

A Large-Scale Foundation Model for RNA Enables Diverse Function and Structure Prediction

Eric Xing

`eric.xing@gmail.com`

GenBio AI; Mohamed bin Zayed University of Artificial Intelligence; Carnegie Mellon University

Shuxian Zou

GenBio AI; Mohamed bin Zayed University of Artificial Intelligence

Tianhua Tao

GenBio AI; University of Washington

Sazan Mahbub

GenBio AI; Carnegie Mellon University

Caleb Ellington

GenBio AI

Robin Algayres

GenBio AI

Dian Li

GenBio AI

Yonghao Zhuang

GenBio AI; Carnegie Mellon University

Hongyi Wang

GenBio AI

Le Song

GenBio AI; Mohamed bin Zayed University of Artificial Intelligence

Article

Keywords:

Posted Date: May 7th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6445344/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

A Large-Scale Foundation Model for RNA Enables Diverse Function and Structure Prediction

Shuxian Zou^{1,2*}, Tianhua Tao^{1,4*}, Sazan Mahbub^{1,3*}, Caleb N. Ellington¹
Robin Algayres[†], Dian Li¹, Yonghao Zhuang^{1,3*}, Hongyi Wang¹
Le Song^{1,2‡}, Eric P. Xing^{1,2,3‡}

¹GenBio AI

²Mohamed bin Zayed University of Artificial Intelligence

³Carnegie Mellon University

⁴University of Washington

April 14, 2025

Accurately predicting RNA structures and functions from nucleotide sequences, or conversely, designing sequences to meet structural and functional requirements, remains a fundamental challenge in RNA biology, largely due to limited annotated data and the poor efficiency of *ab initio* modeling approaches. Here, we introduce AIDO.RNA, a large-scale RNA foundation model that leverages self-supervised pre-training to learn general and effective RNA representations, which can be transferred to tackle a wide range of RNA prediction and design tasks. AIDO.RNA is a 1.6-billion-parameter transformer-based language model, pre-trained on 42 million non-coding RNA (ncRNA) sequences at single-nucleotide resolution. It can be adapted to achieve state-of-the-art performance on 26 out of 28 diverse tasks, including RNA structure and function

*Work done during internship at GenBio AI.

†Work done while at GenBio AI.

‡Corresponding authors: le.song@genbio.ai, eric.xing@genbio.ai.

prediction, mRNA expression modeling, multi-modal RNA isoform expression prediction, and RNA inverse folding, demonstrating its effectiveness and versatility across the board. We find that beyond excelling in ncRNA-related tasks that directly reside in the pre-training data space, AIDO.RNA can be efficiently adapted to new domains with continued domain-specific pre-training to generalize toward untranslated regions and coding regions of mRNA, suggesting a promising pathway to continue to level up biological foundation models in general. We make AIDO.RNA open source and release the utility of the model in AIDO.ModelGenerator, a Python package enabling easy reproduction, application, and extension of our results.

Introduction

Foundation models (FMs) are large-scale deep learning models trained on vast datasets, learning generalizable representations that can be adapted to a wide range of downstream tasks with minimal task-specific modifications. These models have revolutionized natural language processing (NLP) (1, 2) and are now being increasingly applied to biological sequence modeling. In protein science, large-scale pre-trained protein language models (LMs) such as ESM2 (3) and xTrimoPGLM (4) have significantly advanced the prediction of protein structure and function. Similarly, in genomics, large-scale DNA FMs like Nucleotide Transformer (5) and the Evo series (6, 7) have shown impressive performance in genome interpretation and sequence design. With model parameters scaling to 100 billion for proteins (4) and 40 billion for DNA (7), FMs capture fine-grained underlying patterns and possibly causal elements highly related to structure and function, exhibiting strong performance across biological sequence tasks. However, despite RNA’s prominent role in frontier biomedical research (8, 9), FMs for RNA modeling are still orders of magnitude smaller than those for DNA and proteins (10–21).

Understanding sequence-structure/function relationships in RNA is key to both fundamental biology and therapeutic applications. While existing RNA FMs show promise in RNA structure and function prediction, they face three key limitations: 1) **Scope**: Many models are trained on specific RNA subsets (e.g., coding sequences, untranslated regions), restricting their generalizability. 2) **Scale**: RNA models remain orders of magnitude smaller than protein and DNA models, typically containing only millions of parameters, limiting their application and performance potential (**Supplementary Table 1**). 3) **Evaluation**: Most have been tested on a few specific downstream tasks, leaving their broader applicability uncertain.

Here we present AIDO.RNA, the first billion-parameter-scale RNA foundation model designed to capture comprehensive RNA sequence patterns and support diverse RNA prediction tasks. It is a transformer-based LM pre-trained on a carefully curated dataset at nucleotide resolution (**Fig. 1a** and Methods). Emphasizing sequence understanding abilities like ESM2, we adopted a masked language modeling (MLM) objective for pre-training, which aims to reconstruct the randomly masked tokens in the input sequence (1, 22). Special focus was given to the pre-training data, which remains underexplored and lacks consensus in the RNA domain. We explored two RNA sequence databases for pre-training: 1) RNACentral: A curated high-quality dataset containing 42 million ncRNA sequences (23). 2) MARS: A large but noisy dataset with 1.7 billion nucleotide sequences (24). Through extensive experiments, we found that the small but high-quality ncRNA dataset outperforms the large but noisy dataset, leading to improved performance across representative downstream tasks (**Extended Data Table 1** and Methods). Therefore, we pre-trained AIDO.RNA on RNACentral and scaled it to 1.6 billion parameters. To promote wider sharing of utility by the research community, we release, to our knowledge, the largest-scale FM for RNA.

AIDO.RNA achieves strong performance in a wide range of downstream tasks after task-specific fine-tuning (**Fig. 1b**). We conducted comprehensive benchmarking across 28 tasks,

including 26 RNA sequence understanding tasks and two more complex tasks – multi-modal transcript isoform expression prediction and RNA inverse folding, both requiring co-inference of AIDO.RNA with other modality-specific models. As shown in **Fig. 1c**, AIDO.RNA surpasses previous SOTA results on 24 out of 26 sequence understanding tasks, including RNA secondary structure prediction and mRNA translation efficiency prediction, demonstrating its ability to comprehend both ncRNA and mRNA sequences. Additionally, we show that our model can work with DNA and protein FMs to achieve better results for the multi-modal transcript isoform expression prediction task. Finally, in RNA inverse folding, which involves generating sequences from 3D RNA backbone structures, AIDO.RNA outperforms the previous SOTA model gRNAd (25). The strong performance of AIDO.RNA across the board demonstrates it is a powerful and versatile FM for RNA. By establishing a comprehensive benchmark across diverse RNA tasks, AIDO.RNA sets a new standard for RNA FMs, bridging a critical gap in biological sequence modeling.

Results

AIDO.RNA captures ncRNA structural and functional information

RNA secondary structure, formed by reverse complementary base pairing, is more stable and accessible than its tertiary form within cells. Accurate prediction of RNA secondary structure is essential for tasks such as higher-order structure prediction and function prediction (26). To evaluate the efficacy of AIDO.RNA, we first tested it on the RNA secondary structure prediction task. We utilized an RNA secondary structure dataset derived from bpRNA-1m (27), which primarily consists of ncRNAs (Methods). We compared AIDO.RNA with SPOT-RNA (28), UFold (29) and several pre-trained RNA LMs, including RNA-FM (11), RNAErnie (21) and RiNALMo (19). Our model achieved SOTA results on the test set bpRNA-TS0 (**Fig. 2a**), attaining an F1 score of 0.783 and outperforming existing models by large margins. This result

demonstrates that our model effectively captures the intrinsic mapping between ncRNA primary sequence and its secondary structure. To further validate the model’s predictions, we visualized three examples from the test set and observed that the predicted secondary structures closely resembled their corresponding ground truth structures (**Fig. 2c**).

We further investigated how well AIDO.RNA generalizes to predict secondary structures from unseen RNA families. We leveraged the dataset from Szikszai et al. (30), which contains secondary structures from nine distinct RNA families (**Supplementary Table 3**). We split the data into nine subsets based on their families, each time leaving out one family for evaluation while using the remaining families for training and validation. We fine-tuned AIDO.RNA on each subset and evaluated it on the held-out family (Methods). Our model achieved SOTA performance on six out of nine RNA families in terms of F1 score. On average across all nine families, it reached an F1 score of 0.733, surpassing all baseline models (**Fig. 2b**). These results highlight our model’s strong generalization capability to unseen RNA families.

To better understand in which stage AIDO.RNA captures the structural information, we conducted an unsupervised analysis on the pre-trained AIDO.RNA model. We performed a case study on RNaseP:RNA component of the eukaryotic ribosome using nucleotide dependency analysis proposed by Silva et al. (31), which can quantify how nucleotide substitutions at one position affect the probabilities of nucleotides at other positions (Methods). We generated the nucleotide dependency map using AIDO.RNA and found it captured long-distance interactions in both secondary and tertiary structures (**Extended Data Fig. 1**). This result suggests that our model might have captured structural information during the pre-training stage.

Beyond structure prediction, we further explored AIDO.RNA’s capability in predicting ncRNA functional families, which is essential for understanding the regulatory roles of RNA molecules in biological processes. Our model achieved an accuracy of 0.993 in classifying 88 functional families for short ncRNA sequences (Methods), reducing the error rate by half

compared to the previous SOTA model, Uni-RNA (13). When 200% boundary noise was added to the input sequences, our model maintained high performance with an accuracy of 0.994, demonstrating strong robustness to noisy sequence boundaries (Fig. 2d). Additionally, we evaluated AIDO.RNA’s ability to predict splice sites, a crucial aspect of gene structure elucidation and alternative splicing mechanisms. RNA splicing is a fundamental step in eukaryotic gene expression, influencing transcript diversity and regulation. Using the donor and acceptor splice site datasets from Scalzitti et al. (32), we assessed our model on four unseen species: Zebrafish, Fly, Worm, and Plant (Methods). On average across these species, AIDO.RNA achieved an F1 score of 0.960 on the donor dataset and 0.937 on the acceptor dataset (Fig. 1c). Averaged across both datasets, our model attained SOTA F1 scores on 3 out of 4 species, suggesting its generalization ability to unseen species in splice site prediction (Fig. 2e). Furthermore, we investigated AIDO.RNA’s performance in predicting post-transcriptional RNA modifications, which are chemical modifications occurring on RNA molecules after transcription, playing vital roles in stability, translation, and function. AIDO.RNA accurately predicted 12 types of RNA modifications, achieving an average AUROC score of 0.971, outperforming previous models such as MultiRM (33) and Uni-RNA (13) (Fig. 2f and Methods). These results suggest that AIDO.RNA effectively captures underlying relationships between ncRNA sequences and their functional roles.

AIDO.RNA generalizes to mRNA engineering tasks

mRNA medicine is an emerging and rapidly evolving field that utilizes mRNA-based delivery to instruct human cells to produce therapeutic proteins, enabling targeted treatments for various diseases. The central design consideration for an mRNA medicine is its nucleotide sequence, generally with the goal to maximize the production of therapeutic proteins in the cell (34). Protein expression is highly dependent on the relative abundance of the mRNA transcript and the

rate at which mRNA molecules are translated into proteins. To evaluate AIDO.RNA’s potential in facilitating therapeutic mRNA design, we collected a set of tasks from literature, covering mRNA expression level, mRNA translation efficiency, and protein abundance prediction.

We first applied AIDO.RNA to predict mRNA expression level, the relative abundance of the mRNA transcript in the cell. We utilized the human Muscle, PC3, and HEK datasets from Chu et al. (18), which contain mRNA expression measurements associated with 5’ untranslated regions (UTRs) (Methods). AIDO.RNA achieved SOTA performance across all three datasets, with Spearman correlation coefficients (CC) of 0.73, 0.69, and 0.72 on Muscle, PC3, and HEK, respectively. Compared to UTR-LM, a domain-specific model pre-trained directly on 5’ UTR sequences, AIDO.RNA outperformed it by +10.47% on Muscle, +10.00% on PC3, and +10.35% on HEK (**Fig. 3a**). Notably, although AIDO.RNA was pre-trained on ncRNA sequences, it demonstrated a strong understanding of 5’ UTRs, a region within mRNA, highlighting the model’s ability to generalize beyond its original pre-training domain through task-specific fine-tuning.

We further tested AIDO.RNA’s capability in predicting mRNA expression levels using coding sequences (CDSs) as input. We leveraged the datasets preprocessed by Outeiral and Deane (20), which contain transcript abundance measurements for CDSs from seven organisms (Methods). Despite being pre-trained on ncRNA sequences, AIDO.RNA achieved a similar Pearson CC on average across seven organisms as CaLM (0.559 v.s. 0.556), a codon LM that was pre-trained directly on CDSs (**Fig. 3b**). Interestingly, compared with the protein LM ESM2-650M, which achieved an average Pearson CC of 0.482, AIDO.RNA obtained significantly higher performance (one-tailed paired T-test, p-value = 0.01). Given ESM2-650M is a strong protein LM pre-trained on over 60 million protein sequences, the fact that it underperformed AIDO.RNA which was pre-trained on 42 million ncRNA sequences, indicates that sequence modeling at the nucleotide resolution offers richer signal than at the amino acid resolution.

To further explore AIDO.RNA’s capability in CDS-based tasks, we utilized CaLM’s pre-training dataset and conducted a second phase of pre-training to adapt AIDO.RNA to the CDS domain. In NLP, continued pre-training in domains has been proven to be an effective strategy for improving performance under both high- and low-resource settings (35). We initialized the model with pre-trained AIDO.RNA model weights, and trained it on 9 million CDS sequences (Methods). The resulting model, AIDO.RNA-CDS, was evaluated in the same manner as AIDO.RNA on CDS-based mRNA expression level prediction tasks. Across all seven organisms, AIDO.RNA-CDS consistently outperformed AIDO.RNA, demonstrating improved understanding of CDS sequences. On average, AIDO.RNA-CDS achieved a Pearson CC of 0.606, with a relative improvement of +8.50% over AIDO.RNA (Fig. 3b). This significant performance gain (one-tailed paired T-test, p-value < 0.001) underscores the effectiveness of continued pre-training in adapting RNA FMs to new domains, demonstrating their extensibility.

Furthermore, we applied AIDO.RNA to predict mRNA translation efficiency, the rate at which mRNA molecules are translated into proteins, a property of critical importance in mRNA vaccine development. We benchmarked our model on the human Muscle, PC3 and HEK datasets from Chu et al. (18) (Methods). AIDO.RNA achieved SOTA Spearman CC of 0.73, 0.72 and 0.68 on Muscle, PC3 and HEK, respectively. Notably, it outperformed UTR-LM by +9.00%, +11.23% and +13.05% on Muscle, PC3 and HEK, correspondingly (Fig. 3c). The consistently better performance of AIDO.RNA on human 5’ UTR based tasks over baseline models showcases its capability in understanding human 5’ UTR sequences. Additionally, we utilized AIDO.RNA to model mean ribosome loading (MRL), a key indicator of translation efficiency. We used the MRL dataset from Sample et al. (36), where the average number of ribosomes associated to a given mRNA sequence was measured for each 5’ UTR sequence (Method). AIDO.RNA attained a 0.86 R^2 score, reaching the SOTA performance as RiNALMo (Fig. 3d).

Lastly, we applied AIDO.RNA for protein abundance prediction. We utilized datasets from

Outeiral and Deane (20), which encompass five organisms, with each sample containing a CDS and the corresponding protein abundance measurement (Methods). AIDO.RNA-CDS achieved SOTA performance on four out of five organisms, with an average Pearson CC of 0.735 across all datasets, outperforming CaLM by +12.62% and AIDO.RNA by +10.19% (Fig. 3e). When comparing the performance of ESM2-650M and AIDO.RNA-CDS on CDS-based tasks, our model consistently outperforms ESM2-650M, indicating that our model is competent in protein-level tasks, opening new avenues for studying proteins.

AIDO.RNA facilitates multi-modal RNA isoform expression prediction

Not only can AIDO.RNA be used as a standalone FM to facilitate RNA-only sequence understanding tasks, it can also be used together with FMs of other modalities to tackle more complex multi-modal tasks. We explored the interplay between DNA, RNA, and protein FMs, enabling the joint modeling of DNA, RNA, and protein sequences for gene expression prediction. In eukaryotes, a single gene in the DNA sequence can be transcribed into multiple RNA isoforms, each with distinct exon combinations, due to the alternative splicing mechanism in gene expression. This process allows for increased protein diversity, enabling different functional proteins to be produced from the same gene, depending on cell type, developmental stage, or environmental conditions. The RNA isoform expression level in the cell is not only affected by its RNA sequence, but also regulated by the regulatory elements like enhancers and promoters in the DNA sequence.

To understand the impact of different modalities in predicting isoform expression, we investigated RNA isoform expression prediction in human tissues. Following the data preprocessing pipeline proposed by Gurau-Luis et al. (37), we curated a population-level tissue-specific RNA isoform expression dataset from the Genotype-Tissue Expression (GTEx) Portal (Methods). We systematically investigated single-modal, bi-modal, and tri-modal modeling approaches for RNA

isoform expression prediction to gain insights into how different modalities impact performance. We trained a series of models on the dataset and compared their performance on the hold-out chromosomes. We found that the DNA sequence is the most informative modality among the three, possibly due to the integration of regulatory elements for the transcript since we are using a DNA sequence of 197k bases. Interestingly, given the long DNA sequences, further integrating RNA sequences leads to a large performance gain. Besides, when jointly modeling DNA and RNA sequences, protein information seems to be redundant for predicting the RNA isoform expression. This result means that jointly modeling DNA and RNA sequences is essential. When comparing to IsoFormer, a multi-modal model that utilizes DNA, RNA, and protein sequences as input, our fusion models (DNA+RNA and DNA+RNA+Protein) achieved a SOTA Spearman CC of 0.734 on this task, while using 60% fewer trainable parameters (Fig. 4b).

We further evaluated the impact of RNA FMs on RNA isoform expression prediction by comparing AIDO.RNA-CDS, RNA-FM, and RiNALMo under two experimental settings: 1)RNA-only: We trained models using only RNA sequences as input; 2) RNA+DNA: We fixed Enformer as the DNA sequence encoder while varying the RNA sequence encoders among different RNA FMs, using both DNA and RNA sequences as input (Methods). In both settings, AIDO.RNA-CDS consistently outperformed RNA-FM and RiNALMo, regardless of whether Enformer was included in the model. Interestingly, when jointly modeled with Enformer, different RNA FMs yielded quite similar performances and all of them outperformed IsoFormer (Fig. 4c).

Lastly, we explored how our models perform across different human tissues, including blood, muscle, brain, etc. Across all 30 tissues, our fusion model (DNA+RNA+Protein) consistently outperformed IsoFormer (Fig. 4d). We found that its performance is highly correlated with Enformer’s performance, with a Spearman CC of 0.98. For AIDO.RNA-CDS, however, we observed a weak correlation with our fusion model (0.25 Spearman CC). These results indicate that DNA sequence / Enformer is the major contributor to the performance of our fusion model.

Notably, RNA isoform expression prediction marks our initial step toward multi-modal learning aimed at capturing the intricate complexity of biological systems – a key insight underlying the AI-driven Digital Organism (AIDO) vision (38). We demonstrate that this approach outperforms single-modality modeling, highlighting the advantages of integrating diverse biological signals. To support and encourage broader multi-modal learning across FMs within the community, we implemented our code in a model- and task-agnostic manner, enabling flexible application to any combination of FMs across a wide variety of downstream tasks.

AIDO.RNA benefits RNA sequence design

Beyond RNA sequence understanding, we further explored how AIDO.RNA can facilitate RNA sequence design, which is crucial to therapeutic innovation (39), synthetic biology (40, 41), and fundamental molecular biology research (42). The goal is to generate RNA sequences with specific properties (25, 43). To achieve this, we extended AIDO.RNA with a discrete diffusion modeling framework, enabling generative RNA sequence design. Our approach combines AIDO.RNA with gRNAd (25) and probabilistic diffusion (44, 45), iteratively refining sequences through masking and predicting optimal nucleotide compositions (Methods). This general framework supports both unconditional and conditional sequence design.

Specifically, we conducted experiments on RNA inverse folding, a task aiming to generate RNA sequences that fold into the given 3D structure (25, 46). We used the dataset from Das et al. (47), which consists of 4,025 training, 100 validation, and 98 test samples. We evaluated our model in two settings: (1) adaptation with conditional diffusion, where AIDO.RNA is fine-tuned for the inverse folding task (Fig. 5a); and (2) zero-shot generation, where AIDO.RNA remains frozen. Sequence recovery rate (SRR) is adopted as the evaluation metric, which is defined as the average percentage of ground truth nucleotides correctly recovered in the sampled sequences. Under the zero-shot setting, AIDO.RNA achieved 53.16% SRR, surpassing the original gRNAd,

indicating that our pre-trained model can benefit RNA inverse folding. When further fine-tuned with diffusion, AIDO.RNA achieved a higher SRR of 54.41% (**Fig. 5b**), outperforming gRNAd by a large margin. A detailed analysis of different RNA sequence length ranges reveals that AIDO.RNA, when augmented with gRNAd, enhances the SRR of inverse folding across most length ranges (**Fig. 5c**). We visualized the structures of sequences generated by our diffusion method for an example RNA (PDB ID: 3B58) in **Fig. 5d** and observed that they exhibit similar structural characteristics. These results showcase that AIDO.RNA can benefit RNA sequence design.

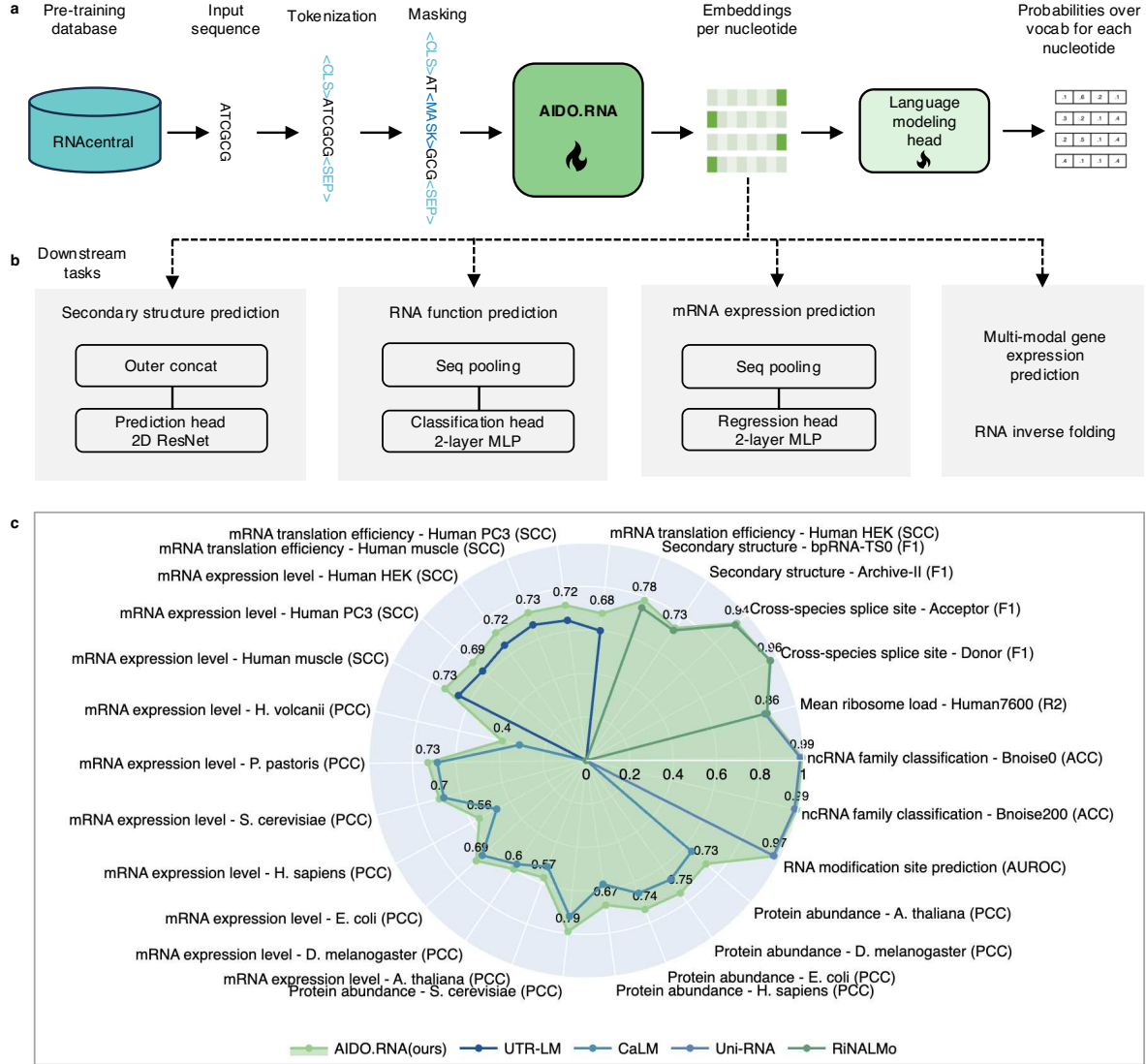


Fig. 1 | Overview of AIDO.RNA. **a**, AIDO.RNA is a transformer-based language model pre-trained on 42M ncRNA sequences from RNAcentral database in single nucleotide resolution using masked language modeling objective. We scale the model size to 1.6B parameters, making it the largest-scale general-purpose RNA FM to date. **b**, After pre-training, AIDO.RNA can be easily applied to a wide range of RNA related tasks by adding task-specific prediction heads and fine-tuning on task-specific datasets. **c**, AIDO.RNA achieves SOTA results on 24 out of 26 RNA sequence understanding tasks, including RNA secondary structure prediction, RNA function prediction and mRNA-related tasks, compared to two domain expert models UTR-LM (18) and CaLM (20) and two general RNA FMs Uni-RNA (13) and RiNALMo (19). PCC: Pearson CC, SCC: Spearman CC.

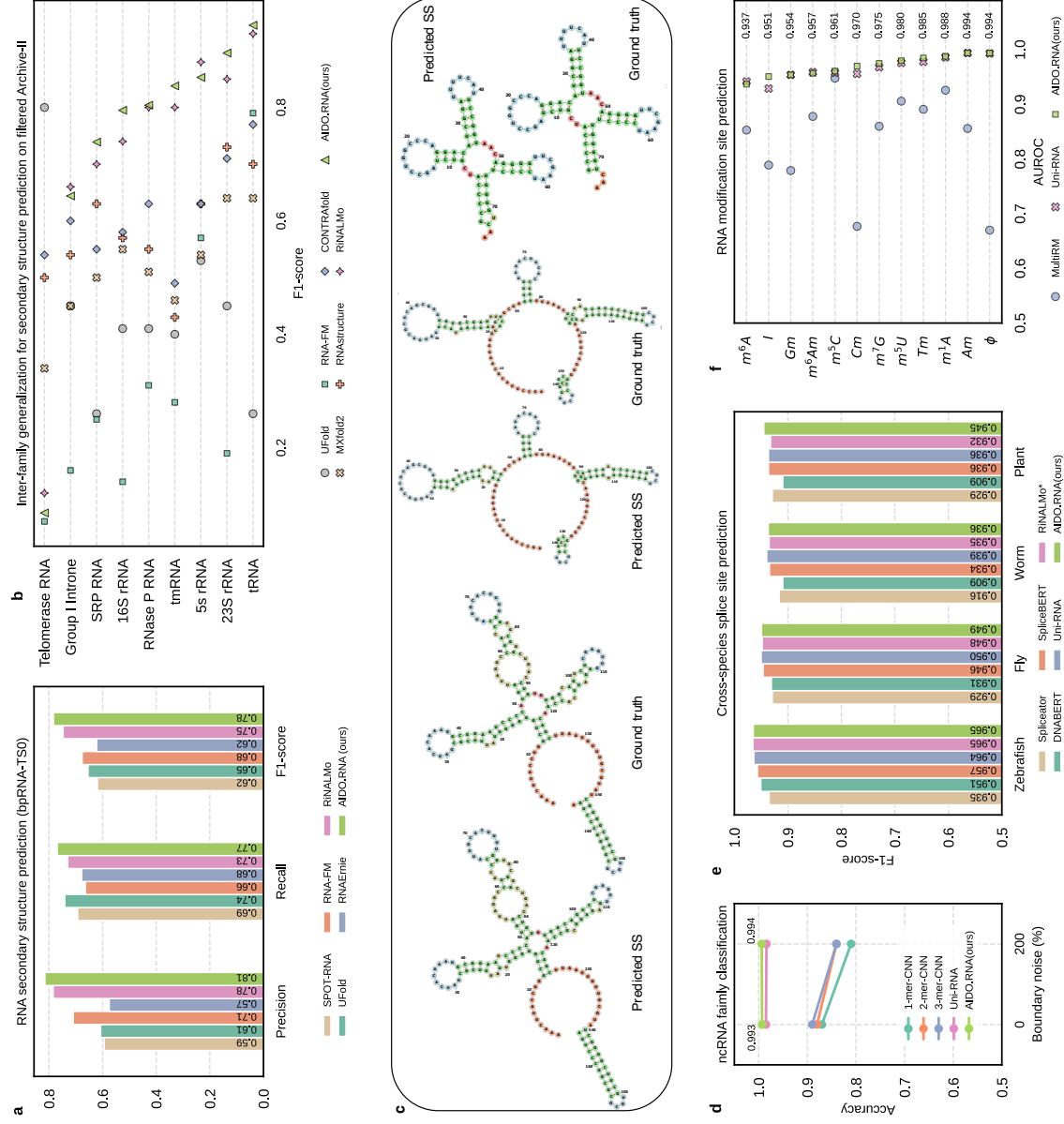


Fig. 2 | AIDO.RNA captures ncRNA structural and functional information. **a**, AIDO.RNA improves upon the SOTA performance for RNA secondary structure prediction on the widely used bpRNA dataset, achieving higher scores in all metrics. **b**, AIDO.RNA shows superior inter-family generalization capability for secondary structure prediction in six out of nine RNA families. **c**, Visualization of secondary structures predicted by AIDO.RNA for three RNA molecules in the bpRNA-TS0 test set. **d**, AIDO.RNA achieves SOTA performance in predicting functional families of short ncRNA sequences and is robust to high sequence boundary noise. **e**, AIDO.RNA generalizes well to predict splice sites of unseen species. Reported are the average F1 scores across the donor and acceptor datasets. Results of RiNALMo are reproduced by us by using their codebase and hyperparameters. **f**, AIDO.RNA excels in predicting 12 types of RNA modification sites.

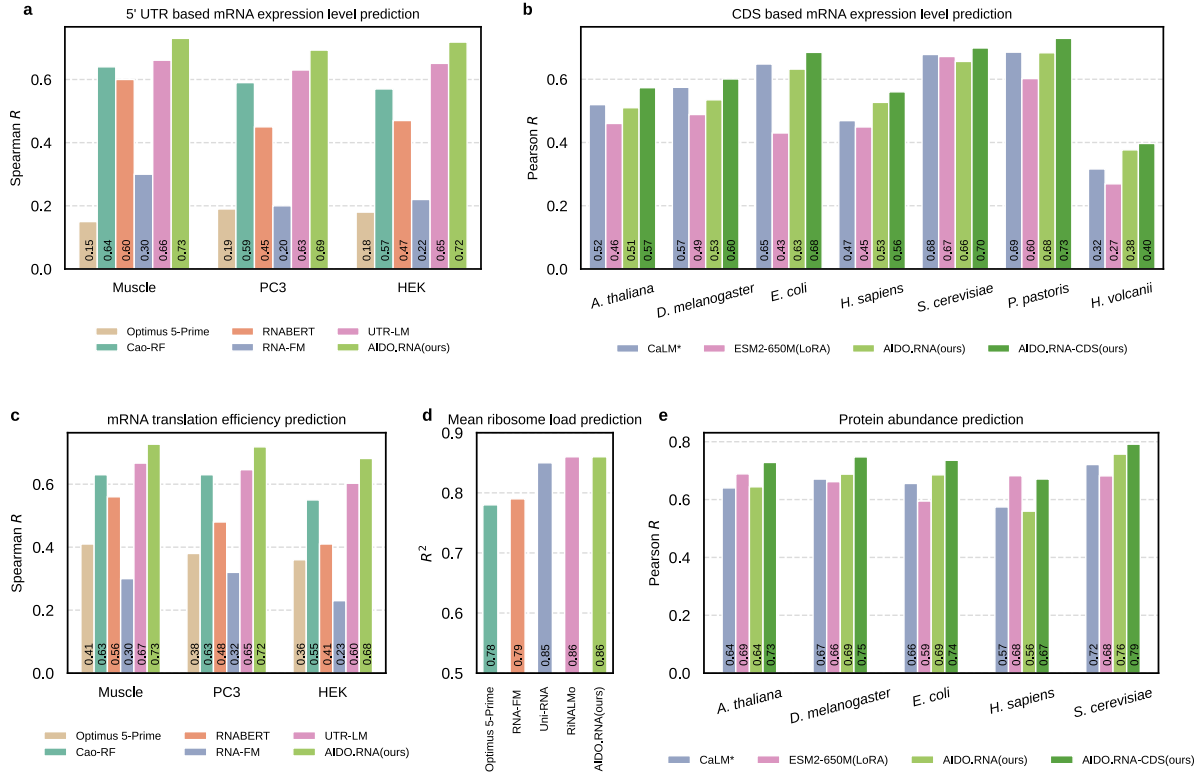


Fig. 3 | AIDO.RNA generalizes to mRNA engineering tasks. **a**, AIDO.RNA outperforms the domain expert model UTR-LM in 5' UTR based mRNA expression level prediction on human tissue and cell lines. **b**, AIDO.RNA-CDS achieves SOTA results in CDS-based mRNA expression level prediction on seven organisms. **c**, AIDO.RNA excels in 5' UTR based mRNA translation efficiency prediction on human tissue and cell lines. **d**, AIDO.RNA achieves the best performance on mean ribosome load prediction. **e**, AIDO.RNA-CDS outperforms CaLM and ESM2 in protein abundance prediction on four out of five organisms.

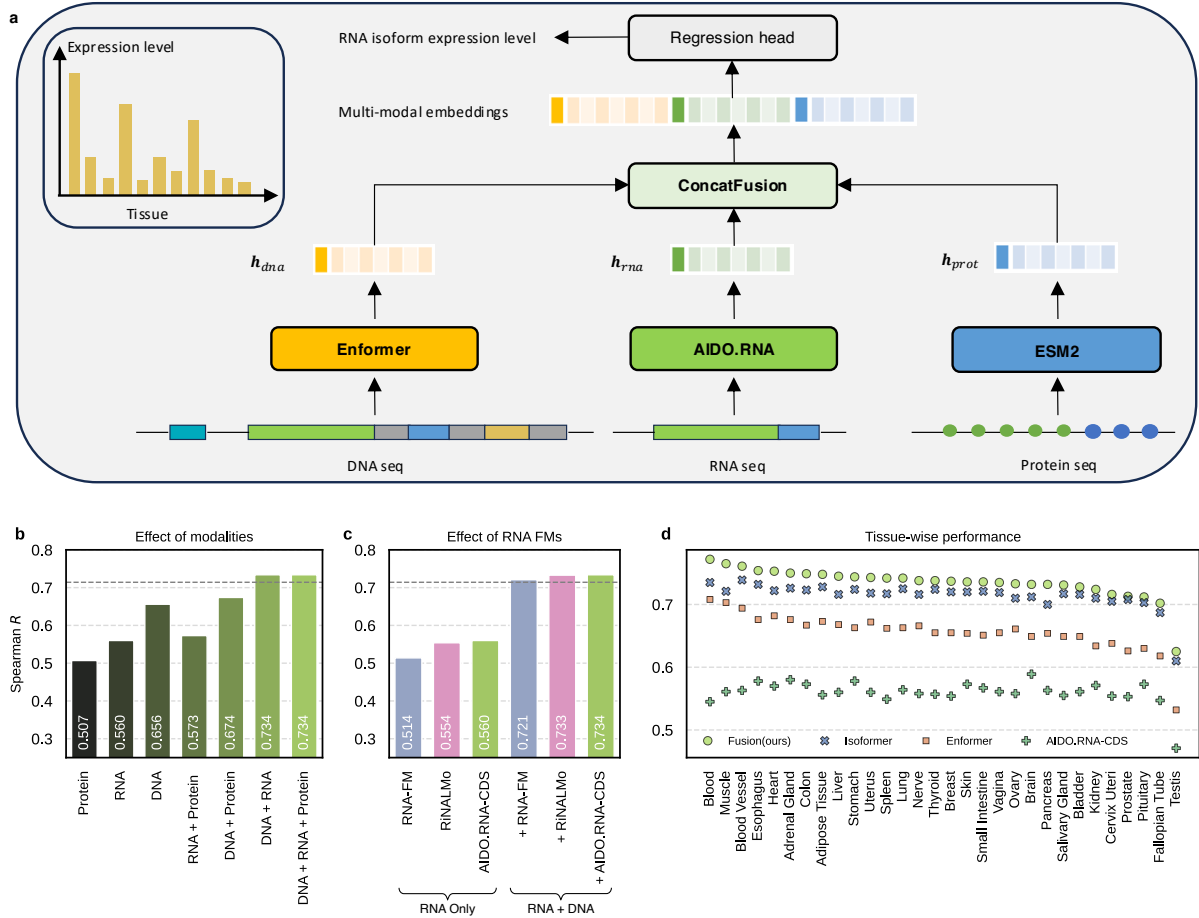


Fig. 4 | AIDO.RNA facilitates multi-modal RNA isoform expression prediction. **a**, For multi-modal RNA isoform expression prediction, we integrate Enformer, AIDO.RNA-CDS, and ESM2 to encode DNA, RNA, and protein sequences, respectively. Each single-modality embedding undergoes mean pooling before being projected into a shared latent space, where embeddings from different modalities are concatenated to form a multi-modal representation. A two-layer MLP regression head is then applied to jointly predict RNA isoform expression levels across 30 human tissues. **b**, We systematically investigate single-modal, bi-modal, and tri-modal modeling approaches for RNA isoform expression prediction to gain insights into how different modalities impact performance. We find that jointly modeling DNA and RNA sequences is essential. We also find that DNA sequence is the most informative modality among the three. The horizontal dashed line represents the performance of IsoFormer (Spearman CC 0.714), a multi-modal model that utilizes DNA, RNA and protein sequences as input. When jointly modeling DNA and RNA, our model outperforms IsoFormer. **c**, To understand how different RNA FMs affect the performance, we fix the DNA sequence encoder to Enformer and change the RNA sequence encoder to be RNA-FM, RiNALMo and AIDO.RNA-CDS. Our AIDO.RNA-CDS achieves the best performance. Notably, when jointly modeling with Enformer, different RNA FMs result in quite similar results. **d**, Our fusion method consistently outperforms IsoFormer across tissues. Its performance is highly correlated with Enformer’s performance (Spearman CC 0.98) while correlates weakly with AIDO.RNA-CDS’s performance (Spearman CC 0.25).

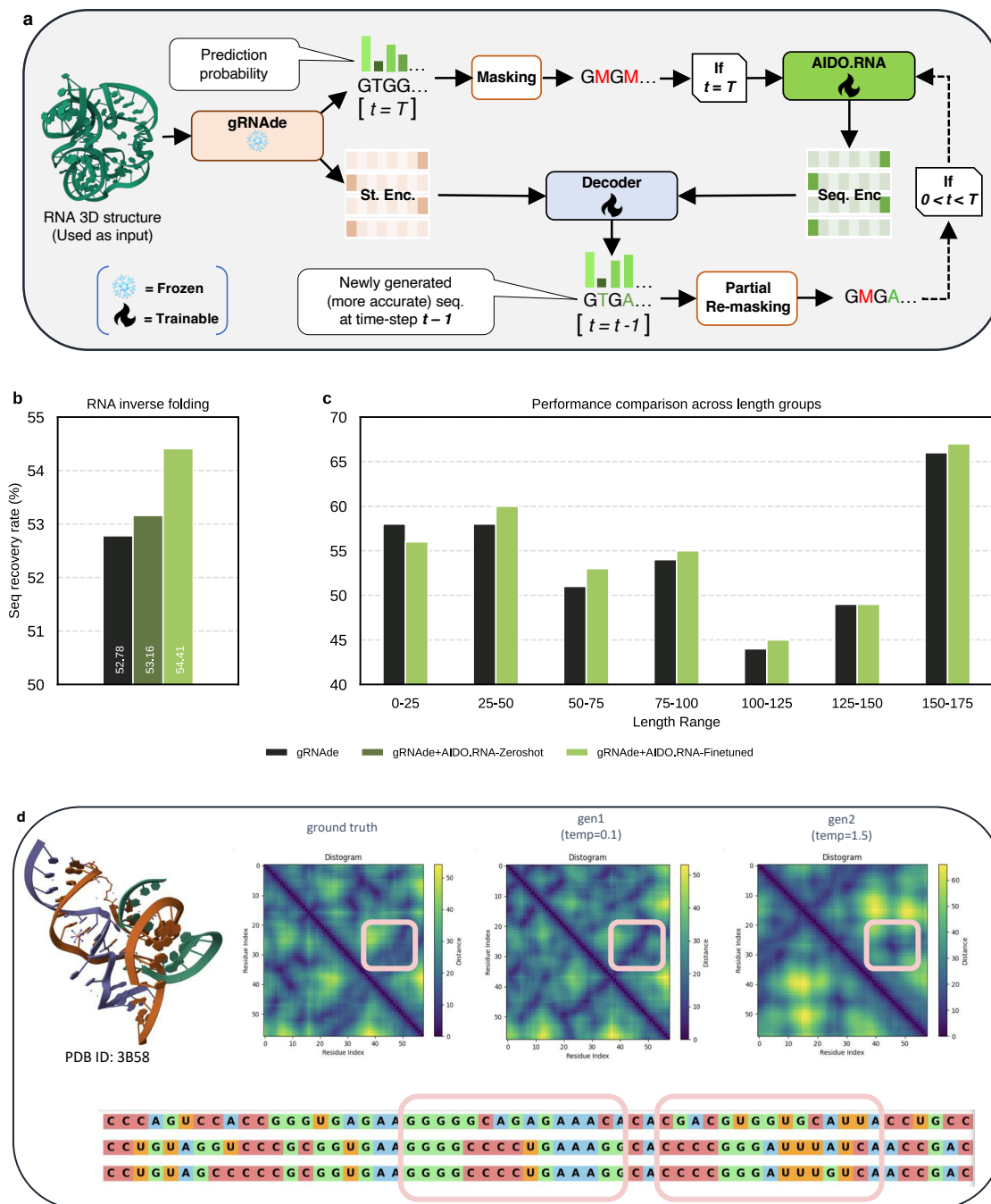


Fig. 5

Fig. 5 | AIDO.RNA benefits RNA sequence design. **a**, Our framework to augment gRNAd with AIDO.RNA for RNA inverse folding, where we fine-tune AIDO.RNA and train a separate decoder. **b**, Quantitative results on how AIDO.RNA enhances RNA inverse folding capabilities of gRNAd while augmented with gRNAd, in both zero-shot and fine-tuned settings. **c**, AIDO.RNA enhances inverse folding performance for most of the ranges of length. **d**, Generated sequence by our inverse folding framework leads to similar structures. Here we show the ground truth histogram (distances between base pairs) of the given RNA 3D structure (PDB ID: 3B58) and the predicted histogram (by RhoFold (48)) of two sequences generated by our framework. **Top row:** At the very left, we show the true structure of the RNA. The second image shows the ground truth histogram. The third image (labeled as *gen1*) shows the predicted histogram of RNA sequences generated by our framework, conditioned on the true structure, with low sampling temperature ($\text{temp}=0.1$). Note that sampling temperature controls how much randomness we allow for the generation (see (25) for details). Even though this histogram is a bit different from the ground truth, it shows structural similarity in several regions. We have marked a region of interest with a rectangle for demonstration. The right-most image in the top row shows how the structure (represented as histogram, denoted at *gen2*) changes when we allow more randomness in sampling, by introducing a high sampling temperature ($\text{temp}=1.5$) (49, 50). As expected, compared to the ground truth and *gen1*, the structural details are lost to a great extent in *gen2* as we allow more randomness. However, it is interesting to see that some structural properties are still preserved, for instance, the region within the marked rectangle. **Bottom row:** We annotate the pair of segments corresponding to the region of interest we discussed above. These segments seem to have contact in 3D space even for the sequence generated with high temperature (*gen2*).

Discussion

In this work, we demonstrate that scaling RNA FMs to billions of parameters enables them to improve efficiency and performance on a diverse range of RNA tasks, including RNA structure and function prediction as well as RNA sequence design. Building on the insights from prior RNA FMs and our exploration on the pre-training datasets, we found that high-quality ncRNA data is crucial for pre-training powerful RNA FMs at scale. Learning conservation dynamics of ncRNA across 30 billion tokens enables AIDO.RNA to capture RNA-specific structural and functional information, which forms a strong inductive bias for improving the performance of downstream tasks with limited data. ncRNA pre-training is essential to achieving this performance at scale, being free from protein-level selection pressures, which can confound the relationship between conservation and function of the RNA molecule itself. Indeed, AIDO.RNA effectively generalizes to mRNA-related tasks, particularly those involving UTRs and even many coding-specific tasks.

Continuing to adapt the model to mRNA coding domains in a second stage of self-supervised learning leads to even more substantial performance improvements over the original pre-trained model on coding-specific tasks. The effectiveness of continued pre-training makes AIDO.RNA truly extensible, particularly for domains where large-scale pre-training data is limited, enabling strong performance without the need to train a FM in the domain from scratch. From another perspective, this result underscores the critical role of pre-training data. In the pursuit of larger, more performant, and more efficient FMs, future works should focus on the influence of different combinations and proportions of RNA sequence types toward downstream performance across diverse tasks.

Since proteins are translated from CDS sequences through many-to-one codon-to-amino acid mappings, CDSs contain richer information than protein sequences about protein structure and

function. For instance, seminal works have shown the effect of codon biases on protein tertiary structures (51). This implies that modeling protein-level tasks at the nucleotide resolution may reveal insights beyond what protein LMs can capture. The findings from our work on protein abundance prediction, where AIDO.RNA-CDS outperforms ESM2, provide empirical support for this hypothesis. Furthermore, AIDO.RNA can work synergistically with DNA FMs to improve gene expression prediction. Our experiments reveal that both DNA and RNA information contribute significantly to transcript isoform expression prediction, and the combination of DNA, RNA, and protein leads to SOTA performance. These findings highlight the potential of multi-modal biological modeling, an essential step in the development of an AI-driven Digital Organism.

For drug discovery applications, we believe that integrating AIDO.RNA into the RNA sequence design pipeline holds great promise for accelerating the discovery of RNA-based therapeutics. By accurately predicting key properties of designed mRNA sequences—such as expression levels and translation efficiency—AIDO.RNA can serve as a unified tool for optimizing many attributes contributing to therapeutic efficacy. Moreover, when combined with a structure encoder and a discrete diffusion algorithm, AIDO.RNA can be used to generate RNA sequences that fold into desired 3D structures, enabling structure-guided RNA design for advanced applications in synthetic biology and medicine.

Data availability

We compiled all the downstream task datasets on Hugging Face. RNA secondary structure prediction dataset is available at <https://huggingface.co/datasets/genbio-ai/rna-secondary-structure-prediction> RNA inverse folding dataset is available at <https://huggingface.co/datasets/genbio-ai/rna-inverse-folding>. We released the RNA isoform expression dataset at <https://huggingface.co/datasets/>

`genbio-ai/transcript_isoform_expression_prediction`.

Datasets for other downstream tasks are available at `https://huggingface.co/datasets/genbio-ai/rna-downstream-tasks`.

Code availability

Code for downstream fine-tuning, evaluation, and dependency mapping analysis is available on AIDO.ModelGenerator at `https://github.com/genbio-ai/ModelGenerator/tree/main/experiments/AIDO.RNA`. Pre-trained model weights and downstream fine-tuning checkpoints are available on Hugging Face at `https://huggingface.co/collections/genbio-ai/aidorna-6747516bb48ed96c847f5dd8`.

References

1. J. D. M.-W. C. Kenton, L. K. Toutanova, *Proceedings of NAACL-HLT* (2019), pp. 4171–4186.
2. A. Radford (2018).
3. Z. Lin, *et al.*, *Science* **379**, 1123 (2023).
4. B. Chen, *et al.*, *arXiv preprint arXiv:2401.06199* (2024).
5. H. Dalla-Torre, *et al.*, *Nature Methods* pp. 1–11 (2024).
6. E. Nguyen, *et al.*, *Science* **386**, eado9336 (2024).
7. G. Bixi, *et al.*, *bioRxiv* pp. 2025–02 (2025).
8. L. A. Rojas, *et al.*, *Nature* **618**, 144 (2023). Publisher: Nature Publishing Group.
9. E. Bettini, M. Locci, *Vaccines* **9**, 147 (2021).
10. M. Akiyama, Y. Sakakibara, *NAR genomics and bioinformatics* **4**, lqac012 (2022).
11. J. Chen, *et al.*, *bioRxiv* pp. 2022–08 (2022).
12. Y. Zhang, *et al.*, *bioRxiv* pp. 2023–03 (2023).
13. X. Wang, *et al.*, *bioRxiv* pp. 2023–07 (2023).
14. K. Chen, *et al.*, *bioRxiv* pp. 2023–01 (2023).
15. S. Li, *et al.*, *bioRxiv* pp. 2023–09 (2023).
16. Y. Yang, *et al.*, *bioRxiv* pp. 2023–09 (2023).

17. N. Boyd, *et al.*, *bioRxiv* pp. 2023–12 (2023).
18. Y. Chu, *et al.*, *Nature Machine Intelligence* pp. 1–12 (2024).
19. R. Josip Penić, T. Vlašić, R. G. Huber, Y. Wan, M. Šikić, *arXiv e-prints* pp. arXiv–2403 (2024).
20. C. Outeiral, C. M. Deane, *Nature Machine Intelligence* **6**, 170 (2024).
21. N. Wang, *et al.*, *Nature Machine Intelligence* pp. 1–10 (2024).
22. A. Vaswani, *Advances in Neural Information Processing Systems* (2017).
23. *Nucleic acids research* **49**, D212 (2021).
24. K. Chen, T. Litfin, J. Singh, J. Zhan, Y. Zhou, *bioRxiv* pp. 2023–02 (2023).
25. C. K. Joshi, *et al.*, *bioRxiv* (2024).
26. J. Singh, K. Paliwal, T. Litfin, J. Singh, Y. Zhou, *Bioinformatics* **38**, 3900 (2022).
27. P. Danaee, *et al.*, *Nucleic acids research* **46**, 5381 (2018).
28. J. Singh, J. Hanson, K. Paliwal, Y. Zhou, *Nature communications* **10**, 5407 (2019).
29. L. Fu, *et al.*, *Nucleic acids research* **50**, e14 (2022).
30. M. Szikszai, M. Wise, A. Datta, M. Ward, D. H. Mathews, *Bioinformatics* **38**, 3892 (2022).
31. P. Tomaz Da Silva, *et al.*, Nucleotide dependency analysis of DNA language models reveals genomic functional elements (2024).
32. N. Scalzitti, *et al.*, *BMC bioinformatics* **22**, 1 (2021).
33. Z. Song, *et al.*, *Nature communications* **12**, 4011 (2021).

34. M. Metkar, C. S. Pepin, M. J. Moore, *Nature Reviews Drug Discovery* **23**, 67 (2024).
35. S. Gururangan, *et al.*, *arXiv preprint arXiv:2004.10964* (2020).
36. P. J. Sample, *et al.*, *Nature biotechnology* **37**, 803 (2019).
37. J. J. Garau-Luis, *et al.*, *Advances in Neural Information Processing Systems* **37**, 78431 (2025).
38. L. Song, E. Segal, E. Xing, *Technical Report* (2024).
39. J. Chery, *Postdoc journal: a journal of postdoctoral research and postdoctoral affairs* **4**, 35 (2016).
40. V. Singh, *Gene* **535**, 1 (2014).
41. F. J. Isaacs, D. J. Dwyer, J. J. Collins, *Nature biotechnology* **24**, 545 (2006).
42. A. Churkin, *et al.*, *Briefings in bioinformatics* **19**, 350 (2018).
43. D. Nori, W. Jin, *arXiv preprint arXiv:2405.18768* (2024).
44. S. S. Sahoo, *et al.*, *arXiv preprint arXiv:2406.07524* (2024).
45. L. Zheng, J. Yuan, L. Yu, L. Kong, *arXiv preprint arXiv:2302.05737* (2023).
46. A. Busch, R. Backofen, *Bioinformatics* **22**, 1823 (2006).
47. R. Das, J. Karanicolas, D. Baker, *Nature methods* **7**, 291 (2010).
48. T. Shen, *et al.*, *arXiv preprint arXiv:2207.01586* (2022).
49. J. Ho, A. Jain, P. Abbeel, *Advances in neural information processing systems* **33**, 6840 (2020).

50. J. Austin, D. D. Johnson, J. Ho, D. Tarlow, R. Van Den Berg, *Advances in Neural Information Processing Systems* **34**, 17981 (2021).
51. A. A. Komar, T. Lesnik, C. Reiss, *FEBS letters* **462**, 387 (1999).
52. J. Su, *et al.*, *Neurocomputing* **568**, 127063 (2024).
53. J. L. Ba, J. R. Kiros, G. E. Hinton, *arXiv preprint arXiv:1607.06450* (2016).
54. I. Loshchilov, *arXiv preprint arXiv:1711.05101* (2017).
55. T. Dao, *arXiv preprint arXiv:2307.08691* (2023).
56. M. Steinegger, J. Söding, *Nature biotechnology* **35**, 1026 (2017).
57. M. F. Sloma, D. H. Mathews, *RNA* **22**, 1808 (2016).
58. D. H. Mathews, *Methods* **162**, 60 (2019).
59. T. M. R. Noviello, F. Ceccarelli, M. Ceccarelli, L. Cerulo, *PLoS computational biology* **16**, e1008415 (2020).
60. I. Kalvari, *et al.*, *Nucleic acids research* **46**, D335 (2018).
61. M. Wang, *et al.*, *Molecular & cellular proteomics* **11**, 492 (2012).
62. J. J. Garau-Luis, *et al.*, *arXiv preprint arXiv:2406.14150* (2024).
63. E. J. Hu, *et al.*, *arXiv preprint arXiv:2106.09685* (2021).
64. X. Wang, *et al.*, *arXiv preprint arXiv:2402.18567* (2024).
65. Z. Zheng, *et al.*, *International conference on machine learning* (PMLR, 2023), pp. 42317–42338.

66. N. Houlsby, *et al.*, *International conference on machine learning* (PMLR, 2019), pp. 2790–2799.

Methods

Pre-training AIDO.RNA

Pre-training dataset RNAcentral database contains a comprehensive ncRNA sequence collection representing all ncRNA types from a broad range of organisms. In specific, we collected sequences from `rnacentral_active.fasta.gz` and `rnacentral_inactive.fasta.gz` in version 24.0 (<https://ftp.ebi.ac.uk/pub/databases/RNAcentral/releases/24.0/sequences/>). We removed duplications using SeqKit toolkit, resulting in a dataset of 42 million unique ncRNA sequences. We observed that the data follows a long tail distribution in terms of RNA types. To better understand the generalizability of our model on different types of RNA, we down-sampled the frequent types, such as rRNA, and up-sampled the infrequent types, such as snoRNA, for validation and testing. The distributions of the training, validation and test sets are shown in **Supplementary Table 2**. The training set consists of 41.5 million distinct ncRNA sequences, comprising a total of 30 billion nucleotides. On average, each sequence has 728 nucleotides. The validation and test sets each contain 80k sequences.

Sequence tokenization We encoded each nucleotide (A, T, C, G) as a token and used N to represent other rare bases (U has been transformed to T in our dataset). The rationale behind using a nucleotide-resolution tokenization scheme is to facilitate structure prediction, which relies on base-level information. We also introduced some special tokens, including [CLS], [SEP], [MASK], [PAD]. The vocabulary size is set to 16. When processing each RNA sequence, we prepended the [CLS] token at the beginning and appended the [SEP] token at the end. This allows the model to separate a full-sized RNA from a cropped one.

Model architecture and training details We adopted an encoder-only transformer as our pre-training model architecture (**Extended Data Fig. 2**). We used Rotary Position Embedding

(RoPE) to allow better position modeling (52). In addition, we used LayerNorm (53) and SwiGLU activation function in our model to make it more expressive and stable in pre-training. We pre-trained two models with different sizes, i.e., 650M, 1.6B, on ncRNA sequences. We used the MLM objective with a masking ratio 0.15. In specific, this involves randomly selecting 15% of the input tokens for each input sequence. For the selected tokens, there are three possible operations: 1) The token is masked with a probability of 0.8; 2) The token is replaced with a random token with a probability of 0.1; 3) The token remains unchanged with a probability of 0.1. Cross-entropy loss was computed on those selected tokens. We trained our models on 30 billion unique nucleotides for 6 epochs. We used the AdamW optimizer with a weight decay of 0.01 (54). The peak learning rate was set to $5e-5$ and gradually decayed to $1e-5$ via the cosine learning rate scheduler. For a comprehensive summary of architecture and training hyperparameters, please refer to **Extended Data Table 2**

Hardware We implemented our pre-trained model using the Megatron-LM framework and pre-trained AIDO.RNA-1.6B on 64 Nvidia A100-80GB GPUs for 4 days. FlashAttention-2 (55) and BFloat16 mixed precision training were used to accelerate the pre-training process.

Continued pre-training of AIDO.RNA on coding sequences To adapt AIDO.RNA to the CDS domain, we continued to pre-train it on 9 million CDS sequences released by Outeiral and Deane (20). We initialized the model weights with the AIDO.RNA and trained it for 13,000 steps with a peak learning rate of $5e-5$ and a mask ratio of 0.25. All the other hyperparameters are the same as in pre-training. The continued pre-trained model is denoted as AIDO.RNA-CDS. We trained AIDO.RNA-1.6B using 8 Nvidia A100-80GB GPUs for 3 days. The computational resources used in continued pre-training are about 9% of the pre-training.

Pre-training AIDO.RNA-1B on MARS database We also pre-trained a 1-billion parameter model on the MARS database, which contains 1.7 billion nucleotide sequences from various biological databases (24). The sequences were aligned to a standardized DNA alphabet and underwent filtering and clustering steps. Extremely long and short sequences (exceeding 4,096 bases or below 10 bases) were excluded. We performed a two-step clustering approach using the MMseqs2 algorithm (56). The first step clustered sequences with at least 90% identity and 80% overlap with the longest sequence in each cluster, resulting in the MARS90 dataset. In the second step, sequences with at least 50% identity and 80% overlap with the longest sequence were clustered, yielding the refined MARS50 dataset. MARS50 comprises 886 million sequences, totaling 344 billion bases, with an average length of 389 bases. Finally, the dataset was randomly split, with 0.2% allocated for validation and another 0.2% for testing purposes. We pre-trained the AIDO.RNA-1B model using the MARS50 dataset. This model has 36 layers, 32 attention heads, and a hidden size of 1,536. The other architecture hyperparameters are the same as our 1.6B model. During pre-training, we trained the model for 145k steps, employing MLM loss with a mask ratio of 15%. We used the AdamW optimizer with a peak learning rate of $1.5e-4$ and a weight decay of 0.02.

Downstream task datasets

RNA secondary structure prediction We employed two benchmark datasets for RNA secondary structure prediction, as developed by Singh et al. (2019) (28) and Szikszai et al. (2022) (30). The former dataset was derived from bpRNA-1m (27) by filtering out sequences longer than 500 bases and applying an 80% sequence similarity cut-off. This preprocessed dataset was divided into three splits: TR0 for training with 10,814 samples, VL0 for validation with 1,300 samples, and TS0 for testing with 1,305 samples. For generalization assessment, we used the dataset by Szikszai et al. (2022) (30), consisting of 3,865 RNAs from nine families.

This dataset was generated from the Archive-II dataset (57) by filtering out sequences longer than 512 nucleotides. It was then split into nine subsets, each time leaving out one family for evaluation while using the remaining families for training and validation. The statistics of each family in the filtered version of the Archive-II dataset are shown in **Supplementary Table 3**. Following RiNALMo, we used the same metric calculation approach proposed by (58). We considered $(i \pm 1, j)$ and $(i, j \pm 1)$ pairings as correct predictions for a nucleotide pairing (i, j) , where i and j denote nucleotide indices in the RNA sequence.

ncRNA family classification We utilized the preprocessed Rfam-novel dataset from Noviello et al. (59), which contains sequences with varying degrees of artificially introduced boundary noise. We leveraged two datasets, one with 0% boundary noise and another with 200% boundary noise in sequences. Sequences with 0% boundary noise represent the original ncRNA sequences, while those with 200% boundary noise contain random nucleotides added to both the 5' and 3' ends, each equivalent to 100% of the original sequence length. The dataset was divided into training, validation, and test sets, comprising 106k, 17k, and 25k samples, respectively, labeled across 88 Rfam classes (60).

Cross-species splice site prediction We utilized data from Spliceator, a comprehensive resource containing experimentally confirmed, error-free splice sites from over 100 eukaryotic species (32). Specifically, we trained and validated our model using the acceptor and donor datasets from the GS-1 subset. To evaluate generalization, we tested the model on data from four species not included in the training set: Zebrafish, Fruit fly, Worm, and Plant. For both the donor and acceptor datasets, we used 18k samples for training, 4k samples for validation, and 20k samples per species for evaluation. Each sample in the dataset is a 400-nucleotide sequence centered on either a splice site or a non-splice site.

RNA modification site prediction We used the processed data from MultiRM (33), which was derived from 20 epi-transcriptome profiles generated using 15 different base-resolution technologies. These profiles encompass 12 types of RNA modifications: *Am*, *Cm*, *Gm*, *Tm*, m^1A , m^5C , m^5U , m^6A , m^6Am , m^7G , Φ , and *I*, with training set sample counts of 1.4k, 1.7k, 1.3k, 2k, 16k, 3k, 3.5k, 65k, 2.2k, 0.8k, and 3k, respectively. Negative sites were randomly selected from unmodified bases within the same transcripts containing the positive sites. Overall, the train/valid/test data contain 305k/3.6k/1.2k samples, with a sequence length of 1,001. For evaluation, we calculated the AUROC score for each modification site and averaged the score across modification sites as an overall metric.

mRNA expression level prediction For the 5' UTR based mRNA expression level prediction, we used the human muscle tissue (Muscle), human prostate cancer cell line PC3 (PC3), and human embryonic kidney 293T cell line (HEK) datasets from Chu et al. (18), each containing 1.3k, 12.6k, and 14.4k samples, respectively. For each sample, it has a 5' UTR sequence with length ≤ 100 and an associated expression level in the corresponding tissue or cell line measured by RNA-Seq RPKM (reads per kilobase of transcript per million mapped reads). For the CDS-based mRNA expression level prediction, we used the transcript abundance prediction data from Outeiral and Deane (20), which contains seven datasets from *A. thaliana*, *D. melanogaster*, *E. coli*, *H. sapiens*, *S. cerevisiae*, *H. volcanii* and *P. pastoris*, respectively. The sample sizes for each dataset are 12.7k, 9.8k, 3.5k, 5.5k, 5.4k, 3.2k, and 4.7k, respectively. For each sample, it has a CDS and an associated expression level in the corresponding organism measured by the natural logarithm of the transcript count per million, ranging between -16 and 16. The sequences in these datasets are generally quite long, with average sequence length ranging from 892 to 1,578.

mRNA translation efficiency prediction We used the Muscle, PC3, HEK datasets from Chu et al. (18), which contain the same 5' UTR sequences as in 5' UTR based mRNA expression

level prediction task but with translation efficiency labels measured by dividing the Ribo-seq RPKM by the RNA-sequencing RPKM.

Mean ribosome load prediction We used the same benchmark dataset used by previous best-performing methods, which is a large-scale synthetic Human 5' UTR library (36) consisting of 83,919 5' UTRs spanning 75 distinct lengths, each paired with its associated MRLs. To ensure balanced validation, 7,600 sequences were uniformly sampled at each length (namely Random7600), while the remaining data is allocated for training purposes. Note that we use the same splitting as the previous studies for a fair comparison (11, 36). Furthermore, an extra dataset containing 7,600 authentic human 5' UTRs, distributed in a similar manner as the synthetic collection, is employed as the test set (namely Human7600).

Protein abundance prediction We used the public data from Outeiral and Deane (20) for protein abundance prediction. It contains five datasets from *A. thaliana*, *D. melanogaster*, *E.coli*, *H. sapiens*, and *S. cerevisiae*, respectively. The sample sizes for each dataset are 11.5k, 10.3k, 3.6k, 11.8k, and 4.9k, respectively. For each sample, it has a CDS and an associated protein abundance in the corresponding organism measured by the estimated number of copies per cell annotated in PAXdb (61), ranging between 0 and 10^5 . The sequences in these datasets are generally quite long, with average sequence length ranging from 952 to 1,654.

Multi-modal RNA isoform expression prediction We curated the human transcript isoform expression dataset from the GTEx portal following the preprocessing pipeline in (62). We downloaded the RNA-seq Transcript TPMs file from the bulk tissue expression in GTEx Analysis V8. The table contains transcript expression collected from 30 non-diseased tissues in nearly 1000 human individuals. We averaged the transcript expression measurements across individuals to obtain population-level tissue-specific transcript expression data. For each transcript, we used

the associated Ensembl transcript id to retrieve a DNA sequence of length 197k centering on the transcript start site using the human reference genome GRCh38. We also retrieved the protein sequence for each transcript if it encodes a protein. For each sample in the dataset, we have a DNA sequence, an RNA sequence, a protein sequence (potentially), and expression levels of the transcript across 30 tissues. Furthermore, we split the dataset into training, validation and test sets based on the chromosomes. In specific, we used chromosomes 1-19 for training, 20 for validation and 21-22 for test. The training, validation, and test sets contain 175k, 4k and 6.5k samples respectively. We performed $\log_2(0.0001 + x)$ transform for the labels.

Downstream fine-tuning and analysis methods

RNA secondary structure prediction

We generated pair representations by applying outer concatenation to the LM’s outputs, concatenating the representation of nucleotide j with that of nucleotide i for each pair (i, j) . In the prediction head, the concatenated representation is first linearly projected to an embedding dimension of 64. This is followed by two bottlenecked ResNet-2D blocks and a convolution layer, all with 64 kernels of size 3 and followed by instance normalization and ReLU activation. The output matrix represents pairing probability logits for nucleotide pairs, with binary cross-entropy loss calculated only for elements above the main diagonal due to the symmetry of secondary structures. To train the model, we used the AdamW optimizer with weight decay 0.01. To prevent the model from overfitting, we regularized both the LM and the prediction head with a dropout rate of 0.1. Fine-tuning involved a gradual unfreezing method, starting with training the prediction head for the first three epochs, then unfreezing three additional layers every three epochs over a total of 60 epochs with a batch size of 4, and a learning rate decreasing from 10^{-4} to 10^{-5} . A greedy algorithm was used to convert base pairing probabilities into secondary structures, prioritizing the highest probabilities and excluding conflicting pairs, avoiding non-

canonical pairings and sharp hairpin loops ($|i - j| < 4$). The classification threshold was tuned on the validation set to balance the pairing ratio.

RNA function and mRNA expression prediction tasks

The ncRNA family classification, cross-species splice site prediction, RNA modification site prediction, and the mRNA-related tasks in Section can be formulated as sequence-level prediction tasks. We used similar fine-tuning settings for these tasks except for mean ribosome load prediction, as summarized in **Extended Data Table 3**. For tasks with short sequences, we applied fully fine-tuning to maximize potential performance. For tasks with long sequences, we used LoRA (63), a parameter-efficient fine-tuning approach that freezes the pre-trained model weights while training only the prediction head and rank decomposition matrices added to each transformer layer. For classification tasks, we used cross-entropy loss. For regression tasks, we performed label normalization and used mean square error as the loss function. For the protein abundance prediction task, we converted the label y to natural logarithm space by using $\log(1 + y)$ before the label normalization. For the mean ribosome load prediction task, we used a different head other than MLP. The prediction head comprises a linear projection into 256 dimensions and nine ResNet-1D blocks. Each block includes two convolution layers with 256 kernels of size 3, followed by instance normalization and an ELU activation function. The model underwent fine-tuning for 60 epochs. During the first 3 epochs, only the prediction head was trained. The learning rate started at 10^{-4} and linearly decreased to 10^{-5} over the first 5,000 training steps before remaining constant. The batch size for training was set to 64. For all the tasks, we trained the model using AdamW with a weight decay of 0.01 and a hidden layer dropout rate of 0.1.

Multi-modal RNA isoform expression prediction

We proposed a simple and effective multi-modal learning method for this task, which leveraged multiple single-modality FMs as sequence encoders and used a simple aggregation module

to integrate the information across modalities to predict the target. Each single-modality embedding underwent mean pooling before being projected into a shared latent space, where embeddings from different modalities were concatenated to form a multi-modal representation (**Fig. 4a**). In specific, let f_{omic} denote an omic (DNA / RNA / protein) sequence encoder, the single-modality embedding for an omic sequence x_{omic} is computed as follows: $\mathbf{h}_{omic} = \text{mean_pooling}(f_{omic}(x_{omic})) \in R^{d_{omic}}$. The single-modality embedding is then projected into a hidden space of dimension d by: $\mathbf{h}'_{omic} = W_{omic}^T \mathbf{h}_{omic} + b$, where $W_{omic} \in R^{d_{omic} \times d}$, b is a bias term. The multi-modal embedding is a simple concatenation of the projected single-modality embeddings, computed as follows: $\mathbf{h}_{mm} = [\mathbf{h}'_{dna}; \mathbf{h}'_{rna}; \mathbf{h}'_{prot}]$. Compared to IsoFormer, which uses cross-attention as the fusion mechanism, our fusion method is more straightforward and allows for symmetric information integration, robust to the change of encoders' order. Notably, our method and code implementation are model-agnostic and can be used for various multi-modal tasks, such as protein-DNA/RNA binding prediction. This flexibility allows seamless application to other tasks and FMs.

In our experiments, we set f_{dna} to be Enformer, f_{rna} to be AIDO.RNA-CDS, f_{prot} to be ESM2-150M, $d = 1024$. We used a two-layer MLP with hidden size 1024 as the regression head. We performed $\log_2(0.0001 + x)$ transform for the labels. We used mean square error as the loss function with label normalization to reduce the effect of outliers. During training, we used fully fine-tuning for Enformer while using LoRA fine-tuning for AIDO.RNA-CDS and ESM2-150M. We used a learning rate of 1e-4 and a batch size of 16. We trained the model for 3 epochs and selected the best checkpoint based on the validation Spearman correlation coefficient. For the bi-modal models (DNA+RNA, DNA+protein, RNA+protein), we used similar training settings. For mono-modal models (DNA, RNA, protein), we used fully fine-tuning for each model.

RNA inverse folding

We augmented AIDO.RNA with a SOTA RNA inverse folding method, gRNAd (25), to further enhance the generated sequence by gRNAd with iterative refinement. We conducted two sets of experiments: (a) Adaptation with conditional diffusion – where we fine-tuned our AIDO.RNA encoder with conditional diffusion for inverse folding, (b) Zero-shot generation – where we directly leveraged AIDO.RNA, without any fine-tuning, to iteratively refine the RNA sequence generated by gRNAd.

Adaptation with conditional diffusion As shown in **Fig. 5a**, we leveraged masked diffusion language modeling (44, 45) to adapt AIDO.RNA for inverse folding. The goal is to learn a marginal distribution $p_\theta(x_{t-1}|x_t)$ that approximates the true transition probability $q(x_{t-1}|x_t, x_0)$ of estimating a less noisy variant of an RNA sequence, x_{t-1} , given a relatively more noisy variant x_t , where x_0 is the native RNA sequence.

For a given 3D RNA structure, we begin by taking the predicted sequence and structural embedding by gRNAd, S_0 and e^{st} , respectively. Note that we also take into consideration the probability P_{S_0} of the nucleotides S_0 , as predicted by gRNAd. During training, we mask out $\lceil \frac{tS_0}{\tau} \rceil$ (where $t \sim \mathcal{U}[1, \tau]$) least probable nucleotides from S_0 based on P_{S_0} , where τ is the maximum number of denoising steps we are considering. This produces x_t , which can be assumed as a noisy variant of the expected RNA sequence. We then pass x_t through AIDO.RNA’s encoder that produces sequence embedding e_t^{seq} . Then e^{st} and e_t^{seq} are processed by an adapter module (64, 65), which in our design is a multi-head self-attention layer (22) with bottleneck (66), that generates a new estimate of the RNA sequence x_{t-1} . Note that here AIDO.RNA and the adapter combined work as the estimated transition function $p_\theta(x_{t-1}|x_t)$ we discussed before. We then optimize the diffusion objective in Equation **1**

$$\mathcal{J} = \mathbb{E}_{t \sim \mathcal{U}[1, \tau], x_{t-1} \sim p_\theta} \left[\frac{\mathcal{L}_{CE}(x_0, x_{t-1})}{t} \right], \quad (1)$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss.

After training, we can generate sequences starting from the initial estimate S_0 and structural embedding e^{str} provided by gRNAd, and applying masking to the M least confident tokens in S_0 (based on predicted class probabilities of the tokens), resulting in a masked sequence $x_t = x_T$ (where $T < M$ is the number of denoising steps to unmask M tokens during inference). We then iteratively denoise this sequence for T steps, where at each step t we unmask only the most probable $\lceil \frac{M}{T} \rceil$ nucleotides (tokens) that were previously masked. This process gives us $\hat{x}_0$, our final estimate.

Zero-shot generation This method utilizes the generation approach in the conditional diffusion discussed above; however, it operates entirely in a zero-shot manner. This means we utilize our pre-trained AIDO.RNA without any fine-tuning or the implementation of an adapter module for decoding. Rather, we directly leverage the language modeling head from our pre-training stage to decode the sequence. Specifically, we begin with the initial estimate S_0 generated by gRNAd. Subsequently, we apply a masking technique to the top M least confident tokens in S_0 , resulting in a masked sequence x_t . Following this, we iteratively denoise the sequence over several steps, gradually refining it until we arrive at $\hat{x}_0$, which serves as our generated sequence. This approach ensures that we can leverage the pre-trained capabilities of AIDO.RNA effectively while maintaining simplicity in the process.

Nucleotide dependency analysis

To understand how nucleotide substitutions at one position affect the probabilities of nucleotides at other positions in our pre-trained AIDO.RNA model, we performed nucleotide dependency analysis using the method proposed by Silva et al. (31). In specific, given an RNA sequence of

length L , we computed the dependency map $E \in R^{L \times L}$ as follows:

$$e_{i,j} = \max_{k,q \in \{A,T,C,G\}} \left| \log_2 \left(\frac{\hat{\text{odds}}(n_j = k \mid n_1, \dots, n_i = q, \dots, n_L)}{\hat{\text{odds}}(n_j = k \mid n_1, \dots, n_L)} \right) \right|,$$

where $e_{i,j}$ is the "pixel" value of the dependency map at (i, j) , k and q are the key and query nucleotides, n_i is the i -th nucleotide in the RNA sequence, and $\hat{\text{odds}}$ are computed based on the output probabilities from our pre-trained AIDO.RNA as follows:

$$\begin{aligned} \hat{\text{odds}}(n_j = k \mid n_1, \dots, n_i = q, \dots, n_L) &= \frac{P((n_j = k \mid n_1, \dots, n_i = q, \dots, n_L))}{P((n_j \neq k \mid n_1, \dots, n_i = q, \dots, n_L))} \\ \hat{\text{odds}}(n_j = k \mid n_1, \dots, n_L) &= \frac{P((n_j = k \mid n_1, \dots, n_L))}{P((n_j \neq k \mid n_1, \dots, n_L))} \end{aligned}$$

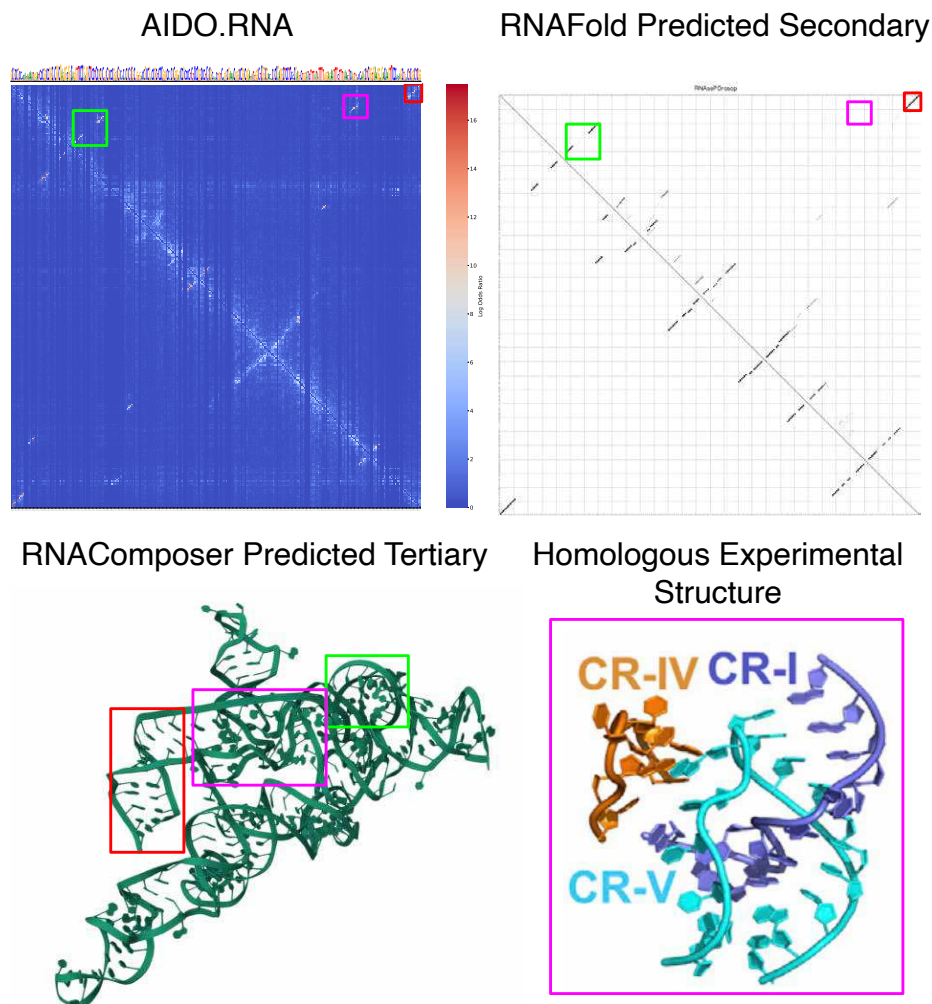
Extended Data

Model	Pre-training data	SS TS0 (F1)	MRL(R^2)	CSP (F1)
Uni-RNA (13)	1B noisy RNA sequences/DNA fragments	-	0.85	0.947
AIDO.RNA-1B(ours)	886M noisy RNA sequences/DNA fragments	0.69	0.86	0.967
RiNALMo (19)	36M high-quality ncRNA sequences	0.75	0.86	0.945
AIDO.RNA-1.6B(ours)	42M high-quality ncRNA sequences	0.78	0.86	0.949

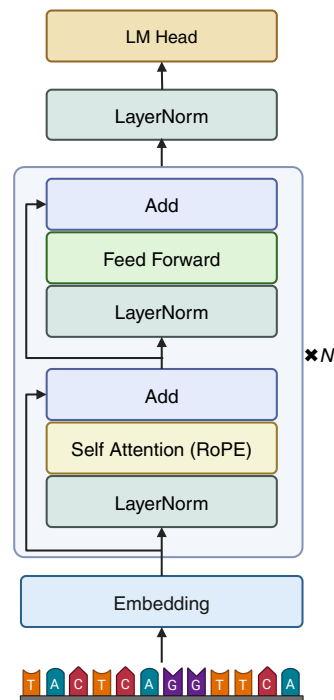
Extended Data Table 1 | High-quality ncRNA data proves more effective for pre-training than large volumes of low-quality data. AIDO.RNA-1B (Methods), trained on a large but lower-quality dataset, underperforms compared to RiNALMo, which is pre-trained on a smaller, high-quality ncRNA dataset. This suggests that simply increasing dataset size does not guarantee improved downstream performance, especially when the data quality is low or the distribution diverges from the target tasks. We reproduced RiNALMo’s performance on cross-species splice site prediction using their publicly available codebase and the hyperparameters specified in their paper.

	AIDO.RNA-650M	AIDO.RNA-1.6B
dataset	RNAcentral (30B tokens)	RNAcentral (30B tokens)
num-layers	33	32
hidden-size	1,280	2,048
num-attn-heads	20	32
ffn-hidden-size	3,392	5,440
tie embedding	TRUE	TRUE
max-seq-len	1,024	1,024
batch size	2M tokens	2M tokens
epochs	6	6
optimizer	AdamW	AdamW
peak learning rate	5e-5	5e-5
min learning rate	1e-5	1e-5
lr scheduler	cosine	cosine
warmup steps	2,000	2,000
weight decay	0.01	0.01
hidden dropout	0.0	0.0
attention dropout	0.1	0.1
residual dropout	0.1	0.1
clipnorm	1.0	1.0
precision	bf16	bf16
use flash attention	No	Yes
distributed backend	DDP	DDP

Extended Data Table 2 | Hyperparameters for training AIDO.RNA models.



Extended Data Fig. 1 | Nucleotide dependencies of AIDO.RNA reveal RNA secondary and tertiary structure contacts. The green and red boxes show short-range and long-range contacts in the secondary structure while the fuchsia purple box reveals contacts in the tertiary structure.



Extended Data Fig. 2 | Model architecture of AIDO.RNA. It is an encoder-only transformer with pre-LayerNorm (53), Rotary Position Embedding (RoPE) (52) and SwiGLU activation function.

Task	Input	Pooling	Pred head	FT setting	LR	BS	Epoch
ncRNA family classification	ncRNA	mean	2 layer MLP (d=512)	LoRA (r=16,alpha=32, lora_dropout=0.1)	4e-4	64	15
Cross-species splice site prediction	pre-mRNA	CLS	3 layer MLP (d=[512,128])	LoRA (r=32,alpha=64, lora_dropout=0.1)	2.5e-4	32	10
RNA modification site prediction	RNA	CLS	2 layer MLP (d=512)	LoRA (r=16,alpha=32, lora_dropout=0.1)	4e-4	64	10
mRNA expression level prediction	5' UTR	mean	2 layer MLP (d=512)	Fully fine-tuning, 10 fold cross-validation	1.e-5	Muscle: 8 PC3/HEK: 32	30
	CDS	mean	2 layer MLP (d=512)	LoRA (r=32,alpha=64, lora_dropout=0.1) 5 fold cross-validation	3e-4	16	15
mRNA translation efficiency prediction	5' UTR	mean	2 layer MLP (d=512)	Fully fine-tuning, 10 fold cross-validation	1e-5	Muscle: 8 PC3/HEK: 32	30
Protein abundance prediction	CDS	mean	2 layer MLP (d=512)	LoRA (r=32,alpha=64, lora_dropout=0.1) 5 fold cross-validation	3e-4	16	15

Extended Data Table 3 | Fine-tuning settings for sequence-level prediction tasks.

Acknowledgements

We would like to thank the authors of gRNAde for providing their models' checkpoints and configuration files for our inverse folding experiment.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInfomationAIDORNA.pdf](#)