WHAT YOUR BENCHMARK CANNOT SEE: NON-IDENTIFIABILITY OF BACKBONE SYMMETRY UNDER INVARIANT READOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmark scores can disguise a structural impossibility as an ordinary performance drop. We show that under a fixed invariant readout, scores cannot distinguish directional information missing from the backbone from information erased at readout, making backbone symmetry non-identifiable. When input transformations exchange target labels, invariant predictions force orbit-paired accuracy to exactly zero on orbit-closed free orbits, regardless of loss, optimisation, decoding, or capacity, even while conventional metrics remain non-zero. We identify joint channel and positional symmetrisation as the mechanism and test its predictions using parameter-matched readouts across three DNA foundation models on a new direction-aware promoter benchmark. Linear probes recover direction from Caduceus representations, while a target-matched equivariant readout restores paired success and improves aggregate performance. A synthetic sweep and an MNIST-derived reflection task reproduce the obstruction beyond DNA; a matched invariant genomic control shows no paired collapse. The problem is not symmetry itself, but a readout action that conflicts with the target—and benchmark scores alone need not expose it.

1 INTRODUCTION

Symmetry is a powerful inductive bias, but its validity depends on how both the input and the target transform. If a transformation rotates a vector, permutes labels, or exchanges directional channels, enforcing invariance can exclude the desired predictor from the hypothesis class. Existing work establishes the benefits of correctly specified equivariance and the risks of incorrect or approximate symmetry (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Elesedy & Zaidi, 2021; Wang et al., 2023; 2022). We ask a complementary evaluation question: *can a standard benchmark detect that the wrong symmetry was imposed?*

Representation benchmarks, including those for DNA language models, typically attach a common lightweight readout to each backbone (Marin et al., 2024; Tang et al., 2025). Their scores therefore belong to the composition of backbone and readout. Under an invariant readout, a low score cannot distinguish directional information that is absent from information erased downstream, while non-zero aggregate metrics do not exclude exact orbit collapse. Probe studies likewise show that representation conclusions depend on the probe (Hewitt & Liang, 2019; Belinkov, 2022). Closest to our readout analysis, Talhi et al. (2026) show that linear permutation-invariant GNN readouts discard non-trivial representation components. Their obstruction concerns separating inputs and so costs discriminative power, and it requires the readout to be linear. We instead consider targets that transform non-trivially and assume only invariance of the score as a function of the input.

Reverse complementation (RC) in genomics makes this ambiguity concrete. RC-aware genomics has used parameter sharing (Shrikumar et al., 2017), conjoined prediction (Zhou et al., 2022), and equivariant architectures (Mallet & Vert, 2021; Schiff et al., 2024). Invariance is appropriate for scalar targets unchanged by RC, but not for direction-resolved outputs. Promoters can initiate transcription rightward, leftward, or in both directions, with direction-dependent sequence and chromatin determinants (Core et al., 2014; Ibrahim et al., 2018). RC exchanges these activities, whereas a

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.

binary "promoter present" label hides the distinction. We therefore construct an orbit-closed benchmark with the explicit two-channel action $\rho(R)(a_{\rightarrow}, a_{\leftarrow}) = (a_{\leftarrow}, a_{\rightarrow})$.

On this benchmark, we prove that every invariant readout has exactly zero *orbit-paired accuracy* on free directional orbits, independent of loss, optimisation, decoding, or capacity, even when conventional metrics remain non-zero. We establish the resulting non-identifiability of backbone symmetry, localise the information loss to joint channel and positional symmetrisation, and derive orbit-level diagnostics and an *a priori* symmetrisation gap. Parameter-matched interventions across HyenaDNA, NTv3, and Caduceus recover the directional signal and repair the collapse with a symmetry-matched equivariant readout, while a matched invariant genomic control shows no structural obstruction. Direction-aware promoter prediction is our auditable test case; the broader message is that a scientific benchmark should verify the target action at the readout before attributing its score to the representation.

Our theory holds for any finite group acting on inputs and labels through a permutation representation, our genomic evidence is confined to the exact $\mathbb{Z}_2$ reverse-complement action, and we verify the ceiling $1 - \frac{1}{2}\pi_A$ across $\pi_A \in [0, 1]$ in a controlled non-genomic instance (Section 3.4). The corresponding exact-zero OPA_A statement applies whenever the free-orbit set is non-empty, i.e. $\pi_A \in (0, 1]$. Settings with approximate, continuous, or uncertain symmetries lie outside the evidence we present.

2 THEORY

We formalize three claims. Imposing a symmetry at the readout restricts the attainable function class, and the restriction depends on where the symmetry is imposed (Theorem 1). When the label space carries a non-trivial representation of the group, the restriction is not a loss of accuracy but a structural prohibition, with an exact and decoding-independent zero (Theorem 2). Because a benchmark score is a property of the composition of backbone and readout, the symmetry properties of a backbone are not identifiable from scores obtained under a fixed invariant readout (Theorem 3). All proofs are deferred to Appendix B.

2.1 SETUP

Group action. Let $\mathcal{X}$ be the space of length-L DNA sequences and let $R : \mathcal{X} \to \mathcal{X}$ be the reverse-complement (RC) map. Since $R^2 = \mathrm{id}$, the set $G = \{e, R\} \cong \mathbb{Z}_2$ is a group and the orbit of x is $\mathcal{O}(x) = \{x, Rx\}$. Labels $y = (a_{\rightarrow}, a_{\leftarrow})$ record initiation activity in the $5' \to 3'$ direction of the presented strand and of its complement, giving $\mathcal{Y} = \{y_{00}, y_{10}, y_{01}, y_{11}\}$. Reverse-complementing exchanges which strand is read $5' \to 3'$, so G acts on $\mathcal{Y}$ through the permutation representation

$$\rho(R) = \mathrm{swap} : \quad y_{00} \mapsto y_{00}, \quad y_{11} \mapsto y_{11}, \quad y_{10} \leftrightarrow y_{01}. \tag{1}$$

We write $\mathcal{Y} = S \sqcup A$ with fixed labels $S = \{y_{00}, y_{11}\}$ and swapped labels $A = \{y_{10}, y_{01}\}$, and set $\pi_A = \Pr[Y \in A]$. Two properties are used throughout.

Assumption 1 (Label equivariance)**.** $\Pr[Y = y \mid X = Rx] = \Pr[Y = \mathrm{swap}(y) \mid X = x]$ for all x, y.

Assumption 2 (Orbit-closed evaluation)**.** $x \in \mathcal{D} \Rightarrow Rx \in \mathcal{D}$.

Assumption 1 is a definitional identity equation, not a modeling hypothesis: $a_{\rightarrow}$ names the activity of whichever strand is presented, so relabelling the presented strand permutes the label exactly as in equation 1. Assumption 2 is a property of the evaluation protocol, which we enforce by construction (Section 3.1).

Orbit closure also makes the input marginal R-invariant, $X \overset{d}{=} RX$; all population statements below are taken with respect to such a distribution, which is what licenses the pairing of risk contributions over $\{x, Rx\}$.

RC palindromes ($Rx = x$, orbits of size one) are consistent with Assumption 1 and deterministic labels only if $y(x) \in S$, since they force $y(x) = \mathrm{swap}(y(x))$; every orbit carrying a label in A therefore has size exactly two, and palindromes affect none of the statements below.

Backbone and isotypic decomposition. A backbone is a map $h : \mathcal{X} \to \mathbb{R}^{T\times D}$. The RC parameter-sharing backbones (Shrikumar et al., 2017; Mallet & Vert, 2021; Schiff et al., 2024) satisfy

$$h(Rx) = \sigma\, h(x), \qquad \sigma = \sigma_L \otimes \sigma_c, \tag{2}$$

where σ_L reverses the length axis and σ_c the channel axis. Writing $h = (h_1, h_2)$ for the two halves of the channel axis and rev for reversal within a half, put $\Pi_+ h = \frac{1}{2}(h_1 + \operatorname{rev} h_2)$ and $\Pi_- h = \frac{1}{2}(h_1 - \operatorname{rev} h_2)$, so $\mathbb{R}^D = V_+ \oplus V_-$ with $\dim V_\pm = D/2$. Then equation 2 gives

$$\Pi_+(\sigma h) = +\,\sigma_L\, \Pi_+ h, \qquad \Pi_-(\sigma h) = -\,\sigma_L\, \Pi_- h, \tag{3}$$

i.e. V_+ carries the trivial and V_- the sign representation of G on the channel axis. Crucially, *both* components are still acted on by σ_L along the length axis; this two-axis structure is used in Proposition 2. A pooling operator is *position-symmetric* if it is invariant to token order (mean, max and content-based attention pooling all are) and *position-aware* otherwise; we write $m_\pm = \mathrm{Pool}(\Pi_\pm h)$.

Symmetry class of a readout. A score function $s : \mathcal{X} \to \mathbb{R}^{\mathcal{Y}}$ is *G-invariant* if $s(Rx) = s(x)$ for all x, and *ρ-equivariant* if $s(Rx) = \mathrm{swap}(s(x))$ for all x. The classifier is $\hat{y}(x) = \arg\max_y s_y(x)$.

2.2 A UNIFIED READOUT FAMILY

All readouts we study are instances of one template. For a shared network q on features z and a representation ρ of G on $\mathbb{R}^{\mathcal{Y}}$,

$$H_\rho(z) = \frac{1}{|G|} \sum_{\gamma\in G} \rho(\gamma)^{-1} q(\gamma \cdot z) = \tfrac{1}{2}\left[\, q(z) + \rho(R)^{-1} q(\sigma z)\,\right], \tag{4}$$

which yields, with P the permutation matrix of swap,

$$H_{\mathrm{free}}(z) = q(z), \qquad H_{\mathrm{inv}}(z) = \tfrac{1}{2}[q(z) + q(\sigma z)], \qquad H_{\mathrm{equi}}(z) = \tfrac{1}{2}[q(z) + P\, q(\sigma z)], \tag{5}$$

where H_{free} omits the symmetrisation. Since swap must fix S pointwise, the permutation representations of G on $\mathcal{Y}$ compatible with equation 1 are exactly $\{\mathrm{id}, \mathrm{swap}\}$: *the design space of symmetrised readouts for this task has exactly two points*, and H_{inv} is one of them. All three instances share the parameters of q, so every comparison among them is exactly parameter-matched and differs only in ρ.

2.3 PARAMETER-FREE DIAGNOSTICS

Let N_A be the number of evaluation samples whose label lies in A, let $\mathcal{P}_A$ be the set of *free orbits*, i.e. orbits whose two elements carry the two distinct labels of A; $|\mathcal{P}_A| = N_A/2$.

Definition 1 (Diagnostics)**.** The *orbit-paired accuracy* and the invariance/equivariance agreement rates are

$$\mathrm{OPA}_A = \frac{1}{|\mathcal{P}_A|} \sum_{\{x,Rx\}\in\mathcal{P}_A} \mathbf{1}[\hat{y}(x) = y(x)] \cdot \mathbf{1}[\hat{y}(Rx) = y(Rx)], \tag{6}$$

$$\mathrm{IAR}_A = \mathbb{E}_{x:\, y(x)\in A}\, \mathbf{1}[\hat{y}(Rx) = \hat{y}(x)], \qquad \mathrm{EAR}_A = \mathbb{E}_{x:\, y(x)\in A}\, \mathbf{1}[\hat{y}(Rx) = \mathrm{swap}(\hat{y}(x))]. \tag{7}$$

The denominator of equation 6 counts orbits, not samples. Both agreement rates are restricted to $y(x) \in A$: on S the map swap is the identity, so the two coincide there and their unrestricted versions are uninformative. We additionally report the action-specific logit residuals

$$\delta_{\mathrm{inv}} = \max_x \|s(Rx) - s(x)\|_\infty, \qquad \delta_{\mathrm{equi}} = \max_x \|s(Rx) - Ps(x)\|_\infty, \tag{8}$$

which certify numerical invariance and equivariance, respectively, rather than agreement only after decoding.

2.4 WHAT INVARIANCE COSTS, AND WHERE

Invariance can be imposed on features (average hidden states, then read out) or on logits (read out twice, then average). These are not equivalent. With $\mathcal{G}$ a readout class closed under precomposition with Π_+, let $\mathcal{F}_{\text{feat}} = \{g \circ \Pi_+ : g \in \mathcal{G}\}$ and $\mathcal{F}_{\text{logit}} = \{\frac{1}{2}(g + g \circ \sigma) : g \in \mathcal{G}\}$, corresponding respectively to the split-and-average downstream protocol of RC parameter-sharing models (Schiff et al., 2024) and to post-hoc conjoining (Zhou et al., 2022). An invariant readout can measure the *magnitude* of the antisymmetric component but never its *sign*, as detailed in Appendix B.1.1.

Theorem 1 (Hierarchy)**.** *Assume equation 2 and position-symmetric pooling, so σ acts on pooled features by $(m_+, m_-) \mapsto (m_+, -m_-)$. Then every element of $\mathcal{F}_{\text{feat}}$ is a function of m_+ alone; every element of $\mathcal{F}_{\text{logit}}$ is an* even *function of m_-; the inclusion $\mathcal{F}_{\text{feat}} \subseteq \mathcal{F}_{\text{logit}}$ is strict whenever $\mathcal{G}$ contains a function depending non-trivially on $\|m_-\|$; and every element of both classes is G-invariant.*

2.5 DIRECTIONAL COLLAPSE

Theorem 2 (Directional collapse)**.** *Let $\hat{y}$ be induced by any G-invariant score function and let Assumptions 1–2 hold with deterministic labels. Then*

$$\mathrm{OPA}_A = 0 \quad \textit{exactly,} \qquad \mathrm{Rec}(y_{10}) + \mathrm{Rec}(y_{01}) \le 1, \qquad \mathrm{Acc} \le 1 - \tfrac{1}{2}\pi_A, \tag{9}$$

the last two bounds being tight. None of the three statements depends on the loss, the optimiser, the decoding rule, or the capacity of the readout.

Corollary 1 (Threshold-free collapse)**.** *Consider a binary directional task whose two labels form a single free orbit, under Assumptions 1–2 with deterministic labels. Then any G-invariant score function attains area under the receiver-operating-characteristic curve $\mathrm{AUC} = \frac{1}{2}$ exactly, with ties counted as one half, and $\mathrm{MCC} = 0$ exactly (Appendix B.1.9). In particular the collapse is visible without choosing a threshold or a decoding rule.*

The exact zero is the sharpest of the three statements; the other two are inequalities. The next lemma explains *why* an invariant model trained by gradient descent lands where it does.

Lemma 1 (Target symmetrisation)**.** *Let ℓ be any proper scoring rule and q any G-invariant probabilistic predictor. Under Assumptions 1–2, the risk of q against P equals its risk against the orbit-averaged target $\bar{P}(y \mid x) = \frac{1}{2}\big(P(y \mid x) + P(\mathrm{swap}\, y \mid x)\big)$. Hence if the invariant class contains $\bar{P}$, the unique minimiser satisfies $q_{y_{10}}(x) = q_{y_{01}}(x)$ for every x.*

An invariant readout therefore does not solve the stated task: it solves the task whose directional information has been averaged away. Failure is not evidence of underfitting, since the model optimises its own objective exactly. Whether *both* directional recalls vanish, rather than one reaching unity and the other zero, depends on the output parameterisation and decoding rule (Appendix B.1.4). Closed-form values of the multiclass MCC (Gorodkin, 2004) under collapse are given in Appendix B.1.9, together with the reason MCC *cannot* separate the two symmetry classes and is therefore reported only for comparability with prior work.

Proposition 1 (Exact signatures)**.** *Let τ be the fraction of free orbits on which the classifier predicts a label in A. Under Assumptions 1–2: if $\hat{y}$ is G-invariant then $\mathrm{IAR}_A = 1$ exactly, and*

$$\mathrm{EAR}_A = 1 - \tau, \qquad \mathrm{Rec}(y_{10}) + \mathrm{Rec}(y_{01}) = \tau = 1 - \mathrm{EAR}_A; \tag{10}$$

if $\hat{y}$ is ρ-equivariant then $\mathrm{EAR}_A = 1$ exactly, $\mathrm{IAR}_A = 1 - \tau$, and $\mathrm{Rec}(y_{10}) = \mathrm{Rec}(y_{01}) = \mathrm{OPA}_A$ exactly.

Each symmetry class is thus certified by *one* exact indicator, while the other merely re-encodes τ. We report the certifying indicator together with τ and note equation 10, rather than presenting the two rates as independent evidence. Identity equation 10 also lets us read the tightness of the recall bound directly off the audit table.

2.6 NON-IDENTIFIABILITY OF BACKBONE SYMMETRY

A benchmark score is a property of the composition $g \circ h$, not of only h. For a readout class $\mathcal{F}$ let $\Psi_{\mathcal{F}}(h) = \sup_{g \in \mathcal{F}} \mathrm{OPA}_A(g \circ h)$ be the best orbit-paired accuracy a benchmark could report for

backbone h, with $\mathcal{F}_{\text{inv}}$ and $\mathcal{F}_{\text{equi}}$ the classes equation 4 under position-symmetric pooling and q ranging over a universal approximator class.

Theorem 3 (Non-identifiability)**.** *Under Assumptions 1–2 with deterministic labels and $\pi_A > 0$, $\Psi_{\mathcal{F}_{\text{inv}}}(h) = 0$ for* every *backbone h: the functional is constant and carries no information about h. By contrast $\Psi_{\mathcal{F}_{\text{equi}}}$ is not constant, since there exist backbones h_1, h_2 with $\Psi_{\mathcal{F}_{\text{equi}}}(h_2) = 0 < \Psi_{\mathcal{F}_{\text{equi}}}(h_1)$.*

A benchmark run under a fixed invariant readout therefore cannot separate a backbone that never encoded the directional component from one that encodes it exactly, whereas the same benchmark run under the target-matched action can. Appendix B.1.6 gives the proof, the risk-level analogue, and the additional condition the latter requires. Theorem 3 applies directly to the common practice of fixing one lightweight prediction head across all models so as to make the comparison fair (Marin et al., 2024; Tang et al., 2025): holding the readout fixed removes one confounder while installing another. We test its operational consequence in Section 3.1: changing only the readout action can change the ordering assigned to the same set of backbones.

2.7 Localising the obstruction

The action equation 2 factors over two axes, and a readout can quotient each independently: channel averaging removes σ_c, position-symmetric pooling removes σ_L. Exact invariance requires both.

Proposition 2 (Two-axis annihilation)**.** *Let $\Phi_{c,\pi}$ be the feature map with channel treatment $c \in \{+, \text{full}\}$ (Π_+ versus retaining both halves) and pooling $\pi \in \{\text{sym}, \text{aware}\}$. Under equation 2,*

$$\begin{aligned} \Phi_{+,\text{sym}}(Rx) &= \Phi_{+,\text{sym}}(x), & \Phi_{+,\text{aware}}(Rx) &= \sigma_L \Phi_{+,\text{aware}}(x), \\ \Phi_{\text{full},\text{sym}}(Rx) &= \big(m_+(x), -m_-(x)\big), & \Phi_{\text{full},\text{aware}}(Rx) &= \big(\sigma_L m_+(x), -\sigma_L m_-(x)\big). \end{aligned} \tag{11}$$

Only the first is G-invariant. Hence a readout on $\Phi_{+,\text{sym}}$ cannot represent any function distinguishing y_{10} from y_{01}, whereas readouts on the other three are unrestricted. In particular, retaining channel averaging while replacing position-symmetric with position-aware pooling breaks exact invariance and admits $\text{OPA}_A > 0$.

Proposition 2 makes two falsifiable predictions that we test in Sections 3.2 and 3.3: directional information is linearly decodable from three of the four feature maps and at chance in exactly one (global-mean Π_+), and the failure is repairable by changing the pooling operator without abandoning the channel symmetry that motivates RC-equivariant backbones.

2.8 Generality and an *priori* diagnostic

Nothing above is specific to $\mathbb{Z}_2$ or to DNA. For a finite group G acting on $\mathcal{X}$ and on $\mathcal{Y}$ through a permutation representation ρ, any G-invariant classifier obeys the orbit-mode ceiling

$$\text{Acc} \;\le\; \mathbb{E}_{\mathcal{O}}\Big[\max_{c \in \mathcal{Y}} \big|\{x \in \mathcal{O} : y(x) = c\}\big| / |\mathcal{O}|\Big], \tag{12}$$

which equals $1 - \frac{1}{2}\pi_A$ for free $\mathbb{Z}_2$-orbits (Appendix B.1.8). Representation-theoretically, an invariant readout confines the attainable functions to the trivial isotypic component of ρ; any target with support on a non-trivial component is therefore unattainable. Section 3.2 tests this localisation by intervening on channel and positional symmetrisation.

Moreover, the cost of invariance is quantifiable before training. Define the *symmetrisation gap* $\Delta_G = \mathbb{E}_x[\max_y P(y \mid x) - \max_y \bar{P}(y \mid x)]$. For deterministic labels on free orbits of size two, $\Delta_G = \pi_A/2$ — exactly the accuracy gap of Theorem 2. Since Δ_G depends only on label statistics, it supports pre-registered predictions: invariant readouts are harmless iff $\Delta_G = 0$ (Section 3.1). One caveat: the theory requires ρ to be a genuine permutation of $\mathcal{Y}$. When it is not, e.g. reverse-complementing a splice donor does not yield an in-frame acceptor. π_A is undefined and only partial degradation is predicted; we flag such tasks rather than force them into the framework.

3 Empirical Results

We test the predictions of Section 2 on a strand-aware promoter benchmark built from K562 GRO-cap initiation calls and divergent TSS pairs (Core et al., 2014). Each 512-bp locus carries the two-

channel label $y = (a_{\rightarrow}, a_{\leftarrow})$ from equation 1; RC re-expresses the same locus in the opposite input orientation and therefore exchanges the channels exactly. The two binary outputs are decoded jointly as one of $\{y_{00}, y_{10}, y_{01}, y_{11}\}$, and the held-out evaluation set is closed under RC by pairing each x with Rx. Full benchmark construction and split controls are provided in Appendix A.

Across HyenaDNA, NTv3, and Caduceus, we compare the parameter-matched H_{inv}, H_{free}, and H_{equi} readouts, which respectively impose invariant, unconstrained, and target-matched equivariant output actions (Figure 1). These representative backbones were selected to cover models with and without an architecturally encoded RC action. *Adapted* means that the readout and the permitted backbone parameters are optimised jointly, whereas *frozen* means that only the readout is trained on the released backbone. Unless stated otherwise, we use global mean pooling and report mean $\pm$ SEM over five seeds, together with conventional task metrics, OPA_A, and the symmetry audits IAR_A and EAR_A. Model-specific adaptation and the operational readout implementation are detailed in Appendix A.

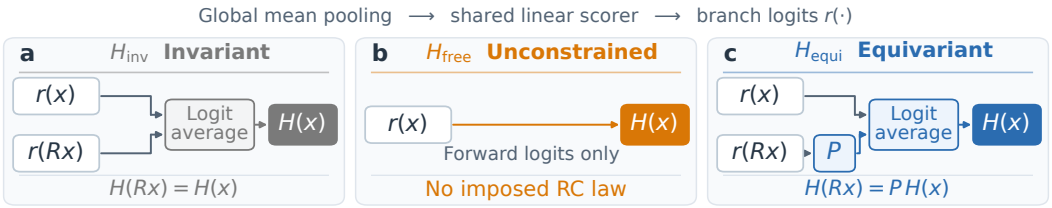


Figure 1: **Global-mean readout family.** Branch logits $r(\cdot)$ follow global mean pooling and a bias-free linear scorer shared within each readout. P swaps right/left logits before logit averaging. Neither averaging nor P adds learned parameters.

With global mean pooling, H_{inv} is tensor-level equivalent to the released Caduceus RCPS sequence-classification head, so the invariant arm of our family is not a constructed baseline but the deployed downstream protocol of an RC-equivariant foundation model (Appendix A). H_{free} and H_{equi} retain that head's scorer and stream handling and change only the output action.

3.1 Invariant readouts induce an exact collapse hidden by aggregate metrics

Every invariant readout failed to classify both members of a directional orbit correctly. Across three backbones, two backbone-training regimes, and five seeds (30 independently trained, mean-pooled H_{inv} models), $\text{OPA}_A = 0$ and $\text{IAR}_A = 1$ (Table 1). Moreover, for every predicted state $c \in \mathcal{Y}$, the orbit-closed confusion counts satisfy $N(y_{10} \to c) = N(y_{01} \to c)$, with zero discrepancy in all 30 runs. This integer identity verifies both orbit closure and the realised decision-level invariance. The full class-recall decomposition for all three readouts, together with the invariant integer orbit audit and action-specific logit-level residuals, is reported in Appendix Table A1.

Table 1: Mean-pooled invariant-readout audit (five-seed mean $\pm$ SEM with rounded). MCC and directional prediction rate τ vary; $\text{OPA}_A = 0$, $\text{IAR}_A = 1$ and identical directional-row counts hold exactly in every run.

Backbone	Regime	MCC	τ	OPA_A	IAR_A
Caduceus	Adapted	0.199±0.011	0.545±0.036	0.000	1.000
	Frozen	0.094±0.005	0.115±0.029	0.000	1.000
HyenaDNA	Adapted	0.257±0.006	0.110±0.028	0.000	1.000
	Frozen	0.157±0.006	0.319±0.018	0.000	1.000
NTv3	Adapted	0.242±0.013	0.246±0.036	0.000	1.000
	Frozen	0.163±0.003	0.318±0.018	0.000	1.000

Conventional metrics obscure this categorical failure. In the adapted regime, state MCC remains between 0.199 and 0.257 (Table 1). These values reflect performance on the fixed-label states and on at most one orientation of each free orbit; they do not alter the exact zero in paired success. An MCC-only comparison would therefore reduce a difference in attainable function class to an ordinary difference in degree.

Nor is the zero explained by a failure to emit directional labels. Across backbones and training regimes, the directional prediction rate τ varies substantially, and several conditions attain non-zero recall on both orientations, whereas OPA_A remains zero (Appendix Table A1). The two recalls sum to τ, as required by Proposition 1, but invariance assigns the same prediction to both members of each orbit and prevents joint correctness. Aggregate scores and prediction marginals thus depend on the backbone and its adaptation, whereas the orbit-level obstruction is fixed by the readout.

Theorem 2 predicts this failure before fitting: the 882 directional sequences among 1,776 held-out examples give $\pi_A \approx 0.4966$, $\Delta_G = \pi_A/2 \approx 0.2483$, and an invariant accuracy ceiling of $1 - \Delta_G \approx 0.7517$. This function-class limit requires failure on at least one member of every free orbit, regardless of optimisation.

3.2 The obstruction lies in the readout and requires joint channel and positional symmetrisation

The exact zero in Table 1 could reflect either an absence of directional information in the backbone or its removal by the readout. We froze the released Caduceus backbone, discarded its task readout, and fitted the same linear probe to four fixed representations crossing Π_+ or Π_- with global-mean or 16-bin position-aware pooling. Only jointly symmetrised global-mean Π_+ is at chance; retaining either channel antisymmetry or positional order permits directional decoding (Figure 2a).

This pattern persists in backbones obtained after end-to-end adaptation with global-mean H_{inv}: its deployed representation remains uninformative, while the discarded Π_- and position-aware Π_+ branches remain predictive (Figure 2b). The failure is thus localised to their joint symmetrisation at readout, not to a backbone that never encoded direction.

Across adaptation actions, all three informative probe AUCs were lowest after H_{inv}, intermediate after H_{free}, and highest after H_{equi} (Figure 2d). The free action exceeded the invariant action in every paired seed; the equivariant action exceeded the free action in four of five seeds for each probe. Because a linear probe measures accessibility rather than total information, this supports attenuation—not proof of erasure—under the mismatched action.

The probe objective and optimisation protocol, complete train/validation/test metrics, and a symmetric-label positive control are provided in Appendix A.2 and Tables A2–A3.

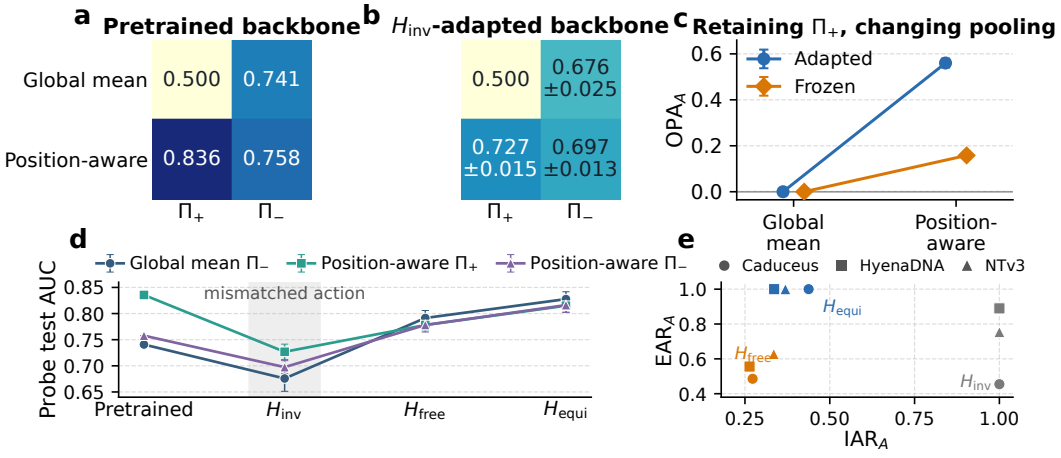


Figure 2: **Two-axis localisation of directional collapse in Caduceus.** (a,b) Right-versus-left probe AUC after crossing channel projection with pooling for the released and H_{inv}-adapted backbones. (c) Replacing global mean by ordered pooling makes OPA_A positive. (d) Informative-probe trajectories across backbone sources. (e) Decision-level action audit for adapted models (colour: readout; marker: backbone). Error bars denote SEM over five seeds; pretrained results use one released checkpoint.

In the corresponding readout intervention, retaining Π_+ but replacing global mean with ordered pooling removed the structural zero in both frozen and adapted settings (Figure 2c). Positional order

therefore exposes direction already present in the channel-symmetric representation, identifying the failure with the joint quotient rather than the RC-equivariant backbone. Complete position-aware results appear in Appendix Table A4.

3.3 A SYMMETRY-MATCHED OUTPUT ACTION REPAIRS THE COLLAPSE ACROSS BACKBONES

Breaking the joint quotient is sufficient to expose direction, but it leaves open how predictions should transform under RC. Here the target supplies that transformation law: the two directional outputs swap. We therefore asked whether encoding this output action is preferable to leaving the readout unconstrained. Under global mean pooling, we compared the parameter-matched family in equation 5; for each backbone, the three readouts share a bias-free scorer and differ only in the fixed action applied to its outputs.

Across adapted models, the orbit audit recovered the intended symmetry class (Table 2; Figure 2e). H_{inv} remained at $\mathrm{OPA}_A = 0$ with exact invariant agreement, H_{equi} attained exact equivariant agreement and positive paired accuracy, and H_{free} occupied an interior regime. Moreover, H_{equi} exceeded H_{free} in both paired accuracy and state macro-F1 on all three backbones (Table 2), showing a benefit from matching the target action beyond merely removing the invariance constraint. Exact equivariance also made correctness orbit-paired: in every H_{equi} condition, $\mathrm{Rec}(y_{10}) = \mathrm{Rec}(y_{01}) = \mathrm{OPA}_A$ (Appendix Table A1).

Table 2: Global-mean directional-promoter results (five-seed mean ± SEM, rounded to three decimals). Bold marks the best MCC, F1 and OPA per backbone/regime, including displayed ties. MCC and F1 are four-state metrics. Exact structural entries omit SEM; IAR/EAR are not ranked.

Backbone	Readout	MCC	Macro-F1	RC F1	OPA_A	IAR_A	EAR_A
Adapted backbone							
Caduceus	H_{inv}	0.199±0.011	0.390±0.006	0.394±0.005	0.000	1.000	0.455±0.036
	H_{free}	0.293±0.004	0.468±0.003	0.477±0.007	0.263±0.010	0.273±0.010	0.486±0.004
	H_{equi}	**0.374±0.012**	**0.528±0.009**	**0.528±0.009**	**0.471±0.011**	0.438±0.010	1.000
HyenaDNA	H_{inv}	0.257±0.006	0.312±0.008	0.320±0.007	0.000	1.000	0.890±0.028
	H_{free}	0.381±0.007	0.531±0.007	0.531±0.006	0.327±0.024	0.263±0.014	0.556±0.022
	H_{equi}	**0.442±0.006**	**0.581±0.004**	**0.581±0.004**	**0.590±0.009**	0.336±0.009	1.000
NTv3	H_{inv}	0.242±0.013	0.340±0.006	0.343±0.006	0.000	1.000	0.754±0.036
	H_{free}	0.347±0.004	0.503±0.004	0.508±0.005	0.288±0.009	0.335±0.006	0.628±0.002
	H_{equi}	**0.399±0.008**	**0.548±0.006**	**0.548±0.006**	**0.542±0.009**	0.369±0.011	1.000
Frozen backbone							
Caduceus	H_{inv}	0.094±0.005	0.240±0.005	0.231±0.002	0.000	1.000	0.885±0.029
	H_{free}	**0.105±0.002**	**0.249±0.003**	**0.249±0.001**	0.001±0.001	0.805±0.005	0.791±0.011
	H_{equi}	**0.105±0.002**	**0.249±0.002**	**0.249±0.002**	**0.051±0.002**	0.917±0.003	1.000
HyenaDNA	H_{inv}	0.157±0.006	0.307±0.004	0.307±0.001	0.000	1.000	0.681±0.018
	H_{free}	0.174±0.004	0.366±0.003	0.368±0.001	0.143±0.004	0.425±0.005	0.616±0.008
	H_{equi}	**0.185±0.001**	**0.370±0.001**	**0.370±0.001**	**0.229±0.003**	0.641±0.004	1.000
NTv3	H_{inv}	0.163±0.003	0.310±0.002	0.319±0.003	0.000	1.000	0.682±0.018
	H_{free}	**0.184±0.002**	**0.365±0.002**	0.353±0.003	0.123±0.005	0.470±0.009	0.648±0.008
	H_{equi}	0.177±0.002	0.360±0.001	**0.360±0.001**	**0.214±0.003**	0.665±0.005	1.000

In the adapted regime, matching the output action increased OPA_A over H_{free} by 0.208, 0.263, and 0.254 for Caduceus, HyenaDNA, and NTv3, respectively; state macro-F1 simultaneously increased by 0.060, 0.050, and 0.045. Thus the exact action audit is accompanied by a consistent predictive gain rather than being satisfied only as a formal constraint.

Across all 60 constrained runs, the prescribed decision action held exactly; Caduceus and NTv3 also had zero action-matched logit residual, while bfloat16-loaded HyenaDNA deviated by at most 9.77×10^{-4} without changing any decoded prediction (Appendix Table A1). The unconstrained H_{free} instead occupied backbone-dependent interior values in Figure 2e.

To separate the readout effect from representation adaptation, we froze each backbone and varied only the output action. On Caduceus, state macro-F1 barely changes from 0.240 to 0.249, yet OPA_A rises from 0.000 to 0.051; aggregate performance therefore almost hides a qualitative repair. The ordering of the reported macro-F1 means also changes from NTv3 > HyenaDNA > Caduceus under H_{inv} to HyenaDNA > NTv3 > Caduceus under H_{equi} (Table 2). Because the features are fixed, this reordering is readout-induced. End-to-end adaptation strengthens it into a complete reversal, oper-

ationalising Theorem 3: benchmark conclusions about a backbone depend on the readout through which it is observed.

3.4 THE OBSTRUCTION EXTENDS BEYOND DNA

We separated the group action from genomic semantics with two CPU-scale controls. In an orbit-closed synthetic $\mathbb{Z}_2$ task, varying π_A from 0 to 1 made the invariant readout follow the predicted ceiling $1 - \pi_A/2$ exactly and kept $\mathrm{OPA}_A = 0$ for every $\pi_A > 0$; the parameter-matched free and equivariant readouts attained unit accuracy and OPA (Appendix Figure 4).

In a real-data control, horizontal reflection moved each MNIST digit (LeCun et al., 1998) between the left and right sides of a canvas and exchanged its label. On this orbit-closed task, the invariant model retained macro-F1 0.497 ± 0.001 but had zero OPA in every seed; free and equivariant readouts solved every held-out orbit (Table 3; Appendix Figure 3).

Table 3: MNIST left–right control (five-seed mean ± SEM; rounded to three decimals).

Readout	Accuracy	Macro-F1	OPA_A	IAR_A	EAR_A
H_{inv}	0.500	0.497 ± 0.001	0.000	1.000	0.000
H_{free}	1.000	1.000	1.000	0.000	1.000
H_{equi}	1.000	1.000	1.000	0.000	1.000

Although the free readout learned equivariant decisions here, its action-matched logit residual was 3.19 ± 1.01, versus exactly zero for H_{equi}; paired predictions alone do not certify the score action.

3.5 A MATCHED INVARIANT TASK EXHIBITS NO STRUCTURAL OBSTRUCTION

H3K27me3 prediction from the revised Nucleotide Transformer benchmark (Dalla-Torre et al., 2025) supplies a matched control with $y(Rx) = y(x)$. Using global-mean Caduceus, we compared H_{inv} and H_{free} with adapted and frozen backbones; pair accuracy records orbits for which both x and Rx are correct.

Here H_{inv} showed no paired collapse: it attained exact invariant agreement, identical original/RC performance, and pair accuracy equal to single-view accuracy in both regimes (Table 4). H_{free} improved adapted original-view macro-F1, but not RC-view or paired performance; when frozen, its macro-F1 was similar while its pair accuracy remained lower.

Table 4: Caduceus on the RC-invariant H3K27me3 control (five-seed mean ± SEM; rounded to three decimals). A displayed 0.000 SEM may be non-zero before rounding.

Regime	Readout	MCC	Macro-F1	RC F1	Accuracy	Pair Acc.	IAR	δ_{inv}
Adapted	H_{inv}	0.550 ± 0.005	0.761 ± 0.004	0.761 ± 0.004	0.766 ± 0.003	0.766 ± 0.003	1.000	0.000
	H_{free}	0.596 ± 0.004	0.784 ± 0.002	0.685 ± 0.003	0.788 ± 0.002	0.600 ± 0.006	0.726 ± 0.006	5.414 ± 0.217
Frozen	H_{inv}	0.457 ± 0.002	0.716 ± 0.005	0.716 ± 0.005	0.722 ± 0.003	0.722 ± 0.003	1.000	0.000
	H_{free}	0.438 ± 0.002	0.715 ± 0.001	0.711 ± 0.001	0.717 ± 0.001	0.652 ± 0.000	0.874 ± 0.002	0.621 ± 0.222

Thus, the same readout that cannot solve free directional orbits is well specified on an RC-fixed target. The issue is not invariance itself, but whether the readout action matches the label action.

4 DISCUSSION

Benchmark scores characterise a backbone–readout composition, not the backbone alone. Our interventions expose how a mismatched readout can hide directional information behind non-zero aggregate scores; matching the target action restores orbit-level success. Evaluation should therefore specify input and target actions and audit paired predictions alongside conventional metrics.

The promoter benchmark, invariant genomic control and two non-genomic tasks test this failure across biological and non-biological settings. Broader scientific domains and approximate or continuous symmetries remain to be examined. Symmetry is a contract among input action, target action and readout.

AI USE STATEMENT

Generative AI tools were used to: (1) to aid and polish the writing, including improving grammar, clarity, and phrasing; (2) for retrieval and discovery, such as searching for and identifying related work; and (3) to assist with research execution, specifically in writing and debugging code for our experiments. All scientific judgements, claims, and final wording were reviewed and decided by the authors. The authors take full responsibility for the final content.

ETHICS STATEMENT

All experiments use publicly available reference-genome sequence and functional-genomics measurements from established cell lines. No new human participants were recruited, and no individual-level or identifiable genomic data were analysed. The benchmark is intended for methodological research, not clinical decision making. Released code and derived split metadata will respect the licences and access terms of the original resources.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs are provided in Appendix B; Appendix A specifies benchmark construction, leakage-controlled splits, backbone adaptation, optimisation, readout and probe definitions, evaluation metrics, and complete results. The anonymised supplementary material accompanying this submission provides code for benchmark construction, training, evaluation, probing, and figure generation; fixed split manifests and configuration files; per-seed metric tables; and software environment specifications. Exact symmetry claims are accompanied by both decision-level and logit-level numerical audits.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K. Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, Javier Mendoza-Revilla, Fatimah Al-Ani, Evan Seitz, Yair Schiff, Yohan Bornachot, Ariana Hernandez, Marie Lopez, Alexandre Laterre, Karim Beguir, Peter Koo, Volodymyr Kuleshov, Alexander Stark, Bernardo P. de Almeida, and Thomas Pierrot. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.22.695963.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016.

Leighton J. Core, André L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel, and John T. Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, 46:1311–1320, 2014. doi: 10.1038/ng.3142.

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, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287–297, 2025. doi: 10.1038/s41592-024-02523-z.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969. PMLR, 2021.

Jan Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5-6):367–374, 2004. doi: 10.1016/j.compbiolchem.2004.09.006.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

Mahmoud M. Ibrahim, Aslihan Karabacak, Alexander Glahs, Ena Kolundzic, Antje Hirsekorn, Alexa Carda, Baris Tursun, Robert P. Zinzen, Scott A. Lacadie, and Uwe Ohler. Determinants of promoter and enhancer transcription directionality in metazoans. *Nature Communications*, 9: 4472, 2018. doi: 10.1038/s41467-018-06962-z.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2747–2755. PMLR, 10–15 Jul 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Vincent Mallet and Jean-Philippe Vert. Reverse-complement equivariant networks for DNA sequences. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13511–13523, 2021.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen A. Baccus. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201, 2023. doi: 10.52202/075280-1872.

Yair Schiff, Chia Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43632–43648. PMLR, 2024.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Reverse-complement parameter sharing improves deep learning models for genomics. *bioRxiv*, pp. 103663, 2017. doi: 10.1101/103663.

Mouad Talhi, Arne Wolf, and Anthea Monod. Breaking symmetry bottlenecks in GNN readouts. *arXiv preprint arXiv:2602.05950*, 2026. doi: 10.48550/arXiv.2602.05950.

Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K Koo. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. *Genome Biology*, 26(1):203, 2025. doi: 10.1186/s13059-025-03674-8.

Dian Wang, Xupeng Zhu, Jung Yeon Park, Mingxi Jia, Guanang Su, Robert Platt, and Robin Walters. A general theory of correct, incorrect, and extrinsic equivariance. *Advances in Neural Information Processing Systems*, 36:40006–40029, 2023.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23078–23091. PMLR, 2022.

Hannah Zhou, Avanti Shrikumar, and Anshul Kundaje. Towards a better understanding of reverse-complement equivariance for deep learning models in genomics. In *Proceedings of the 16th Machine Learning in Computational Biology Meeting*, volume 165 of *Proceedings of Machine Learning Research*, pp. 1–33. PMLR, 2022.

APPENDIX

A EXPERIMENTAL DETAILS AND COMPLETE RESULTS

Benchmark construction. We built the benchmark in GRCh37/hg19 genomic coordinates from the official strand-specific GRO-cap transcription-start-site (TSS) calls and divergent TSS pairs released for K562 (Core et al., 2014). Positive anchors were required to lie within 250 bp of a GENCODE v19 promoter locus. A single unpaired plus- or minus-strand call defined the rightward- or leftward-only state, respectively; the bidirectional state required an official divergent pair, and loci with unpaired evidence on both strands were excluded as ambiguous. Each anchor was centred in a 512-bp sequence retained in genomic orientation, so RC exchanges the two directional channels by definition.

Backgrounds were GC-content-matched hard negatives drawn from promoter-distal DNase-accessible regions without nearby GRO-cap or annotated TSS evidence and from GENCODE promoter candidates without a GRO-cap call in either K562 or GM12878. ENCODE blacklist intervals and sequences containing more than 1% ambiguous bases were removed. Before splitting, overlapping 512-bp windows and exact sequence duplicates were assigned to common leakage groups. A stratified group split of the approximately class-balanced K562 pool yielded 9,345 training, 1,187 validation, and 1,776 held-out test sequences; the main experiments use this within-cell-line test set. Canonical sequence/RC keys confirmed zero RC-orbit overlap between any pair of splits.

Backbones and optimisation. We used HyenaDNA-small-32k (Nguyen et al., 2023), the pretrained NTv3-8M checkpoint (Boshar et al., 2025), and Caduceus-PS-131k ($d_{\text{model}} = 256$, 16 layers) (Schiff et al., 2024). Each 512-bp sequence was centred in a 1,024-nt input and symmetrically padded with N; direct single-nucleotide tokenisation omitted special tokens so that tokenisation itself preserved the RC action. For NTv3, the representation extractor used the embedding stem, the first convolutional-tower block, and the Transformer torso, with average-pooling downsampling disabled; this same extractor was used for every readout and training regime.

End-to-end HyenaDNA and NTv3 runs used LoRA with rank 16, scaling parameter $\alpha = 32$, dropout 0.15, and no bias. Adapters targeted `in_proj`, `out_proj`, `fc1`, and `fc2` in HyenaDNA, and `linear`, `mha_output`, `fc1`, and `fc2` in NTv3; Caduceus instead updated all backbone weights. Frozen controls attached no adapters and optimised only the readout. We trained with two-channel binary cross-entropy and AdamW for 30 epochs, using backbone/readout learning rates of $2 \times 10^{-4}/5 \times 10^{-4}$, weight decay 0.05, batches of 16 with four-step gradient accumulation, and cosine decay after 10% linear warmup. The checkpoint with the highest validation state macro-F1 was retained.

Operational readout implementation. The theory writes a four-state score for clarity, whereas the implementation emits two Bernoulli logits, one per directional channel. All reported heads follow the split–restore–pool–score structure of the released Caduceus RCPS sequence-classification head: paired streams are placed in a common orientation, pooled independently, and passed through one shared `Linear(D, 2, bias=False)` scorer. With global mean pooling, H_{inv} is tensor-level equivalent to the native Caduceus head; H_{free} and H_{equi} retain the same scorer and stream-handling protocol while changing only the final output action. We therefore refer to this parameter-matched family as *official-derived*, rather than claiming that all three variants are unmodified repository heads.

Writing $r(x) = q(\text{Pool}(h(x)))$, we use $H_{\text{free}}(x) = r(x)$, $H_{\text{inv}}(x) = \frac{1}{2}[r(x) + r(Rx)]$, and $H_{\text{equi}}(x) = \frac{1}{2}[r(x) + Pr(Rx)]$; when the backbone carries the RC action, these are exactly Equation equation 5. HyenaDNA and NTv3 obtain the ordered pair by evaluating the shared backbone on x and Rx; Caduceus exposes the corresponding native RC-related streams in one forward pass, which we restore to a common channel convention before applying the same readouts. Componentwise thresholding maps the two logits to $y_{00}, y_{10}, y_{01}, y_{11}$ and commutes with the channel swap P. Thus the two-logit parameterisation carries the same label action as Equation equation 1, and the decision-level collapse and orbit diagnostics apply verbatim. Training uses the mean of the two channelwise binary cross-entropies; because binary cross-entropy is strictly proper for each Bernoulli marginal, the target-symmetrisation argument also applies componentwise.

Non-genomic controls. The synthetic sweep used 1,000 training and 1,000 test orbits at each of 11 equally spaced values of π_A, with seeds 40–44. Eight-dimensional inputs contained separately transformed even and odd coordinates, and R negated only the odd coordinates. Labels comprised two fixed states and one exchanged directional pair. A shared bias-free linear scorer was trained for 450 full-batch Adam steps at learning rate 0.05; the invariant, free, and equivariant constructions differed only by the output action.

For the MNIST-derived control, we repurposed the official images to isolate a known reflection action rather than evaluate standard digit classification. Digit identity was ignored and the target recorded only whether the digit occupied the left or right side of a 28×64 canvas. Each input was paired with its horizontal reflection in both training and evaluation. We used 10,000 base images from the official training split and 2,000 orbit representatives from the official test split, preserving the official split provenance. The shared scorer was an MLP with one 64-unit GELU hidden layer and two output logits (114,882 trainable parameters), trained for eight epochs with AdamW, learning rate 2×10^{-3}, weight decay 10^{-4}, and batch size 256. All runs used CPU execution and seeds 40–44. The logit audits were computed on the paired official test subset.

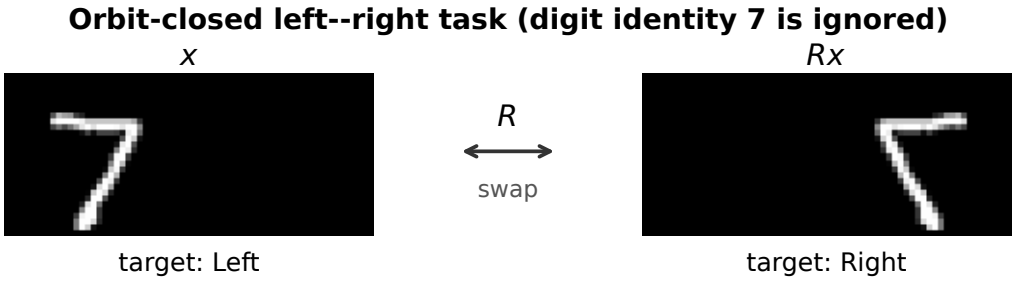


Figure 3: **MNIST orbit construction.** Digit identity is ignored; horizontal reflection moves the digit across the canvas and exchanges the left/right target. Every training and test representative is accompanied by its reflected partner.

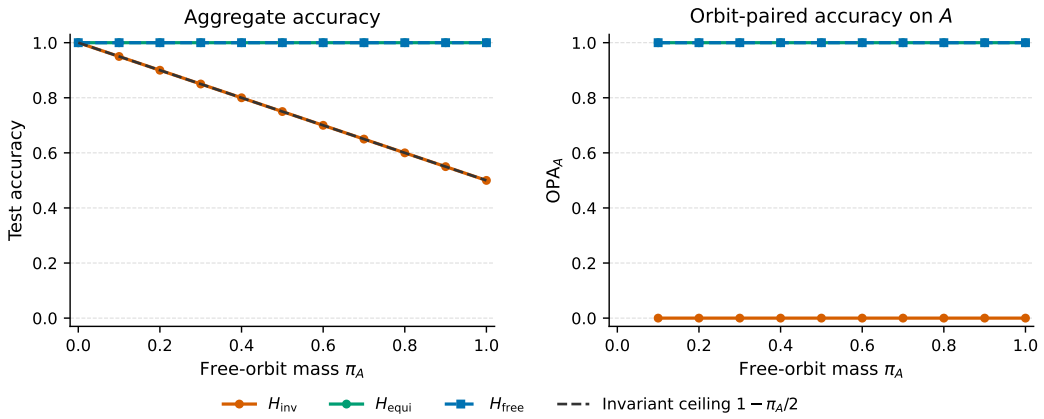


Figure 4: **Synthetic free-orbit-mass sweep.** Test accuracy (left) and orbit-paired accuracy on the exchanged-label subset (right) for parameter-matched invariant, free, and equivariant linear readouts. Curves show means over five seeds; error bars are smaller than the markers. The invariant curve coincides with $1 - \pi_A/2$ and has $\mathrm{OPA}_A = 0$ for every $\pi_A > 0$.

A.1 Global-mean directional benchmark

Throughout the appendix, *adapted backbone* denotes the jointly trained setting defined in Section 3, whereas *frozen backbone* denotes the released pretrained checkpoint with only the readout optimised. For every original held-out test sequence x, evaluation appends Rx and applies the label

action in equation 1. "Original view" and "RC view" refer to evaluation on x and on Rx with its transformed label, respectively. Conventional metrics are evaluated on individual views, while pair metrics are evaluated once per orbit. Table 2 reports the complete parameter-matched readout comparison under global mean pooling in the main text. Table A1 reports class-specific recalls and decision-level symmetry for all three readouts. We do not bold columnwise maxima in these diagnostic tables: IAR and EAR audit different transformation laws rather than forming a common monotone performance objective, and absolute performance is not the central comparison.

Table A1: Complete directional-recall and decision-level symmetry audit under global mean pooling. All three parameter-matched readouts are reported; τ is the directional prediction rate. Values are mean ± SEM over five seeds unless exact. Values are rounded to three decimals.

Backbone	Readout	$\mathrm{Rec}_{\rightarrow}$	$\mathrm{Rec}_{\leftarrow}$	τ	OPA_A	IAR_A	EAR_A
Adapted backbone							
Caduceus	H_{inv}	0.276±0.047	0.269±0.028	0.545±0.036	0.000	1.000	0.455±0.036
	H_{free}	0.460±0.030	0.449±0.019	0.618±0.024	0.263±0.010	0.273±0.010	0.486±0.004
	H_{equi}	0.471±0.011	0.471±0.011	0.562±0.010	0.471±0.011	0.438±0.010	1.000
HyenaDNA	H_{inv}	0.076±0.040	0.034±0.023	0.110±0.028	0.000	1.000	0.890±0.028
	H_{free}	0.551±0.024	0.446±0.045	0.599±0.017	0.327±0.024	0.263±0.014	0.556±0.022
	H_{equi}	0.590±0.009	0.590±0.009	0.664±0.009	0.590±0.009	0.336±0.009	1.000
NTv3	H_{inv}	0.215±0.060	0.030±0.030	0.246±0.036	0.000	1.000	0.754±0.036
	H_{free}	0.429±0.014	0.425±0.013	0.515±0.010	0.288±0.009	0.335±0.006	0.628±0.002
	H_{equi}	0.542±0.009	0.542±0.009	0.631±0.011	0.542±0.009	0.369±0.011	1.000
Frozen backbone							
Caduceus	H_{inv}	0.115±0.029	0.000	0.115±0.029	0.000	1.000	0.885±0.029
	H_{free}	0.107±0.013	0.014±0.004	0.102±0.008	0.001±0.001	0.805±0.005	0.791±0.011
	H_{equi}	0.051±0.002	0.051±0.002	0.083±0.003	0.051±0.002	0.917±0.003	1.000
HyenaDNA	H_{inv}	0.187±0.077	0.132±0.082	0.319±0.018	0.000	1.000	0.681±0.018
	H_{free}	0.232±0.018	0.260±0.024	0.405±0.007	0.143±0.004	0.425±0.005	0.616±0.008
	H_{equi}	0.229±0.003	0.229±0.003	0.359±0.004	0.229±0.003	0.641±0.004	1.000
NTv3	H_{inv}	0.187±0.078	0.131±0.080	0.318±0.018	0.000	1.000	0.682±0.018
	H_{free}	0.212±0.025	0.210±0.015	0.349±0.005	0.123±0.005	0.470±0.009	0.648±0.008
	H_{equi}	0.214±0.003	0.214±0.003	0.335±0.005	0.214±0.003	0.665±0.005	1.000

Across all 30 H_{inv} runs and all 30 H_{equi} runs, the prescribed decision-level action held exactly, and the invariant confusion-row discrepancy was zero in every H_{inv} run. Caduceus and NTv3 had zero action-matched logit residual in every constrained run. The only nonzero residuals occurred for HyenaDNA, which our loader instantiated with bfloat16 weights and evaluated under bfloat16 autocast; their binary-quantised magnitudes were at most 9.77×10^{-4}, consistent with finite-precision rounding, and did not alter any decoded prediction or decision-level identity.

A.2 Linear-probe specification and controls

We distinguish the *backbone source* from the probe fitted to it. *Pretrained* uses the released Caduceus checkpoint before training on the directional-promoter benchmark. *Adapted* H_ρ uses the backbone weights from the validation-selected end-to-end checkpoint trained jointly with readout H_ρ on that benchmark. For every source, the trained task readout is discarded, the backbone is frozen, and a newly initialised linear probe is fitted; the source labels therefore identify which backbone checkpoint is being probed, not different probe architectures.

Let $\phi(x)$ denote one of the fixed probe representations obtained by crossing Π_+ or Π_- with global-mean or ordered 16-bin pooling. Each probe uses an orbit-closed binary design containing both x and Rx. The target is $z = 1$ for rightward and $z = 0$ for leftward promoters, and is flipped for the RC member. If D is the diagonal matrix of feature standard deviations computed on the training design, we divide by D without centring and fit the zero-intercept logistic probe using binary cross-entropy (BCE),

$$p_w(z = 1 \mid x) = \sigma\big(w^\top D^{-1}\phi(x)\big), \qquad \widehat{w} = \arg\min_w \frac{1}{N}\sum_{i=1}^{N} \mathrm{BCE}(z_i, p_w(x_i)) + \frac{\|w\|_2^2}{2CN}, \quad (13)$$

with $C = 1$. No intercept or feature centring is used, preserving the exact sign action of Π_-. Ordered pooling concatenates masked means from 16 contiguous bins, preserving their left-to-right order. Probes are fitted with single-precision PyTorch L-BFGS, strong-Wolfe line search, a maximum of 5,000 iterations, gradient tolerance 10^{-5}, and parameter-change tolerance 10^{-9}; all reported fits terminated before the iteration cap.

For the right-versus-left probe, Test OPA applies equation 6 to the two binary predictions on each (x, Rx) pair. For the symmetric-label control, paired accuracy instead requires both fixed-label members of the pair to be correct.

Table A2 reports the full directional probe audit for the pretrained backbone and for backbones adapted with the global-mean H_{inv}, H_{free}, and H_{equi} readouts. As a pipeline control, we apply the same position-aware Π_+ features and optimisation to background-versus-bidirectional prediction, where the binary label is fixed rather than swapped under RC (Table A3).

Table A2: Complete right-versus-left linear-probe audit for Caduceus. Pretrained probes the released checkpoint; Adapted H_ρ probes the backbone extracted from the validation-selected end-to-end checkpoint trained with H_ρ. In every row the task readout is discarded, the source backbone is frozen, and a fresh zero-intercept linear probe is fitted. The pretrained row has one source checkpoint, whereas adapted rows aggregate five readout-training seeds. SD(p) is the standard deviation of test probabilities and verifies that the chance-level global-mean Π_+ probe is constant rather than merely misranked. Values are rounded to three decimals. Rounding to 0.000 SEM or 1.000 scores does not imply exact equality.

Backbone source	Pool	Rep.	Train AUC	Val AUC	Test AUC	Test MCC	Test OPA	SD(p)
Pretrained	Mean	Π_+	0.500	0.500	0.500	0.000	0.000	0.000
	Mean	Π_-	0.799	0.739	0.741	0.356	0.678	0.252
	16-bin	Π_+	0.986	0.831	0.836	0.514	0.755	0.424
	16-bin	Π_-	0.980	0.781	0.758	0.397	0.698	0.409
Adapted H_{inv}	Mean	Π_+	0.500	0.500	0.500	0.000	0.000	0.000
	Mean	Π_-	0.745±0.010	0.694±0.023	0.676±0.025	0.237±0.032	0.618±0.016	0.210±0.010
	16-bin	Π_+	0.918±0.008	0.734±0.015	0.727±0.015	0.333±0.019	0.666±0.010	0.355±0.009
	16-bin	Π_-	0.924±0.004	0.718±0.018	0.697±0.013	0.290±0.017	0.645±0.008	0.359±0.011
Adapted H_{free}	Mean	Π_+	0.500	0.500	0.500	0.000	0.000	0.000
	Mean	Π_-	0.995±0.002	0.789±0.009	0.791±0.014	0.424±0.024	0.712±0.012	0.431±0.010
	16-bin	Π_+	1.000±0.000	0.778±0.011	0.778±0.011	0.405±0.015	0.702±0.007	0.456±0.005
	16-bin	Π_-	1.000±0.000	0.773±0.008	0.778±0.014	0.417±0.021	0.708±0.011	0.463±0.007
Adapted H_{equi}	Mean	Π_+	0.500	0.500	0.500	0.000	0.000	0.000
	Mean	Π_-	0.991±0.004	0.834±0.014	0.828±0.014	0.505±0.023	0.752±0.012	0.422±0.010
	16-bin	Π_+	0.998±0.001	0.823±0.011	0.815±0.012	0.482±0.021	0.741±0.011	0.444±0.007
	16-bin	Π_-	0.998±0.001	0.828±0.011	0.816±0.014	0.497±0.020	0.749±0.010	0.448±0.007

Table A3: Symmetric-label positive control using the same position-aware Π_+ features and optimisation pipeline. Backbone-source conventions follow Table A2. The probe distinguishes background from bidirectional promoters, whose labels are fixed under RC. Pair Acc. requires both members of an RC pair to be correct. Its non-chance performance rules out a degenerate feature-extraction or probe-fitting pipeline. Values are rounded to three decimals. Rounding to 0.000 SEM or 1.000 scores does not imply exact equality.

Backbone source	Train AUC	Val AUC	Test AUC	Test MCC	Pair Acc.	SD(p)
Pretrained	0.990	0.876	0.875	0.620	0.810	0.435
Adapted H_{inv}	1.000	0.846±0.008	0.851±0.008	0.538±0.015	0.767±0.008	0.471±0.004
Adapted H_{free}	1.000±0.000	0.843±0.006	0.859±0.002	0.549±0.010	0.774±0.005	0.473±0.003
Adapted H_{equi}	0.996±0.002	0.857±0.008	0.859±0.009	0.584±0.022	0.791±0.011	0.450±0.008

A.3 Complete position-aware intervention

Table A4 extends the Caduceus intervention in Figure 2c to every backbone and readout under both training regimes. Because RC reverses the ordered bins, the Π_+ construction labelled H_{inv} for engineering consistency is not an invariant classifier in this setting; the table therefore audits its realised action rather than assuming invariance from its name.

Table A4: Complete ordered 16-bin readout results on the directional promoter benchmark (mean ± SEM over five seeds). MCC and Macro-F1 are four-state metrics on the original view; RC F1 is four-state macro-F1 on the reverse-complement view. Under position-aware pooling, the engineering name H_{inv} denotes the Π_+ channel projection but no longer defines an invariant classifier because RC reverses the ordered bins. Values are rounded to three decimals.

Backbone	Readout	MCC	Macro-F1	RC F1	OPA_A	IAR_A	EAR_A
Adapted backbone							
Caduceus	H_{inv}	0.461 ± 0.009	0.594 ± 0.007	0.596 ± 0.006	0.560 ± 0.019	0.327 ± 0.022	0.947 ± 0.005
	H_{free}	0.418 ± 0.003	0.560 ± 0.003	0.564 ± 0.006	0.397 ± 0.011	0.280 ± 0.015	0.674 ± 0.013
	H_{equi}	0.464 ± 0.009	0.597 ± 0.007	0.597 ± 0.007	0.565 ± 0.010	0.366 ± 0.008	1.000
HyenaDNA	H_{inv}	0.453 ± 0.008	0.588 ± 0.006	0.588 ± 0.006	0.534 ± 0.010	0.282 ± 0.009	0.859 ± 0.009
	H_{free}	0.420 ± 0.008	0.563 ± 0.007	0.549 ± 0.005	0.400 ± 0.008	0.233 ± 0.008	0.613 ± 0.012
	H_{equi}	0.459 ± 0.008	0.593 ± 0.006	0.593 ± 0.006	0.598 ± 0.010	0.334 ± 0.014	1.000
NTv3	H_{inv}	0.452 ± 0.007	0.589 ± 0.005	0.581 ± 0.002	0.510 ± 0.004	0.288 ± 0.007	0.824 ± 0.007
	H_{free}	0.397 ± 0.010	0.547 ± 0.008	0.549 ± 0.002	0.385 ± 0.004	0.265 ± 0.007	0.643 ± 0.006
	H_{equi}	0.450 ± 0.005	0.587 ± 0.004	0.587 ± 0.004	0.580 ± 0.008	0.355 ± 0.006	1.000
Frozen backbone							
Caduceus	H_{inv}	0.249 ± 0.002	0.393 ± 0.002	0.393 ± 0.002	0.158 ± 0.002	0.659 ± 0.007	0.877 ± 0.008
	H_{free}	0.245 ± 0.003	0.413 ± 0.003	0.422 ± 0.002	0.147 ± 0.002	0.418 ± 0.004	0.573 ± 0.003
	H_{equi}	0.259 ± 0.002	0.424 ± 0.003	0.424 ± 0.003	0.276 ± 0.009	0.622 ± 0.009	1.000
HyenaDNA	H_{inv}	0.284 ± 0.002	0.454 ± 0.001	0.450 ± 0.001	0.288 ± 0.002	0.454 ± 0.007	0.821 ± 0.012
	H_{free}	0.268 ± 0.004	0.444 ± 0.004	0.463 ± 0.002	0.269 ± 0.004	0.354 ± 0.002	0.668 ± 0.007
	H_{equi}	0.293 ± 0.002	0.463 ± 0.002	0.463 ± 0.002	0.383 ± 0.003	0.502 ± 0.004	1.000
NTv3	H_{inv}	0.292 ± 0.005	0.457 ± 0.004	0.461 ± 0.003	0.277 ± 0.006	0.472 ± 0.006	0.803 ± 0.012
	H_{free}	0.287 ± 0.002	0.457 ± 0.001	0.459 ± 0.002	0.255 ± 0.005	0.388 ± 0.003	0.674 ± 0.006
	H_{equi}	0.306 ± 0.003	0.470 ± 0.002	0.470 ± 0.002	0.386 ± 0.004	0.519 ± 0.005	1.000

A.4 RC-invariant control: dataset and evaluation details

The H3K27me3 control uses the revised public release of the Nucleotide Transformer downstream benchmark (Dalla-Torre et al., 2025): a binary K562 ENCODE ChIP-seq task with 1,000-bp sequences. We canonicalised each sequence together with its RC, removed duplicate orbits within each split, and removed every orbit shared across splits with priority test $>$ validation $>$ train. The resulting train/validation/test splits contained 26,871/3,000/2,947 sequences and had zero cross-split RC-orbit overlap. Each sequence was centred and N-padded to the same 1,024-nt model input used in the directional benchmark; the remaining Caduceus and optimisation settings were unchanged.

The conventional, RC-view, orbit-paired, and numerical symmetry metrics are reported in Table 4.

B Deferred proofs

B.1 Supporting derivations

B.1.1 Proof of Theorem 1

That every element of $\mathcal{F}_{\mathrm{feat}}$ depends only on m_+ is immediate from the definition. For $\mathcal{F}_{\mathrm{logit}}$, let $f = \frac{1}{2}(g + g\circ\sigma)$; then $f(m_+, -m_-) = \frac{1}{2}\big(g(m_+, -m_-) + g(m_+, m_-)\big) = f(m_+, m_-)$, so f is even in m_-, and conversely every even function of m_- is fixed by the symmetrisation. Inclusion holds because $g \circ \Pi_+ \in \mathcal{G}$ by assumption and is already σ-invariant, hence equals its own symmetrisation; strictness follows by taking $g(m_+, m_-) = \|m_-\|^2$, which lies in $\mathcal{F}_{\mathrm{logit}}$ but is not a function of m_+. Invariance of both classes holds since in either case $s(Rx)$ and $s(x)$ evaluate the same function at σ-related arguments on which it is constant. □

Remark 1 (Why post-hoc conjoining is a strong baseline)**.** Strictness of the inclusion offers an explanation for the finding of (Zhou et al., 2022) that a standard model made conjoined after training is difficult to beat with RC parameter sharing. Feature averaging retains only the trivial isotypic component; logit symmetrisation additionally retains every non-linear invariant of the sign component. On genuinely RC-invariant targets the latter is therefore strictly more expressive at equal parameter count.

B.1.2 TIGHTNESS IN THEOREM 2

Both inequalities in equation 9 are attained by the invariant rule that predicts y_{10} on every input whose label lies in A and the correct label otherwise. This rule has $\mathrm{Rec}(y_{10}) = 1$, $\mathrm{Rec}(y_{01}) = 0$ and $\mathrm{Acc} = 1 - \frac{1}{2}\pi_A$, while still satisfying $\mathrm{OPA}_A = 0$. Note that it breaks the label symmetry by fiat and is never selected by a proper scoring rule, by Lemma 1. □

B.1.3 PROOF OF LEMMA 1

Pair the population risk across orbits,

$$\mathcal{L}(q) = -\tfrac{1}{2}\,\mathbb{E}_x\Big[\textstyle\sum_y P(y \mid x)\log q_y(x) + \sum_y P(y \mid Rx)\log q_y(Rx)\Big].$$

Invariance gives $q(Rx) = q(x)$ and Assumption 1 gives $P(y \mid Rx) = P(\mathrm{swap}\, y \mid x)$, so the second sum equals $\sum_y P(\mathrm{swap}\, y \mid x)\log q_y(x)$. Adding the two sums yields $\bar{P}$. Properness of ℓ identifies $\bar{P}$ as the unique minimiser, and $\bar{P}(y_{10} \mid x) = \bar{P}(y_{01} \mid x)$ by construction. The argument applies verbatim to any proper scoring rule, the log loss being one instance. □

B.1.4 WHEN BOTH DIRECTIONAL RECALLS VANISH EXACTLY

Corollary 2. *Let $q = \bar{P}$ as in Lemma 1 and let $\hat{y}$ use* $\arg\max$ *decoding. Then $\hat{y}(x) \in A$ if and only if*

$$\tfrac{1}{2}\big(P(y_{10} \mid x) + P(y_{01} \mid x)\big) > \max\big(P(y_{00} \mid x), P(y_{11} \mid x)\big). \tag{14}$$

If equation 14 fails almost everywhere then $\mathrm{Rec}(y_{10}) = \mathrm{Rec}(y_{01}) = 0$ *exactly.*

Corollary 2 is strictly weaker than Theorem 2: halving the directional confidence caps each directional score at one half of the model's directional mass, which is often insufficient to outrank the two fixed classes, but this depends on the class prior and on decoding. We therefore report a low-directional-prediction regime ($\tau = 0.1102$ for adapted HyenaDNA) and a regime in which directional predictions are common ($\tau = 0.5454$ for adapted Caduceus). Both satisfy $\mathrm{OPA}_A = 0$ and the recall identity equation 10. The latter has rightward and leftward recalls of 0.2762 and 0.2692, excluding the absence of directional predictions or a one-sided class collapse as explanations for the orbit-level zero.

B.1.5 PROOF OF PROPOSITION 1

Invariance gives $\hat{y}(Rx) = \hat{y}(x) = c$, so $\mathrm{IAR}_A = 1$, and the EAR_A event holds iff $\mathrm{swap}(c) = c$, i.e. iff $c \in S$, which occurs on a $(1 - \tau)$ fraction of free orbits. On each free orbit, exactly one element is correct when $c \in A$ and neither when $c \in S$, so the number of correct predictions on A is $\tau|\mathcal{P}_A|$; dividing by $N_A/2 = |\mathcal{P}_A|$ gives equation 10. Equivariance gives $\hat{y}(Rx) = \mathrm{swap}(\hat{y}(x))$ directly, and the pairs (x, Rx) contribute symmetrically to the two directional recalls.

B.1.6 PROOF OF THEOREM 3, AND THE RISK-LEVEL ANALOGUE

For any h and any $g \in \mathcal{F}_{\mathrm{inv}}$ the composition $g \circ h$ is a G-invariant score function, so Theorem 2 gives $\mathrm{OPA}_A(g \circ h) = 0$; taking the supremum yields $\Psi_{\mathcal{F}_{\mathrm{inv}}}(h) = 0$, and the argument uses no property of h.

For non-constancy, let h_1 satisfy equation 2 with a sign component that determines the direction and set $h_2 = \Pi_+ h_1$. By equation 3, $h_2(Rx) = \sigma_L h_2(x)$, and position-symmetric pooling maps σ_L to the identity, so $r(x) = q(\mathrm{Pool}(h_2(x)))$ satisfies $r(Rx) = r(x)$ for every q. Hence

$$H_{\mathrm{equi}}(x) = \tfrac{1}{2}\big[r(x) + P\, r(Rx)\big] = \tfrac{1}{2}\,(I + P)\, r(x), \tag{15}$$

and $\frac{1}{2}(I+P)$ is the orthogonal projection onto the swap-fixed subspace of $\mathbb{R}^{\mathcal{Y}}$, so the two directional coordinates coincide at every input and no decoding rule can assign y_{10} to one member of a free orbit and y_{01} to the other; thus $\Psi_{\mathcal{F}_{\mathrm{equi}}}(h_2) = 0$. For h_1 the sign component survives pooling up to a global sign and the equivariant readout attains $\mathrm{OPA}_A > 0$, as verified on all three backbones in Section 3.3.

Remark 2 (Why the statement is placed on OPA_A)**.** By Theorem 1, an invariant readout on h_1 realises the even functions of (m_+, m_-) whereas one on $h_2 = \Pi_+ h_1$ realises only functions of

m_+; the former class is strictly larger, so the two backbones need not attain equal *risk* under $\mathcal{F}_{\mathrm{inv}}$. Invariant readouts can exploit $\|m_-\|$ on the fixed-label block S, and aggregate metrics may reflect this. What collapses to an exact, backbone-independent zero is the orbit-level functional on free orbits, which is why Theorem 3 is stated there.

Corollary 3 (Risk-level analogue)**.** *Let h_1, h_2 be as above, write $V(h, \mathcal{F}) = \inf_{g \in \mathcal{F}} \mathcal{R}(g \circ h)$ for a proper risk, and assume in addition that $\bar{P}$ is measurable with respect to $m_+ = \mathrm{Pool}(\Pi_+ h_1)$. Then*

$$V(h_1, \mathcal{F}_{\mathrm{inv}}) = V(h_2, \mathcal{F}_{\mathrm{inv}}) \quad \textit{while} \quad V(h_1, \mathcal{F}_{\mathrm{equi}}) < V(h_2, \mathcal{F}_{\mathrm{equi}}). \tag{16}$$

Proof. Under $\mathcal{F}_{\mathrm{inv}}$ both compositions are G-invariant, so by Lemma 1 the risk of either equals its risk against $\bar{P}$; the measurability condition places $\bar{P}$ in the attainable class of both backbones, so both attain its Bayes risk. Under $\mathcal{F}_{\mathrm{equi}}$, h_1 retains the sign component and attains the unrestricted Bayes risk, whereas h_2 has discarded it and, by equation 15, remains at the symmetrised risk; the inequality is strict whenever $\bar{P} \neq P$, i.e. whenever $\pi_A > 0$. □

Without the measurability condition only $V(h_1, \mathcal{F}_{\mathrm{inv}}) \leq V(h_2, \mathcal{F}_{\mathrm{inv}})$ holds, by Remark 2.

B.1.7 Proof of Proposition 2

Immediate from equation 3 together with the definition of position-symmetric pooling, which commutes with σ_L and maps it to the identity on pooled features. In the second case the length reversal survives as a permutation of pooled windows; in the third the sign of m_- survives. The final claim follows by applying Theorem 2 to the first case.

B.1.8 Orbit-mode ceiling for general finite groups

Proposition 3. *For a finite group G acting on $\mathcal{X}$ and on $\mathcal{Y}$ through a permutation representation ρ, with $\Pr[Y = y \mid \gamma x] = \Pr[Y = \rho(\gamma)^{-1} y \mid x]$, deterministic labels, and any G-invariant classifier, equation 12 holds, with equality attained by predicting an orbit-wise modal label.*

Proof. An invariant classifier is constant on orbits, so on each orbit it can do no better than the modal label; taking expectations over orbits, weighted by probability mass, gives equation 12. For a free $\mathbb{Z}_2$-orbit whose two labels are exchanged by ρ, the bracketed quantity is $1/2$, and averaging over the π_A fraction of such orbits yields $1 - \frac{1}{2}\pi_A$. □

B.1.9 Closed-form MCC values, and why MCC cannot adjudicate

Let $\mathrm{MCC} = \big(Nc - \sum_k t_k p_k\big) \big/ \sqrt{(N^2 - \sum_k p_k^2)(N^2 - \sum_k t_k^2)}$ with t_k true counts, p_k predicted counts and c correct predictions (Gorodkin, 2004).

Corollary 4. *(i) For a two-class orbit-closed task whose two labels form a single free orbit, every G-invariant classifier attains $\mathrm{MCC} = 0$ and $\mathrm{Acc} = \frac{1}{2}$ exactly, for any decoding rule. (ii) For the balanced four-class task ($t_k = N/4$), the collapse pattern sending all directional inputs to y_{00} gives $\mathrm{MCC} = \sqrt{2}/3 \approx 0.4714$, and the maximum over all invariant classifiers is $\mathrm{MCC} = 2\sqrt{2/15} \approx 0.7303$.*

Proof. (i) Orbit closure and Theorem 2 force $c = N/2$, and $t_1 = t_2 = N/2$ gives $\sum_k t_k p_k = (N/2)(p_1 + p_2) = N^2/2 = Nc$, so the numerator vanishes identically. (ii) Balance gives $\sum_k t_k^2 = N^2/4$ and $\sum_k t_k p_k = (N/4)\sum_k p_k = N^2/4$. For the collapse pattern, $c = N/2$ and $(p_k) = (3N/4, 0, 0, N/4)$, so $\sum_k p_k^2 = 5N^2/8$ and $\mathrm{MCC} = (N^2/4)/\sqrt{(3N^2/8)(3N^2/4)} = \sqrt{2}/3$. For the maximum, the recall bound gives $c \leq 3N/4$; attaining it requires directional predictions to lie in A, which fixes $p_{00} = p_{11} = N/4$ and $p_{10} + p_{01} = N/2$, and $\sum_k p_k^2$ is maximised at $(N/4, N/2, 0, N/4)$, giving $\sum_k p_k^2 = 3N^2/8$ and $\mathrm{MCC} = (N^2/2)/\sqrt{(5N^2/8)(3N^2/4)} = 2\sqrt{2/15}$. Enumerating the remaining count patterns yields smaller values. □

The maximiser in (ii) is the degenerate invariant rule of Appendix B.1.2: it breaks the label symmetry by fiat, achieving unit recall on one directional class and zero on the other. Such a rule is never selected by a proper scoring rule (Lemma 1) but it *is* a member of the invariant class. Consequently

the invariant ceiling on MCC can exceed the MCC actually attained by an unconstrained readout, and MCC admits no valid separation between the two symmetry classes; the same holds for accuracy and macro-F_1. This is a property of aggregate metrics rather than a limitation of the analysis, and it is the reason we adjudicate the central claim with OPA_A, for which Theorem 2 gives an exact, decoding-independent zero. We report MCC throughout for comparability with prior work only.