Nüwa.RNA: An RNA Foundation Model for Unified Representation with Deep Structure Infusion

Kun Huang[1] **Jiyang Li**[1] **Xin Guo**[1,2]* **Limei Han**[1,2] **Yuan Cheng**[1,2]*
[1]Shanghai Academy of Artificial Intelligence for Science, Shanghai 200003, China
[2]Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, China

Abstract

RNA function arises from a hierarchical folding process in which primary sequences form secondary structures that govern biological activity. While foundation models have proven effective for learning transferable RNA representations, most existing approaches rely on primary sequence alone or incorporate structural information only superficially during pre-training. Three problems limit current approaches: experimentally determined secondary structures are scarce, computational pseudo-labels are noisy, and no existing model integrates structure at multiple levels across masking, architecture, and supervision. Here, we present **Nüwa.RNA**, a pre-trained foundation model built on a *structure-driven multi-level pre-training framework* that integrates secondary structure at all three stages of pre-training simultaneously. The framework has three components. First, dual-granularity masking combines single-token masking with structure motif masking, treating entire structural elements as indivisible units. Second, a noisy structural attention bias injects curriculum-corrupted contact maps into every Transformer layer to model pseudo-label uncertainty. Third, a structure denoising objective recovers clean base-pairing states from corrupted input, closing the denoising loop. Pre-trained on a large-scale non-coding RNA corpus, Nüwa.RNA achieves state-of-the-art performance on 10 of 13 BEACON benchmark tasks, with up to $+7.83$ F1 improvement on secondary structure prediction, and attains the highest zero-shot secondary structure prediction accuracy among single-sequence RNA foundation models.

1 Introduction

Beyond their classical role as intermediaries in gene expression, RNA molecules regulate chromatin remodeling, catalyze biochemical reactions, and sense small-molecule ligands [1, 2]. These functions arise from a hierarchical folding process: primary sequences form local base pairs that define secondary structures, which further collapse into compact tertiary folds. Because each level of this hierarchy constrains the next, accurately modeling the sequence-to-structure relationship is a prerequisite for understanding RNA function, with implications for drug target discovery, RNA-based therapeutics, and synthetic biology.

Self-supervised pre-training on large-scale RNA sequence databases is now the standard approach for learning transferable RNA representations. General-purpose models such as RNA-FM [3], RiNALMo [4] (650M parameters), and AIDO.RNA [5] (1.6B parameters) have shown that scaling masked language modeling to millions of non-coding RNA sequences yields strong transfer to both structure and function prediction tasks. Specialized models further target specific RNA types, including plant transcriptomes [6], splice sites [7], untranslated regions [8], and coding sequences [9]. Still, most RNA foundation models rely exclusively on primary sequence during pre-training and treat

*Correspondence to: guoxin@sais.org.cn, cheng_yuan@fudan.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

structure as a purely downstream target. Several recent models incorporate structural information: ERNIE-RNA [10] through pairwise attention biases, structRFM [11] through paired-base co-masking, and OmniGenome [12] through dot-bracket input tokens. Three challenges remain.

(1) **Structure integration is single-pathway and shallow.** Each existing structure-aware model incorporates structural information through only one mechanism: attention bias [10], mask construction [11], or input representation [12]. This single-pathway design limits how much structure can inform the learned representations. ERNIE-RNA lacks explicit structural loss supervision; structRFM does not condition the attention computation on structural priors; OmniGenome relies on dot-bracket tokens appended to the input sequence, which introduces a lossy discretization of the contact map and risks contaminating the sequence embedding space.

(2) **Pseudo-label noise is ignored rather than modeled.** Because experimentally determined secondary structures remain scarce, models must rely on computationally predicted pseudo-labels, whose error rates remain substantial on non-canonical or complex topologies. Existing structure-aware methods directly trust these predictions without modeling their uncertainty, and systematic folding errors propagate into the learned representations.

(3) **Masking strategies are agnostic to RNA structural semantics.** Standard token-level or span-level masking treats every position as equally informative, ignoring cooperative dependencies within secondary structure. RNA function often depends on complete structural elements such as entire stems or hairpin loops. While RNAErnie [13] introduced motif-aware masking, its motifs are one-dimensional sequence patterns (RNA-binding protein consensus sites and frequent k-mers) that carry no secondary structure semantics. No existing method masks at the granularity of two-dimensional structural elements parsed from contact maps, where entire stems, loops, and junctions are treated as indivisible units.

To address these challenges, we present **Nüwa.RNA**, a pre-trained foundation model built on a *structure-driven multi-level pre-training framework* that integrates secondary structure at all three stages of pre-training: mask construction, model architecture, and loss supervision. The central observation is that pseudo-label noise can serve as a training signal: by deliberately corrupting structural input and training the model to reconstruct the original base-pairing annotations, the model internalizes canonical base-pairing dependencies, which acts as a structural regularizer that improves robustness to noisy input.

At the masking level, **dual-granularity masking** combines random single-token masking with structure motif masking that treats entire structural elements as indivisible units. At the architectural level, a **noisy structural attention bias** injects curriculum-corrupted contact maps into every Transformer layer with noise that progressively increases over training. At the supervision level, a **structure denoising objective** recovers clean base-pairing states from corrupted input. The noise injection and denoising objective are coupled: the same corrupted contact map that biases attention serves as the structural input from which the model must reconstruct clean base pairs, forming a closed denoising loop. Because structure enters only through attention biases and loss functions rather than input tokens, the full pairwise information of the contact map is preserved.

On the BEACON benchmark, Nüwa.RNA achieves the best performance on 10 of 13 tasks, with the largest gains on structure-related evaluations ($+7.83$ F1 on secondary structure prediction). It also attains the highest zero-shot secondary structure prediction accuracy among single-sequence RNA foundation models. Ablation experiments confirm that the three levels are complementary, each providing information that the other two cannot fully recover. Our main contributions are:

(1) Nüwa.RNA is the first RNA foundation model to integrate secondary structure at all three stages of pre-training, mask construction, model architecture, and loss supervision, through a unified structure-driven multi-level framework.

(2) We propose a noise-injection–denoising mechanism that converts pseudo-label uncertainty into a training signal, so the model learns robust structural representations from imperfect computational annotations.

(3) Nüwa.RNA achieves state-of-the-art performance on 10 of 13 BEACON benchmark tasks, with up to $+7.83$ F1 improvement on structure-related evaluations, and establishes the best zero-shot secondary structure prediction among single-sequence RNA foundation models.

2 Related Work

General-purpose RNA language models. Self-supervised pre-training on nucleotide sequences is now the standard approach for learning transferable RNA representations. RNA-FM [3] first adapted BERT-style masked language modeling (MLM) to RNA with a 100M-parameter encoder. Subsequent models have scaled substantially: RNAErnie [13] introduced motif-aware masking over sequence-level patterns (RBP binding sites and frequent k-mers), RiNALMo [4] trained a 650M-parameter encoder with strong generalization on the BEACON benchmark [14], and AIDO.RNA [5] further scaled to 1.6 billion parameters. However, none of these models incorporate structural information during pre-training, so the folding constraints that govern RNA function are absent from the learned embeddings.

Specialized RNA language models. Several models target specific RNA types or functions, including PlantRNA-FM [6] for plant transcriptomes, SpliceBERT [7] for splice sites via cross-species transfer, UTR-LM [8] for mRNA $5'$ untranslated regions with RNAfold-predicted structure as auxiliary supervision, and CaLM [9] for codon-level representations. While effective in their respective domains, their narrow focus limits generalizability to other RNA types and tasks.

Structure-aware RNA models. A few recent models incorporate structural information into pre-training, though each does so through a single pathway. ERNIE-RNA [10] encodes base-pairing rules as fixed pairwise attention biases but lacks structural loss supervision. structRFM [11] co-masks paired bases during MLM yet directly trusts BPfold [15] pseudo-labels without modeling their noise. OmniGenome [12] appends dot-bracket tokens to the input, but this lossy discretization of the contact map risks contaminating the sequence embedding space and discards pairwise distance information. In contrast, Nüwa.RNA injects structure at all three stages of pre-training (mask construction, model architecture, and loss supervision) while modeling pseudo-label uncertainty, and represents structure through contact-map-derived attention biases to preserve both the embedding space and full pairwise structural information.

3 Methods

3.1 Pre-training Data

Following Zhu et al. [11], we construct a sequence–structure corpus from RNAcentral [16], the largest public non-coding RNA database. We choose RNAcentral over larger but noisier alternatives such as MARS, since Zou et al. [5] show that curated ncRNA sequences from RNAcentral outperform a much larger MARS corpus at the same model scale. Secondary structures are predicted with BPfold [15], to obtain canonical base pair annotations (A–U, G–C, G–U) with per-pair confidence scores. The pre-training length is set to 1,536 in the first phase, which retains 93.4% of sequences in full (vs. 62.7% at 512 and 76.1% at 1,024), and extended to 8,192 in the second phase (Section 3.4). The resulting corpus contains approximately 32 million sequences totaling 20 billion nucleotides.

Each secondary structure is represented as a binary contact map $\mathbf{C} \in \{0,1\}^{L\times L}$, where $C(i,j) = 1$ indicates that positions i and j form a canonical base pair. Because these annotations are computationally predicted rather than experimentally determined, we treat them as *pseudo-labels* with inherent noise and design our training to be robust to prediction errors (Section 3.3.2). To prevent data leakage, we remove from the pre-training corpus any sequence with $\geq$80% identity to sequences in downstream test sets, following the deduplication protocol of CD-HIT at the 80% threshold. From each contact map, we parse six types of structural elements (stems, hairpin loops, internal loops, bulges, multi-branch junctions, and external loops) using the algorithm described in Appendix A. These parsed elements serve as indivisible masking units for structure motif masking (Section 3.3.1).

3.2 Model Architecture

Nüwa.RNA is an encoder-only Transformer designed for long RNA sequences (Figure 1, top left). It combines single-nucleotide tokenization, hybrid local-global attention, and a noisy structural attention bias that injects secondary structure priors into the attention computation.

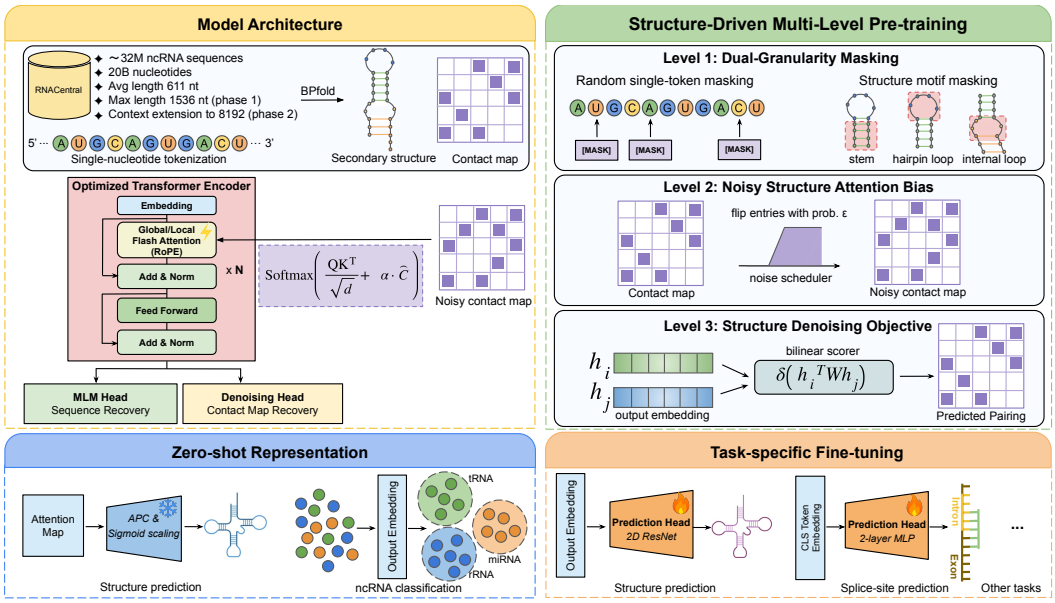


Figure 1: Overview of Nüwa.RNA. **Top left**: Encoder-only Transformer with single-nucleotide tokenization, hybrid local/global attention, and noisy contact-map attention biases (learnable per-layer scalar $\alpha^{(l)}$). **Top right**: Multi-level pre-training: Level 1 dual-granularity masking, Level 2 curriculum-noised attention bias, Level 3 structure denoising. **Bottom left**: Zero-shot capabilities. **Bottom right**: Task-specific fine-tuning heads.

Single-nucleotide tokenization. Each base (A, U, G, C) is an individual token (vocabulary size 9). Unlike k-mer or BPE tokenization, this preserves exact positional correspondence, avoids segmentation artifacts from point mutations, and enables nucleotide-resolution predictions of base-pairing interactions.

Hybrid attention. Following ModernBERT [17], the backbone combines sliding window attention (window size 128) with global attention layers interleaved at every third layer. This hybrid scheme is well suited to RNA, where local secondary structure motifs coexist with distal base pairs that connect nucleotides hundreds of positions apart. All layers use Flash Attention [18] for efficient processing of sequences up to 8,192 nucleotides, and Rotary Positional Embeddings (RoPE) [19].

Structural attention bias. At every Transformer layer, a contact-map-derived bias term is added to the attention logits before the softmax, modulated by a learnable per-layer scalar $\alpha^{(l)}$ that controls how strongly each layer relies on the structural prior. Unlike ERNIE-RNA [10], which propagates the preceding layer's attention map as the next layer's bias, Nüwa.RNA provides the same explicit contact map signal to all layers while allowing each layer to calibrate its reliance independently. Details of noise injection and curriculum scheduling are described in Section 3.3.2.

Prediction heads and inference. Two heads supervise pre-training: masked language modeling for sequence recovery and contact map denoising for structure refinement (Section 3.3.3). Structure enters only through attention biases and the denoising loss, so the sequence embedding space remains intact and the model can operate without structural input at inference time. The architecture supports four model sizes (20M, 100M, 650M, 1.6B; Appendix C), with scaling behavior analyzed in Section 4.4.

3.3 Structure-Driven Multi-Level Pre-training

We inject secondary structure information at three complementary levels: mask construction (Level 1), attention computation (Level 2), and loss supervision (Level 3). The overall training objective combines a masked language modeling loss with a structure denoising loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MLM}} + \lambda \mathcal{L}_{\text{denoise}} \tag{1}$$

where $\mathcal{L}_{\text{MLM}}$ supervises nucleotide recovery at masked positions and $\mathcal{L}_{\text{denoise}}$ trains the model to reconstruct clean base-pairing states from corrupted structural input. We describe each level below.

3.3.1 Level 1: Dual-Granularity Masking

Standard masked language modeling masks tokens uniformly at random, ignoring the cooperative dependencies within RNA secondary structure. We design two complementary masking strategies at different granularities, both contributing to $\mathcal{L}_{\text{MLM}}$ (Figure 1, top right).

Random single-token masking. Individual nucleotides are masked with probability p_{mask}, following the standard BERT protocol. This captures position-independent sequence regularities.

Structure motif masking. Instead of masking random contiguous spans [20], we mask *entire structural elements* as indivisible units. Let $\mathcal{E} = \{E_1, E_2, \ldots\}$ be the set of structural elements parsed from the contact map (Appendix A), where each E_k is the set of positions constituting a stem, hairpin loop, internal loop, bulge, junction, or external loop. We randomly sample a subset for masking:

$$\mathcal{M}_{\text{motif}} = \bigcup_{k \in \text{Sample}(\mathcal{E})} E_k \tag{2}$$

For stems, all paired positions on both the $5'$ and $3'$ strands are masked together; for loops, all unpaired positions are masked as a unit. The model must therefore learn cooperative base-pairing rules and composition constraints of structural elements.

Sequence recovery loss. Each sequence is processed in two forward passes per step, one per masking strategy, and the losses are summed (Section 3.4). Writing $\mathcal{M} = \mathcal{M}_{\text{rand}} \cup \mathcal{M}_{\text{motif}}$ for compactness:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log p_\theta(x_i^* \mid \mathbf{h}_i) \tag{3}$$

where x_i^* is the ground-truth nucleotide and $\mathbf{h}_i$ is the encoder output at position i.

3.3.2 Level 2: Noisy Structural Attention Bias

Level 2 injects secondary structure priors into the attention mechanism with deliberate noise to model pseudo-label uncertainty.

Noise injection. Given the contact map $\mathbf{C}$, we construct a noisy version $\tilde{\mathbf{C}}$ by applying asymmetric flipping probabilities to preserve sparsity: true base pairs ($C(i,j)$=1) are flipped to 0 with probability ϵ, while unpaired entries ($C(i,j)$=0) are flipped to 1 with probability $\epsilon \cdot \rho$, where ρ is the positive-to-negative ratio of the contact map:

$$\tilde{C}(i,j) = \begin{cases} 1 - C(i,j), & \text{w.p. } \epsilon \text{ (if } C(i,j) = 1) \text{ or } \epsilon \cdot \rho \text{ (if } C(i,j) = 0) \\ C(i,j), & \text{otherwise} \end{cases} \tag{4}$$

Attention bias. The noisy contact map is injected as an additive bias at every layer l. In sliding window layers the bias acts only within the local window; long-range structural guidance is thus provided exclusively through the global layers:

$$\text{Attn}^{(l)}(i,j) = \text{softmax}\left(\frac{\mathbf{Q}_i^{(l)} {\mathbf{K}_j^{(l)}}^\top}{\sqrt{d}} + \alpha^{(l)} \cdot \tilde{C}(i,j) \right) \tag{5}$$

where $\alpha^{(l)}$ is a learnable per-layer scalar controlling how much each layer relies on the structural prior.

Noise curriculum. We schedule $\epsilon(t)$ to increase over training:

$$\epsilon(t) = \begin{cases} \epsilon_{\min}, & t < 0.2T \\ \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot \frac{t-0.2T}{0.6T}, & 0.2T \le t \le 0.8T \\ \epsilon_{\max}, & t > 0.8T \end{cases} \tag{6}$$

where T is the total training steps. Early training uses low noise for reliable structural guidance; late training increases noise, which forces the model to rely more on learned sequence patterns and less on the external structural signal. At inference, the bias can be zeroed or populated with any folding tool's predictions.

3.3.3 Level 3: Structure Denoising Objective

Level 3 closes the denoising loop from Level 2: given the noisy contact map $\tilde{\mathbf{C}}$ injected as attention bias, the model is trained to recover the clean contact map $\mathbf{C}$, which provides pairwise supervision for base-pairing interactions.

Supervision is restricted to motif-masked positions for efficiency. Let $\mathcal{P} = \{(i,j) : i \in \mathcal{M}_{\text{motif}}, C(i,j) = 1\}$ be the positive set of true base pairs involving masked positions. For each positive pair $(i,j) \in \mathcal{P}$, we uniformly sample N_{neg} negatives $\hat{\mathcal{N}}(i) \subset \{k : C(i,k) = 0\}$. Pairing scores are computed via a bilinear function:

$$\hat{C}(i,j) = \sigma(\mathbf{h}_i^\top \mathbf{W}_{\text{denoise}}\, \mathbf{h}_j) \tag{7}$$

Because base pairing is symmetric ($C(i,j) = C(j,i)$), $\mathbf{W}_{\text{denoise}}$ is symmetrized as $(\mathbf{W}_{\text{denoise}} + \mathbf{W}_{\text{denoise}}^\top)/2$ before computing scores. The loss is:

$$\mathcal{L}_{\text{denoise}} = -\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[\log \hat{C}(i,j) + \frac{1}{N_{\text{neg}}} \sum_{k\in\hat{\mathcal{N}}(i)} \log(1 - \hat{C}(i,k)) \right] \tag{8}$$

Restricting supervision to mask-related pairs keeps cost at $O(|\mathcal{P}| \cdot N_{\text{neg}})$ rather than $O(L^2)$. Purely unpaired masked positions (e.g., hairpin loops) contribute no entries to $\mathcal{P}$ and are supervised only by $\mathcal{L}_{\text{MLM}}$; if $|\mathcal{P}| = 0$, $\mathcal{L}_{\text{denoise}}$ is set to zero. Together with the noisy attention bias from Level 2, this forms a closed denoising loop.

3.4 Pre-training Setup

Masking protocol. Both masking strategies use $p_{\text{mask}} = 30\%$. In each training step, each sequence is processed twice: one copy undergoes random single-token masking (standard BERT protocol [21]: 80% `[MASK]`, 10% random substitution, 10% unchanged) and the other undergoes structure motif masking (all selected positions replaced with `[MASK]`). The MLM losses from both copies are summed. The denoising loss weight λ is set to 0.1.

Optimizer and schedule. All models are trained with AdamW [22] (β_1=0.9, β_2=0.98, ϵ=10^{-6}, weight decay 10^{-5}) using a linear decay schedule with 6% warmup, decaying to 2% of the peak learning rate. Gradient norms are clipped at 1.0. Following ModernBERT [17], bias terms are disabled in all linear layers and normalization modules. Detailed hyperparameters are provided in Appendix D.

Two-phase context extension. The first phase trains on sequences up to 1,536 nucleotides (93.4% of all sequences retained in full), accounting for 90% of total compute. The second phase extends the context to 8,192 nucleotides by increasing the RoPE base frequency from 10,000 to 160,000, with shorter sequences interleaved to preserve short-context performance.

Infrastructure. Training uses 4 machines with 32 NVIDIA A100 80 GB GPUs in total, fully sharded data parallelism (FSDP), and activation checkpointing. Phase 1 runs for 10 epochs, and Phase 2 adds one context-extension epoch. Phase 1 peak learning rates range from 8×10^{-4} (20M) to 2×10^{-4} (1.6B); effective batch sizes are 2,048 for 20M/100M and 4,096 for 650M/1.6B, with Phase 2 settings in Appendix D. All models use BF16 mixed precision and Flash Attention [18].

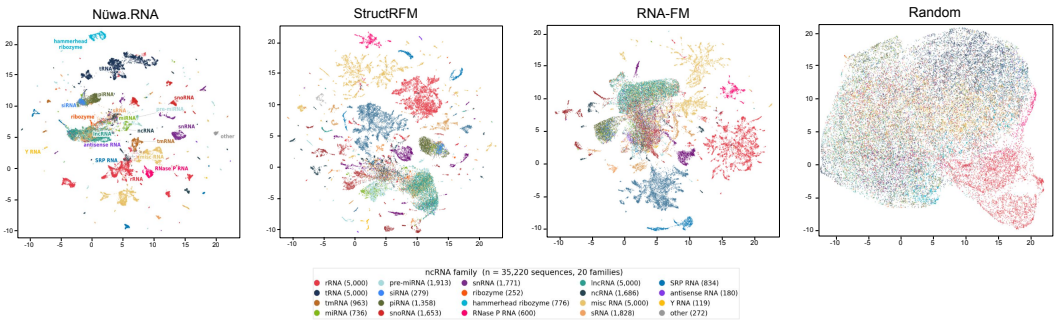


Figure 2: UMAP projections of mean-pooled last-layer representations from 35,220 ncRNA sequences across 20 families (RNAcentral). Nüwa.RNA yields the tightest clusters; StructRFM and RNA-FM show partial separation with more mixing; a randomly initialized model produces no structure.

4 Results

We first examine whether pre-training captures biologically meaningful representations, then quantify internalized structural knowledge via zero-shot secondary structure prediction. We benchmark on all 13 BEACON tasks [14], spanning structure prediction, gene regulation, and functional classification, against 18 baselines from BEACON and RNAGenesis [23]. Scaling experiments and ablation studies complete the evaluation. Unless otherwise noted, all reported results use Nüwa.RNA-1.6B; fine-tuning details are in Appendix F. For all downstream tasks, the structural attention bias is zeroed out at inference time, so the model operates from sequence input alone without requiring an external folding tool. This ensures a fair comparison against sequence-only baselines.

4.1 Representation Analysis

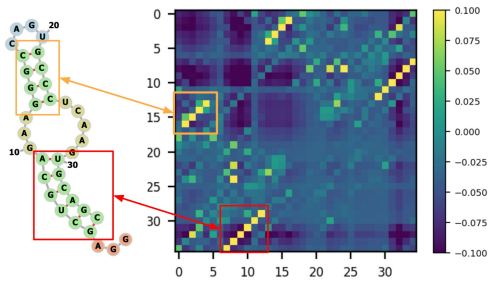
UMAP projections of 35,220 ncRNA sequences across 20 families (Figure 2) show that Nüwa.RNA produces well-separated clusters for distinct ncRNA families, with functionally related types projected into neighboring regions, whereas StructRFM and RNA-FM show partial separation with more overlap and a randomly initialized model yields no structure. At the nucleotide level, attention maps (Figure 3(a)) exhibit anti-diagonal patterns at stem base-pair positions, which means that attention heads learn to track canonical base pairings without task-specific supervision. Taken together, these results show that Nüwa.RNA encodes both family-level functional semantics and nucleotide-level structural interactions.

4.2 Zero-Shot Secondary Structure Prediction

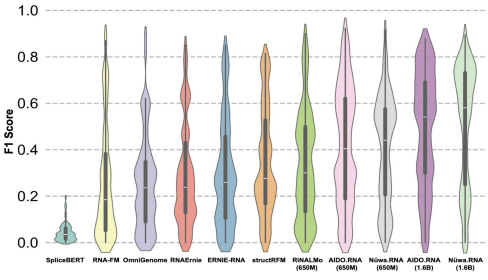
We measure internalized structural knowledge through zero-shot secondary structure prediction. Attention matrices are symmetrized, corrected via average product correction (APC), and thresholded into base-pairing predictions; the head and threshold are selected on 40 held-out PDB structures and applied unchanged to 70 test RNAs (full PDB identifiers in Appendix H). As shown in Figure 3(b), Nüwa.RNA achieves the highest median F1 among single-sequence RNA foundation models, outperforming AIDO.RNA [5] (1.6B) and structRFM [11], while mRNA-focused models (e.g., SpliceBERT) show negligible capability. On an NPSL2 stem-loop RNA containing bulge and hairpin motifs (PDB: 7UGA, 43 nt; Figure 4), fine-tuned Nüwa.RNA recovers all 16 base pairs (F1 = 1.0), and zero-shot mode achieves 0.83, approaching the supervised BPfold (0.87) while substantially outperforming ERNIE-RNA zero-shot (0.64).

4.3 Benchmark Evaluation

Table 1 reports results on all 13 BEACON tasks. Nüwa.RNA achieves the best performance on 10 of 13 tasks and remains competitive on the remaining three, with gains across structure prediction, gene regulation, and functional classification.



(a) Attention maps capturing stem structures. Colored boxes on the secondary structure diagram (left) highlight base-pairing interactions, with corresponding high-attention patterns visible at matching positions in the attention map (right).



(b) Zero-shot RNA secondary structure prediction via attention map probing. Each violin shows the distribution of per-sequence F1 scores on a test set of 70 experimentally resolved RNA 3D structures. Models are ordered by increasing median F1.

Figure 3: Structural knowledge encoded in pre-trained representations. (a) Attention heads learn to track canonical base-pairing interactions, producing anti-diagonal patterns aligned with stem regions. (b) Zero-shot secondary structure prediction from attention maps: Nüwa.RNA achieves the highest median F1 among single-sequence RNA foundation models.

Structure prediction. The largest gains appear on structure-related tasks: SSP +7.83, CMP +5.95, and DMP +4.56 over the strongest baseline, which shows that multi-level structural integration transfers to downstream structural evaluation. Comparing with RiNALMo [4] (650M) and AIDO.RNA [5] (1.6B), the two largest sequence-only RNA foundation models, Nüwa.RNA (81.2%) surpasses both AIDO.RNA (78.3%) and RiNALMo (74.7%), setting a new state of the art on SSP.

Gene regulation. Nüwa.RNA improves SSI (R^2 43.77% vs. 42.36%) and VDP (MCRMSE 0.284 vs. 0.291) over the strongest foundation model baseline. On SPL, where evaluation targets vertebrate pre-mRNA splicing, Nüwa.RNA (39.00%) outperforms all general-purpose ncRNA models; the gap to SpliceBERT-MS510 is consistent with the expected domain shift from ncRNA to pre-mRNA. On APA and MRL, Nüwa.RNA achieves competitive performance, suggesting that structure-aware representations generalize beyond the ncRNA training domain.

Functional classification. Nüwa.RNA achieves 99.40% accuracy on ncRNA classification and 97.34% AUC on modification site prediction, 1.58 and 2.18 points above the strongest baseline. CRISPR on-target and off-target efficacy show marginal gains (ρ 36.13% vs. 35.46% and 45.60% vs. 44.93%), consistent with the primary signal being gRNA–DNA complementarity rather than RNA structure.

4.4 Scaling Behavior

Figure 5 plots pre-training loss versus epochs for all four model sizes (20M, 100M, 650M, 1.6B). All variants converge smoothly, and each capacity increase yields a consistent reduction in final loss, suggesting that multi-level structural integration scales predictably (Appendix E).

4.5 Ablation Study

We ablate each component from the full Nüwa.RNA-650M model (chosen for computational efficiency; the same trends hold at other scales) on four representative tasks spanning structural, regulatory, and functional categories (Table 2). The study addresses three questions.

Q1 : **Does each level of structural integration contribute independently?**

Removing any single level degrades performance, with the attention bias (Level 2) being the most influential: its removal causes the largest single-component drop (SSP -4.9, MRL -3.4, ModSite -2.1). Dropping the denoising loss (Level 3) reduces all four metrics (SSP -2.2, ModSite -1.2). The no-structure baseline (SSP 72.1) falls 7.1 F1 points below the full model, which shows that the three levels are complementary.

Q2 : **Are dual-granularity masks complementary?**

Table 1: BEACON benchmark results [14]. VDP: MCRMSE (↓); all others higher-is-better. Baselines from BEACON and RNAGenesis [23]. **Bold**: best; underline: second best.

	SSP↑ F1%	CMP↑ P@L%	DMP↑ R^2%	SSI↑ R^2%	SPL↑ Acc%	APA↑ R^2%	NcRNA↑ Acc%
CNN	$49.95_{\pm 0.82}$	$43.89_{\pm 5.53}$	$27.76_{\pm 5.00}$	$34.36_{\pm 0.12}$	$8.43_{\pm 0.38}$	$50.93_{\pm 0.17}$	$88.62_{\pm 0.71}$
ResNet	$57.26_{\pm 3.14}$	$59.59_{\pm 0.68}$	$30.26_{\pm 1.81}$	$37.74_{\pm 0.16}$	$21.15_{\pm 1.56}$	$56.45_{\pm 0.94}$	$88.33_{\pm 1.22}$
LSTM	$58.61_{\pm 0.21}$	$40.41_{\pm 1.67}$	$44.77_{\pm 0.47}$	$35.44_{\pm 1.13}$	$36.66_{\pm 1.83}$	$67.03_{\pm 0.86}$	$88.78_{\pm 0.10}$
RNABERT [24]	$57.27_{\pm 0.30}$	$45.21_{\pm 10.87}$	$48.19_{\pm 0.64}$	$31.62_{\pm 0.64}$	$0.18_{\pm 0.18}$	$57.66_{\pm 2.11}$	$68.95_{\pm 7.29}$
RNA-FM [3]	$68.50_{\pm 0.54}$	$47.56_{\pm 6.73}$	$51.45_{\pm 0.51}$	$\underline{42.36}_{\pm 0.24}$	$34.84_{\pm 0.87}$	$70.32_{\pm 0.97}$	$96.81_{\pm 0.06}$
RNA-MSM [25]	$57.98_{\pm 0.47}$	$57.26_{\pm 15.38}$	$37.49_{\pm 4.10}$	$39.22_{\pm 0.23}$	$38.33_{\pm 0.76}$	$70.40_{\pm 1.12}$	$84.85_{\pm 0.27}$
SpliceBERT-H510 [7]	$64.93_{\pm 0.84}$	$45.80_{\pm 6.03}$	$55.56_{\pm 1.00}$	$38.91_{\pm 0.07}$	$44.80_{\pm 1.93}$	$58.65_{\pm 2.34}$	$95.92_{\pm 0.67}$
SpliceBERT-MS510 [7]	$43.24_{\pm 28.64}$	$52.64_{\pm 7.56}$	$10.27_{\pm 0.20}$	$38.58_{\pm 0.50}$	$\mathbf{50.55}_{\pm 0.49}$	$52.46_{\pm 17.36}$	$95.87_{\pm 0.36}$
SpliceBERT-MS1024 [7]	$68.26_{\pm 0.20}$	$47.32_{\pm 3.16}$	$55.89_{\pm 0.48}$	$39.22_{\pm 0.02}$	$\underline{48.52}_{\pm 0.49}$	$60.03_{\pm 3.42}$	$96.05_{\pm 0.78}$
UTR-LM-MRL [8]	$59.71_{\pm 0.30}$	$45.51_{\pm 23.51}$	$55.21_{\pm 2.91}$	$39.52_{\pm 0.36}$	$36.20_{\pm 1.84}$	$64.99_{\pm 4.90}$	$89.97_{\pm 0.62}$
UTR-LM-TE&EL [8]	$59.57_{\pm 0.20}$	$60.32_{\pm 7.27}$	$54.94_{\pm 2.54}$	$40.15_{\pm 0.11}$	$37.35_{\pm 5.48}$	$72.09_{\pm 0.82}$	$81.33_{\pm 8.55}$
UTRBERT-3mer [26]	$60.37_{\pm 0.47}$	$51.03_{\pm 21.48}$	$50.95_{\pm 0.44}$	$34.31_{\pm 0.00}$	$44.24_{\pm 0.53}$	$69.52_{\pm 4.56}$	$92.88_{\pm 0.38}$
UTRBERT-4mer [26]	$59.41_{\pm 0.45}$	$44.91_{\pm 27.56}$	$47.77_{\pm 2.08}$	$33.22_{\pm 0.00}$	$42.04_{\pm 0.53}$	$72.71_{\pm 0.85}$	$94.32_{\pm 0.95}$
UTRBERT-5mer [26]	$47.92_{\pm 8.75}$	$44.71_{\pm 7.64}$	$48.67_{\pm 1.70}$	$31.27_{\pm 0.00}$	$39.19_{\pm 0.37}$	$72.70_{\pm 1.77}$	$93.04_{\pm 0.37}$
UTRBERT-6mer [26]	$38.56_{\pm 28.76}$	$51.56_{\pm 20.30}$	$50.02_{\pm 1.05}$	$29.93_{\pm 0.17}$	$38.58_{\pm 2.72}$	$71.17_{\pm 2.30}$	$93.12_{\pm 0.17}$
BEACON-B [14]	$64.18_{\pm 0.44}$	$60.81_{\pm 1.70}$	$56.28_{\pm 0.41}$	$38.78_{\pm 0.18}$	$37.43_{\pm 1.43}$	$70.59_{\pm 0.91}$	$94.63_{\pm 0.16}$
BEACON-B512 [14]	$58.75_{\pm 3.72}$	$61.20_{\pm 2.11}$	$56.82_{\pm 0.63}$	$39.13_{\pm 0.08}$	$37.24_{\pm 1.09}$	$72.00_{\pm 0.17}$	$94.99_{\pm 0.21}$
RNAGenesis [23]	$\underline{73.37}_{\pm 0.25}$	$\underline{72.10}_{\pm 0.23}$	$\underline{70.77}_{\pm 0.33}$	$40.75_{\pm 0.08}$	$38.14_{\pm 1.39}$	$\mathbf{89.03}_{\pm 0.02}$	$\underline{97.82}_{\pm 0.02}$
Nüwa.RNA	$\mathbf{81.20}_{\pm 0.28}$	$\mathbf{78.05}_{\pm 0.37}$	$\mathbf{75.33}_{\pm 0.25}$	$\mathbf{43.77}_{\pm 0.15}$	$39.00_{\pm 0.55}$	$\underline{87.32}_{\pm 0.28}$	$\mathbf{99.40}_{\pm 0.05}$

	Modif↑ AUC%	MRL↑ R^2%	VDP↓ MCRMSE	PRS↑ R^2%	CRI-On↑ ρ%	CRI-Off↑ ρ%
CNN	$70.87_{\pm 0.40}$	$74.13_{\pm 0.58}$	$0.361_{\pm 0.003}$	$45.40_{\pm 0.66}$	$29.69_{\pm 2.52}$	$11.40_{\pm 0.10}$
ResNet	$71.03_{\pm 0.32}$	$74.34_{\pm 0.22}$	$0.349_{\pm 0.003}$	$55.21_{\pm 0.28}$	$28.55_{\pm 2.42}$	$11.50_{\pm 0.22}$
LSTM	$94.83_{\pm 0.31}$	$83.94_{\pm 0.08}$	$0.329_{\pm 0.002}$	$55.45_{\pm 0.71}$	$26.83_{\pm 1.32}$	$8.60_{\pm 0.13}$
RNABERT [24]	$82.82_{\pm 19.09}$	$29.79_{\pm 20.15}$	$0.378_{\pm 0.003}$	$54.60_{\pm 0.23}$	$29.77_{\pm 3.98}$	$4.27_{\pm 1.05}$
RNA-FM [3]	$94.98_{\pm 0.04}$	$79.47_{\pm 0.47}$	$0.347_{\pm 0.003}$	$55.98_{\pm 0.09}$	$31.62_{\pm 1.16}$	$2.49_{\pm 1.56}$
RNA-MSM [25]	$94.89_{\pm 0.18}$	$83.48_{\pm 0.18}$	$0.330_{\pm 0.001}$	$56.94_{\pm 0.38}$	$34.92_{\pm 1.99}$	$3.85_{\pm 0.99}$
SpliceBERT-H510 [7]	$62.57_{\pm 1.92}$	$83.49_{\pm 0.47}$	$0.321_{\pm 0.000}$	$54.90_{\pm 3.45}$	$26.61_{\pm 1.30}$	$4.00_{\pm 1.13}$
SpliceBERT-MS510 [7]	$55.87_{\pm 5.25}$	$\underline{84.98}_{\pm 0.28}$	$0.315_{\pm 0.003}$	$50.98_{\pm 7.46}$	$27.13_{\pm 0.27}$	$3.49_{\pm 2.12}$
SpliceBERT-MS1024 [7]	$53.45_{\pm 6.25}$	$67.15_{\pm 30.54}$	$0.313_{\pm 0.000}$	$57.72_{\pm 0.45}$	$27.59_{\pm 4.61}$	$5.00_{\pm 0.71}$
UTR-LM-MRL [8]	$56.41_{\pm 2.90}$	$77.78_{\pm 6.03}$	$0.325_{\pm 0.002}$	$57.28_{\pm 0.10}$	$28.49_{\pm 1.37}$	$4.28_{\pm 0.15}$
UTR-LM-TE&EL [8]	$59.70_{\pm 10.52}$	$82.50_{\pm 1.45}$	$0.319_{\pm 0.001}$	$53.37_{\pm 3.54}$	$32.49_{\pm 4.14}$	$2.91_{\pm 1.18}$
UTRBERT-3mer [26]	$95.14_{\pm 0.11}$	$83.89_{\pm 0.13}$	$0.337_{\pm 0.002}$	$56.83_{\pm 0.26}$	$29.92_{\pm 1.95}$	$4.48_{\pm 1.12}$
UTRBERT-4mer [26]	$95.10_{\pm 0.12}$	$82.90_{\pm 0.75}$	$0.341_{\pm 0.002}$	$56.43_{\pm 0.67}$	$23.20_{\pm 1.10}$	$3.11_{\pm 1.10}$
UTRBERT-5mer [26]	$94.78_{\pm 0.07}$	$75.64_{\pm 4.70}$	$0.343_{\pm 0.001}$	$57.16_{\pm 0.08}$	$25.74_{\pm 0.00}$	$3.93_{\pm 0.24}$
UTRBERT-6mer [26]	$95.08_{\pm 0.17}$	$83.60_{\pm 0.39}$	$0.340_{\pm 0.001}$	$57.14_{\pm 0.12}$	$28.60_{\pm 1.55}$	$4.90_{\pm 0.57}$
BEACON-B [14]	$94.74_{\pm 0.20}$	$72.29_{\pm 0.28}$	$0.320_{\pm 0.001}$	$54.67_{\pm 0.36}$	$26.01_{\pm 1.81}$	$4.42_{\pm 0.33}$
BEACON-B512 [14]	$94.92_{\pm 0.07}$	$72.35_{\pm 0.28}$	$0.320_{\pm 0.001}$	$55.20_{\pm 0.26}$	$28.17_{\pm 1.81}$	$3.82_{\pm 1.04}$
RNAGenesis [23]	$\underline{95.16}_{\pm 0.10}$	$\mathbf{85.83}_{\pm 0.05}$	$\underline{0.291}_{\pm 0.001}$	$\underline{58.19}_{\pm 0.11}$	$\underline{35.46}_{\pm 1.23}$	$\underline{44.93}_{\pm 3.77}$
Nüwa.RNA	$\mathbf{97.34}_{\pm 0.06}$	$83.50_{\pm 0.18}$	$\mathbf{0.284}_{\pm 0.002}$	$\mathbf{58.22}_{\pm 0.10}$	$\mathbf{36.13}_{\pm 0.78}$	$\mathbf{45.60}_{\pm 1.15}$

Removing motif masking drops SSP by 2.7 points, while removing random masking instead degrades ncRNA classification (-1.5) and MRL (-2.1), which shows that the two granularities capture distinct information. Replacing motif masking with structure-agnostic span masking [20] also reduces SSP, which means structure-aware boundaries provide information beyond contiguous spans.

Q3 : **Does noise injection and curriculum scheduling matter?**

Using clean contact maps without noise underperforms the noisy variant (SSP 76.8 vs. 79.2), and a fixed noise rate also underperforms the curriculum schedule (78.3 vs. 79.2), so both deliberate corruption and gradual noise escalation contribute to robustness.

Table 2: Ablation study on Nüwa.RNA-650M. Red indicates the full model. Orange rows correspond to Q1, blue to Q2, and green to Q3. Best results in **bold**.

Motif mask	Random mask	Attn bias	Noise curric.	Denoise loss	SSP (F1)	ncRNA (Acc)	MRL (R^2)	ModSite (AUC)
✓	✓	✓	✓	✓	**79.2**	**99.1**	**83.5**	**97.3**
	✓	✓	✓	✓	76.5	98.6	82.7	96.7
✓		✓	✓	✓	78.0	97.6	81.4	95.9
✓	✓			✓	74.3	97.8	80.1	95.2
✓	✓	✓	✓		77.0	98.4	82.2	96.1
	✓				72.1	97.0	78.6	94.5
span	✓	✓	✓	✓	77.3	98.8	83.1	97.0
✓	✓	✓		✓	76.8	98.5	81.9	96.2
✓	✓	✓	fixed	✓	78.3	98.9	82.8	96.8

5 Conclusion

We presented Nüwa.RNA, a foundation model that integrates secondary structure at three complementary pre-training stages: mask construction, model architecture, and loss supervision. By injecting structure exclusively through contact-map-derived attention biases and a denoising loss, the framework encodes robust structural representations without contaminating the sequence embedding space. Nüwa.RNA achieves the best performance on 10 of 13 BEACON tasks, and ablation experiments confirm that the three levels are complementary.

Limitations and future work. The framework relies on computationally predicted secondary structures as pseudo-labels; incorporating experimentally determined structures from chemical probing could further improve pre-training quality. Although the model operates without structural input at inference time by default (zeroing the attention bias), supplying BPfold-predicted contact maps can further improve performance at the cost of an additional folding preprocessing step. The corpus consists primarily of non-coding RNA, which may limit gains on mRNA-related tasks. Extending the framework to tertiary structure and RNA–protein interactions is a natural next step.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 82394432 and 92249302), and the Shanghai Municipal Science and Technology Major Project (Grant No. 2023SHZDZX02). This research was supported by AI for Science Program, Shanghai Municipal Commission of Economy and Information (No. 2025-GZL-RGZN-BTBX-01008). This research was supported by the NovaInspire platform of Shanghai Academy of Artificial Intelligence for Science. The computations in this research were performed using the CFFF platform of Fudan University. The authors declare no competing interests.

References

[1] Tim R Mercer, Marcel E Dinger, and John S Mattick. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009.

[2] Yiran Zhu, Liyuan Zhu, Xian Wang, et al. RNA-based therapeutics: an overview and prospectus. *Cell Death & Disease*, 13(7):644, 2022.

[3] Jiayang Chen, Zhihang Hu, Siqi Sun, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

[4] Rafael Josip Penić, Tin Vladušić, Zoran Bobić, et al. RiNALMo: General-purpose RNA language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043*, 2024.

[5] Shuxian Zou, Tianhua Tao, Sazan Mahbub, et al. A large-scale foundation model for RNA function and structure prediction. *bioRxiv*, 2024.

[6] Haopeng Yu et al. An interpretable RNA foundation model for exploring functional RNA motifs in plants. *Nature Machine Intelligence*, 6:1616–1625, 2024.

[7] Ken Chen et al. Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. *Briefings in Bioinformatics*, 25:bbae163, 2024.

[8] Yuting Chu et al. A 5' UTR language model for decoding untranslated regions of mRNA and function predictions. *Nature Machine Intelligence*, 6:449–460, 2024.

[9] Carlos Outeiral and Charlotte M Deane. Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence*, 6:170–179, 2024.

[10] Weijie Yin, Zhaoyu Zhang, Shuo Zhang, et al. ERNIE-RNA: an RNA language model with structure-enhanced representations. *Nature Communications*, 16:6246, 2025.

[11] Heqin Zhu, Ruifeng Li, et al. A fully open structure-guided RNA foundation model for robust structural and functional inference. *Research Square (preprint)*, 2024.

[12] Heng-cai Zhang, Ke Chen, Yunxiang Liu, Yifan Gong, Jian Hou, et al. OmniGenome: Aligning RNA sequences with secondary structures in genomic foundation models. *arXiv preprint arXiv:2407.11242*, 2024.

[13] Ning Wang, Jiang Bian, Yuchen Li, et al. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 2024.

[14] Yuchen Ren, Zhiyuan Chen, Lifeng Qiao, et al. BEACON: Benchmark for comprehensive RNA tasks and language models. In *Advances in Neural Information Processing Systems*, 2024.

[15] Heqin Zhu, Peng Xiong, et al. Deep generalizable prediction of RNA secondary structure via base pair motif energy. *Nature Communications*, 16:5856, 2025.

[16] Bethany A Sweeney, Anton I Petrov, Boris Burkov, et al. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49(D1):D212–D220, 2021.

[17] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[18] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pages 16344–16359, 2022.

[19] Jianlin Su, Yu Lu, Shengfeng Pan, et al. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[20] Mandar Joshi, Danqi Chen, Yinhan Liu, et al. SpanBERT: Improving pre-training by representing and predicting spans. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 64–77, 2020.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[23] Zaixi Zhang, Ruofan Jin, Linlin Chao, et al. RNAGenesis: A generalist foundation model for functional RNA therapeutics. *bioRxiv*, 2024.

[24] Masato Akiyama and Yasubumi Sakakibara. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genomics and Bioinformatics*, 4(1):lqac012, 2022.

[25] Yikun Zhang, Mei Lang, Jiuhong Jiang, et al. Multiple sequence-alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3, 2024.

[26] Yuning Yang, Gen Li, Kexin Qian, et al. Deciphering 3' UTR mediated gene regulation using interpretable deep representation learning. *bioRxiv*, 2023.

[27] Jaswinder Singh, Jack Hanson, Kuldip Paliwal, et al. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, 10(1):5407, 2019.

[28] Marcell Szikszai, Michael Wise, Amitava Datta, et al. Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. *Bioinformatics*, 38(16):3892–3899, 2022.

[29] David H Mathews. How to benchmark RNA secondary structure prediction accuracy. *Methods*, 162:60–67, 2019.

[30] Paul J Sample, Ban Wang, David W Reid, et al. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nature Biotechnology*, 37(7):803–809, 2019.

[31] Nicolas Scalzitti, Arnaud Kress, Romain Orhand, et al. Spliceator: multi-species splice site prediction using convolutional neural networks. *BMC Bioinformatics*, 22(1):561, 2021.

[32] Teresa Maria Rosaria Noviello, Francesco Ceccarelli, Michele Ceccarelli, et al. Deep learning predicts short non-coding RNA functions from only raw sequence data. *PLoS Computational Biology*, 16(11):e1008415, 2020.

[33] Zitao Song, Daiyun Huang, Bowen Song, et al. Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. *Nature Communications*, 12(1):4011, 2021.

A Structure Element Parsing

Given a binary contact map $\mathbf{C} \in \{0,1\}^{L\times L}$, we parse six types of structural elements by tracing paired and unpaired regions:

- **Stems**: consecutive stacked base pairs, i.e., maximal runs of positions $(i,j),(i{+}1,j{-}1),\ldots$ where all entries in $\mathbf{C}$ are 1. Both the $5'$ and $3'$ strands of each stem are grouped together.
- **Hairpin loops**: contiguous unpaired positions enclosed at the tip of a stem, connecting the two strands of the innermost base pair.
- **Internal loops**: unpaired positions flanked by exactly two stems, appearing on both strands between two consecutive base pairs in the nested structure.
- **Bulges**: a special case of internal loops where unpaired positions appear on only one strand.
- **Multi-branch junctions**: unpaired regions connecting three or more stems, forming the branching points of the secondary structure.
- **External loops**: single-stranded regions at the $5'$ and $3'$ termini and unpaired segments connecting independent sub-structures that are not enclosed by a higher-level stem.

These six element types partition all positions of a structured RNA into non-overlapping units, each treated as an indivisible masking unit during structure motif masking (Section 3.3.1 of the main paper).

B Secondary Structure Prediction Visualization

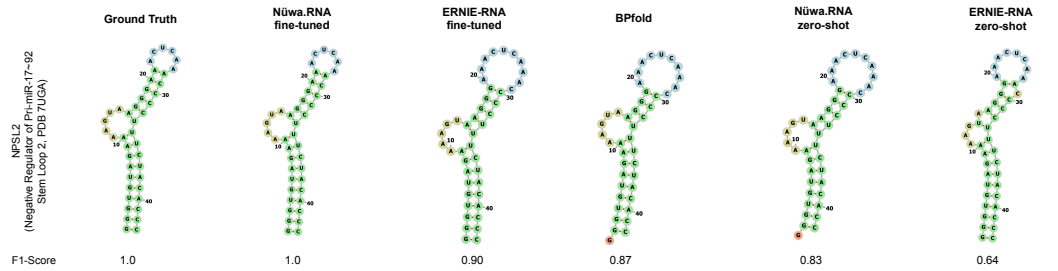


Figure 4: Secondary structure predictions on an NPSL2 stem-loop RNA (PDB: 7UGA, 43 nt) containing bulge and hairpin motifs. Ground truth (left) is compared with fine-tuned Nüwa.RNA (F1 = 1.0), fine-tuned ERNIE-RNA (0.90), BPfold (0.87), zero-shot Nüwa.RNA (0.83), and zero-shot ERNIE-RNA (0.64). Binary F1 scores are reported below each structure.

C Model Configurations

Table 3 summarizes the architectural hyperparameters of the four Nüwa.RNA variants. All models share the same architectural components (hybrid sliding-window/global attention, RoPE, noisy structural attention bias) and differ only in the number of layers and hidden dimension.

D Training Hyperparameters

Table 4 lists the pre-training hyperparameters for each model size across both training phases. All models use the same optimizer configuration and learning rate schedule; only the peak learning rate and batch size vary with model capacity.

E Scaling Behavior

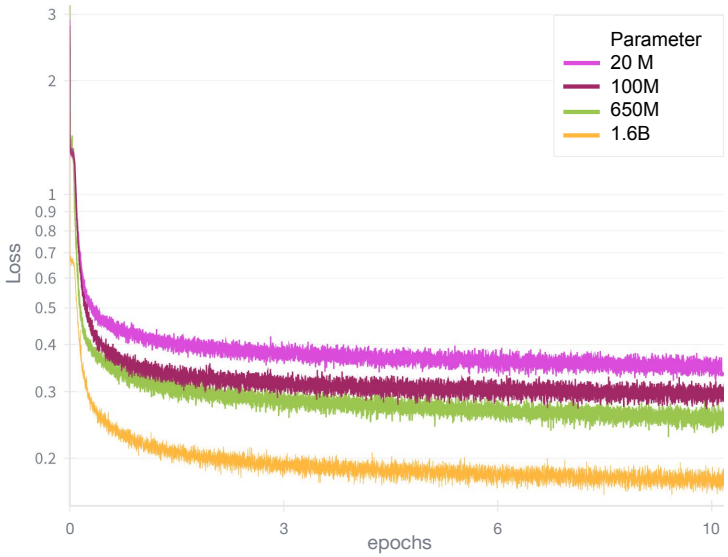
Figure 5 shows the Phase 1 pre-training loss trajectories for all four Nüwa.RNA model sizes. Three points stand out. First, all models converge smoothly with no loss spikes or instability, which shows

Table 3: Model configurations for the Nüwa.RNA family.

	20M	**100M**	**650M**	**1.6B**
Layers	12	12	33	32
Hidden dim	256	768	1,280	2,048
Attention heads	8	12	20	32
FFN dim	1,024	3,072	5,120	8,192
Global attn interval	3	3	3	3
Sliding window size	128	128	128	128
Max sequence length	8,192			
Pre-training length	1,536			

Table 4: Pre-training hyperparameters for the Nüwa.RNA family.

	20M	**100M**	**650M**	**1.6B**
Phase 1: 1,536 nt context				
Peak learning rate	8×10^{-4}	5×10^{-4}	3×10^{-4}	2×10^{-4}
Batch size (sequences)	2,048	2,048	4,096	4,096
Epochs	10 (90% of total compute)			
Phase 2: 8,192 nt context extension				
Peak learning rate	3×10^{-4}	2×10^{-4}	1×10^{-4}	5×10^{-5}
Batch size (sequences)	256	256	512	512
Epochs	1 (10% of total compute)			
RoPE θ (global attn)	160,000			
Shared settings				
Optimizer	AdamW (β_1=0.9, β_2=0.98, ϵ=10^{-6})			
Weight decay	10^{-5} (fully decoupled)			
LR schedule	Linear decay (6% warmup, decay to 2% of peak)			
Precision	BF16 mixed precision			
λ (denoising weight)	0.1			
p_{mask} (random masking)	30%			
p_{motif} (motif masking)	∼30%			
$\epsilon_{\min}$ / $\epsilon_{\max}$ (noise)	0.05 / 0.30			
N_{neg} (denoising negatives)	16			



Pretraining loss across various model sizes in the first phase.

Figure 5: Pre-training loss across model sizes during Phase 1. All four Nüwa.RNA variants (20M, 100M, 650M, 1.6B) show stable convergence over 10 epochs, with larger models achieving consistently lower final loss.

that the three-level structural integration (dual-granularity masking, noisy attention bias, and structure denoising) does not interfere with standard Transformer optimization dynamics. Second, the gap between adjacent model sizes remains roughly uniform on a log scale: scaling from 20M to 100M, 100M to 650M, and 650M to 1.6B each yields a comparable reduction in final loss (approximately $0.37 \rightarrow 0.31 \rightarrow 0.25 \rightarrow 0.16$). This regular spacing suggests that the structural pre-training objectives remain effective at larger scales. Third, the loss curves stay smooth throughout the 10-epoch Phase 1 run, which supports the optimization stability of the combined masking and denoising objectives.

F Downstream Fine-tuning Details

For BEACON benchmark tasks (Table 1), we fine-tune the full model using the AdamW optimizer with a cosine learning rate schedule. Learning rates are selected from $\{1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$ via validation performance. Sequence-level tasks use mean-pooled encoder outputs fed into a task-specific linear head, while nucleotide-level and pair-level tasks use per-position or pairwise outputs, respectively. All experiments are conducted with three random seeds and we report the mean $\pm$ standard deviation.

For the extended comparison with AIDO.RNA (Appendix G), we adopt task-specific fine-tuning protocols that vary in architecture (linear head vs. multi-layer MLP), tuning strategy (full fine-tuning vs. LoRA), and learning rate, as detailed in Appendix G, Section G.2.

G Comparison with AIDO.RNA on Downstream Tasks

To provide a direct comparison with AIDO.RNA [5], a 1.6B-parameter encoder-only RNA foundation model pre-trained on 42 million ncRNA sequences from RNAcentral, we evaluate Nüwa.RNA on all downstream fine-tuning tasks reported in Zou et al. [5]. This section describes the datasets, evaluation protocols, and per-dataset results.

G.1 Datasets

Secondary structure prediction. Two benchmark datasets are used: (1) bpRNA-TS0 [27], the test split from bpRNA-1m filtered by a 500-nucleotide length cutoff and 80% sequence similarity, containing 1,305 test sequences; (2) the filtered Archive-II dataset [28] with 3,865 RNAs from nine families, evaluated via 9-fold cross-validation with one family held out per fold. Following Mathews [29], we consider $(i \pm 1, j)$ and $(i, j \pm 1)$ pairings as correct predictions for a true base pair (i, j).

mRNA translation efficiency and expression level. We use the UTR-LM datasets [8] with sequences from three cell lines: human muscle tissue (Muscle; 1,257 samples), human prostate cancer cell line PC3 (12,579 samples), and human embryonic kidney 293T (HEK; 14,410 samples). Each sample includes a $5'$ UTR sequence paired with a translation efficiency label (Ribo-seq RPKM divided by RNA-seq RPKM) and an expression level label (RNA-seq RPKM). Both labels are in natural logarithm space. Evaluation uses 10-fold cross-validation with Spearman correlation.

Mean ribosome load. The dataset from Sample et al. [30] contains a large-scale synthetic human $5'$ UTR library of 83,919 sequences spanning 75 distinct lengths. The test set (Human7600) consists of 7,600 authentic human $5'$ UTRs. Evaluation uses R^2 score.

Transcript abundance. We use the CaLM datasets [9] containing coding sequences from seven organisms: A. thaliana, D. melanogaster, E. coli, H. sapiens, S. cerevisiae, P. pastoris, and H. volcanii. Labels are the natural logarithm of transcript counts per million. Evaluation uses 5-fold cross-validation with Pearson correlation.

Protein abundance. Five organism datasets from CaLM [9] (A. thaliana, D. melanogaster, E. coli, H. sapiens, S. cerevisiae) with protein copy-per-cell counts from PAXdb. Evaluation follows the same 5-fold cross-validation protocol with Pearson correlation.

Splice site prediction. The Spliceator dataset [31] contains confirmed error-free splice sites from over 100 eukaryotic species. Each sample is a 400-nucleotide sequence centered on a splice or non-splice site. Training uses the GS-1 subset; testing uses four unseen species (zebrafish, fly, worm, plant) with separate acceptor and donor datasets. Evaluation uses F1 score.

ncRNA family classification. The Rfam-novel dataset [32] contains 105,864 training, 17,324 validation, and 25,342 test ncRNA sequences across 88 Rfam families. Evaluation uses classification accuracy.

Modification site prediction. The MultiRM dataset [33] covers 12 types of RNA modifications (Am, Cm, Gm, Tm, m^1A, m^5C, m^5U, m^6A, m^6Am, m^7G, Ψ, I) with 304,661 training and 1,200 test samples. Each sample is a 1,001-nucleotide sequence. Evaluation uses AUROC averaged across modification types.

G.2 Fine-tuning Protocols

Secondary structure prediction. Pair representations are generated by outer concatenation of encoder outputs. The prediction head consists of a linear projection to dimension 64, two ResNet-2D blocks with 64 kernels of size 3 and instance normalization, and a final convolution layer. Binary cross-entropy loss is computed on elements above the main diagonal. Training uses a gradual unfreezing strategy: the prediction head is trained alone for the first three epochs, then three additional encoder layers are unfrozen every three epochs, over a total of 60 epochs. The learning rate decreases from 10^{-4} to 10^{-5}. A greedy post-processing algorithm converts predicted base-pairing probabilities into secondary structures by selecting the highest-probability pairs and excluding conflicts and sharp hairpin loops ($|i - j| < 4$).

Translation efficiency and expression level. Mean-pooled encoder outputs are fed into a two-layer MLP with hidden dimension 512. We fully fine-tune the model with peak learning rate 1×10^{-5}, weight decay 0.01, and dropout 0.1. Batch size is 8 for Muscle and 32 for PC3 and HEK. Training runs for 30 epochs with MSE loss.

Mean ribosome load. The prediction head comprises a linear projection to 256 dimensions followed by nine ResNet-1D blocks (256 kernels of size 3, instance normalization, ELU activation). The model is fine-tuned for 60 epochs with the prediction head trained alone for the first 3 epochs, using a learning rate decaying from 10^{-4} to 10^{-5}. Batch size is 64.

Transcript and protein abundance. Mean-pooled encoder outputs are fed into a two-layer MLP with hidden dimension 512. We use LoRA fine-tuning with rank 32 and $\alpha = 64$. Sequences are truncated to 1,024 nucleotides. Training runs for 15 epochs with MSE loss, peak learning rate 3×10^{-4}, weight decay 0.01, batch size 16, and dropout 0.1. For protein abundance, labels are transformed to $\log(1 + y)$ space.

Splice site prediction. The [CLS] embedding is fed into a three-layer MLP with hidden dimensions 512 and 128. We use LoRA fine-tuning with rank 32 and $\alpha = 64$. Training runs for 10 epochs with cross-entropy loss, peak learning rate 2.5×10^{-4}, and batch size 32.

ncRNA family classification. Mean-pooled encoder outputs are fed into a two-layer MLP with hidden dimension 512. We use LoRA fine-tuning with rank 16 and $\alpha = 32$. Training runs for 15 epochs with cross-entropy loss, peak learning rate 4×10^{-4}, and batch size 64.

Modification site prediction. The [CLS] embedding is fed into a two-layer MLP with hidden dimension 512. We use LoRA fine-tuning with rank 16 and $\alpha = 32$. Training runs for 10 epochs with binary cross-entropy loss, peak learning rate 4×10^{-4}, and batch size 64.

G.3 Per-Dataset Results

Per-dataset results are provided in Table 5.

Table 5: Full per-dataset results on all downstream tasks from Zou et al. [5]. Both models use 1.6B parameters. †Reproduced value. Best results in **bold**. Standard deviations are omitted to match the reporting format of Zou et al. [5].

Task	Dataset	Metric	AIDO.RNA	Nüwa.RNA
Secondary Structure	bpRNA	F1	0.783	**0.812**
	Archive-II	F1	0.733	**0.740**
mRNA Translation Eff.	Muscle	Spearman	0.730	0.730
	PC3	Spearman	0.820	**0.823**
	HEK	Spearman	0.680	0.680
mRNA Expression	Muscle	Spearman	**0.730**	0.720
	PC3	Spearman	0.680	**0.694**
	HEK	Spearman	0.720	0.720
Mean Ribosome Loading	Human 5′UTR	R^2	0.825†	**0.827**
Transcript Abundance	A. thaliana	Pearson	0.510	**0.543**
	D. melanogaster	Pearson	0.535	**0.567**
	E. coli	Pearson	0.632	**0.638**
	H. sapiens	Pearson	0.527	**0.551**
	S. cerevisiae	Pearson	0.656	**0.684**
	P. pastoris	Pearson	0.684	**0.702**
	H. volcanii	Pearson	**0.377**	0.336
Protein Abundance	A. thaliana	Pearson	0.644	**0.695**
	D. melanogaster	Pearson	0.688	**0.715**
	E. coli	Pearson	0.685	**0.706**
	H. sapiens	Pearson	0.560	**0.618**
	S. cerevisiae	Pearson	0.757	**0.784**
Splice Site	Acceptor	F1	0.940	**0.941**
	Donor	F1	0.960	**0.969**
ncRNA Classification	Rfam-novel	Acc	0.993	**0.995**
Modification Site	MultiRM	AUROC	0.971	**0.972**

H Zero-Shot Secondary Structure Prediction: Dataset Details

The zero-shot evaluation in Section 4.2 uses experimentally determined structures from the Protein Data Bank (PDB). The 40 held-out validation structures (sequence lengths 33–349 nt) are: 1KPZ_A, 1R7Z_A, 1RMN_A, 1U63_A, 1XJR_A, 1YMO_A, 1Z43_A, 2A43_A, 2GDI_A, 2K96_A, 2L3E_A, 2LHP_A, 2LK3_A, 2LUB_A, 2M8K_A, 2MXS_A, 2N6X_A, 2QUS_A, 3G78_A, 3OWZ_A, 3PHP_A, 3R1C_A, 4ENC_A, 4FRG_A, 4GMA_A, 4KQY_A, 4LGT_A, 4P8Z_A, 4QLM_A, 4TNA_D, 4V9F_BA, 4WFL_A, 5BJO_A, 5DH8_A, 5KH8_A, 5ML7_A, 6DME_A, 6SY6_A, 7EAF_A, 7MLX_R. The 70 test structures (sequence lengths 33–451 nt) are: 1I6U_C, 1X8W_A, 1XHP_A, 1Y0Q_A, 1YSV_A, 2DR8_A, 2F4X_A, 2FQN_A, 2G1W_A, 2GCS_A, 2GIS_A, 2HOJ_A, 2IL9_A, 2K95_A, 2KD8_A, 2KGP_A, 2KPC_A, 2KZL_A, 2L8F_A, 2LC8_A, 2LDL_A, 2LI4_A, 2LPA_A, 2LQZ_A, 2LX1_A, 2M12_A, 2M21_A, 2M24_A, 2M4W_A, 2M57_A, 2M8K_B, 2MEQ_A, 2MHI_A, 2MIS_A, 2MNC_A, 2MTJ_A, 2N3R_A, 2N4L_A, 2NBY_A, 2NC0_A, 2NOK_A, 2Y95_A, 3D0U_A, 3DIG_A, 3IWN_A, 3NDB_A, 3OFC_A, 3P22_A, 3PDR_A, 3Q3Z_A, 3SJ2_A, 3SKI_A, 3SUX_A, 4ENB_A, 4JF2_A, 4KQY_B, 4L81_A, 4LVZ_A, 4P95_A, 4PQV_A, 4QK8_A, 4R4V_A, 4RZD_A, 4TS0_A, 4TZX_A, 4V9F_B9, 5F9G_A, 5KH8_B, 7UGA_A, 7VNV_A; We further curate 15 additional structures as a TS-Hard set (sequence lengths 40–320 nt) to stress-test performance on complex topologies: 2D1B_A, 2L1F_B, 2LBJ_A, 2LBL_A, 2LDT_A, 2LPS_A, 2LU0_A, 2LV0_A, 2M5U_A, 3E5C_A, 3EP2_A, 3NJ6_A, 4A4T_A, 5X2G_A, 7UMC_A. The attention head and threshold are selected on the validation set and applied unchanged to the test set.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim state-of-the-art results on the majority of BEACON benchmark tasks; Table 1 confirms best performance on 10 of 13 tasks. All technical contributions (dual-granularity masking, noisy structural attention bias, structure denoising objective) are fully described in Section 3 and validated in Sections 4.3 and 4.5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses three limitations: reliance on computationally predicted pseudo-labels, ncRNA-biased pre-training corpus, and the restriction to secondary structure (with tertiary structure and RNA–protein interactions noted as future work).

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results; all contributions are empirical.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Model configurations are in Appendix C, pre-training hyperparameters in Appendix D, the noise curriculum in Section 3.3.2, masking protocol in Section 3.4, and fine-tuning details in Appendix F. The pre-training data source (RNAcentral) and structure prediction tool (BPfold) are publicly available.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and pre-trained model weights will be released upon publication. The pre-training data (RNAcentral) and benchmark (BEACON) are publicly available.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3.4 specifies the optimizer, learning rate schedule, masking protocol, and noise schedule. Appendix D provides per-model-size hyperparameters. Appendix F describes the fine-tuning protocol, including learning rate search range and number of random seeds.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: All downstream experiments are conducted with three random seeds and we report the mean (Appendix F). Standard deviations are reported alongside means in Table 1 for all 13 BEACON tasks. The primary claims rest on tasks with large margins (e.g., SSP +7.83, CMP +5.95), where the improvements substantially exceed the observed variance across seeds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.4 states that models are trained on 32 NVIDIA A100 80 GB GPUs across 4 machines with BF16 mixed precision and Flash Attention. Appendix D reports the number of training epochs for both phases. The total wall-clock training time for the 1.6B model is approximately 5 days.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available biological sequence data and does not involve human subjects, private data, or deceptive practices.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a foundational research contribution for RNA representation learning. The method improves general understanding of RNA biology with applications in drug discovery and therapeutics. We are not aware of direct negative societal impacts, as the model operates on non-coding RNA sequences and does not generate novel biological agents.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The model predicts RNA secondary structure and functional properties from existing sequences. It does not generate novel sequences or organisms, and poses no foreseeable misuse risk beyond standard bioinformatics tools.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: RNAcentral, BPfold, and BEACON are cited in Sections 3.1 and 4. These assets are publicly released under CC0 1.0, MIT, and Apache 2.0 licenses, respectively, and are used only as public research resources.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

Justification: Pre-trained model weights will be released with documentation upon publication but are not included with the submission. Model configurations and training details sufficient for reproduction are provided in the appendix.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methodology. Any LLM usage was limited to writing assistance and does not impact the scientific content.