RNA FOUNDATION MODELS RESPOND TO BIOLOGY THEIR EXPLANATIONS MOSTLY MISS

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying which nucleotides drive an RNA model's prediction enables biological discovery beyond prediction alone. Attribution methods provide this insight by scoring each nucleotide's contribution. Their genomic validation has mostly been on task-specific convolutional networks (CNNs) trained from scratch, yet these methods are now applied to pretrained transformer RNA foundation models (FMs). Whether they expose what an RNA FM responds to has not been rigorously tested against labelled biological signal. We use in-silico mutagenesis (ISM), an exhaustive single-substitution scan of the model, to score against the same labels as every method, which we term grounding. ISM shows how much of the labelled signal each model demonstrably uses, a reference every method can be read against. Across six conventional attribution methods, a transformer-specific propagation method (AttnLRP), six RNA FMs (20M–650M parameters) and six RNA tasks, ISM substantially recovers the biological signals on four tasks, yet the conventional methods recover this signal only partially, and recovery varies substantially across FMs of near-identical task performance. We also show that RNA FMs' own substitution responses identify structural base pairs that the conventional methods do not consistently expose. These findings motivate validating each attribution method on the particular fine-tuned model before drawing biological conclusions.

RNA models predict sequence function and structure with increasing accuracy (de Sainte Agathe et al., 2023; Sample et al., 2019), but a prediction alone does not reveal which parts of the sequence drove it. Attribution methods address this by assigning each nucleotide a contribution to the model output. In genomics these per-position maps allow for biological insight into new data, including recovering motifs, identifying regulatory grammar and prioritising variants (Alipanahi et al., 2015; Shrikumar et al., 2018; Avsec et al., 2021b; Zhou & Troyanskaya, 2015; Movva et al., 2019; Novakovsky et al., 2023). RNA foundation models (FMs), frequently transformer-encoders pretrained on large sequence corpora and fine-tuned downstream (Chen et al., 2022; 2024; Yin et al., 2025; Yang et al., 2025; Penić et al., 2025; Zou et al., 2024), are becoming a popular approach for modelling genomics (Consens et al., 2025; Wang et al., 2025; Shu et al., 2026).

However, the validation of conventional attribution methods and toolkits (Kokhlikyan et al., 2020; Schreiber, 2025) has focussed on CNN architectures (Koo & Eddy, 2019; Koo & Ploenzke, 2021; Majdandzic et al., 2023); whether the methods reliably expose which nucleotides drive a transformer-based RNA FM's prediction has not been rigorously tested. Furthermore, early evidence suggests their application faces challenges for genomic transformer models (Zhou et al., 2026). Common method evaluation grounds an attribution map against experimentally validated biological signals within a sequence, such as motifs (Koo & Ploenzke, 2021; Majdandzic et al., 2023), but annotations are incomplete: a low grounding score may mean the model relies on unmeasured signal, or that the method missed signal the model uses. Evaluation on FMs must therefore distinguish these.

We separate these possibilities using in-silico mutagenesis (ISM) (Zhou & Troyanskaya, 2015; Alipanahi et al., 2015), an exhaustive scan of all single-nucleotide substitutions, grounded against the same annotations as the attribution methods. Because each substitution directly intervenes on the model, we do not use ISM as ground truth but as a reference for the quantity of measured signal the model demonstrably responds to. Both ISM and the methods are read against a positional floor: the grounding value a sequence-blind explanation achieves when only the average location of annota-

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.

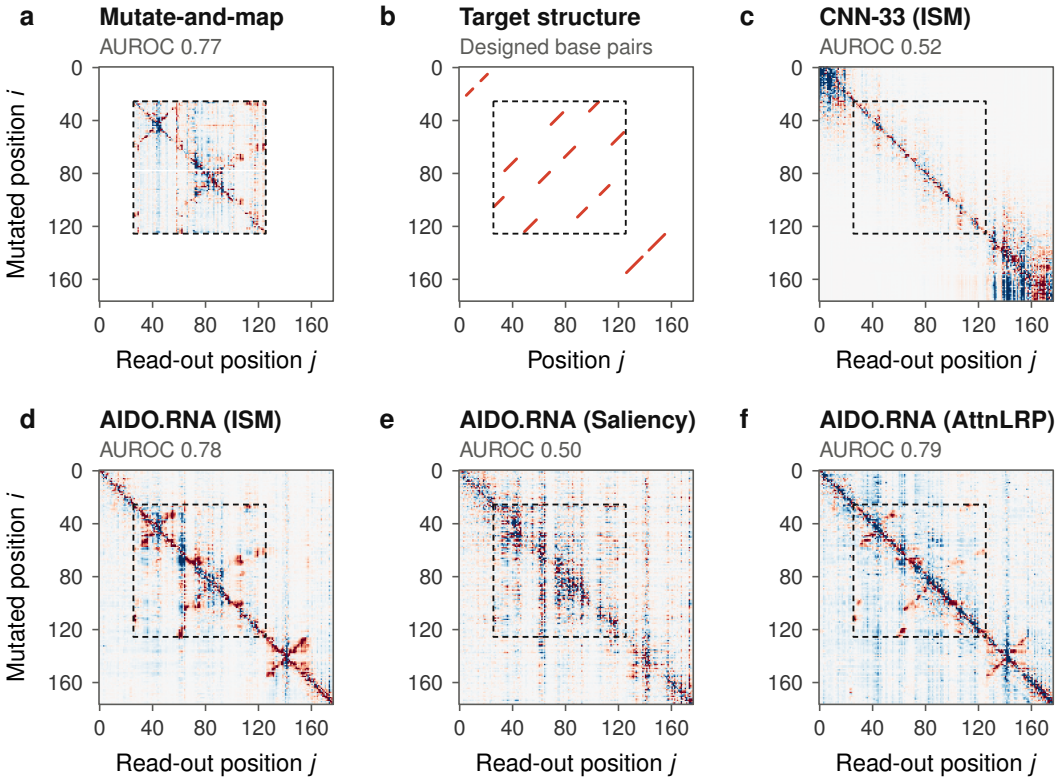


Figure 1: RNA FMs learn structure that their attributions miss. Each map relates positions i and j of a designed RNA; base pairs appear as the off-diagonal stripes of the target structure (*b*). The lab experiment (*a*) and the FM's own mutation scan (ISM, *d*) and AttnLRP (*f*) recover these pairs. A standard attribution method (Saliency, *e*) does not. A CNN trained on the same task recovers very little of them (*c*). AUROC measures pair recovery against non-pairs matched on distance and partner letter.

tions in the held-out data is scored. A method's deficit between its own grounding and ISM's reflects biological signal that the model uses but the explanation fails to recover. Where ISM's grounding is lower than the floor, the signal is uninformative for that model.

We evaluate seven attribution methods, six conventional gradient- and backpropagation-based methods and one with relevance-propagation rules for attention and normalisation layers, AttnLRP (Achtibat et al., 2024). We evaluate across six fine-tuned transformer-based RNA FMs and five functional and one structural RNA task, each with measured biological signal and a positional floor. On the structure task, we also read each FM's attention weights as a map. Because practitioners rarely have annotations for new data, we also measure faithfulness by comprehensiveness and sufficiency (DeYoung et al., 2020). Both are measured on real sequences and on dinucleotide-shuffled nulls that preserve local composition while disrupting the biological signal (Altschul & Erickson, 1985; Schreiber, 2025).

Of the six tasks, ISM exceeds the positional floor on four, for every FM; the attribution methods recover this biology only partially and inconsistently. Across the four tasks, the median over method–FM–task combinations of each method's grounding as a share of the same FM's ISM grounding is 60%, and its share of ISM's gain over the positional floor is 11%. While AttnLRP is the most grounded method on three of the tasks, it is the least grounded on another, falling below the floor for half of the models. The grounding of an attribution method therefore strongly depends on the model it explains, and a method validated on one model cannot be assumed to transfer to another.

A method that misses the labelled motifs might still recover relational signal, such as base pairing, that no position-wise annotation can express. The structure task, which scores designed base pairs,

tests this directly (Figure 1). The FMs' own ISM maps recover the base pairs, as does last-layer attention on five of the six; AttnLRP exceeds the model's own ISM on two. Yet the conventional methods recover substantially less, the best reaching $\sim$0.60 AUROC. We also identify practical challenges when applying conventional methods to transformers. For five of six FMs, the reference implementations of DeepLIFT (Shrikumar et al., 2017; Kokhlikyan et al., 2020; Schreiber, 2025) attach no propagation rules to the encoders' activations and consequently reduce to Saliency (Appendix D).

We summarise our contributions below:

1. We distinguish two failures of attribution evaluation by employing ISM as a grounding reference. We demonstrate that six encoder RNA FMs respond to biological signals, and that conventional and transformer-specific methods miss a substantial share.
2. We show that RNA FMs' learned base-pairing structure can be identified with ISM and AttnLRP, while conventional methods recover far less, three of them at chance.
3. To support model-method evaluation, we release all our processed public data, with labelled biological signal, and fine-tuned checkpoints with evaluation code.

1 Related Work

Genomic Attribution. In genomics, an attribution method's explanation is read as biology. Examples include substitution scans of a CNN scoring the regulatory effect of noncoding variants (Zhou & Troyanskaya, 2015), DeepLIFT maps recovering transcription-factor motifs (Shrikumar et al., 2017; Avsec et al., 2021b), or gradient methods for gene expression (Avsec et al., 2021a). A wrong explanation can therefore become a wrong biological claim. An explanation can be faithful to the model's computation or plausible to a human, but the two are independent (Jacovi & Goldberg, 2020). In genomics, where no human judge of plausibility is available, explanations are grounded against labelled biological signals in the sequences, either with known motifs or synthetic sequences with planted motifs (Koo & Ploenzke, 2021; Prakash et al., 2022).

Explanation failures. However, biological annotation is incomplete, so poor grounding does not uniquely identify a failure of the attribution method, as the model may instead rely on signals outside the available annotations. Attribution maps can also vary substantially between models with similar predictive performance (Majdandzic et al., 2022), making explanations from one model an unreliable proxy for another. Faithfulness provides a complementary evaluation by measuring how model outputs change when highly attributed positions are removed or retained (DeYoung et al., 2020), although both attribution methods and faithfulness metrics have exhibited failures under sanity checks (Adebayo et al., 2018; Tomsett et al., 2020). Together, these results leave unresolved whether poor biological grounding reflects the explanation method, the particular model being explained, or both. Perturbation-based approaches such as CREME interrogate what a model has learned through controlled sequence interventions (Toneyan & Koo, 2024).

Transformer attribution. Transformer-based foundation models are replacing task-specific CNNs across genomics (Consens et al., 2025), including for RNA. The explanations released with them are often attention maps (Yu et al., 2024; Ji et al., 2021), although attention weights are not generally reliable accounts of feature importance (Jain & Wallace, 2019), and gradient or backpropagation-based explanations can become unreliable around attention and layer normalisation, which motivated propagation rules designed for transformers (Ali et al., 2022; Achtibat et al., 2024). Whether standard genomic attribution transfers reliably to pretrained transformer architectures has so far been studied primarily in DNA, where attribution maps of pretrained models have been found to be diffuse (Tang et al., 2025), and different interpretability methods can disagree and fail to recover validated regulatory motifs (Zhou et al., 2026). These studies expose failures of genomic attribution, but do not distinguish signal missed by the explanation from signal absent from the model's own response. We address these gaps by using exhaustive single-nucleotide perturbations grounded against the same annotations as the measured attribution method. By also measuring against a positional floor, we can identify the minimum grounding score a method must achieve to qualify as evidence a model has learnt the biological signal and did not depend on position alone. We extend the evaluation to an RNA structure-based task, measuring the capacity to recover pair-interactions within a sequence.

2 BACKGROUND

2.1 RNA REPRESENTATION

Each sequence $s \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{U}\}^L$ is one-hot encoded as $X_0 = \phi(s) \in \mathbb{R}^{L\times 4}$, and every attribution method returns a map $M \in \mathbb{R}^{L\times 4}$, with $M_{i,a}$ the score for letter a at position i (Appendix B). Following Majdandzic et al. (2023), every map, including ISM, is mean-centred across letters at each position, removing a component that no real substitution can probe (Appendix G). Each position is scored using the centred map at its observed letter s_i. For a gradient method, the contrast $M_{i,a} - M_{i,s_i}$ is then a first-order estimate of the effect of substituting a at i, which ISM measures exactly.

2.2 ATTRIBUTION METHODS

We evaluate six widely used, gradient and backpropagation-based attribution methods and compare them with one method designed specifically for transformers. All seven are formalised in Appendix C and described briefly below.

Saliency, also known as vanilla gradients, is the gradient of the model's output with respect to its input, $\nabla_X f$. **Input** $\times$ **Gradient**, $X_0 \odot \nabla_X f$, weights the gradient by its input.

SmoothGrad (Smilkov et al., 2017) adds Gaussian noise to multiple versions of the input and averages their gradients, which smooths the map.

GradientSHAP (Lundberg & Lee, 2017) approximates SHAP values by calculating the gradient of the model's output with respect to sampled points from a straight-line path between dinucleotide-shuffled baseline values and the input data, with added Gaussian noise.

Integrated Gradients (Sundararajan et al., 2017) addresses saturation by integrating gradients along a straight path to the input from a uniform baseline that gives each letter $1/4$.

DeepLIFT (Shrikumar et al., 2017) propagates contributions backward with layer-wise rules, comparing each activation with its value on the same uniform baseline.

AttnLRP (Achtibat et al., 2024) is a layer-wise relevance propagation method with rules written for transformers. AttnLRP propagates relevance through the attention products, normalisers and the element-wise activations with a rule for each.

2.2.1 IN-SILICO MUTAGENESIS (ISM)

ISM is the computational analogue of saturation mutagenesis (Patwardhan et al., 2009). It measures the change in a model's prediction for each single-base substitution (Appendix F). ISM can itself serve as a model-agnostic attribution method, but it needs $3L + 1$ forward passes per sequence, far more than any gradient method.

2.3 RNA FOUNDATION MODELS AND TASKS

We select six RNA tasks where their labelled signals are sparse (unlike, e.g., secondary-structure prediction, where every position matters (Lorenz et al., 2011)) and known to drive the label, so that the model is likely to learn the signals. Selected tasks are binding-site prediction (BBP), mean ribosome load (MRL), splice-site prediction (SPL), polyadenylation signal (PAS), N^6-methyladenosine (m6A) and chemical reactivity (CRP). Detailed descriptions are available in Appendix J. We use six encoder-based RNA foundation models spanning a range of depths and parameter counts (20M–650M): SpliceBERT (Chen et al., 2024), RNA-FM (Chen et al., 2022), ERNIE-RNA (Yin et al., 2025), OmniGenome (Yang et al., 2025), RiNALMo (Penić et al., 2025) and AIDO.RNA (Zou et al., 2024) (Table 8). For each task, each model was fine-tuned end-to-end from its pretrained weights with a randomly initialised task-specific linear head. Attributions were evaluated on held-out data. Further model and training details can be found in Appendix M. Across tasks the six models' scores were tightly clustered. Across models, the standard deviation within a task had a median of 0.015 and never exceeded 0.059 (PAS). On four of the six tasks, all models achieved substantial scores, but performance was lower on PAS and m6A, as shown in Table 11.

Table 1: At the median across backbones, every method grounds below ISM, and on four of five tasks the most faithful is not the most grounded. Medians over six backbones on the 400 held-out sequences per task. Grounding (G; AUPRC, or for CRP base-pair AUROC over 131 designs with non-pairs matched on distance and partner letter) is read against ISM's grounding and the task's positional floor (CRP 0.5). SPL's floor is 1.000, so SPL enters no aggregate. On CRP, Integrated Gradients, GradientSHAP and SmoothGrad are scored on a fixed subset of 36 designs (Appendix J). Faithfulness (F) is comprehensiveness. Most faithful per task in bold, most grounded underlined.

	BBP		MRL		SPL		CRP		PAS	
Method	$G\uparrow$	$F\uparrow$	$G\uparrow$	$F\uparrow$	$G\uparrow$	$F\uparrow$	$G\uparrow$	$F\uparrow$	$G\uparrow$	$F\uparrow$
ISM (reference)	0.750	1.435	0.757	1.481	0.420	0.973	0.755	1.395	0.225	2.030
Floor / random ranking	0.275	0.383	0.210	0.470	1.000	0.323	0.500	0.445	0.138	0.359
Integrated Gradients	0.590	1.181	<u>0.592</u>	**1.178**	<u>0.364</u>	0.823	0.605	0.998	0.082	1.155
GradientSHAP	0.504	**1.255**	0.449	1.170	0.362	**0.848**	0.599	**1.111**	0.088	**1.750**
SmoothGrad	0.531	1.013	0.349	0.931	0.205	0.728	0.549	0.781	0.119	1.100
Input $\times$ Gradient	0.374	0.949	0.298	0.713	0.179	0.468	0.515	0.456	0.095	0.842
DeepLIFT	0.337	0.680	0.277	0.717	0.157	0.434	0.511	0.520	0.069	0.640
Saliency	0.327	0.638	0.276	0.702	0.157	0.434	0.511	0.521	0.069	0.640
AttnLRP	<u>0.617</u>	1.136	0.242	0.665	0.265	0.826	<u>0.635</u>	0.600	<u>0.150</u>	1.561

3 EXPERIMENTS

Our evaluation asks whether attribution methods recover biological signals to which the model demonstrably responds, whether their recovery is consistent across similarly performing models, and whether faithfulness can identify explanations that recover those signals. For the tasks BBP, MRL, SPL and PAS, grounding was measured as AUPRC between a method's attributions and labelled signals. For CRP the positive and negative classes were equally balanced and thus AUROC was used. Because the labelled signals are incomplete, we ground ISM maps against the same labelled signals to provide a reference for how each model demonstrably responds to substitutions in the sequence. Because a model can score well by relying on position alone, we also report a positional floor: the grounding of a sequence-blind scorer that assigns every sequence the average signal rate at each position (Appendix I). A task contributes to the aggregate comparison, the medians over method–FM–task combinations, only where ISM clears its floor. For SPL and m6A, we found that the labelled signal cannot ground explanations above the tasks' positional floor, and thus exclude them from aggregation (Appendix J).

We measure faithfulness by comprehensiveness and sufficiency (DeYoung et al., 2020). We rank positions by a method's signed attribution, then either delete the top 1–75% or retain only those positions and record the resulting change in output. When a base is deleted it undergoes shuffled-base replacement: it is replaced by the base at that position in a dinucleotide shuffle of the sequence, averaged over three shuffles. Both scores are divided by the change in output when the whole sequence is shuffled, so that tasks with different output units can be pooled; a random ranking is scored alongside as the reference. Faithfulness is measured on a fixed, stratified draw of 400 held-out sequences per task (the 400-sequence draw), shared by every model and method (split-half reliability 0.995), and on each sequence's own dinucleotide-shuffled null, which disrupts the biological signal (Table 3, Appendix H).

3.1 ATTRIBUTIONS RECOVER ONLY A FRACTION OF SIGNAL COMPARED TO ISM

On the four tasks where ISM exceeds the positional floor for every FM, attribution methods recover substantially less of the signal than the models' own substitution responses. Across the seven methods, median grounding is 60% of ISM's grounding, while the median recovery of ISM's gain over the floor is only 11%. Thus, low attribution grounding cannot be explained solely by an absence of model responses at labelled positions. m6A was designed as a negative control where both classes carry the central $DRACH$, so its presence should be uninformative. Because the motif sits at the same positions in every window, a sequence-blind scorer already reaches 0.83 and no method or ISM scores higher. The control also caught a flaw in the data: a U at the motif's first position occurs

only in decoys, and all six models learned it. On PAS the signal is informative, narrowly: ISM clears the floor on all six FMs, while most conventional methods fall to or below it (Table 1).

Where the models demonstrably respond to the labelled signal, the methods return only a fraction of ISM's grounding. Across the four tasks the seven methods recover a median 60% of ISM's grounding and 11% of its gain over the floor, and 59 of the 168 method–FM–task combinations score at or below the floor itself (Figure 3, Table 1). The size of the deficit from ISM's grounding depends on the specific method (Table 1): Integrated Gradients achieves the highest grounding out of conventional methods on BBP and MRL, while Saliency and DeepLIFT recover the least and sit at or below the floor most often. Transformer-specific propagation rules do not remove the deficit: AttnLRP recovers 80% of ISM's grounding and 50% of its gain over the floor, and is the most grounded method on BBP, yet it is substantially less grounded than Integrated Gradients on MRL. This deficit leaves two questions: whether the methods instead identify unlabelled signals, and whether their grounding varies because the models differ in their responses to the labelled biology.

3.1.1 Attributions of different FMs do not agree on where unlabelled signal lies

If differences in attribution grounding primarily reflect differences in the models' use of labelled biology, grounding should vary alongside the models' own substitution responses. A method correctly explaining a model may appear to have poor grounding simply because we don't measure the explained signal. To test this, we measure where attributions fall along each sequence and how well the six FMs' attributions agree outside the labelled signal, on the 400-sequence draw. If the methods were correctly attributing to unmeasured signal that the models share, the FMs' attributions should agree outside the labelled positions.

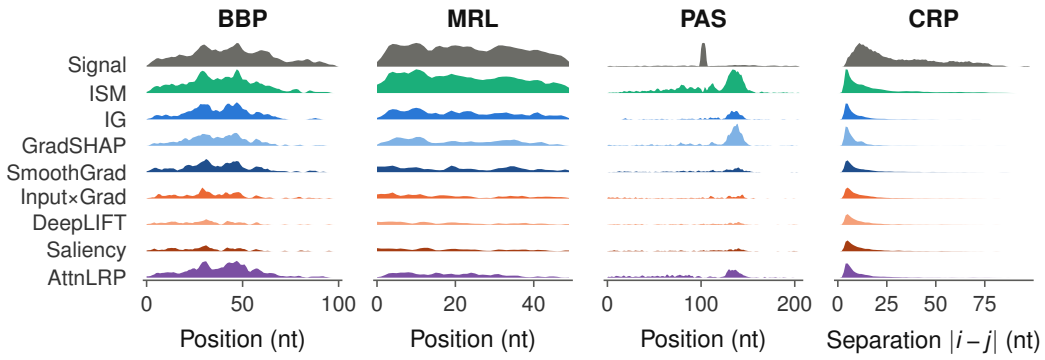


Figure 2: Consensus positions (in the top 10% of at least four of the six FMs) of each method, per sequence of the 400-sequence draw of each task; CRP uses the 36-design subset, where every method's pair maps exist. Signal: the labelled positions of the same sequences; on PAS it includes the control hexamer at positions 100–105. Method rows share one vertical scale per task, the signal row has its own. CRP's axis is the separation $|i - j|$ of a pair.

For each method and sequence, we take each model's 10% of positions with the largest attribution magnitude. A position that at least four of the six models pick is a consensus position: a place where the models agree, whether or not it is labelled. To measure agreement outside the signal, we repeat the selection over the unlabelled positions only and, for every pair of models, measure the Jaccard overlap of their two sets; two random sets of this size overlap by 0.05. On CRP a position is a pair (i, j) of the $L \times L$ map, and the signal is the target base pairs.

On BBP and MRL, the methods' consensus follows the labelled signal along the sequence (Figure 2). On unlabelled positions, the FMs rarely agree: top attribution overlap by only 0.07–0.13 (Table 4), and ISM's overlap is higher on every task. ISM computed on the same FMs overlaps far more, and on BBP and MRL the few unlabelled positions the methods share sit within 3 nucleotides (nt) of a labelled position far more often than unlabelled positions overall (Table 4). Models relying on different unlabelled signals would show the disagreement in ISM as well; they do not. This makes that reading less plausible without excluding it. On PAS every method's consensus peaks 30–40 nt

from the sequence's hexamer, no nearer the labels than chance: FMs' explanations agree on what causes the prediction but the labels do not cover it.

3.1.2 Attribution grounding varies across models despite similar task performance

A method may be explaining a model correctly, but if the model does not use the labelled biological signal, even an optimal method need not exceed the positional floor. We examine grounding compared to per-task performance. On BBP and MRL model performance varied by no more than 2.4% (coefficient of variation (CoV) ≤ 0.008, Appendix A), and on CRP MAE spans only 0.22–0.29. ISM also recovers signal consistently across all four tasks. Yet the methods' grounding varies across these FMs far more than ISM's does: CoV is 7.6 times ISM's (Figure 3). While AttnLRP is comparatively tightly clustered on BBP, similar variability is observed on the remaining tasks. A method's grounding on one model therefore does not reliably transfer to another of similar task performance; what needs validating is the method–model pair.

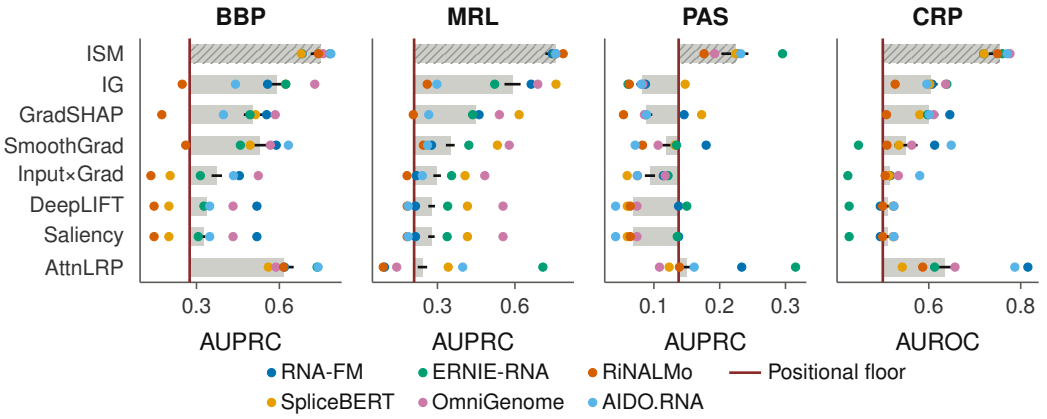


Figure 3: Methods' explanations only recover a fraction of ISM's identified signal. Dark red: positional-prior floor. Bars run from the floor to the median over six models. Circles are the individual models. SPL and m6A are excluded, as neither ISM nor the models' grounding passed their tasks' positional floor.

3.2 Faithfulness Identifies Better Methods, Not Better Explanations

Without labelled signal or ground truths, faithfulness is one of few quantitative checks available to a practitioner to evaluate whether to trust an attribution method's explanation. We test this application of faithfulness by evaluating whether the metric can predict a better grounded method or certify whether a sequence's explanation is grounded (Figure 4).

Across methods, the relation holds; across sequences, it is weak. On a fixed sequence, methods rank similarly by faithfulness and by grounding, and for two-thirds of method pairs the more faithful method is also the more grounded; on the sequences where ISM shows the model uses the labelled biology, this rises to 72%. A more faithful explanation is only slightly more likely to be grounded (Spearman $\rho = 0.27$), and this does not improve on ISM-grounded sequences (Figure 4). Faithfulness can guide the choice of method, but cannot certify an explanation. Moreover, when each sequence is paired with its own dinucleotide-shuffled null, normalised faithfulness on the shuffled inputs matches or exceeds that on real sequences for most method–FM–task groups (Table 3). Faithfulness measures the model's response, not whether that response is biological.

3.3 RNA FMs Learn Structure That CNNs and Conventional Attributions Miss

The structure task provides a direct test of whether explanations recover a biologically meaningful relationship expressed in the model's substitution responses. We trained four models from scratch, three CNNs and a small transformer, to within 10% of the weakest FM's

performance on BBP, MRL and CRP, and examined all methods' grounding and faithfulness against 400 held-out sequences per model (Appendix M). On BBP, MRL and CRP conventional methods better explain the three CNNs trained from scratch: the attributions are more faithful (higher comprehensiveness on 18/18 task–method pairs, lower sufficiency on 15/18) and, on BBP and MRL, better grounded than on any FM (Figure 5a, Table 9).

CRP grounds each model's cross-position substitution responses against designed base pairs, with non-pairs matched on distance and partner letter, so a scorer that reads only pairing rules scores exactly at chance (Appendix J). Read this way, the FMs' own ISM maps recover the designed pairs at the level of the mutate-and-map assay: level with it on four of the six FMs and slightly below on the remaining two (Figure 5). The models trained from scratch sit just above chance (0.52–0.58) and recover little of the structure present in their training data. Base pairing is therefore learned by the FMs and is demonstrably expressed in their substitution responses.

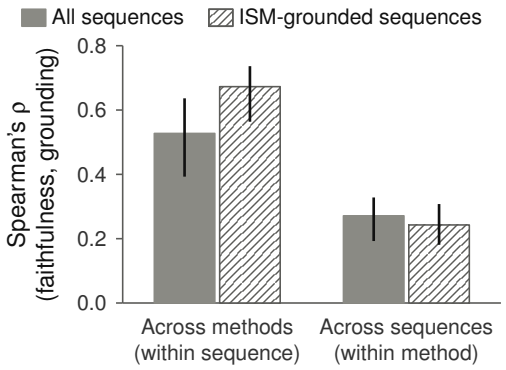


Figure 4: Median Spearman ρ between comprehensiveness and grounding over the seven methods on BBP, MRL and PAS, with 95% bootstrap intervals over sequences (Appendix A). ISM-grounded sequences are the third of sequences on which ISM's own map correlates best with the labels.

The conventional methods expose less of it. Saliency, DeepLIFT and Input $\times$ Gradient score at chance, and Integrated Gradients and GradientSHAP score below the model's own ISM (Table 1). AttnLRP recovers the pairs on RNA-FM and AIDO.RNA (0.82 and 0.79), where it exceeds that model's own ISM, yet scores near chance on SpliceBERT and below ISM on the other three. Attention read directly from the last-layer estimator recovers the pairs on five of six FMs, although the four estimators disagree substantially (Appendix L). Even before fine-tuning, last-layer attention recovers pairing on four FMs, indicating that much of the structural signal predates fine-tuning. The recovery belongs to the method–model pair, not to the method.

4 Discussion and Limitations

Our results show why biological validation of an explanation must be specific to the model being explained. On the four tasks where ISM exceeds the positional floor, the evaluated RNA FMs respond to labelled signals that attribution methods recover incompletely. Similar predictive performance does not imply similar attribution grounding, and faithfulness provides limited evidence that an individual explanation recovers those signals. Biological interpretation therefore requires validation of the method–model pair. Using a positional floor also diagnosed the labels themselves. SPL's labels are fixed offsets from the site being scored, and m6A's motif sits at the same positions in every window, behind which ISM exposed a class-separating position.

Gradient methods read the local geometry of a particular fine-tuned model. That geometry may depend on architecture, pretraining and fine-tuning. Yet a method's grounding moves far less between its two fine-tuning seeds than it spreads across the four smaller models (Table 12): the variation lies in the models themselves, not in fine-tuning noise, motivating validation of each method–checkpoint pair. On the structure task, the models' substitution responses recover designed base pairs that the conventional methods largely miss: the biology is present, demonstrably used, and missing from the explanations. We did not test the method–model relationship beyond the chosen architectures and encourage a wider exploration of models in future work.

Faithfulness tracks the model's computation, regardless of the presence of biological signal. It separates weaker methods from stronger ones but cannot certify an explanation. ISM mitigates this as a black-box check of what the model responds to under single-nucleotide substitutions, but it requires

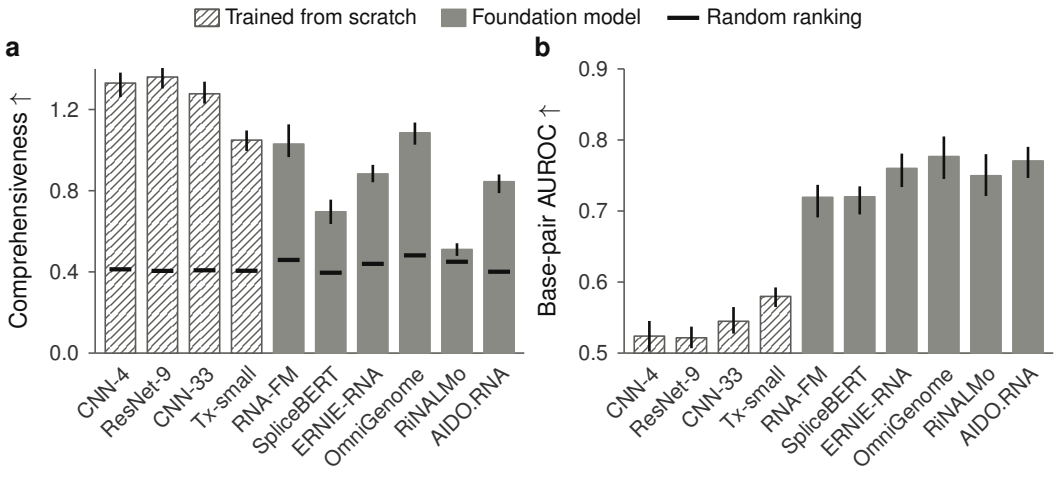


Figure 5: RNA FMs' substitution responses reveal base pairing, but their attributions are less faithful than those of models trained from scratch. Models are ordered from left to right: CNNs, a small transformer, and RNA FMs. (a) Median comprehensiveness of the six conventional methods on BBP, MRL and CRP. Ticks mark a random ranking. (b) Base-pairs identified by each model's ISM map on CRP. Vertical black lines: 95% bootstrap intervals.

$3L + 1$ forward passes per sequence. In practice, this may be computationally prohibitive. Therefore, prior to selecting an attribution method we recommend that first, a practitioner should verify that an implementation attaches its rules to the architecture, since DeepLIFT reduced to Saliency on five of six FMs. Use faithfulness to discard weak methods and read the model's own substitution response on a subset of data, against the positional floor where annotation exists, and repeat this for every method–model pair, since a method's grounding did not carry over between FMs.

4.1 Limitations

- RNA FMs may learn biological mechanisms we neither know of nor annotate, so a low grounding score is weak negative evidence as a method may miss what we annotated while recovering real signal we did not.
- ISM probes only single substitutions, so it can miss epistatic effects and under-read redundant motifs. Pairwise effects are tested only on CRP.
- CRP's grounding uses synthetic sequences designed to fold into target structures; whether the result holds for natural RNA is untested.
- All RNA FMs share a similar BERT-style encoder architecture. Although this is the prevalent design for RNA FMs, whether our results hold for other architectures, such as decoders or state-space models, is untested.

AI use statement

In this work, we used generative AI tools to support literature discovery, code generation, workflow management and manuscript preparation. We did not use AI for data generation or to conduct exhaustive literature searches.

Ethics statement

This work performs no new human- or animal-subject experiments and evaluates computational methods using previously released RNA datasets and models. A primary concern motivating this study is the risk of interpreting model explanations as biological evidence when those explanations do not reliably reflect the signals used by the model. Our results therefore support validating attri-

bution methods on each model before drawing biological conclusions. We make no clinical claims from the evaluated models or explanations.

Reproducibility statement

We provide the methodological and experimental details required to reproduce our evaluation throughout the main text and appendices. The appendices specify the input representation and attribution methods; implementation and validation details for the evaluated methods; in-silico mutagenesis, gauge correction, null construction and grounding procedures; the six RNA tasks and their biological annotations; attention-based analyses; and the architectures, fine-tuning procedure, random seeds and held-out performance of all evaluated models. We release the datasets, model checkpoints and code required to support reproduction and further evaluation.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *ICML'24: Proc. of the 2024 International Conference on Machine Learning*, pp. 135–168. PMLR, 2024.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS'18: Proc. of the 2018 Conference on Neural Information Processing Systems*, pp. 9525–9536, 2018.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for transformers: Better explanations through conservative propagation. In *ICML'22: Proc. of the 2022 International Conference on Machine Learning*, pp. 435–451. PMLR, 2022.

Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, 33 (8):831–838, 2015.

S F. Altschul and B W. Erickson. Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.*, 2 (6):526–538, 1985.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, 18(10):1196–1203, 2021a.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription factor binding reveal soft motif syntax. *Nat. Genet.*, 53:354–366, 2021b.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF Int. Con. on Comp. Vis. (ICCV)*, pp. 387–396, 2021.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *bioRxiv*, 2022. doi: 10.1101/2022.08.06.503062.

Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, and Yuedong Yang. Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. *Brief. Bioinform.*, 25(3), 2024.

Micaela E. Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J. Theis, Alan Moses, and Bo Wang. Transformers and genome language models. *Nat. Mach. Intell.*, 7(3):346–362, 2025.

Jean-Madeleine de Sainte Agathe, Mathilde Filser, Bertrand Isidor, Thomas Besnard, Paul Gueguen, Aurélien Perrin, Charles Van Goethem, Camille Verebi, Marion Masingue, John Rendu, Mireille Cossée, Anne Bergougnoux, Laurent Frobert, Julien Buratti, Élodie Lejeune, Éric Le Guern, Florence Pasquier, Fabienne Clot, Vasiliki Kalatzis, Anne-Françoise Roux, Benjamin Cogné, and David Baux. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. *Human Genomics*, 17(1):7, 2023.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *ACL'20: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Tanja Gesell and Stefan Washietl. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinform.*, 9:248, 2008.

Shujun He, Rui Huang, Jill Townley, Rachael C. Kretsch, Thomas G. Karagianes, David BT. Cox, Hamish Blair, Dmitry Penzar, Valeriy Vyaltsev, Elizaveta Aristova, Arsenii Zinkevich, Artemy Bakulin, Hoyeol Sohn, Daniel Krstevski, Takaaki Fukui, Fumiya Tatematsu, Yusuke Uchida, Donghoon Jang, Jun Seong Lee, Roger Shieh, Tom Ma, Eduard Martynov, Maxim V. Shugaev, Habib Bukhari, Kazuki Fujikawa, Kazuki Onodera, Christof Henkel, Shlomo Ron, Jonathan Romano, John J. Nicol, Grace P Nye, Yuan Wu, Christian Choe, Walter Reade, Eterna participants, and Rhiju Das. Ribonanza: Deep learning of RNA structure through dual crowdsourcing. *bioRxiv*, 2024. doi: 10.1101/2024.02.24.581671.

htm042, Jack Cole, keli1985, Ma Wanli233, and YuanLi. OmniRNA challenge. `https://www.kaggle.com/competitions/omni-rna-challenge`, 2026. Kaggle. Accessed 20 July 2026.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *ACL'20: Proc. of the 2020 Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. ACL, 2020.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL'19: Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 3543–3556. ACL, 2019.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, 2002.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Peter K Koo and Sean R Eddy. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput. Biol.*, 15(12):e1007560, 2019.

Peter K. Koo and Matt Ploenzke. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nat. Mach. Intell.*, 3(3):258–266, 2021.

Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA package 2.0. *Algorithms Mol. Biol.*, 6:26, 2011.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS'17: Proc. of the 2017 Conference on Neural Information Processing Systems*, pp. 4765–4774, 2017.

Arturo Magana-Mora, Manal Kalkatawi, and Vladimir B Bajic. Omni-polyA: a method and tool for accurate recognition of poly (A) signals in human genomic DNA. *BMC Genomics*, 18(1):620, 2017.

Antonio Majdandzic, Chandana Rajesh, Amber Tang, Shushan Toneyan, Ethan Labelson, Rohit Tripathy, and Peter K Koo. Selecting deep neural networks that yield consistent attribution-based interpretations for genomics. In *Machine Learning in Computational Biology*, pp. 131–149. PMLR, 2022.

Antonio Majdandzic, Chandana Rajesh, and Peter K. Koo. Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biol.*, 24:109, 2023.

Tim R. Mercer, Michael B. Clark, Stacey B. Andersen, Marion E. G. Brunck, Wilfried Haerty, Joanna Crawford, Ryan J. Taft, Lars K. Nielsen, Marcel E. Dinger, and John S. Mattick. Genome-wide discovery of human splicing branchpoints. *Genome Res.*, 25(2):290–303, 2015.

Rajiv Movva, Peyton Greenside, Georgi K Marinov, Surag Nair, Avanti Shrikumar, and Anshul Kundaje. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One*, 14(6):e0218073, 2019.

Jonathan M Mudge, Sílvia Carbonell-Sala, Mark Diekhans, Jose Gonzalez Martinez, Toby Hunt, Irwin Jungreis, Jane E Loveland, Carme Arnan, If Barnes, Ruth Bennett, Andrew Berry, Alexandra Bignell, Daniel Cerdán-Vélez, Kelly Cochran, Lucas T Cortés, Claire Davidson, Sarah Donaldson, Cagatay Dursun, Reham Fatima, Matthew Hardy, Prajna Hebbar, Zoe Hollis, Benjamin T James, Yunzhe Jiang, Rory Johnson, Gazaldeep Kaur, Mike Kay, Riley J Mangan, Miguel Maquedano, Laura Martínez Gómez, Nourhen Mathlouthi, Ryan Merritt, Pengyu Ni, Emilio Palumbo, Tamara Perteghella, Fernando Pozo, Shriya Raj, Cristina Sisu, Emily Steed, Dulika Sumathipala, Marie-Marthe Suner, Barbara Uszczynska-Ratajczak, Elizabeth Wass, Yucheng T Yang, Dingyao Zhang, Robert D Finn, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Anshul Kundaje, Benedict Paten, Michael L Tress, Ewan Birney, Fergal J Martin, and Adam Frankish. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.*, 53(D1): D966–D975, 2025.

Van Nostrand, Peter Freese, Gabriel A. Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, Jia-Yu Chen, Neal A. L. Cody, Daniel Dominguez, Sara Olson, Balaji Sundararaman, Lijun Zhan, Cassandra Bazile, Louis Philip Benoit Bouvrette, Julie Bergalet, Michael O. Duff, Keri E. Garcia, Chelsea Gelboin-Burkhart, Myles Hochman, Nicole J. Lambert, Hairi Li, Michael P. McGurk, Thai B. Nguyen, Tsultrim Palden, Ines Rabano, Shashank Sathe, Rebecca Stanton, Amanda Su, Ruth Wang, Brian A. Yee, Bing Zhou, Ashley L. Louie, Stefan Aigner, Xiang-Dong Fu, Eric Lécuyer, Christopher B. Burge, Brenton R. Graveley, and Gene W. Yeo. A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818):711–719, 2020.

Gherman Novakovsky, Nick Dexter, Maxwell W. Libbrecht, Wyeth W. Wasserman, and Sara Mostafavi. Obtaining genetic insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.*, 24:125–137, 2023.

Rupali P. Patwardhan, Choli Lee, Oren Litvin, David L. Young, Dana Pe'er, and Jay Shendure. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.*, 27:1173–1175, 2009.

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. *Nat. Commun.*, 16(1), 2025.

Eva I. Prakash, Avanti Shrikumar, and Anshul Kundaje. Towards more realistic simulated datasets for benchmarking deep learning models in regulatory genomics. In *Machine Learning in Computational Biology*, pp. 58–77. PMLR, 2022.

Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O. F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid D. Morris, and Timothy R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.

Paul J. Sample, Ban Wang, David W. Reid, Vlad Presnyak, Iain J. McFadyen, David R. Morris, and Georg Seelig. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.*, 37(7):803–809, 2019.

Jacob Schreiber. tangermeme: A toolkit for understanding cis-regulatory logic using deep learning models. *bioRxiv*, 2025. doi: 10.1101/2025.08.08.669296.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML'17: Proc. of the 2017 International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. *CoRR*, abs/1811.00416, 2018.

Liyuan Shu, Jiao Tang, Xiaoyu Guan, and Daoqiang Zhang. A comprehensive survey of genome language models in bioinformatics. *Brief. Bioinform.*, 27(1), 2026.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL `https://arxiv.org/abs/1706.03825`.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML'17: Proc. of the 2017 International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K Koo. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. *Genome Biol.*, 26:203, 2025.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6021–6029, 2020.

Shushan Toneyan and Peter K. Koo. Interpreting cis-regulatory interactions from large-scale deep neural networks. *Nat. Genet.*, 56(11):2517–2527, 2024.

Jill Townley, Wipapat Kladwang, David Baker, Hamish M Blair, Christian A Choe, Gina El Nesr, Andrew Favor, Eli Fisker, Daniel B Haack, Shujun He, Jason Hingey, Po-Ssu Huang, Rui Huang, Chaitanya K Joshi, Thomas Karagianes, Andrew Kubaney, Pietro Liò, Adamo Mancino, Jonathan Romano, Boris Rudolfs, Nicholas Spellmon, Navtej Toor, Jigyasa Verma, Vivian Wu, Zhiheng Yu, Eterna Participants, and Rhiju Das. De novo design of RNA pseudoknots with deep learning. *Science*, 393(6814):931–937, 2026.

He Wang, Yikun Zhang, Jie Chen, Jian Zhan, and Yaoqi Zhou. A comparative review of RNA language models. *arXiv preprint arXiv:2505.09087*, 2025.

Christopher Workman and Anders Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, 27(24):4816–4822, 1999.

Zhihao Xia, Yu Li, Bin Zhang, Zhongxiao Li, Yuhui Hu, Wei Chen, and Xin Gao. DeeReCT-PolyA: A robust and generic deep learning method for PAS identification. *Bioinformatics*, 35(14):2371–2379, 2019.

Heng Yang, Renzhi Chen, and Ke Li. Bridging sequence-structure alignment in RNA foundation models. In *AAAI'25: Proc. of the 2025 AAAI Conference on Artificial Intelligence*, pp. 21929–21937. AAAI Press, 2025.

Weijie Yin, Zhaoyu Zhang, Shuo Zhang, Liang He, Ruiyang Zhang, Rui Jiang, Gan Liu, Jingyi Wang, Xuegong Zhang, Tao Qin, and Zhen Xie. ERNIE-RNA: An RNA language model with structure-enhanced representations. *Nat. Commun.*, 16(1), 2025.

Haopeng Yu, Heng Yang, Wenqing Sun, Zongyun Yan, Xiaofei Yang, Huakun Zhang, Yiliang Ding, and Ke Li. An interpretable RNA foundation model for exploring functional RNA motifs in plants. *Nat. Mach. Intell.*, 6:1616–1625, 2024.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, 12(10):931–934, 2015.

Shasha Zhou, Mingyu Huang, and Ke Li. Position: Genomic model research must move beyond anecdotal evaluation of interpretability methods. In *ICML'26: Proc. of the 2026 International Conference on Machine Learning*. PMLR, 2026.

Shuxian Zou, Tianhua Tao, Sazan Mahbub, Caleb N. Ellington, Robin Algayres, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P. Xing. A large-scale foundation model for RNA function and structure prediction. *bioRxiv*, 2024. doi: 10.1101/2024.11.28.625345.

A Statistics

All intervals are 95% percentile bootstrap intervals (B=2000, fixed seed) over the independent unit of each task: sequence (BBP stratified by RNA-binding protein), site (SPL) or design target with designs nested (CRP). Aggregates over backbones are medians of per-backbone values, with six backbones; every reported number uses fine-tuning seed 0. Coefficient of variation (CoV) is the sample standard deviation divided by the mean of raw grounding across the six backbones, for one method on one task; the methods' CoV is the median over method–task pairs and ISM's the median over tasks.

Two-sided Wilcoxon signed-rank tests over the 24 task–FM pairs (Table 14), comparing each method's grounding with the same FM's ISM (its 36-design value where the method is scored on that subset), reject equality for all seven methods at Holm-adjusted $p \le 2.5 \times 10^{-5}$; the six conventional methods ground below ISM on 23–24 of the 24 pairs and AttnLRP on 20, and ISM exceeds the positional floor on all 24 (sign test $p = 1.2 \times 10^{-7}$).

B Input Representation

A sequence $s = (s_1, \ldots, s_L) \in \mathcal{A}^L$, with $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$, is one-hot encoded as $\phi(s)_{i,:} = e_{s_i}$, a vertex of the simplex,

$$\Delta(\mathcal{A}) := \{p \in \mathbb{R}^{\mathcal{A}}_{\ge 0} : \sum_a p_a = 1\}$$

so that $\phi(s) \in \mathcal{X} := \Delta(\mathcal{A})^L \subset \mathbb{R}^{L\times 4}$.

We regard $\mathcal{X}$ as a continuous relaxation of sequence space, in which real sequences occupy only the vertices and $\mathcal{X}$ is their convex hull. For a model whose input is a learned nucleotide embedding $E \in \mathbb{R}^{4\times d}$, the scalar read-out f (a target logit, or a per-position output at the evaluated site) factors as $f(X) = g(XE)$, and the chain rule carries embedding gradients back to the simplex, $\nabla_X f = (\nabla_Z g)\, E^\top$ with $Z = XE$.

C Attribution Method Formalisations

Below, $X_0 = \phi(s)$ is the input, X' a baseline, s_i the observed letter at position i, and $P = I_4 - \frac{1}{4}\mathbf{1}\mathbf{1}^\top$ the gauge projection applied to every map (Appendix G).

Saliency. The raw input gradient; the local linear sensitivity of the prediction to an infinitesimal change in each coordinate.

$$M = \nabla_X f(X_0)$$

Input $\times$ Gradient. The gradient masked by the input, so only the wild-type coordinate of each position is nonzero.

$$M = X_0 \odot \nabla_X f(X_0)$$

At a one-hot input this carries a single scalar per position: after gauge correction it assigns $\frac{3}{4}g_i$ to the wild-type letter and $-\frac{1}{4}g_i$ to each of the three alternatives, where $g_i = \nabla_X f(X_0)_{i,s_i}$.

Integrated Gradients. Accumulates gradients along the straight-line path from baseline to input, evaluated by the midpoint rule with $n = 32$ steps, escalated to $n = 64$ where the completeness error $|\sum M - (f(X_0) - f(X'))|$ exceeds 5% of $|f(X_0) - f(X')|$ at the median or 10% at the 95th percentile. On RiNALMo and AIDO.RNA (every task) and OmniGenome (PAS) the error does not converge by 512 steps. These seven method–FM–task results are kept in every reported number; excluding them would move the methods' median share of ISM's grounding (Section 3.1) from 60% to 62%.

$$M = (X_0 - X') \odot \frac{1}{n}\sum_{k=0}^{n-1} \nabla_X f\big(X' + \alpha_k(X_0 - X')\big), \qquad \alpha_k = \tfrac{k+1/2}{n}$$

where X' is the uniform baseline, $X'_{i,a} = \frac{1}{4}$.

SmoothGrad. Averages the gradient over Gaussian perturbations of the input, suppressing high-frequency gradient noise; we use $S = 25$ samples at $\sigma = 0.15$ with $\varepsilon^{(j)} \sim \mathcal{N}(0, I)$, applied only at nucleotide positions.

$$M = \frac{1}{S}\sum_{j=1}^{S} \nabla_X f\big(\Pi(X_0 + \sigma\varepsilon^{(j)})\big)$$

Perturbed samples are returned to the simplex by the rectify-and-renormalize retraction

$$\Pi(V)_{i,a} = \frac{\max(V_{i,a},\, 0)}{\max\big(\sum_b \max(V_{i,b},\, 0),\; 10^{-8}\big)},$$

where the denominator floor guards the degenerate case in which a row is everywhere non-positive. Π is a retraction rather than a projection and is biased ($\mathbb{E}[\Pi(X_0 + \sigma\varepsilon)] = (0.859, 0.047, 0.047, 0.047)$ at a vertex); an unbiased tangent-space kernel would leave the simplex, so the bias is retained.

GradientSHAP. Interpolates between noisy inputs and randomly drawn baselines, giving a sampled expectation over the SHAP baseline distribution; we use $S = 25$ samples, $\sigma = 0.1$, $\alpha_j \sim \mathcal{U}(0, 1)$, and $D^{(j)} = X_0 + \sigma\varepsilon^{(j)} - X'_{(j)}$. Each $X'_{(j)}$ is a distinct dinucleotide shuffle of the sequence itself, so the S paths start from S composition-matched references rather than from a single point; the noise $\varepsilon^{(j)}$ is applied only at nucleotide positions.

$$M = \frac{1}{S}\sum_{j=1}^{S} D^{(j)} \odot \nabla_X f\Big(X'_{(j)} + \alpha_j D^{(j)}\Big)$$

Unlike SmoothGrad, no retraction is applied, since the interpolation already places evaluation points off the vertex set.

DeepLIFT. Difference-from-reference multipliers m propagated by the rescale rule (tangermeme v1.4.0, `deep_lift_shap`, `hypothetical=True`) against the uniform reference X' that Integrated Gradients uses. We report hypothetical contributions (Shrikumar et al., 2017; 2018),

$$M_{i,a} = m_{i,a} - \sum_{b \in \mathcal{A}} X'_{i,b}\, m_{i,b},$$

whose subtracted term is constant in a and lies in the kernel of P, so the gauged map is mP. The reference therefore reaches the gauged map only through m, which equals the input gradient unless a rescale rule is applied.

C.1 AttnLRP

We use AttnLRP (Achtibat et al., 2024) in its gradient form: the one-hot input times the gradient under a modified backward pass,

$$M = X_0 \odot \tilde{\nabla}_X f(X_0),$$

on the same checkpoints and target scalar as every other method, and gauged in the same way (Appendix G). Each rule returns the stock operation's output, so the forward pass is unchanged to the bit; only the backward pass differs:

- **LayerNorm:** the normaliser is held constant, so the layer is linear in its input.
- **Elementwise activations** (GELU, SiLU, RiNALMo's swish): the identity rule, with gradient $\varphi(x)/x$ (its limit, $\frac{1}{2}$, at $x = 0$).
- **Products of two input-dependent tensors** ($QK^\top$, AV, the SwiGLU gate of AIDO.RNA and RiNALMo, and ERNIE-RNA's bilinear pairing bias): the uniform rule, halving the gradient so each operand receives half the relevance.
- **Softmax and linear operations:** unchanged. In this form AttnLRP's softmax rule is the softmax gradient, and a linear layer's gradient is its ε-rule as $\varepsilon \to 0$.

The rules are attached to module instances for the duration of the call, and a `TorchFunctionMode` intercepts `matmul`, `einsum` and `softmax`, since three backbones compute attention inline. With the parameters frozen during the call, a tensor requires a gradient exactly when it depends on X, which is how products are recognised as bilinear.

D DeepLIFT and Saliency

We computed DeepLIFT with tangermeme (v1.4.0) (Schreiber, 2025), using hypothetical contributions and a single uniform reference; the gauge correction cancels the reference term, leaving the multiplier, which equals the input gradient unless a rescale rule is applied. On five of the six models DeepLIFT and Saliency agree to a median per-unit $\max|D - S|/\max|S|$ below 3×10^{-5} on every positional task (Figure 6). On their fine-tuned m6A models, the multipliers of tangermeme and Captum (v0.9.0) (Schreiber, 2025; Kokhlikyan et al., 2020) match the autograd gradient to within 3×10^{-6}; a backward-hook census finds no module on the gradient path to which either library attaches a rule (activations are custom classes or function calls); and on BBP a dinucleotide reference moves the maps by a median of about 10^{-7}.

ERNIE-RNA is the exception, for tangermeme only: our differentiable form of its base-pairing bias puts the stock `torch.nn.GELU` of its pairwise-bias projection on the gradient path, which tangermeme registers and Captum does not. Its maps differ from Saliency by a median 15–21% on each task and move by 13% under a dinucleotide reference (BBP), whereas Captum's multipliers match the gradient exactly (fine-tuned m6A and BBP). On the pretrained trunk, disabling that rule or keeping the original integer lookup reduces the difference below 10^{-7}; for CRP the two maps are identical on five FMs and agree to within 6×10^{-6} on ERNIE-RNA. In small models trained from scratch, swapping custom for stock activation modules with identical weights switches the agreement on and off, and with rules on every activation and softmax the contributions sum to $f(X_0) - f(X')$ to within 2×10^{-5} for an MLP but not for any transformer we tested (Ali et al., 2022; Achtibat et al., 2024).

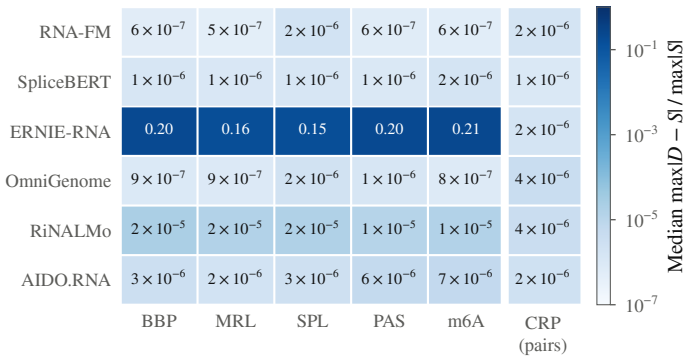


Figure 6: DeepLIFT departs from Saliency only for ERNIE-RNA on the positional tasks. Median per-unit $\max|D - S| / \max|S|$ between the two maps, real sequences; CRP is computed on the base-pair maps (median over designs).

E Alternative Attribution Method Baselines

The reference point is a free parameter of path-based attribution, and we make two different choices: the uniform-simplex point for IG, and 25 dinucleotide shuffles for GradientSHAP. We substitute each reference and re-score the result for faithfulness and grounding, on the same 400 sequences used throughout.

A single shuffled reference is inadequate. IG assigns exactly zero wherever the shuffle restores the observed base, at 30% of positions, and faithfulness and grounding both deteriorate (Table 2). Averaging IG over 25 shuffles eliminates these zeros (2%) and attains the highest comprehensiveness of the configurations tested, in 23 of 24 task–model combinations, at 25× the cost. With a single reference point, GradientSHAP reduces to a stochastic form of IG, correlating with our IG maps at $\rho = 0.92$, against $\rho = 0.35$ between two seeds of GradientSHAP itself, and improves no measure.

Table 2: Each alternative reference minus the reference we use for that method (the uniform point for IG, 25 shuffles for GradientSHAP): median over task–model combinations with a 95% bootstrap interval (2,000 resamples). Faithfulness on all 400 sequences of BBP, MRL, PAS and CRP (24 combinations); grounding (AUPRC) on BBP, MRL and PAS, over sequences carrying an annotation (18). The reseeded row gives GradientSHAP's own seed-to-seed variation.

Alternative	ΔComp. ↑	ΔSuff. ↓	ΔG ↑
IG, 1 shuffle	-0.116 $[-0.197, -0.050]$	$+0.106$ $[+0.010, +0.254]$	-0.096 $[-0.209, -0.017]$
IG, 25 shuffles (mean)	$+0.248$ $[+0.210, +0.385]$	-0.162 $[-0.275, -0.117]$	$+0.010$ $[-0.003, +0.057]$
GradientSHAP, uniform	-0.023 $[-0.126, +0.011]$	$+0.041$ $[-0.033, +0.105]$	$+0.032$ $[-0.004, +0.105]$
GradientSHAP, reseeded	$+0.002$ $[-0.010, +0.004]$	$+0.004$ $[-0.015, +0.007]$	$+0.004$ $[-0.004, +0.006]$

F In-silico mutagenesis

ISM is the computational form of saturation mutagenesis (Patwardhan et al., 2009; Alipanahi et al., 2015). Writing $s^{i \to a}$ for s with position i set to letter a,

$$\Delta_{i,a} = f\big(\phi(s^{i \to a})\big) - f\big(\phi(s)\big), \qquad \Delta_{i,s_i} = 0,$$

which costs $3L + 1$ forward passes and is scored on the same target scalar the gradient methods differentiate. Because the observed base scores exactly zero, the gauge correction P shifts each row by $c_i = -\frac{1}{4}\sum_a \Delta_{i,a} = -\frac{3}{4}\bar{\Delta}_i$, with $\bar{\Delta}_i$ the mean over the three substitutions; the observed base therefore scores $-\frac{3}{4}\bar{\Delta}_i$, and a base whose mutation lowers the output scores positive, the sign convention used throughout. The per-position value used for grounding and faithfulness is the gauged map read at the observed base, $\tilde{M}_{i,s_i}$. For Integrated Gradients and GradientSHAP, whose maps multiply the

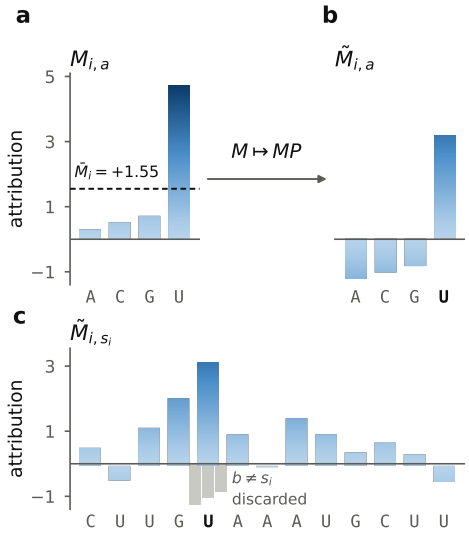


Figure 7: Demonstration of per-base attribution being corrected and extracted into per-position values for a sequence. **a** shows the raw per-base attributions for one position in a sequence. **b** shows the same attributions after Majdandzic's gauge fix. **c** shows the final per-position attributions for the entire sequence, with the per-base attributions for one position shown as discarded in grey.

gradient by $X_0 - X'$, gauging does not remove the component along $\mathbf{1}$, so their per-position value is instead $\sum_a M_{i,a}$, except for SPL, CRP faithfulness and IG's PAS grounding in Table 1 and for the appendix comparisons.

G Gauge correction

Every substitution at position i is a displacement $e_a - e_{s_i}$ in the zero sum subspace $\{v \in \mathbb{R}^4 : \mathbf{1}^\top v = 0\}$, so the component of $M_{i,:}$ along $\mathbf{1}$ cannot affect any within-simplex perturbation, and all contrasts $M_{i,a} - M_{i,b}$ are invariant under $M_{i,:} \mapsto M_{i,:} + c_i\mathbf{1}$. Off the simplex the model is unconstrained by training data, which makes that component unreliable (Majdandzic et al., 2023). We remove it by the orthogonal projection

$$\tilde{M} = MP, \qquad P = I_4 - \tfrac{1}{4}\mathbf{1}\mathbf{1}^\top,$$

applied to every map, ISM included, before any comparison (Figure 7). For Saliency and SmoothGrad, $M_{i,a}$ is a directional derivative in the ambient coordinate (i, a), not a substitution effect; the first-order estimate of $\mathrm{ISM}_{i,a}$ is the contrast $\hat{\Delta}_{i,a} = M_{i,a} - M_{i,s_i}$, which is invariant under P and is the gauge-free content of any map. The per-position value used for grounding and faithfulness is the gauged map read at the observed base, $\tilde{M}_{i,s_i}$.

H Dinucleotide Shuffle

Originally developed to test whether evolutionary similarity between two sequences can be described by dinucleotide composition alone, the algorithm keeps the local composition that drives folding energies and low-order sequence models, and destroys arrangement, so it is the standard null for asking whether an algorithm responds to composition or to sequence (Altschul & Erickson, 1985; Gesell & Washietl, 2008; Workman & Krogh, 1999). We use the implementation in tangermeme (Schreiber, 2025).

Let $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$ and $x = x_1 \cdots x_L$, with dinucleotide counts $n_{ab}(x) = |\{i < L : x_i = a,\ x_{i+1} = b\}|$. A shuffle is a draw from

$$\mathcal{S}(x) = \{\, y \in \Sigma^L : n_{ab}(y) = n_{ab}(x)\ \forall a, b,\ y_1 = x_1,\ y_L = x_L \,\},$$

Table 3: Faithfulness on real and on dinucleotide-shuffled sequence, paired: each real sequence is compared with its own shuffled null. Medians over the 30 task × model groups (BBP, MRL, PAS, CRP and m6A; SPL, with 10 pairs, is left out) and the Pearson correlation of the 30 pairs. m6A's nulls shuffle the flanks and keep the central motif.

	Comprehensiveness ↑			Sufficiency ↓		
Method	real	shuffled	r	real	shuffled	r
ISM	1.466	1.934	0.72	0.048	−0.145	0.67
Integrated Gradients	1.133	1.148	0.92	0.303	0.201	0.93
GradientSHAP	1.196	1.229	0.93	0.278	0.132	0.91
SmoothGrad	0.895	0.863	0.94	0.598	0.539	0.96
Input × Gradient	0.680	0.680	0.95	0.721	0.693	0.97
DeepLIFT	0.562	0.545	0.95	0.800	0.741	0.97
Saliency	0.562	0.539	0.94	0.800	0.750	0.96
AttnLRP	1.022	0.972	0.94	0.464	0.325	0.93
Random ranking	0.388	0.364	0.42	0.871	0.822	0.41

Table 4: Outside the labelled signal, the attribution methods overlap little more than chance. Overlap: Jaccard overlap of two backbones' top-10% unlabelled positions, averaged over sequences and backbone pairs; random sets overlap by 0.05, maps of different sequences by 0.06–0.10. Unlabelled consensus: positions per sequence that at least four of six backbones rank in their top 10% and that are not labelled (CRP: pairs per design, of about 3,800 candidates), of which ≤3 nt away: the share of these within 3 nt of a labelled position (All: every unlabelled position). Methods: median over the seven attribution methods; CRP uses the 36-design subset (Appendix J). ISM does not clear the positional floor on SPL and m6A.

	Overlap		Unlabelled consensus		Of which ≤3 nt away		
Task	Methods	ISM	Methods	ISM	Methods	ISM	All
BBP	0.10	0.18	0.52	1.06	22%	15%	8%
MRL	0.10	0.63	0.27	1.80	73%	35%	18%
SPL	0.13	0.37	2.24	9.08	12%	5%	2%
CRP	0.12	0.24	65	162	24%	52%	13%
PAS	0.07	0.15	0.81	4.35	7%	8%	7%
m6A	0.08	0.13	0.04	0.13	47%	25%	26%

so length, base and dinucleotide composition and both sequence ends are conserved while the interior ordering is disrupted.

I Positional Floor

A scorer that has learned nothing but position can score well against an annotation whose elements sit at fixed offsets (branch points 18–40 nt upstream of the acceptor, the PAS elements at fixed distances from the hexamer), and the dinucleotide-shuffled null does not test this, since a shuffled sequence carries no annotation. The floor is a model-free scorer that never reads the sequence: over the U scored held-out units of a task, with $B_u \in \{0,1\}^L$ the annotation and $\mathcal{M}_u$ the positions the task scores (the whole window, less the ± 3 self-mask on SPL),

$$\mathrm{prior}(i) = \frac{\sum_{u \in U} [\, i \in \mathcal{M}_u \,]\, B_u(i)}{\sum_{u \in U} [\, i \in \mathcal{M}_u \,]},$$

assigned to every unit and scored with the identical metric, polarity and mask. Because it is estimated on the units it is compared against it is slightly optimistic as a floor, which is the conservative direction for a claim of excess. On CRP the negatives are matched on distance and partner letter, so the floor is chance, 0.5. Only the part of a score above the floor is signal.

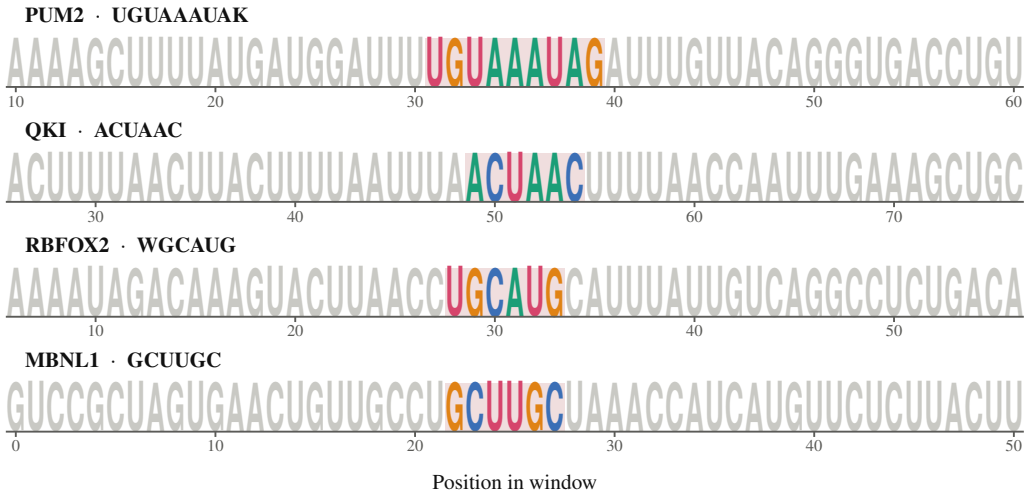


Figure 8: **BBP.** One bound window per RNA-binding protein: 51 of its 101 nt, centred on the delivered CisBP-RNA motif hit (shaded). The consensus each PWM encodes is given beside the protein.

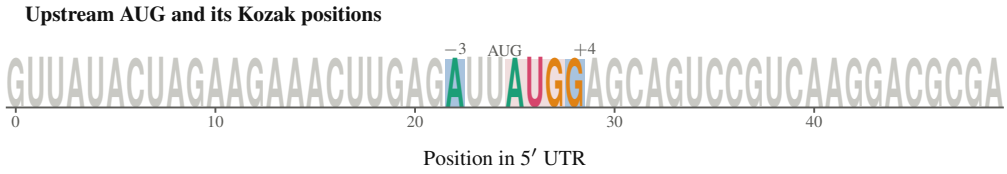


Figure 9: **MRL.** One 5′ UTR carrying a single upstream AUG, with the codon and the two Kozak determinants that set how strongly it competes: a purine at -3 and a G at $+4$.

J RNA TASKS

Sources, construction and splits are given in Appendix K. Attribution and ISM target one scalar per task: the pre-sigmoid logit for BBP (of the row's protein), PAS and m6A, the pre-softmax logit of the true class at each site for SPL, the predicted load for MRL, and the mean predicted 2A3 reactivity over the design region for CRP.

J.1 BINDING-SITE PREDICTION (BBP)

Binary classification of whether a named RNA-binding protein (PUM2, QKI, RBFOX2 or MBNL1) binds a 101-nt window: positives are centred on ENCODE K562 eCLIP peaks, and negatives are the same windows shifted 10 kb along the same chromosome and strand, clear of that protein's peaks. The annotation marks hits of the protein's own CisBP-RNA motif, measured *in vitro* and so independent of the eCLIP labels, which occurs in 55% of bound and 7% of unbound windows (Figure 8).

J.2 MEAN RIBOSOME LOAD (MRL)

Regression of the mean ribosome load of a randomised 50-nt 5′ UTR from a polysome-profiling MPRA (Sample et al., 2019); the reporter's coding sequence starts immediately after the window, so every AUG in it is an upstream AUG (uAUG) competing for initiation. The annotation marks the uAUG codons and, separately, the Kozak -3 and $+4$ positions of uAUGs with a purine and a G there; UTRs carrying such a uAUG average a load of 4.17, against 7.70 for UTRs without a uAUG (Figure 9).

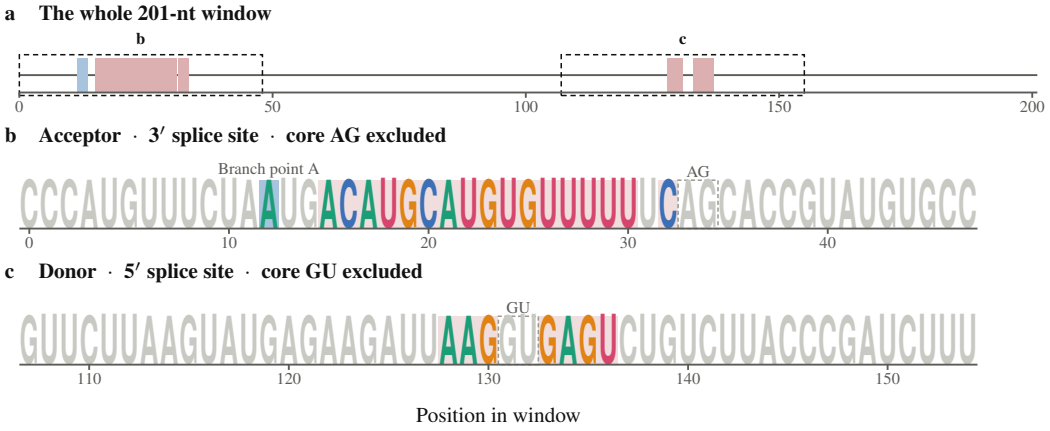


Figure 10: **SPL.** One window carrying both splice sites. Shaded: the scored mask, the acceptor's polypyrimidine tract and the donor's flanking consensus. Dashed: the core GU and AG, excluded from scoring as trivially learnable. The branch point is a second, separate annotation.

J.3 Splice-site prediction (SPL)

Per-position classification of 201-nt windows into donor, acceptor or neither, with sites taken from GENCODE v50 protein-coding exon boundaries and each labelled site scored separately; the annotation marks the splice consensus flanking each donor and the polypyrimidine tract upstream of each acceptor, excluding the GU and AG cores as trivially learnable, and lariat-sequenced branch points (Mercer et al., 2015) form a separate annotation (Figure 10). Because these elements lie at fixed offsets from the queried site, a scorer knowing only the site's position and class reaches AUPRC 1.000 against 0.420 for ISM, so SPL does not contribute to aggregate comparisons (Appendix I).

J.4 Polyadenylation signal (PAS)

Binary classification of a 206-nt window as a real cleavage site or a decoy; every window carries a candidate poly(A) hexamer (AAUAAA or a variant) at positions 100–105, which carries almost no label information, so discrimination must come from the flanks. The annotation marks the upstream UGUA element in $[0, 100)$ and the downstream U/GU-rich element, $\texttt{U}\{5,\}\cup\texttt{UGUGU}\cup(\texttt{GU})\{3,\}$, in $[106, 206)$, with the hexamer kept only as a control (Figure 11).

J.5 N^6-methyladenosine (m6A)

Binary classification of whether the central adenosine of a 41-nt window, set in a DRACH motif at positions 18–22, is methylated; the annotation marks the DRACH positions as a motif-presence control, since both classes carry the motif (Figure 12). In the dataset as obtained, however, a U at the motif's first position occurs only in decoys, in every split, and all six models learned it: substituting that U for A or G raises the methylation logit by 7.1–10.2, against at most 1.2 elsewhere in the motif.

J.6 Chemical reactivity (CRP)

Per-position regression of 2A3 and DMS chemical reactivity, low where a base is paired and high where it is unpaired, trained on Stanford Ribonanza (He et al., 2024) and attributed on OpenKnot OK7a (Townley et al., 2026): 160 designs for 20 pseudoknotted targets, each probed by mutate-and-map with a single mutant per position of its design region. Each map relates positions as an $L \times L$ response, the change in predicted reactivity at j when i is mutated, scored by AUROC against the designed base pairs, pseudoknots included, with non-pairs matched on distance and partner letter so that a complementarity-only scorer sits at exactly 0.500; excluding six designs that overlap Ribonanza training rows leaves 131 of the 137 scorable designs, on which the assay itself reaches a median of 0.770 (Figure 13). Integrated Gradients, GradientSHAP and SmoothGrad, whose pair maps are 25–32 times the cost of Saliency's, are scored on a fixed subset of 36 designs, two per

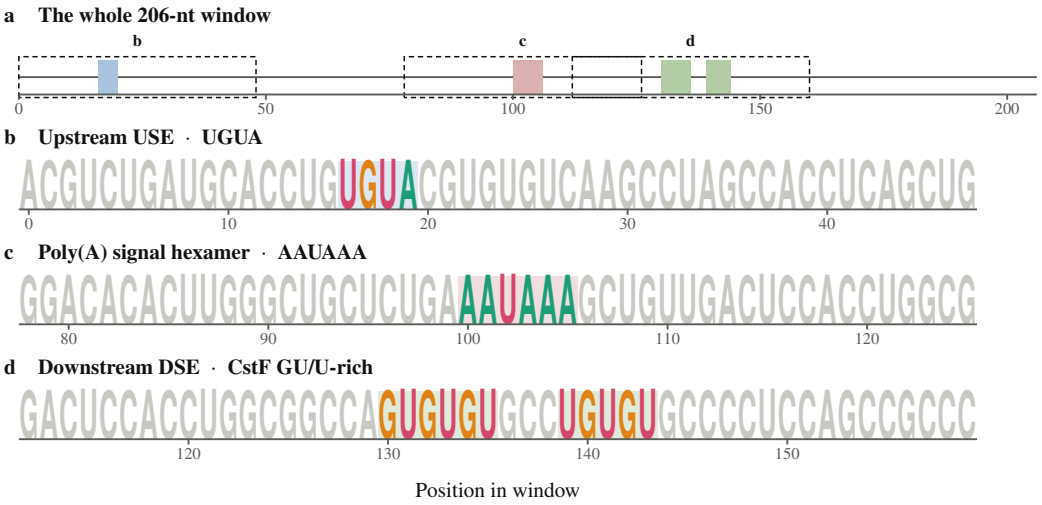


Figure 11: **PAS.** One window and its three labelled elements: the upstream UGUA, the poly(A) hexamer, and the downstream GU/U-rich CstF element. Every window carries a hexamer by construction, so discrimination must come from the flanking elements.

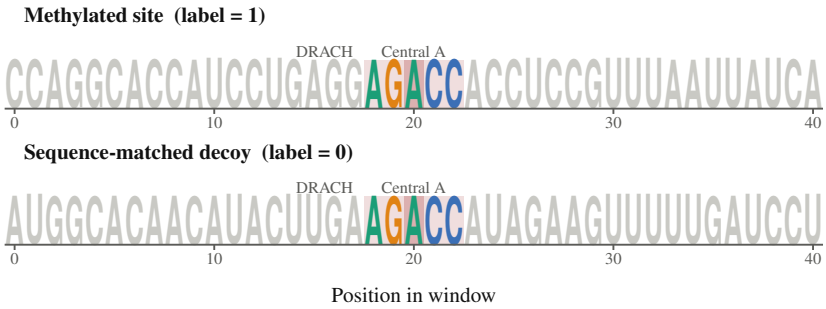


Figure 12: **m6A.** Two held-out windows, a methylated site and its sequence-matched decoy. Both carry the central $DRACH$ consensus and the adenosine it is read at, so the motif was meant to mark the two classes alike, as a negative control; in the data, a U at its first position occurs only in decoys.

target; ISM scores 0.750 on this subset against 0.755 on all 131, and these three methods are read against its subset value.

K RNA Data

Data. Each task is built from public raw data into a processed dataset with a fixed train/validation/test split, summarised in Table 6. The datasets, splits and grounding annotations are released with the notebooks that build them; rebuilding from the raw files reproduces every dataset and split. All raw data are public releases, listed in Table 5.

Construction. Each dataset holds one row per labelled unit (Table 6), in the RNA alphabet and in transcript orientation; the genomic windows of BBP and SPL are cut from hg38 and reverse-complemented on the minus strand. Exact duplicates were collapsed to one copy, and sequences with contradictory labels (m6A 20, PAS 2,144, SPL 11) or, for PAS, occurring in more than one of its three collections (3,425) were removed, so every sequence is unique within its dataset.

Splits. Each task has its own split rule; row counts are in Table 10.

- **BBP**: Whole chromosomes, 80/10/10, so that overlapping windows and each positive's shifted negative share a split.

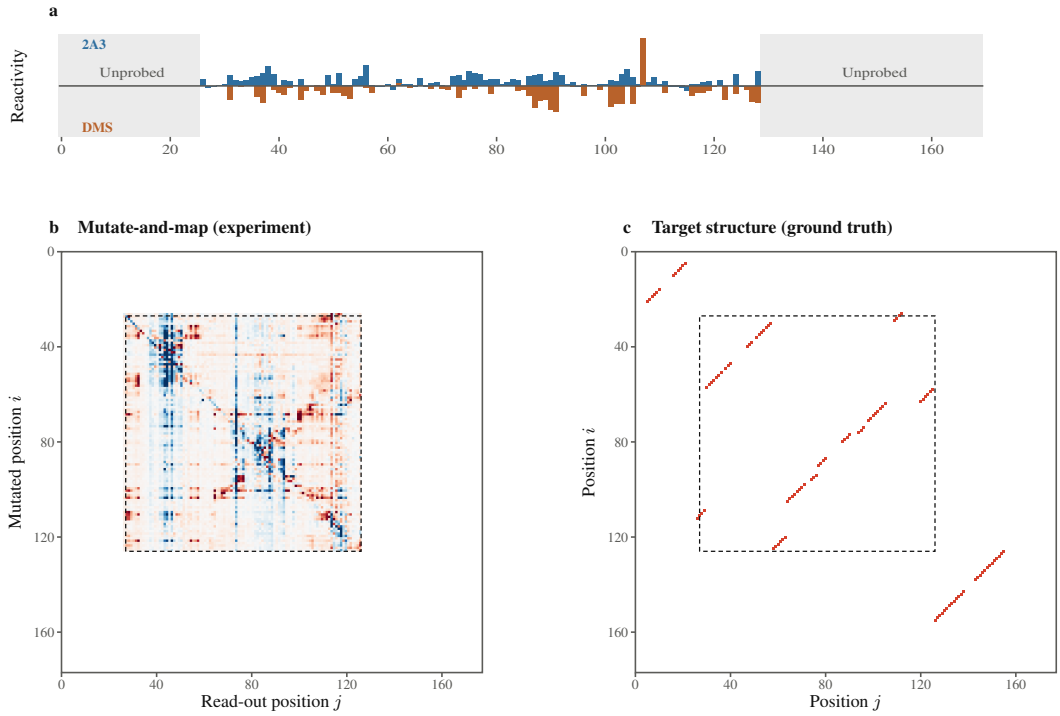


Figure 13: CRP. **(a)** What the head is trained on: per-nucleotide chemical reactivity of one held-out design, 2A3 above the line and DMS below. The flanks are never probed. **(b)** The mutate-and-map experiment: entry (i, j) is the change in reactivity at j when i is mutated, so a helix appears off the diagonal. **(c)** The design's target secondary structure. Nested and pseudoknotted pairs are drawn alike, both being scored as true pairs. Only the design region, dashed, is mutated.

- **MRL**: Rank holdout on read depth; the 20,000 deepest-read UTRs form the test set and the next 10,000 the validation set, where ribosome load is measured most reliably.
- **SPL**: Whole chromosomes (validation chr18 and chr21, test chr20), since alternative isoforms of a gene share splice sites.
- **PAS**: Cross-species; both mouse strains are held out for testing (94.6% of test), and the human sequences are split 70/20/10 by groups sharing 31-mers (Jaccard ≥ 0.5), stratified by hexamer.
- **m6A**: Seeded random 80/10/10, as near-duplicate groups are small (at most eight sequences).
- **CRP**: Ribonanza sequences split 90/10 into training and validation by groups sharing 11-mers (containment ≥ 0.7); OpenKnot OK7a (2A3 only) is the held-out set. The OpenKnot pilot libraries (PK50 and PK90, 16,492 sequences) and 11 sequences containing an OK7a design region verbatim were removed before fine-tuning; near-copies of six OK7a designs remain in training, and those designs are excluded from base-pair scoring.

L Attention

We read the attention weights of every layer from the same forward pass used for attribution and averaged them over heads. Four standard estimators combine the layers: their mean (Jain & Wallace, 2019), the last layer, attention rollout (Abnar & Zuidema, 2020), and attention weighted by its gradient with respect to the output (Chefer et al., 2021). Each nucleotide is scored by the attention it receives, averaged over the positions the output reads; attention has no sign, so positions are ranked by attention alone. On CRP, the nucleotide-by-nucleotide attention map is scored for base pairs in place of ISM's. Grounding and faithfulness are otherwise computed as for the attribution

Table 5: Public Data sources.

Task	Source	Organism, cell line and version
BBP	ENCODE eCLIP IDR peaks (Nostrand et al., 2020); motifs from CisBP-RNA (Ray et al., 2013)	Human K562 cells, hg38: ENCFF880MWQ, ENCFF206RIM, ENCFF786UOW, ENCFF603WDI; CisBP-RNA 2.00
MRL	Polysome-profiling MPRA (Sample et al., 2019)	Human HEK293T cells; eGFP reporter library, GEO GSM3130435
SPL	GENCODE annotation (Mudge et al., 2025) on UCSC hg38 (Kent et al., 2002); branch points (Mercer et al., 2015)	Human; GENCODE v50 (GRCh38.p14) protein-coding transcripts; branch points lifted from hg19
PAS	DeeReCT-PolyA collections (Xia et al., 2019), human set from Omni-PolyA (Magana-Mora et al., 2017)	Human; mouse C57BL/6J and SPRET/EiJ fibroblasts
m6A	Kaggle OmniRNA Challenge (htm042 et al., 2026)	Competition train and test files, test labels from the post-close solution
CRP	Stanford Ribonanza (He et al., 2024); Eterna OpenKnot (Townley et al., 2026)	Synthetic and natural RNAs, 2A3 and DMS chemical mapping in vitro; OpenKnot OK7a v4.5.2 with its target structures

Table 6: Processed datasets. Rows and size (millions of nucleotides) count all splits.

Task	L (nt)	Each row	Rows	Size	Held-out set
BBP	101	Window centred on an eCLIP peak, or shifted 10 kb from it as background	21,278	2.1	chr6, chr18, chr22, one unplaced contig
MRL	50	Randomised 5′ UTR and its mean ribosome load	326,033	16.3	20,000 deepest-read UTRs
SPL	201	Window labelled per base: a donor or acceptor at index 60–140, or no site at all	772,347	155.2	chr20
PAS	206	Window around a real or decoy poly(A) hexamer at 100–105	103,746	21.4	Both mouse strains and 10% of human
m6A	41	Window centred on a DRACH adenosine, methylated or not	50,362	2.1	Random 10%
CRP	170/177/206	RNA with per-base 2A3 and DMS reactivity	165,441	29.3	OK7a: 160 designs, 14,136 rows, 2A3 only

methods (Table 7). On CRP we also score the pretrained models before fine-tuning: their last-layer attention recovers pairing on ERNIE-RNA, RiNALMo, OmniGenome and AIDO.RNA (0.84, 0.75, 0.70, 0.68) but not on RNA-FM or SpliceBERT (0.50, 0.51).

M Models and training

Models. Six encoder RNA foundation models (Table 8), all with single-nucleotide tokens. Two changes make the one-hot input differentiable without altering any forward pass: ERNIE-RNA's attention bias, gathered by integer token id, is replaced by the equivalent bilinear form of the one-hot input, and for RiNALMo and OmniGenome the one-hot embedding is substituted before the ESM-style token-dropout rescale.

Fine-tuning. One recipe for all 36 task $\times$ model combinations: the trunk is unfrozen and trained end to end with a `Linear(hidden,` n_{out}`)` head on the mean-pooled nucleotide states (per-position for SPL and CRP), head weights $\sim \mathcal{N}(0, 0.02^2)$ from a fixed seed with zero bias. AdamW (weight decay 0.01), global-norm clipping at 1.0, learning rate 2×10^{-5} for the trunk and 10^{-3} for the head, 1,000 linear warm-up steps then constant, batch $\min(32,\ 8192/L)$ sequences, fp32, seed 0, one GPU per run. Validation every 500–1,200 steps (per task) on a fixed 1,500-row sample; a run stops when the selection metric has not improved for 20 consecutive evaluations or at the task's step cap, and the best-validation state, re-scored on the whole split, is the checkpoint (Table 10). 33 runs stopped on patience; SpliceBERT on CRP reached the cap; RiNALMo's MRL and SPL runs

Table 7: Grounding (G) and faithfulness (F, comprehensiveness) of attention. Medians over six models on the 400-sequence draw; CRP G is base-pair AUROC, with Integrated Gradients and GradientSHAP on the 36-design subset (Appendix J). Three attribution methods and four attention estimators. Most faithful per task in **bold**, most grounded underlined, excluding ISM.

	BBP		MRL		CRP	
Method	$G\uparrow$	$F\uparrow$	$G\uparrow$	$F\uparrow$	$G\uparrow$	$F\uparrow$
ISM (reference)	0.750	1.435	0.757	1.481	0.755	1.395
Floor / random ranking	0.275	0.383	0.210	0.470	0.500	0.445
Integrated Gradients	0.590	1.181	0.592	**1.178**	0.605	0.998
GradientSHAP	0.504	**1.255**	0.449	1.170	0.599	**1.111**
AttnLRP	0.617	1.136	0.242	0.665	0.635	0.600
Attention (raw)	0.347	0.675	0.599	0.911	0.654	0.706
Attention (last layer)	0.229	0.545	0.502	0.884	0.786	0.611
Attention rollout	0.162	0.471	0.344	0.775	0.501	0.382
Attention $\times$ gradient	0.595	0.779	0.594	0.858	0.608	0.627

Table 8: The six RNA foundation models: public checkpoint, parameters, depth and hidden width.

Model	Checkpoint	Params (M)	Layers	Hidden
SpliceBERT	`multimolecule/splicebert`	19.7	6	512
RNA-FM	`multimolecule/rnafm`	99.5	12	640
ERNIE-RNA	`multimolecule/ernierna`	85.7	12	768
OmniGenome	`yangheng/OmniGenome-186M`	185.9	32	720
RiNALMo	`multimolecule/rinalmo-giga`	650.9	33	1280
AIDO.RNA	`genbio-ai/AIDO.RNA-650M`	646.7	33	1280

diverged and its m6A run collapsed after their best checkpoint, which those runs use. The same recipe with a second seed refits all six backbones on BBP, MRL, PAS and CRP; held-out scores move by at most 0.04 between seeds (median 0.002). Held-out scores are in Table 11: macro AUC over the four RBP heads (BBP), Pearson r (MRL), macro F1 over donor, acceptor and neither (SPL), MCC (PAS, m6A) and MAE on 2A3 reactivity (CRP, reported as $1-\text{MAE}$).

Models trained from scratch. Four models were trained from scratch (Table 13), with a 33-layer deep CNN to match the depth of RiNALMo and AIDO.RNA, with a receptive field longer than every task. We also train a small transformer, labelled *Tx-small*. All are 128 wide, tokenise $\{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{U}\}$ with no special tokens and end in the same head as the models. CNN-4 stacks Conv–BatchNorm–ReLU layers; ResNet-9 and CNN-33 use pre-activation residual blocks (BatchNorm–ReLU–Conv, twice); Tx-small is four pre-norm blocks of 4-head self-attention with a learned relative-position bias and a GELU feed-forward. Training follows the recipe above with a single learning rate of 10^{-3} for the whole network, and CRP trains on its 177-nt rows only (121,765 of 136,413), the length of every attributed design (Table 9).

Table 9: Models trained from scratch beside the six RNA FMs: held-out score (BBP and MRL on the test split; CRP on the OK7a designs, 2A3 only, the same rows for every model), positional grounding (median AUPRC over the six methods), faithfulness (median comprehensiveness and sufficiency over tasks and methods) and base-pair recovery of the model's own substitution map (median AUROC over 131 designs, non-pairs matched on distance and partner letter; the assay scores 0.77). Integrated Gradients and GradientSHAP enter the grounding and faithfulness medians with the earlier position score for all ten models.

	Held-out Score			Grounding		Faithfulness		
Model	BBP AUC	MRL r	CRP MAE	BBP	MRL	comp.	suff.	CRP AUROC
CNN-4	0.874	0.922	0.312	0.61	0.84	1.33	0.11	0.52
ResNet-9	0.883	0.972	0.307	0.66	0.65	1.36	0.11	0.52
CNN-33	0.879	0.974	0.304	0.61	0.69	1.28	0.19	0.54
Tx-small	0.881	0.965	0.316	0.35	0.45	1.05	0.50	0.58
RNA-FM	0.892	0.969	0.278	0.55	0.26	1.03	0.44	0.72
SpliceBERT	0.893	0.974	0.292	0.37	0.46	0.70	0.77	0.72
ERNIE-RNA	0.892	0.974	0.249	0.42	0.39	0.88	0.64	0.76
OmniGenome	0.888	0.976	0.218	0.51	0.54	1.09	0.46	0.78
RiNALMo	0.905	0.966	0.259	0.15	0.18	0.51	0.85	0.75
AIDO.RNA	0.883	0.974	0.257	0.37	0.26	0.84	0.69	0.77

Table 10: Per-task fine-tuning settings. Rows are sequences (SPL: windows). CRP's validation split pools the 2A3 and DMS channels; its held-out set (OpenKnot OK7a) is 2A3 only, so the two MAEs are not the same quantity.

Task	L (nt)	Head	Selection metric	Train / val / test rows	Step cap
BBP	101	Linear(h, 4)	macro AUC	17,022 / 2,128 / 2,128	40,000
MRL	50	Linear(h, 1)	Pearson r	296,033 / 10,000 / 20,000	150,000
SPL	201	Linear(h, 3)	macro F1	733,583 / 19,565 / 19,199	150,000
PAS	206	Linear(h, 1)	MCC	26,255 / 7,499 / 69,992	70,000
m6A	41	Linear(h, 1)	MCC	40,289 / 5,036 / 5,037	80,000
CRP	170/177/206	Linear(h, 2)	$-$MAE	136,413 / 14,892 / 14,136	88,000

Table 11: Held-out task performance of the six fine-tuned backbones. Per task metrics are shown in Table 10.

Backbone	BBP	MRL	SPL	PAS	m6A	CRP
RNA-FM	0.892	0.969	0.859	0.415	0.382	0.722
SpliceBERT	0.893	0.974	0.869	0.340	0.371	0.708
ERNIE-RNA	0.892	0.974	0.866	0.431	0.378	0.751
OmniGenome	0.888	0.976	0.868	0.324	0.329	0.782
RiNALMo	0.905	0.966	0.860	0.286	0.370	0.741
AIDO.RNA	0.883	0.974	0.860	0.311	0.395	0.743

Table 12: Grounding varies far more between models than between fine-tuning seeds, on the four smaller models. Between seeds: the change in a model's grounding when it is refitted with the second seed (median absolute change over the models). Between models: the standard deviation of the seed-averaged grounding across the models (three for DeepLIFT on PAS). The between-model spread is the larger in 19 of 21 method–task pairs; every reported result elsewhere uses seed 0. CRP has second-seed checkpoints but no second-seed attribution maps, so it is not shown.

	Between seeds			**Between models**		
Method	BBP	MRL	PAS	BBP	MRL	PAS
ISM (reference)	0.008	0.004	0.060	0.050	0.003	0.060
Integrated Gradients	0.020	0.038	0.042	0.071	0.087	0.041
GradientSHAP	0.015	0.036	0.024	0.036	0.085	0.035
SmoothGrad	0.023	0.033	0.039	0.065	0.141	0.022
Input $\times$ Gradient	0.022	0.021	0.043	0.133	0.119	0.058
DeepLIFT	0.018	0.019	0.023	0.130	0.126	0.062
Saliency	0.024	0.025	0.045	0.132	0.127	0.046
AttnLRP	0.014	0.031	0.034	0.075	0.317	0.084

Table 13: Four models trained from scratch. ResNet-9 and CNN-33 use dilated residual convolutions while Tx-small has standard self-attention with a learned relative-position bias.

Model	Layers	Kernel / heads	Receptive field	Params (M)
CNN-4	4	$k = 9$	33 nt	0.59
ResNet-9	9	$k = 5$	125 nt	0.74
CNN-33	33	$k = 5$	485 nt	2.72
Tx-small	4	4 heads, FF 256	whole sequence	0.53

Table 14: Grounding of every method on every fine-tuned backbone, on the 400-sequence draw; on CRP, Integrated Gradients, GradientSHAP and SmoothGrad are scored on the 36-design subset (Appendix J). Each task's header gives its metric and positional floor.

Method	RNA-FM	SpliceBERT	ERNIE-RNA	OmniGenome	RiNALMo	AIDO.RNA	Median
BBP — AUPRC; positional floor 0.275							
ISM (reference)	0.682	0.679	0.783	0.758	0.742	0.786	0.750
Integrated Gradients	0.557	0.622	0.623	0.728	0.248	0.441	0.590
GradientSHAP	0.554	0.514	0.494	0.585	0.174	0.398	0.504
SmoothGrad	0.589	0.494	0.459	0.568	0.261	0.633	0.531
Input × Gradient	0.455	0.205	0.314	0.524	0.135	0.434	0.374
DeepLIFT	0.519	0.200	0.327	0.432	0.146	0.348	0.337
Saliency	0.519	0.200	0.306	0.432	0.146	0.348	0.327
AttnLRP	0.618	0.560	0.736	0.587	0.615	0.740	0.617
MRL — AUPRC; positional floor 0.210							
ISM (reference)	0.745	0.756	0.756	0.759	0.787	0.759	0.757
Integrated Gradients	0.663	0.759	0.522	0.689	0.260	0.298	0.592
GradientSHAP	0.461	0.616	0.436	0.540	0.206	0.266	0.449
SmoothGrad	0.277	0.532	0.421	0.578	0.245	0.262	0.349
Input × Gradient	0.218	0.407	0.355	0.483	0.182	0.241	0.298
DeepLIFT	0.215	0.416	0.339	0.554	0.183	0.186	0.277
Saliency	0.215	0.416	0.338	0.554	0.183	0.186	0.276
AttnLRP	0.095	0.342	0.708	0.142	0.090	0.398	0.242
PAS — AUPRC; positional floor 0.138							
ISM (reference)	0.226	0.224	0.295	0.192	0.176	0.232	0.225
Integrated Gradients	0.087	0.148	0.061	0.083	0.063	0.081	0.082
GradientSHAP	0.146	0.172	0.087	0.086	0.054	0.089	0.088
SmoothGrad	0.179	0.132	0.134	0.106	0.083	0.072	0.119
Input × Gradient	0.114	0.059	0.121	0.117	0.075	0.075	0.095
DeepLIFT	0.137	0.059	0.150	0.074	0.064	0.042	0.069
Saliency	0.137	0.059	0.135	0.074	0.064	0.042	0.069
AttnLRP	0.233	0.123	0.315	0.108	0.139	0.161	0.150
CRP — base-pair AUROC; positional floor 0.500							
ISM (reference)	0.719	0.720	0.760	0.777	0.750	0.770	0.755
Integrated Gradients	0.607	0.602	0.639	0.637	0.527	0.596	0.605
GradientSHAP	0.646	0.580	0.597	0.611	0.508	0.601	0.599
SmoothGrad	0.613	0.535	0.447	0.563	0.509	0.649	0.549
Input × Gradient	0.517	0.514	0.424	0.534	0.505	0.580	0.515
DeepLIFT	0.495	0.522	0.427	0.524	0.500	0.524	0.511
Saliency	0.495	0.522	0.427	0.524	0.500	0.524	0.511
AttnLRP	0.816	0.542	0.613	0.657	0.587	0.787	0.635