# PARTNER-CONDITIONED RNA REPRESENTATIONS FOR FULL-SEQUENCE SNORNA–TARGET INTERACTION PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Predicting small nucleolar RNA (snoRNA) interactions requires identifying both the target RNA and its interaction site among many sequence-compatible candidates. Existing supervised snoRNA formulations largely score preselected local regions, while interaction predictors based on RNA language models (RNA-LMs) typically combine representations obtained independently for each RNA. We introduce **SnoBench**, an experimentally grounded benchmark that extends evaluation from controlled ranking of predefined candidate windows to target identification and site localization from complete target sequences. Built from transcriptome-wide RNA–RNA contact measurements, SnoBench contains 9,156 snoRNA–target pairs and 31,790 observed interaction regions, preserves variable-width annotations, keeps closely related targets within the same data split, and samples cross-target controls from a genome-wide target pool. We further develop a partner-conditioned interaction model that adapts frozen pretrained RNA representations to the candidate partner before interaction scoring. Bidirectional cross-attention allows the two RNA representations to inform one another, while a signed nucleotide-pair attention bias uses biologically motivated base-pairing information to guide cross-RNA attention. Granularity-matched prediction heads separately model target-level compatibility and site-level interaction evidence. Across representative thermodynamic, snoRNA-specific, and RNA-LM-based baselines, our model performs best in both evaluation settings. In full-sequence prediction with 100 candidate targets, it improves Target Recall@5 from $10.34\%$ to $15.72\%$ and Target MRR from 0.0911 to 0.1183 over the strongest baseline, while more than doubling Joint Recall@5, which requires correct target identification and site localization. SnoBench further reveals that strong discrimination among predefined local candidates does not necessarily transfer to complete sequences.

## 1 INTRODUCTION

Small nucleolar RNAs (snoRNAs) are short non-coding RNAs that regulate other RNA molecules by binding to specific sites within them. They have been increasingly implicated in diverse regulatory processes and human disease. Characterizing their interaction landscape is therefore important for understanding RNA regulation and disease-associated mechanisms. Predicting these interactions, however, is challenging. Candidate binding sites span the transcriptome, yet only a small fraction are observed in vivo. Sequence complementarity is informative but insufficient: functional snoRNA–target duplexes often contain wobble pairs, mismatches, and bulges, so matching criteria permissive enough to capture them also admit many competing sequence matches. Moreover, the accessibility and compatibility of a candidate site depend on the broader sequence and structural context of both RNAs. Accurate prediction therefore requires distinguishing biologically relevant interactions from sequence-compatible candidates using signals beyond local complementarity alone.

Recent advances in pretrained RNA language models (RNA-LMs) provide contextual nucleotide representations learned from large RNA sequence collections. These representations capture surrounding sequence information and structural regularities (Yin et al., 2025; Zhang et al., 2024). For interaction prediction, such representations offer a richer alternative to fixed local matching. However, RNA-LMs



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

are pretrained on individual sequences, whereas RNA–RNA interaction prediction requires modeling relationships between two molecules.

Two complementary gaps remain in bringing pretrained RNA representations to snoRNA interaction prediction. First, existing supervised snoRNA formulations, exemplified by snoGloBe (Deschamps-Francoeur et al., 2022), operate on short, preselected candidate regions. This differs from practical target discovery, where the interaction site is unknown and must be localized within complete target sequences. Although local scorers can be applied across a transcript by sliding windows, no standardized benchmark evaluates joint target identification and site localization in this setting. Second, in RNA–RNA interaction prediction, RNA-LM-based models such as RIME obtain pretrained representations for each RNA independently and combine them only downstream (Setti et al., 2026), leaving the pretrained representations themselves unaware of the interaction partner.

Recent transcriptome-wide RNA–RNA contact measurements provide experimental observations at sufficient scale to address these limitations (Liu et al., 2025). Leveraging these measurements, we introduce **SnoBench**, an experimentally grounded benchmark containing 9,156 snoRNA–target pairs and 31,790 observed interaction regions. In addition to **controlled candidate ranking** over preselected local windows, SnoBench introduces **full-sequence target identification and site localization**, requiring models to identify both the interacting target and its site from complete candidate target sequences. Together, the two settings separate local interaction discrimination from target identification and site localization over complete sequences.

We further develop a **partner-conditioned interaction model** that adapts pretrained RNA representations to the candidate partner before interaction scoring. Bidirectional cross-attention conditions each RNA representation on its candidate partner, while a signed nucleotide-pair attention bias uses biologically motivated base-pairing information to guide cross-RNA attention. Granularity-matched dual heads separately model target-level compatibility and site-level interaction evidence.

We compare against representative thermodynamic, snoRNA-specific, and RNA-LM-based predictors, with learned baselines retrained on the SnoBench training split. Our model leads in controlled candidate ranking, improving MRR from 0.3688 to 0.3918 over retrained RIME, the strongest baseline in both settings. More importantly, in full-sequence prediction with 100 candidate targets, it improves Target R@5 from $10.34\%$ to $15.72\%$ and Target MRR from 0.0911 to 0.1183, while more than doubling Joint R@5, which requires correct target identification and site localization. Results across the two settings further show that strong discrimination among supplied local candidates does not necessarily translate into full-sequence target identification and site localization: snoGloBe, the second-strongest baseline in controlled candidate ranking, performs close to random on full-sequence target-identification metrics.

Our contributions are:

- We introduce **SnoBench**, an experimentally grounded benchmark for controlled candidate ranking and full-sequence target identification with site localization. It preserves variable-width interaction regions, uses genome-wide cross-target controls, and enforces target-component-disjoint splits.
- We develop a **partner-conditioned interaction model** that combines bidirectional cross-attention, a learned signed nucleotide-pair attention bias, and granularity-matched dual heads to condition pretrained RNA representations on the candidate partner and model interaction evidence at both target and site levels.
- Our model outperforms all evaluated baselines in both settings, improving full-sequence Target R@5 from $10.34\%$ to $15.72\%$ over the strongest baseline and more than doubling Joint R@5. A matched-backbone comparison using the same frozen ERNIE-RNA encoder further supports the effectiveness of the proposed interaction architecture.

## 2 Related Work

**RNA–RNA interaction prediction.** Computational approaches to RNA–RNA interaction prediction span both physics-based and learned models. Thermodynamic methods such as RNAplex and IntaRNA score candidate duplexes by complementarity and stability, with IntaRNA also modeling site accessibility (Tafer & Hofacker, 2008; Mann et al., 2017). PLEXY and RNAsnoop add class-specific

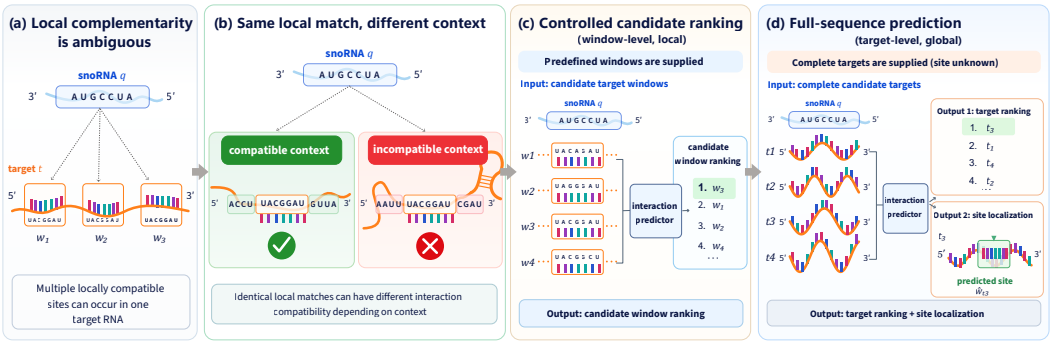


Figure 1: Overview of the interaction-prediction task. Local sequence complementarity may support several candidate regions, while sequence context and the candidate partner affect interaction compatibility. Full-sequence prediction requires identifying both the interacting target and interaction region.

constraints for box C/D and H/ACA snoRNA targets, respectively (Kehr et al., 2011; Tafer et al., 2010).

Recent learned approaches increasingly build on representations from pretrained nucleotide and RNA language models, including the Nucleotide Transformer (Dalla-Torre et al., 2025), RNA-FM (Chen et al., 2022), RiNALMo (Penić et al., 2025), and ERNIE-RNA (Yin et al., 2025). Representative interaction predictors based on these models obtain pretrained representations for the two RNAs independently before combining them downstream. RIME constructs pairwise interaction features from Nucleotide Transformer representations (Setti et al., 2026), while MiRInter-Trans combines sequence-level RNA-FM representations (Nicolini et al., 2026).

Cross-RNA conditioning has also been explored in miRNA target prediction, but with task-specific encoders rather than pretrained RNA language models. Mimosa uses cross-attention to condition target-site representations on the miRNA (Bi et al., 2024), while MiRformer similarly conditions dual sequence encoders through cross-attention (Gu et al., 2026). Our model combines pretrained RNA representations with cross-RNA conditioning. Inspired by ERNIE-RNA's use of nucleotide-pair information in single-sequence self-attention, we introduce a learned signed nucleotide-pair bias that guides attention between the two interacting RNAs.

**snoRNA interaction data and evaluation.** Existing supervised snoRNA predictors have been trained on relatively small labeled datasets. snoGloBe, for example, was developed using a collection of 594 positive interactions assembled from curated annotations and high-throughput interaction data (Deschamps-Francoeur et al., 2022). Its experimentally observed interactions are processed into RNAplex-derived duplexes, split at bulges, and represented as fixed-width local windows. As a result, contacts spanning bulges are broken into separate fragments, and the full extent and variable width of the original interaction regions are not preserved. The prediction task is thus defined over preselected local windows.

More recently, transcriptome-wide assays such as snoKARR-seq have mapped thousands of snoRNA–RNA contacts across diverse transcript targets, including interactions beyond canonical snoRNA modification sites (Liu et al., 2025). snoKARR-seq, however, is an experimental mapping resource rather than a benchmark for interaction prediction. SnoBench builds on these measurements, creating a benchmark with defined training examples, data splits, and evaluation tasks.

## 3 SnoBench: Benchmark Construction and Evaluation

### 3.1 Experimental Interaction Data

We build on the published snoKARR-seq supplementary interaction records (Liu et al., 2025). The source collection contains 50,902 reproducible observations from five human cell lines, with snoRNA and target interaction regions reported for each observation. These measurements capture transcriptome-wide snoRNA–RNA contacts, including interactions beyond canonical snoRNA-guided

modification sites, and provide experimental supervision for both target identification and interaction-site localization. We use the reported interaction coordinates directly rather than converting observed contacts into a fixed local alignment representation. Observations are resolved to fixed strand-oriented reference sequences and consolidated at the query–target sequence-pair level. Repeated observations with identical target coordinates are represented by a single observed region while retaining their provenance, whereas distinct regions remain separate even when overlapping or touching. The resulting benchmark contains 9,156 query–target pairs and 31,790 observed interaction regions, with a median region width of 46 nt and an interquartile range of 24–73 nt.

### 3.2 Benchmark Construction

**Target-component splits.** To reduce leakage between related targets, we group targets into components based on shared gene or locus identity, identical sequences, and sequence similarity or containment, and assign each component to a single split. This prevents closely related targets from crossing training and evaluation splits. Grouping thresholds and implementation details are provided in Appendix A.

**Controlled candidate lists.** For controlled candidate ranking, we construct fixed lists containing windows from both the observed target and alternative targets. Because experimental annotations are variable-width regions, we define a coverage-based match between a candidate interval $w$ and an observed region $r$ as

$$H_{\mathrm{cov}}(w, r) = \mathbf{1}[2|w \cap r| \geq \min(|r|, 2|w|)]\,. \tag{1}$$

When $w$ is at least half as wide as $r$, this requires $w$ to cover at least half of $r$; when $w$ is shorter than half of $r$, it must lie entirely within $r$.

One designated positive is selected from candidate windows satisfying equation 1. Every remaining candidate window in the fixed list must fail the coverage criterion against all observed regions for the corresponding query–target pair. Same-target controls are sampled from alternative windows on the observed target under this requirement.

Cross-target controls are sampled from the split-local eligible target universe after excluding the true target, known interaction partners for the query, exact target aliases or sequence duplicates, and related genes or target components. Candidate scales, list composition, positive selection, and sampling details are provided in Appendix A.

**Full-sequence target sets.** Distractor targets are drawn from the split-local eligible target universe and matched hierarchically on target-level covariates, including gene biotype, target length, scan count, GC content, and sequence complexity, to reduce reliance on target-level covariate shortcuts. Progressively relaxed fallback tiers are used when tightly matched distractors are unavailable; the procedure is detailed in Appendix A.

### 3.3 Prediction Tasks and Evaluation

Let $q$ denote a full snoRNA sequence, $t$ a candidate target sequence, and $\mathcal{G}_{q,t}$ the set of experimentally observed regions for the pair. We evaluate two complementary settings: controlled candidate ranking and full-sequence target identification with site localization. The controlled setting isolates discrimination among supplied local candidates, whereas the full-sequence setting more closely reflects practical interaction prediction, in which both the interacting target and its site are unknown and must be identified from complete sequences.

**Controlled candidate ranking.** In the controlled setting, a method is given a fixed candidate list $\mathcal{C} = \{(t_s, w_s)\}_{s=1}^{S}$, where each $(t_s, w_s)$ specifies a candidate target and predefined window for the same snoRNA query $q$. The method returns a ranking over $\mathcal{C}$, and we evaluate the rank of the designated positive using mean reciprocal rank (MRR) and Recall@$K$.

**Full-sequence prediction.** Given a query $q$ and complete candidate targets $\mathcal{T} = \{t_1, \ldots, t_N\}$ containing the annotated positive target $t^\star$ together with distractors, a method must identify both the interacting target and an interaction site within it. A method returns a target ranking $\pi_q$ and a top predicted region $\hat{w}_t$ for each target $t$. Candidate regions are generated and scored using each

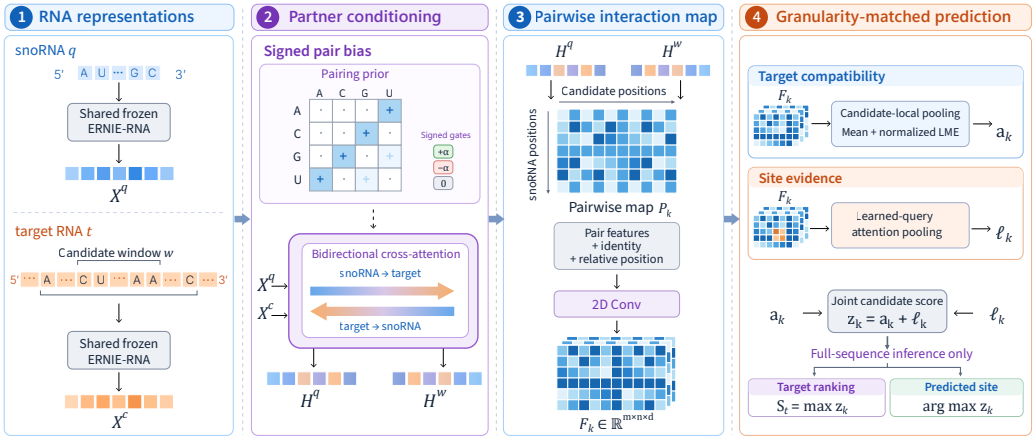


Figure 2: Architecture of the partner-conditioned interaction model. A frozen RNA encoder supplies nucleotide representations, which are conditioned through bidirectional cross-attention and summarized by pairwise interaction modeling and separate target and site heads.

method's native, gold-independent inference procedure; prediction widths, strides, and internal search representations may therefore differ across methods. Method-specific inference configurations are reported in Appendix C. Target identification is evaluated using MRR and Recall@$K$ of $t^\star$ in $\pi_q$.

**Localization and joint evaluation.** Localization is evaluated on the annotated positive target $t^\star$ independently of its position in the target ranking. Because methods may predict intervals of different widths, we report two complementary criteria. The center-based criterion is width-agnostic and evaluates whether the midpoint of the predicted interval falls within an observed region, while the coverage-based criterion measures whether the predicted interval sufficiently covers an observed region:

$$H_{\text{center}}(w, \mathcal{G}) = \mathbf{1}[\exists r \in \mathcal{G} : c(w) \in r], \qquad H_{\text{cov}}(w, \mathcal{G}) = \max_{r \in \mathcal{G}} H_{\text{cov}}(w, r), \tag{2}$$

where $c(w)$ denotes the midpoint of $w$, retained without rounding for even-width intervals. We report center-based and coverage-based localization Recall@1 for the top predicted region $\hat{w}_{t^\star}$.

Joint evaluation also requires successful target retrieval. For either criterion $H \in \{H_{\text{center}}, H_{\text{cov}}\}$, Joint Recall@$K$ counts an instance as correct when

$$\text{rank}_{\pi_q}(t^\star) \le K \quad \text{and} \quad H(\hat{w}_{t^\star}, \mathcal{G}_{q,t^\star}) = 1. \tag{3}$$

Additional localization analyses are provided in Appendix A.

## 4 Partner-Conditioned snoRNA Interaction Model

Our model conditions pretrained snoRNA and target representations on each other before scoring their interaction (Figure 2). A frozen ERNIE-RNA encoder provides contextual nucleotide representations, bidirectional cross-attention with a signed nucleotide-pair attention bias conditions each RNA on its candidate partner, and a convolutional module summarizes the resulting pairwise interaction map. Granularity-matched dual heads produce target-compatibility and site-localization evidence for each candidate window. Their combined candidate score is used directly for controlled ranking and reduced to a target score by taking the maximum over searched regions in full-sequence inference.

### 4.1 Contextual RNA Representations

For a complete snoRNA sequence $q$ and candidate window $w$ on target $t$, let $\tilde{t}_w$ denote the candidate-centered target context. A shared frozen ERNIE-RNA encoder $E$ (Yin et al., 2025) produces $X^q = E(q)$ and $X^c = E(\tilde{t}_w)$. The target context contains the candidate window together with

surrounding sequence on both sides. After encoding, learned segment embeddings distinguish the candidate window from its flanking context before partner conditioning; all snoRNA positions use the in-region embedding.

### 4.2 Bidirectional Partner Conditioning

We condition each RNA on its candidate partner using bidirectional cross-attention. Query nucleotides attend to the target context, while target-context nucleotides attend back to the snoRNA. For attention head $h$,

$$\mathrm{Attn}_h(Q_h, K_h, V_h) = \mathrm{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_h}} + \alpha_h B^{(h)} + M\right) V_h, \tag{4}$$

where $Q_h$ is computed from one RNA and $K_h, V_h$ from its partner, $d_h$ is the head dimension, and $M$ masks padded key positions. The two directions use distinct attention and feed-forward parameters.

The signed nucleotide-pair bias is defined by $B_{ij}^{(h)} = T_h[x_i, y_j]$ for snoRNA nucleotide $x_i$ and target nucleotide $y_j$, where $T_h$ is a trainable head-specific lookup table. Each table is initialized from the same pairing prior: Watson–Crick pairs receive value $3.0$, G–U and U–G wobble pairs receive $1.5$, and all remaining pairs receive $0$. The reverse attention direction uses the transpose of the corresponding bias matrix.

The signed gates $\alpha_h$ are initialized into positive, negative, and neutral groups, inducing different initial uses of the pairing prior across attention heads. Both the gates and pairwise lookup tables remain unconstrained and trainable. After bidirectional conditioning, we retain the partner-conditioned snoRNA representations $H^q$ and the central candidate-window slice $H^w$ from the target branch. Surrounding target context can therefore influence $H^w$, while subsequent interaction modeling remains candidate-local.

### 4.3 Pairwise Interaction Modeling

We construct a two-dimensional interaction map over snoRNA and candidate-window positions by combining separate learned projections with an element-wise product,

$$P_{ij} = \phi_q(H_i^q) \odot \phi_w(H_j^w). \tag{5}$$

Each spatial cell therefore represents features associated with a particular snoRNA–target nucleotide pair. The map is augmented with nucleotide-pair identity and relative-position information. Before convolution, the valid target axis is oriented antiparallel to the snoRNA, placing local interaction patterns in a consistent spatial orientation. A masked two-dimensional convolutional network then integrates neighboring pair features while excluding padded spatial positions.

Let $F_k$ denote the convolutional feature map for candidate $k$ and $\Omega_k$ its valid spatial cells. The target-compatibility branch combines channel-wise normalized log-mean-exp pooling with spatial mean pooling, whereas the site-localization branch uses learned-query attention pooling over $\Omega_k$. The resulting representations are

$$\begin{aligned} R_k^{\mathrm{target}} &= \psi([\mathrm{LME}_{\Omega_k}(F_k); \mathrm{Mean}_{\Omega_k}(F_k)]), \\ R_k^{\mathrm{site}} &= \psi(\mathrm{AttnPool}_{\Omega_k}(F_k)). \end{aligned} \tag{6}$$

Here $\mathrm{AttnPool}$ denotes learned-query pooling followed by the dimension-matching projection, so that both branches use the shared projection $\psi$. The normalized log-mean-exp definition is provided in the Appendix.

Separate prediction heads produce target and site logits, which are combined at the candidate level:

$$a_k = f_{\mathrm{target}}(R_k^{\mathrm{target}}), \qquad \ell_k = f_{\mathrm{site}}(R_k^{\mathrm{site}}), \qquad z_k = a_k + \ell_k. \tag{7}$$

Here $a_k$ represents window-conditioned target compatibility and $\ell_k$ represents site-level evidence for candidate window $k$.

### 4.4 Candidate-Local Target and Site Prediction

**Target supervision.** Consider a controlled training list containing $J$ distinct targets, with annotated positive target $t^\star$ and candidate set $I_t$ for target $t$. We use label-smoothed target labels $y_t =$

$(1 - \epsilon_{\text{target}})\mathbf{1}[t = t^\star] + \epsilon_{\text{target}}/J$, with $\epsilon_{\text{target}} = 0.05$, and apply supervision directly to the candidate-local target logits $a_k$:

$$\mathcal{L}_{\text{target}} = \frac{2}{J} \sum_t \frac{1}{|I_t|} \sum_{k \in I_t} \text{BCE}_\omega(a_k, y_t), \qquad \omega = J - 1. \tag{8}$$

Averaging first within each target ensures that targets contribute equally regardless of how many candidate windows they contain. The positive-class weight balances the positive target against the $J - 1$ negatives. The fixed factor 2 is a gradient-scale normalization; its derivation is given in Appendix B.

**Site and joint supervision.** Site supervision uses unsmoothed cross-entropy to distinguish the designated positive window from alternative windows on the annotated positive target. The combined candidate score $z_k$ is additionally supervised over the complete candidate list using label-smoothed cross-entropy with $\epsilon_{\text{joint}} = 0.05$. The complete objective is

$$\mathcal{L} = 0.25\,\mathcal{L}_{\text{target}} + 0.25\,\mathcal{L}_{\text{site}} + 0.50\,\mathcal{L}_{\text{joint}}. \tag{9}$$

Complete loss definitions and optimization details are provided in Appendix B.

### 4.5 Controlled and Full-Sequence Inference

For controlled candidate ranking, every candidate window is scored independently using $z_k$ from equation 7, and candidates are ranked directly by this score.

For full-sequence inference, candidate regions are generated across each complete target according to the declared search procedure (Appendix C) and scored using the same candidate-local function. Let $I_t$ denote the regions searched on target $t$. Its retrieval score is

$$S_t = \max_{k \in I_t} z_k, \tag{10}$$

and targets are ranked by $S_t$. For each target, its predicted interaction region is the candidate attaining this maximum. When the maximum is not unique, all tied regions are retained and localization is evaluated in expectation under uniform tie-breaking, as for other exact-score ties (Appendix A.4).

## 5 Experiments

### 5.1 Experimental Setup

**Evaluation settings.** Both settings use the same 986 held-out SnoBench query–positive-target pairs. Controlled ranking uses $S = 17$ candidate windows per pair: one designated positive and 16 controls, with six 21-nt, six 42-nt, and five 70-nt windows. Full-sequence evaluation uses $N = 100$ complete candidate targets per pair, comprising the annotated positive target and 99 distractors, with no candidate windows supplied. Within each setting, all methods share the same supplied lists. Candidate construction and evaluation metrics follow Section 3.

**Model and training.** We use a frozen ERNIE-RNA encoder selected by validation performance and one bidirectional partner-conditioning block with 12 attention heads. Pair-bias gates are initialized with four heads each at $\alpha_h \in \{+1, -1, 0\}$. The downstream modules are trained with a peak learning rate of $2 \times 10^{-5}$, weight decay of $10^{-3}$, 10% linear warmup, and cosine decay. Our model and retrained RIME are evaluated over three independent runs in controlled ranking, with checkpoints selected by validation MRR; reported $\pm$ values denote standard deviations across runs. Full-sequence evaluation, which scans all 100 complete candidate targets per pair, is run once per method because of its computational cost. Additional architectural and training settings are provided in Appendix B.

### 5.2 Baselines

For controlled ranking, we evaluate random and Watson–Crick complementarity controls, RNAplex (Tafer & Hofacker, 2008), IntaRNA (Mann et al., 2017), PLEXY (Kehr et al., 2011), and both released and SnoBench-retrained versions of snoGloBe (Deschamps-Francoeur et al., 2022) and RIME (Setti et al., 2026). Full-sequence evaluation includes IntaRNA, both snoGloBe variants, and retrained RIME, each using its native, gold-independent inference procedure.

Table 1: Held-out controlled candidate ranking on SnoBench test pairs. All methods are evaluated on the same 17-candidate lists. MRR and Recall@$K$ are micro-averaged over evaluation instances.

| Method | MRR | R@1 (%) | R@5 (%) |
|---|---|---|---|
| Random | 0.2023 | 5.88 | 29.41 |
| Watson–Crick | 0.2414 | 9.09 | 35.97 |
| RNAplex | 0.2183 | 6.95 | 32.45 |
| IntaRNA | 0.2577 | 9.23 | 38.64 |
| PLEXY | 0.2127 | 6.83 | 30.77 |
| snoGloBe (released) | 0.2588 | 9.03 | 40.94 |
| RIME (released) | 0.1798 | 3.75 | 25.56 |
| snoGloBe (retrained) | 0.3059 | 12.52 | 50.46 |
| RIME (retrained) | 0.3688 $\pm$ 0.0028 | 17.92 $\pm$ 0.21 | 59.70 $\pm$ 0.41 |
| **Ours** | **0.3918** $\pm$ 0.0028 | **19.73** $\pm$ 0.97 | **63.32** $\pm$ 0.94 |

Table 2: Full-sequence prediction on SnoBench pairs with 100 candidate targets each. Localization is evaluated on the positive target independently of its rank; joint metrics require both target identification and localization. Subscripts $\mathrm{ctr}$ and $\mathrm{cov}$ denote center- and coverage-based criteria. Recall is in %. For 100 candidate targets, random target ranking has expected MRR of $0.0519$. All displayed estimates are pair-micro. Ours and retrained RIME share the same search grid; IntaRNA and snoGloBe use their native site definitions.

| | Target identification | | | Site localization | | Joint target + site | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | R@1 | R@5 | MRR | R@1$_{\mathrm{ctr}}$ | R@1$_{\mathrm{cov}}$ | R@1$_{\mathrm{ctr}}$ | R@5$_{\mathrm{ctr}}$ | R@1$_{\mathrm{cov}}$ | R@5$_{\mathrm{cov}}$ |
| snoGloBe (released) | 1.040 | 5.327 | 0.0552 | 8.950 | 7.835 | 0.101 | 0.735 | 0.101 | 0.634 |
| snoGloBe (retrained) | 1.065 | 6.085 | 0.0575 | 8.722 | 8.012 | 0.220 | 0.811 | 0.220 | 0.811 |
| IntaRNA | 1.826 | 7.201 | 0.0667 | 9.229 | 8.925 | 0.203 | 0.811 | 0.203 | 0.913 |
| RIME (retrained) | 2.637 | 10.345 | 0.0911 | 11.968 | 10.041 | 1.217 | 2.738 | 1.217 | 2.535 |
| **Ours** | **3.245** | **15.720** | **0.1183** | **13.438** | **13.590** | **2.738** | **5.832** | **2.586** | **5.375** |

# 6 Results

## 6.1 Controlled Candidate Ranking

Table 1 reports controlled candidate ranking. Classical complementarity and energy-based baselines provide only modest gains over random. The released RIME checkpoint performs below the random reference, likely reflecting a mismatch between its original training distribution and SnoBench rather than a limitation of the architecture itself. Our model achieves the best performance across all three metrics, reaching an MRR of $0.3918 \pm 0.0028$ compared with $0.3688 \pm 0.0028$ for retrained RIME.

## 6.2 Full-Sequence Target Identification and Site Localization

**Controlled ranking does not directly predict full-sequence identification.** Retrained snoGloBe is well above random in controlled ranking but close to random in full-sequence target identification ($0.0575$ vs. $0.0519$); the released model shows the same pattern. This is consistent with its local window-scoring formulation: discriminating among supplied candidate regions does not necessarily yield reliable target rankings when sites must first be discovered across complete sequences.

**Our model improves both target identification and localization.** Relative to retrained RIME, our model improves Target MRR from $0.0911$ to $0.1183$ ($+30\%$) and Target Recall@5 from $10.345\%$ to $15.720\%$ ($+52\%$). On the positive target, center-based localization Recall@1 increases from $11.968\%$ to $13.438\%$, while coverage-based Recall@1 increases from $10.041\%$ to $13.590\%$. Pair-resampled and target-component-clustered pair-micro 95% intervals exclude zero for Target MRR, Target Recall@5, and coverage-based localization Recall@1.

Table 3: Ablation study on controlled SnoBench ranking.

| Model | MRR | R@1 | R@5 |
|---|---|---|---|
| RIME (retrained, NT backbone) | 0.3688 ± 0.0028 | 17.92 ± 0.21 | 59.70 ± 0.41 |
| RIME (retrained, ERNIE-RNA backbone) | 0.3778 ± 0.0065 | 18.53 ± 1.07 | 61.39 ± 0.66 |
| *Partner conditioning* | | | |
| Ours w/ partner conditioning w/o pair prior | 0.3814 ± 0.0096 | 18.70 ± 1.05 | 61.73 ± 1.38 |
| Ours w/o partner conditioning | 0.3834 ± 0.0052 | 19.06 ± 0.73 | 61.46 ± 0.71 |
| *Prediction architecture* | | | |
| Ours w/ dual heads w/o matched pooling | 0.3882 ± 0.0112 | 19.68 ± 1.76 | 61.95 ± 2.05 |
| Ours w/o dual heads | 0.3804 ± 0.0163 | 18.57 ± 2.30 | 61.07 ± 1.81 |
| **Ours (full)** | **0.3918** ± 0.0028 | **19.73** ± 0.97 | **63.32** ± 0.94 |

**The gap is largest in joint prediction.** Center-based Joint Recall@5 increases from $2.738\%$ to $5.832\%$, while coverage-based Joint Recall@5 increases from $2.535\%$ to $5.375\%$, both more than double those of retrained RIME. The coverage-based Joint Recall@5 gain has pair-resampled and target-component-clustered pair-micro 95% intervals excluding zero; complete bootstrap results are reported in Appendix Tables 10 and 11. Together, these results show that strong performance in controlled candidate ranking does not necessarily translate to full-sequence target identification and site localization.

### 6.3 Ablation Studies

We conduct two controlled-ranking analyses to separate backbone and architectural effects. First, we replace RIME's original Nucleotide Transformer (NT) backbone with ERNIE-RNA while keeping its downstream architecture fixed. Second, we use nested ablations of our model, evaluating the nucleotide-pair prior within cross-attention and matched pooling within the dual-head architecture.

Replacing NT with ERNIE-RNA improves retrained RIME, indicating that the RNA-specific backbone explains part of the gain over the original configuration. However, matched-backbone RIME remains below our full model, showing that encoder substitution alone does not account for the overall improvement.

The partner-conditioning ablations reveal an interaction between cross-attention and the nucleotide-pair prior. Cross-attention without the pair prior does not improve performance over removing partner conditioning altogether, but the complete pair-prior-guided conditioning block performs best. We hypothesize that, with limited supervised interaction data, unconstrained cross-attention is more prone to fitting spurious cross-RNA correlations, whereas the signed pairing prior provides biologically grounded guidance, allowing attention to focus on pairing evidence during cross-RNA conditioning.

For the prediction architecture, dual-head prediction provides most of the improvement, while matched pooling yields a smaller additional gain. Together, the ablations indicate that the improvement over matched-backbone RIME reflects downstream interaction modeling rather than backbone choice alone, with the nucleotide-pair prior playing an important role in effective partner conditioning.

## 7 Conclusion

We introduced SnoBench to evaluate snoRNA interaction prediction on full-sequence target identification and site localization, beyond preselected local candidates. We also developed a partner-conditioned interaction model that adapts pretrained RNA representations to the candidate partner through bidirectional cross-attention, nucleotide-pair attention biases, and granularity-matched prediction heads. Our model outperforms evaluated baselines in both controlled and full-sequence evaluation, with particularly large gains when target identification and site localization must succeed jointly. Matched-backbone comparisons and nested ablations further separate these gains from the choice of pretrained encoder and support the contributions of partner conditioning and dual-head prediction. SnoBench reveals that strong discrimination among supplied local candidates does not necessarily transfer to complete sequences, highlighting the importance of evaluating RNA–RNA interaction models under settings in which both the interacting target and its site must be identified.

## REFERENCES


Yue Bi, Fuyi Li, Cong Wang, Tong Pan, Chen Davidovich, Geoffrey I Webb, and Jiangning Song. Advancing microrna target site prediction with transformer and base-pairing patterns. *Nucleic Acids Research*, 52(19):11455–11465, 2024.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P De Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature methods*, 22(2):287–297, 2025.

Gabrielle Deschamps-Francoeur, Sonia Couture, Sherif Abou-Elela, and Michelle S Scott. The snoglobe interaction predictor reveals a broad spectrum of c/d snorna rna targets. *Nucleic acids research*, 50(11):6067–6083, 2022.

Jiayao Gu, Can Chen, and Yue Li. Mirformer: a dual-transformer-encoder framework for predicting microrna-mrna interactions from paired sequences. *Bioinformatics*, 42(Supplement_1):btag222, 2026.

Stephanie Kehr, Sebastian Bartschat, Peter F Stadler, and Hakim Tafer. Plexy: efficient target prediction for box c/d snornas. *Bioinformatics*, 27(2):279–280, 2011.

Bei Liu, Tong Wu, Bernadette A. Miao, Fei Ji, Shun Liu, Pingluan Wang, Yutao Zhao, Yuhao Zhong, Arunkumar Sundaram, Tie-Bo Zeng, Marta Majcherska-Agrawal, Robert J. Keenan, Tao Pan, and Chuan He. snoRNA-facilitated protein secretion revealed by transcriptome-wide snoRNA target identification. *Cell*, 188(2):465–483.e22, 2025. doi: 10.1016/j.cell.2024.10.046. URL https://doi.org/10.1016/j.cell.2024.10.046.

Martin Mann, Patrick R Wright, and Rolf Backofen. Intarna 2.0: enhanced and customizable prediction of rna–rna interactions. *Nucleic acids research*, 45(W1):W435–W439, 2017.

Marco Nicolini, Federico Stacchietti, Francisco Javier Molina, Carlos Cano, Jesus Alcala-Fdez, Alberto Paccanaro, Elena Casiraghi, and Giorgio Valentini. Mirinter-trans: a transformer-based framework for microrna interaction prediction. *Bioinformatics Advances*, 6(1):vbag073, 2026.

Rafael Josip Penić, Tin Vlašić, Roland G Huber, Yue Wan, and Mile Šikić. Rinalmo: general-purpose rna language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, 2025.

Adriano Setti, Giorgio Bini, Flaminia Pellegrini, Valentino Maiorca, Gabriele Proietti, Dimitrios-Miltiadis Vrachnos, Angelo D'Angelo, Alexandros Armaos, Julie Martone, Michele Monti, et al. The role of low-complexity repeats in rna–rna interactions and a deep learning framework for duplex prediction. *Nature communications*, 17(1):1637, 2026.

Hakim Tafer and Ivo L Hofacker. Rnaplex: a fast tool for rna–rna interaction search. *Bioinformatics*, 24(22):2657–2663, 2008.

Hakim Tafer, Stephanie Kehr, Jana Hertel, Ivo L Hofacker, and Peter F Stadler. Rnasnoop: efficient target prediction for h/aca snornas. *Bioinformatics*, 26(5):610–616, 2010.

Weijie Yin, Zhaoyu Zhang, Shuo Zhang, Liang He, Ruiyang Zhang, Rui Jiang, Gan Liu, Jingyi Wang, Xuegong Zhang, Tao Qin, and Zhen Xie. ERNIE-RNA: an RNA language model with structure-enhanced representations. *Nature Communications*, 16:10076, 2025. doi: 10.1038/s41467-025-64972-0. URL https://www.nature.com/articles/s41467-025-64972-0.

Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, et al. Multiple sequence alignment-based rna language model and its application to structural inference. *Nucleic acids research*, 52(1):e3–e3, 2024.

### AI USE STATEMENT

Generative AI tools were used to assist with software development, literature search, and editing of the manuscript for clarity and presentation. AI-assisted code and textual suggestions were reviewed and verified by the authors. The authors take full responsibility for the methodology, experiments, results, claims, and final content of the submission.

### REPRODUCIBILITY STATEMENT

Implementation and training details are described in Appendix B, with full-sequence inference and computational requirements in Appendix C and statistical analyses in Appendix D. The supplementary artifact includes benchmark-construction code, fixed data splits and candidate definitions, training and evaluation scripts, and selected model configurations supporting reproduction of the reported experiments.

## A BENCHMARK CONSTRUCTION AND EVALUATION DETAILS

### A.1 SOURCE OBSERVATIONS AND TARGET-COMPONENT SPLITS

SnoBench is constructed from 50,902 human reproducible duplex-group observations in sheets 4–9 of the published snoKARR-seq supplementary workbook (Liu et al., 2025). References comprise the GRCh38 primary assembly, GENCODE v39 annotations, and the fixed snoDB GRCh38 export. Source genomic coordinates are converted to 0-based half-open intervals; minus-strand sequences are reverse complemented and offsets reversed. A target representative is a contiguous strand-oriented genomic reference span, including intronic sequence where present.

Only identical target start/end coordinates are deduplicated within a model-input pair. Distinct overlapping, touching, and separated observed regions remain separate. The cumulative source filtering is 50,902 source observations $\rightarrow$ 49,872 with target arm at least 5 nt $\rightarrow$ 48,771 with resolved query $\rightarrow$ 35,358 with resolved target $\rightarrow$ 34,427 canonical representative-span-eligible members. These members yield 9,156 model-input pairs and 31,790 observed target intervals after pair construction and within-pair coordinate deduplication. Table 4 reports the SnoBench split counts before restricting cross-method reporting to the shared evaluation cohort.

To reduce target leakage, targets are linked through same-gene/locus relationships, exact sequence identity, global sequence similarity, and local containment. The same-locus rule requires at least 0.80 reciprocal genomic overlap. Global sequence edges require at least 0.90 identity and at least 0.90 coverage of both sequences; containment edges require at least 0.95 identity and at least 0.95 coverage of the shorter sequence. Union-find transitive closure defines target components, and each final component is assigned wholly to one split.

### A.2 CONTROLLED CANDIDATE CONSTRUCTION

Each controlled list contains one designated positive and 16 controls, with six 21-nt, six 42-nt, and five 70-nt windows. A positive must satisfy the coverage rule in equation 1 against at least one individual observed interval. Training assigns one positive width to each pair by a constrained quota; that width does not change across passes or negative identity groups. Eligible gold intervals and legal positive starts are ordered deterministically, and successive training passes cycle through them.

Up to eight controls are drawn from alternative positions on the true target, including up to two near-site controls. Same-target controls must fail the coverage rule against every observed region for that pair. Remaining slots are divided approximately equally between matched and uniform cross-target controls, subject to width quotas. Cross-target sampling excludes known partners of the query, target aliases or identical sequences, the same gene, related target components, previously used genes/components, targets too short for the sampled width, and sampled windows containing non-ACGU bases. Matched sampling relaxes from exact covariates to biotype plus length, biotype alone, and finally the split-local eligible universe. Four deterministic negative-identity groups (R1–R4) are rotated during training; validation and the fixed test view use R1.

Full snoRNA sequences are supplied to the scorer, while gold coordinates, labels, positive ordinals, and sampling categories are withheld from scoring inputs.

Table 4: SnoBench split counts before restricting cross-method reporting to the shared evaluation cohort.

| Counting unit | Train | Validation | Test |
|---|---|---|---|
| Model-input sequence pairs | 7,131 | 868 | 1,157 |
| Distinct positive-target sequences | 3,778 | 384 | 649 |
| Positive-target entities/genes | 3,782 | 388 | 649 |
| Pair-bearing target components | 3,496 | 365 | 637 |
| Observed gold intervals | 24,618 | 3,065 | 4,107 |

### A.3 Full-sequence target sets and common evaluation support

For each held-out pair, the N100 setting contains the annotated positive target and 99 distractors drawn from the eligible split-local genome-wide gene registry, rather than only from genes appearing as positives elsewhere. Distractors are matched hierarchically using GENCODE gene biotype, representative-span length $L$, number of shared-grid search windows $C(L)$, GC content, and low complexity. Low complexity is $1 - H(A, C, G, U)/2$, where $H$ is base-frequency Shannon entropy in bits. The exact tier matches biotype together with $\lfloor \log_2 \max(L, 1) \rfloor$, $\lfloor \log_2 \max(C(L), 1) \rfloor$, $\mathrm{int}(10\, GC)$, and $\mathrm{int}(10\, LC)$. Fallback tiers match biotype plus length, biotype alone, and finally the split-local universe.

The achieved matching tiers are shown in Table 5.

For cross-method reporting, we use a shared held-out cohort of 986 box C/D query–target pairs. The class restriction provides common method support for snoRNA-specific baselines with class-specific applicability, including snoGloBe and PLEXY. Of the 1,157 held-out SnoBench pairs, 987 satisfy the shared C/D and token-support requirements; one list contains a 20-nt distractor whereas the shared neural grid begins at 21 nt, yielding 986 complete N100 lists under the common-support protocol. The same 986 pair identities are used for the controlled and full-sequence comparisons.

### A.4 Exact ties and localization definitions

For exact score ties, let $g$ candidates have strictly larger scores and let $e$ candidates, including the positive, share its score. The positive occupies ranks $g + 1, \ldots, g + e$ under uniform tie-breaking, giving

$$\mathrm{R@}\,K = \frac{\min\{\max(K - g, 0), e\}}{e}, \qquad \mathrm{RR} = \frac{1}{e} \sum_{r=g+1}^{g+e} \frac{1}{r}. \tag{11}$$

Final reporting compares stored scores without decimal rounding. The same expected-rank treatment is used for target retrieval in N100.

For half-open predicted and observed intervals $w = [s_w, e_w)$ and $r = [s_r, e_r)$, the midpoint is retained without rounding. With multiple observed regions, a prediction is a hit if any individual region satisfies the declared criterion; separated regions are never replaced by their outer envelope. When several windows share the exact maximum score on the true target, localization is averaged uniformly over those tied windows. Localization is evaluated on the annotated positive target regardless of its retrieval rank; Joint R@$K$ additionally requires the positive target to lie within the target top $K$.

**Localization criteria under heterogeneous prediction widths.** The benchmark retains variable-width observed regions, while the evaluated methods return regions at different native resolutions: snoGloBe reports 13-nt sites, Ours and retrained RIME search 21/42/70-nt windows, and IntaRNA returns variable-width sites. A single IoU threshold would therefore mix localization success with agreement in interval width. For a prediction completely contained in an observed interval,

$$\mathrm{IoU}(w, r) = \frac{|w|}{|r|}. \tag{12}$$

Table 5: Achieved matching tiers among the 97,614 distractor occurrences in the shared N100 evaluation cohort. A list may contain distractors from more than one tier.

| Achieved matching tier | Distractor occurrences | Share (%) |
|---|---|---|
| Exact covariate bins | 52,150 | 53.425 |
| Biotype and length bin | 36,483 | 37.375 |
| Biotype only | 8,662 | 8.874 |
| Split-wide uniform fallback | 319 | 0.327 |

Table 6: Localization robustness on the shared N100 evaluation cohort. Recall and IoU hit rates are percentages; center distance is measured in nucleotides.

| Method | Center R@1 | Coverage R@1 | IoU$\geq 0.25$ | IoU$\geq 0.50$ | Median center distance |
|---|---|---|---|---|---|
| Ours | **13.44** | **13.59** | **13.39** | **6.69** | **2760.5** |
| RIME (retrained) | 11.97 | 10.04 | 10.04 | 5.07 | 2793.5 |
| snoGloBe (released) | 8.95 | 7.83 | 4.28 | 1.32 | 3850.5 |
| snoGloBe (retrained) | 8.72 | 8.01 | 4.46 | 1.32 | 2842.0 |
| IntaRNA | 9.23 | 8.92 | 7.40 | 3.65 | 4072.0 |

A 13-nt or 21-nt prediction lying entirely inside a representative 46-nt observed region therefore has IoU 0.283 or 0.457, respectively, both below a 0.5 threshold despite complete containment within the observed interval.

The two reported criteria provide complementary views of localization. Center-based evaluation emphasizes positional agreement across native prediction widths, while coverage additionally incorporates span overlap: a prediction shorter than half of an observed region must be fully contained, and a wider prediction must cover at least half of that observed region. IoU and center-distance are reported as supplementary boundary-sensitive diagnostics.

The shared 21/42/70-nt neural search grid attains 100% coverage of annotated positive regions under the benchmark coverage criterion across all 986 pairs in the shared evaluation cohort. The declared shared search space therefore contains at least one coverage-valid candidate window for every positive target.

# B Training and Implementation Details

## B.1 Training objective

Let a 17-window training list contain $J$ distinct target entity IDs, with candidate set $I_t$ for target $t$ and designated positive $k^\star$. With $y_t = 0.95\,\mathbf{1}[t = t^\star] + 0.05/J$, the target loss is

$$\mathcal{L}_{\text{target}} = \frac{2}{J} \sum_t \frac{1}{|I_t|} \sum_{k \in I_t} \text{BCE}_{J-1}(a_k, y_t), \tag{13}$$

where $\text{BCE}_{J-1}$ is binary cross-entropy with positive weight $J-1$, computed separately for each list. Let $c^\star$ denote the target component containing $t^\star$, and let $I_{c^\star}$ denote the candidates whose targets belong to $c^\star$. By construction, cross-target controls from this component are excluded, so $I_{c^\star} = I_{t^\star}$ in the training lists. Site supervision uses unsmoothed cross-entropy over $I_{c^\star}$,

$$\mathcal{L}_{\text{site}} = -\log \frac{\exp(\ell_{k^\star})}{\sum_{k \in I_{c^\star}} \exp(\ell_k)}, \tag{14}$$

and the joint loss applies 0.05 label smoothing to the 17 candidate scores $z_k = a_k + \ell_k$:

$$\mathcal{L}_{\text{joint}} = -\sum_{k=1}^{17} \widetilde{y}_k \log \frac{\exp(z_k)}{\sum_{j=1}^{17} \exp(z_j)}, \qquad \widetilde{y}_k = 0.95\,\mathbf{1}[k = k^\star] + \frac{0.05}{17}. \tag{15}$$

The final objective is $0.25\mathcal{L}_{\text{target}} + 0.25\mathcal{L}_{\text{site}} + 0.50\mathcal{L}_{\text{joint}}$.

Table 7: Implementation settings for the final model.

| Setting | Value |
|---|---|
| Encoder / trainability | ERNIE-RNA / frozen |
| Encoder width | 768 |
| Target context | candidate window $\pm 64$ nt, boundary-clipped |
| Conditioning blocks / heads / head width | 1 / 12 / 64 |
| Pair-prior initialization | WC: 3.0; GU/UG: 1.5; other: 0 |
| Pair-prior gates | four each initialized at $+1$, $-1$, and 0 |
| Pair-map input channels | 4 product + pair identity + relative position |
| Convolution channels | 32 / 64 / 64 |
| Target pooling | masked mean + normalized LME |
| Site pooling | learned-query attention pooling |
| Attention/FFN dropout / downstream dropout | 0.15 / 0.25 |
| Total / trainable parameters | 107,662,101 / 14,313,116 |

**Gradient-scale normalization.** For weighted BCE,

$$\frac{\partial \,\mathrm{BCE}_{J-1}(a,y)}{\partial a} = (1-y)\sigma(a) - (J-1)y[1-\sigma(a)].$$

At $a = 0$ this derivative is $(1 - Jy)/2$. Consequently, for $k \in I_t$,

$$\left.\frac{\partial \mathcal{L}_{\mathrm{target}}}{\partial a_k}\right|_{a=0} = \frac{1/J - y_t}{|I_t|}. \tag{16}$$

This matches the candidate-level gradient obtained by distributing the zero-logit gradient of a smoothed target multiclass cross-entropy equally across a target's windows. The factor therefore provides a fixed analytical gradient-scale normalization at the symmetric initialization point.

### B.2 Optimization and checkpoint selection

Training uses AdamW with $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. Trainable new modules use learning rate $2 \times 10^{-5}$; pair-prior parameters use zero weight decay and other trainable modules use weight decay $10^{-3}$. The schedule allows 14,272 updates over at most 16 passes, with 1,427 warmup updates followed by cosine decay. Training uses BF16 autocast, gradient clipping at 1.0, four complete parent lists per microbatch, and two accumulation steps, for an effective batch of eight lists.

Early stopping requires at least 12 completed passes, patience four, and a validation MRR improvement greater than $10^{-4}$. Checkpoints are selected by clean validation pair-micro 17-way MRR of the candidate-local $a + \ell$ score. The full-sequence scan uses seed 0/pass 9, selected by validation MRR before held-out evaluation. All model, checkpoint, decoder, and localization-setting selection is based exclusively on validation data; N100 test outcomes play no role in selection.

### B.3 Model implementation

Additional implementation settings are summarized in Table 7. The final model uses frozen ERNIE-RNA nucleotide representations (Yin et al., 2025); each candidate target window is encoded with up to 64 nt of flanking context on each side, clipped at target boundaries. For valid spatial cells $\Omega$ and feature channel $c$, the normalized log-mean-exp pooling used by the target branch is

$$\mathrm{LME}(F_c;\Omega) = \log \sum_{(i,j)\in\Omega} \exp(F_{cij}) - \log|\Omega|. \tag{17}$$

Pooling is candidate-local. Cross-window reduction is applied only when candidate scores are converted to a full-sequence target score. The final candidate score is exactly $z_k = a_k + \ell_k$.

### B.4 Ablation configurations

Controlled-ranking ablations evaluate four aspects of the proposed design: cross-RNA conditioning, the signed attention pair prior, granularity-matched pooling, and the dual prediction architecture. The

Table 8: Scoring and search configurations for the methods reported in the main N100 full-sequence comparison.

| Method | Controlled score | Full-sequence search | Predicted span | Target reduction |
|---|---|---|---|---|
| Ours | $a + \ell$ | 21/42/70 grid, strides 10/21/35 | 21/42/70 nt | max window |
| RIME (retrained) | class-logit difference | same 21/42/70 grid | 21/42/70 nt | max window |
| snoGloBe | max native probability | native 13-mer scan | 13 nt | max site |
| IntaRNA | negative minimum energy | native full-sequence search | variable | best energy |

cross-RNA-conditioning control replaces partner exchange with matched-capacity within-sequence refinement. The attention-prior ablation removes the signed attention bias. The shared-pooling variant supplies the same mean/LME representation to the target and site predictors, and the single-head comparison uses one candidate-level predictor in place of the dual prediction architecture. All variants use the same SnoBench splits, controlled candidate construction, training budget, and validation-based checkpoint-selection procedure.

# C Full-Sequence Inference and Computational Scope

## C.1 Baseline training and search configurations

Controlled methods score each supplied candidate independently. Because complete-target scanning is substantially more expensive than controlled ranking, the main N100 comparison focuses on representative high-performing methods from the major baseline families evaluated in the controlled setting: IntaRNA for thermodynamic interaction prediction, snoGloBe for snoRNA-specific learning, and retrained RIME for RNA-LM-based interaction modeling. All full-sequence methods receive the same candidate target identities and use their declared gold-independent search procedures. Table 8 summarizes these search configurations.

For retrained RIME, the official RIME head is trained on the 7,131 SnoBench training parents using the same fixed positive schedule and rotating R1–R4 controlled lists, with 17-way cross-entropy on its native class-logit difference. The Nucleotide Transformer encoder is frozen, and checkpoints are selected by validation MRR. The matched-backbone comparison changes only the frozen input representation from Nucleotide Transformer to ERNIE-RNA while retaining the RIME head and training protocol. Retrained snoGloBe retains its native 127-feature tree-based formulation and uses RNAplex-derived weak labels on canonical 13-mer examples; tree-count checkpoints are likewise selected on validation. These retraining procedures preserve the native modeling assumptions of the baselines while adapting their trainable components to SnoBench.

Ours and retrained RIME use the same shared grid. For a target of length $L \geq w$, starts advance by $\lfloor w/2 \rfloor$ for $w \in \{21, 42, 70\}$ and an additional tail window at $L - w$ is included when needed so that the target end is covered. Each window receives up to 64 nt of clipped flanking context. The target score is the maximum candidate score and all exact maximizing ties are retained for localization.

snoGloBe uses its native 13-mer search (query stride one, target stride two), while IntaRNA uses native full-query/full-target site discovery with Turner04 energies at 37°C, a 7-bp seed, and maximum interaction length 70. Preserving native inference keeps each baseline's intended search procedure, while the candidate target identities remain fixed across methods. The localization criteria in Appendix A.4 therefore support comparison across these native prediction resolutions.

## C.2 Scope of full-sequence evaluation

Full-sequence evaluation searches every candidate target for interaction sites before reducing candidate scores to a target-level score. The N100 setting uses one annotated target together with 99 matched genome-wide distractors, providing a large-candidate retrieval evaluation that can be applied consistently across heterogeneous learned and native baselines.

The evaluation pipeline also supports N1000 target sets. Moving from N100 to N1000 increases the number of complete targets processed by approximately $10\times$, and repeating full-sequence inference for three trained seeds adds another factor of three. Retrained RIME further uses the same dense

Table 9: Compute summary for final-model training and N100 full-sequence inference. Inference GPU-hours report active scoring time.

| Stage | Model | Hardware | Wall time | GPU-hours |
|---|---|---|---|---|
| Training / seed | Ours | 1× A100-80GB | ∼9.4 h | ∼9.4 |
| Training / 3 seeds | Ours | 1× A100-80GB per run | – | 28.2 |
| N100 full scan | Ours | 8× A100-80GB | 13.3 h | 103.8 |
| N100 full scan | RIME (NT) | 4× L40S | 114.5 h | 450.7 |

Table 10: Paired N100 bootstrap comparison of Ours and retrained RIME using the pair-micro estimand reported in the main paper. Recall point estimates are percentages; differences and confidence limits are percentage points. MRR uses raw units.

| Endpoint | Criterion | Ours | RIME | Difference | 95% CI lower | 95% CI upper |
|---|---|---|---|---|---|---|
| Target | R@1 | 3.25 | 2.64 | +0.61 | -0.71 | +1.93 |
| Target | R@5 | 15.72 | 10.34 | +5.38 | +3.04 | +7.81 |
| Target | MRR | 0.1183 | 0.0911 | +0.0271 | +0.0144 | +0.0401 |
| Localization | center R@1 | 13.44 | 11.97 | +1.47 | -0.56 | +3.50 |
| Localization | coverage R@1 | 13.59 | 10.04 | +3.55 | +1.47 | +5.63 |
| Joint | center R@1 | 2.74 | 1.22 | +1.52 | +0.46 | +2.59 |
| Joint | center R@5 | 5.83 | 2.74 | +3.09 | +1.67 | +4.56 |
| Joint | coverage R@1 | 2.59 | 1.22 | +1.37 | +0.30 | +2.48 |
| Joint | coverage R@5 | 5.38 | 2.54 | +2.84 | +1.52 | +4.26 |

21/42/70-nt grid with the substantially larger Nucleotide Transformer backbone; its reported N100 scan requires 450.7 GPU-hours, compared with 103.8 GPU-hours for our final scan. We therefore use three independent training runs for controlled ranking and the validation-selected checkpoint for the full-sequence evaluation. All checkpoint selection is completed before N100 test evaluation.

### C.3 Computational cost

Final-model training uses one NVIDIA A100-SXM4-80GB GPU per seed and requires approximately 9.4 GPU-hours per seed, or 28.2 GPU-hours across the three reported training runs. The final N100 scan of our model used eight A100-80GB GPUs and required approximately 13.3 hours of active wall-clock scoring time, corresponding to 103.8 GPU-hours.

Retrained RIME is evaluated on the same 21/42/70-nt search grid but uses the substantially larger Nucleotide Transformer backbone. Its final N100 scan used four NVIDIA L40S GPUs and required approximately 114.5 hours of wall-clock time, corresponding to 450.7 GPU-hours. These measurements quantify the compute required by the two learned models under the reported shared-grid N100 evaluation.

The table reports the final training and inference runs used for the paper's reported results.

## D Supplementary Statistical Analysis

### D.1 Paired N100 bootstrap

For each N100 endpoint, Ours and retrained RIME are evaluated on the same 986 query–positive-target pairs and the same fixed target lists. The reported N100 point estimates use pair-micro aggregation. We quantify evaluation-set uncertainty in two complementary ways while keeping the fitted checkpoints fixed. First, paired bootstrap resamples the 986 evaluation pairs with replacement and recomputes the pair-micro difference. Second, target-component-clustered bootstrap resamples the positive-target components with replacement, retains all constituent pairs, and again recomputes the pair-micro difference. The latter preserves the reported pair-level estimand while accounting for dependence among pairs that share a target component.

Table 11: Target-component-clustered paired-bootstrap 95% intervals for the same pair-micro Ours-minus-RIME differences. Recall limits are percentage points; MRR limits use raw units.

| Endpoint | Criterion | CI lower | CI upper |
|---|---|---|---|
| Target | R@1 | -1.74 | +3.06 |
| Target | R@5 | +0.11 | +11.73 |
| Target | MRR | +0.0008 | +0.0592 |
| Localization | center R@1 | -1.04 | +4.16 |
| Localization | coverage R@1 | +0.93 | +6.31 |
| Joint | center R@1 | -0.21 | +3.50 |
| Joint | center R@5 | +0.05 | +7.04 |
| Joint | coverage R@1 | -0.41 | +3.44 |
| Joint | coverage R@5 | +0.23 | +5.95 |

The pair-resampled and target-component-clustered intervals both exclude zero for Target MRR, Target R@5, coverage localization R@1, and coverage Joint R@5. Together, the two analyses assess the reported pair-micro differences under pair-level and target-component-clustered resampling. Model and checkpoint selection are completed using validation data before N100 evaluation.

## E Reproducibility and Availability

Training uses seeds 0/1/2. The executed training environment is Python 3.8.18, PyTorch 1.10.0, and CUDA 11.3. An anonymized artifact accompanying the submission includes benchmark-construction code, fixed split and candidate-list definitions, training and evaluation scripts, and selected model configurations supporting reproduction of the reported experiments. Upon publication, we will publicly release the code, processed SnoBench benchmark data, trained model checkpoints, and evaluation artifacts. Third-party source data remain governed by their original access and redistribution terms.