the contact map of the input are fed into the encoder of a transformer. The transformer consists of both an encoder and a decoder, and this architecture has been appreciated to capture long-range dependences and relationships within the encoded information[25]. The encoder, which processes the input representation (including the GVP-encoded structural features and the contact map), employs attention mechanisms to focus on relevant structural information. Notably, the attention mechanism may utilize the contact map to capture spatial dependences between base pairs. The decoder then processes the output of the encoder for sequence generation, operating in a step-by-step manner to produce output nucleotides.
- *Training details*. Model training was performed using eight NVIDIA GeForce RTX 3090 graphics processing units, and hyperparameters were fine-tuned to optimize model performance. The GVP module was configured with the following parameters: 15 top *K* neighbors, node hidden dimensions of 256 (vector) and 512 (scalar), edge hidden dimensions of 32 (scalar) and 1 (vector), three encoder layers and a dropout rate of 0.1. The transformer was configured with the following parameters: three encoder and decoder layers, each comprising four attention heads, and an attention dropout of 0.1. Both the encoder and decoder were defined using an embedding dimension of 512. The cross-entropy loss function was used.
- *Data*. As discussed above, we leveraged experimentally determined structures from the PDB. We utilized DSSR[42] with its default settings to extract contact maps from the PDB structures. These contact maps provide information about the spatial arrangement of base pairs within RNA molecules, augmenting model learning of structural features. Additionally, as discussed above, to address the limited availability of PDB data for training our models, we leveraged RhoFold-predicted structures for our model training. For these structures, the corresponding contact maps were directly generated by RhoFold.

Comparison with other models

LEARNA[20], Meta-LEARNA[20], RiboLogic[21], MCTS-RNA[27], gRNAde[28], RDesign[29] and eM2dRNAs[30] represent different models that have been developed for the task of RNA sequence generation.

- LEARNA leverages deep reinforcement learning to optimize RNA sequences by considering both secondary-structure and sequence constraints.

- Meta-LEARNA extends LEARNA by leveraging meta-learning techniques to enhance the model's adaptability and efficiency, enabling it to learn from previous RNA design tasks.
- RiboLogic uses riboswitch data to predict RNA sequences that optimize translation efficiency, leveraging secondary-structure and function information.
- MCTS-RNA performs inverse folding by employing a tree search algorithm to search RNA sequence space and a scoring function that incorporates structural and functional information.
- gRNAde is a multistate graph neural network that generates candidate RNA sequences conditioned on one or more 3D backbone structures.
- RDesign uses a hierarchical representation learning framework that learns 3D structural representations through contrastive learning at both the cluster and sample levels.
- eM2dRNAs is a multiobjective meta-heuristic algorithm for designing RNA sequences relying on recursive decomposition of a target structure.

These models differ in their algorithms, target applications and the manner in which they integrate structural information to design RNA sequences as outputs.

Benchmarking metrics

To benchmark the performance of RhoDesign against that of other models, we used the following evaluation metrics.

- *Recovery rate*. The recovery rate is a quantitative measure used to evaluate the effectiveness of a design method. It is calculated by comparing the designed RNA sequence to the natural RNA sequence and finding the fraction of correctly recovered nucleotides. A higher recovery rate indicates a more accurate and effective design method. By definition, the optimal value of recovery rate is 1.
- *Perplexity*. Perplexity is calculated as the exponentiated average negative log-likelihood per sequence. Given a sequence of nucleotides $s = n_1 n_2 n_3 \ldots n_N$, the perplexity is calculated using the following formula: $\mathrm{perplexity}(s) = P(n_1 n_2 n_3 \ldots n_N)^{-1/N}$, where $P(n_1 n_2 n_3 \ldots n_N)$ is the probability assigned by the model to the entire sequence relative to the sequences in the reference database. Perplexity is a metric used in natural language processing (NLP) to measure how well a language model can predict a given sequence of words. The metric reflects how surprised the model is when it encounters the next word in a sequence: the lower the perplexity, the better the model is at predicting the sequence. By definition, the optimal value of perplexity is 1.
- *TM score*. The TM score[44] is a scalar quantity between 0 and 1 which measures the global similarity between two 3D structures. It compares the structural alignment of two structures, with a higher score indicating a more similar structure. The key components of the TM score include the alignment of corresponding residues in the compared structures, the RMSD between their positions and a length-dependent normalization factor. In our experiments, TM score is used to evaluate the quality of the designed RNA sequences by comparing the predicted 3D structure of the designed RNA sequence with the target input 3D structure. Because RhoFold is used to predict 3D structures, there may be a non-zero average value of TM score arising from the difference between RhoFold-predicted 3D structures and PDB-deposited structures. Empirically, we found that this value was 0.247 for the benchmarking experiments shown in Fig. 1c. Our finding that the optimal TM score is low across all our benchmarking experiments, including those shown in Supplementary Table 1, suggests that 3D-structure prediction methods remain limited.
- *RMSD*. The RMSD is a scalar quantity with typical units of angstroms, which measures the average distance between atoms of aligned structures. As with the TM score, here the RMSD is used to evaluate the quality of the designed RNA sequences by comparing the predicted 3D structure of the designed RNA sequence with the target input 3D structure. By definition, the optimal value of RMSD is 0 Å; because RhoFold is used to predict 3D structures, there may be a non-zero average value of RMSD arising from the difference between RhoFold-predicted 3D structures and PDB-deposited structures. Empirically, we found that this value was 17.45 Å for the benchmarking experiments shown in Fig. 1c.

Benchmarking RhoDesign

For benchmarking, 3,435 structures and 276 structures from the PDB were used for training and testing, respectively. Structures with sequence similarity of >0.6 and structural similarity (TM score) of >0.2 were removed from the test set. Additionally, to ensure non-redundancy between the RhoFold-predicted training data and the test set, we used CD-HIT[45] with the lowest threshold (sequence similarity of 0.8) to remove duplicates. The maximum sequence similarity between the training and test sets was 0.514 (minimum 0.06, average 0.399), while the maximum TM score between the training and test sets was 0.172 (minimum 0.013, average 0.050). Thus, our benchmarking enabled us to evaluate RhoDesign's ability to generalize. As shown in Fig. 1c and Supplementary Table 1, when trained and tested on these sets RhoDesign exhibits a recovery rate of 52.9%, outperforming other models trained and tested on the same sets (whose recovery rates ranged from 23.7% to 26.1%).

In an orthogonal benchmark, we performed fivefold cross-validation after dividing the dataset on the basis of sequence and structural similarity. As shown in Supplementary Table 1, two sets of experiments were performed where we (1) employed PSI-CD-HIT[45] for sequence clustering with a sequence similarity threshold of 0.6, and (2) calculated the TM-score matrix using US-align[46] for structure similarity clustering with a structural similarity threshold of 0.5. Subsequently, the dataset of 3,711 instances was partitioned into five subsets, and fivefold cross-validation was performed with the training and test sets comprising 80% and 20% of the data, respectively. Notably, RhoDesign was the top-performing model on the basis of recovery rate and RMSD, with average sequence recovery rates of 61.7% and 63.5% and RMSDs of 12.704 and 13.013 Å under the sequence similarity and structure similarity conditions considered here, respectively. Intriguingly, although underperforming in terms of recovery rate, gRNAde modestly outperformed RhoDesign in terms of TM score under the structure similarity condition considered here.

Ablation studies for RhoDesign

In our first ablation study (Supplementary Table 1), we evaluated the performance of RhoDesign on the fixed-backbone sequence design tasks[47]—in which the sequence is unknown, the structure is known and the output is the designed sequence—described in the main text using the different indicated conditions. The data used and the train–test splits were identical to those described in 'Benchmarking RhoDesign', with the exception that we also investigated whether RhoFold-augmented data and secondary-structure information could enhance model performance. First, we compared RhoDesign under three conditions: using RhoFold-augmented data only ('predicted'), experimental data only ('PDB') or a combination of the two ('PDB + predicted'). Incorporating RhoFold-augmented data into the training process improved the recovery rate when compared with using experimental data only (Fig. 1d). The RhoDesign model trained with RhoFold-augmented data (predicted) achieved a recovery rate of 41.3%, while the model trained with experimental data (PDB) achieved a recovery rate of 30% (Fig. 1d). Additionally, combining RhoFold-augmented data and experimental data resulted in the best performance, in terms

https://doi.org/10.1038/s43588-024-00720-6

of recovery rate. The RhoDesign model trained with both PDB and predicted data achieved a recovery rate of 52.9% (Fig. 1d), indicating that augmentation effectively enhances performance. Finally, we compared RhoDesign models trained with only tertiary-structure information ('3D information only') and a combination of both tertiary-structure and secondary-structure information ('All components'). Here, we used the combination dataset comprising predicted and PDB data to train models with these two conditions. The results demonstrated that incorporating both secondary- and tertiary-structure information into the model yielded better performance in terms of all metrics considered. Specifically, the All components model achieved the highest recovery rate (52.9%) of all models considered, the highest TM score of 0.212 and the lowest perplexity score of 2.428.

The second ablation study (Supplementary Table 2) was performed similarly to the above, but with RhoDesign models trained on PDB data only, and with the model architecture modifications indicated.

RNA 3D-structure prediction

To expand the training data of RhoDesign, we utilized RhoFold[16] to predict 3D structures from the RNAcentral[45] database. Specifically, we initially predicted 3D structures for one million randomly selected RNAcentral sequences, focusing on single sequences (sequences not represented in higher-order structures such as RNA–protein complexes) for computational tractability. To enhance the reliability of the predictions, we filtered RhoFold-predicted structures on the basis of the pLDDT values (a measure of local confidence), choosing RNAcentral sequences with RhoFold-predicted pLDDT values greater than 0.6. This resulted in the dataset of 369,499 predicted structures used to train RhoDesign, as detailed above. RhoFold additionally produces predicted secondary structures as part of its initial output, and these secondary structures were used to augment training where indicated, as detailed above.

We further note that, in this work, we focus on using RhoFold-predicted structures with pLDDT > 60 to balance prediction confidence with the size of the resulting dataset for training RhoDesign. To study the effects of different pLDDT thresholds, we trained RhoDesign models using (1) all RhoFold-predicted structures with pLDDT > 80 and (2) all RhoFold-predicted structures with pLDDT > 40, two additional thresholds that subdivide the empirical distribution of all 1 million pLDDT values from RhoFold (Supplementary Fig. 1). We found that the performance of RhoDesign, as measured by recovery rate, TM score and RMSD, was worse in both cases (1) and (2) when trained models were applied to the same benchmarking task as in Fig. 1c (Supplementary Fig. 1). In case (1), the decreased performance might arise due to the smaller number of structures available (34,565 as opposed to 369,499 when pLDDT > 60). In case (2), the decreased performance might arise due to the lower quality of the structures hampering the training.

While RhoFold was inconclusive in predicting G-quadruplexes in these aptamers, the presence of G-quadruplexes in several of aptamers **1**–**4** may be consistent with their AlphaFold 3-predicted 3D structures as well as their RhoFold-predicted secondary structures (Extended Data Fig. 2). Indeed, the recent release of AlphaFold 3 details how AlphaFold 3 can be used to predict 3D RNA structures[19]. Benchmarks of AlphaFold 3 on CASP15 natural RNA targets suggest that it performs similarly well to RhoFold for these targets[19]. Furthermore, AlphaFold 3's prediction for the 3D structure of Mango-III is nearly identical to 6UP0, and it is possible that 6UP0 could be found in AlphaFold 3's training set (Extended Data Fig. 1). Nevertheless, using AlphaFold 3 to predict the 3D structures of aptamers **1**–**4** suggests the presence of G-quadruplexes in aptamers **1**, **2** and **4**, as visualized in Extended Data Fig. 2. In all cases, consistent with the ions found in 6UP0, we generated AlphaFold 3 models using specific ion constraints, which included one potassium ion (K^+) and three magnesium ions (Mg^{2+}).

Consistent with the notion that our approach captures structural information that is useful for Mango-like fluorescence activity, structural predictions across eight tertiary-structure prediction methods and four secondary-structure prediction methods—including methods not relying on deep learning—indicated that predicted structures were largely robust across different prediction methods, with aptamer **1** consistently predicted to have relatively high TM scores and low RMSD values with respect to Mango-III. These predictions also support the view that aptamer **1** is more structurally similar to Mango-III than are aptamers **2**–**4**, and that other aptamers designed by RhoDesign largely exhibit structural similarity to Mango-III, as measured by the TM score and RMSD (Supplementary Table 3). Here, to analyze the generality of structural predictions, 3D structures were predicted using SimRNA[48] (https://genesilico.pl/SimRNAweb), DRfold[49] (https://github.com/leeyang/DRfold), trRosettaRNA[17] (https://yanglab.qd.sdu.edu.cn/trRosettaRNA/), RNAComposer[50] (https://rnacomposer.cs.put.poznan.pl/) and RoseTTAFoldNA[51] (https://github.com/uw-ipd/RoseTTAFold2NA), in addition to RhoFold (https://github.com/ml4bio/RhoFold), AlphaFold 3 (https://alphafoldserver.com/) and AlphaFold 3 with the specific ion constraints noted above (Supplementary Table 3). SimRNA uses a physics-based model that simulates the RNA folding process, DRfold leverages end-to-end learning and deep geometrical potentials, trRosettaRNA combines deep learning with Rosetta energy minimization, RNAComposer is a knowledge-based method that assembles RNA tertiary structures from secondary-structure motifs and RoseTTAFoldNA is a nucleic-acid-specific adaptation of RoseTTAFold[15] that uses deep learning and homology modeling. These packages were installed from their respective repositories and run using default settings where applicable.

While the results shown in Supplementary Table 3 indicate the consistency of different sequence-to-structure prediction methods, we note that the quantitative values of structural similarity found in our benchmarks are limited by the performance of structural prediction methods, and our observation that optimal values of TM score and RMSD are substantially less than 1 and greater than 0 using RhoFold (Supplementary Table 1) suggest that improvements in the accuracy of structural predictions are needed. Importantly, RhoDesign can be used and developed independently of any 3D structural prediction method, and RhoDesign outperforms other methods in different benchmarks including the recovery rate—perhaps the most appropriate metric for the task of structure-guided design.

RNA secondary-structure prediction

As detailed in RNA 3D-structure prediction, secondary structures are predicted as part of RhoFold's output. Nevertheless, to optimize the comparison with the secondary structure of PDB 6UP0, we used RhoFold's implementation of Amber22 and AmberTools23[52] to perform additional default relaxation steps for the predicted 3D structures for each of aptamers **1**–**4**. The secondary structures shown in Extended Data Fig. 2 were then generated on the basis of the corresponding RhoFold-predicted relaxed 3D structures or PDB 6UP0 using the RNApdbee 2.0 webserver[53] (http://rnapdbee.cs.put.poznan.pl/). RNApdbee annotates secondary structures of knotted and unknotted RNAs on the basis of PDB files, utilizing a specialized algorithm to analyze 3D coordinates and derive structural annotations.

To analyze the generality of structural predictions, 2D structures were predicted using UFold[54] (https://github.com/uci-cbcl/UFold), RNAfold[55] (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi), MXfold2[56] (https://github.com/mxfold/mxfold2) and RNA-FM[57] (https://github.com/ml4bio/RNA-FM), in addition to RhoFold (Supplementary Table 3). UFold is a deep learning method trained directly on annotated data and base-pairing rules, RNAfold predicts minimum free-energy structures and base-pair probabilities, MXfold2 integrates a deep neural network with Turner's nearest-neighbor free-energy parameters and RNA-FM is a foundation model. These packages were installed from their respective repositories and run using default settings where applicable.

Generation of RNA sequences using RhoDesign
The 3D structure of PDB 6UP0 was provided as input to three different RhoDesign models, each generating 20 sequences: (1) a RhoDesign model trained on RNA structures excluding 6UP0 and those with a sequence similarity greater than 0.5, fixing the known 'GUACGAAGG' and 'GUAC' of the input sequence as a template motif, (2) a RhoDesign model trained on RNA structures including 6UP0 without the motif constraint and (3) a RhoDesign model trained on RNA structures excluding 6UP0 and those with a sequence similarity greater than 0.5 and without the motif constraint. A subset of resulting aptamers filtered to have predicted RMSD < 7.5 Å and sequence similarity < 0.65 with respect to 6UP0 was selected for further testing, resulting in the synthesis of five, seven and six aptamers, respectively, from each of the RhoDesign models (a total of 18 aptamers, as described in Supplementary Data 1).

For the spectral clustering based on predicted 3D structures shown in Fig. 2e, 4,000 sequences generated on the basis of the 3D structure of PDB 6UP0 were sampled from the above set of generated sequences. To account for variation in Mango aptamer structure, we similarly sampled 6,000 RhoDesign-generated sequences, which were generated using the RhoFold-predicted structures of Mango-I and Mango-III as inputs.

For the generation of additional aptamer candidates based on aptamer **1**, the full RhoDesign model was provided with the RhoFold-predicted structure of aptamer **1** as the backbone input. 5,000 sequences were generated using RhoDesign (Supplementary Data 1), and these sequences were downsampled for synthesis and evaluation as described in the main text.

Generation of RNA sequences using RaptGen
The sequences 'GUACGAAGGGACGGUGCGGAGAGGAGAGUAC' and 'GUACGAAGGAAGGUUUGGUAUGUGGUAUAUUCGUAC' (truncated variants of Mango-I and Mango-III A10U) were provided as training inputs. Random latent space coordinates were generated using the rand() function in MATLAB, and RaptGen[33] was used to produce RNA sequences with distributions similar to those of the training sequences. RMSDs with respect to PDB 6UP0 were predicted using the corresponding RhoFold-predicted structures for each generated sequence, and of 100 sequences generated all 22 unique sequences—which had predicted RMSD < 12 Å and sequence similarity between 0.57 and 0.87 with respect to 6UP0—were selected for further testing. It is important to note that RaptGen was originally developed to be targeted toward SELEX applications, and the small number of sequences used to train RaptGen in this study may deviate from its ideal use for SELEX (in which the training set may be large, comprising thousands of sequences or more).

Generation of RNA sequences using other structure-to-sequence models
Sequences were generated from each of the seven models indicated in the main text (LEARNA, Meta-LEARNA, RiboLogic, MCTS-RNA, gRNAde, RDesign and eM2dRNAs) using the provided model checkpoints in their respective repositories, that is, https://github.com/automl/learna for LEARNA and Meta-LEARNA, https://github.com/wuami/RiboLogic for RiboLogic, https://github.com/tsudalab/MCTS-RNA for MCTS-RNA, https://github.com/chaitjo/geometric-rna-design for gRNAde, https://github.com/A4Bio/RDesign for RDesign and https://github.com/iARN-unex/eM2dRNAs for eM2dRNAs. Default parameters for each model were used without any modifications, and no additional tuning or parameter adjustments were performed. One hundred sequences were sampled for each of gRNAde, RDesign, eM2dRNAs and MCTS-RNA; LEARNA, Meta-LEARNA and RiboLogic are based on reinforcement learning and/or converged after sampling about 10 sequences. Therefore, only 10 sequences were sampled from each of these methods. For each set of generated sequences, where applicable, sequences were downsampled to consider only those with the lowest RhoFold-predicted RMSD with respect to Mango-III in 6UP0, as well as those with at most 8 'C' repeats to improve synthesizability, resulting in a final set of 70 sequences (10 sequences corresponding to each model).

Sequences and sequence similarity
Sequence similarity was calculated using the pairwise2.align.globalxx in Biopython, as the number of matches in the highest-scoring alignment divided by the length of the aligned sequence. The reference sequences used for calculating sequence similarity were as follows: GCUACGAAGGAAGGAUUGGUAUGUGGUAUAUUCGUAGC for 6UP0 (a truncated variant of Mango-III A10U), GUACGAAGGGACGGUGCGGAGAGGAGAGUACGUGC for Mango-I, GUACGAAGGAGAGGAGAGGAAGAGGAGAGUACGUGC for Mango-II, GUACGAAGGAAGGAUUGGUAUGUGGUAUAUUCGUACGUGC for Mango-III, GUACGAAGGAAGGUUUGGUAUGUGGUAUAUUCGUACGUGC for Mango-III A10U and GUACCGAGGGAGUGGUGAGGAUGAGGCGAGUACGUGC for Mango-IV. We note that Mango-III A10U is referred to as 'Mango-III' in this work, and the sequence for Mango-III noted above (without the A10U mutation) was used only as a reference sequence for calculating sequence similarity. The sequences used for aptamers **1**–**4** were GUUACGGGGAAGGAGCUAAUGCUGUGUGCGUUCGUGGU, GCUACGGGGAGAGGACUAAUGCUGUAUGCGUUCGCGGC, GUUACGGGGAAGGAAUCAAUGCUGUUCGUGCUUACGGC and GUUAUGGGGAAGGAGUCGAUGCUGUGCGCGCUUGUGGU, respectively.

Structural clustering
RhoFold was used to predict RNA 3D structures. US-align[46] was used for structural alignment, producing a matrix of TM scores. Spectral clustering in scikit-learn was performed on the resulting matrix, and *t*-distributed stochastic neighbor embedding using default parameters was used to visualize the clustering.

Synthesis and experimental testing of generated RNAs
Generated RNAs were synthesized by Integrated DNA Technologies and resuspended in TE buffer (pH 8.0). In vitro reactions were performed on a black 384-well microplate with each RNA at a final concentration of 25 μM and TO1–biotin (TO1–3PEG–biotin, G955, Applied Biological Materials) at a final concentration of 10 μM in reaction buffer (TE buffer containing 40 mM HEPES, 100 mM KCl and 1 mM $MgCl_2$). Reactions were incubated at 37 °C for 30 min, and fluorescence was measured at Ex/Em = 475/500–550 using a SpectraMax M3 plate reader (Molecular Devices) or a Promega GloMax plate reader (Promega). A length-controlled random 36-nucleotide aptamer was used as a negative control. A list of all RNA sequences synthesized for testing and corresponding experimental data can be found in Supplementary Data 1.

Binding predictions
As discussed in the main text, in contrast to other generated sequences based on Mango-I and Mango-III, aptamer **1** clustered more similarly to Mango-III than did Mango-I, suggesting that it was sampling a structural space associated with high fluorescence and that 3D structural features conserved with Mango-III's may underlie its activity. Despite this suggestion, using sequence- and secondary-structure-based deep learning approaches to predict TO1–biotin binding activity as an indicator of fluorescence intensity did not result in a substantial correlation between the resulting scores and the empirically observed fluorescence intensities across the 110 synthesized and tested aptamers (Supplementary Fig. 2). It is possible that limitations in these methods, as opposed to substantive differences in 3D structure, underlie the lack of predictive ability, as has been suggested for molecular docking using AlphaFold-predicted protein structures[58]. Specifically, to leverage all available information, we compiled available SELEX data for Mango aptamers (Supplementary Dataset 1 of ref. 59). We used these data to train four distinct models: PrismNet, DLPRB_cnn, DLPRB_rnn and

MPBind. For PrismNet, DLPRB_cnn and DLPRB_rnn, we used default settings of model architecture and model parameters provided in their respective GitHub repositories (https://github.com/kuixu/PrismNet and https://github.com/ilanbb/dlprb). For MPBind, we used the default settings provided in the official user guide[34]. We found that PrismNet, DLPRB_cnn and DLPRB_rnn performed worse than MPBind using SELEX data when performance was measured using the Pearson correlation coefficient with respect to relative fluorescence intensity values normalized by that of Mango-I (Supplementary Data 1). We note that MPBind was originally developed to predict SELEX-derived binding aptamers: because MPBind is a meta-motif-based statistical framework, it may be particularly applicable and predictive when trained using SELEX data, which typically provide large amounts of binding information relevant to specific sequence motifs.

Aptamer optimization
5,000 RNA sequences were generated using the predicted structure of aptamer **1** as input with a fully trained RhoDesign model, as described in the main text. Sequences were further filtered for MPBind scores of >0.4 (approximately that of aptamer **1**) and for higher sequence similarity with respect to aptamer **1** than any Mango aptamer (Supplementary Data 1).

Pharmacologic inhibition of fluorescence activity
Three different G-quadruplex-selective small-molecule binders, TMPyP4, pyridostatin and carboxypyridostatin[37–39], were procured from Cayman Chemical Company and MedChemExpress and dissolved in dimethyl sulfoxide to generate working solutions. Compounds or vehicle (1% dimethyl sulfoxide) were then added at the indicated final concentrations to reactions containing 10 µM of RNA aptamer and 5 µM of TO1–biotin, under conditions similar to those described above. Reactions were incubated at 37 °C for 30 min, and fluorescence was measured at Ex/Em = 475/500–550 using a Promega GloMax plate reader.

QUMA-1 dye fluorescence
For direct detection of G-quadruplexes, BioTracker QUMA-1 RNA G-quadruplex live cell dye was obtained from MilliporeSigma (SCT056), dissolved in dimethyl sulfoxide, and added to a final concentration of 10 µM to reaction buffer containing 10 µM of RNA aptamer and no TO1–biotin. Reactions were incubated at 37 °C for 30 min, and fluorescence was measured at Ex/Em = 520/660–720 using a Promega GloMax plate reader.

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

Data availability
The numerical data supporting the findings of this paper are provided in the Source Data, and the sequences can be generated by running RhoDesign. Source Data for Figs. 1 and 2 are available. Sequences generated from RhoDesign and accompanying data are available as Supplementary Data 1. The training dataset and model checkpoints for RhoDesign are available from Zenodo[60]. The PDB structure for Mango-III (A10U), 6UP0, is available from the PDB[61].

Code availability
RhoDesign is available at https://github.com/ml4bio/RhoDesign and from Zenodo[60].

References

1. Cech, T. R., Zaug, A. J. & Grabowski, P. J. In vitro splicing of the ribosomal RNA precursor of *Tetrahymena*: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**, 487–496 (1981).
2. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857 (1983).
3. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
4. Dinger, M. E., Mercer, T. R. & Mattick, J. S. RNAs as extracellular signaling molecules. *J. Mol. Endocrinol.* **40**, 151–159 (2008).
5. Keefe, A. D., Pai, S. & Ellington, A. Aptamers as therapeutics. *Nat. Rev. Drug. Discov.* **9**, 537–550 (2010).
6. Tuerk, C., MacDougal, S. & Gold, L. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl. Acad. Sci. USA* **89**, 6988–6992 (1992).
7. Pardee, K. et al. Rapid, low-cost detection of Zika virus using programmable biomolecular components. *Cell* **165**, 1255–1266 (2016).
8. Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G. & Collins, J. J. A deep learning approach to programmable RNA switches. *Nat. Commun.* **11**, 5057 (2020).
9. Valeri, J. A. et al. Sequence-to-function deep learning frameworks for engineered riboregulators. *Nat. Commun.* **11**, 5058 (2020).
10. Takahashi, M. K. et al. A low-cost paper-based synthetic biology platform for analyzing gut microbiota and host biomarkers. *Nat. Commun.* **9**, 3347 (2018).
11. Green, A. A., Silver, P. A., Collins, J. J. & Yin, P. Toehold switches: de-novo-designed regulators of gene expression. *Cell* **159**, 925–939 (2014).
12. Paige, J. S., Wu, K. Y. & Jaffrey, S. R. RNA mimics of green fluorescent protein. *Science* **333**, 642–646 (2011).
13. Miao, Z. & Westhof, E. RNA structure: advances and assessment of 3D structure prediction. *Annu. Rev. Biophys.* **46**, 483–503 (2017).
14. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
15. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
16. 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).
17. Wang, W. et al. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nat. Commun.* **14**, 7266 (2023).
18. Pearce, R., Li, Y., Omenn, G. S. & Zhang, Y. Fast and accurate *ab initio* protein structure prediction using deep learning potentials. *PLoS Comput. Biol.* **18**, e1010539 (2022).
19. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
20. Das, R. et al. Assessment of three-dimensional RNA structure prediction in CASP15. *Proteins* **91**, 1747–1770 (2023).
21. Runge, F., Stoll, D., Falkner, S. & Hutter, F. Learning to design RNA. In *International Conference on Learning Representations 2019* https://openreview.net/pdf?id=ByfyHh05tQ (ICLR, 2019).
22. Wu, M. J., Andreasson, J. O. L., Kladwang, W., Greenleaf, W. & Das, R. Automated design of diverse stand-alone riboswitches. *ACS Synth. Biol.* **8**, 1838–1846 (2019).
23. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
24. Jing, B. et al. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations* https://openreview.net/pdf?id=1YLJDvSx6J4 (ICLR, 2021).
25. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* 5998–6008 (NIPS, 2017).
26. Hsu, C. et al. Learning inverse folding from millions of predicted structures. *Proc. Mach. Learn. Res.* **162**, 8946–8970 (2022).

27. Yang, X., Yoshizoe, K., Taneda, A. & Tsuda, K. RNA inverse folding using Monte Carlo tree search. *BMC Bioinform.* **18**, 468 (2017).
28. Joshi, C. K. & Liò, P. gRNAde: a geometric deep learning for 3D RNA inverse design. *Methods Mol. Biol.* **2847**, 121–135 (2025).
29. Tan, C. et al. RDesign: hierarchical data-efficient representation learning for tertiary structure-based RNA design. In *The Twelfth International Conference on Learning Representations* (ICLR, 2024).
30. Rubio-Largo, Á., Lozano-García, N., Granado-Criado, J. & Vega-Rodríguez, M. A. Solving the RNA inverse folding problem through target structure decomposition and multiobjective evolutionary computation. *Appl. Soft Comput.* **147**, 110779 (2023).
31. Autour, A. et al. Fluorogenic RNA Mango aptamers for imaging small non-coding RNAs in mammalian cells. *Nat. Commun.* **9**, 656 (2018).
32. Jeng, S. C. Y. et al. Fluorogenic aptamers resolve the flexibility of RNA junctions using orientation-dependent FRET. *RNA* **27**, 433–444 (2021).
33. Iwano, N. et al. Generative aptamer discovery using RaptGen. *Nat. Comput. Sci.* **2**, 378–386 (2022).
34. Jiang, P. et al. MPBind: a meta-motif-based statistical framework and pipeline to predict binding potential of SELEX-derived aptamers. *Bioinformatics* **30**, 2665–2667 (2014).
35. Jeng, S. C., Chan, H. H., Booy, E. P., McKenna, S. A. & Unrau, P. J. Fluorophore ligand binding and complex stabilization of the RNA Mango and RNA Spinach aptamers. *RNA* **22**, 1884–1892 (2016).
36. Trachman, R. J. III et al. Structural basis for high-affinity fluorophore binding and activation by RNA Mango. *Nat. Chem. Biol.* **13**, 807–813 (2017).
37. Liu, L. Y., Ma, T. Z., Zeng, Y. L., Liu, W. & Mao, Z. W. Structural basis of pyridostatin and its derivatives specifically binding to G-quadruplexes. *J. Am. Chem. Soc.* **144**, 11878–11887 (2022).
38. Han, F. X., Wheelhouse, R. T. & Hurley, L. H. Interactions of TMPyP4 and TMPyP2 with quadruplex DNA. Structural basis for the differential effects on telomerase inhibition. *J. Am. Chem. Soc.* **121**, 3561–3570 (1999).
39. Rocca, R. et al. Molecular recognition of a carboxy pyridostatin toward G-quadruplex structures: why does it prefer RNA? *Chem. Biol. Drug Des.* **90**, 919–925 (2017).
40. Chen, X. C. et al. Tracking the dynamic folding and unfolding of RNA G-quadruplexes in live cells. *Angew. Chem. Int. Ed. Engl.* **57**, 4702–4706 (2018).
41. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
42. Lu, X. J., Bussemaker, H. J. & Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**, e142 (2015).
43. The RNAcentral Consortium. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.* **47**, D221–D229 (2019).
44. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
45. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
46. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115 (2022).
47. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of *de novo* protein design. *Nature* **537**, 320–327 (2016).
48. Boniecki, M. J. et al. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **44**, e63 (2016).
49. Li, Y. et al. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nat. Commun.* **14**, 5745 (2023).
50. Biesiada, M. et al. Automated RNA 3D structure prediction with RNAComposer. *Methods Mol. Biol.* **1490**, 199–215 (2016).
51. Baek, M. et al. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **21**, 117–121 (2024).
52. Case, D. A. et al. AmberTools. *J. Chem. Inf. Model.* **63**, 6183–6191 (2023).
53. Zok, T. et al. RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res.* **46**, W30–W35 (2018).
54. Fu, L. et al. UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **50**, e14 (2022).
55. Lorenz, R. et al. ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
56. Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **12**, 941 (2021).
57. 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).
58. Wong, F. et al. Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Mol. Syst. Biol.* **18**, e11081 (2022).
59. Trachman, R. J. III et al. Structure and functional reselection of the Mango-III fluorogenic RNA aptamer. *Nat. Chem. Biol.* **15**, 472–479 (2019).
60. Wong, F. et al. Supporting code for: Deep generative design of RNA aptamers using structural predictions. *Zenodo* https://doi.org/10.5281/zenodo.13892413 (2024).
61. Trachman, R. J. & Ferre-D'Amare, A. R. Structure of the Mango-III fluorescent aptamer bound to YO3-biotin. *Protein Data Bank* https://doi.org/10.2210/pdb6UP0/pdb (2019).

Acknowledgements

This work was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award K25AI168451 (to F.W.), the Swiss National Science Foundation under grant number SNSF_ 203071 (to A.K.), the National Science Foundation Graduate Research Fellowship (to A.Z.W.), the Research Grants Council of the Hong Kong Special Administrative Region, China (projects CUHK 14222922 and RGC GRF 2151185 to I.K. and project CUHK 24204023 to Y.L.), a grant from the Innovation and Technology Commission of the Hong Kong Special Administrative Region, China (projects GHP/065/21SZ, IDBF24ENG06 and ITS/247/23FP to Y.L.), the National Key R&D Program of China (project 2022ZD0160101 to Y.L.) and the Broad Institute of MIT and Harvard (to J.J.C.). This work is part of the Antibiotics-AI Project, which is directed by J.J.C. and supported by the Audacious Project, Flu Lab, LLC, the Sea Grape Foundation, R. Zander and H. Wyss for the Wyss Foundation, and an anonymous donor. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

F.W. conceived research, performed or directed all experiments, wrote the paper and supervised research. D.H. and L.H. developed RhoDesign and performed computational analyses, with contributions from J.W., Z.H., Q.Y. and I.K. A.K. and A.Z.W. conceived research and performed experiments and analyses. S.O. and A.L. performed experiments. J.R., W.J., T.Z., K.I. and J.X.C. performed analyses. S.Z. conceived research and performed analyses. Y.L. conceived research, performed or directed all analyses and supervised research. J.J.C. conceived and supervised research. All authors assisted with manuscript editing.

Competing interests

J.J.C. is the founding scientific advisory board chair of Integrated Biosciences. F.W. is a co-founder of Integrated Biosciences. S.O. has an equity interest in Integrated Biosciences. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43588-024-00720-6.

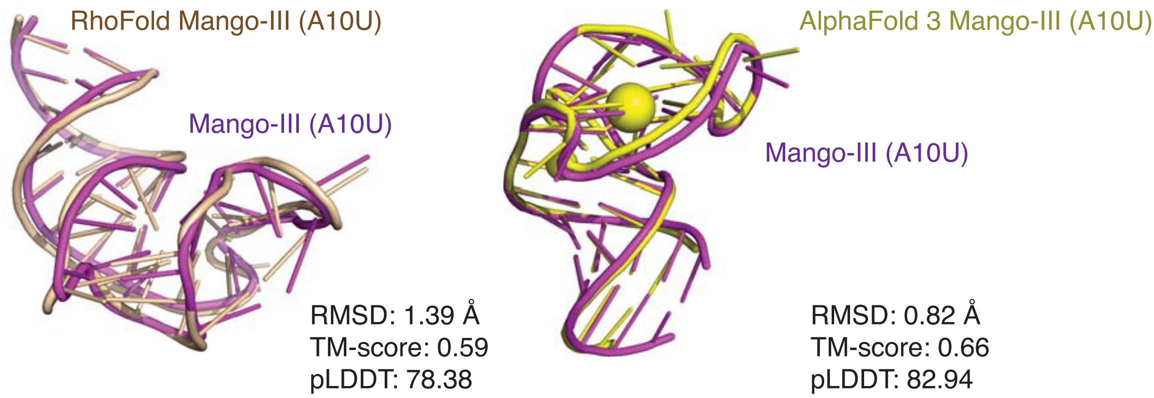
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-024-00720-6.

Correspondence and requests for materials should be addressed to Yu Li or James J. Collins.

Peer review information *Nature Computational Science* thanks Jianyi Yang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jie Pan, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

 https://doi.org/10.1038/s43588-024-00720-6

a

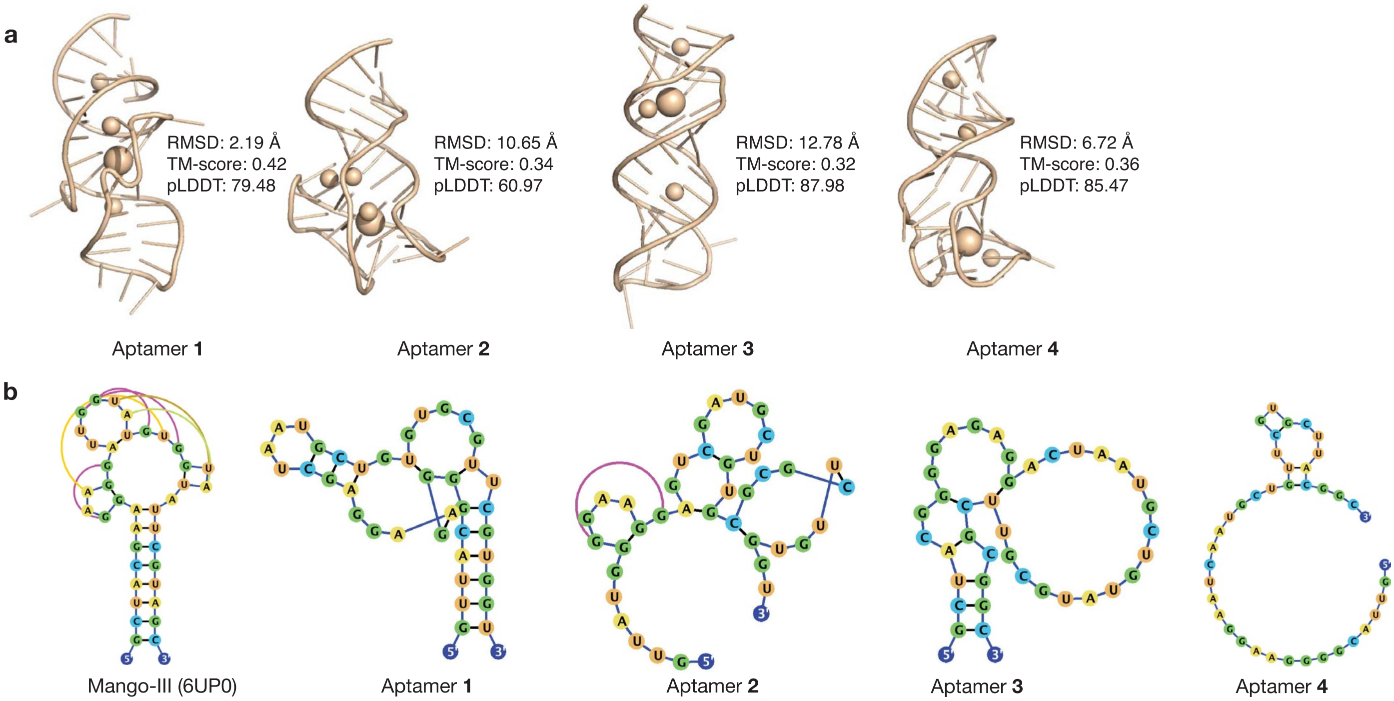


b

Aptamer	Sequence	Length
Mango-I	–**GUAC**–**GAAGG**GACGGUGCGG–AGAGGA––––GA**GUAC**GUGC–	35
Mango-II	–**GUAC**–**GAAGG**AGAGGAGAGGAAGAGGA––––GA**GUAC**GUGC–	36
Mango-III	–**GUAC**–**GAAGG**AAGGAUUGGUAUGUGGUAUAUUC**GUAC**GUGC–	40
Mango-III (A10U)	–**GUAC**–**GAAGG**AAGGUUUGGUAUGUGGUAUAUUC**GUAC**GUGC–	40
Mango-IV	–**GUAC**C**GAAGG**AGUGGUGAGGAUGAGGC––––GA**GUAC**GUGC–	37
Aptamer **1**	GUUAC–GGGGAAGGA––GCUAAUG––CUGUGUGCGUUCGUGGU	38
Aptamer **2**	GCUAC–GGGGAGAGG––ACUAAUG––CUGUAUGCGUUCGCGGC	38
Aptamer **3**	GUUAC–GGGGAAGGA––AUCAAUG––CUGUUCGUGCUUACGGC	38
Aptamer **4**	GUUAU–GGGGAAGGA––GUCGAUG––CUGUGCGCGCUUGUGGU	38

Extended Data Fig. 1 | Structural fidelity of RhoFold's Mango-III prediction, conserved sequence motifs in Mango aptamers, and comparison to aptamers 1-4. a, (Left) RhoFold-predicted 3D structure of Mango-III, aligned to the ground truth structure for Mango-III in PDB 6UP0. (Right) AlphaFold 3-predicted 3D structure of Mango-III, aligned to the ground truth structure for Mango-III in PDB 6UP0. **b**, Comparison of aptamers **1**-**4**'s sequences against Mango sequences. Here, conserved sequence motifs in Mango aptamers are indicated in red.

 https://doi.org/10.1038/s43588-024-00720-6



Extended Data Fig. 2 | AlphaFold 3-predicted 3D and RhoFold-predicted secondary structures. **a**, Predicted 3D structures for aptamers **1-4** generated using AlphaFold 3, as detailed in the *Methods–RNA 3D structure prediction*. RMSD, TM-score, and pLDDT values for each structure as compared to the ground truth structure for Mango-III in PDB 6UP0 are shown. **b**, Secondary structures for Mango-III and aptamers **1-4**, as generated based on the corresponding PDB structure (6UP0; Mango-III) or RhoFold predictions (aptamers **1-4**), as detailed in the *Methods–RNA secondary structure prediction*.

nature portfolio

Corresponding author(s): James J. Collins

Last updated by author(s): Oct 4, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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	RhoDesign is available at https://github.com/ml4bio/RhoDesign and from https://zenodo.org/records/13892413 (no version number). RhoFold is available at https://github.com/ml4bio/RhoFold (no version number). RaptGen is available at https://github.com/hmdlab/raptgen (no version number). BioPython (v.1.83) is available at https://biopython.org/wiki/Download. The RNApdbee 2.0 webserver is available at http://rnapdbee.cs.put.poznan.pl. DSSR is available at https://x3dna.org/ (no version number). US-align is available at https://zhanggroup.org/US-align/ (no version number). Scikit-learn (v.1.4) is available at https://scikit-learn.org/stable/. PrismNet is available at https://github.com/kuixu/PrismNet (no version number). DLPRB is available at https://github.com/ilanbb/dlprb (no version number). Additional models used for benchmarking can be found at the following: https://github.com/automl/learna for LEARNA and Meta-LEARNA (no version number), https://github.com/wuami/RiboLogic for Ribologic (no version number), https://github.com/tsudalab/MCTS-RNA for MCTS-RNA (no version number), https://github.com/chaitjo/geometric-rna-design for gRNAde (no version number), https://github.com/A4Bio/RDesign for RDesign (no version number), and https://github.com/iARN-unex/eM2dRNAs for eM2dRNAs (no version number). 3D structures were predicted using SimRNA (https://genesilico.pl/SimRNAweb; no version number), DRfold (https://github.com/leeyang/DRfold; no version number), trRosettaRNA (https://yanglab.qd.sdu.edu.cn/trRosettaRNA/; no version number), RNAComposer (https://rnacomposer.cs.put.poznan.pl/; no version number), and RoseTTAFoldNA (https://github.com/uw-ipd/RoseTTAFold2NA; no version number), in addition to RhoFold (https://github.com/ml4bio/RhoFold; no version number), AlphaFold 3 (https://alphafoldserver.com/), and AlphaFold 3 with the specific ion constraints. 2D structures were predicted using UFold (https://github.com/uci-cbcl/UFold; no version number), RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi; no version number), MXfold2 (https://github.com/mxfold/mxfold2; no version number), and RNA-FM (https://github.com/ml4bio/RNA-FM; no version number), in addition to RhoFold. Amber22 and AmberTools23 are included as part of RhoFold (no relevant version numbers to note). PrismNet, DLPRB_cnn, and DLPRB_rnn are from their respective GitHub repositories (https://github.com/kuixu/PrismNet and https://github.com/ilanbb/dlprb; no version numbers for all). MPBind (v2.1) is available from https://morgridge.org/research/regenerative-biology/software-resources/mpbind/. SoftMax® Pro 7.2 was used to collect data on the SpectraMax M3 plate reader. GloMax® Discover System Software v4.1 was used to collect data on the Promega GloMax plate reader.

Data analysis	MATLAB (R2023b) was used for data analysis and is available from Mathworks (Natick, MA, https://www.mathworks.com/products/matlab.html).

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The numerical data supporting the findings of this paper are provided in the Source Data, and the sequences can be generated by running RhoDesign. Source Data for Figures 1 and 2 are available. Sequences generated from RhoDesign and accompanying data are available as Supplementary Data 1. The training dataset and model checkpoints for RhoDesign are available from https://zenodo.org/records/13892413. The PDB structure for Mango-III (A10U), 6UP0, is available from https://www.rcsb.org/structure/6up0.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	Sample sizes were not predetermined and were chosen a priori of each experiment. All measurements performed were reported. Sample sizes were chosen to approximately reflect the minimal number of samples for which robust differences in fluorescence can be measured. These sample sizes were deemed to sufficient given the high reproducibility of the in vitro measurements when biological duplicate measurements were performed, as relevant for the experimental measurements shown in Fig. 2c,j (and provided in Supplementary Data 1).
Data exclusions	No data were excluded.
Replication	All data were representative of at least one biological replicate, and the numbers of replication of each experiment are indicated as relevant. All attempts at replication were successful.
Randomization	There were no preallocation considerations, and samples were allocated into experimental groups randomly.
Blinding	As we were not aware of any potential sources of bias in our experiments, we were not blinded to allocation during experiments and outcome assessment, and data collection and analysis were not performed blind to the conditions of each experiment.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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