

Composition-Controlled Evaluation of RNA Structure Awareness in Foundation Models

Elliot Tower
Mechanistic Research
elliott@elliotttower.ai

July 2026

Abstract

We introduce a composition-controlled evaluation framework for measuring RNA secondary structure awareness in foundation models and apply it across four preregistered phases to ten models spanning attention-based, SSM-based, and hybrid architectures across four orders of magnitude in parameter count (1.2M–7B). The framework addresses a composition confound inherent to RNA secondary structure: stems are 68.5% GC while loops are 43.9%, so any embedding metric sensitive to nucleotide identity inherits a $\sim 3\times$ bias toward stem positions. Our nucleotide-stratified permutation null absorbs this confound, reducing the apparent structure signal from $5.5\times$ to $1.8\times$ (mean across 52 families). A max-over-layers correction ensures the null is computed under the same best-layer selection as the real data.

The framework reveals a comprehensive null: no model encodes secondary structure beyond what composition and architecture explain. NT v2 (ESM+GLU, 56M, DNA-pretrained) shows high attention-contact correlation ($\rho = 0.322$), but untrained weights produce the same signal ($\rho = 0.328$), exposing this as an architectural artifact of the ESM attention mechanism. Phase 3 adds two RNA-pretrained models: RiNALMo (650M) shows genuine but modest mutation sensitivity (17/52 families exceed the composition null, sign test 94%), while UTR-LM (1.2M) shows no structure awareness despite structure-aware pretraining objectives. Phase 4 adds three more models: ERNIE-RNA (86M, RNA-pretrained) and SpliceBERT (19.4M, RNA-pretrained) both confirm the null pattern on mutation and attention; DNABERT-2 (117M, DNA, BPE tokenization) shows the only clear learned attention signal ($\rho = 0.257$ trained vs 0.000 untrained), though this excludes ALiBi position biases. Evo (StripedHyena, 7B) shows the weakest mutation sensitivity despite being $500\times$ larger than Caduceus (14M), indicating that pretraining domain and tokenization dominate scale. No family survives both nucleotide and dinucleotide composition controls. The directional effect is real—trained exceeds untrained in 50/52 families for RNA-FM and 52/52 for ERNIE-RNA—but the magnitude remains below the preregistered threshold for practical significance.

The framework generalizes to any domain where feature composition correlates with the target structure.

1 Introduction

Foundation models pretrained on nucleotide sequences have been proposed for RNA biology, but evaluating whether they encode secondary structure—the pattern of intramolecular base pairing that governs RNA stability and function—requires metrics that distinguish learned representations from sequence-composition artifacts.

We initially adopted the Grassmannian geodesic distance (d_g), a geometric metric for comparing embedding subspaces between wild-type and mutant sequences. A zero-parameter 3-mer frequency

baseline matched the trained models’ d_g scores, falsifying the metric: it captured k -mer composition, a property that random linear maps preserve under the Johnson–Lindenstrauss lemma (?), rather than learned structure. Similar failure modes have been reported for geometric transfer metrics in other domains (?).

The d_g collapse motivated a replacement metric with explicit composition controls. We designed a mutation sensitivity ratio under complement swap ($A \leftrightarrow U$, $C \leftrightarrow G$) with a nucleotide-stratified permutation null. The key observation: stems are enriched for G and C (68.5% vs 43.9% in loops), and complement swaps at G/C positions produce inherently larger embedding perturbations than at A/U positions regardless of structural context. Any naive stem-vs-loop comparison inherits this 24.7 percentage-point enrichment as a confound.

The contribution is the evaluation framework and an honest report of what it reveals: a comprehensive null on structure encoding across ten models, an informative scaling observation (7B parameters do not compensate for domain mismatch), and a corrected attention analysis showing that NT v2’s high attention-contact correlation is architectural rather than learned. Structure probing of RNA models has previously focused on downstream task performance (??) or attention visualization rather than composition-controlled metrics. ? showed that BERT embeddings become increasingly anisotropic through layers, a phenomenon we observe in RNA-FM on repeat sequences. Vig et al. demonstrated that protein language model attention heads correlate with contact maps (?); our attention-contact analysis extends this to RNA, finding strong learned correlation for NT v2 and near-zero for RNA-FM. The mutation sensitivity approach relates to *in silico* mutagenesis for variant effect prediction (?), applied here to the representation layer with structure-defined mutation categories (stem vs. loop). Composition confounds in genomic analyses have a long history: dinucleotide frequencies affect codon usage statistics (?), CpG content confounds methylation studies, and GC content correlates with gene density and chromatin accessibility. Our stratified permutation framework adapts these insights to the evaluation of neural sequence models.

2 Methods

2.1 RNA families and models

Phase 1 evaluated 12 RNA families with experimentally determined secondary structures from PDB crystal structures and Rfam consensus annotations (Table ??). Phase 2 expands to $N = 52$ families drawn from Rfam 14.10, spanning ribozymes, riboswitches, snRNAs, cis-regulatory elements, miRNA precursors, CRISPR repeats, and other ncRNAs (Appendix A). Positions are labeled stem (base-paired) or loop (unpaired) from dot-bracket annotation. Minimum criteria: ≥ 20 nt, ≥ 5 stem and 5 loop positions.

Phases 1–2 evaluate five models spanning attention-based, SSM-based, and hybrid architectures (Table ??): RNA-FM (BERT, 99M, character, RNA-pretrained; ?), NT v2 (ESM+GLU, 56M, 6-mer, DNA; ?), HyenaDNA (Hyena SSM, 5.4M, character, DNA; ?), Caduceus (BiMamba, 14M, character, DNA; ?), and Evo (StripedHyena, 7B, byte-level, DNA; ?). Phase 3 adds RiNALMo (BERT+RoPE+SwiGLU, 650M, character, RNA; ?) and UTR-LM (transformer encoder, 1.2M, character, UTR-focused; ?). Phase 4 adds ERNIE-RNA (BERT, 86M, character, RNA; ?), SpliceBERT (BERT, 19.4M, character, RNA; ?), and DNABERT-2 (MosaicBERT+ALiBi, 117M, BPE, DNA; ?). Six of the ten models are RNA-pretrained; parameter counts span four orders of magnitude.

Table 1: Phase 1 RNA families. Source: PDB crystal structures and Rfam consensus annotations. Stem: number of base-paired positions.

RNA family	Class	Length	Stem	Source
tRNA ^{Phe} (yeast)	tRNA	76	42	PDB 1EHZ
tRNA ^{Ala} (human)	tRNA	76	42	Rfam RF00005
5S rRNA (<i>E. coli</i>)	rRNA	120	72	PDB 1C2X
Hammerhead ribozyme	Ribozyme	48	16	PDB 2OEU
SAM riboswitch	Riboswitch	94	52	PDB 2GIS
TPP riboswitch	Riboswitch	80	40	PDB 2CKY
SRP RNA helix 8	Structural	77	44	PDB 1HQ1
miR-21 precursor	Regulatory	72	64	Rfam RF00001
IRES HCV domain II	Regulatory	77	50	PDB 1P5O
HDV ribozyme	Ribozyme	76	48	PDB 1DRZ
RNase P specificity	Structural	78	42	PDB 1NBS
U2 snRNA stem	Spliceosomal	76	46	PDB 3K1V

Table 2: Models under test. Phases 1–2 evaluate the first five; Phase 3 adds RiNALMo and UTR-LM; Phase 4 adds the final three.

Model	Architecture	Params	Tokenization	Pretraining
RNA-FM	BERT encoder	99M	Character	RNA
NT v2	ESM + GLU attention	56M	6-mer	DNA
HyenaDNA	Hyena SSM	5.4M	Character	DNA
Caduceus	BiMamba SSM	14M	Character	DNA
Evo	StripedHyena	7B	Byte-level	DNA
RiNALMo	BERT + RoPE + SwiGLU	650M	Character	RNA
UTR-LM	Transformer encoder	1.2M	Character	UTR (RNA)
ERNIE-RNA	BERT encoder	86M	Character	RNA
SpliceBERT	BERT encoder	19.4M	Character	RNA
DNABERT-2	MosaicBERT + ALiBi	117M	BPE	DNA

2.2 Structure sensitivity metrics

Mutation sensitivity ratio. For each position in each RNA sequence, we perform a complement swap ($A \leftrightarrow U$, $C \leftrightarrow G$) and measure the cosine distance between wild-type and mutant embeddings at the mutated position. The structure sensitivity ratio R is the mean cosine distance at stem positions divided by the mean at loop positions. We compute R at every layer and report the maximum; the null models below are computed under the same max-over-layers selection.

Attention-contact correlation. For models with accessible attention matrices (RNA-FM, NT v2, RiNALMo, UTR-LM, ERNIE-RNA, SpliceBERT, DNABERT-2), we compute the Spearman rank correlation between symmetrized attention weights and a binary contact matrix (1 for base-paired positions, 0 otherwise). For NT v2’s 6-mer and DNABERT-2’s BPE tokenization, contacts are aggregated to the token level. All models except DNABERT-2 are loaded with `attn_implementation=“eager”` to ensure attention weights are returned. DNABERT-2’s MosaicBERT architecture uses a fused Wqkv projection; we extract content-only attention by hooking the Wqkv output, splitting Q and K, and computing $\text{softmax}(QK^T/\sqrt{d_k})$ without ALiBi position biases. We compare trained attention to an untrained baseline with randomized weights.

Structure probing. Logistic regression on per-position embeddings predicting stem vs loop,

using GroupKFold cross-validation by RNA family to test cross-family generalization. Balanced accuracy is reported against the majority baseline.

2.3 Composition null models

The nucleotide-stratified null shuffles stem/loop labels independently within each nucleotide type (A, U, G, C), preserving per-nucleotide composition while destroying the structure-label association. For each of 100 permutations, a single shuffled label assignment is generated and R is computed at every layer; the maximum across layers is taken. The 95th percentile of these 100 max-over-layers values defines the null threshold. This controls for first-order composition (per-nucleotide GC enrichment) but not for dinucleotide context—stem-G occurs predominantly in GC base pairs, while loop-G occurs in diverse dinucleotide contexts.

The dinucleotide-stratified null extends this to second-order context: labels are shuffled within each dinucleotide type (previous nucleotide + current nucleotide: AA, AC, AG, ...). The same max-over-layers procedure applies. This null is run on all families exceeding the first-order null, plus tRNA-Ala and SAM riboswitch (Phase 1 survivors) regardless of Phase 2 result.

Both phases were preregistered with frozen code and analysis SHAs (Phase 1: 694b43b; Phase 2: bd4b3fd). Phase 1 ($N = 12$) tested whether RNA-FM’s trained/untrained ratio exceeds 2.0 (H1), whether $\geq 6/12$ families exceed the null (H1b), and whether probing exceeds the majority baseline (H5). Phase 2 ($N \geq 50$) replicated these at scale (H6, H6b, H8), added a dinucleotide null test (H7: ≥ 1 family survives), a confirmatory sign test promoting the Phase 1 exploratory unanimity (H10: trained > untrained in $\geq 75\%$ of families), and an NT v2 attention test (H11: trained $\rho > 0.15$, untrained < 0.05). A methodological correction applied to both phases: the Phase 1 null had been computed per-layer independently, while the metric used max-over-layers. Phase 2 corrects this mismatch.

3 Results

3.1 Attention-contact correlation is architectural

NT v2 trained attention correlates with base-pairing contacts at mean Spearman $\rho = 0.322$ across 52 families (Table ??). An earlier analysis using the default SDPA attention backend reported $\rho = 0.000$ for untrained weights, but SDPA silently returns `None` for attention weights, producing a spurious zero. With eager attention computation, untrained NT v2 produces $\rho = 0.328$ —slightly *higher* than trained. The per-family delta (trained – untrained) averages -0.006 with no systematic direction: 22 families show higher trained correlation, 30 show higher untrained. The entire attention-contact signal is an architectural property of the ESM attention mechanism and 6-mer tokenization, present from random initialization.

Among character-tokenized models, RiNALMo and ERNIE-RNA show the strongest suggestions of learned attention: RiNALMo trained $\rho = 0.101$ versus untrained 0.082 ($\Delta = +0.019$), ERNIE-RNA trained $\rho = 0.099$ versus untrained 0.075 ($\Delta = +0.024$). Both deltas are small. SpliceBERT (0.064 vs 0.056), RNA-FM (0.051 vs 0.060), and UTR-LM (0.052 vs 0.053) show no learned signal. DNABERT-2’s content-only attention (computed from Wqkv projections without ALiBi position biases; Section ??) shows the largest trained/untrained gap (0.257 vs 0.000), but this metric excludes the position-dependent biases that dominate the model’s actual attention computation.

NT v2’s 6-mer tokenization explains the architectural correlation. Base-pairing partners—typically 3–30 nucleotides apart—are aggregated into adjacent tokens, producing structured attention

Table 3: Attention-contact Spearman correlation ($N = 52$). NT v2’s high correlation is entirely architectural (untrained $\geq$ trained). DNABERT-2 shows learned content attention but excludes ALiBi position biases (*).

Model	Trained	Untrained	Interpretation
NT v2	0.322	0.328	Architectural
DNABERT-2*	0.257	0.000	Learned (content-only)
RiNALMo	0.101	0.082	Weak ($\Delta = +0.019$)
ERNIE-RNA	0.099	0.075	Weak ($\Delta = +0.024$)
SpliceBERT	0.064	0.056	None ($\Delta = +0.008$)
RNA-FM	0.051	0.060	None
UTR-LM	0.052	0.053	None

Table 4: Mutation sensitivity summary, Phase 1 ($N = 12$).

Model	Mean ratio	Families > null	Probing Δ
RNA-FM (99M, RNA)	1.774	1/12	+0.013
Caduceus (14M, DNA)	1.257	3/12	+0.041
HyenaDNA (5.4M, DNA)	1.223	0/12	+0.034
Evo (7B, DNA)	1.170	0/12	+0.028
NT v2 (56M, DNA)	1.094	6/12	+0.016

patterns from positional proximity alone. Random weights with this tokenization already capture enough pairwise structure to match trained attention.

3.2 Scale does not compensate for domain mismatch

Evo (7B parameters, byte-level tokenization, DNA-pretrained) produces the weakest mutation sensitivity of all trained models (Table ??). Two families (tRNA-Phe: 0.927, tRNA-Ala: 0.899) show ratios below 1.0—Evo is less sensitive to stem mutations than loop mutations for these structures. Probing accuracy (0.615) peaks at layer 0 (embedding) and falls below Caduceus (0.628, 14M) and HyenaDNA (0.621, 5.4M). Despite being $500\times$ larger than Caduceus and $70\times$ larger than RNA-FM, Evo shows no advantage on any structure metric. This observation is consistent across mutation sensitivity, probing, and the null comparison, though $n = 5$ models is too few to support a general scaling claim.

3.3 Composition controls absorb the embedding-level signal

The nucleotide-stratified null reduced the apparent RNA-FM signal from $\sim 5.5\times$ (naive, before composition control) to $1.8\times$ (composition-controlled). The 24.7 percentage-point GC enrichment in stems was the dominant contributor. Per-family results (Table ??) show that only one family exceeded the first-order null for RNA-FM: tRNA-Ala (3.208), with high-GC stems and low-GC loops where residual dinucleotide-context effects contribute. NT v2 shows 6/12 families exceeding the null despite having the lowest mean ratio (1.094), suggesting a more consistent per-family signal than RNA-FM’s high-mean, high-variance pattern.

Phase 1 preregistered hypotheses confirmed this pattern. H1 (trained/untrained ratio ≥ 2.0) failed at $1.69\times$. H1b ($\geq 6/12$ families exceed null) failed at 1/12. H5 (probing > baseline) failed,

Table 5: Per-family mutation sensitivity ratio, Phase 1 ($N = 12$).

Family	RNA-FM	Untrained	NT v2	HyenaDNA	Caduceus	Evo
tRNA-Phe (yeast)	2.758	1.050	1.180	1.185	1.374	0.927
tRNA-Ala (human)	3.208	1.091	1.207	1.284	1.328	0.899
5S rRNA	1.157	1.008	0.963	1.012	1.075	1.048
Hammerhead	1.328	1.076	1.118	1.292	1.456	1.147
SAM riboswitch	2.535	1.056	1.042	1.074	1.269	1.283
TPP riboswitch	1.442	1.093	1.058	1.227	1.324	1.888
SRP RNA	1.261	1.035	1.168	1.029	1.136	0.925
miR-21	1.299	1.049	1.086	1.868	1.341	1.156
IRES HCV dom. II	1.710	1.045	0.928	1.190	1.225	0.994
HDV ribozyme	1.135	1.021	1.175	1.150	1.291	1.316
RNase P	1.513	1.034	0.998	1.217	1.086	1.217
U2 snRNA	2.200	1.024	1.203	1.144	1.173	1.234
Mean	1.774	1.048	1.094	1.223	1.257	1.170
> null	1/12	—	6/12	0/12	3/12	0/12

Table 6: Phase 2 mutation sensitivity ($N = 52$).

Model	Mean ratio	Families > nuc null	Probing Δ
RNA-FM (99M, RNA)	1.774	1/52	-0.010
NT v2 (56M, DNA)	1.169	28/52	-0.039
Caduceus (14M, DNA)	1.197	5/52	-0.021
HyenaDNA (5.4M, DNA)	1.138	8/52	-0.058
Evo (7B, DNA)	1.352	6/52	-0.042
RNA-FM untrained	1.038	3/52	—

with margins of +0.013 to +0.041 within noise at $N = 12$. RNA-FM trained did exceed untrained in 12/12 families (Wilcoxon $p = 0.000244$), but this per-family unanimity was not preregistered and is reported as exploratory.

3.4 Phase 2 confirms and refines at $N = 52$

Phase 2 evaluates the original five models on 52 Rfam families under the corrected max-over-layers null. RNA-FM has the highest mean ratio (1.774) but exceeds the null in only 1/52 families (2%), barely above the untrained control’s 3/52 chance rate (Table ??). The sole RNA-FM survivor is tRNA-Ala (3.208 vs null 3.179), a borderline exceedance absorbed by the dinucleotide null (Section ??). All five models fall *below* the 58.1% majority baseline on probing at $N = 52$; the positive Phase 1 margins were noise amplified by small sample size.

H6 (ratio ≥ 2.0) failed at $1.77\times$. H6b (≥ 9 exceed null) failed at 1/52. H7 (≥ 1 family survives dinucleotide null) failed: the sole first-order survivor (tRNA-Ala) is absorbed by the dinucleotide null. H8 (probing > baseline +0.02) failed at -0.010. H10 (sign test) passed: RNA-FM trained exceeds untrained in 50 of 52 families (96%), confirming the Phase 1 exploratory observation as preregistered-confirmatory. The two exceptions are families with fewer than 10 stem positions where ratio estimates are noisy. The magnitude ($1.77\times$) remains below the 2.0 threshold for practical claims. H11 (NT v2 attention) **failed**: trained $\rho = 0.322$ across 52 families, but untrained = 0.328 after correcting for the SDPA backend bug that had produced a spurious 0.000. The correlation is

Table 7: Dinucleotide null results. No family survives both composition controls.

Family	Ratio	Nuc 95th	Dinuc 95th	> dinuc
tRNA-Ala	3.208	3.179	4.070	No
SAM riboswitch	2.535	3.022	3.600	No

Table 8: Phase 3 results ($N = 52$). RiNALMo shows a consistent directional signal (49/52 sign test) but a modest effect size. UTR-LM shows no structure sensitivity.

Model	Mean ratio	> nuc null	Sign test	Probing	Attn ρ	Attn ρ (unt.)
RiNALMo	1.109	17/52	49/52 (94.2%)	0.561	0.101	0.082
UTR-LM	1.045	3/52	37/52 (71.2%)	0.577	0.052	0.053

architectural (Section ??).

3.5 Dinucleotide null absorbs all survivors

The dinucleotide-stratified null was run on the sole first-order survivor (tRNA-Ala) plus SAM riboswitch (a Phase 1 survivor, included regardless of Phase 2 result). Both fall below the second-order control (Table ??). tRNA-Ala (3.208) is absorbed by a dinucleotide null of 4.070, confirming that its signal reflected GC dinucleotide enrichment rather than learned structure. SAM riboswitch (2.535) is similarly absorbed (3.600).

No RNA family survives both orders of composition control. The primary evidence for a learned structure effect is the directional result (H10: 50/52 trained > untrained), which does not depend on any individual family exceeding any threshold. The effect is consistent but small ($1.77\times$), and entirely attributable to sequence composition once stratified nulls are applied.

3.6 Phase 3: RNA-specialized models

Phase 3 adds two RNA-pretrained models—RiNALMo (650M, BERT+RoPE+SwiGLU) (?) and UTR-LM (1.2M, transformer encoder) (?)—bringing the total to seven models across five architectures. Both are evaluated on all 52 families with the same metrics as Phase 2, plus trained/untrained comparisons for mutation sensitivity and attention.

RiNALMo produces a higher mutation sensitivity ratio than any DNA-pretrained model except RNA-FM (Table ??). The sign test (49/52 trained > untrained, 94.2%) exceeds the preregistered H15 threshold of 75%, confirming a directional effect. However, the mean ratio (1.109) falls well below the 2.0 practical threshold (H14: confirmed), and only 17/52 families exceed the nucleotide-stratified null. RiNALMo’s attention-contact correlation (0.101 trained vs 0.082 untrained) is the only trained-exceeds-untrained attention result across all seven models, though the per-family sign test (35/52, 67.3%) falls short of strong directional evidence.

UTR-LM shows the weakest mutation sensitivity of all seven models (1.045), with only 3/52 families exceeding the null—comparable to the RNA-FM untrained baseline (3/52). UTR-LM’s 1.2M parameter count and UTR-focused pretraining produce no detectable structure signal on full-length RNA families. Its probing accuracy (0.577) is the highest of all models, consistent with the baseline probing result reflecting embedding dimensionality rather than learned structure (Section ??). UTR-LM attention (0.052 trained vs 0.053 untrained) shows no learned structure.

Table 9: Phase 4 results ($N = 52$). ERNIE-RNA and SpliceBERT confirm the null pattern. DNABERT-2 mutation and probing are not extractable due to BPE tokenization misalignment.

Model	Mean ratio	> nuc null	Sign test	Probing	Attn ρ	Attn ρ (unt.)
ERNIE-RNA	1.149	22/52	52/52 (100%)	0.604	0.099	0.075
SpliceBERT	1.045	4/52	48/52 (92.3%)	0.580	0.064	0.056
DNABERT-2	—	—	—	—	0.257*	0.000

*Content-only attention from Wqkv hooks; excludes ALiBi position biases.

H12 (RiNALMo attention $\rho < 0.10$) **rejected**: trained $\rho = 0.101$, barely exceeding the threshold. H13 (UTR-LM attention $\rho > 0.15$) **rejected**: $\rho = 0.052$. H14 (RiNALMo ratio < 2.0) **confirmed**: ratio = 1.109. H15 (RiNALMo sign test $\geq 75\%$) **confirmed**: 49/52 (94.2%).

3.7 Phase 4: Expanded model coverage

Phase 4 adds ERNIE-RNA (86M, BERT, RNA-pretrained; ?), SpliceBERT (19.4M, BERT, RNA-pretrained; ?), and DNABERT-2 (117M, MosaicBERT with ALiBi, DNA-pretrained; ?), bringing the total to ten models.

ERNIE-RNA produces the strongest sign test of all ten models: trained exceeds untrained in 52/52 families (100%), with a mean ratio of 1.149 (Table ??). Despite this perfect directional consistency, the mean ratio remains well below the 2.0 practical threshold, and 22/52 families exceed the nucleotide null—a rate driven by the consistent small effect rather than large individual exceedances. ERNIE-RNA’s probing accuracy (0.604) is the highest of all models, though this reflects embedding dimensionality rather than structure discrimination.

SpliceBERT (19.4M) shows the weakest mutation sensitivity among RNA-pretrained models (1.045), comparable to UTR-LM (1.045) and the untrained baseline. Only 4/52 families exceed the null. The sign test (48/52, 92.3%) exceeds the 75% threshold, confirming a directional effect that is consistent but negligible in magnitude.

DNABERT-2 uses BPE tokenization and MosaicBERT architecture with ALiBi position biases, which prevent standard metric extraction. BPE tokens do not align with nucleotide positions, making per-position mutation sensitivity and probing undefined. MosaicBERT’s fused Wqkv projection and ALiBi biases prevent standard attention extraction; we compute content-only attention by hooking the Wqkv output, splitting into Q and K, and computing $\text{softmax}(QK^T/\sqrt{d_k})$ without ALiBi. The resulting trained $\rho = 0.257$ versus untrained 0.000 shows learned content attention that correlates with base pairing—unlike NT v2, whose correlation is architectural. The untrained model produces uniform content attention (all Spearman correlations zero) because random Wqkv weights yield unstructured query-key products. With ALiBi biases included, the true attention pattern would be dominated by local position bias, so 0.257 is an upper bound on the full model’s attention-contact correlation.

H16 (ERNIE-RNA attention < 0.10) **confirmed**: $\rho = 0.099$, a marginal pass. H17 (ERNIE-RNA sign test $\geq 75\%$) **confirmed**: 52/52 (100%). H18 (SpliceBERT attention < 0.10) **confirmed**: $\rho = 0.064$. H19 (SpliceBERT ratio < 2.0) **confirmed**: 1.045. H20 (DNABERT-2 |trained – untrained| < 0.05) **rejected**: $\Delta = 0.257$ (content-only). H21 (DNABERT-2 trained $\rho < 0.20$) **rejected**: $\rho = 0.257$ (content-only).

4 Discussion

The composition confound we identified—GC enrichment in stems inflating embedding-level structure metrics—is an instance of a general problem: whenever the target structure correlates with feature composition, any embedding metric sensitive to feature identity inherits the correlation. The nucleotide-stratified permutation null is a template: stratify by the confounding feature, shuffle labels within strata, and recompute the metric under the same selection procedure used on real data. The same principle applies to protein secondary structure (amino acid composition differs between helices, sheets, and coils), chromatin accessibility (GC content correlates with open chromatin), and any biological structure correlated with sequence composition. The Johnson–Lindenstrauss lemma guarantees that random projections of dimension $d \geq O(\epsilon^{-2} \log n)$ preserve pairwise distances with distortion $\leq \epsilon$ (?), meaning any subspace-distance metric computed on embeddings of dimension $d \gg \log n$ produces non-trivial scores from random representations. Composition controls are a prerequisite for interpreting geometric structure metrics.

NT v2’s attention-contact correlation ($\rho = 0.322$) was initially reported as fully learned (untrained = 0.000), a finding that appeared to demonstrate structure encoding from DNA pretraining. The corrected analysis (untrained = 0.328) reverses this interpretation: the correlation is an architectural property of the ESM attention mechanism combined with 6-mer tokenization, present from random initialization. Among the seven character-tokenized attention models (Table ??), all show trained attention within ± 0.025 of untrained. DNABERT-2 is the exception: its content-only attention ($\rho = 0.257$ trained vs 0.000 untrained) shows learned structure-aware content patterns, though the practical attention includes ALiBi position biases that dominate the computation. The SDPA bug that produced the spurious zero is a cautionary example: HuggingFace’s default attention backend silently returns `None` when `output_attentions=True`, and downstream code that interprets absent attention as zero correlation produces a qualitatively wrong result.

The two-order composition cascade validates the framework’s design. The first-order null absorbs $\sim 3\times$ of the naive signal; the second-order null absorbs what remains. The Phase 1 survivors tRNA-Ala and SAM riboswitch are both absorbed by second-order controls. No family survives both orders of composition control, confirming that the embedding-level structure signal is entirely attributable to nucleotide and dinucleotide composition.

The 2.0 threshold for H1/H6 was chosen to demand an effect large enough for practical downstream use. H6 fails while H10 passes, separating magnitude from direction: RNA-FM’s representations are consistently more structure-sensitive than random weights across 50/52 families, but the $1.77\times$ effect size is insufficient for applications requiring reliable stem-loop discrimination.

Ten models spanning attention-based, SSM-based, and hybrid architectures provide broad coverage, though representation remains thin per architecture. Phase 4 extends the null attention result to five RNA-pretrained BERT-style models: RNA-FM, RiNALMo, UTR-LM, ERNIE-RNA, and SpliceBERT all show trained attention within ± 0.025 of untrained. ERNIE-RNA’s mutation sign test (52/52, 100%) is the strongest directional result, followed by RNA-FM (50/52, 96.2%) and RiNALMo (49/52, 94.2%), consistent with RNA pretraining conferring a small directional advantage that does not cross practical thresholds. SpliceBERT and UTR-LM show minimal mutation sensitivity despite RNA pretraining, suggesting that model capacity (19.4M and 1.2M respectively) limits structure encoding. DNABERT-2’s BPE tokenization prevents per-position metrics but illustrates a broader limitation: subword tokenization breaks the nucleotide-level assumptions of the evaluation framework. The evaluation targets secondary structure only and cannot assess tertiary contacts or pseudoknots. The observation that Evo underperforms smaller models rests on a single model per size class.

5 Conclusion

A composition-controlled evaluation framework reveals that GC enrichment in stems inflates naive structure-sensitivity metrics by $\sim 3\times$. After controlling at one or two orders of sequence composition, the residual embedding-level structure signal falls below a preregistered threshold for practical significance.

No positive result survives all controls. NT v2’s attention-contact correlation ($\rho = 0.322$) is architectural: untrained weights produce $\rho = 0.328$, and the earlier report of 0.000 was an artifact of the SDPA attention backend. Seven character-tokenized attention models show trained attention indistinguishable from untrained; DNABERT-2 shows learned content attention ($\rho = 0.257$) that excludes ALiBi position biases. Phase 2 ($N = 52$) confirms: the directional effect is real (50/52, H10 PASS), the magnitude is modest (H6 FAIL at $1.77\times$), and composition controls absorb every first-order survivor. Phases 3–4 add five more models: ERNIE-RNA achieves the strongest sign test (52/52, 100%) with a mean ratio of 1.149, RiNALMo confirms a directional signal (49/52) at 1.109, and SpliceBERT and UTR-LM show no structure sensitivity (≤ 1.045).

The framework—stratified permutation nulls, max-over-layers matching, preregistered thresholds, untrained baselines—provides a reusable template for evaluating structure awareness in sequence foundation models.

Data availability

Phase 1 results: `data/gpu_results/batch3_structure_metrics/`. Phase 2 results: Modal volume `prereg-experiment-results`, path `structure_metrics/expanded_rfam/`. Phase 3 results: `data/gpu_results/expanded_rfam/` (`rinalmo_expanded*.json`, `utrlm_expanded*.json`, `nt_expanded_20260713_193419.json` with eager attention fix). Phase 4 results: `data/gpu_results/expanded_rfam/` (`ernierna_expanded_20260713_193419.json`, `splicebert_expanded_20260713_193419.json`, `dnabert2_expanded_20260713_201316.json`). RNA structures, analysis code, and preregistration documents are in the `causal-rna` and `factorization-unified` repositories. Phase 1 prereg SHA: 694b43b/a7d10f5. Phase 2 prereg SHA: bd4b3fd/74c8f49. Model checkpoints: RNA-FM (m14bio/RNA-FM), NT v2 (InstaDeepAI/nucleotide-transformer-v2-50m-multi-hyena), HyenaDNA (LongSafari/hyena-dna-small-32k-seqlen), Caduceus (kuleshov-group/caduceus-ps-seqlen-131k-base), Evo (togethercomputer/evo-1-131k-base), RiNALMo (m14bio/RiNALMo), UTR-LM (m14bio/UTR-LM), ERNIE-RNA (multimolecule/ernierna), SpliceBERT (multimolecule/splicebert), DNABERT-2 (zhihan1996/DNABERT-2-117M).

References

- Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., Zong, L., Hong, L., Xiao, J., King, I., et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv:2204.00300*, 2022.
- Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., et al. The Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv:2023.01.11.523679*, 2023.
- Nguyen, E., Poli, M., Faber, M., et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *NeurIPS*, 2023.
- Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., Kuleshov, V. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *ICML*, 2024.

- Nguyen, E., Poli, M., Durrant, M. G., et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723), 2024.
- Johnson, W. B. & Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.
- Tower, E. Boundary conditions for geometric evaluation of foundation model representations. *Working paper*, 2026.
- Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *EMNLP*, 2019.
- Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., & Rajani, N. F. BERTology meets biology: Interpreting attention in protein language models. In *ICLR*, 2020.
- Frazer, J., Notin, P., Dias, M., et al. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.
- Karlin, S. & Buré, C. Dinucleotide relative abundance extremes: A genomic signature. *Trends in Genetics*, 14(10):398–402, 1998.
- Penić, R. J., Vlah, D., Huber, R., Heider, D., & Šikić, M. RiNALMo: General-purpose RNA language models can generalize well on structure prediction tasks. *arXiv:2403.00043*, 2024.
- Chu, Y., Yu, D., Li, Y., Huang, K., Shen, Y., Cong, L., Zhang, J., & Wang, M. A 5' UTR language model for decoding untranslated regions of mRNA and function predictions. *Nature Machine Intelligence*, 6:449–460, 2024.
- Wang, N., Bian, J., Li, J., et al. ERNIE-RNA: An RNA language model with structure-enhanced representations. *arXiv:2305.14196*, 2024.
- Chen, K., Zhou, Y., Ding, M., Wang, Y., Ren, Z., & Yang, Y. Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction. *Briefings in Bioinformatics*, 25(1), 2024.
- Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. V., & Liu, H. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. In *ICLR*, 2024.

A RNA family details ($N = 52$)

Phase 1 families (12, from original preregistration):

Expanded families (40 additional):

[TODO: Add Rfam accession IDs and GC% columns from family metadata.]

Family	Stem	Loop	RNA-FM ratio	> nuc null
tRNA-Phe (yeast)	42	34	2.758	No
tRNA-Ala (human)	42	34	3.208	Yes
5S rRNA	72	48	1.158	No
Hammerhead	16	32	1.328	No
SAM riboswitch	52	42	2.535	No
TPP riboswitch	40	40	1.442	No
SRP RNA	44	33	1.261	No
miR-21	64	8	1.299	No
IRES HCV dom. II	50	27	1.710	No
HDV ribozyme	48	28	1.135	No
RNase P	42	36	1.513	No
U2 snRNA	46	30	2.200	No

Family	Class	Stem	Loop	RNA-FM ratio	> nuc null
6S RNA	ncRNA	42	34	1.823	No
7SK RNA	ncRNA	38	31	1.593	No
Bacterial SRP	ncRNA	43	22	2.179	No
c-di-GMP ribosw.	Riboswitch	36	27	1.826	No
Cobalamin ribosw.	Riboswitch	38	36	2.214	No
Corona 5' UTR	Cis-reg	51	33	1.252	No
Corona s2m	Cis-reg	56	27	1.268	No
CRISPR leader	CRISPR	25	30	1.121	No
CrPV IRES	IRES	40	24	1.575	No
Fluoride ribosw.	Riboswitch	36	28	1.473	No
FMN riboswitch	Riboswitch	52	26	1.889	No
glmS ribozyme	Ribozyme	44	38	1.252	No
Glycine ribosw.	Riboswitch	26	32	1.167	No
Group I intron P4P6	Intron	46	31	1.827	No
Group II intron D5	Intron	38	21	1.011	No
Hatchet ribozyme	Ribozyme	26	31	1.865	No
HCV IRES III	IRES	40	24	1.480	No
Histone 3' stem	Cis-reg	10	10	4.078	No
IRE stem-loop	Cis-reg	21	9	2.006	No
Lysine ribosw.	Riboswitch	40	38	1.581	No
Manganese ribosw.	Riboswitch	40	24	1.008	No
miR-122	miRNA	44	28	1.252	No
miR-155	miRNA	44	22	2.529	No
miR-let-7	miRNA	48	27	0.952	No
Pistol ribozyme	Ribozyme	28	38	1.673	No
preQ1 ribosw.	Riboswitch	26	25	2.757	No
Purine ribosw.	Riboswitch	36	21	2.353	No
SAH riboswitch	Riboswitch	32	28	1.252	No
SECIS element	Cis-reg	30	28	1.166	No
T-box leader	Cis-reg	38	26	2.651	No
THF riboswitch	Riboswitch	52	30	1.150	No
tmRNA	ncRNA	56	29	2.682	No
Twister ribozyme	Ribozyme	30	37	1.584	No
U1 snRNA	snRNA	44	31	1.721	No
U4 snRNA	snRNA	38	34	2.101	No
U5 snRNA	snRNA	44	18	2.864	No
U6 snRNA	snRNA	40	30	1.091	No
Vault RNA	ncRNA	44	29	1.264	No
Y RNA	ncRNA	36	32	1.979	No
ZMP riboswitch	Riboswitch	30	26	1.961	No