GEOMETRY IS NOT ENOUGH: RANKING RNA STRUCTURE CANDIDATES AGAINST A SEQUENCE PRIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Computational RNA structure prediction produces many candidate conformations per sequence, and selecting among them remains a bottleneck, especially for long RNAs, whose candidates often share correct base-paired stems but differ in how those stems pack in three dimensions. Geometry alone can confirm that a packing is physically plausible, but not that it is the one the sequence adopts. We therefore frame evaluation as sequence–geometry alignment and introduce SIRGE, an SE(3)-invariant evaluator that conditions the geometry of every nucleotide pair on attention maps from a frozen RNA language model, used as a sequence prior. A contact-recovery objective trains these pair representations to recognise native tertiary contacts, so SIRGE judges each candidate's packing against what its sequence implies. Against six established evaluators under two distribution shifts, SIRGE ranks candidates best. On CLS-1, a benchmark of held-out, longer RNAs with broader candidate quality, it improves Kendall-τ alignment by 70% over the strongest baseline; on candidates from six unseen deep-learning predictors, it almost doubles the strongest baseline's top-1 hit rate. Ablations show that the prior and the contact objective are complementary, and that alignment helps most on long RNAs. An interpretability study shows that the prior carries tertiary-contact information beyond base pairing, and that SIRGE's scores track native tertiary packing where a geometry-only counterpart does not. These results suggest that for long RNAs geometric plausibility is necessary but not sufficient, and that pretrained sequence models can supply the missing reference.

1 INTRODUCTION

RNA molecules regulate gene expression, catalyse biochemical reactions, and are an expanding class of therapeutic targets (Cech & Steitz, 2014). Their functions depend critically on three-dimensional structure (Churkin et al., 2018; Ponce-Salvatierra et al., 2019), yet determining RNA conformations remains difficult and costly (Renaud et al., 2018; Jin & Chen, 2025). Computational pipelines therefore often produce multiple candidates of varying quality for the same sequence (Li et al., 2025a; Tarafder & Bhattacharya, 2024), making reliable evaluation essential, since errors at this stage can negate gains made by the predictor itself (Townshend et al., 2021; Li et al., 2025b). RNA structure evaluation is typically approached primarily as a geometric problem. Given a candidate structure, an evaluator reasons over its atoms, distances, orientations, and higher-order geometric features to estimate its quality. We argue that this view is incomplete and that candidate evaluation is fundamentally a problem of **sequence–geometry alignment**. Our argument is that the quality of a candidate structure is not determined by its geometry in isolation, but by whether that geometry is compatible with the particular RNA sequence that is expected to adopt it. Geometry describes the interactions realised by a candidate conformation, while the sequence provides context about which structural arrangements are plausible for that RNA.

Recent methods have made geometric reasoning for RNA evaluation increasingly expressive (Liu et al., 2026; Li et al., 2025a; Townshend et al., 2021; Tarafder & Bhattacharya, 2024; Li et al., 2025b), but they do not condition these geometric representations on contextual nucleotide embeddings pretrained across large RNA sequence collections. RNA language models (Chen et al., 2022; Abramson et al., 2024; Penić et al., 2025) provide a natural source of such information. Through self-supervised pretraining, they learn nucleotide representations that capture dependencies across

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.

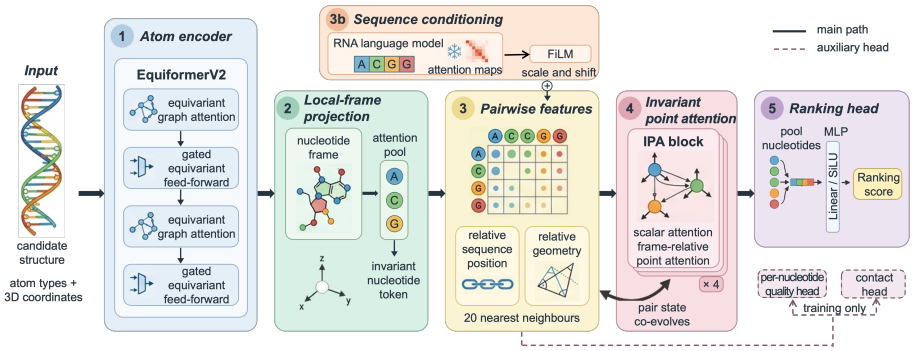


Figure 1: **Overview of SIRGE.** Atom features from an equivariant encoder ① are pooled into nucleotide-local frames ②. Pair geometry ③ is modulated by attention maps from a frozen RNA language model via FiLM ③b, then refined jointly with nucleotide states by invariant point attention ④. A ranking head scores the candidate ⑤. Dashed auxiliary heads are used only in training.

RNA sequences and transfer to structure-related tasks. RiNALMo, for example, learns contextual nucleotide embeddings from millions of non-coding RNA sequences and demonstrates strong transfer to RNA structure prediction tasks (Penić et al., 2025). This raises a natural research question: *can a sequence prior help identify the best RNA structure candidate?*

We introduce **SIRGE**, a **S**equence-**I**nformed **R**NA **G**eometric **E**valuator for ranking 3D structure candidates. As illustrated in Fig. 1, SIRGE first encodes the candidate at atomic resolution with an $\mathrm{SE}(3)$-equivariant atom transformer from the Equiformer family (Liao et al., 2024) and then projects the resulting atom representations into nucleotide-local frames (Tarafder & Bhattacharya, 2024). Attention maps from a frozen RNA language model (Penić et al., 2025) modulate the geometric pair features between nucleotides through FiLM (Perez et al., 2018), which conditions the structure on a base-pairing prior derived from the sequence. Dense invariant point attention, biased by these conditioned pair features, propagates information across the nucleotide frames (Jumper et al., 2021; Tarafder & Bhattacharya, 2024). Finally, a pooled head produces the candidate score, while auxiliary heads that predict tertiary contacts and per-nucleotide lDDT (Mariani et al., 2013) supervise the intermediate representations during training only. We compare SIRGE with six evaluators spanning diverse inductive biases under two distribution shifts. CLS-1 (Li et al., 2025a) holds out sequence clusters containing longer RNAs and a broader range of candidate quality, whereas Gen-OOD (Martinović et al., 2024) contains predictions from six deep-learning structure predictors, none of which generated the training candidates. SIRGE outperforms every baseline in both settings, improving global Kendall–τ by 70% on CLS-1 and almost doubling the top-1 hit rate on Gen-OOD. Because the RNA sequence is always known at scoring time, these gains require no additional data. Interpretability analyses further show that the improvement arises from the joint mechanism formed by the sequence prior and the contact-recovery objective: the prior carries information about tertiary helix packing beyond canonical base pairing, while the objective trains SIRGE to recover this signal and use it as a reference when evaluating candidate structures. The benefit is most pronounced for RNAs longer than 150 nucleotides, where candidates often share similar secondary structure but differ substantially in tertiary packing (Bahai et al., 2024; Li et al., 2025b; Miao et al., 2020).

Our main **contributions** are as follows. **(i)** We formulate RNA 3D structure evaluation as sequence–geometry alignment and introduce SIRGE, which conditions pairwise geometric representations on attention maps from a frozen RNA language model. **(ii)** We show that SIRGE outperforms established evaluators under shifts in sequence length, candidate quality, and candidate generator, and we isolate the contribution of each component through ablation. **(iii)** Through an interpretability study, we show that the sequence prior carries tertiary packing information that SIRGE learns to use as a reference, with the largest gains on long RNAs.

2 Related Work

Early RNA structure evaluators relied on hand-crafted objectives, either combining knowledge-based potentials with physics-inspired energy terms (Xu & Chen, 2015) or deriving distance- and torsion-

dependent potentials from experimental structures (Wang et al., 2015). However, their accuracy is limited by how well hand-crafted terms approximate RNA folding physics (Tan et al., 2019; Li et al., 2025a). Learned evaluators instead derive representations from candidate geometry. RNA3DCNN applies convolutions to voxelised atomic environments (Li et al., 2018), while ARES and PAMNet use geometric graph networks (Townshend et al., 2021; Zhang et al., 2023). Recent methods introduce richer structural organisation. lociPARSE uses nucleotide-local frames and invariant point attention (Tarafder & Bhattacharya, 2024) and EquiRNA models RNA through hierarchical equivariant representations (Li et al., 2025b). HomRank formulates evaluation as within-RNA learning to rank (Li et al., 2025a). These methods improve geometric reasoning but do not condition it on contextual representations learned through large-scale RNA sequence pretraining. RNA language models provide such representations. RNA-FM (Chen et al., 2022) and RiNALMo (Penić et al., 2025) learn contextual nucleotide embeddings from unlabelled RNA sequences and transfer them to structural and functional tasks. Their use in structure modelling has focused primarily on predicting structural properties rather than ranking alternative conformations. SIRGE connects these lines of work by using nucleotide embeddings to condition a symmetry-aware geometric evaluator.

3 SEQUENCE–GEOMETRY ALIGNMENT FOR RNA STRUCTURE EVALUATION

SIRGE scores a candidate structure by how well its tertiary helix packing agrees with the nucleotide interactions its sequence implies. A frozen RNA language model supplies these expected interactions as a prior (Section 3.1), and SIRGE learns to judge each candidate's geometry against it. Let $S = (s_1, \ldots, s_T)$, with $s_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$, be an RNA sequence with candidate structures $X_1, \ldots, X_M$. Candidate X_m has N_m atoms; atom a has position $\mathbf{x}_{m,a} \in \mathbb{R}^3$, element one-hot $\mathbf{c}_{m,a} \in \{0,1\}^5$ and nucleotide $\rho_m(a) \in \{1, \ldots, T\}$. We learn a scoring function $f_\theta(X_m, S) \in \mathbb{R}$ with parameters θ.

3.1 DEFINING SEQUENCE–GEOMETRY ALIGNMENT

RNA folds hierarchically (Brion & Westhof, 1997; Tinoco Jr & Bustamante, 1999): the chain first folds back on itself so that segments whose bases can pair form canonical base pairs (A–U, G–C and G–U), creating short helices called stems that make up the secondary structure. The stems are then arranged relative to one another in three dimensions, held in place by tertiary contacts between nucleotides that lie far apart in the sequence. This arrangement is known as helix packing (Cate et al., 1996; Doherty et al., 2001); it gives the molecule its overall shape, the tertiary structure. In longer RNAs, stems form largely independently of their helix packing. In blind settings, where the experimental structure is not yet known, predictors tend to reproduce the stems correctly but misplace them relative to one another (Miao et al., 2020). Long RNAs are correspondingly hard to evaluate (Li et al., 2025b): geometry alone can confirm that stems are well formed and their helix packing is physically plausible, but not which of the many plausible arrangements the sequence adopts. Base pairing alone does not settle this, as sequence changes that preserve every base pair can still alter how the stems pack (Yesselman et al., 2019). Judging helix packing therefore requires a sequence-specific expectation of which nucleotides should interact.

We obtain this expectation from RiNALMo (Penić et al., 2025), an RNA language model pretrained by masked-nucleotide prediction on 36 million sequences. Its self-attention maps provide pairwise signals between sequence positions that are learned without structural labels and available for any RNA at test time, including RNAs unseen in training. To identify heads that encode structure, we probe each against annotated base pairs on validation RNAs and retain the three that recover them well. These heads mark the stems whose arrangement SIRGE must judge, and on held-out RNAs they also detect tertiary contacts above chance (Appendix B), which is the information helix packing requires. Denoting the extraction and preprocessing of these maps (symmetrisation and normalisation; Appendix A.4) by ψ, the sequence prior is $A = \psi(S) \in \mathbb{R}^{T \times T \times K}$, where $a_{ij} = A_{ij,:} \in \mathbb{R}^K$, and $K = 3$ features for each nucleotide pair (i, j). Because all candidates for the same RNA share A, the prior alone is identical across them; thus, sequence information can affect within-target ranking only through its interaction with candidate-dependent geometry. We therefore define *sequence–geometry alignment* as the agreement between a candidate's pairwise geometry, in particular its tertiary contacts, and the interactions its sequence prior expects. SIRGE learns to measure it by conditioning the geometric description of every nucleotide pair on a_{ij} (Section 3.2), with a contact objective that trains these pair representations to recognise native tertiary contacts (Section 3.3).

3.2 LEARNING SEQUENCE–GEOMETRY ALIGNMENT

SIRGE first encodes each nucleotide's atomic environment, then compares the arrangement of nucleotide pairs with the sequence prior. This hierarchy connects local molecular geometry to helix packing, the level at which sequence–geometry alignment is defined (Section 3.1).

From atoms to nucleotides. A compact Equiformer-style encoder Φ_{atom} (Liao et al., 2024) operates on a graph connecting spatially neighbouring atoms. Internally, its features and edge spherical harmonics span orders $\ell \in \{0, 1, 2\}$ (scalars, vectors and second-order tensors), and attention mixes the second-order channels into the lower orders at every layer, so they still shape the output. We then read out only the scalar and vector features of each atom, $\mathbf{a}^{(0)}_{m,a}$ and $\mathbf{a}^{(1)}_{m,a}$. To make the vectors independent of the candidate's global orientation, we express them in local frames. Biologically, each nucleotide consists of a base attached to a sugar element, and the sugars, linked by phosphate groups, form the backbone. We mirror this biological construction and, following AlphaFold2 and lociPARSE (Jumper et al., 2021; Tarafder & Bhattacharya, 2024), build a frame $F_{m,i} = (\mathbf{R}_{m,i}, \mathbf{o}_{m,i})$ for nucleotide i from its sugar, phosphate and base atoms, with origin $\mathbf{o}_{m,i}$ on the sugar and rotation $\mathbf{R}_{m,i} \in \mathrm{SO}(3)$ (Appendix A.2). Each atom a of nucleotide $i = \rho_m(a)$ is then described in its nucleotide's frame,

$$\mathbf{d}_{m,a} = \big[\mathbf{a}^{(0)}_{m,a},\, \mathbf{R}^{\top}_{m,i}\mathbf{a}^{(1)}_{m,a},\, \mathbf{R}^{\top}_{m,i}(\mathbf{x}_{m,a} - \mathbf{o}_{m,i})\big], \tag{1}$$

where brackets denote concatenation and $\mathbf{R}^{\top}_{m,i}$ acts on each vector channel. Attention pooling over each nucleotide's atoms, combined with its base identity, gives the nucleotide representation $\mathbf{h}^{(0)}_{m,i} \in \mathbb{R}^{d_h}$ (Appendix A.3).

Aligning pair geometry with the sequence prior. The nucleotide representations $\mathbf{h}^{(0)}_{m,i}$ summarise the atomic environment of each nucleotide. To describe how nucleotides are arranged relative to one another, SIRGE also constructs a separate descriptor $\mathbf{g}_{m,ij} \in \mathbb{R}^{d_g}$ for every ordered pair (i, j). Its components are computed directly from the two nucleotide frames: the distance between the origins, the direction from i to j in the frame of i, their relative orientation, and an embedding of the signed sequence separation $i - j$ (Appendix A.5). Combining spatial features with sequence separation allows SIRGE to tell tertiary contacts, i.e. pairs far apart in sequence but close in space, from chain neighbours and base pairs within stems. This is important, as only tertiary contacts reveal the helix packing. SIRGE conditions each descriptor on the prior features a_{ij} of the same pair (Section 3.1) with feature-wise linear modulation (FiLM) (Perez et al., 2018),

$$\widetilde{\mathbf{g}}_{m,ij} = [\mathbf{1} + \boldsymbol{\gamma}(a_{ij})] \odot \mathbf{g}_{m,ij} + \boldsymbol{\beta}(a_{ij}), \tag{2}$$

where the learned affine maps $\boldsymbol{\gamma}, \boldsymbol{\beta} : \mathbb{R}^K \to \mathbb{R}^{d_g}$ scale and shift each feature, and a multilayer perceptron gives the initial pair representation $\mathbf{z}^{(0)}_{m,ij} = \phi_{\mathrm{pair}}(\widetilde{\mathbf{g}}_{m,ij}) \in \mathbb{R}^{d_z}$. Because the prior is the same for every candidate while the geometry varies, this is where SIRGE compares each pair's arrangement with what the sequence expects, for instance whether two distant nucleotides that the prior links are in contact in this candidate.

Propagating alignment across the molecule. Each pair representation describes a single pair, yet whether a pair's geometry agrees with the sequence prior depends on the rest of the structure. SIRGE uses $B = 4$ dense invariant point attention (IPA) blocks (Jumper et al., 2021), in which every nucleotide attends to every other. In each block and head (superscripts omitted), the attention logit is

$$e_{m,ij} = \mathbf{q}^{\top}_{m,i}\mathbf{k}_{m,j}/\sqrt{2d_c} - \eta D_{m,ij} + \xi(\mathbf{z}_{m,ij}), \tag{3}$$

which combines three signals: (i) the similarity of d_c-dimensional queries and keys projected from the nucleotide representations; (ii) spatial proximity in this candidate, where $D_{m,ij}$ is the summed squared distance between learned points placed in the two nucleotide frames and $\eta > 0$ is learned; and (iii) what the sequence expects of the pair, through an affine bias ξ computed from the sequence-conditioned pair representation. Each nucleotide can thus draw on nucleotides that are close in the candidate, linked by the sequence, or both, so that disagreements between the two can shape its update. A softmax over j turns the logits into weights that aggregate scalar and point values into $\mathbf{h}^{(b+1)}_{m,i}$ (Appendix A.6). Each pair representation is then updated from its two nucleotides,

$$\mathbf{z}^{(b+1)}_{m,ij} = \mathbf{z}^{(b)}_{m,ij} + U_b\big([\mathbf{z}^{(b)}_{m,ij}, \pi(\mathbf{h}^{(b+1)}_{m,i}), \pi(\mathbf{h}^{(b+1)}_{m,j})]\big), \tag{4}$$

with a shared affine projection π and a multilayer perceptron U_b, so that it reflects the wider context of both nucleotides rather than the pair alone. The contact objective (Section 3.3) trains the final pair representations $\mathbf{z}_{m,ij}^{(B)}$ to recognise native tertiary contacts. Finally, attention pooling over the final nucleotide representations and a multilayer perceptron give the score (Appendix A):

$$f_\theta(X_m, S) = \mathrm{MLP}_{\mathrm{score}}\big(\mathrm{Pool}_{\mathrm{RNA}}(\{\mathbf{h}_{m,i}^{(B)}\}_{i=1}^{T})\big) \tag{5}$$

3.3 Training objectives

As in HomRank (Li et al., 2025a), each training batch holds candidates of a single RNA, with labels computed from its native structure X_*, so SIRGE learns to compare structures of the same sequence. The main objective ranks whole candidates and produces the score, whilst two auxiliary objectives localise structural errors to individual nucleotides and to nucleotide pairs. For ranking, ListNet (Cao et al., 2007) matches, over a list $\mathcal{I}$ of candidates, the predicted distribution $\widehat{p}_m \propto \exp(f_\theta(X_m, S)/\tau)$ to a target $p_m \propto \exp(-s_m/\tau)$, where s_m is the RMSD of X_m from X_* after rigid superposition, divided by sequence length T. We minimise the cross-entropy

$$\mathcal{L}_{\mathrm{rank}} = \mathbb{E}_{\mathcal{I}}\big[-\sum_{m\in\mathcal{I}} p_m \log \widehat{p}_m\big]. \tag{6}$$

Building the target from RMSD values lets it reflect the relative quality of each candidate compared to each other candidate. RMSD, however, is dominated by poorly modelled regions (Zhang & Skolnick, 2004; Tarafder & Bhattacharya, 2024) and its differences grow with RNA length (Hajdin et al., 2010), so dividing by T keeps the target from becoming much sharper for long RNAs, without changing the order within an RNA.

Learning local structure accuracy. A whole-candidate score cannot localise errors: a stem can keep its internal geometry yet sit in the wrong place. A linear head therefore maps each final nucleotide representation $\mathbf{h}_{m,i}^{(B)}$ to a local score, trained against per-nucleotide lDDT (Mariani et al., 2013), the fraction of interatomic distances around nucleotide i that X_m preserves from X_*, which requires no superposition. The loss $\mathcal{L}_{\mathrm{local}}$ applies ListNet at each nucleotide position, comparing the candidates of the batch with one another rather than the nucleotides of one candidate. With the local score $\widehat{y}_{m,i}^{\mathrm{local}} = \mathbf{w}_{\mathrm{local}}^{\top}\mathbf{h}_{m,i}^{(B)} + b_{\mathrm{local}}$ and the label $y_{m,i}^{\mathrm{local}}$, the predicted and target distributions over a list $\mathcal{I}$ at position i are $\widehat{p}_{m,i} \propto \exp(\widehat{y}_{m,i}^{\mathrm{local}}/\tau_{\mathrm{local}})$ and $p_{m,i} \propto \exp(y_{m,i}^{\mathrm{local}}/\tau_{\mathrm{local}})$, and the cross-entropy is averaged over positions,

$$\mathcal{L}_{\mathrm{local}} = \mathbb{E}_{\mathcal{I}}\big[\frac{1}{T}\sum_{i=1}^{T}\big(-\sum_{m\in\mathcal{I}} p_{m,i}\log \widehat{p}_{m,i}\big)\big], \tag{7}$$

where $\mathbf{w}_{\mathrm{local}}$ and b_{local} are learned. Higher lDDT is better, so both distributions put more mass on better-modelled candidates at that position. Within an RNA, nucleotide position alone explains most of the variance in per-nucleotide lDDT (median 80.7% across our training RNAs, interquartile range 67.7–85.8%), so this variation says nothing about which candidate is better. Because the head reads the same features that the ranking readout pools, it shapes the representations behind the score.

Learning contact recovery. Whilst per-nucleotide lDDT locates errors, it checks only distances found in the native, so it is blind to spurious contacts that a candidate adds. Tertiary contacts hold the stems together, so recognising which of them a candidate correctly reproduces is direct evidence of whether its fold is right. We call a pair a tertiary contact if its nucleotides are close in space, at least five positions apart in the sequence, and not a canonical base pair. For every pair that is a contact in the candidate, the native or both, a head reads the symmetrised final pair representation $\mathbf{z}_{m,ij}^{(B)} + \mathbf{z}_{m,ji}^{(B)}$ and predicts whether the candidate shares that contact with the native, trained with binary cross-entropy ($\mathcal{L}_{\mathrm{contact}}$). With $\widehat{y}_{m,ij}^{\mathrm{contact}} = \sigma(\mathrm{MLP}_{\mathrm{contact}}(\mathbf{z}_{m,ij}^{(B)} + \mathbf{z}_{m,ji}^{(B)}))$, a symmetric head predicts contact recovery from the final pair states:

$$\mathcal{L}_{\mathrm{contact}} = \frac{1}{|\mathcal{W}|}\sum_{(m,i,j)\in\mathcal{W}} \ell_{\mathrm{BCE}}\big(y_{m,ij}^{\mathrm{contact}}, \widehat{y}_{m,ij}^{\mathrm{contact}}\big), \tag{8}$$

Table 1: Global and band-specific Kendall-τ alignment on CLS-1. Mean $\pm$ std over three seeds.

Model	Global	D1 (2–5 Å)	D2 (5–10 Å)	D3 (10–15 Å)
ARES	0.133 ± 0.016	0.306 ± 0.037	0.186 ± 0.024	0.047 ± 0.021
PAMNet	0.063 ± 0.011	0.267 ± 0.012	0.051 ± 0.008	-0.062 ± 0.021
RNA3DCNN	0.080 ± 0.018	0.193 ± 0.052	0.095 ± 0.024	-0.011 ± 0.008
EquiRNA	0.069 ± 0.016	0.157 ± 0.054	0.080 ± 0.026	0.029 ± 0.018
lociPARSE	0.144 ± 0.012	0.335 ± 0.014	0.157 ± 0.017	0.044 ± 0.023
HomRank	0.159 ± 0.017	0.298 ± 0.033	0.229 ± 0.023	0.092 ± 0.014
SIRGE	$\mathbf{0.270} \pm 0.011$	$\mathbf{0.453} \pm 0.017$	$\mathbf{0.347} \pm 0.010$	$\mathbf{0.209} \pm 0.007$

where ℓ_{BCE} is binary cross-entropy and $\mathcal{W}$ contains the supervised candidate–pair triples. Summing the directed states makes the prediction independent of pair order.
Supervising both contact sets penalises contacts a candidate misses as well as contacts it adds. These are the pair representations through which the RiNALMo-derived sequence prior enters SIRGE (Section 3.2). The objective therefore trains the model to judge a candidate's tertiary contacts, and hence its helix packing, against what the sequence leads it to expect: the sequence–geometry alignment defined in Section 3.1.

The complete objective is $\mathcal{L} = \mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{local}}\mathcal{L}_{\mathrm{local}} + \lambda_{\mathrm{contact}}\mathcal{L}_{\mathrm{contact}}$. Native structures supply training labels only: at inference, SIRGE scores a candidate from its coordinates and sequence alone, and both auxiliary heads are discarded.

4 Experimental Validation

Dataset and experimental setup We train SIRGE on the training split of Li et al. (2025a): 131 non-redundant, single-chain RNAs whose candidates come from molecular-dynamics simulations and FARFAR2 (Appendix C). We compare it with six evaluators spanning a range of inductive biases and training objectives. Every method, SIRGE included, is trained for five epochs on the same data, with the checkpoint selected on a held-out validation cluster, and is reported as mean $\pm$ standard deviation over three runs. To isolate sequence–geometry alignment, we also train SIRGE-base, which shares SIRGE's architecture, data and its two ranking objectives, over whole candidates and over nucleotides, but lacks the two components of alignment (Section 3.1): the sequence prior and the contact objective. We evaluate SIRGE on two held-out datasets, representing two complementary distribution shifts: CLS-1 (Li et al., 2025a) changes sequence length and candidate quality while keeping the training generators (Section 4.1), whereas Gen-OOD (Dataset 3 of Martinović et al. (2024)) draws candidates from six unseen deep-learning predictors (Section 4.2). We then ablate each component (Section 4.3) and examine the signal behind SIRGE's rankings (Section 4.4). Dataset, baseline and evaluation details are in Appendices C and D.

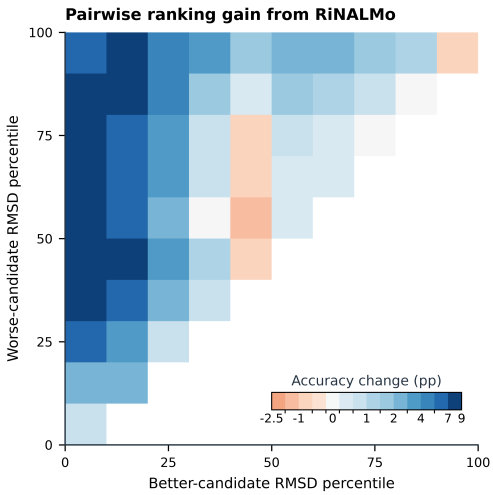


Figure 2: Pairwise-accuracy change on CLS-1 of SIRGE vs SIRGE-base. Blue indicates a gain from conditioning.

4.1 Generalisation across sequence length and candidate quality

We first test generalisation under two shifts from training: longer sequences, which admit more possible nucleotide interactions, and more varied candidate quality (Figure 10). The CLS-1 dataset was specifically designed to introduce both while keeping the training generators: molecular dynamics, which evolves atomic coordinates under a force field, and FARFAR2, which assembles candidates from fragments of solved RNA structures. Our primary metric, global Kendall-τ alignment

Table 2: Top-3 ranking and top-k accuracy on CLS-1, reported as mean $\pm$ std over three seeds.

Model	Top-3	Hit@10	Hit@50	Hit@100
ARES	0.257 ± 0.009	0.270 ± 0.027	0.397 ± 0.073	0.714 ± 0.082
PAMNet	0.212 ± 0.002	0.270 ± 0.055	0.333 ± 0.048	0.397 ± 0.073
RNA3DCNN	0.216 ± 0.001	0.333 ± 0.048	0.460 ± 0.027	0.571 ± 0.048
EquiRNA	0.233 ± 0.012	0.254 ± 0.073	0.444 ± 0.120	0.635 ± 0.099
lociPARSE	0.254 ± 0.008	0.270 ± 0.027	0.413 ± 0.073	0.524 ± 0.048
HomRank	0.265 ± 0.003	0.476 ± 0.095	0.619 ± 0.095	0.746 ± 0.027
SIRGE	$\mathbf{0.308} \pm 0.013$	$\mathbf{0.571} \pm 0.067$	$\mathbf{0.738} \pm 0.034$	$\mathbf{0.810} \pm 0.027$

Table 3: Kendall–τ alignment, top-1 retrieval and top-3 ranking accuracy on Gen-OOD

Model	Kendall–τ	Hit@1	Top-3
ARES	-0.171 ± 0.043	0.111 ± 0.007	0.179 ± 0.031
PAMNet	-0.160 ± 0.019	0.095 ± 0.031	0.179 ± 0.031
RNA3DCNN	-0.108 ± 0.018	0.095 ± 0.012	0.179 ± 0.021
EquiRNA	-0.059 ± 0.009	0.118 ± 0.020	0.183 ± 0.023
lociPARSE	-0.019 ± 0.021	0.135 ± 0.036	0.206 ± 0.027
HomRank	-0.198 ± 0.008	0.091 ± 0.018	0.190 ± 0.012
SIRGE	$\mathbf{0.110} \pm 0.054$	$\mathbf{0.246} \pm 0.018$	$\mathbf{0.317} \pm 0.041$

$-\tau(f_\theta, \mathrm{RMSD})$ over each RNA's full evaluation pool, tests whether SIRGE's quality features stay informative across levels of folding error. We also report it on sampled ten-candidate lists, stratified by lowest RMSD into D1 (2–5 Å), D2 (5–10 Å) and D3 (10–15 Å); each band covers only imperfect structures, the regime relevant to computational RNA design, and exposes failures that the full-pool average could hide. SIRGE outperforms every baseline globally and in all three bands (Table 1), and its relative lead grows as candidates become less accurate: in D3, it more than doubles the best baseline. Figure 2 isolates the part of this advantage due to sequence–geometry alignment: relative to SIRGE-base, SIRGE orders candidate pairs more accurately, most of all when the better candidate is among the best in the pool. Because evaluators are used in practice to shortlist candidates for costly downstream steps, Table 2 also reports two selection metrics: Top-3 ranking accuracy, whether the three lowest-RMSD candidates in each sampled list are ordered correctly, and Hit@k ($k \in \{10, 50, 100\}$), the fraction of RNAs whose lowest-RMSD candidate appears among the k highest-scored candidates. SIRGE leads on both, so its ranking advantage carries over to selection.

4.2 Selecting the top candidates amongst modern deep learning generators

In practice, an evaluator must choose among the outputs of modern deep-learning predictors, none of which generated our training candidates. Gen-OOD tests this setting directly: 84 RNAs, each with one candidate from each of AlphaFold 3, DRfold, DeepFoldRNA, RhoFold, RoseTTAFoldNA and trRosettaRNA (504 in total). Three properties suit it to this test. First, it was built to benchmark predictors on RNAs they had not seen: every RNA postdates a conservative training cutoff and has no close sequence relative from before that cutoff. Its candidates are therefore genuine predictions, with the errors predictors make in practice, rather than recollections of training structures. Second, no predictor dominates: each supplies the best structure for some RNAs, and the benchmark's authors found that established scoring functions could not reliably pick it out. The choice must therefore be made RNA by RNA, which is exactly an evaluator's task. Selection here is greedy: the aim is to pick the best prediction, not to order the poor ones, so Kendall-τ alignment is reported for completeness only. Our main metric is Hit@1, whether the top-scored candidate is the best of the six; with six candidates per RNA, the larger shortlists used for CLS-1 would be trivial. Because RMSD is dominated by poorly modelled regions (Section 3.3), small differences between the best candidates need not reflect real differences in quality. To avoid resting our conclusions on a single pick, we also report Top-3 ranking accuracy, treating RMSD differences below 0.1 Å as ties. SIRGE leads on both selection metrics and is the only evaluator whose ranking agrees with candidate quality (Table 3).

4.3 Ablation study

Table 4 ablates SIRGE's components on CLS-1. *No-ct* drops the contact objective, *No-cond* drops the sequence prior, *No-loc* drops the per-nucleotide lDDT objective, *Perm* keeps the prior but randomly shuffles its nucleotide positions during training, so the map keeps its values but no longer matches the sequence; SIRGE-base (Section 4) drops both the prior and the contact objective. Every component helps, and removing both alignment components costs the most (-19.6%, against -17.7% for No-loc). Their effects are not additive: relative to SIRGE-base, the prior alone adds 0.028 and the contact objective alone 0.013, yet together they add 0.053, more than the sum of their parts. The contact objective is thus worth almost twice as much when the prior is present, which is what we expect if the prior supplies the expectation and the contact objective the test (Section 3.1).

Table 4: Ablation of SIRGE on CLS-1. Results are reported as mean $\pm$ std over three seeds; relative decrease w.r.t. SIRGE is in brackets.

Model	Kendall–τ
SIRGE	$\mathbf{0.270} \pm 0.011$
SIRGE-base	0.217 ± 0.018 (-19.6%)
No-ct	0.245 ± 0.008 (-9.3%)
No-cond	0.230 ± 0.005 (-14.8%)
No-loc	0.222 ± 0.013 (-17.7%)
Perm	0.232 ± 0.017 (-14.1%)

4.4 Anatomy of the sequence–geometry alignment signal

Having shown that the sequence prior and contact objective improve ranking, we now show empirically that this gain comes from sequence–geometry alignment: SIRGE judges candidates by their helix packing and, as a result, organises them by quality in its learned representation. Alignment should matter most for long RNAs, whose candidates tend to share correct stems but differ in packing (Section 3.1), and SIRGE's gain over SIRGE-base indeed grows with RNA length across CLS-1 (Figure 3). The 150-nucleotide split is the median length of CLS-1 and matches the length band beyond which prior benchmarks report the sharpest loss of predictor accuracy (Martinović et al., 2024). We therefore study one such RNA in detail: the 154-nucleotide specificity domain of *Bacillus subtilis* RNase P (PDB 1NBS), whose fold is held together by long-range helix packing. Its molecular-dynamics candidates (4–16.5 Å) keep the native stems but differ in how they pack, making 1NBS representative of long RNAs and a direct test of sequence–geometry alignment.

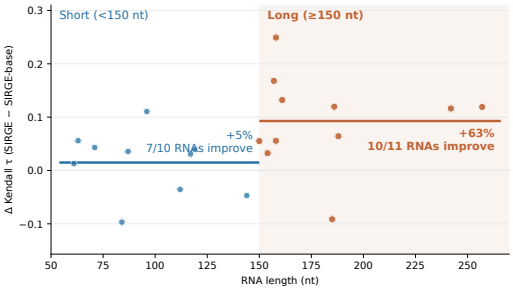


Figure 3: Kendall-τ alignment gain of SIRGE over SIRGE-base for each CLS-1 RNA, against sequence length. Mean over 3 seeds. Lines show the mean gain for RNAs of shorter than and at least 150 nucleotides; $+x\%$ represents SIRGE improvement compared to SIRGE-base.

Figure 4 compares candidate representations with and without the sequence prior and contact objective. SIRGE-base already captures how the candidates vary: its embedding traces a continuous path through the pool, consistent with conformations sampled along the dynamics trajectories. Yet it lacks direction: candidates equally far from the native lie at many points along the path, so geometry alone tells the candidates apart without telling which is closer to the native. SIRGE's embedding follows a path of the same kind, but RMSD now falls steadily along it towards the native. Alignment does not create this structure, but rather it orients it, turning variation in the helix packing into an axis of quality. The rank-density maps show the consequence for selection: SIRGE's ranks agree with the true order most tightly among the best candidates, whereas for the other three models predicted rank says little about quality. This orientation has a biological basis. The 1NBS candidates keep the native stems (Figure 5a) but lose tertiary contacts as RMSD grows (Figure 5b), so their quality is a matter of helix packing. How stems pack is itself encoded in the sequence (Yesselman et al., 2019), and the prior brings this information into SIRGE from sequence alone: it recovers most native base pairs on the typical held-out RNA and detects tertiary contacts even between nucleotides 96 or more positions apart, where simple controls fall to chance (Figure 6; Appendix B). The prior need not predict the fold by itself, only supply the reference against which SIRGE judges each candidate's packing. Accordingly, SIRGE-base, without the prior or the contact objective, largely ignores packing

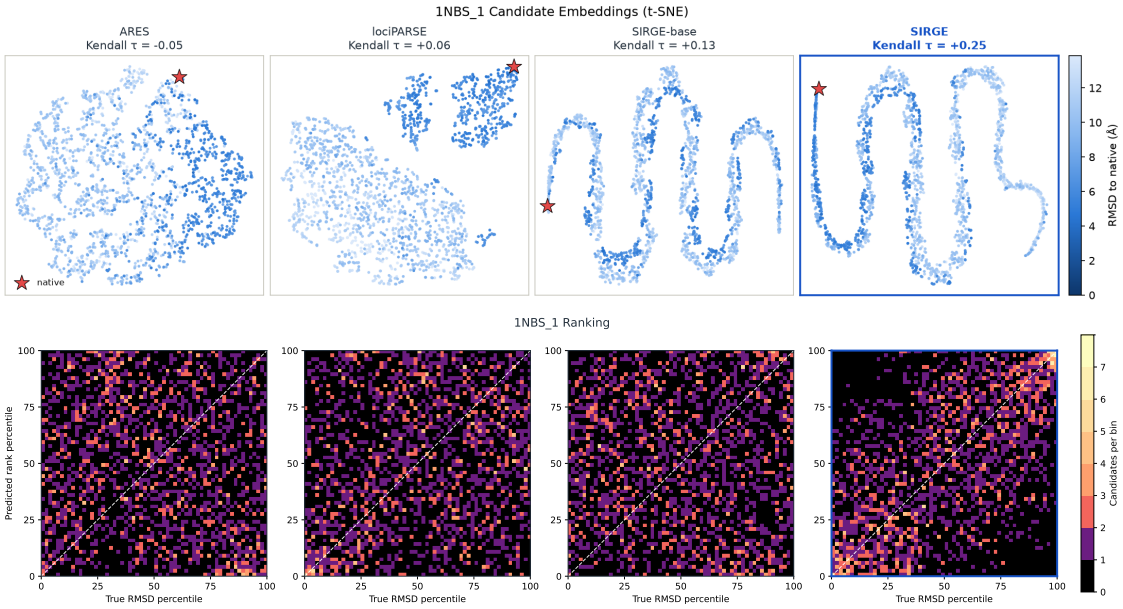
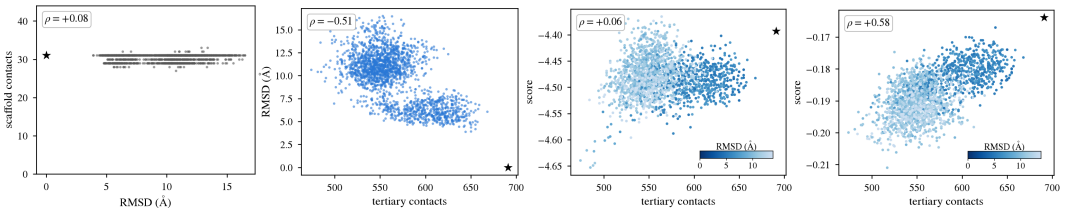


Figure 4: The 1NBS candidate pool as organised by four evaluators. **Top**: t-SNE of candidate embeddings, coloured by RMSD to the native (star); titles give Kendall-τ alignment mean over 3 seeds. **Bottom**: density of predicted rank percentile against true RMSD percentile (0 is best); mass on the dashed diagonal indicates agreement.



(a) Stem contacts (b) Tertiary contacts (c) SIRGE-base (d) SIRGE

Figure 5: Tertiary packing, not base pairing, separates good from bad 1NBS candidates. Each dot is a candidate and the star is the native; ρ denotes Spearman's rank correlation, and contacts are defined as in Section 3. (a) Stem contacts remain constant across the pool. (b) Tertiary contacts decrease as RMSD increases. (c) SIRGE-base scores are largely unrelated to tertiary contacts. (d) SIRGE scores increase with them. In (c,d), colour denotes RMSD to the native.

(Figure 5c), whereas SIRGE's score rises with tertiary contacts (Figure 5d). For long RNAs, whose candidates cannot be reliably ordered by geometry alone, SIRGE therefore ranks candidates by how well their packing matches what their sequence implies, which is precisely where blind predictions go wrong. More generally, these results suggest that for long RNAs geometric plausibility is necessary but not sufficient: ranking their candidates requires judging helix packing against a sequence-specific reference, which a pretrained RNA LLM can supply.

5 Conclusion

We introduced SIRGE, an RNA structure evaluator that frames candidate ranking as sequence–geometry alignment: whether a candidate's helix packing agrees with the interactions its sequence implies. Attention maps from a frozen RNA language model supply this expectation, and a contact objective teaches SIRGE to test each candidate against it. SIRGE outperforms established evaluators under shifts in sequence length, candidate quality and candidate generator, and alignment helps most on long RNAs, where geometry alone cannot tell which plausible packing the sequence adopts.

AI USE STATEMENT

We used large language models (Claude, including Claude Code, and ChatGPT) as general-purpose assistants. Of the tasks for which the ICLR 2027 policy requires disclosure, the only one they helped with was implementing our methods: the authors designed the methods and directed the implementation, and LLMs helped write the code. Beyond that, they helped create and edit software code, draft parts of the paper and edit it for readability, summarize existing literature, identify relevant literature, search for information, and brainstorm. The authors reviewed and checked LLM outputs and take full responsibility for the contents of this paper.

ETHICS STATEMENT

This work involves no human subjects or personal data. All structures come from the Protein Data Bank and from publicly released benchmarks, used under their original terms. We see no direct path from ranking RNA structure candidates to harmful applications.

REPRODUCIBILITY STATEMENT

The architecture is specified in full in Appendix A, the datasets and splits in Appendix C, and the baselines, training settings, checkpoint selection and evaluation metrics in Appendix D. Code and test datasets are provided through an anonymised repository at https://anonymous.4open.science/r/SIRGE-FB53/README.md.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

Akash Bahai, Chee Keong Kwoh, Yuguang Mu, and Yinghui Li. Systematic benchmarking of deep-learning methods for tertiary RNA structure prediction. *PLOS Computational Biology*, 20 (12):e1012715, 2024.

Philippe Brion and Eric Westhof. Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, 2007.

Jamie H Cate, Anne R Gooding, Elaine Podell, Kaihong Zhou, Barbara L Golden, Craig E Kundrot, Thomas R Cech, and Jennifer A Doudna. Crystal structure of a group i ribozyme domain: principles of RNA packing. *Science*, 273(5282):1678–1685, 1996.

Thomas R Cech and Joan A Steitz. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell*, 157(1):77–94, 2014.

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. *BioRxiv*, pp. 2022–08, 2022.

Alexander Churkin, Matan Drory Retwitzer, Vladimir Reinharz, Yann Ponty, Jérôme Waldispühl, and Danny Barash. Design of RNAs: comparing programs for inverse RNA folding. *Briefings in bioinformatics*, 19(2):350–358, 2018.

Elizabeth A Doherty, Robert T Batey, Benoît Masquida, and Jennifer A Doudna. A universal mode of helix packing in RNA. *Nature structural biology*, 8(4):339–343, 2001.

Stanley D Dunn, Lindi M Wahl, and Gregory B Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3): 333–340, 2008.

Christine E Hajdin, Feng Ding, Nikolay V Dokholyan, and Kevin M Weeks. On the significance of an RNA tertiary structure prediction. *RNA*, 16(7):1340–1349, 2010.

Lei Jin and Shi-Jie Chen. Computational modeling of cotranscriptional RNA folding. *Computational and Structural Biotechnology Journal*, 2025.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873):583–589, 2021.

Jun Li, Wei Zhu, Jun Wang, Wenfei Li, Sheng Gong, Jian Zhang, and Wei Wang. RNA3dcnn: Local and global quality assessments of RNA 3d structures using 3d deep convolutional neural networks. *PLoS computational biology*, 14(11):e1006514, 2018.

Liang Li, Yinghua Yao, Ivona Martinović, Mandar Kulkarni, Yuangang Pan, Roland G Huber, and Mile Šikić. Homrank: Homogeneous RNA ranking for 3d structure evaluation. *bioRxiv*, pp. 2025–08, 2025a.

Zongzhao Li, Jiacheng Cen, Wenbing Huang, Taifeng Wang, and Le Song. Size-generalizable RNA structure evaluation by exploring hierarchical geometries. In *International Conference on Learning Representations*, volume 2025, pp. 42496–42521, 2025b.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations*, volume 2024, pp. 39282–39309, 2024.

Xiaocheng Liu, Wenkai Wang, Zongyang Du, Ziyi Wang, Zhenling Peng, and Jianyi Yang. Quality assessment of RNA 3d structure models using deep learning and intermediate 2d maps. *Communications Biology*, 9(1):293, 2026.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Ivona Martinović, Tin Vlašić, Yang Li, Bryan Hooi, Yang Zhang, and Mile Šikić. A comparative review of deep learning methods for RNA tertiary structure prediction. *bioRxiv*, pp. 2024–11, 2024.

Zhichao Miao, Ryszard W Adamiak, Maciej Antczak, Michał J Boniecki, Janusz Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, Yi Cheng, Fang-Chieh Chou, Rhiju Das, et al. RNA-puzzles round iv: 3d structure predictions of four ribozymes and two aptamers. *RNA*, 26(8):982–995, 2020.

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.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI conference on artificial intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Almudena Ponce-Salvatierra, Astha, Katarzyna Merdas, Chandran Nithin, Pritha Ghosh, Sunandan Mukherjee, and Janusz M Bujnicki. Computational modeling of RNA 3d structure based on experimental data. *Bioscience reports*, 39(2):BSR20180430, 2019.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *Biorxiv*, pp. 2020–12, 2020.

Jean-Paul Renaud, Ashwin Chari, Claudio Ciferri, Wen-ti Liu, Hervé-William Rémigy, Holger Stark, and Christian Wiesmann. Cryo-em in drug discovery: achievements, limitations and prospects. *Nature reviews Drug discovery*, 17(7):471–492, 2018.

Ya-Lan Tan, Chen-Jie Feng, Lei Jin, Ya-Zhou Shi, Wenbing Zhang, and Zhi-Jie Tan. What is the best reference state for building statistical potentials in RNA 3d structure evaluation? *RNA*, 25(7): 793–812, 2019.

Sumit Tarafder and Debswapna Bhattacharya. lociparse: a locality-aware invariant point attention model for scoring RNA 3d structures. *Journal of Chemical Information and Modeling*, 64(22): 8655–8664, 2024.

Ignacio Tinoco Jr and Carlos Bustamante. How RNA folds. *Journal of molecular biology*, 293(2): 271–281, 1999.

Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, Ramya Rangan, Masha Karelina, Rhiju Das, and Ron O Dror. Geometric deep learning of RNA structure. *Science*, 373(6558): 1047–1051, 2021.

Jian Wang, Yunjie Zhao, Chunyan Zhu, and Yi Xiao. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3d RNA structures. *Nucleic acids research*, 43(10):e63–e63, 2015.

Andrew Martin Watkins, Ramya Rangan, and Rhiju Das. Farfar2: improved de novo rosetta prediction of complex global RNA folds. *Structure*, 28(8):963–976, 2020.

Xiaojun Xu and Shi-Jie Chen. Physics-based RNA structure prediction. *Biophysics reports*, 1(1): 2–13, 2015.

Joseph D Yesselman, Sarah K Denny, Namita Bisaria, Daniel Herschlag, William J Greenleaf, and Rhiju Das. Sequence-dependent RNA helix conformational preferences predictably impact tertiary structure formation. *Proceedings of the National Academy of Sciences*, 116(34):16847–16855, 2019.

Shuo Zhang, Yang Liu, and Lei Xie. A universal framework for accurate and efficient geometric deep learning of molecular systems. *Scientific Reports*, 13(1):19171, 2023.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

A SIRGE ARCHITECTURE

This appendix specifies the architecture summarised in Section 3. SIRGE combines a frozen sequence prior with two learned representations: nucleotide features $\mathbf{h}_{m,i}^{(b)} \in \mathbb{R}^{d_h}$ and directed pair features $\mathbf{z}_{m,ij}^{(b)} \in \mathbb{R}^{d_z}$. We use $d_h = 64$, $d_z = 16$, and $B = 4$ nucleotide attention blocks. The superscript b indexes these blocks; ℓ is reserved for the geometric order of the atom features. Atomic coordinates and nucleotide frames are fixed during evaluation.

A.1 ATOM GRAPH AND GEOMETRIC ENCODER

The atom graph keeps heavy atoms only; $\mathbf{c}_{m,a}$ is a one-hot over C, N, O, P and S. Nucleotides are indexed in sequence order, so $\rho_m(a)$ and the prior A address the same positions. Each atom attends to its 32 nearest atoms, itself included.

The encoder is a compact EquiformerV2 (Liao et al., 2024): three blocks of equivariant attention and gated feed-forward layers, four heads each, with $C_0 = 32$ scalar, $C_1 = 8$ vector and $C_2 = 4$ second-order channels, written as $\mathbf{a}_{m,a}^{(\ell)} \in \mathbb{R}^{(2\ell+1)\times C_\ell}$. Edge distances are encoded by 16 Gaussians with centres spanning 0 to 30 Å and no hard cutoff.

A.2 NUCLEOTIDE FRAMES AND RIGID-MOTION INVARIANCE

Frame construction. The frame $F_{m,i} = (\mathbf{R}_{m,i}, \mathbf{o}_{m,i})$ of nucleotide i in candidate X_m is built from three of its atoms: C4′, a carbon of the sugar ring; P, the phosphorus atom of the phosphate group on its 5′ side, which links it to nucleotide $i-1$; and the nitrogen through which the base bonds to the sugar, N9 for A and G and N1 for C and U. Writing $\mathbf{x}_{m,i}^{\mathrm{C4'}}$, $\mathbf{x}_{m,i}^{\mathrm{P}}$ and $\mathbf{x}_{m,i}^{\mathrm{N}}$ for their coordinates, the origin is $\mathbf{o}_{m,i} = \mathbf{x}_{m,i}^{\mathrm{C4'}}$ and the displacements of the other two atoms from it are

$$\boldsymbol{\Delta}_{m,i}^{\mathrm{P}} = \mathbf{x}_{m,i}^{\mathrm{P}} - \mathbf{o}_{m,i}, \qquad \boldsymbol{\Delta}_{m,i}^{\mathrm{N}} = \mathbf{x}_{m,i}^{\mathrm{N}} - \mathbf{o}_{m,i}. \tag{9}$$

The first nucleotide of a chain, and of each segment after a chain break, has no phosphate of its own; for it, $\mathbf{x}_{m,i}^{\mathrm{P}}$ denotes instead the coordinate of C1′, the sugar carbon bonded to the base. One Gram–Schmidt step turns the two displacements into orthonormal axes,

$$\begin{aligned} \mathbf{e}_{m,i}^{(1)} &= \frac{\boldsymbol{\Delta}_{m,i}^{\mathrm{P}}}{\|\boldsymbol{\Delta}_{m,i}^{\mathrm{P}}\|_2}, \qquad \boldsymbol{\Delta}_{m,i}^{\mathrm{N}\perp} = \boldsymbol{\Delta}_{m,i}^{\mathrm{N}} - \big((\mathbf{e}_{m,i}^{(1)})^\top \boldsymbol{\Delta}_{m,i}^{\mathrm{N}}\big)\mathbf{e}_{m,i}^{(1)}, \\ \mathbf{e}_{m,i}^{(2)} &= \frac{\boldsymbol{\Delta}_{m,i}^{\mathrm{N}\perp}}{\|\boldsymbol{\Delta}_{m,i}^{\mathrm{N}\perp}\|_2}, \qquad \mathbf{e}_{m,i}^{(3)} = \mathbf{e}_{m,i}^{(1)} \times \mathbf{e}_{m,i}^{(2)}. \end{aligned} \tag{10}$$

The first axis points from the origin towards P. The second is the direction from the origin towards the base with its component along the first axis removed, so the two are orthogonal. The cross product completes a right-handed triple. The matrix $\mathbf{R}_{m,i} = [\mathbf{e}_{m,i}^{(1)}, \mathbf{e}_{m,i}^{(2)}, \mathbf{e}_{m,i}^{(3)}]$ therefore has orthonormal columns and determinant $+1$, so it is a rotation: $\mathbf{R}_{m,i}$ maps local coordinates to global ones and $\mathbf{R}_{m,i}^\top$ the reverse, which is how it enters equation 1. The construction requires $\boldsymbol{\Delta}_{m,i}^{\mathrm{P}}$ and $\boldsymbol{\Delta}_{m,i}^{\mathrm{N}}$ to be nonzero and non-collinear, which holds for every nucleotide of every structure used in this work.

Effect of a global rigid motion. Let $\mathbf{Q} \in \mathrm{SO}(3)$ be a rotation and $\mathbf{t} \in \mathbb{R}^3$ a translation applied to the whole candidate, $\mathbf{x}'_{m,a} = \mathbf{Q}\mathbf{x}_{m,a} + \mathbf{t}$, where a prime marks a transformed quantity. Coordinate differences drop the translation, so both displacements in equation 9 become $\mathbf{Q}\boldsymbol{\Delta}$. Rotation preserves norms and inner products, so the Gram–Schmidt step in equation 10 commutes with it, and $\det \mathbf{Q} = +1$ gives $(\mathbf{Q}\mathbf{e}_{m,i}^{(1)}) \times (\mathbf{Q}\mathbf{e}_{m,i}^{(2)}) = \mathbf{Q}(\mathbf{e}_{m,i}^{(1)} \times \mathbf{e}_{m,i}^{(2)})$. The frames therefore move with the candidate,

$$\mathbf{o}'_{m,i} = \mathbf{Q}\mathbf{o}_{m,i} + \mathbf{t}, \qquad \mathbf{R}'_{m,i} = \mathbf{Q}\mathbf{R}_{m,i}. \tag{11}$$

A reflection would flip $\mathbf{e}_{m,i}^{(3)}$, so SIRGE is invariant to rigid motions but not to reflections, as an evaluator of chiral molecules should be. Interatomic distances are preserved, so the 32-nearest-neighbour graph is unchanged; the encoder sees only displacements and is equivariant by construction,

so $\mathbf{a}'^{(0)}_{m,a} = \mathbf{a}^{(0)}_{m,a}$ and $\mathbf{a}'^{(1)}_{m,a} = \mathbf{Q}\mathbf{a}^{(1)}_{m,a}$, the rotation acting on each column. Writing $\mathbf{I}_3$ for the 3×3 identity and using $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_3$, the two projections in equation 1 obey

$$\begin{aligned}(\mathbf{R}'_{m,i})^\top \mathbf{a}'^{(1)}_{m,a} &= \mathbf{R}^\top_{m,i}\mathbf{Q}^\top\mathbf{Q}\mathbf{a}^{(1)}_{m,a} = \mathbf{R}^\top_{m,i}\mathbf{a}^{(1)}_{m,a},\\ (\mathbf{R}'_{m,i})^\top(\mathbf{x}'_{m,a} - \mathbf{o}'_{m,i}) &= \mathbf{R}^\top_{m,i}\mathbf{Q}^\top\mathbf{Q}(\mathbf{x}_{m,a} - \mathbf{o}_{m,i}) = \mathbf{R}^\top_{m,i}(\mathbf{x}_{m,a} - \mathbf{o}_{m,i}),\end{aligned} \tag{12}$$

so $\mathbf{d}'_{m,a} = \mathbf{d}_{m,a}$. Pooling (Appendix A.3) reads only $\mathbf{d}_{m,a}$ and the base identity, so $\mathbf{h}^{(0)}_{m,i}$ is invariant.

The pair descriptor (Appendix A.5) is invariant by the same cancellation: $\delta'_{m,ij} = \|\mathbf{Q}\mathbf{r}_{m,ij}\|_2 = \delta_{m,ij}$, and $\widehat{\mathbf{r}}'_{m,ij}$ and $\mathbf{O}'_{m,ij} = \mathbf{R}^\top_{m,i}\mathbf{Q}^\top\mathbf{Q}\mathbf{R}_{m,j}$ reduce as in equation 12, while sequence separation and the prior a_{ij} do not depend on coordinates. FiLM and the pair MLP therefore give an invariant $\mathbf{z}^{(0)}_{m,ij}$. In each IPA block (Appendix A.6; block and head superscripts omitted), the scalar logit term and the pair bias are functions of invariant features, and a learned point $\mathbf{p}_{m,j}$ of nucleotide j, which sits at $\mathbf{R}_{m,j}\mathbf{p}_{m,j} + \mathbf{o}_{m,j}$ in global coordinates, moves with the candidate, so the point distances $D_{m,ij}$ and with them the attention weights $\alpha_{m,ij}$ are unchanged. The aggregated scalar values are invariant. For the point output, the weights sum to one, so the translation enters once with each sign and cancels,

$$\sum_{j=1}^{T}\alpha_{m,ij}\big(\mathbf{R}'_{m,j}\mathbf{p}_{m,j} + \mathbf{o}'_{m,j}\big) - \mathbf{o}'_{m,i} = \mathbf{Q}\Big[\sum_{j=1}^{T}\alpha_{m,ij}\big(\mathbf{R}_{m,j}\mathbf{p}_{m,j} + \mathbf{o}_{m,j}\big) - \mathbf{o}_{m,i}\Big], \tag{13}$$

and $(\mathbf{R}'_{m,i})^\top = \mathbf{R}^\top_{m,i}\mathbf{Q}^\top$ removes the remaining $\mathbf{Q}$ when the output is returned to the receiving frame. Hence $\mathbf{h}^{(b+1)}_{m,i}$ and $\mathbf{z}^{(b+1)}_{m,ij}$ are invariant whenever $\mathbf{h}^{(b)}_{m,i}$ and $\mathbf{z}^{(b)}_{m,ij}$ are. By induction over the B blocks, so are $\mathbf{h}^{(B)}_{m,i}$, the pooled representation of Appendix A.7, and the score $f_\theta(X_m, S)$.

A.3 Attention pooling within each nucleotide

The descriptor $\mathbf{d}_{m,a}$ of equation 1 has 59 entries: 32 scalar channels, the 24 components of the eight local vectors, and the three local coordinates. An MLP of widths $59 \to 128 \to 64$ with SiLU activation maps it to a per-atom feature, and four attention-pooling heads, each with a learned query and 16-dimensional keys and values, pool these features over the atoms of each nucleotide. The four head outputs are concatenated with the one-hot base identity (A, C, G, U) and projected by an affine map from 68 to 64 dimensions to give $\mathbf{h}^{(0)}_{m,i}$. Pooling is confined to the atoms of one nucleotide, but their encoder features already carry information from neighbouring atoms, including those of other nucleotides.

A.4 The frozen sequence prior

Head selection. We use RiNALMo-giga (650M parameters) (Penić et al., 2025) and probe each of its 660 attention heads (33 blocks of 20 heads) for base-pairing signal on the held-out validation RNAs. Base pairs are annotated on the native by the geometric rule of Appendix A.8; 21 of the 23 RNAs have at least one canonical pair and enter the average. For one head and one RNA, the nucleotide pairs with $|i - j| \geq 3$ are ranked by the symmetrised, average-product-corrected map of equation 14, and the score is the precision among the n_{bp} top-ranked pairs, where n_{bp} is the number of annotated base pairs of that RNA, so that precision equals recall. Three heads stand apart (Table 5): $(32, 13)$, $(32, 1)$ and $(33, 12)$, written as (block, head) with blocks numbered from 1 and heads from 0, reach 0.52, 0.52 and 0.50, against 0.09 for the next head and a mean of 0.02 over all heads. We keep these three. The prior therefore needs no structural labels at test time. Its heads were chosen with labels from the validation split, meaning the test splits played no part.

Preprocessing ψ. For a chain of length T, let $A^{\mathrm{raw}}_k \in \mathbb{R}^{T\times T}$ be the attention map of selected head $k \in \{1, 2, 3\}$ with the start and end tokens removed. It is symmetrised, corrected for positions that attend broadly by average-product correction (Dunn et al., 2008; Rao et al., 2020), and standardised:

$$A^{\mathrm{sym}}_k = A^{\mathrm{raw}}_k + (A^{\mathrm{raw}}_k)^\top, \quad A^{\mathrm{apc}}_{ij,k} = A^{\mathrm{sym}}_{ij,k} - \frac{\overline{A}^{\mathrm{sym}}_{i\cdot,k}\,\overline{A}^{\mathrm{sym}}_{\cdot j,k}}{\overline{A}^{\mathrm{sym}}_{\cdot\cdot,k}}, \quad A_{ij,k} = \begin{cases}\dfrac{A^{\mathrm{apc}}_{ij,k} - \mu_k}{\sigma_k}, & |i-j| \geq 3,\\ 0, & \text{otherwise,}\end{cases} \tag{14}$$

Table 5: Head selection. Base-pair precision at n_{bp}, averaged over the 21 annotated validation RNAs, for the ten best RiNALMo-giga heads and the mean over all 660 heads. Retained heads are marked.

Head (block, head)	Precision at n_{bp}
(32, 13) ✓	0.522
(32, 1) ✓	0.517
(33, 12) ✓	0.498
(32, 9)	0.092
(10, 0)	0.085
(31, 5)	0.079
(1, 2)	0.075
(14, 0)	0.074
(33, 5)	0.073
(19, 8)	0.070
Mean over all heads (median)	0.023 (0.018)

where the bars denote the row, column and overall means of A_k^{sym}, and μ_k, σ_k are the mean and standard deviation of A_k^{apc} over the entries with $|i-j| \geq 3$ of that chain. Zeroing the near-diagonal entries stops the prior from emphasising sequence neighbours, whose separation the pair descriptor already supplies. Where a structure has more than one chain, each chain is processed separately and A is block-diagonal, with zeros between chains. Entries of A can be negative as they are features, not probabilities of pairing or contact.

A.5 Pair geometry and sequence conditioning

The descriptor of the ordered pair (i, j) describes nucleotide j as seen from the frame of nucleotide i. With $\mathbf{r}_{m,ij} = \mathbf{o}_{m,j} - \mathbf{o}_{m,i}$,

$$\delta_{m,ij} = \|\mathbf{r}_{m,ij}\|_2, \qquad \widehat{\mathbf{r}}_{m,ij} = \frac{\mathbf{R}_{m,i}^{\top}\mathbf{r}_{m,ij}}{\delta_{m,ij}} \quad (\mathbf{0} \text{ for } i = j), \qquad \mathbf{O}_{m,ij} = \mathbf{R}_{m,i}^{\top}\mathbf{R}_{m,j} \tag{15}$$

are the distance between the two origins, the unit direction towards j in the frame of i, and the rotation of frame j relative to frame i. The descriptor is

$$\mathbf{g}_{m,ij} = \Big[\mathrm{emb}\big(\mathrm{clip}_{[-32,32]}(i-j)\big),\ \mathrm{rbf}(\delta_{m,ij}),\ \delta_{m,ij}/\delta_{\max},\ \widehat{\mathbf{r}}_{m,ij},\ \mathrm{vec}_6(\mathbf{O}_{m,ij})\Big] \in \mathbb{R}^{42}, \tag{16}$$

where emb is a learned 16-dimensional embedding of the signed sequence separation, clipped to $[-32, 32]$; rbf encodes the distance with 16 Gaussians centred evenly on $[0, \delta_{\max}]$ with $\delta_{\max} = 32$ Å, and the scalar $\delta_{m,ij}/\delta_{\max}$ keeps the descriptor sensitive beyond the last centre; and vec_6 takes the first two columns of the rotation matrix, which determine the third. Although $a_{ij} = a_{ji}$, the descriptor is directed, $\mathbf{g}_{m,ij} \neq \mathbf{g}_{m,ji}$ in general, and so are the pair states built from it.

The FiLM maps $\boldsymbol{\gamma}, \boldsymbol{\beta}$ of equation 2 are affine, $\mathbb{R}^3 \to \mathbb{R}^{42}$, and initialised to zero, so training starts from the unconditioned descriptor. The pair MLP ϕ_{pair} has widths $42 \to 32 \to 16$ with SiLU activation.

A.6 Invariant point attention and pair updates

The $B = 4$ blocks follow the invariant point attention (IPA) of AlphaFold2 (Jumper et al., 2021) with two differences: the nucleotide frames are fixed, since SIRGE evaluates a structure rather than refines it, and the pair states enter only through the attention bias, not as aggregated values. Each block has $H = 4$ heads with $d_c = 16$ scalar channels and $P = 4$ query, key and value points per head, and attends over all T^2 ordered pairs, self-pairs included. For block b and head r, affine maps of $\mathbf{h}_{m,i}^{(b)}$ produce scalar queries, keys and values $\mathbf{q}_{m,i}^{(b,r)}, \mathbf{k}_{m,i}^{(b,r)}, \mathbf{v}_{m,i}^{(b,r)} \in \mathbb{R}^{d_c}$ and local points $\mathbf{p}_{m,i,u}^{q,(b,r)}, \mathbf{p}_{m,i,u}^{k,(b,r)}, \mathbf{p}_{m,i,u}^{v,(b,r)} \in \mathbb{R}^3$ for $u \in \{1, \ldots, P\}$, learned locations in the frame of nucleotide i rather than atom coordinates. A local point is placed in global coordinates by

$$\mathcal{T}_{m,i}(\mathbf{p}) = \mathbf{R}_{m,i}\mathbf{p} + \mathbf{o}_{m,i}. \tag{17}$$

Attention logits. The logit $e^{(b,r)}_{m,ij}$ is equation 3 with block and head indices restored. Its spatial term is the summed squared distance between the query points of i and the key points of j in global coordinates,

$$D^{(b,r)}_{m,ij} = \sum_{u=1}^{P} \big\| \mathcal{T}_{m,i}(\mathbf{p}^{q,(b,r)}_{m,i,u}) - \mathcal{T}_{m,j}(\mathbf{p}^{k,(b,r)}_{m,j,u}) \big\|_2^2, \tag{18}$$

with coefficient $\eta_{b,r} = \mathrm{softplus}(\zeta_{b,r})/(6\sqrt{P})$, where $\zeta_{b,r}$ is a learned scalar and the constant absorbs the fixed factors of the standard IPA point term. The pair bias $\xi_{b,r}$ is an affine map $\mathbb{R}^{d_z} \to \mathbb{R}$ of $\mathbf{z}^{(b)}_{m,ij}$, initialised to zero. The logits are normalised over all nucleotides j,

$$\alpha^{(b,r)}_{m,ij} = \frac{\exp(e^{(b,r)}_{m,ij})}{\sum_{j'=1}^{T} \exp(e^{(b,r)}_{m,ij'})}. \tag{19}$$

Aggregation. The weights aggregate the scalar values and the value points,

$$\begin{aligned} \overline{\mathbf{v}}^{(b,r)}_{m,i} &= \sum_{j=1}^{T} \alpha^{(b,r)}_{m,ij}\, \mathbf{v}^{(b,r)}_{m,j}, \\ \overline{\mathbf{p}}^{(b,r)}_{m,i,u} &= \mathbf{R}^{\top}_{m,i} \Big[\sum_{j=1}^{T} \alpha^{(b,r)}_{m,ij}\, \mathcal{T}_{m,j}(\mathbf{p}^{v,(b,r)}_{m,j,u}) - \mathbf{o}_{m,i} \Big], \end{aligned} \tag{20}$$

where the point output is returned to the frame of the receiving nucleotide, which is what makes it invariant to rigid motions (Appendix A.2).

Nucleotide update. Across heads, the scalar outputs, the three coordinates of every point output and the norm of every point output are concatenated into $H(d_c + 3P + P) = 128$ features, and the update is

$$\begin{aligned} \mathbf{h}^{(b+\frac{1}{2})}_{m,i} &= \mathrm{LN}_{b,1}\Big(\mathbf{h}^{(b)}_{m,i} + \mathbf{W}^{(b)}_{\mathrm{IPA}} \big[\overline{\mathbf{v}}^{(b,r)}_{m,i},\ \overline{\mathbf{p}}^{(b,r)}_{m,i,u},\ \|\overline{\mathbf{p}}^{(b,r)}_{m,i,u}\|_2 \big]_{r\le H,\ u\le P} \Big), \\ \mathbf{h}^{(b+1)}_{m,i} &= \mathrm{LN}_{b,2}\Big(\mathbf{h}^{(b+\frac{1}{2})}_{m,i} + \mathrm{FFN}_b\big(\mathbf{h}^{(b+\frac{1}{2})}_{m,i}\big) \Big), \end{aligned} \tag{21}$$

where $\mathbf{W}^{(b)}_{\mathrm{IPA}} : \mathbb{R}^{128} \to \mathbb{R}^{64}$ is affine, LN denotes layer normalisation and FFN_b is an MLP of widths $64 \to 128 \to 64$ with SiLU activation.

Pair update. After each nucleotide update, equation 4 refines the pair states from the two nucleotides they connect. The shared projection π is affine, $\mathbb{R}^{64} \to \mathbb{R}^{16}$, and U_b is an MLP of widths $48 \to 32 \to 16$ with SiLU activation whose last layer is initialised to zero, so that the pair states start from their FiLM-conditioned initialisation. The updated state $\mathbf{z}^{(b+1)}_{m,ij}$ supplies the bias ξ of block $b+1$; the final state $\mathbf{z}^{(B)}_{m,ij}$ is read by the contact head (Appendix A.8), and the ranking readout uses $\mathbf{h}^{(B)}_{m,i}$ (Appendix A.7).

A.7 Candidate pooling and ranking readout

The final nucleotide features $\mathbf{h}^{(B)}_{m,i}$ are pooled by four attention heads with learned queries and 16-dimensional keys and values. The head outputs are concatenated, linearly projected, and scored by an MLP of widths $64 \to 256 \to 1$ with ELU activation, which gives $f_\theta(X_m, S)$ of Section 3.2. RiNALMo contributes through pair conditioning only, no language-model feature is concatenated at readout.

A.8 Tertiary-contact labels

Two nucleotides are in contact when their C1′ atoms lie within 12 Å; the contact is tertiary if they are at least five positions apart and not a canonical base pair. Base pairs are annotated geometrically: a pair is G–C, A–U or G–U if at least two of its canonical hydrogen-bond atom pairs (N1–N3,

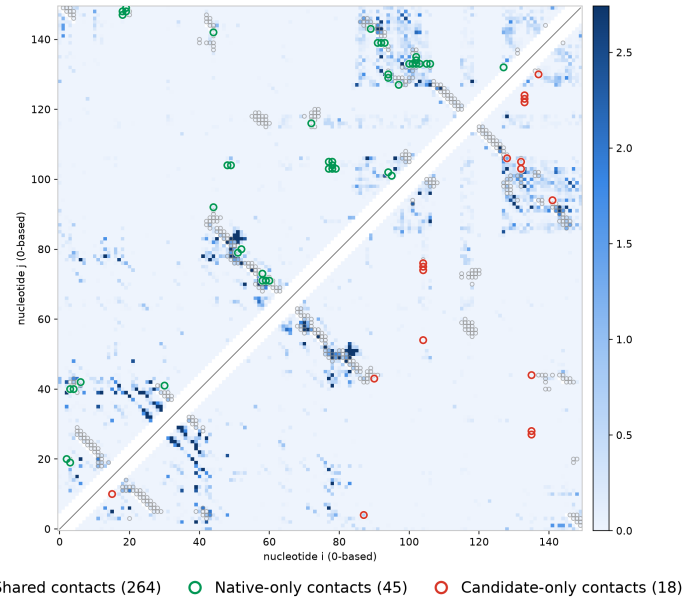


Figure 6: Sequence prior and tertiary-contact differences for 1NBS. The background shows the processed attention map of a representative RiNALMo head (33, 12). Native tertiary contacts are shown in the upper triangle and contacts of candidate 1_221 (6.93 Å RMSD) in the lower triangle. Grey circles denote contacts shared by the native and candidate, green circles native contacts missed by the candidate, and red circles spurious contacts introduced by the candidate.

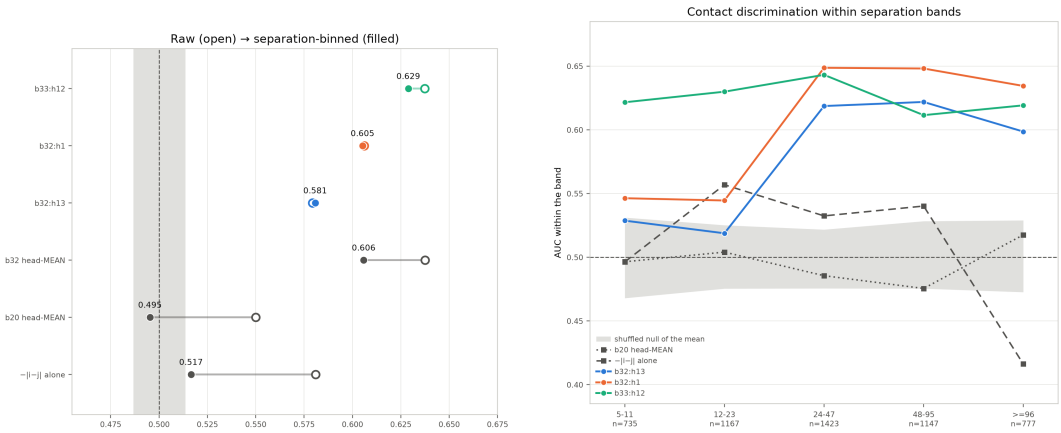
N2–O2, O6–N4 for G–C; N1–N3, N6–O4 for A–U; O6–N3, N1–O2 for G–U) lie within 3.5 Å, with $|i-j| \geq 3$ and one-to-one matching of partners. Contact sets are computed independently for the candidate and the native. The contact head is an MLP of widths $16 \to 32 \to 1$ with SiLU activation, applied to $\mathbf{z}^{(B)}_{m,ij} + \mathbf{z}^{(B)}_{m,ji}$.

B TERTIARY-CONTACT SIGNAL IN THE SEQUENCE PRIOR

The three RiNALMo heads retained in Section 3.1 were selected for the base pairs they recover (Appendix A.4), and base pairs define the stems. Helix packing, however, is determined by tertiary contacts: interactions between nucleotides distant in the sequence that fix the relative arrangement of the stems. For the prior to serve as a reference for packing, the same heads must also carry information about such contacts, and must do so on RNAs outside the selection set. This section tests that requirement on the 21 native structures of CLS-1, a held-out sequence cluster that played no part in head selection.

Set-up. For each native, the positives are its tertiary contacts by the rule of Appendix A.8: C1′ atoms within 12 Å, at least five positions apart, not a canonical base pair. The negatives are all other pairs at least five positions apart that are not canonical pairs. Each pair is scored by the preprocessed map A of one head, and the AUC of separating positives from negatives is computed per RNA and averaged. Chance is set by maps whose nucleotide positions are randomly permuted (20 permutations per head). Two controls test what a head would score without any structural information: the mean over the 20 heads of block 20, a mid-network block, and sequence separation alone, $-|i-j|$, because contacts are more common between nearby nucleotides.

Removing the effect of sequence distance. Separation alone reaches an AUC of 0.58, so a raw AUC proves little. We remove its effect in two ways. First, the AUC is computed within five separation bands (5–11, 12–23, 24–47, 48–95 and ≥ 96 nucleotides) and averaged. Second, each contact is compared only with non-contacts at the same sequence separation and the same distance in the secondary-structure graph, whose edges are the backbone and the canonical pairs, restricted to



(a) Overall vs. separation-controlled AUC.

(b) AUC across sequence-separation bands.

Figure 7: Tertiary-contact discrimination of the three selected RiNALMo attention heads on the 21 CLS-1 RNAs. Heads are denoted by block and head index, e.g., b32:h13 for head 13 in block 32. (a) Overall tertiary-contact AUC across all nucleotide pairs is shown by open markers, while filled markers show the mean AUC computed within sequence-separation bins, controlling for the effect of $|i - j|$. The grey band denotes the 95% range of the corresponding separation-controlled mean under position-shuffled attention maps; $|i - j|$-only and block-wise head-mean baselines are shown for comparison. (b) AUC within each sequence-separation band, macro-averaged over RNAs containing contacts in that band; n gives the number of tertiary contacts in each band. The selected heads retain above-chance discrimination across separation ranges, including long-range pairs (≥ 96 nt), indicating that their signal is not explained by sequence separation alone.

pairs more than four steps apart. The graph distance matters biologically: a nucleotide next to a base pair lies within 12 Å of its partner's neighbour, so a base-pairing head detects such contacts for free. Matching on both distances leaves only contacts that a predictor built from sequence separation and secondary structure scores at exactly 0.5.

Results. Binning by separation removes the lift that sequence locality gives any predictor: separation alone and the block-20 mean fall to chance, while the three heads retain their signal (Figure 7a), on 19 or 20 of the 21 RNAs depending on the head (Figure 9a). Within bands (Figure 7b), head $(33, 12)$ exceeds the shuffled range at every separation, and all three heads do so from 24 nucleotides upwards, including pairs at least 96 positions apart, the separations that helix packing involves and at which both controls sit at or below chance.

The matched comparison (Figure 8) confirms that this is not an artefact of proximity to stems: on the 1,817 contacts more than four steps from a stem that have matched non-contacts (20 RNAs; 4U7U_6 has none), all three heads remain above chance with confidence intervals clear of 0.5. The block-32 head mean retains signal as well, whereas the block-20 mean does not: the information is concentrated in the last blocks, which supply the selected heads.

The same heads recover the stems (Figure 9b): the best head exceeds a precision of 0.5 on 14 of the 21 RNAs, 1NBS among them. The five RNAs on which recovery fails are those on which the tertiary-contact AUC sits at chance in panel (a): where the prior does not resolve the stems, it does not resolve their packing either. For 1NBS, Figure 6 shows the map of head $(33, 12)$ concentrating on the stems and on several of the contacts a poor candidate loses.

An AUC near 0.6 is a modest signal, and SIRGE does not use the prior as a contact predictor. It is the reference against which the contact objective (Section 3.3) teaches SIRGE to judge a candidate's packing, and these results show that the reference carries information about long-range contacts, which geometry alone cannot supply, on RNAs the model has never seen.

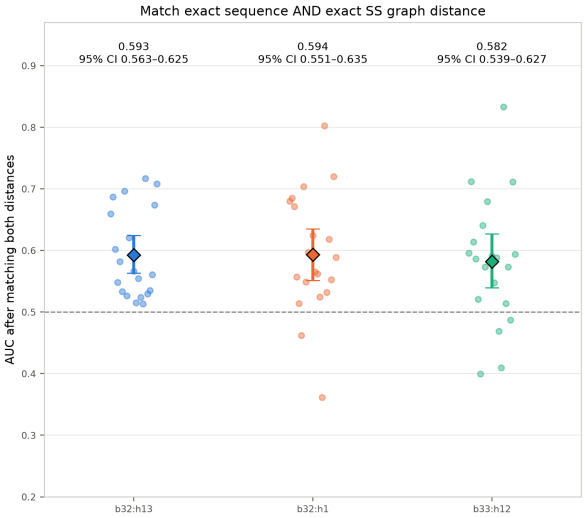


Figure 8: Tertiary-contact AUC under matched sequence and secondary-structure distances. Each tertiary contact is matched with non-contacts having the same sequence separation and secondary-structure graph distance, for pairs separated by more than four positions. The dashed line marks AUC$= 0.5$, the value for predictors based only on the two matched distances. Dots denote per-RNA AUCs for the 20 RNAs with matched contacts; diamonds denote means with RNA-bootstrap 95% confidence intervals.

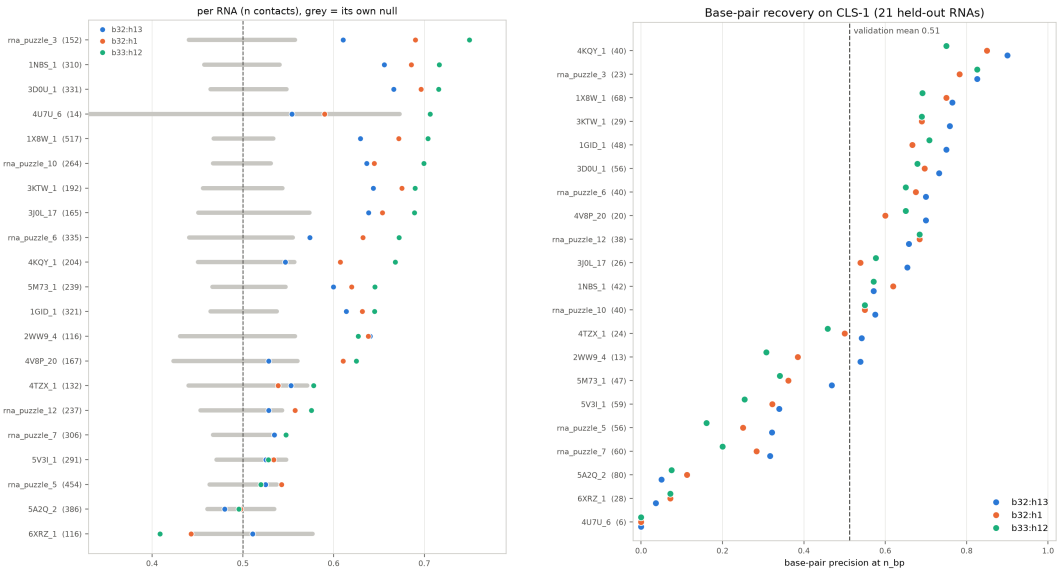
C DATASETS

HomRank data: CLS-0, CLS-2 and CLS-1. Training, validation and the first test set are drawn from the dataset of HomRank (Li et al., 2025a), which comprises 190 RNA targets from three sources: 151 structures collected from the Protein Data Bank, for each of which HomRank generated 1,000 candidates by molecular-dynamics simulation; the 18 targets of the ARES dataset (Townshend et al., 2021), each with 1,000 candidates generated by Rosetta FARFAR2 (Watkins et al., 2020); and 21 RNA-Puzzles targets with 451,629 FARFAR2 candidates in total. Candidate quality is the RMSD to the experimental structure after optimal superposition, as provided by HomRank. HomRank clusters the 190 targets by pairwise sequence similarity (`Bio.Align`, hierarchical clustering) into three groups: CLS-0 (146 RNAs: 123 from the PDB, 8 from ARES and 15 RNA-Puzzles), CLS-1 (21: 15 from the PDB and 6 RNA-Puzzles) and CLS-2 (23: 13 from the PDB and 10 from ARES). We exclude the 15 RNA-Puzzles targets of CLS-0 from training, since RNA-Puzzles is an established benchmark, and train on the remaining 131 RNAs. CLS-2 serves for validation and checkpoint selection, CLS-1 for evaluation (Table 6).

The two generators are unevenly distributed across splits. Molecular dynamics, which perturbs a starting structure under a force field, supplies 94% of the training candidates; FARFAR2, which assembles candidates from fragments of solved structures, supplies the remaining 6%, all from the 8 ARES targets. In CLS-1 the six RNA-Puzzles targets contribute 90% of the full pool and 44% of the candidates scored after subsampling. CLS-1 therefore shifts the generator mix as well as sequence length and candidate quality (Figure 10).

For each CLS-1 RNA we score at most 2,000 candidates, drawn uniformly at random from its pool, and every method scores the same subset.

Leakage controls. The split is made at the level of RNA targets, so an RNA's sequence, native structure and complete candidate pool fall in one partition, and the three partitions belong to distinct sequence-similarity clusters. No candidate or label of CLS-1 contributes to training or checkpoint selection; all reported predictions use the checkpoint selected on CLS-2. RiNALMo is frozen and receives only the sequence.



(a) Per-RNA tertiary-contact AUC.

(b) Per-RNA base-pair recovery.

Figure 9: Per-RNA analysis of the three selected RiNALMo attention heads on the 21 CLS-1 RNAs. (**a**) Separation-binned tertiary-contact AUC for each RNA and each selected head. RNAs are ordered by the AUC of head $(33, 12)$; numbers in brackets give the number of native tertiary contacts, and grey bars show the range obtained from position-shuffled attention maps for the corresponding RNA. (**b**) Base-pair recovery on the same RNAs, measured as precision among the top-n_{bp} ranked nucleotide pairs, where n_{bp} is the number of native base pairs shown in brackets. RNAs are ordered by the best-performing head, and the dashed line marks the mean precision on the validation RNAs used for head selection.

Table 6: Composition of the HomRank splits. Candidate counts are for the complete pools; CLS-1 is subsampled to at most 2,000 candidates per RNA at evaluation.

			Candidates by generator		
Partition	RNAs	Length (nt)	MD	FARFAR2	Total
Training (CLS-0)	131	22–144	123,000	8,000	131,000
Validation (CLS-2)	23	12–49	13,000	10,000	23,000
Test (CLS-1)	21	61–257	15,000	137,917	152,917

Gen-OOD. Gen-OOD is drawn from Dataset 3 of Martinović et al. (2024), a benchmark built to test whether deep-learning predictors generalise to RNAs released after their training data. Its authors extracted all 20,320 RNA chains in the Protein Data Bank, clustered them with MMseqs2 at 90% sequence identity and 80% coverage, and kept the clusters in which every member was released after 1 April 2022, taken as a conservative training cutoff for all six predictors (AlphaFold 3's training data end in September 2021). Chains shorter than 10 nucleotides, with resolution worse than 9 Å or with fewer than 90% of nucleotides resolved were removed, one representative was chosen per cluster, and the 190 RNAs for which all six predictors returned a structure form Dataset 3. Predictions used published weights and default settings, one shared multiple-sequence alignment for the four predictors that require one, and the first model where a predictor returns several. Every chain was predicted on its own, from its sequence alone, although 173 of the 190 come from complexes, 166 with proteins and 7 with other RNAs, so that the experimental reference is the chain's conformation as bound.

We use the 84-RNA subset of Dataset 3 assembled by HomRank (Li et al., 2025a) for evaluating candidates from several generators. Each RNA has one candidate from each of AlphaFold 3, DRfold, DeepFoldRNA, RhoFold, RoseTTAFoldNA and trRosettaRNA, giving 504 candidates. We refer to this subset as Gen-OOD, since it tests evaluators under a shift of candidate generator. Gen-OOD

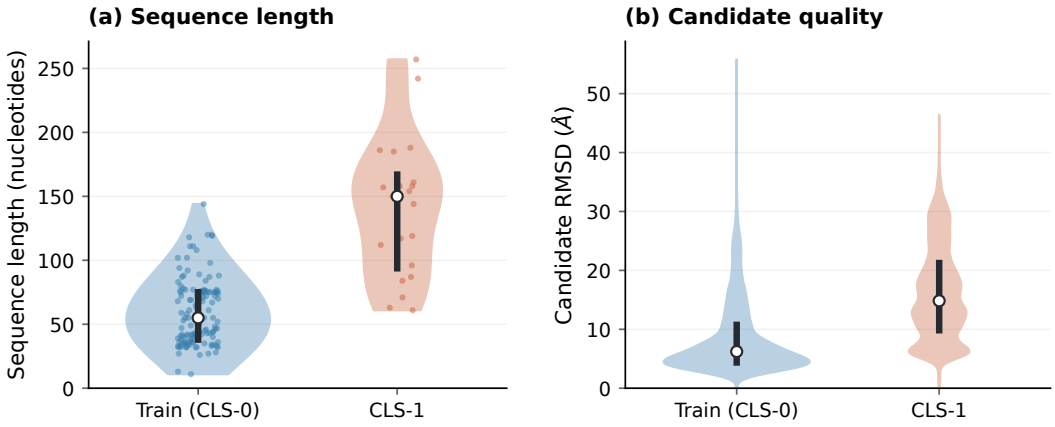


Figure 10: Distribution shifts from train to CLS-1. (a) Sequence length. (b) Candidate RMSD across the complete released pools. CLS-1 shifts towards longer sequences and a broader range of structural errors.

is used for evaluation only: none of its RNAs, candidates or labels enters training or checkpoint selection, and its composition was fixed independently of this work.

D Experimental Protocol

Baselines. We compare SIRGE with six learned evaluators spanning the main approaches to geometric RNA evaluation: ARES (Townshend et al., 2021), an equivariant network over atoms; RNA3DCNN (Li et al., 2018), a 3D convolutional network over voxelised atomic environments; PAMNet (Zhang et al., 2023), a physics-aware graph network; lociPARSE (Tarafder & Bhattacharya, 2024), which uses nucleotide frames and invariant point attention; EquiRNA (Li et al., 2025b), a hierarchical equivariant network; and HomRank (Li et al., 2025a), which trains ARES's architecture with a listwise ranking objective. Each baseline keeps its published architecture and objective.

Training and model selection. Every learned method, SIRGE and its ablations included, is trained on the CLS-0 training set for five epochs with three random seeds, using the optimiser settings of its published implementation. SIRGE uses AdamW with learning rate 3×10^{-3}, weight decay 10^{-4}, gradient clipping at 5.0, a 100-step linear warm-up and cosine decay. Each batch holds 16 candidates of one RNA, from which 32 lists of 16 are sampled for the ranking loss, and gradients are accumulated over eight batches. The ranking loss uses $\tau = 0.01$ with RMSD in Å; the per-nucleotide loss uses $\tau_{\text{local}} = 0.02$, with λ_{local} rising from 0 to 1 over the first 200 optimisation steps. For every method, the checkpoint that is best under its validation criterion on CLS-2 is kept and evaluated on CLS-1 and Gen-OOD without further training. SIRGE-base and the ablations of Section 4.3 follow the same protocol and seeds.

Evaluation. All scores are oriented so that larger is better. On CLS-1, each RNA is evaluated on its subsampled pool of at most 2,000 candidates (Appendix C): Kendall-τ alignment and Hit@k over the whole pool, and the banded results and Top-3 over 4,000 lists of ten candidates drawn uniformly at random from the pool, identical for every method, a band's value being the mean over its lists. On Gen-OOD, every metric is computed over the six candidates of each RNA. Kendall-τ alignment is $-\tau(f_\theta, \text{RMSD})$, so $+1$ is perfect agreement with candidate quality, 0 is chance, and negative values mean an evaluator ranks candidates against their quality; every τ reported in tables and figures is this alignment, never the raw correlation. Top-3 treats RMSD differences below 0.1 Å as ties on both benchmarks. Every metric is computed per RNA, macro-averaged over RNAs and then over seeds; the reported standard deviation is across seeds.

Reproducibility and hardware. Results are reported as mean $\pm$ standard deviation over the three training seeds. All experiments ran on four NVIDIA RTX PRO 6000 Blackwell GPUs with 96 GB of memory each.