LOOKING IS NOT DEPENDING: TESTING POSITIONAL ATTENTION WITH AN ADDRESSABLE k-MER MEMORY IN DNA MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained DNA models are widely used to predict regulatory activity, chromatin state, and splicing. As accuracy improves, it remains unclear whether predictions rely on task-relevant biological signals or on incidental correlations in the data. One common form of evidence, attention concentrated at functional elements, shows where the model looks but not what the prediction depends on. Testing this distinction requires connecting attention at annotated positions with interventions on the corresponding internal representations. To enable this, we present LIND, a frozen Mamba backbone with an explicit k-mer memory and a local read head. Attention in LIND corresponds to memory sources selected by genomic address, so matched interventions and controls test positional explanations while the input sequence and parameters remain unchanged. To avoid the influence of dataset artifacts, the aforementioned tests were conducted under conditions where LIND achieved relative improvements across 16 Nucleotide Transformer tasks. At splice acceptors, attention is enriched 1.54-fold at the canonical AG, yet removing the underlying sources leaves AUROC essentially unchanged, opposite to the prespecified direction, while full memory removal costs 0.281. Attention looks at the AG in this Mamba case, yet silencing these sources does not reduce AUROC. In this model, looking is not depending. Our code can be found in the Supplementary Materials accompanying this paper.

1 INTRODUCTION

Predicting gene regulatory activity, chromatin state, and splicing from nucleotide sequence is a fundamental problem in genomics. These predictions support functional annotation of noncoding regions and the interpretation of genetic variants (Zhou & Troyanskaya, 2015; Kelley et al., 2016; Jaganathan et al., 2019; Avsec et al., 2021a). Building on pretraining methods from natural language processing, nucleotide sequence models such as DNABERT (Ji et al., 2021), Nucleotide Transformer (Dalla-Torre et al., 2025), Evo (Nguyen et al., 2024), and SpliceBERT (Chen et al., 2024) have expanded the range of sequence-based predictions. As predictive accuracy improves, a further question becomes important. Do these predictions rely on task-relevant biological signals or on incidental correlations in the data? Testing whether models have learned established biological relationships requires evidence about how they make their predictions (Consens et al., 2025).

For a positional explanation, we need to identify which sequence positions a prediction depends on and whether they coincide with known functional elements, such as splice sites and transcription-factor binding sites. One common form of evidence is attention alignment. DNABERT uses attention to identify informative sequence regions (Ji et al., 2021); SpliceBERT reports elevated attention between splice donor and acceptor sites (Chen et al., 2024); and Consens *et al.* associate attention heads with biological annotations (Consens et al., 2025). Such alignment connects a model's representation to biological knowledge, but attention concentrated at a position does not by itself establish that prediction depends on it. The distinction is particularly important in genomics because a functional element is also a specific sequence. The canonical acceptor AG is a splice signal (Burset et al., 2000), yet the same dinucleotide also occurs outside acceptor sites. Elevated attention at a true splice-site AG might merely reflect the model's local recognition of this dinucleotide (k-mer identity), rather than a genuine dependence on that specific position for the final prediction. If the model relies instead on

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.

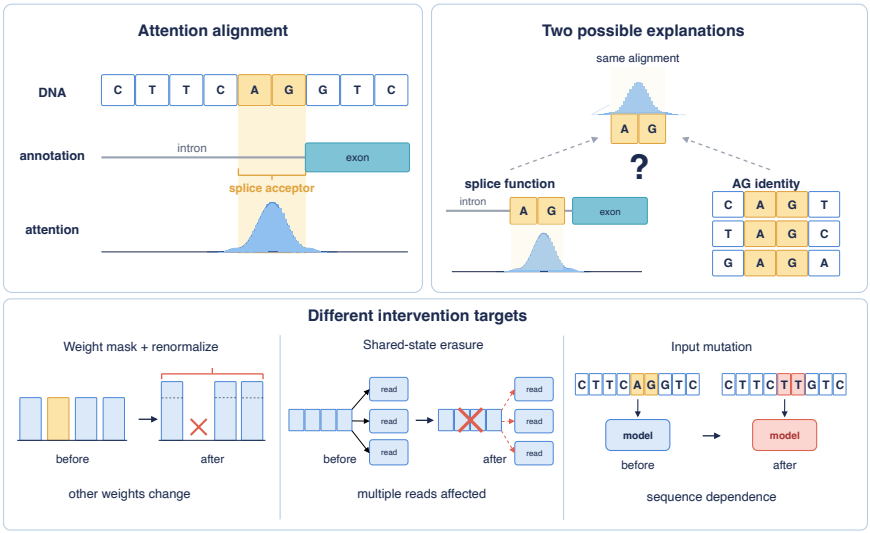


Figure 1: Attention alignment and the intervention question (schematic). Top left: attention can align with an annotated splice-acceptor AG. Top right: the same alignment alone cannot distinguish sensitivity to splice context from recognition of the AG sequence. Bottom: masking weights, erasing a shared state, and input mutation act on different computational objects (Section 1).

the broader sequence context to make its decision, removing access to this highly attended AG would surprisingly not harm performance. This highlights why alignment alone cannot establish predictive dependence. The NLP debate on attention as explanation has already highlighted the need to test the link between attention and predictive importance (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). For genomic interpretation, the corresponding question is whether an enriched attention map identifies memory sources of information whose removal harms prediction. Figure 1 illustrates why alignment alone cannot answer this question.

Intervention provides a way to test this question, but the object being changed determines what the result can establish. Attention-free Mamba and convolutional backbones do not expose attention weights for this analysis (Gu & Dao, 2023; Bo et al., 2025a). In attention-based models, masking selected weights and renormalizing redistributes the remaining attention (Serrano & Smith, 2019). Erasing a shared internal state can affect multiple downstream reads, so its effect depends on where the operation is applied (Vig et al., 2020; Meng et al., 2022). Input mutagenesis tests sequence changes, while input attribution assigns contributions to bases; neither directly specifies which internal read produced the prediction (Zhou & Troyanskaya, 2015; Sundararajan et al., 2017; Shrikumar et al., 2017). Existing genomic studies have recognized the need for perturbation, but they intervene at scales that cannot isolate specific positional reads. For instance, Enformer compares attention to sequence-replacement scores at enhancer windows (Avsec et al., 2021a); however, mutating the input sequence triggers cascading changes throughout the computational graph, failing to isolate the specific internal representation read at that exact site. Similarly, Consens *et al.* ablate entire attention heads associated with biological annotations (Consens et al., 2025). While this establishes a head's global utility, this coarse intervention shuts down reads across the entire sequence, making it impossible to confirm if the prediction truly depended on a localized read at a specific biological signal (e.g., a single splice-acceptor).

Our question, in contrast, concerns the precise internal representation read at a single annotated position. Addressing this requires a highly targeted intervention: we must intervene strictly on the specific computational object the attention was measured on, holding both the input sequence and the rest of the model parameters intact. Therefore, establishing an explicit correspondence between the observed attention and the intervened content becomes the fundamental design challenge.

A test of this correspondence needs three properties. First, each observed read source must map to a genomic coordinate. Second, the intervention must select sources at those addresses while holding the input sequence and model parameters fixed. Third, the test must support matched position and

sequence controls, a prespecified intervention direction, and replication on independent samples. The design challenge is to expose the read content as a selectable computational object and define its mapping to an annotation. A contextual representation may combine information from several positions, so its sequence index alone does not specify which read content an intervention removes. We therefore make position-indexed memory sources explicit, allowing the same source addresses to connect attention measurement with intervention.

We introduce LIND, a Mamba-based DNA model with an explicit k-mer memory and a local attention *read head* (Figure 2). The memory contains learned vectors indexed by the k-mer at each position, and the read head queries nearby entries within a 33-bp window. Each read source carries a genomic address, allowing the corresponding identifiers to be selected for intervention while the input sequence and model parameters remain fixed. PROTOCOL-F connects the predictive gain from adding the read head, attention enrichment at annotations, and the effect of intervening on the corresponding sources. Position and sequence controls test specificity, and disjoint samples test replication. The case list and intervention directions were fixed before the first intervention. Section 2.4 defines the intervention operations and the four evaluation criteria.

The results separate looking from depending through a progressive logical chain. As a necessary baseline to avoid the influence of dataset artifacts that could compromise downstream interpretability, we confirmed the read head's functional efficacy. It improves all 16 retained Nucleotide Transformer (NT) tasks with a mean AUROC gain of $+0.1020$, which holds across twelve tasks without test-in-train overlap. Building on this validated module, we evaluated the interpretability of attention weights. We observed that attention concentrated at functional elements superficially satisfies conventional hallmarks of biological relevance. Specifically, attention at canonical splice acceptors is enriched 1.54-fold at the AG dinucleotide, replicates at 1.34-fold on 2,000 disjoint sequences, and collapses to background when the AG is mutated to TT or shuffled. Yet, a matched intervention contradicts these attention-based explanations: while silencing the whole memory costs 0.281 in AUROC, silencing all memory sources under the AG leaves the acceptor AUROC essentially unchanged rather than reduced as prespecified. Extending beyond splice sites, descriptive comparisons on other functional elements point in the same direction. In pooled codon and transcription-factor comparisons, attention weights vary far more with local sequence identity than annotation context, yielding identity-to-context spread ratios from 8.5 to 36.8 (Appendix I.2). We call this pattern the **identity–dependence gap**. Ultimately, these findings address our core problem regarding the interpretability of attention: high and sequence-sensitive attention weights merely show where the model looks, not what the prediction depends on. Our contributions are as follows:

- **An addressable read head for DNA models.** A local attention read head over explicit k-mer memory that improves every one of 16 retained NT tasks under a paired protocol with a shared frozen backbone, and whose attention can be mapped to position-indexed sources (Sections 2 and 3.2).
- **PROTOCOL-F.** A prespecified, locally locked four-criterion protocol covering enrichment, intervention direction, specificity against controls, and replication, that scores a positional attention claim on the object the attention was measured on (Section 2.4).
- **The identity–dependence gap.** At splice acceptors, replicated and sequence-specific enrichment coexists with a source intervention two orders of magnitude below the whole-memory effect and in the non-predicted direction; pooled codon and TF-motif comparisons show larger identity than context spreads, while the other annotation families delimit the scope of this observation (Sections 3.3 and 3.5).

2 METHOD: ADDRESSABLE MEMORY AND READ HEAD

Overview. LIND combines four bidirectional Mamba2 blocks with convolutional mixing, an explicit k-mer memory before blocks 0 and 2, and a local read head (Figure 2). Each block averages forward and backward Mamba2 outputs and then applies a multi-scale convolutional mixer. Pretraining supplies a common frozen backbone and memory; downstream training updates the classifier and, in the read-head arm, the added head. Masked mean pooling produces the sequence representation. The Mamba2 backbone (Dao & Gu, 2024) provides the fixed predictive context for the addressable read. Its implicit attention interpretation is distinct from the explicit memory-source weights measured here, which are tied to genomic addresses.

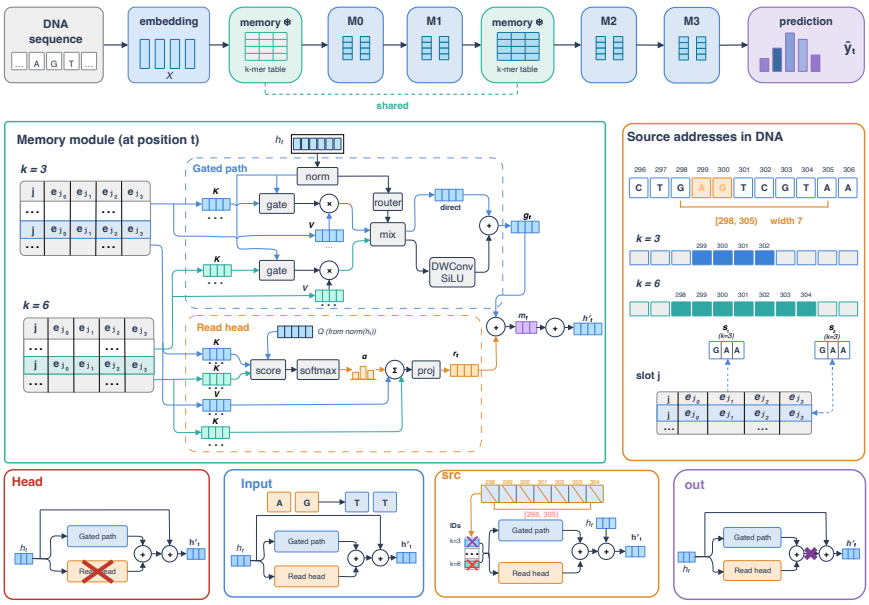


Figure 2: Model and intervention objects. Memory is reused at two backbone locations. Both arms contain the gated pathway; one also contains the attention read head. Source and output operations act at different points in this computation. Section 2.3 defines the exact covering span.

2.1 Explicit Memory and Shared Pathway

A contextual hidden state mixes information from many positions and offers no selectable object tied to a single genomic coordinate; to address this, LIND stores sequence content in an explicit memory whose sources carry genomic addresses. For each order $o \in \{3, 6\}$, a hash table stores learned embeddings of canonical k-mers. The source at address p covers $[p - \lfloor o/2 \rfloor, p - \lfloor o/2 \rfloor + o)$. A *slot* indexes sequence identity; an *address* specifies an occurrence. The same slot can therefore appear at multiple addresses. Source IDs are computed from the input sequence, independently of the hidden state. A gated pathway shared by both model arms conditions per-order memory values on the hidden state, combines them through a router, and adds a depthwise-convolution term. It uses the same source identifiers as the read head. Writing its contribution as $\mathbf{g}_t$ and the read-head contribution as $\mathbf{r}_t$, the memory update is $\mathbf{h}'_t = \mathbf{h}_t + \mathbf{g}_t + \mathbf{r}_t$. The hidden state supplies the gates and read query as well as the residual bypass; IDs select embeddings rather than entering the residual sum directly. Appendix B further specifies the complete forward computation, the exact memory geometry, and the full training configuration.

2.2 Local Read and Received Attention

Attention can be attributed to genomic positions only if its weights are defined over addressed sources; to address this, we add a local attention read head over the memory. At position t, the head attends jointly over valid order–address pairs $\mathcal{N}(t) = \{(o, s) : |s - t| \le 16\}$. Queries $\mathbf{q}_t$ are projected from normalized hidden states; keys $\mathbf{k}_s^{(o)}$ and values $\mathbf{v}_s^{(o)}$ come from the memory embeddings:

$$\alpha_{t,o,s} = \frac{\exp(\mathbf{q}_t^\top \mathbf{k}_s^{(o)} / \sqrt{d_a})}{\sum_{(o',s') \in \mathcal{N}(t)} \exp(\mathbf{q}_t^\top \mathbf{k}_{s'}^{(o')} / \sqrt{d_a})}, \qquad \mathbf{r}_t = W_{\text{out}} \sum_{(o,s) \in \mathcal{N}(t)} \alpha_{t,o,s} \mathbf{v}_s^{(o)} + \mathbf{b}_{\text{out}}. \tag{1}$$

where $d_a = 32$ is the attention dimension, and W_{out} and $\mathbf{b}_{\text{out}}$ are the output projection and bias. The read adds to the residual stream. Its zero-initialized output projection makes the head inactive when the pretrained checkpoint is loaded. The 16,672 head parameters are shared across the two memory applications of the model.

We measure attention from the source side, summing the mass received from queries within reach and averaging over memory applications $j \in \{0, 2\}$:

$$A(p) = \tfrac{1}{2} \sum_{j \in \{0,2\}} \sum_{t:|t-p| \le 16} \sum_{o \in \{3,6\}} \alpha^{(j)}_{t,o,p}. \tag{2}$$

Enrichment is mean $A(p)$ at annotations divided by mean background $A(p)$ in the same sequences. This allocation is not a decomposition of prediction: values and downstream computation also matter. Erasing what an attention layer reads has prior precedent (Serrano & Smith, 2019); the address map defines our genomic intervention object.

2.3 Matched Read Interventions

Testing whether the prediction depends on an attended position requires changing only the attended object (Section 1); to address this, our operations select memory sources by genomic address while the input sequence and parameters stay fixed.

Source removal. The source operation (`src`) invalidates identifiers at selected addresses for both orders. Removed sources leave the attention normalization set, so other weights renormalize; the shared gated pathway also receives modified identifiers. It holds bases and parameters fixed while changing which sources are available. For AG at $[300, 302)$, overlapping order-3 addresses are $\{299, \ldots, 302\}$ and order-6 addresses are $\{298, \ldots, 304\}$. Applying their union $[298, 305)$ to both orders selects 14 entries: all 11 overlaps plus three additional order-3 entries.

Read-head values under fixed attention. To test the read head separately from this shared-source operation, we additionally zero selected projected read-head values, after the value projection and its bias, while replaying the clean attention weights separately at each memory application. For a source set S, the intervened read is

$$\mathbf{r}_t^{(-S)} = W_{\text{out}} \sum_{(o,s) \in \mathcal{N}(t)} \alpha^{\text{clean}}_{t,o,s} \mathbb{1}[(o, s) \notin S] \mathbf{v}_s^{(o)} + \mathbf{b}_{\text{out}}. \tag{3}$$

This operation leaves the shared branch's identifiers and parameters intact; later hidden states and gates can still respond to earlier changes. Clean weights are held fixed at both applications, including an unmasked application in the single-call controls. Thus the test measures value contributions under the observed allocation, not unrestricted adaptation of later attention. Zeroing all values retains $\mathbf{b}_{\text{out}}$; zeroing the entire read also removes that bias.

Other intervention objects. The output operation (`out`) masks the combined memory residual after smoothing and the read. It changes what the memory writes *at* a position, whereas `src` changes what other positions can read *from* it. Same-width offsets, input substitutions, slot interventions, and whole-memory operations place these effects in context.

2.4 Evaluation Protocol and Criteria

We test whether attention aligned with an annotated position identifies internal read sources that support predictive performance. For a trained DNA predictor $f_\theta(x)$, let $A(p)$ denote the attention received by internal sources at genomic position p, and let $S(B)$ map annotated bases B to the source addresses whose k-mers overlap them. Attention enrichment compares $A(p)$ at the annotation with background positions. A matched intervention instead changes the sources in $S(B)$, keeps x and θ fixed, and measures the resulting predictive change. We report $\Delta\text{AUROC} = \text{AUROC}_{\text{intervention}} - \text{AUROC}_{\text{intact}}$; the specified claim of positive support for predictive performance predicts a decrease. Dependence of individual predictions is a separate quantity, measured through probability changes and prediction class flips.

These measurements answer distinct questions: whether adding the read head improves prediction (L1), whether its attention aligns with an annotation (L2), and how prediction responds to changing the attended sources (L3). A small AUROC change concerns ranking on the evaluated sample, not invariance of every individual prediction. PROTOCOL-F connects these measurements through four criteria: **enrichment**, elevated attention at annotations relative to within-sequence background;

intervention direction, a loss in predictive performance when the specified sources are removed; **specificity**, contrasts against the designated position, sequence, and source controls; and **replication**, enrichment on a disjoint sample using the same checkpoint. Section 2.3 defines the operations, and Appendix E gives the tests and correction family. An untested criterion is recorded separately from one not met. These criteria evaluate the stated positional claim; they do not define biological causality.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Data and comparison. We first establish the predictive utility of the read head, then test its positional interpretation at acceptor AG and its transfer to a second backbone. Both arms already contain memory and freeze the same 1,709,800 pretrained parameters. The control trains 4,902 classifier parameters; LIND trains these plus the 16,672 read-head parameters. The comparison measures adding the head under a fixed recipe, with unequal trainable capacity and one training seed. We evaluate all 18 NT tasks (Dalla-Torre et al., 2025) and retain 16 for the headline comparison: the two enhancer tasks have 25% exact test-in-train overlap, compared with at most 7.27% for the others. Appendix A.1 reports every task and alternative screens. Both arms use 30 epochs, learning rate 2×10^{-4}, batch size 32, a 10% validation split (seed 1337), and training seed 1; the best validation checkpoint is tested once.

Metrics and samples. Table 1 uses AUROC throughout: class-1 probabilities for binary tasks and macro one-versus-rest AUROC for the three-class combined splice-site task. We re-evaluate the same 32 selected checkpoints on their complete test splits in bf16, without retraining or changing checkpoint selection. Native benchmark metrics (MCC, F1, and accuracy), validation results, and overlap remain in Appendix A.1. AUROC also measures the intervention effects. Acceptor enrichment uses 500 discovery and 2,000 site-disjoint RefSeq replication windows; interventions use all 2,218 NT acceptor test sequences. The two analyses share the checkpoint and annotation-to-source mapping, rather than individual examples. Six position controls have source centers 120, 180, 240, 360, 420, and 480 and match the intervention width. The case list and expected directions were fixed before the first intervention; Appendix D documents the local lock, the test-split analysis, and later corrections. Intervals are 95% bootstrap CIs when available. Corrected tests use the recorded Benjamini–Hochberg family (Benjamini & Hochberg, 1995); offset z is reported as a descriptive contrast, not as another significance test (Appendix E).

3.2 PREDICTIVE PERFORMANCE

First, the addressable read head improves prediction on every retained task. The read-head arm raises test AUROC on all 16 tasks (Table 1), from a task mean of 0.7630 to 0.8650 (+0.1020); individual gains range from +0.0339 to +0.2947. **Second, the gains are largest on the splice-site tasks.** Mean family gains are 0.2757 for splice sites, against 0.0702 for histone marks and 0.0345 for promoters. **Third, the improvement is robust to overlap screening and metric choice.** All twelve zero-overlap tasks improve, and the original native-metric validation comparison agrees in direction on all 16 tasks (Appendix A.1).

Within the trained read-head arm, zeroing the head's output projection lowers AUROC on all 18 tasks by 0.088–0.493 (Appendix C). This operation measures sensitivity to removing a term used during training; it does not give a capacity-controlled gain or identify the responsible positions. The splice acceptor task therefore supplies a precise next test: a known canonical dinucleotide coordinate and an observed attention enrichment.

3.3 REPLICATED ATTENTION ALIGNMENT AT ACCEPTOR AG

The stored discovery analysis reports 1.54-fold received-attention enrichment at AG (95% CI $[1.38, 1.70]$, $q = 7.9 \times 10^{-10}$, $n = 500$), replicating at 1.34-fold on 2,000 disjoint sequences (CI $[1.30, 1.39]$, $q = 4.0 \times 10^{-46}$). The 1.42 fold used for the sequence controls is a separate AG-conditioned site summary, not a re-estimate of the registered 1.54-fold statistic. AG has the largest

Table 1: Test AUROC for the same seed-1 checkpoints on all retained tasks. Binary tasks use class 1; combined splice sites use macro one-versus-rest AUROC. ΔAUROC is read head minus control, computed before rounding; mean is unweighted across tasks. Bold marks the higher score. The improvement is statistically significant under a two-sided paired t-test across the 16 tasks ($p = 3.4 \times 10^{-4}$). Native metrics and overlap are in Appendix A.1.

Task	Control	Read head	ΔAUROC
Histone marks			
H3	0.8672	**0.9492**	+0.0820
H3K4me1	0.6967	**0.7695**	+0.0728
H3K4me2	0.6528	**0.6900**	+0.0373
H3K4me3	0.6251	**0.6938**	+0.0688
H3K9ac	0.7514	**0.8284**	+0.0770
H3K14ac	0.6866	**0.7818**	+0.0952
H3K36me3	0.7478	**0.8186**	+0.0708
H3K79me3	0.8038	**0.8627**	+0.0589
H4	0.8983	**0.9488**	+0.0505
H4ac	0.6716	**0.7604**	+0.0888
Promoters			
Non-TATA	0.9574	**0.9919**	+0.0345
All	0.9590	**0.9929**	+0.0339
TATA	0.9525	**0.9874**	+0.0349
Splice sites			
Acceptor	0.6513	**0.9131**	+0.2618
Donor	0.6412	**0.9117**	+0.2705
All (macro OvR)	0.6453	**0.9400**	+0.2947
Mean, retained	0.7630	**0.8650**	+0.1020

excess of site fold over its own sham-offset fold among the 16 dinucleotides (+0.30, $q = 1.4 \times 10^{-7}$), although raw folds do not identify AG as uniquely preferred. In the matched sequence-control analysis, mutating AG to TT lowers the fold from its own baseline of 1.42 to 1.03; shuffling while preserving base composition lowers it to 1.01. CC retains 1.39, indicating sensitivity to local k-mer content rather than to the AG letters alone. Other task checkpoints show weaker enrichment (H3K4me3 1.17; donor 1.10). These observations establish replicated, sequence-sensitive alignment. They do not yet establish that the aligned sources support predictive performance.

3.4 Prediction Under Matched Read Interventions

Do the aligned sources support predictive performance? Whole-memory source removal reduces acceptor AUROC by 0.2809 and mean predicted probability on true acceptors from 0.774 to 0.349. Removing sources in the AG covering span instead changes AUROC by +0.0027 (recomputed CI $[+0.0011, +0.0044]$), opposite to the specified decrease (Table 2). This is a small positive effect, not an estimate of zero. The six same-width source controls average −0.0010 (SD 0.0010), yielding a descriptive site contrast $z = +3.66$. Thus the site differs from the control mean in the unpredicted direction. The whole-memory loss provides context for operation scale, not a test that the local effect must be proportionally large.

Read-head values under fixed attention. Source removal changes both memory paths and their attention allocation. To check the read head separately, we additionally intervene on exactly the four order-3 and seven order-6 sources overlapping AG, using Equation 3. The identity replay reproduces all 2,218 baseline logits exactly. Removing the selected values increases AUROC by 0.0026 (95% paired bootstrap CI $[+0.0008, +0.0044]$), whereas zeroing all read-head values reduces it by 0.2899. The effect is not an absence of predictive influence: mean absolute change in positive-class probability is 0.0225, and 54 of 2,218 predicted classes change (2.43%). Alignment therefore does not establish a beneficial contribution under this operation, even though individual predictions depend on the intervened values. Applying the covering union to both orders gives +0.0023; single-application and six offset controls are in Appendix G. Additional training-run checks are reported in Appendix G.1; these are distinct from the independent-sample enrichment replication.

Table 2: Acceptor interventions on 2,218 test sequences. ΔAUROC is intervention minus intact. Head-only values use fixed clean attention; source removal changes both paths. Intervals are 95% paired bootstrap CIs. Input effects are across 16 substitutions.

Operation	Object and extent	ΔAUROC	Interval or context
Local head values	Exact AG overlaps (4/7)	$+0.0026$	$[+0.0008, +0.0044]$
All head values	Values, output bias retained	-0.2899	$[-0.3141, -0.2663]$
Whole read head	Entire head output	-0.3736	$[-0.3989, -0.3462]$
Whole source	All identifiers	-0.2809	Global operation
Local source	$[298, 305)$, both orders	$+0.0027$	$[+0.0011, +0.0044]$
Local output	Combined output, same span	-0.0418	$[-0.0508, -0.0324]$
Input	Bases $[300, 302)$	—	$[-0.0032, +0.0070]$

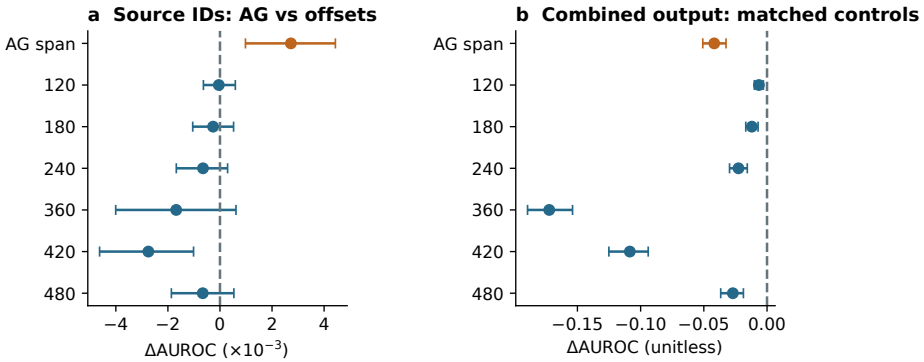


Figure 3: Shared-memory source-ID and combined-output interventions. Both use the corrected width-7 span and six controls, labelled by absolute source centre. Error bars are 95% CIs; horizontal scales differ. ΔAUROC is intervention minus intact on the acceptor test split. The separate read-head value test is reported in Table 2.

Masking the combined memory output at the same span reduces AUROC by 0.0418 ($q = 0.0015$), but output controls also have varying losses (mean -0.058, SD 0.067). The AG window has no excess loss relative to that control summary ($z = +0.24$). Figure 3 shows the distinct source and output responses. The width sweep reduces the source contrast from $z = +3.66$ at width 7 to $+1.70$ at width 11; larger widths provide no comparably distinctive site contrast (Appendix F).

Input substitutions at the two AG coordinates range from -0.0032 to $+0.0070$ AUROC; 12 of 16 increase it. Matched substitutions 150 bp downstream range from -0.0016 to $+0.0010$. None reaches the prespecified $|\Delta P| > 0.02$ on true acceptors carrying AG, so the planned mediation fraction is undefined, not zero. Slot interventions also fail the specificity criterion against frequency-matched and uniform controls; only the low-frequency comparison passes (Appendices E and H.1). The acceptor case meets enrichment and replication, but the tested local read operations do not satisfy the predicted direction. The slot analyses also fail the designated specificity comparisons.

3.5 Identity and Annotation Context

To examine the sequence–function ambiguity beyond AG, we compare attention across sequence identities and annotation contexts. The other families also define the observation's scope. Donor GT (fold 0.93), poly(A) AATAAA (0.89), and CTCF (1.00) do not meet the enrichment prerequisite; their full intervention chains were not run. Codon and TF comparisons hold identity fixed while varying annotated role. No codon comparison detects a context difference (Welch $p = 0.08$–0.93); 1 of 14 eligible TF motif–class comparisons is positive at $q < 0.05$. Pooled identity-to-context spread ratios are 8.5 for codons and 36.8 for distal-enhancer motifs, using distinct case-specific attention measures. These descriptive ratios are neither fractions of explained variance nor evidence that contexts are equivalent. Appendix I.2 retains all pooled and per-order estimates and the coverage table; exploratory poly(A) spatial results are in Appendix H.2.

3.6 Transfer to a Convolutional Backbone

To test whether the read interface requires Mamba, we reuse a pretrained ConvNova gated-convolution backbone (Bo et al., 2025b) with the same explicit memory and add the same head. On the acceptor task, three paired training seeds improve AUROC from 0.6497 ± 0.0027 to 0.7872 ± 0.0020 (mean $\pm$ sample SD), a mean gain of 0.1375; binary F1 improves from 0.6935 to 0.7482. The split, training budget, and validation-selection rule match the primary comparison. **The read interface and predictive gain transfer to ConvNova.** This comparison uses our pretrained implementation with memory in both arms. Architecture, training details, variability across runs, and the additional AG analyses are reported in Appendix J. The Mamba acceptor case supplies the main alignment–intervention finding; transfer of the design does not require the same sequence-specific attention pattern in both backbones. Limitations and future work are discussed in Appendix K.

4 Related Work

DNA sequence models. Transformer models pre-trained on genomes provide transferable representations across regulatory and splicing tasks (Ji et al., 2021; Zhou et al., 2024; Dalla-Torre et al., 2025); long-convolution and state-space models extend the context at lower cost (Nguyen et al., 2023; Schiff et al., 2024; Bo et al., 2025b). Mamba has implicit attention interpretations (Ali et al., 2024; Dao & Gu, 2024). LIND exposes an explicit k-mer read and tests its genomic sources; the paired benchmark evaluates this addition to a fixed backbone.

Sequence interpretation. DeepSEA and Basset related regulatory predictions to sequence features using input mutagenesis (Zhou & Troyanskaya, 2015; Kelley et al., 2016); SpliceAI and Enformer predict splicing and long-range expression, respectively (Jaganathan et al., 2019; Avsec et al., 2021a), and attribution methods such as Integrated Gradients and DeepLIFT assign input contributions relative to a reference (Sundararajan et al., 2017; Shrikumar et al., 2017). BPNet combined DeepLIFT-based contribution scores with TF–MoDISco for motif discovery (Avsec et al., 2021b). Attention-based accounts examine the model's allocation of weight: SpliceBERT reports elevated attention between donor and acceptor sites (Chen et al., 2024), and Consens *et al.* associate genomic attention heads with biological features and ablate head sets against layer-matched controls (Consens et al., 2025). Building on these associations and perturbation tests, Protocol-F evaluates an explicit memory source with an annotation-aligned address and a prespecified intervention direction.

Attention and internal interventions. Jain and Wallace showed that alternative attention distributions and input erasure can leave predictions unchanged (Jain & Wallace, 2019); Wiegreffe and Pinter replied that the test must be model-consistent and the notion of explanation stated (Wiegreffe & Pinter, 2019); Serrano and Smith erased the representations an attention layer reads and renormalized the rest (Serrano & Smith, 2019). Analyses of attention identifiability and of the norms of the attended vectors explain why weights alone need not name a prediction's contributors (Brunner et al., 2019; Kobayashi et al., 2020), and causal mediation and causal tracing intervene on internal variables with explicit effect definitions (Vig et al., 2020; Meng et al., 2022); randomization tests check that an explanation depends on the model at all (Adebayo et al., 2018). These studies test distinct objects, including inputs, attention distributions, intermediate representations, heads, and internal states at selected positions and layers. Our study connects a genomic annotation to explicit memory-source addresses and evaluates the resulting intervention against a specified direction and matched controls. It does not claim that internal or position-specific interventions are new in themselves.

5 Conclusion

LIND connects positional attention to explicit, addressable memory sources, enabling matched interventions with the sequence and parameters held fixed. The read head achieves relative improvements across the 16 Nucleotide Transformer tasks, and its predictive benefit transfers to ConvNova. In the Mamba acceptor case, replicated attention enrichment at AG coexists with source and read-value interventions that do not reduce AUROC. This finding separates attention alignment from support for predictive performance under the tested operations. Positional explanations therefore benefit from testing the attended sources with explicit interventions and matched controls.

AI USE STATEMENT

The authors used generative AI assistants for experiment and analysis code, training and evaluation workflows, aggregation of stored outputs, figure-generation code and diagram drafts, literature-search assistance, and manuscript organization, drafting, and revision. This assistance extended beyond language polishing. Scientific interpretation and the final selection of claims remain the authors' responsibility. The manuscript distinguishes stored experimental measurements from schematic illustrations, and the supplement records analysis scripts and outputs. Bibliographic sources require author verification; AI-assisted work does not substitute for checking the underlying code, measurements, and cited papers.

ETHICS STATEMENT

This work studies computational models on public sequence benchmarks and genomic annotations. It does not present a clinical validation or infer causal biological mechanisms from attention alone. Public availability does not remove data-use restrictions; dataset licenses and access conditions apply to reuse. Attention maps and internal interventions should not be used as standalone evidence for clinical decisions or variant pathogenicity.

REPRODUCIBILITY STATEMENT

Sections 2 and 3.1 define the model, operations, and evaluation. The appendix reports all 18 tasks and overlap checks (Appendix A.1), training and architecture settings (Appendix B), protocol provenance and analysis splits (Appendix D), and complete acceptor statistics and controls (Appendices E–H). Code and stored analysis outputs are supplied anonymously. The acquisition scripts identify public datasets; fine-tuned checkpoints and code are intended for release at publication.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Ameen Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of mamba models. *arXiv preprint arXiv:2403.01590*, 2024. URL https://arxiv.org/abs/2403.01590.

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. *Nature 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, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics*, 53(3):354–366, 2021b.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yu Bo, Weian Mao, Daniel Shao, Weiqiang Bai, Peng Ye, Xinzhu Ma, Junbo Zhao, Hao Chen, and Chunhua Shen. Revisiting convolution architecture in the realm of dna foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 37848–37870, 2025a.

Yu Bo, Weian Mao, Yanjun Shao, Weiqiang Bai, Peng Ye, Xinzhu Ma, Junbo Zhao, Hao Chen, and Chunhua Shen. Revisiting convolution architecture in the realm of DNA foundation models. *arXiv preprint arXiv:2502.18538*, 2025b. URL https://arxiv.org/abs/2502.18538.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*, 2019.

Moisés Burset, Igor A Seledtsov, and Victor V Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic acids research*, 28(21):4364–4375, 2000.

Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440, 2022.

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. *Briefings in bioinformatics*, 25(3):bbae163, 2024.

Micaela E Consens, Ander Diaz-Navarro, Vivian Chu, Lincoln Stein, Housheng Hansen He, Alan Moses, and Bo Wang. Interpreting attention mechanisms in genomic transformer models: a framework for biological insights. *bioRxiv*, pp. 2025–06, 2025.

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

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. URL https://arxiv.org/abs/2405.21060.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548, 2019.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 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.

David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, et al. Jaspar 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 52(D1):D174–D182, 2024.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proceedings of machine learning research*, 235:43632, 2024.

Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, 27(5):849–864, 2017.

Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2931–2951, 2019.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

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

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *International Conference on Learning Representations*, volume 2024, pp. 41642–41665, 2024.

Technical Appendix

Table of Contents

A Data and Evaluation Splits **15**
A.1 Exact-Sequence Overlap . . . 15
A.2 Annotation and Checkpoint Mapping . . . 15

B Model and Training Details **15**
B.1 Backbone and Pre-training . . . 15
B.2 Read Head . . . 16
B.3 Fine-tuning . . . 16
B.4 Trainable Parameters . . . 16
B.5 Memory Geometry and Forward Computation . . . 16

C Read-Head Output Zeroing **17**

D Analysis Protocol and Provenance **17**
D.1 Locked Document . . . 17
D.2 Analysis Split . . . 18

E Acceptor Statistics **18**
E.1 Reading the Estimates . . . 18

F Memory Interventions **18**
F.1 Offset Profile . . . 18
F.2 Span Correction . . . 19
F.3 Width Sweep . . . 19

G Read-Head Value Interventions and Uniform-Metric Reevaluation **19**
G.1 Additional Training Seeds for the Primary Backbone . . . 21

H Input and Annotation Controls **21**
H.1 Sequence Intervention . . . 21
H.2 Exploratory Poly(A) Spatial Analysis . . . 21

I Identity and Annotation Context **22**
I.1 Coverage Across Annotation Families . . . 22
I.2 Identity and Context Spreads . . . 23

J Transfer to a ConvNova Convolutional Backbone **24**

K Interpretation, Limitations, and Future Work **25**

L Concepts Definition **25**

A Data and Evaluation Splits

A.1 Exact-Sequence Overlap

For each of the 18 NT tasks we counted the test sequences that appear verbatim in that task's own training split (Table 3). The threshold separating retained from excluded tasks is not a round number chosen in advance: the empirical distribution has a gap between all splice sites at 7.3% and the two enhancer tasks at 25.0%, with nothing in between. Under alternative screens the original native-metric arm comparison gives: strict zero overlap, 12 tasks, mean ΔScore $+0.132$, 12 of 12 improved; overlap at most 1%, 15 tasks, $+0.118$, 15 of 15; no screen, 18 tasks, $+0.120$, 17 of 18, the exception being enhancers at -0.019 (control 0.4333, read head 0.4147; enhancer types 0.2905 and 0.3304, $+0.040$).

Table 3: Test-in-train overlap per task, with test-set size and both arms on the validation and the test split. Scores use MCC for histone/enhancer tasks, F1 for promoters and acceptor/donor, and accuracy for combined splice sites. ΔScore is read-head minus control. The last two rows are the excluded enhancer tasks, reported for completeness.

			Validation score			Test score			
Task	n_{test}	Overlap	Ctrl	Read	ΔScore	Ctrl	Read	ΔScore	Status
H3	1497	0.0000	0.6014	0.7727	+0.1713	0.5850	0.7702	+0.1851	retained
H3K4me1	3168	0.0000	0.3357	0.4263	+0.0906	0.2905	0.4195	+0.1290	retained
H3K4me2	3069	0.0000	0.2421	0.2984	+0.0563	0.1991	0.2830	+0.0839	retained
H3K4me3	3680	0.0000	0.1810	0.3288	+0.1478	0.1952	0.2738	+0.0785	retained
H3K9ac	2779	0.0000	0.3353	0.5088	+0.1735	0.3631	0.5025	+0.1393	retained
H3K14ac	3305	0.0000	0.2902	0.4195	+0.1293	0.2464	0.4204	+0.1741	retained
H3K36me3	3488	0.0000	0.4139	0.4793	+0.0654	0.3885	0.4954	+0.1068	retained
H3K79me3	2884	0.0000	0.4766	0.5955	+0.1189	0.4720	0.5768	+0.1049	retained
H4	1461	0.0000	0.6442	0.7921	+0.1479	0.6438	0.7873	+0.1435	retained
H4ac	3410	0.0000	0.3031	0.4272	+0.1241	0.2382	0.3719	+0.1337	retained
Splice acc.	2218	0.0000	0.7083	0.8561	+0.1478	0.6795	0.8415	+0.1620	retained
Splice don.	2198	0.0000	0.6992	0.8529	+0.1537	0.7066	0.8469	+0.1403	retained
Prom. non-TATA	5299	0.0028	0.9009	0.9583	+0.0574	0.8942	0.9584	+0.0642	retained
Prom. all	5920	0.0035	0.9052	0.9649	+0.0596	0.8968	0.9600	+0.0632	retained
Prom. TATA	621	0.0097	0.8872	0.9442	+0.0570	0.8943	0.9560	+0.0617	retained
Splice, all	3000	0.0727	0.4641	0.8100	+0.3459	0.4460	0.8173	+0.3713	retained
Enhancers	400	0.2500	0.5496	0.7772	+0.2276	0.4333	0.4147	−0.0186	excluded
Enhancer types	400	0.2500	0.4330	0.7043	+0.2713	0.2905	0.3304	+0.0399	excluded

A.2 Annotation and Checkpoint Mapping

The splice analyses use their corresponding acceptor or donor checkpoints. The poly(A), CTCF, codon, and transcription-factor analyses use the H3K4me3 read-head checkpoint unless explicitly stated otherwise for the exploratory poly(A) follow-up. Sequence windows and annotation positions are reconstructed using the annotation-processing scripts in the code supplement. The resource download script and checksum lists identify the input files. Transcription-factor motifs come from JASPAR (Rauluseviciute et al., 2024), with functional intervals defined using ENCODE candidate cis-regulatory elements (Moore et al., 2020). The attention analysis compares annotated positions with background within the same sequences; downstream acceptor interventions instead use the benchmark test records.

B Model and Training Details

B.1 Backbone and Pre-training

The LIND backbone uses bidirectional Mamba2 with $d_{\text{model}} = 128$ and four blocks, with the k-mer memory module (orders 3 and 6, embedding width 64, hash table of 4,099 slots per order) applied before blocks 0 and 2 (zero-based). Each block averages independent forward and reversed-sequence Mamba2 outputs, adds the residual, and applies a multi-scale convolutional mixer with

kernel size 5 and dilations 1, 4, and 16. This is a sequential convolutional branch after bidirectional Mamba2, not the separate parallel-hybrid configuration. The Mamba2 internal convolution uses kernel size 4; the memory pathway also has its own depthwise convolution. Pre-training optimizes a masked-reconstruction objective (masked tokens are predicted from the unmasked context; "masked-reconstruction checkpoint" below refers to this stage) and ran for 10^6 steps on a three-corpus mixture—the GRCh38.p14 human reference assembly (Schneider et al., 2017), human sequence with variants injected from the high-coverage 1000 Genomes cohort (Byrska-Bishop et al., 2022), and the 849-genome multi-species collection of the Nucleotide Transformer (Dalla-Torre et al., 2025)—across two GPUs, about 27.8 GPU-hours. Both downstream arms initialize from the same pre-trained checkpoint.

B.2 Read Head

Half-window $w = 16$ (33-bp span), $d_{\text{attn}} = 32$, a single head, one softmax jointly over both orders and all offsets, zero-initialized output projection. Parameters: query projection $128 \times 32 + 32$; key and value projections $2 \times (64 \times 32 + 32)$ each; output projection $32 \times 128 + 128$; total 16,672, shared by the two memory layers.

B.3 Fine-tuning

Both arms fine-tune for 30 epochs at learning rate 2×10^{-4}, batch size 32, in bf16, with a 10% validation split (split seed 1337), selecting the checkpoint with the best validation macro-average and evaluating it once on the test split. No class weighting is applied. The classifier uses masked mean pooling over sequence positions. Canonical reverse-complement collapsing of k-mer identities does not imply reverse-complement-invariant sequence pooling. The original 18-task comparison uses training seed 1; additional acceptor runs are reported in Appendices G.1 and J. All 36 runs (2 arms $\times$ 18 tasks) have stored final evaluation outputs.

B.4 Trainable Parameters

Both arms freeze the same 1,709,800 backbone parameters.

Arm	Read head	Task heads	Total trainable
Read-head arm	16,672	4,902	21,574
Control arm	—	4,902	4,902

B.5 Memory Geometry and Forward Computation

The address of a k-mer is a sequence position; the embedding slot identifies its canonical sequence content. An address intervention is local to selected occurrences, whereas a slot intervention affects occurrences sharing that slot. For AG at $[300, 302)$, the covering address span is $[298, 305)$. Because this union is applied to both orders, it also covers extra order-3 entries. The shared gated branch and the read head use the same identifiers. Invalid sources are excluded from attention; changing them can also affect the shared branch and later hidden states. An output mask instead multiplies the combined memory residual contribution after smoothing and the attention read. These implementation details determine the scope of the two interventions in Section 2.3.

For completeness, let $i_t^{(o)}$ be the source ID computed from the input k-mer and $\mathbf{e}_t^{(o)} = E_o[i_t^{(o)}]$. The shared branch uses

$$\mathbf{z}_t = \mathrm{LN}_h(\mathbf{h}_t), \qquad \mathbf{K}_t^{(o)} = \mathrm{LN}_K(W_{K,o}\mathbf{e}_t^{(o)} + \mathbf{b}_{K,o}), \tag{4}$$

$$\gamma_t^{(o)} = \sigma(\mathbf{z}_t^\top \mathbf{K}_t^{(o)}/\sqrt{d}), \qquad \mathbf{V}_t^{(o)} = W_{V,o}\mathbf{e}_t^{(o)} + \mathbf{b}_{V,o}, \tag{5}$$

$$\boldsymbol{\pi}_t = \mathrm{softmax}(W_R\mathbf{z}_t + \mathbf{b}_R), \qquad \mathbf{u}_t = \sum_o \pi_t^{(o)}\gamma_t^{(o)}\mathbf{V}_t^{(o)}, \tag{6}$$

$$\mathbf{g}_t = \mathbf{u}_t + \mathrm{SiLU}\big(\mathrm{DWConv}(\mathrm{LN}_O(\mathbf{u}))_t\big), \qquad \mathbf{h}'_t = \mathbf{h}_t + \mathbf{g}_t + \mathbf{r}_t. \tag{7}$$

Here $d = 128$, and DWConv has kernel size 3. The read-head query is another projection of $\mathbf{z}_t$; its key and value projections are separate from $W_{K,o}, W_{V,o}$. The control sets $\mathbf{r}_t = 0$. Corpus-feature and

router-prior branches are disabled in these runs. Padding IDs select the zero memory embedding, but affine projection biases in the gated branch can remain; source invalidation is therefore not identical to zeroing the full memory output. The read head explicitly excludes invalid IDs from its normalization set. IDs never enter the residual sum directly.

C Read-Head Output Zeroing

This ablation is within the read-head arm and is separate from the two-arm comparison of Table 1. For each of the 18 fine-tuned read-head checkpoints, the read head's output projection (W_{out} and $\mathbf{b}_{\text{out}}$ in Eq. 1) is set to zero at test time, which removes the read $\mathbf{r}_t$ from the residual stream while leaving the shared gated memory pathway and every other parameter unchanged, and the checkpoint is re-evaluated on the task's test split. AUROC falls on all 18 tasks, by 0.088 to 0.493 (Table 4); the 95% intervals from 2,000 bootstrap resamples of the test split exclude zero in every case. The classifiers were trained with the read term present and are evaluated on inputs without it, changing the distribution of representations received by the classifier. The drop measures sensitivity to this test-time operation, not a mathematical bound on the head's contribution, and is not compared with the arm difference, which measures a different quantity.

Table 4: Test AUROC of each read-head checkpoint before and after zeroing the read-head output projection. ΔAUROC is zeroed minus intact; the interval is a 95% bootstrap CI over 2,000 resamples of the test split. Sorted by ΔAUROC.

Task	n_{test}	Intact	Zeroed	ΔAUROC	95% CI
H4	1461	0.9489	0.4559	-0.4930	$[-0.5267, -0.4602]$
Prom. TATA	621	0.9874	0.5085	-0.4789	$[-0.5237, -0.4337]$
Splice, all	3000	0.9396	0.4622	-0.4774	$[-0.4936, -0.4611]$
Prom. non-TATA	5299	0.9919	0.5550	-0.4369	$[-0.4528, -0.4219]$
Splice don.	2198	0.9117	0.4862	-0.4255	$[-0.4516, -0.4008]$
Splice acc.	2218	0.9130	0.5394	-0.3736	$[-0.3999, -0.3489]$
Prom. all	5920	0.9928	0.6934	-0.2994	$[-0.3137, -0.2858]$
H3K9ac	2779	0.8285	0.5322	-0.2964	$[-0.3218, -0.2720]$
H3K14ac	3305	0.7817	0.5203	-0.2614	$[-0.2850, -0.2381]$
Enhancer types	400	0.7143	0.4661	-0.2482	$[-0.3089, -0.1890]$
H4ac	3410	0.7604	0.5651	-0.1954	$[-0.2165, -0.1736]$
H3K4me3	3680	0.6938	0.5389	-0.1549	$[-0.1788, -0.1310]$
Enhancers	400	0.7825	0.6396	-0.1430	$[-0.1968, -0.0915]$
H3	1497	0.9492	0.8151	-0.1341	$[-0.1543, -0.1145]$
H3K4me2	3069	0.6901	0.5585	-0.1315	$[-0.1528, -0.1090]$
H3K79me3	2884	0.8626	0.7340	-0.1286	$[-0.1455, -0.1104]$
H3K36me3	3488	0.8184	0.6922	-0.1262	$[-0.1419, -0.1112]$
H3K4me1	3168	0.7694	0.6816	-0.0877	$[-0.1031, -0.0726]$

D Analysis Protocol and Provenance

D.1 Locked Document

The case library (five cases, priority order, control families, direction of every intervention, and the allowed strength of conclusion per case) is a single file whose in-document lock note, SHA-256 digest, and modification time are:

File	`docs/protocol_f/case_library.md` (37,641 bytes)
SHA-256	`8f63c1f2168521ab98a879b64073a4ee adf42b0bda632d29dac1e58630abb84c`
Modification time	2026-09-03 21:22
Earliest acceptor intervention output	2026-09-03 23:36

The prespecified priority order was poly(A), splice donor/acceptor, codons, CTCF, and transcription-factor motifs; TATA and CpG were excluded before the fact. Cases without a corresponding NT task

(poly(A), CTCF, codons, transcription-factor motifs) were prespecified to use the H3K4me3 readhead checkpoint as a generic donor and to carry weaker conclusions than cases with a corresponding task. The prespecified direction for criterion (b) was negative in every case. The local lock note and filesystem times are not independent public timestamps. They document the available ordering evidence, not a public preregistration of all analyses.

D.2 Analysis Split

The locked case library fixes the cases, control families, and intervention directions; it does not name an analysis split. A working draft of the protocol, which was not locked, proposed using validation splits for analysis and reserving test splits for reported scores. The acceptor interventions (the offset profile, the width sweep, the slot-level intervention, and the sequence substitutions) were evaluated on the acceptor test split ($n = 2{,}218$) and are recorded as such in the output files; the case list and directions were fixed before they ran. The fixed case list limits one form of case selection, but evaluating on the test split does not remove analysis-selection risk, and we state the split wherever an intervention estimate is reported. Enrichment and replication use separately drawn samples ($n = 500$, seed 1337; $n = 2{,}000$, seed 2024). The later span correction and exploratory poly(A) offset analysis are identified separately below.

E Acceptor Statistics

The disjoint replication chromosome strata agree to three decimals (1.342, $n = 279$; 1.343, $n = 1{,}721$). Raw discovery site folds are 1.49 for CG and 1.42 for AG, with overlapping intervals. The AG-to-TT and mononucleotide-preserving shuffled-sequence folds of 1.03 and 1.01 do not show significant enrichment over background ($q = 0.26$ and 0.76); this is not an equivalence test. Reverse complementation gives fold 1.13. The whole-memory output operation lowers AUROC by 0.410. For the final slot intervention, 25 frequency-matched, 9 uniform, and 0 low-frequency control sets among 200 draws are at least as harmful as the selected acceptor slots ($q = 0.19$, 0.095, and 0.011 respectively).

All 25 recorded claims for the acceptor case appear in Table 5. The summary assigns 24 to the correction family; A2b is descriptive and has no p-value, so the Benjamini–Hochberg calculation uses the 23 available tests. A permutation test with m control draws cannot produce a p-value below $1/(m+1)$, so the six $m = 10$ rows have a floor of 0.0909 and cannot reach $\alpha = 0.05$; they are superseded by the $m = 200$ rows (floor 0.00498) and listed because they exist in the run files. D2 is registered but outside the family because a null is the desired outcome; absence of a difference is not equivalence, and no equivalence test was prespecified.

E.1 Reading the Estimates

The registered discovery estimate (A1, fold 1.5355) and the dinucleotide-control site estimate (A2b, fold 1.4245) are distinct stored summaries. Positional excess compares each dinucleotide's site fold with its own sham-offset fold. The sequence-intervention estimate in B2 is the selected substitution statistic recorded in the claim file, not the mean of all substitutions. Replication applies to the enrichment estimate. Neither a nonsignificant comparison nor D2's small point estimate establishes equivalence.

F Memory Interventions

F.1 Offset Profile

The offset profile evaluates both modes at 58 width-7 windows in the 600-bp acceptor sequence, evaluated at the corrected span $[298, 305)$ to cover the AG-overlapping order-6 entry at position 304. The starts combine a coarse grid (21 to 567, step 14) with a local fine grid (283 to 321, step 2), with duplicate starts removed and the site start 298 included. For `src`, the site value is $+0.0027$, the profile lies within $[-0.0087, +0.0031]$, and the site ranks 57th of 58 by drop. For `out`, the site value is -0.0418, the profile lies within $[-0.1892, -0.0006]$, the site ranks 23rd of 58 by drop, and

Table 5: Acceptor-case estimates and recorded correction family. D2 is outside the family; the 24 other records include the six superseded low-resolution permutation tests below the rule. A2b is descriptive and has no reported test; the correction is computed over the other 23 family members. Available intervals are shown; dashes do not indicate zero uncertainty.

Claim	Criterion	Estimate	95% CI	q
A1 site fold	(a)	1.5355	[1.3767, 1.7041]	7.9×10^{-10}
A2 AG positional excess	(a)	$+0.3005$	$[+0.1968, +0.4041]$	1.41×10^{-7}
A2b AG fold at site	(a), descriptive	1.4245	—	—
B1 `src`, width 7, ΔAUROC	(b)	$+0.0027$	$[+0.0010, +0.0044]$	0.0114
B2 sequence subst., ΔP	(b)	-0.0064	$[-0.0094, -0.0035]$	0.0588
C slot mass, freq.-matched	(c)	-0.0315	—	0.1859
C slot mass, uniform	(c)	-0.0393	—	0.0954
C slot mass, low-frequency	(c)	-0.0472	—	0.0114
C slot per-occ., freq.-matched	(c)	-0.0037	—	0.2975
C slot per-occ., uniform	(c)	-0.0001	—	0.3051
C slot per-occ., low-freq.	(c)	-0.0128	—	0.0114
C sequence, AG→CC	(c)	1.3928	[1.3166, 1.4729]	2.06×10^{-18}
C sequence, AG→TT	(c)	1.0285	[0.9835, 1.0740]	0.2621
C sequence, base-composition shuffle	(c)	1.0099	[0.9502, 1.0766]	0.7555
C sequence, reverse complement	(c)	1.1280	[1.0697, 1.1904]	1.50×10^{-4}
C model, H3K4me3 checkpoint	(c)	1.1742	[1.1351, 1.2154]	5.53×10^{-16}
C model, donor checkpoint	(c)	1.0989	[1.0577, 1.1436]	4.45×10^{-5}
D1 replication fold	(d)	1.3427	[1.2993, 1.3885]	3.98×10^{-46}
D2 chromosome stratification	(d)	-0.0006	—	outside family
Superseded $m = 10$ permutation rows (floor 0.0909)				
C slot mass, freq.-matched	(c)	-0.0390	—	0.2323
C slot mass, uniform	(c)	-0.0352	—	0.1394
C slot mass, low-frequency	(c)	-0.0472	—	0.1394
C slot per-occ., freq.-matched	(c)	-0.0064	—	0.2323
C slot per-occ., uniform	(c)	$+0.0069$	—	0.4752
C slot per-occ., low-freq.	(c)	-0.0127	—	0.1394

the minimum is at window start 371 in sequence coordinates. The harness noise floor, measured by re-running the baseline, is $|\Delta P| \leq 0.0004$.

F.2 Span Correction

An initial configuration used span $[297, 304)$, which shifted one base to the left and missed the order-6 entry at position 304 overlapping the AG. With that span the profile had 57 windows and the site values were $+0.0017$ (`src`, rank 54th of 57) and -0.0458 (`out`, rank 22nd of 57). The profile above was recomputed with the corrected 7-bp span $[298, 305)$; the width sweep and the headline values in Table 2 use the same corrected span together with six offset controls centered at positions 120, 180, 240, 360, 420, and 480.

F.3 Width Sweep

Table 6 gives the site effect and its z-score against the six controls for both modes at five widths. For `src`, the standardized site contrast weakens with width and is near zero at widths 21 and above. For `out`, the site is not distinguished from the tested controls at any width. These comparisons do not establish equivalence between arbitrary windows. Zeroing the whole memory in `out` costs 0.410 AUROC, against 0.281 in `src`.

G Read-Head Value Interventions and Uniform-Metric Reevaluation

Checkpoint and metric preservation. The uniform-AUROC table reevaluates the original validation-selected checkpoint for each of the 16 retained tasks and both arms. No checkpoint

Table 6: Width sweep at the acceptor site for both intervention modes. z is the site effect relative to the mean and standard deviation of six offset controls.

	src			out		
Width	Site ΔAUROC	Control mean	z	Site ΔAUROC	Control mean	z
7	$+0.0027$	-0.0010	$+3.66$	-0.0418	-0.0582	$+0.24$
11	$+0.0014$	-0.0018	$+1.70$	-0.0910	-0.0993	$+0.08$
21	-0.0045	-0.0044	-0.03	-0.1876	-0.1693	-0.12
41	-0.0183	-0.0118	-0.54	-0.2586	-0.2300	-0.18
81	-0.0314	-0.0282	-0.11	-0.3061	-0.2733	-0.22

is selected again using AUROC. Binary AUROC uses the class-1 score; combined splice sites uses the macro one-versus-rest average over all three classes. Forward passes use the original BF16 evaluation precision, batch size 32, and masked mean pooling without added reverse-complement ensembling. The read-head arm improves over control on the same 16 tasks under the native metric and AUROC; the unweighted mean AUROC increment is 0.1020. This is variation across tasks for the original training seed, not a multi-seed estimate.

Operation and numerical verification. The additional acceptor interventions use FP32, batch size 64, and all 2,218 original test sequences. The baseline AUROC is 0.913037 (the BF16 performance-table estimate is 0.913059). We record the clean attention tensor independently at the two shared memory applications, replay both tensors, and mask projected values after their biases. Replay without a mask gives exactly identical logits on the entire test set. Masking a value changes neither the input bases nor the memory IDs used by the gated branch. The subsequent hidden state can change, so later gates may change indirectly. The two clean attention tensors remain fixed even in the single-application conditions. Whole-value removal retains the read's output bias; whole-read removal does not.

The exact site mask contains addresses 299–302 for order 3 and 298–304 for order 6. The covering union applies $[298, 305)$ to both orders. Offset masks follow the same geometry, with the motif translated to $[c-1, c+1)$ for $c \in \{120, 180, 240, 360, 420, 480\}$. Shared-source conditions replace IDs with the invalid-ID sentinel and recompute attention, matching the earlier source operation. All probability changes and class flips compare the same samples in the same order. Intervals use 2,000 paired sequence-bootstrap resamples (seed 1337); they do not represent variability across independently trained models. These are additional checks of the previously selected acceptor case, not additions to the original local lock.

Table 7: Fixed-attention read-head interventions. Effects are intervention minus intact. The probability column is mean absolute change in positive-class probability.

Operation	ΔAUROC	95% CI	Mean $\lvert\Delta P\rvert$
Identity replay	0.0000	$[0.0000, 0.0000]$	0.0000
Exact site, both applications	$+0.0026$	$[+0.0008, +0.0044]$	0.0225
Union site, both applications	$+0.0023$	$[+0.0006, +0.0039]$	0.0205
Exact site, application 0	$+0.0024$	$[+0.0014, +0.0034]$	0.0107
Exact site, application 2	$+0.0001$	$[-0.0013, +0.0016]$	0.0172
All values	-0.2899	$[-0.3141, -0.2663]$	0.3687
Whole read output	-0.3736	$[-0.3989, -0.3462]$	0.3880
Exact offset 120	$+0.0000$	$[-0.0005, +0.0006]$	0.0051
Exact offset 180	-0.0008	$[-0.0015, -0.0002]$	0.0072
Exact offset 240	-0.0028	$[-0.0038, -0.0017]$	0.0115
Exact offset 360	$+0.0013$	$[-0.0010, +0.0037]$	0.0317
Exact offset 420	-0.0052	$[-0.0075, -0.0029]$	0.0225
Exact offset 480	-0.0011	$[-0.0024, +0.0001]$	0.0118

For the exact site, 54/2,218 predictions change class; the mean absolute probability change is 0.022496. Thus a small positive AUROC effect is compatible with nonzero individual effects. Exact shared-source removal gives $+0.002700$ AUROC and covering shared-source removal gives $+0.002725$, reproducing the earlier direction and scale. The latter changes 30/2,218 predicted classes

and has mean absolute probability change 0.018004. All-source removal gives -0.280935 AUROC. These operations answer different questions; their magnitudes are not causal bounds on one another.

Artifacts. The runner is `scripts/protocol_f_revision_eval.py`; its batch artifacts retain labels, manifest row indices, sequence hashes, logits, probabilities, checkpoint hashes, and intervention specifications. New results reside under `results/protocol_f/iclr2027_revision_20260918/full_v1/`. Separate directories record BF16 task comparisons and FP32 acceptor interventions. Previously recorded results are retained.

G.1 Additional Training Seeds for the Primary Backbone

We additionally train read-head seeds 2 and 3 using the same frozen Mamba pretraining checkpoint, split seed 1337, 30-epoch recipe, and validation-F1 selection rule. Each run completes 16,830 optimizer steps. Only the classifier and added read head are trained. These runs test variability of the read-head intervention response; they do not provide additional paired controls for the 16-task performance table. Their BF16 test AUROCs are 0.9091 and 0.9039.

Table 8: Primary-backbone exact-site value intervention across training seeds. All intervention evaluations use the same 2,218 test sequences, FP32, and batch size 64. Intervals are paired-sequence bootstrap intervals within each model.

Seed	Intact AUROC	Exact-site ΔAUROC	95% CI	Class flips	All-value ΔAUROC
1	0.9130	$+0.0026$	$[+0.0008, +0.0044]$	54	-0.2899
2	0.9092	$+0.0038$	$[+0.0020, +0.0056]$	45	-0.3106
3	0.9039	$+0.0011$	$[-0.0007, +0.0028]$	50	-0.3220

The exact-site point estimates are positive for all three seeds; seed 3's interval includes both a small decrease and an increase. This is not evidence that every trained model has a positive population effect. The added runs use the same 29 conditions and identity-replay verification as the transfer evaluation.

On test-set true acceptors carrying AG ($n = 1,105$), seeds 2 and 3 have received-attention folds of 1.152 and 1.002, respectively, against the background specified in Appendix J; seed 3's fold is close to the background level. These benchmark-test measurements are distinct from the original model's discovery and replication on independent RefSeq sequences. Thus the direction check does not establish cross-seed replication of the full alignment-and-intervention chain. High AUROC alone does not ensure enrichment of this particular attention statistic at AG. Artifacts are under `training/splice_sites_acceptors/readhead/` and `transfer/mamba/splice_sites_acceptors/readhead/` in the revision result directory.

H Input and Annotation Controls

H.1 Sequence Intervention

Each of the 16 dinucleotides was written over positions $[300, 302)$ of every acceptor test sequence (site) and over positions $[450, 452)$ (sham, $+150$ bp), and the model was re-evaluated; the AG-to-AG condition is a no-op on the 1,105 true acceptors that carry AG and on the 424 decoys that carry it, which is the harness check. ΔP is the mean change in predicted acceptor probability over the true acceptors that carry AG; ΔAUROC is over the whole split. The prespecified threshold for an effective substitution on ΔP was $|\Delta P| > 0.02$; no substitution reached it, so the number of effective substitutions is zero and the mediation fraction that was to be computed over them is undefined, which is why no mediation value is quoted in this paper. Twelve of the sixteen substitutions raise ΔP; the mean over all sixteen is $+0.0059$ (Table 9).

H.2 Exploratory Poly(A) Spatial Analysis

The window-averaged analysis does not establish enrichment for the poly(A) hexamer AATAAA (fold 0.89, 11% below background), and length-matched controls place AATAAA third of ten, so criterion

Table 9: All 16 dinucleotide substitutions at the acceptor site and at the sham offset.

	Site $[300, 302)$		Sham $[450, 452)$	
Written dinucl.	ΔP	ΔAUROC	ΔP	ΔAUROC
AA	-0.00566	-0.00155	-0.00009	$+0.00024$
AC	$+0.00723$	$+0.00235$	$+0.00341$	$+0.00049$
AG	0.00000	$+0.00161$	$+0.00263$	$+0.00100$
AT	$+0.00386$	$+0.00078$	-0.00082	$+0.00041$
CA	$+0.00232$	$+0.00196$	-0.00134	-0.00074
CC	$+0.01305$	$+0.00554$	$+0.00043$	-0.00003
CG	$+0.00902$	-0.00213	$+0.00456$	-0.00161
CT	$+0.01598$	$+0.00695$	-0.00188	$+0.00022$
GA	$+0.00376$	$+0.00043$	$+0.00085$	-0.00093
GC	$+0.00535$	$+0.00214$	$+0.00033$	$+0.00054$
GG	-0.00255	-0.00054	$+0.00057$	$+0.00048$
GT	$+0.00915$	$+0.00219$	$+0.00489$	$+0.00069$
TA	-0.00645	-0.00315	$+0.00301$	$+0.00073$
TC	$+0.01776$	$+0.00526$	-0.00033	0.00000
TG	$+0.00781$	$+0.00197$	$+0.00145$	$+0.00063$
TT	$+0.01319$	$+0.00300$	-0.00123	$+0.00054$
Range	$[-0.0064, +0.0178]$	$[-0.0031, +0.0069]$	$[-0.0019, +0.0049]$	$[-0.0016, +0.0010]$

(a) fails and criterion (c) is not met. A spatially resolved analysis with a masked-reconstruction read head—the read head trained under the masked-reconstruction pre-training objective (Appendix B), rather than the fine-tuned classification head used elsewhere—, run after that result was seen, gives one suggestive observation: read strength at the query's own position exceeds all 46 control hexamers (layer-2 enrichment 4.10, rank-test $p = 0.021$). Because the offset was chosen post hoc from 33 candidates, the correct summary is the maximum-over-offsets version that absorbs the multiplicity, which gives $p = 0.064$. We therefore report the observation as a lead requiring prespecified independent replication and do not count it toward any criterion.

I Identity and Annotation Context

I.1 Coverage Across Annotation Families

The case library contains five families. Acceptor and donor are two branches of the splice-site family and use their corresponding checkpoints. The other families use the H3K4me3 checkpoint except for the separately labelled exploratory poly(A) follow-up. Donor GT has fold 0.93 (95% CI $[0.90, 0.96]$); CTCF has fold 1.00 at the trained 33-bp window ($p = 0.92$) and 0.99 at a 65-bp inference-time window. Poly(A) AATAAA has fold 0.89 (CI $[0.87, 0.90]$), ranking third of ten length-matched hexamers in the same 3′ UTRs, whose folds range from 0.85 to 1.33. These cases did not meet the enrichment prerequisite, and the full intervention chain was not run.

Table 10: Coverage of the five annotation families, one row per landmark; the two splice-site landmarks (acceptor AG and donor GT) form a single family. E: enrichment; D: prespecified intervention direction; S: specificity; R: enrichment replication. Y/N mean met/not met; a dash means not evaluated. Counts in S do not establish equivalence for other pairs.

Annotation	Checkpoint task	E	D	S	R	Observation
Acceptor AG	Acceptor	Y	N	N	Y	Enrichment replicates
Donor GT	Donor	N	—	—	—	Fold 0.93
Poly(A)	H3K4me3	N	—	N	—	Rank 3/10 hexamers
CTCF	H3K4me3	N	—	—	—	Fold 1.00
Codons	H3K4me3	mixed	—	0/4	—	Context contrast not detected
TF motifs	H3K4me3	mixed	—	1/14	—	Limited positive contrasts

I.2 Identity and Context Spreads

For codons (C-4), identity spread is the maximum minus minimum fold enrichment across the four annotated start and stop trinucleotides; function spread is the mean absolute difference between each annotated codon's fold and that of the same trinucleotide at a 3′-UTR position. For transcription-factor motifs (C-5), the underlying measurement is instead incoming attention mass at the motif minus its own background, not fold enrichment. Identity spread is the range of this measurement's motif means inside an ENCODE candidate regulatory-element class; function spread is the mean absolute difference between each motif's mean there and its mean in shadow intervals 10 kb away on the same chromosome that overlap no element. Ratios divide the two spreads within each case, class, and order; absolute spreads from C-4 and C-5 have different units. Nine JASPAR motifs were chosen by a rule fixed before any enrichment was computed (consensus length 6 to 10 bp, ranked by mean information content, CTCF excluded because C-3 covers it); a pair was analyzed only if both arms held at least 500 sites, which admitted 14 of 18 pairs and excluded the four pairs involving the longest motifs. Table 11 lists the spreads and their ratio by case, class, and order, and Figure 4 plots them. The main text quotes the pooled ratios, 8.5 for codons and 36.8 for distal enhancer motifs; the complete pooled and per-order rows range from 1.8 to 36.8, including promoter-like order-3 and order-6 ratios of 1.8 and 6.8.

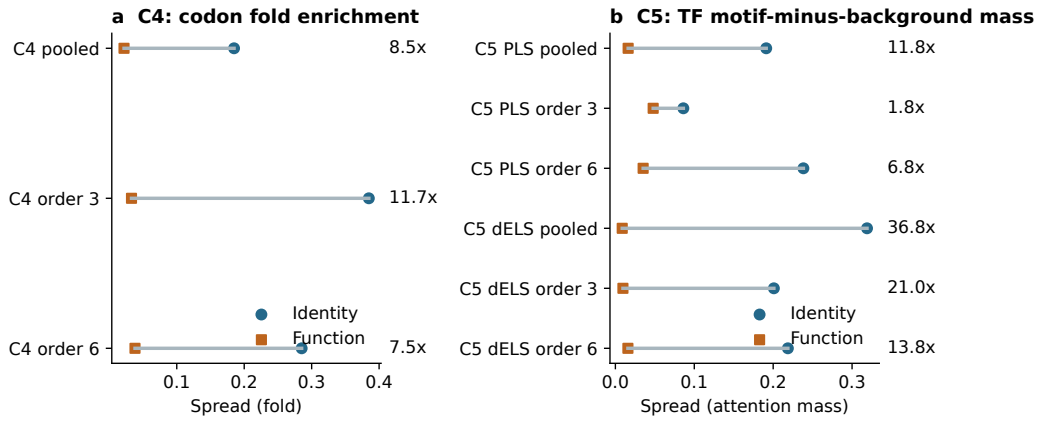


Figure 4: Identity spread against function spread for the codon (C-4) and transcription-factor (C-5) cases, by k-mer order and pooled. Ratios range from 1.8 to 36.8 across all pooled and per-order rows of Table 11. C-4 uses fold enrichment; C-5 uses motif-minus-background attention mass. These are descriptive spread ratios, not explained variance.

Table 11: Identity spread against function spread by case, class, and k-mer order. C-4 uses fold enrichment; C-5 uses motif-minus-background attention mass.

Case	Class, order	Identity spread	Function spread	Ratio
C-5	distal enhancer, pooled	0.31897	0.00866	36.8
C-5	distal enhancer, order 3	0.20090	0.00957	21.0
C-5	distal enhancer, order 6	0.21864	0.01588	13.8
C-5	promoter-like, pooled	0.19119	0.01617	11.8
C-5	promoter-like, order 3	0.08616	0.04790	1.8
C-5	promoter-like, order 6	0.23832	0.03518	6.8
C-4	pooled	0.18500	0.02180	8.5
C-4	order 3	0.38480	0.03290	11.7
C-4	order 6	0.28520	0.03810	7.5

These ratios quantify the specified spread summaries in C-4 and C-5. They are not fractions of explained variance or causal contributions. The codon comparisons do not detect a difference between contexts at the reported tests, but no equivalence test establishes that the contexts have identical effects.

J TRANSFER TO A CONVNOVA CONVOLUTIONAL BACKBONE

Purpose and scope. This additional experiment tests whether the addressable read head can be used with a non-Mamba backbone. We reuse our existing ConvNova gated-convolution implementation and its one-million-step pretraining checkpoint. This is not an official ConvNova checkpoint, an official-recipe reproduction, or a comparison against a claimed state of the art. Both arms already contain explicit memory; this experiment tests adding the read head, not adding memory to a memory-free model.

Architecture and training. The backbone has five gated-convolution blocks of width 128, kernel size 9, and dilations $(1, 1, 1, 4, 16)$. The 64-dimensional k-mer memory is applied before blocks 0 and 2. Both arms freeze the same 2,252,040 pretrained parameters and train 4,902 classifier parameters; the read arm additionally trains the same 16,672 head parameters as the primary model. We use the original 600-bp NT acceptor manifest, the same stratified 10% validation split (seed 1337), and training seeds 1, 2, and 3. Each run uses 30 epochs, batch size 32, learning rate 2×10^{-4}, 500 warmup steps, AdamW with weight decay 0.01, gradient clipping at 1, and BF16 training. The native task sampler and learning-rate rule are retained; there is no added reverse-complement augmentation or ensembling. Validation binary F1 selects each checkpoint, without selecting again on test AUROC. No backbone pretraining is repeated.

Table 12: Convolutional transfer on the acceptor task. Test evaluation uses BF16 and batch size 32 for both arms. SD is sample SD across three training seeds.

Seed	Control AUROC	Read AUROC	ΔAUROC	Control F1	Read F1
1	0.6478	**0.7894**	+0.1417	0.6947	**0.7543**
2	0.6528	**0.7859**	+0.1331	0.6945	**0.7461**
3	0.6484	**0.7862**	+0.1379	0.6914	**0.7442**
Mean	0.6497	**0.7872**	+0.1375	0.6935	**0.7482**
SD	0.0027	0.0020	0.0043	0.0018	0.0054

Matched evaluation. For each selected read checkpoint, the same 29-condition evaluation uses all 2,218 test sequences, FP32, and batch size 64. Conditions include exact AG-overlap values, the covering union, six offsets for each geometry, separate memory applications, all values, the whole read output, and shared-source operations. Clean attention is replayed at both applications. Attention enrichment uses Equation 2, positions $[300, 302)$, and background positions $[32, 568) \setminus [250, 352)$. Results are stored for all sequences and separately by class and AG presence. These benchmark-test profiles are an additional exploratory analysis, not a replication on the independent RefSeq samples used in the primary enrichment experiment. Within-model intervals use 2,000 paired sequence-bootstrap resamples; they do not substitute for variation across training seeds.

Attention and intervention outcomes. No-op replay reproduces all logits exactly for every seed. Among true acceptors carrying AG at the site ($n = 1,105$), attention folds are 1.049, 1.063, and 1.004 for seeds 1–3; the corresponding 95% intervals are $[1.033, 1.066]$, $[1.046, 1.080]$, and $[0.991, 1.018]$. For negative sequences carrying AG ($n = 424$), folds are 1.059, 1.030, and 1.017. These descriptive comparisons show that attention concentration at the AG varies across training runs on this backbone; consistent with our central point, alignment is a run-specific property to be measured rather than assumed.

Table 13: Convolutional read interventions on all 2,218 test sequences. Exact site uses four order-3 and seven order-6 values, with both clean attention tensors fixed. Intervals are within-seed paired-bootstrap intervals.

Seed	Exact-site ΔAUROC	95% CI	Class flips	All-value ΔAUROC
1	−0.0014	[−0.0027, +0.00005]	29/2,218	−0.1484
2	−0.0006	[−0.0021, +0.0010]	28/2,218	−0.1561
3	−0.0020	[−0.0033, −0.0007]	32/2,218	−0.1457

Exact-site value removal has a small negative effect for all seeds, with an interval below zero for seed 3. Removing all values has a larger negative effect in each case (Table 13), but this difference in scale does not demonstrate that local values are unused. Exact shared-source removal gives $+0.0015$, $+0.0006$, and -0.0003; its covering union gives -0.0009, -0.0017, and -0.0023. The difference reinforces the need to specify the intervention object and geometry. The read interface and predictive gain transfer to this backbone; the AG-specific alignment analysis remains a property of the primary Mamba case, where the full discovery–replication protocol was conducted.

Artifacts and recovery. `scripts/train_protocol_f_revision.py` trains the paired arms; `scripts/protocol_f_transfer_eval.py` evaluates their selected checkpoints. Outputs reside in the `training/conv/` and `transfer/conv/` subdirectories of `results/protocol_f/iclr2027_revision_20260918/`. Training snapshots preserve the optimizer, scheduler, random states, sampler state, and completed-step cursor. Evaluation retains immutable per-batch predictions and attention profiles. Resuming validates configurations and input hashes before continuing; existing experiment artifacts are retained.

K Interpretation, Limitations, and Future Work

Interpretation and limitations. The evidence connects a useful read head to replicated attention alignment and local intervention responses that oppose the specified performance decrease. The fixed-attention value test separates the read head from direct changes to the shared pathway, whereas source removal renormalizes attention and changes that pathway's inputs. Individual predictions do change: the finding concerns positive support for predictive performance, not an absence of causal influence. The explanation remains operation-specific. Redundancy and downstream compensation are possible accounts, not established mechanisms. The primary task table has one training seed; the three-seed convolutional check supports transfer on one task. Unequal-capacity arms do not establish capacity-controlled superiority. The complete intervention chain concerns one acceptor task; the other families and the local protocol lock do not extend that result to all DNA models. The practical implication is to state the attention quantity, intervened object, expected direction, and matched controls whenever alignment supports a positional explanation.

Future work. Several directions follow from these limitations: capacity-controlled comparisons and additional training seeds for the primary task table; completing the intervention chains for the annotation families that did not meet the enrichment prerequisite; and applying the same address-matched enrichment–intervention protocol to other backbones and to models whose memory sources are extracted rather than built in.

L Concepts Definition

To formalize this correspondence and clarify our experimental design, we define several core concepts used throughout this work:

- **Explicit k-mer memory:** A dedicated structural component that stores learned vector representations indexed exclusively by the local k-mer sequence at each genomic position.
- **Memory sources (or "read sources"):** The individual, position-indexed vectors retrieved from this memory. Because each source maps to a specific genomic coordinate, they serve as the precise computational objects we target for intervention.
- **Local read head:** An attention mechanism that queries nearby memory sources within a restricted local window of exactly 33 bp ($|s - t| \leq 16$) to gather information.
- **PROTOCOL-F:** Our pre-specified evaluation framework. It tests positional explanations by requiring attention enrichment, a matched intervention direction, specificity against controls, and replication on independent samples.
- **The identity–dependence gap:** The empirical gap we observe where strong, sequence-specific attention enrichment at a biological signal ("looking") does not align with a drop in performance when those exact sources are silenced ("depending").

- **Local sequence identity:** The literal, character-level nucleotide sequence or k-mer composition at a given position (e.g., the specific dinucleotide `AG`), irrespective of its broader genomic location or biological function.
- **Annotation context:** The functional biological status assigned to a position by genomic annotations (e.g., residing at an active splice acceptor site versus an unannotated background region), holding the local sequence identity constant.
- **Identity-to-context spread ratio:** A descriptive summary statistic, not an evaluation metric: the ratio of the spread of a case-specific attention measurement across sequence identities to its spread across annotation contexts (Appendix I.2). These ratios are not fractions of explained variance.
- **Slot and address:** A slot indexes the canonical sequence identity of a k-mer; an address specifies one occurrence at a genomic position. The same slot can therefore appear at multiple addresses: a slot intervention affects every occurrence sharing that identity, whereas an address intervention is local to the selected occurrences.
- **Received attention $A(p)$ and enrichment:** $A(p)$ sums the attention mass that the source at position p receives from all queries within reach, averaged over the two memory applications. Enrichment is the mean $A(p)$ at annotated positions divided by the mean background $A(p)$ within the same sequences.
- **Source removal (`src`) and output masking (`out`):** The two intervention operations. `src` invalidates source identifiers at selected addresses, so the remaining attention renormalizes and the shared gated pathway also receives the modified identifiers; `out` instead masks the combined memory residual written at a position. `src` changes what other positions can read *from* a location; `out` changes what the memory writes *at* it.