ATTRIBUTION EVALUATION NEEDS AN ABLATION, NOT AN ANNOTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature attributions are evaluated by comparing what they highlight against an annotation. We show that such comparisons do not measure what they are read as measuring.

On sequences where ablation demonstrates that a model uses neither the annotated feature nor the region the attribution itself identifies, attributions still locate the annotation at 5–14$\times$ the empirically measured chance level. The effect holds across 15 transcription factors, two architectures and three attribution methods, disappears when the annotated interval is compositionally rewritten, and a variance decomposition puts $\sim$40% of the score's variance on sequence identity against under 2% on the choice of method.

We then measure the contaminating term rather than infer it. Scoring a model against the annotation of a factor it was never trained on sets model usage to zero by construction, and what survives is measurable: $0.22\times$ chance without training, $1.69\times$ after training on the benchmark, $5.45\times$ for a byte-pair architecture. Carrying the decomposition through the argmax the metric performs identifies both terms from measured recovery alone, and the usage term is stable at 3.22 ± 0.26 across factors whose reported scores range from 0.31 to 0.58. The architecture that appears to produce better attributions confers more recovery on annotations it never learned, and shows less usage.

The protocol reproduces the effect on protein binding sites, text rationales and natural textures at 1.5–$2.2\times$, $2.25\times$ and $11.2\times$, an ordering the decomposition predicts — the protein range spanning two negative-set designs, both far below the genomic effect; images are not exempt where a composition-preserving manipulation exists. Ten corrections to the score fail, and the identification says why. A rank-based alternative improves run-to-run reproducibility on every factor in all three sequence domains and costs less than the attribution it evaluates. We derive the runs a comparison needs — four paired, fourteen unpaired — and the second kind is what this literature reports from single runs.

1 INTRODUCTION

A feature attribution is judged by comparing what it highlights against an annotation. The pointing game and IoU do this in vision, rationale F1 in NLP, motif recovery in genomics: in each case a score is computed from how far the attribution's highest-weighted region overlaps a region a human or an experiment marked, and the score is read as a measure of the attribution. Such a comparison conflates three questions. **(a)** Is the feature findable in the input — does it stand out from its background? **(b)** Does the model use it? **(c)** Does the attribution point at it? A recovery score is read as an answer to (c), but it answers (c) only if (a) and (b) are held fixed, and in practice neither is.

Why this is measured on sequences. Separating the three requires two things the domain has to supply: a chance level that can be written down rather than guessed, and a substitution that destroys the annotated feature while leaving the input where the model expects it. Genomic sequences supply both — a motif occupies 2–4% of a window at a known position and can be replaced by a composition-matched surrogate that is still DNA. Object photographs supply neither. We therefore

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.

build the measurement where it can be built, then carry it to proteins, to text, and — on textures, where a pixel shuffle preserves the intensity distribution exactly — to images. The claim is about a class of metric, not about a domain.

This distinction is invisible to the perturbation tests that currently serve as faithfulness checks. Deletion and insertion curves rank positions *by the attribution itself*, so they presume the ranking is meaningful in order to evaluate it. Ablating the annotated feature instead supplies a model-side reference point that no attribution enters.

Contributions.

- A decomposition of ground-truth attribution evaluation into findability, model usage and attribution accuracy, with four falsifiable predictions. Carrying it through the argmax identifies both terms from measured recovery alone (Proposition 1) and shows why every correction to the absolute value must fail (Corollary 1) (Section 3).
- A direct measurement of the contaminating term: scoring a model against a factor it was never trained on sets model usage to zero by construction. The floor is $0.22\times$ chance untrained, $1.69\times$ after training on the benchmark and $5.45\times$ for a byte-pair architecture; it is conferred by training and tracks the architecture (Section 5.5).
- Six independent controls, including a negative-sequence baseline validating the closed-form chance level and a symmetric ablation foreclosing the incomplete-annotation objection (Section 5).
- Replication on proteins, text and natural images, the ratio differing by domain as the decomposition predicts. On textures the full protocol returns the genomic result — $11.2\times$ chance on examples the ablation identifies as unused — locating the difficulty in vision in the manipulation rather than the phenomenon (Section 5.6).
- Ten corrections that fail and one that does not, at 144-fold less cost than the attribution itself (Section 7).

2 Related Work

The setting, for readers outside it. A *transcription factor* binds DNA at a *motif* of 8–20 bases, written as a position-probability matrix over $\{A, C, G, T\}$ since binding tolerates substitutions. The task is binary: positives contain a validated site, negatives are composition-matched windows without one. This is a sparse-annotation detection task of the kind pointing games and IoU evaluate in vision, with the annotation localised exactly and the background characterised well enough to substitute for it.

Auditing ground-truth evaluations. Zhou et al. (2026) survey 3,575 papers applying interpretable machine learning in genomics and audit five attribution methods, reporting that methods disagree and that attributed regions frequently miss validated motifs. Their faithfulness test is a sequential deletion curve, which *presumes what it sets out to establish*: a ranking tracking nothing about the model still produces a steep curve if the positions it ranks highly are ones the model is sensitive to for unrelated reasons, and we show this is what occurs. They report the score raw, so a value near zero cannot be told from one below chance — on SP1 and TBP the published values are 0.1–0.2$\times$ chance — and each model is trained once, so the variation we measure across nine retrainings is invisible. We use their benchmark and their published `perception_score`, which agrees with ours to five decimal places.

Three other lines. Motif recovery as a check on what a sequence model learned goes back to Koo & Eddy (2019), and the benchmarks that formalise it — Patel et al. (2024) among them — report it as a headline number, beside perturbation tests of the kind Chen et al. (2024) propose. Zhang et al. (2023) make the observation that motivates this work — a network trained on designed features need not use all and only those features — and address it by hand-setting weights, removing the confound at the cost of the model. Reynolds & Pan (2025) benchmark recall and stability on UK Biobank genotypes against decoys that assume the absence of signal rather than verifying it; theirs is the closest prior treatment of the reproducibility axis of Section 5.1. Adebayo et al. (2018) and

Kindermans et al. (2019) established that attributions can be insensitive to model and data randomisation and Duan et al. (2024) that method rankings reverse across protocols; our concern is what ground-truth *evaluations* score, and we show rankings reverse across datasets under a fixed protocol. In NLP the observational overlap we measure has been superseded by causal tests (Atanasova et al., 2023; Siegel et al., 2024), and our ablation plays the same role on the model side. Our attributions use a fixed integration baseline; methods averaging over a background (Erion et al., 2021) may carry a different α, which we do not measure.

Positioning. Four conditions define this work: real trained models; model usage measured independently of any attribution; chance level verified empirically; and run-to-run reproducibility. No published work satisfies all four (Appendix B, Table 2). Zhou et al. satisfy the first alone; Zhang et al. (2023) fix usage by construction rather than measuring it; Reynolds & Pan add reproducibility but estimate precision against decoys that assume the absence of signal. Concurrent work (Anonymous, 2026) derives the chance level of Equation 1 and shows raw recovery is not comparable across factors; it corrects the *origin* of the score but does not measure usage, and we show the correction is not sufficient — a model that learned nothing about the annotation still recovers it at $1.69\times$ the corrected chance level.

3 What a Recovery Score Measures

3.1 Setup

For a sequence of length N, an annotated interval M of length L, and an attribution $a \in \mathbb{R}^N$,

$$W^*(a) = \arg\max_w \sum_{i=w}^{w+L-1} |a_i|, \qquad \text{rec} = |W^*(a) \cap M|/L. \tag{1}$$

If W^* were uniform the expected recovery would be $\pi_0 = L/(N - L + 1)$, and we report $\text{APS} = (\text{rec} - \pi_0)/(1 - \pi_0)$.[1]

3.2 A decomposition

Assumption 1 (Additive decomposition)**.**

$$|a_i| \;=\; \beta\, u_i \;+\; \alpha\, s_i \;+\; \varepsilon_i, \tag{2}$$

where $u_i \geq 0$ is the extent to which the model uses position i, $s_i \geq 0$ the extent to which position i attracts attribution for reasons other than this model having learned this annotation, ε_i exchangeable noise, and α, β scalars set by the method and the model.

What s is not. It is not a property of the input alone: Section 5.5 shows it grows with training, from $0.22\times$ chance untrained to $1.69\times$ after training on a different factor. The decomposition requires only that s_i be independent of *this* annotation having been learned — a condition the cross-factor construction imposes by design.

This is a decomposition assumption, not a derivation. It asserts additivity, not independence: u and s may covary, and Section 6 measures how much they do and what it costs. It is justified only by the predictions it makes.

From the decomposition to a measurable quantity. Equation 2 is a statement about positions; the metric returns an argmax over windows. Carrying one to the other identifies the two terms, and this is what makes the assumption testable rather than decorative. Write $S(w) = \sum_{i \in w} |a_i|$, $U = \sum_{i \in M} u_i$ and $\Sigma_w = \sum_{i \in w} s_i$; with u supported on M, the annotated window scores $\beta U + \alpha\Sigma_M + E_M$ and any disjoint window $\alpha\Sigma_w + E_w$.

[1] The closed form and the corrected score follow Anonymous (2026), which establishes the origin; the present work shows it is not sufficient.

Proposition 1 (Identification)**.** *If the window noise E_w is i.i.d. Gumbel with scale σ, then the probability that the top window is the annotated one is*

$$q \;=\; \Pr[W^* = M] \;=\; \mathrm{logistic}\left(\frac{\beta U + \alpha\Delta}{\sigma}\right), \qquad \Delta \;=\; \Sigma_M - \tfrac{\sigma}{\alpha}\log\textstyle\sum_{w\neq M} e^{\alpha\Sigma_w/\sigma}, \quad (3)$$

where $\mathrm{logistic}(x) = (1+e^{-x})^{-1}$ *and σ throughout is the Gumbel scale. Since q is estimated by* APS*, both terms are recoverable from measured recovery alone:*

$$\frac{\alpha\Delta}{\sigma} = \mathrm{logit}(q_{\mathrm{floor}}), \qquad \frac{\beta U}{\sigma} = \mathrm{logit}(q_{\mathrm{diag}}) - \mathrm{logit}(q_{\mathrm{floor}}), \quad (4)$$

where q_{floor} is recovery on a factor the model was not trained on (so $U = 0$ by construction) and q_{diag} on the factor it was.

The Gumbel assumption turns an argmax into a logit; it is standard in discrete choice and we do not test it directly.[2] What we test is whether the quantities it identifies behave as quantities should. On HyenaDNA $\beta U/\sigma = 3.22 \pm 0.26$ across three factors whose reported APS spans 0.309 to 0.584 — **the spread in the score is almost entirely** $\alpha\Delta$. SP1 and TBP have negative APS, so Equation 4 is undefined: the identification excludes them rather than a filter we chose. And on CTCF, DNABERT-2 scores higher than HyenaDNA with under half the usage term, because its background windows sit closer to the annotated one, so less usage suffices to win the argmax. **Higher recovery can mean less model usage**, and Equation 3 says when (Appendix C).

Corollary 1 (Why corrections to the absolute value fail)**.** *q is a logistic function of the margin, so a difference of recoveries $q_{\mathrm{diag}} - q_{\mathrm{floor}}$ is not proportional to βU except near $q = 1/2$. Any correction that shifts or rescales the score without passing through the logit inherits the nonlinearity, which is why the ten corrections of Section 7 all fail and why the rank-based alternative, which is invariant to monotone reparameterisation, does not. Equation 5 is itself such a correction and is accordingly not offered as an estimator of βU, only as a common origin across datasets.*

Operationalisation. u is measured by ablation: replacing M with a composition-matched surrogate and recording the prediction drop, with a control-position replacement subtracted so that the effect of substitution per se cancels. s is not directly observed, but a manipulation that rewrites the content of M while preserving its position, length and composition destroys it.

Predictions. Figure 1d states what Equation 2 implies in four conditions; Figure 1c shows why the two terms are hard to separate even in principle — both peak on the annotation. The central consequence follows immediately: recovery does not separate βu from αs, so where αs is large, high recovery is observed *whether or not the model uses the feature*. A corollary is that the ratio α/β should differ across datasets, which Section 5.6 tests.

4 METHODS

Equation 2 has two terms and the protocol estimates each with its own intervention: Section 4.2 estimates u by removing the feature, Section 4.3 estimates s by removing u instead. Section 4.1 fixes the origin both are read against, and Section 4.4 forecloses the objection that a low u means the annotation was wrong rather than unused.

4.1 CHANCE LEVEL, VERIFIED EMPIRICALLY

Equation 2 assumes W^* is uniform in the absence of signal. Negative sequences carry no annotated feature, so this holds by construction; assigning each a random interval of length L and applying

[2] Three caveats. Adjacent windows overlap, so the noise terms are not strictly independent and the logistic form is an approximation rather than an identity; and q is an exact-match probability while APS is a normalised expected overlap, so APS over-estimates q and the identified $\beta U/\sigma$ is a lower bound. For these two the ordering is unaffected. The third is the assumption under test: Equation 4 subtracts two logits, which isolates $\beta U/\sigma$ only if $\alpha\Delta/\sigma$ is the same at the diagonal and the floor, and since Δ depends on s and s is model-dependent, that is an assumption rather than an identity — one that, if it failed, could reorder the identified terms. Its support is indirect and has a scope: $\beta U/\sigma$ is stable at 3.22 ± 0.26 on HyenaDNA across factors whose raw scores span 0.31 to 0.58, which would not follow if Δ varied freely, but on DNABERT-2 it is 1.50 ± 1.40.

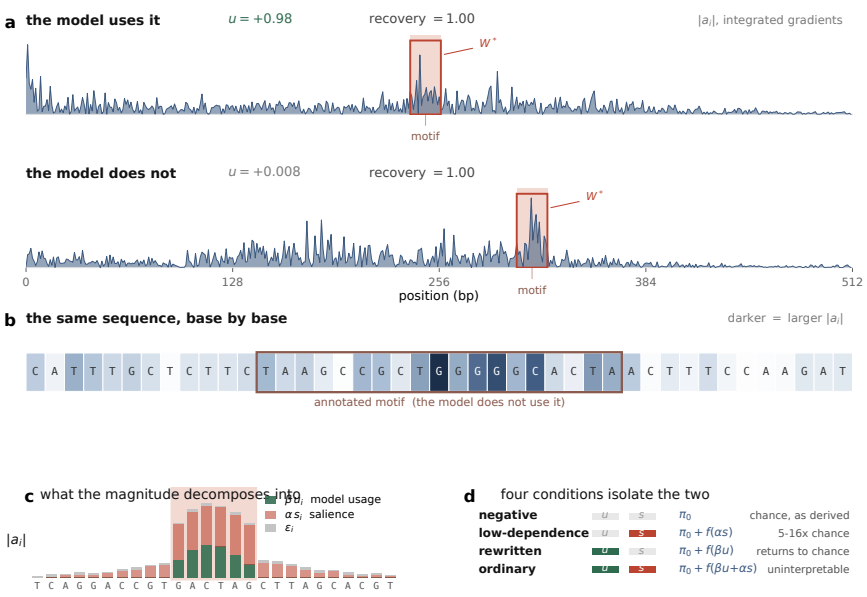

Figure 1: **(a)** Two held-out CTCF sequences, both at recovery 1.00; ablation puts usage at $u = +0.98$ and $u = +0.008$. **(b)** The second base by base — the attribution concentrates on a motif the model does not use. **(c)** The decomposition under Equation 2: both terms peak on the annotation, which is why the score cannot separate them and why their difference (Section 6) does not either. **(d)** Four conditions removing one or both terms, with the measured outcome.

the identical procedure gives $\mathrm{APS}_{\mathrm{null}} \in [-0.015, +0.009]$ across four factors, seed-to-seed standard deviation 0.003–0.009. The closed form also assumes a uniform admissible start, exact only away from the boundaries: enumerating admissible windows per example gives a closed-form-to-exact ratio of 1.001–1.008 across five factors, with an aggregate shift below 1% (Appendix I).

4.2 Estimating model usage

Each positive sequence yields three conditions: the original; the *motif* condition, in which M is replaced by a surrogate drawn from the mononucleotide frequency outside M; and a *control* condition replacing an equal-length interval disjoint from M. The usage estimate is $u(M) = P_{\mathrm{control}} - P_{\mathrm{motif}}$, which cancels the effect of substitution itself. Intervals are computed by paired bootstrap.

Three properties of this estimator — that the substitution removes the feature, that the $u < 0.05$ threshold is not arbitrary, and that a single measurement is reliable enough to threshold — are verified in Appendix I.

4.3 Estimating salience

Ablation gives u; it does not give s. The obvious route — subtract the rewritten condition from the intact one and read what remains as αs — does not work, for reasons Section 6 sets out. We use a different construction, which removes βu rather than subtracting it.

Score a model trained on factor A against the annotation of a factor $B \neq A$. The model has never been trained to recognise B's motif, so $\beta u = 0$ by construction and whatever recovery remains is αs. We compute this for every ordered pair over the five benchmark factors, holding the attribution method, the window size and the scoring pipeline fixed. The diagonal $A = B$ gives $\beta u + \alpha s$; the off-diagonal gives αs alone, and we call it the *floor* of the dataset. Their difference estimates βu.

The construction has one precondition. Where two factors share a binding motif it fails — the model has seen that motif, under another name. We screen pairs by the maximum column-wise

correlation between their JASPAR position-probability matrices over all offsets and both strands, and exclude any pair above 0.60. This removes USF1–MAX (0.712), USF1–JUN (0.725) and USF1–FLI1 (0.655), all E-box factors; no benchmark pair is excluded. Given a per-factor floor f_B we report the corrected score

$$\mathrm{CAS}(B) \;=\; \frac{r_B - f_B\,\pi_0(B)}{1 - f_B\,\pi_0(B)}, \tag{5}$$

which is zero when recovery equals the floor and one at perfect recovery.

4.4 Symmetric ablation

The judgement "the model does not use the annotated feature" leaves open that it uses something else that the attribution correctly identifies. We therefore ablate the attribution's own top window under the same control, giving a two-by-two partition whose cells Appendix I names and sizes. The cell in which the model uses a region the attribution found — *sufficient but not necessary* — varies from 0.09 to 0.55 across factors and seeds, so no single value characterises it; we therefore restrict all low-dependence baselines to sequences where *neither* intervention changes the prediction. On DNABERT-2 the same fraction is 0.17–0.37, matching HyenaDNA's range.

4.5 Datasets, models and procedures

Two genomic settings. The *benchmark* is the five factors of the source study (CTCF, GATA1, MAX, SP1, TBP), chromosome 9 held out, $n = 65$–$1{,}599$; its window construction cannot be reproduced without the peak files. The *purpose-built* set is twelve factors from UniBind (Gheorghe et al., 2019) hg38, four from each tertile of compositional separability, with the motif placed uniformly — making π_0 exact — and GC-matched negatives. Proteins (UniProt `BINDING`) and text (ERASER Movies, DeYoung et al., 2020) are described in Appendix F.

HyenaDNA (Nguyen et al., 2023) and DNABERT-2 (Zhou et al., 2024) fine-tuned per factor under up to nine seeds; integrated gradients (Sundararajan et al., 2017) unless stated, with gradient×activation and DeepLIFT in the variance decomposition. Throughout we distinguish *between-run* standard deviation from *within-run* standard error: the former is invariant to evaluation sample size while the latter falls as $1/\sqrt{n}$, so the latter must not be reported alone. Appendices B and J give the rest.

5 Results

5.1 The model is stable; the metric is not

Across nine retrainings differing only in seed, held-out classification AUC has a standard deviation of 0.0010 (0.9592–0.9623) and the ablation-measured motif dependence 0.0093, intervals overlapping across all nine. The recovery score, on the same sequences from the same models, has a standard deviation of 0.062 (0.522–0.694) — **sixty times larger**. Two measurements agree the models are equivalent; the third does not. This does not shrink with more evaluation sequences: from $n = 1599$ down to 100 it stays at 0.060–0.066 while the within-run standard error rises from 0.009 to 0.039, and across 17 factors the effective sample size is 19–175, or 0.9–6% of the nominal.

5.2 Recovery is dominated by sequence identity

Decomposing the variance of per-sequence recovery over three attribution methods, up to nine training runs and 348–1,599 sequences puts 38–44% on sequence identity, under 4% on the choice of method and under 2% on the run (Table 4) — roughly twenty times the method term and twenty-five times the run term. **A metric whose variance is 40% sequence and 2% method cannot be used to compare methods.** The large residual is consistent with the ranking reversals of Section 5.4.

5.3 How many runs a comparison needs

The same decomposition determines how a comparison should be powered. For a mean over S runs and n sequences, $\mathrm{SE}(\bar{r})^2 = \sigma^2_{\mathrm{run}}/S + \sigma^2_{\mathrm{seq}}/n + \sigma^2_{\varepsilon}/(Sn)$: the first term is unaffected by evaluation

set size and the second by repetition, since repeated runs share the held-out sequences. *The bootstrap intervals conventionally reported estimate the middle term alone.* Two methods compared on the same trained models share β_s, which cancels, leaving the interaction at $\sigma_{\alpha\beta} = 0.024$: detecting $\Delta = 0.05$ needs $S \geq 4$. Comparisons across datasets, architectures or training conditions are not paired, $\sigma_{\text{run}} = 0.062$ applies, and detecting $\Delta = 0.10$ at 80% power needs seven runs per condition — fourteen in total (Appendix D). **The comparisons in this literature are overwhelmingly of the second kind, and they are reported from single runs.** The requirement bites where the effect claimed is of the same order as the variation; between-factor differences, which span 0.47 to 14.14, are not threatened by it.

5.4 Six controls

Six controls establish that the model-independent component exists and is large; Appendix E gives each in full. **Negative sequences** return $\text{APS}_{\text{null}} \in [-0.015, +0.009]$. **Low-dependence positives** — where neither the annotated motif nor the attributed window affects the prediction — stand at 5–14× chance in ten of fifteen factors, and restricting to the doubly-ablated subset *raises* the baseline (CTCF 3.1→7.1, GATA1 5.6→13.4, MAX 5.0→9.3). **Rewriting removes the effect**, collapsing APS to near chance (CTCF $0.555 \rightarrow 0.080$, GATA1 $0.491 \rightarrow 0.045$) and to $[-0.013, +0.047]$ across all twelve purpose-built factors. **Uniform placement** does not remove it (1.5–9.7× in ten of twelve), and **three methods** give thirty of thirty-six factor×method combinations above 2×, rankings reversing across factors.

Two architectures: the better model is worse. DNABERT-2's ratios exceed HyenaDNA's in all five shared factors (CTCF 11.2 vs 3.1; GATA1 19.5 vs 5.6; MAX 22.0 vs 5.0), with a low-dependence baseline of 0.43 against a whole-set APS of 0.73 — **more than half the score is independent of the model** — while its collapse rate is an order of magnitude lower. The architecture whose attributions are better behaved carries the larger model-independent component. These controls establish that it exists; the next section measures it.

5.5 The floor, and what confers it

Applying the construction of Section 4.3 to the benchmark gives the floor directly, and shows it is a property of the training rather than of the data. Table 1 gives the five layers.

Table 1: Recovery as a multiple of π_0 under decreasing amounts of relevant training. The floor is the mean over off-diagonal pairs after the PWM screen, over five HyenaDNA runs and three DNABERT-2 runs.

	what the model learned	recovery	n
random init	nothing	0.22	5
own 12-TF	a different factor, uniform placement	0.44	60
benchmark	a different factor, centre-biased	1.69 ± 0.38	5 runs
DNABERT-2	a different factor, BPE tokenizer	5.45 ± 0.23	3 runs
diagonal	the factor being scored	15.05	5 runs
diagonal, DNABERT-2	the factor being scored	16.53	3 runs

An untrained model recovers the annotation at $0.22\times$ chance — below chance. **The floor is therefore not a property of the input, the annotation or the attribution method. It is conferred by training**, and the amount depends on what was trained on. **Dataset design changes it**: models trained with the annotation placed uniformly recover their own at $12.72\times$ against a floor of $0.44\times$, a ratio of 29, where on the benchmark the same ratio is 9 (Figure 2). **The positional regularity of a dataset, not the model, supplies three quarters of the contamination.**

The floor tracks the architecture, not the class. A byte-pair tokeniser gives a floor of 5.45 ± 0.23 against 1.69 ± 0.38 for a single-token one, over three and five runs, the intervals not overlapping; one such model recovers a *different* class's annotation at $11.99\times$ while recovering its own at $2.84\times$. **The architecture that appears to produce better attributions confers more recovery on annotations it never learned.**

Correcting for it, and what it costs. Equation 5 brackets the score between π_0 and $f_B\pi_0$, which no model can fall below having learned anything at all. On CTCF the raw 24% advantage of DNABERT-2 is eliminated — both land at 0.57 read against their own floors — while on MAX it survives; on SP1 and TBP it returns a negative value for HyenaDNA, which recovers its own annotation less often than models that never saw it (Table 3). This differs from the model-randomisation check of Adebayo et al. (2018), which reports $0.22\times$ where the cross-factor measurement reports $1.69\times$. The floor is $k(k-1)$ attribution passes computed once — 20 on the benchmark against the $9 \times 1599 \times 3$ behind the scores it corrects.

5.6 Proteins, text, and images

The protocol requires sparse ground truth, demonstrable model usage, and a manipulation that destroys the pattern while preserving position and composition. Three further domains admit it, the third being images. On proteins, ablating annotated residues drops the predicted binding probability by 0.369–0.558; recovery reaches $7.4\times$ chance overall but only $1.5\times$ on low-dependence sequences, and rebuilding the task with negatives from other proteins raises that to $2.2\times$ (Appendix F). On text rationales a fine-tuned sentiment classifier gives $2.25\times$.

Images admit the protocol where the manipulation exists. Whether vision admits the protocol is a property of the visual content. In an object photograph the arrangement of pixels is what makes the object, so shuffling destroys s and u together; in a texture the arrangement is one sample from a stationary process, and a pixel shuffle preserves the intensity distribution exactly while destroying it — the image remains a plausible texture, as a shuffled genomic interval remains DNA. We run the full protocol on DTD (Cimpoi et al., 2014): a 32×32 patch of one texture on a natural-texture background (6.3% of the image), a second texture in every image, and the task of detecting the first.

	genomics	DTD textures
untrained model	0.22	2.62 ± 1.41
trained on another pattern (floor)	1.69 ± 0.38	6.49 ± 0.13
trained on the pattern scored	15.05	11.67 ± 0.31
on low-dependence examples	**5–14**	$\mathbf{11.21 \pm 0.86}$
after the composition-preserving shuffle	≈ 1	2.34 ± 0.45

Figures are over the four runs of ten that learned the task (accuracy ≥ 0.94); Appendix G reports all ten. The pattern is the genomic one throughout: on the 269 examples with $u < 0.05$ recovery still stands at $11.21\times$ chance — the decisive condition of Section 5.4 reproduced on images. What differs is the proportion: the floor is 56% of the diagonal here against 11% on the benchmark. The attribution method matters here as it does not in genomics — with integrated gradients the *untrained* recovery exceeds the trained floor, because a continuous-valued image is not equidistant from the integration baseline at every position as a one-hot sequence is; Grad-CAM takes no baseline and the effect disappears, a plausible reason the floor has not surfaced in the vision literature.

What vision does not admit. Object recognition stays outside the protocol because the annotated object is not sparse (27% of the image against 6.3% here) and no substitution preserves its composition while destroying it — conditions on the manipulation, not the phenomenon.[3] The four domains order as Equation 2 predicts, since the ratio measures $\alpha s/\beta u$ and that is a property of the data: genomic motifs and textures are short and distinctive, protein binding residues scattered and compositionally ordinary.

6 A Prediction That Fails

Before the cross-factor construction of Section 5.5 we tried the direct route: αs should not survive the rewrite, so intact minus rewritten should track ablation more closely than intact alone. It does not — correlations with the ablation drop are $+0.38$–$+0.59$ for intact, $+0.18$–$+0.35$ for rewritten, $+0.27$–$+0.33$ for the difference.

[3] A different class of manipulation — generative inpainting, which replaces the object's region with synthesized content rather than rearranging the original pixels — lies outside the scope of this paper.

The first of two explanations is measurable. Correlating the ablation-measured u with per-sequence recovery gives $+0.56$ (CTCF, $n = 400$), $+0.48$ (GATA1, $n = 348$) and $+0.32$ (MAX, $n = 400$), Spearman $+0.49$, $+0.48$, $+0.34$: sequences whose motif the model uses are also sequences whose motif is salient, as one would expect of real binding sites. Equation 2 treats the terms as additive, not independent, and does not require independence — but a *difference* between two positively correlated quantities cancels signal along with the term it is meant to remove. A second explanation is available: the rewrite preserves composition, so a model responding to k-mer content rather than arrangement would be scored as unaffected. We cannot separate the two with the manipulations we have, and report the discrepancy rather than adjust the model.

A natural explanation for αs is the integration baseline, but replacing `[PAD]` with `N` changes recovery by $+0.013$ to $+0.064$ — the wrong direction, and below the seed-to-seed variation. **The salience term is not a baseline artefact.** Correcting position-wise makes matters worse: $|a|-|a_{\text{rewritten}}|$ raises the low-dependence baseline from $4.1\times$ to $17.8\times$.

7 What to Report Instead

7.1 Ten corrections that fail

We evaluated ten: window-jitter filtering, a distance-based score, excluding collapsed cases, stratifying by dependence, switching method, top-k windows, boundary filters, a sample-size adjustment, and the score- and position-level differences of Section 6. Every one adjusts the *absolute value* and four make at least one factor worse — what Equation 2 predicts and Corollary 1 explains (Appendix H).

7.2 A rank-based alternative

If contamination shifts scores without reordering sequences, ranks survive it. Taking the ablation drop as a label and per-sequence recovery as a score, we report $\mathrm{AUC}(\mathbf{1}[u > \tau], \mathrm{rec})$. It needs both classes populated at the threshold, which excludes six of seventeen factors: SP1 and TBP, whose median drop is 0.002–0.003, and WT1, ZNF143, PDX1 and TFAP2C, which leave fewer than twenty sequences above τ. All eleven for which it is computable improve (Figure 3; 1.6–10$\times$, AUC 0.61–0.76), the score is stable for $\tau \in [0.2, 0.8]$, and the ranking reversals of Section 5.4 disappear. Text improves $1.8\times$ and proteins $1.9\times$. It does not transfer to the texture task (AUC 0.54 ± 0.05), where the ablation drop is bimodal — reported by the diagnostic of Section 7.4.

The AUC value is itself a diagnostic. Its magnitude reports how far the attribution tracks model usage — proteins 0.872, genomics 0.61–0.76 — inverting the low-dependence ratios of Section 5.6 (proteins 1.5, genomics 5–14). **Where the ratio is high the AUC is low.**

7.3 The ablation is cheap, and a screen is cheaper

At $n = 100$ the AUC holds to within ± 0.03 of its full-sample value and one trained model suffices (one-seed ablation differs from the nine-seed value by 0.005–0.026): **the reference point costs less than the attribution it evaluates** (Appendix J). Where even that is too much, the attribution collapse rate — computable from the attributions alone — predicts the low-dependence ratio across fifteen factors ($r = -0.771$, $p = 0.0008$), and **the sign runs counter-intuitively**: those that collapse *least* carry the *largest* model-independent component.

7.4 The floor, the pairing, and where it does not apply

Report the floor for the dataset in use. π_0 is what an attribution attains with no information, not what a *model* attains with no relevant information. Measure the second over the $k(k-1)$ ordered pairs after excluding pairs sharing a motif, per dataset and architecture, since it varies by both and cannot be estimated from a subset (Appendix C). **State whether the comparison is paired**: four runs suffice for two methods on the same models, fourteen for two models or datasets (Section 5.3). **Where it does not apply**: on SP1 and TBP the AUC sits at chance, the median ablation drop being 0.002–0.003. Prior audits report that attributions *miss* annotations; we report that they *hit* what the model does not use (Appendix A).

REFERENCES

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

X. Chen, R. Cai, Z. Huang, Y. Zhu, J. Horwood, Z. Hao, Z. Li, and J. M. Hernández-Lobato. Feature attribution with necessity and sufficiency via dual-stage perturbation test for causal explanation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pages 6486–6502, 2024.

Anonymous. Title and author details omitted for double-blind review. Preprint, 2026.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, 2020.

J. Duan, H. Li, H. Zhang, H. Jiang, M. Xue, L. Sun, M. Song, and J. Song. On the evaluation consistency of attribution-based explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 206–224, 2024.

P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, and I. Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, 2023.

G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3:620–631, 2021.

M. Gheorghe, G. K. Sandve, A. Khan, J. Cheneby, B. Ballester, and A. Mathelier. A map of direct TF–DNA interactions in the human genome. *Nucleic Acids Research*, 47(4):e21, 2019.

P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un)reliability of saliency methods. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 267–280. Springer, Cham, 2019.

P. K. Koo and S. R. Eddy. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Computational Biology*, 15(12):e1007560, 2019.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

E. Nguyen, M. Poli, M. Faizi, A. Thomas, M. Wornow, C. Birch-Sykes, S. Massaroli, A. Patel, C. Rabideau, Y. Bengio, S. Ermon, C. Ré, and S. Baccus. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, 2023.

A. Patel, A. Singhal, A. Wang, A. Pampari, M. Kasowski, and A. Kundaje. DART-Eval: A comprehensive DNA language model evaluation benchmark on regulatory DNA. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. arXiv:2412.05430.

J. Reynolds and W. Pan. Benchmarking interpretability of deep learning for predictive genomics: recall, precision, and variability of feature attribution. *PLoS Computational Biology*, 21(12):e1013784, 2025.

N. Siegel, O.-M. Camburu, N. Heess, and M. Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, 2017.

Y. Zhang, Y. Li, H. Brown, M. Rezaei, B. Bischl, P. Torr, A. Khakzar, and K. Kawaguchi. AttributionLab: Faithfulness of feature attribution under controllable environments. In *NeurIPS 2023 Workshop on XAI in Action*, 2023. arXiv:2310.06514.

Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. In *International Conference on Learning Representations*, 2024.

S. Zhou, M. Huang, and K. Li. Position: Genomic model research must move beyond anecdotal evaluation of interpretability methods. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *PMLR*, 2026. arXiv:2606.07607.

Acknowledgements

Omitted for double-blind review.

Reproducibility Statement

Code, dataset construction scripts and exact seeds accompany the paper at `https://anonymous.4open.science/r/attribution-floor-9EB7`. The twelve-factor dataset is regenerable from UniBind hg38 and GRCh38, both public; the five-factor benchmark is taken unmodified from Zhou et al. (2026). We release per-sequence recovery, argmax positions, annotated motif starts and ablation probabilities for every factor and run; the cross-factor matrices for all four settings; the PWM similarity matrix and the exclusion list; and the per-factor reliability estimates with the two independent batches of draws beneath them. Appendix J states what is not released and why. Section 5.3 gives the number of runs a comparison of each kind requires, and every comparison we make is reported with the number behind it.

Ethics Statement

This work uses only public data: the GRCh38 reference genome, UniBind, JASPAR, UniProt, the ERASER movie-review corpus and the Describable Textures Dataset. No human subjects, no personally identifiable information and no newly collected data are involved. The findings concern how a class of evaluation metric behaves and are intended to make claims about model interpretability more conservative, not less; we see no application of this work that increases risk of harm.

Use of Large Language Models

Large language models were used in this work: to draft and edit prose throughout the paper; to write analysis and figure-generation scripts against specifications the authors set; and to check consistency between the numbers reported in the text and the source data. Every experimental design, every intervention, the decomposition of Section 3, the identification of Proposition 1 and all claims made from the results are the authors' own, and the authors have verified each reported figure against the underlying outputs.

A Discussion, in full

Prior audits report that attributions *miss* annotated features; we report that they *hit* features the model does not use. The first is a failure of the method, repairable by a better one; the second survives every method and architecture we tested and **grows in the architecture whose attributions are otherwise better behaved**. Three practices let it go unnoticed: ablation is treated as expensive, so usage is rarely measured; chance-level correction is not standard, so a score near zero cannot be told from one below chance; and run-to-run variation is almost never reported, so a metric moving sixty times more than the model it describes passes unremarked. Zhou et al. (2026) find that 90% of the papers they survey report no failure cases at all.

Nothing in the argument turns on which domain supplies the annotation. Images are not an exception: on textures the protocol runs and returns the same result. What object recognition lacks is the manipulation — a scope condition on the protocol, not on the claim — and that the ratio differs across domains is the sharpest test rather than a weakness, since Equation 2 makes the size of the effect a property of α/β and α/β a property of the data.

B Datasets, models and procedures

B.1 Rewriting the annotated interval

The genomic *shuffle* permutes bases within M, preserving position, length and composition. Because DNA has four symbols, permuting order is equivalent to rewriting the content. The manipulation therefore isolates *order* from *composition* but not pattern from composition: a model responding to the k-mer content of M rather than to its arrangement would be scored as unaffected. This is the intended scope — s as we define it is the salience of the arranged pattern — but it is a scope condition, not a general separation.

Table 2: Prior work against the four conditions. Zhang et al. (2023) hand-set the network weights, so usage is fixed by construction rather than measured; Reynolds & Pan estimate precision against decoys that assume the absence of signal rather than verifying its use. Concurrent work (Anonymous, 2026) verifies the chance level and reports run-to-run variation but does not measure usage independently of the attribution.

	Real trained model	Usage measured independently	Chance level verified	Run-to-run reproducibility
Zhou et al. (2026)	yes	no	no	no
Zhang et al. (2023)	no	by design	no	no
Reynolds & Pan (2025)	yes	no	no	yes
Concurrent work	yes	no	yes	yes
This work	**yes**	**yes**	**yes**	**yes**

B.2 Datasets

Benchmark (5 factors). CTCF, GATA1, MAX, SP1, TBP from Zhou et al. (2026), chromosome 9 held out ($n = 65$–$1{,}599$). The window construction cannot be reproduced without the peak files, so this benchmark cannot be extended.

Purpose-built (12 factors). From UniBind (Gheorghe et al., 2019) hg38. Of 268 factors we retained those with $\geq 20{,}000$ sites and motif length 8–20 (159 candidates), then sampled four from each tertile of compositional separability, giving 12 factors with 3-mer AUC 0.616–0.762. *Windows are constructed so the motif falls uniformly*: the start is $m - U(0, N - L)$, making π_0 exact (verified: USF1 quartiles $[124, 249, 373]$ against an expectation of $[125, 249, 373]$). Negatives are drawn from hg38 matched to the positive GC histogram in 0.02-wide bins. Chromosome 9 holds out 1,944–3,000 sequences per factor.

Proteins. 2,209 reviewed human proteins from UniProt carrying `BINDING` annotations, median 9 annotated residues in a median length of 346 — a ground-truth fraction of 2.56%. No classification

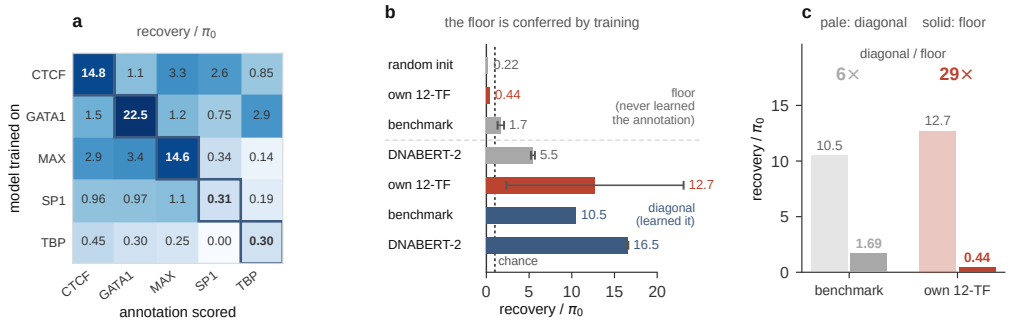

Figure 2: **(a)** Recovery as a multiple of π_0 for every ordered pair of the five benchmark factors (HyenaDNA, one run). Boxed cells are the diagonal. **(b)** The same quantity under decreasing amounts of relevant training. Error bars are the standard deviation over runs where more than one is available. **(c)** Diagonal against floor for the two genomic datasets: uniform motif placement lowers the floor without lowering the signal.

task is given, so we construct one from the ablation itself: positives are original sequences, negatives the same sequences with annotated residues replaced by random amino acids.

Text. ERASER Movies (DeYoung et al., 2020); rationales span 8.4% of a document (median 39.8 words in 679). BERT covers roughly 477 words, so we restrict to the 65% of documents whose rationale falls inside — a selection bias we state rather than correct.

B.3 Models, attributions and seeds

HyenaDNA (single-nucleotide tokenisation) and DNABERT-2 (BPE) with a mean-pooled linear head, 3 epochs; ESM2-35M (Lin et al., 2023) and BERT-base (Devlin et al., 2019) fine-tuned for 3 epochs across three seeds. Attributions: Integrated Gradients (32 steps, `[PAD]` baseline), gradient×activation and DeepLIFT, via Captum on the embedding layer. For DNABERT-2 we map token attributions to nucleotides two ways — dividing by token length and replicating — and report the former; the difference is -0.023 to $+0.008$, smaller than the seed-to-seed variation.

Seeds for `random`, `numpy`, `torch` and `torch.cuda` are fixed with a dedicated data-loader generator. Retraining under a fixed seed reproduces weights bitwise in 8 of 8 checks.

B.4 Statistical procedures

We distinguish *between-run* standard deviation, over models trained with different seeds, from *within-run* standard error, over a bootstrap of the evaluation sequences. The former is invariant to evaluation sample size while the latter falls as $1/\sqrt{n}$, so their ratio scales as $\sqrt{n}$ and *must not be reported alone*. We report an effective sample size $n^* = (c/\mathrm{SD}_{\text{between}})^2$ where $\mathrm{SE} = c/\sqrt{n}$ — the sample size at which the two coincide.

C Cross-factor: full results

Identification of the two terms. Applying Equation 4 to the cross-factor measurements gives $\beta U/\sigma$ and $\alpha\Delta/\sigma$ per factor and architecture. The floor here, and everywhere in the paper, is computed one way: for each training run, the mean recovery over off-diagonal pairs surviving the PWM screen; then the mean over runs. Every number we report passes through a single implementation of that definition.

Table 3: DNABERT-2 relative to HyenaDNA, before and after correcting each against its own per-factor floor. The correction is selective: it removes the apparent gap where the gap was floor and leaves it where it was not. A negative value means recovery falls below the floor — the model recovers its own annotation less often than models that never saw it.

	raw ($\times$ chance)			after Eq. 5		
factor	DB2/Hyena	floor Hyena	floor DB2	Hyena	DB2	interpretation
CTCF	1.24	2.00	10.39	0.566	0.576	**eliminated**
GATA1	1.41	2.46	7.93	0.392	0.517	reduced
MAX	1.74	1.86	2.80	0.297	0.535	survives
SP1	17.30	0.93	3.58	-0.019	0.058	Hyena below floor
TBP	7.82	1.01	2.54	-0.022	-0.006	both below floor

	HyenaDNA			DNABERT-2		
	APS	$\beta U/\sigma$	$\alpha\Delta/\sigma$	APS	$\beta U/\sigma$	$\alpha\Delta/\sigma$
CTCF	0.584	3.52	-3.18	0.734	1.52	-0.51
GATA1	0.414	3.04	-3.39	0.593	2.07	-1.69
MAX	0.309	3.09	-3.90	0.552	3.48	-3.27
SP1	-0.022	undefined		0.134	0.57	-2.44
TBP	-0.022	undefined		0.042	-0.14	-2.99
mean		3.22 ± 0.26	-3.49 ± 0.37		1.50 ± 1.40	-2.18 ± 1.11

APS is the diagonal; the floor enters through $\mathrm{logit}(q_{\mathrm{floor}})$. On TBP the DNABERT-2 usage term is -0.14: the model recovers the factor it was trained on slightly *less* often than factors it was not, which is the same verdict the composition screen reaches for that task.

The floor cannot be estimated from a subset. Taking three of the four available off-diagonal pairs per factor rather than all four gives a mean absolute error of 0.295 against a floor of 1.26 — the value for a single HyenaDNA run, against the 1.69 of Table 1 averaged over five; two pairs give 0.586 and one gives 0.885. On DNABERT-2 the same errors are 0.601, 1.108 and 1.803 against a floor of 5.45. The full $k(k-1)$ set is required, which on the benchmark is 20 attribution passes.

Correcting increases run-to-run variation rather than reducing it. Over five independent runs the coefficient of variation of raw recovery is 0.172 and that of the corrected score 0.211, because the floor is itself estimated from the same runs and enters both numerator and denominator. The correction is for comparability across datasets and architectures, not for stability; where stability is what is needed, the ablation-prediction AUC of Section 7 remains the better quantity.

The floor appears within one epoch. Checkpointing three factors at each epoch over three seeds and scoring them cross-factor gives $1.84\times$ after a single epoch, against $0.22\times$ at initialisation; the remaining epochs move it little ($2.58\times$, $2.57\times$). Most of the floor is conferred by the first epoch of training.

The floor is not predicted by motif properties. On the four benchmark factors with a JASPAR matrix — TBP has none, and its motif length is set by hand — the per-factor floor appears to correlate with information density — total information content divided by motif length — at $r = -0.998$, and not with length ($r = +0.283$) or total information ($r = -0.323$). It does not replicate. Measuring the floor on twelve further factors over all 12×11 ordered pairs and pooling gives $r = +0.12$ ($n = 16$, $p = 0.65$) for density — the sign reverses — with -0.31 for length and -0.24 for total information, none of them significant. We report the benchmark observation and its failure to replicate, and offer no motif property as a predictor of the floor.

D VARIANCE DECOMPOSITION

Table 4 in Section 5.3 gives the decomposition. Two points about it belong here.

Table 4: Variance decomposition of per-sequence recovery over three attribution methods (IG, gradient×activation, DeepLIFT) and up to nine runs. SP1 and TBP have 65 and 81 held-out sequences; their sequence term is correspondingly small and their residual large.

	runs	sequence	run	method	run×method	residual
CTCF	9	39.7%	1.6%	1.7%	0.2%	56.7%
GATA1	9	37.8%	0.7%	3.7%	1.2%	56.7%
MAX	4	43.7%	0.1%	0.6%	0.9%	54.8%
SP1	4	10.0%	0.2%	0.6%	0.9%	88.3%
TBP	4	13.8%	0.6%	1.0%	0.6%	84.0%

The two factors with the smallest held-out sets, SP1 ($n = 65$) and TBP ($n = 81$), have a correspondingly small sequence term and a residual above 84%. With so few sequences the between-sequence variance is estimated poorly and much of it falls into the residual; the three larger factors, where it is estimated from 348–1,599 sequences, are the ones on which Section 5.3 rests.

Power calculations. Writing the recovery of method m on sequence q under run s as

$$r_{msq} \;=\; \mu + \alpha_m + \beta_s + \gamma_q + (\alpha\beta)_{ms} + \varepsilon_{msq}, \tag{6}$$

the standard error of a mean over S runs and n sequences is

$$\mathrm{SE}(\bar{r}) \;=\; \sqrt{\frac{\sigma^2_{\mathrm{run}}}{S} + \frac{\sigma^2_{\mathrm{seq}}}{n} + \frac{\sigma^2_{\varepsilon}}{Sn}}. \tag{7}$$

For a paired comparison the run effect cancels and the requirement is

$$S \;\geq\; \frac{(z_{1-\alpha/2} + z_{1-\eta})^2 \cdot 2\sigma^2_{\alpha\beta}}{\Delta^2}, \tag{8}$$

giving $S \geq 4$ at $\Delta = 0.05$ and $S \geq 1$ at $\Delta = 0.10$. For an unpaired one it is

$$S \;\geq\; \frac{(z_{1-\alpha/2} + z_{1-\eta})^2 \cdot 2\sigma^2_{\mathrm{run}}}{\Delta^2}, \tag{9}$$

giving $S \geq 7$ runs per condition at $\Delta = 0.10$ and 80% power with $\sigma_{\mathrm{run}} = 0.062$ — fourteen in total.

Pooling the interaction with the residual, as a two-way decomposition would, gives 56.9% and 57.9% for CTCF and GATA1. Separating the run×method term leaves the sequence and method terms unchanged and puts the interaction at 0.2% and 1.2% — small enough that method comparisons on a fixed set of models need one or two runs, which is the point of Equation 8.

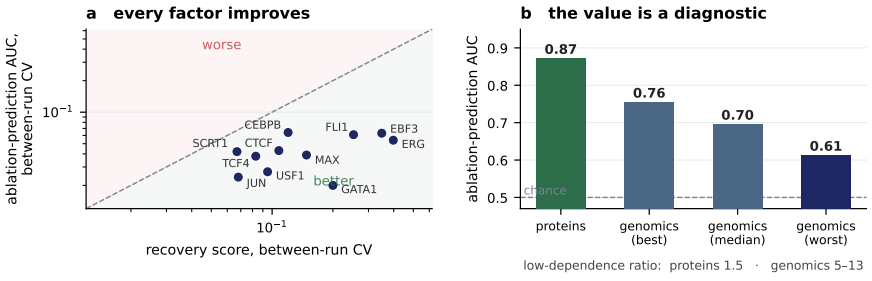

Figure 3: **(a)** Between-run CV, recovery against ablation-prediction AUC; every factor lies below the diagonal. **(b)** The AUC value reports how far attributions track usage, inverting the ratios of Section 5.6.

E Six controls in full

Negative sequences. $\mathrm{APS}_{\mathrm{null}} \in [-0.015, +0.009]$; the closed form holds where no feature is present.

Low-dependence positives. On sequences where neither the annotated motif nor the attributed window affects the prediction, recovery stands at 5–14× chance in ten of fifteen factors, each averaged over nine training seeds (GATA1 14.1, JUN 12.6, USF1 9.8, MAX 9.3, CTCF 8.3, FLI1 7.9, SCRT1 7.6, CEBPB 6.4, TCF4 5.8, PDX1 5.5). Eleven of fifteen exceed 2×. The two below chance — TFAP2C 0.7 and ZNF143 0.5 — have the highest attribution collapse rates, 0.44 and 0.52.

Refinement raises the ratio. Restricting to the doubly-ablated subset increases the baseline (CTCF 3.1→7.1, GATA1 5.6→13.4, MAX 5.0→9.3): the sequences removed are those where the model uses a region the attribution found, which by construction pulls recovery of the *annotated* feature down.

Rewriting removes the effect. Permuting bases within the interval collapses APS to near chance: CTCF $0.555 \to 0.080$, GATA1 $0.491 \to 0.045$, MAX $0.357 \to 0.013$. On the purpose-built dataset shuffled APS falls to $[-0.013, +0.047]$ across all twelve factors. Restricting the manipulation to low-dependence sequences gives the same result: across 26 factor×seed combinations, intact APS ranges from -0.015 to $+0.668$ and rewritten APS from -0.014 to $+0.029$.

Uniform placement. The effect is not an artefact of the benchmark's centre bias: on the purpose-built dataset with uniformly placed motifs, ten of twelve factors show baselines of 1.5–9.7×.

Three methods. Thirty of thirty-six factor×method combinations exceed 2× (IG 0.7–13.0, gradient×activation 1.0–11.5, DeepLIFT 1.4–9.9). Rankings reverse across factors: IG > DeepLIFT > grad×act on CTCF, the reverse on GATA1. Gradient×activation has the lowest collapse rate everywhere (MAX $0.215 \to 0.003$) yet this does not improve reproducibility; the two are independent.

Two architectures: the better model is worse. On DNABERT-2 the ratios are larger than on HyenaDNA in all five shared factors (CTCF 11.2 vs 3.1; GATA1 19.5 vs 5.6; MAX 22.0 vs 5.0), with a low-dependence baseline of 0.43 against a whole-set APS of 0.73: **more than half the score is independent of the model**. Its collapse rate is an order of magnitude lower (0.003–0.067 against 0.28–0.41). The architecture whose attributions are better behaved carries the larger model-independent component. Improving the model does not fix this.

Collapse itself is architecture-dependent: when an attribution finds nothing to point at it falls onto a fixed position — position 0 for the causal HyenaDNA, the final `[SEP]` token for bidirectional BERT.

These six controls establish that the model-independent component exists and is large. They do not measure it. The next section does.

F TWO FURTHER DOMAINS: MEASUREMENTS

Running the protocol requires five conditions: sparse ground truth; demonstrable model usage; well-defined ablation; a manipulation that destroys the pattern while preserving position and composition; and a metric matched to the structure of the ground truth.

F.1 PROTEINS

Fine-tuning ESM2 on the constructed task gives validation accuracies of 0.758, 0.767 and 0.794: the model uses the annotated residues. Ablating them drops the predicted probability by 0.369–0.538, matching genomics (CTCF 0.47) and exceeding text (0.24). Recovery reaches $+0.253$ (CV 0.052) for Integrated Gradients and $+0.230$ for gradient×activation; replacing the residues collapses it to -0.004 and $+0.045$ respectively — as cleanly as the genomic rewrite.

Applying the same symmetric ablation as in genomics, the fraction of sequences in which the model uses a region the attribution found is 0.14–0.40 across three seeds, matching the genomic range. On the doubly-ablated subset the ratio is 1.51, 1.47 and 1.49 ($n = 61$–114), consistent across seeds but with a bootstrap interval that includes one in each.

The negative-set design affects the protein ratio. The task above constructs negatives by replacing the annotated residues of the same sequence, so the background carries no discriminative information and the model has no reason to attend to it — a difference from the genomic setting, where negatives are independent windows matched on GC content. We rebuilt the task with negatives drawn from reviewed human proteins carrying no `BINDING` annotation, matched on length and amino-acid composition (composition-only AUC 0.618, inside the 0.616–0.762 range of the genomic 3-mer screen). Validation accuracy rises from 0.77 to 0.83, the ablation-measured dependence halves ($0.369 \rightarrow 0.174$), the low-dependence fraction rises from 0.35 to 0.58, and the low-dependence ratio rises from 1.49 to 2.21 (2.27, 2.08, 2.27 over three seeds; overall ratio 4.36 against 7.4). **The design accounts for part of the low protein ratio and not for the gap to genomics: under either design the protein ratio stays an order of magnitude below the genomic 5–14×, and lands where text does.**

F.2 TEXT

Masking the rationale drops the predicted probability by 0.242 (CI_{95} 0.181–0.306) against a matched random-token control of -0.011. Recovery reaches 2.14× chance overall and 2.25× on low-dependence documents — again *higher* on the low-dependence subset.

Word-order shuffling is not the analogue of the genomic manipulation. Permuting word order within a rationale leaves APS unchanged (IG $0.249 \rightarrow 0.205$) or raises it (grad×act $0.236 \rightarrow 0.266$), because the vocabulary survives: a sentiment-bearing word carries its signal regardless of position. Masking the rationale tokens instead — each replaced with `[MASK]`, against a control masking the same number of tokens at random positions — halves APS ($0.249 \rightarrow 0.098$).

The substitution does not change the conclusion. We ran the intervention three ways: replacing rationale tokens with `[MASK]`, with words drawn uniformly from the corpus vocabulary, and with words drawn in proportion to corpus frequency. The measured dependence differs sharply — the prediction falls from 0.957 to 0.701 under masking but only to 0.900 and 0.853 under the two word substitutions, so `[MASK]` removes content and shifts the input distribution together, as the genomic shuffle does not. The low-dependence *set* grows accordingly, from 95 documents to 130 and 121. The low-dependence *ratio* does not move: 2.27, 2.06 and 2.10, each over three seeds. The excess we report is a property of the attribution, not of how the rationale was removed.

A position artefact that the protocol catches. Integrated Gradients on BERT places 83.6% of its top windows within the final 50 tokens, driving APS below chance (-0.014). The cause is the trailing `[SEP]` and padding, which differ maximally from the integration baseline and receive large attributions regardless of content. Excluding special tokens moves APS to $+0.119$ and the top window to a median position of 26 (IG) and 47 (grad×act) out of a 512-token window, with 0.7% now falling in the final 50. We report both numbers: the uncorrected one is what a study not checking for this would publish, and the correction is a single line of code.

F.3 THE RATIO DIFFERS BY DOMAIN, AS PREDICTED

Domain	Ground truth	Overall ratio	Low-dependence ratio
Genomics	2–4%	∼14	**5 – 13**
Text	8.4%	2.14	**2.25**
Proteins (paired negatives)	2.56%	7.4	**1.5 (n.s.)**
Proteins (matched negatives)	2.56%	4.4	**2.2**

The effect is not uniform, and Equation 2 says it should not be. The ratio measures $\alpha s/\beta u$: genomics is the extreme in which pattern salience dominates model usage, proteins the case in which it does not — there, the metric measures what it is read as measuring. Genomic motifs are short and compositionally distinctive; binding residues are scattered and compositionally similar to their surroundings.

F.4 THE METRIC IS BOUND TO THE STRUCTURE OF THE GROUND TRUTH

Applying both a maximum-window score and a top-k score to the same attributions, the two favour opposite structures. On CTCF, whose ground truth is one contiguous interval, the window score is 0.568 and top-k 0.302: attribution mass spreads around the interval, and a window captures what individual-position ranks miss. On proteins, whose annotated residues are scattered, a single window can cover only one cluster and 74.7% of sequences score exactly zero, against 26.5% under top-k.

The perception score of prior work presumes a contiguous ground truth. On scattered annotations it reports systematically low values that reflect the metric rather than the attribution. We therefore use a window score in genomics — where it matches both the ground truth and prior work — and top-k in proteins, defining π_0 accordingly in each case.

G TEXTURES: THE PROTOCOL ON IMAGES

Construction. Images are 128×128 crops of a DTD `marbled` photograph. A 32×32 patch of `banded` texture (the target) is present in half of them at a uniformly random position; a `grid` patch (the distractor) is present in all of them, also at random. Both patches are mean- and contrast-normalised so that neither is distinguishable from the background by first-order statistics. The task is to detect the target. The ground truth occupies 6.3% of the image, within the 2–4% of a genomic motif and far from the 27% of an annotated object in ImageNet.

A four-block convolutional network is trained from scratch for 25 epochs on 8,000 images, ten times with different seeds. Attributions are Grad-CAM on the last convolutional block, upsampled to the input; recovery is the overlap between the annotation and the highest-summing 32×32 window, against $\pi_0 = 32^2/128^2 = 0.0625$.

Conditions. The diagonal scores the target in a model trained to detect it. The floor scores the *distractor*, which no model was trained to detect, in the same models — the cross-factor construction of Section 5.5. The shuffle condition scores the target after permuting the pixels within the patch, preserving the intensity distribution exactly while destroying the arrangement. Untrained models are randomly initialised networks of the same architecture.

seed	acc	diag.	floor	shuf.	untr. (t)	untr. (d)	med. u	low-dep.
2	0.948	11.66	6.31	2.85	1.63	1.56	+0.67	9.94
3	0.953	11.28	6.52	1.91	1.42	3.87	+0.53	11.71
6	0.952	12.05	6.51	2.60	1.78	3.82	+0.33	11.41
7	0.973	11.69	6.62	2.02	1.81	1.25	+0.36	11.79
mean		11.67	6.49	2.34	1.66	2.62	+0.47	**11.21**
sd		0.31	0.13	0.45	0.18	1.41		0.86
not learned (accuracy < 0.80, median $u \approx 0$):								
1	0.638	9.57	7.19	1.77	0.88	2.00	+0.00	—
4	0.705	13.13	5.46	2.27	0.98	1.60	+0.00	—
5	0.547	11.11	5.04	2.44	1.05	1.48	+0.00	—
8	0.795	12.67	6.38	2.50	1.76	1.56	+0.01	—
9	0.772	11.33	5.90	1.65	2.44	3.50	+0.85	8.23
10	0.748	12.69	6.13	2.12	0.92	1.91	+0.00	—

Median u is the ablation-measured drop: the target patch replaced with background texture, minus the same replacement at a control position. The low-dependence column is recovery on the examples with $u < 0.05$. *untr. (t)* and *untr. (d)* are randomly initialised networks of the same architecture scored against the target and the distractor annotation respectively. Seed 9 is grouped with the runs that did not learn the task on accuracy (0.772) although its median u is not near zero.

The rank-based score of Section 7 does *not* transfer here. Its AUC is 0.54 ± 0.05, at chance, and the reason is visible in the median-u column: a run either uses the patch throughout ($u \approx 0.5$) or not at all ($u \approx 0$), so thresholding at the median of a bimodal distribution separates nothing. This is the condition Section 7.4 describes, and the AUC reports it correctly rather than failing silently.

Which runs are interpretable. Six of ten runs do not reach 0.80 accuracy on a task the other four solve at 0.95, and in four of those the median ablation drop is 0.000 — the model does not use the target patch at all, so every example is nominally "low-dependence" and the condition carries no information. We report those runs for completeness and read the protocol only on the four that learned the task. The floor is unaffected by this split (5.04–7.19 across all ten against 6.31–6.62 on the four), which is itself consistent with the claim: the floor comes from having learned *something*, not from having learned the target.

Reading the four. The floor is 6.49 ± 0.13 chance, the shuffle brings recovery to 2.34 ± 0.45, and on the 269 examples the ablation identifies as unused ($u < 0.05$) recovery is 11.21 ± 0.86 — indistinguishable from the diagonal at 11.67 ± 0.31. That last figure is the image analogue of Section 5.4: the attribution locates the annotation just as often on examples where removing it does not move the prediction.

Why the integration baseline matters here. Running the same experiment with Integrated Gradients rather than Grad-CAM gives an untrained recovery of 13–29$\times$ across four choices of baseline (gray, sampled noise, a pixel shuffle of the input, and zero), which exceeds the trained floor and makes the comparison uninformative. One-hot sequences are equidistant from the integration baseline at every position, which is why the untrained value is $0.22\times$ in genomics; a continuous-valued image has no such property, and a region whose local statistics differ from the background is far from any fixed baseline whether or not the model uses it. Grad-CAM takes no baseline and the effect disappears. This is a property of the attribution method, not of the domain, and it is why the earlier literature on natural images may not have surfaced the floor.

The floor does not require having learned the target. Seed 5 reaches 0.547 accuracy — it does not learn the task — yet its floor, 5.04, sits inside the range the four learned runs give (6.31–6.62) and inside the range across all ten (5.04–7.19). A model that has not learned the target still confers a floor on the distractor, which is the same conclusion the genomic layers reach from the other direction.

H NEGATIVE RESULTS

We report the manipulations that did not work, in the form in which we ran them.

Ten corrections to the absolute value. Window-jitter filtering; a distance-based continuous score in place of binary overlap; excluding cases whose attribution collapses to position 0; stratifying by ablation-measured dependence; switching attribution method; top-k windows in place of a maximum window; boundary filters excluding motifs within L of either end; a sample-size adjustment; and the score- and position-level differences below. Each adjusts where the score sits, and four leave at least one factor worse than before, measured as the ratio of between-run standard deviation after correction: window jitter (MAX 1.49), excluding collapsed cases (MAX 1.12), and switching method (MAX 1.59). Equation 2 predicts this: the contaminating term is in the absolute value, and shifting the absolute value cannot remove it.

Position-level difference. Applying $|a| - |a_{\text{rewritten}}|$ position by position raises the low-dependence baseline from $4.1\times$ to $17.8\times$ on CTCF over four runs. Normalising the difference and taking a ratio behave the same way. Where both the original and the rewritten sequence produce low, flat attributions, the difference is dominated by noise in two small quantities rather than by the signal being isolated.

Correcting against the floor does not stabilise the score. The coefficient of variation rises from 0.172 to 0.211 over five runs, because the floor is estimated from the same runs and enters both numerator and denominator of Equation 5; Appendix C gives the detail.

The floor cannot be estimated from a subset. Subsetting the off-diagonal pairs inflates the error steeply in both architectures; Appendix C gives the figures.

I INTERVENTION VALIDATION

The results rest on three interventions. Each was checked against a way it could have failed.

The chance level is not an artefact of the closed form. The derivation assumes a uniform admissible start, which is exact away from the boundaries. Enumerating admissible windows per example instead gives a ratio of closed form to exact mean of 1.001–1.008 across the five benchmark factors. Individual boundary-loaded examples differ by up to $1.8\times$, but the aggregate shift is below 1%. Independently, scoring negative sequences — in which the annotated feature is absent by construction — through the identical pipeline returns $[-0.015, +0.009]$, consistent with the derived value.

The salience term is not an integration-baseline artefact. A natural explanation for αs is that positions differing most from the reference input receive large attributions regardless of use. Replacing the `[PAD]` baseline with the `N` token changes recovery by $+0.013$ to $+0.064$ — in the direction opposite to the one that explanation predicts, and smaller than the run-to-run variation of 0.062.

The ablation measurement is reliable enough to threshold. The four-way partition of Section 4.4 applies a threshold to a quantity measured with noise: the replacement is drawn at random, so a single draw is one sample from an intervention distribution rather than a property of the sequence. Running two fully independent batches of three draws each per sequence, with the pixel pool, control position and random stream regenerated every draw, and estimating

$$\lambda \;=\; \frac{\mathrm{Cov}(A, B)}{\mathrm{Var}\left((A + B)/2\right)}$$

gives $\lambda = 0.976$ (CTCF, $n = 400$), 0.972 (GATA1, $n = 348$) and 0.967 (MAX, $n = 400$). Single-batch reliabilities are lower (0.905–0.970), which is why the pooled measurement is used. At these values the attenuation of a threshold-selected contrast is under 3%, so the partition is not an artefact of measurement noise.

The substitution removes the feature. Scanning replaced intervals with the corresponding position weight matrix, the best-match score falls to background in every factor tested: expressed as a fraction of the original-to-background distance, 94–99% of the motif signal is removed (CTCF 99.4%, MAX 95.3%, GATA1 94.2%, SP1 94.2%). The substitution does not incidentally create a functional site.

The threshold is not arbitrary. Among sequences classified as low-dependence at $u < 0.05$, the median u is 0.006–0.011 — below a fifth of the threshold — and the lower quartile is negative: for nearly half of them the prediction *rises* after ablation.

The measurement is reliable. A threshold on a noisy measurement selects a contaminated subset, so we estimated the reliability of u with two fully independent batches of three draws each, regenerating the surrogate, control position and random stream per draw. Under a parallel-replicate model $\mathrm{Cov}(A, B)$ estimates $\mathrm{Var}(u)$: the reliability is 0.967–0.976 across CTCF, GATA1 and MAX ($n = 348$–400), with between-batch correlation 0.94–0.95.

Ablation measures necessity; sufficiency bounds the redundant-path objection. Removing a feature and observing no change is weaker than the computation not involving it: a redundant path would be scored as non-use. We bound how often this occurs by running the complementary intervention. Keeping the annotated interval and replacing everything else with background gives $P(\text{motif only})$; replacing everything except a control interval of the same length gives $P(\text{control only})$.

A sequence is *necessary* if ablating the motif drops the prediction and *sufficient* if the motif alone retains it. Over the five benchmark factors at one run:

	n	necessary	sufficient	both	neither
CTCF	400	375	344	330	11 (2.8%)
GATA1	348	332	316	303	3 (0.9%)
MAX	400	325	376	308	7 (1.8%)
SP1	65	40	63	38	0
TBP	81	49	77	45	0

A redundant path places a sequence in the *sufficient but not necessary* cell: ablating the annotation does not drop the prediction because another route compensates, while the annotation alone retains it. Restricted to this one run and to the sufficiency intervention, that cell holds 3.5% (CTCF), 3.7% (GATA1) and 17.0% (MAX) on the three factors whose task requires the motif, and 38.5% (SP1) and 39.5% (TBP) on the two whose median ablation drop is 0.002–0.003 — where the model uses the motif so little that the necessity test has nothing to detect. The bound is therefore 3.5–17% where the measurement is informative, against the 0.09–0.55 that Section 4.4 reports for the same cell over all factors and seeds under the symmetric ablation. Reading the *neither* cell instead gives under 3% and understates the limitation.

The sufficiency gap $P(\text{motif only}) - P(\text{control only})$ separates the factors the same way ablation does: 0.451 (CTCF), 0.185 (GATA1), 0.114 (MAX) against 0.011 (SP1) and 0.003 (TBP). On the two factors whose task does not require the motif, keeping the motif alone recovers nothing the control does not.

Cross-factor pairs sharing a motif are excluded. The cross-factor design assumes the model has not been trained on the annotation it is scoring. That fails where two factors bind the same motif. We screen every pair by the maximum column-wise correlation between their JASPAR position-probability matrices, taken over all offsets and both strands, and exclude pairs above 0.60. Three pairs are excluded, all E-box factors: USF1–JUN (0.725), USF1–MAX (0.712) and USF1–FLI1 (0.655). The highest surviving pair is CTCF–SP1 at 0.582. No pair among the five benchmark factors is excluded.

J REPRODUCIBILITY

Seeds. Seeds for `random`, `numpy`, `torch` and `torch.cuda` are fixed, with a dedicated generator for the data loader. Retraining under a fixed seed reproduces weights bitwise in 8 of 8 checks. Where we report variation across runs, the runs differ only in this seed.

Cost of the reference point. On the benchmark the ablation over $n = 100$ takes 300 forward passes against the 43,173 a full-set ablation would, and against the 51,168 Integrated Gradients needs over the same sequences.

What is released. Per-sequence recovery, argmax position, annotated motif start and ablation probabilities for every factor and run; the cross-factor matrices for all four settings; the PWM similarity matrix and the exclusion list; per-factor λ estimates with the underlying two-batch draws. Scripts regenerate every figure and table from these files.

What is not released. The five-factor benchmark windows are redistributed under the source study's terms. The purpose-built twelve-factor set is constructed from UniBind hg38 and GRCh38, both public; we release the construction script rather than the sequences.

K LIMITATIONS

The ablation is an input–output measurement: removing the feature not changing the prediction is weaker than the computation not involving it, and a redundant path would be scored as non-use. Appendix I bounds how often this occurs at 3.5–17% of sequences on the three factors where the necessity test is informative, but it does not eliminate it. The rewrite preserves composition, so pattern salience is not separated from compositional salience — the ambiguity under which Section 6's

prediction fails. Text results rest on 65% of documents, protein ratios on 61–114 sequences per seed, and Equation 2 is an assumption warranted by four predictions, three of which hold.