CROSSREFEVAL: PAIRED EVALUATION OF GENOMIC MODELS ACROSS REFERENCE ASSEMBLIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic models are usually evaluated on a single reference genome, leaving open how the same predictor responds when corresponding regions are represented by another assembly. This comparison requires explicit correspondence: for position-resolved prediction, corresponding bases can occur at different output indices, so the rule used to pair predictions can itself affect the measured comparison. We introduce **CrossRefEval**, a paired framework that holds the fitted predictor and decision rule fixed across reference representations and evaluates task performance, cross-reference consistency, and paired correctness at corresponding outputs. We instantiate CrossRefEval on the maternal haplotype assembly of T2T-YAO v2.0 and GRCh38, constructing 254,366 paired windows across five genomic prediction tasks. We evaluate five frozen DNA encoders with readouts trained separately on YAO and GRCh38. Holding prediction arrays fixed, changing only the output correspondence alters the comparison between readout variants: across ten encoder–training-reference combinations, the absolute correspondence-induced shift exceeds the magnitude of the readout difference measured under mapped correspondence in nine cases and reverses the direction of the three-seed mean comparison in six. Better task performance can also accompany greater cross-reference disagreement, while increased agreement can include shared errors. These results show that per-reference task scores alone do not characterize how predictions change across reference representations. CrossRefEval makes reference representation and output correspondence explicit parts of genomic model evaluation rather than fixed background assumptions.

1 INTRODUCTION

Genomic model benchmarks typically evaluate predictions within a fixed reference assembly. The reference determines not only the sequence presented to the model, but also the coordinate system used to define examples, annotations, and prediction outputs. GRCh38 remains the basis for much current human-genome analysis. Increasingly complete assemblies, including diploid references for HG002 and I002C, provide alternative representations of corresponding genomic regions (Nurk et al., 2022; Hansen et al., 2026; Sarashetti et al., 2025). The same fitted predictor may therefore produce different outputs when a corresponding region is represented by a different reference.

Cross-reference evaluation asks how predictions change across paired representations of corresponding genomic regions. Such comparisons require explicit correspondence. For window-level tasks, paired regions provide a natural unit of comparison. For position-resolved tasks, insertions, deletions, and other local sequence differences can shift corresponding bases to different indices, so the output at index j in one assembly need not correspond to the output at index j in another. The correspondence rule is therefore part of the evaluation.

CrossRefEval is a paired framework for evaluating genomic models across reference assemblies. The same fitted predictor and decision rule are applied unchanged to both representations of each paired region. The framework reports task performance on each reference, **cross-reference consistency** at corresponding outputs, and paired correctness under a shared task definition. Higher consistency is not always better: genuine sequence differences may justify different predictions, while high consistency can also reflect shared errors.

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.

The first instantiation of CrossRefEval uses GRCh38 and the maternal haplotype assembly of T2T-YAO v2.0 (He et al., 2023; Chu et al., 2026). T2T-YAO v2.0 is a near-perfect diploid East Asian reference genome. Its high-quality genome annotations motivated this choice and provide the basis for defining the benchmark tasks and shared labels. Existing mappings to GRCh38 establish the corresponding regions and output positions. The benchmark contains 254,366 pairs of windows across five tasks, spanning window-level and position-resolved prediction, with chromosome-based splits. YAO–GRCh38 is one instantiation of CrossRefEval; the same protocol applies to other reference pairs given corresponding regions, compatible task labels, and output maps for position-resolved tasks.

Experiments with five frozen DNA encoders show that output correspondence can affect model comparisons even when the prediction arrays are fixed. Across ten encoder–training-reference combinations, the absolute correspondence-induced shift exceeds the magnitude of the readout difference under mapped correspondence in nine cases, and equal-index matching reverses the sign of the three-seed mean comparison in six. Across tasks, better predictive performance can also accompany greater cross-reference disagreement, while increased agreement can include shared errors. Per-reference task scores alone therefore do not describe how predictions depend on reference representation and output correspondence.

2 Related Work

2.1 Genomic model benchmarks under a fixed reference

Existing genomic benchmarks primarily measure predictive performance on predefined sequence tasks. GUE evaluates DNA representations across classification tasks and species, BEND includes biologically meaningful human-genome tasks and position-level prediction, and DART-Eval focuses on regulatory prediction under zero-shot, probing, and fine-tuning settings (Zhou et al., 2024; Marin et al., 2024; Patel et al., 2024). Nucleotide Transformer similarly evaluates genomic foundation models across a broad collection of downstream prediction tasks (Dalla-Torre et al., 2025), while Feng et al. (2025) show that pooling choices can affect the downstream performance of frozen genomic representations.

These studies broaden genomic model evaluation across tasks, model classes, and training settings, but the evaluation examples themselves are generally defined within a fixed genomic representation. They do not treat paired predictions from alternative reference assemblies as the object of evaluation.

2.2 Reference-dependent effects and cross-assembly correspondence

Reference representation is already known to affect genomic analysis. The complete T2T-CHM13 reference improves read mapping and variant analysis relative to earlier human references, while the HPRC draft pangenome captures sequence and structural diversity absent from GRCh38 and improves variant discovery in short-read analyses (Aganezov et al., 2022; Liao et al., 2023). Macias-Velasco et al. (2026) extend this question to functional genomics by comparing analyses based on standard references, individual assemblies, and a pangenome.

A benchmark of six liftover tools found that epigenomic interval conversion between GRCh37 and GRCh38 can differ by tool and region (Luu et al., 2020). CrossRefEval uses existing correspondence to pair independently generated outputs, not to transfer predictions between assemblies. For position-resolved tasks, this pairing accounts for differences in the indices of corresponding bases.

2.3 Prediction differences across related sequences

Genomic models are also used to compare related but biologically distinct sequences. Sasse et al. (2023) and Huang et al. (2023) evaluate sequence models using personal genomes and expression measurements, and both report limitations in predicting inter-individual expression differences from sequence. Huang et al. further show that models often fail to recover the direction of expression effects associated with cis-regulatory variation. AlphaGenome instead compares predictions from reference and alternate sequences to estimate regulatory variant effects (Avsec et al., 2026).

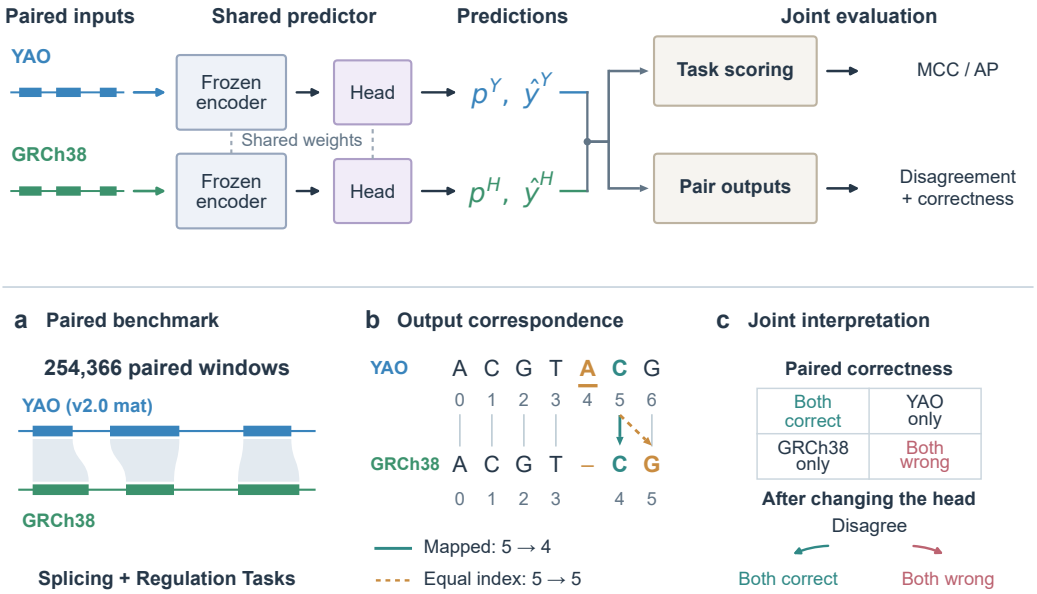

Figure 1: **CrossRefEval framework and YAO–GRCh38 benchmark.** The same fitted predictor is evaluated on both reference representations. (a) Corresponding YAO and GRCh38 windows are paired and share task labels. (b) Position maps match output positions when corresponding bases occur at different sequence indices. (c) CrossRefEval reports task performance, cross-reference consistency, and paired correctness under the shared labels. Resolving a disagreement can produce either correct agreement or a shared error.

These studies aim to predict biological differences between alleles or individuals. CrossRefEval examines prediction changes across corresponding reference regions under shared task labels. Sequence differences may justify different predictions, whereas agreement can include shared errors. Cross-reference consistency is therefore interpreted alongside task performance and paired correctness.

3 CrossRefEval

CrossRefEval applies each fitted predictor and decision rule unchanged to both reference inputs. It reports per-reference task performance, cross-reference consistency, and paired correctness. Position-resolved outputs are paired through an explicit position map (Figure 1).

3.1 YAO–GRCh38 benchmark instantiation

The YAO–GRCh38 benchmark contains 254,366 pairs of DNA windows from corresponding genomic regions. Each pair consists of a YAO window and its corresponding GRCh38 window; both windows share a sample identity and task labels. Records store the coordinates, orientation, split assignment, and sequence hash of both representations. The position-resolved task also uses a within-window map between corresponding output positions.

Using existing coordinate mappings, we pair corresponding autosomal regions of GRCh38 with the maternal haplotype assembly of T2T-YAO v2.0. Accepted mappings define which regions and positions are eligible for paired evaluation. Task-specific construction and filtering then select the final examples. Reference provenance and source assignment are documented in Appendix A.3; mapping construction, filtering rules, and mapping-support analyses are described in Appendices B.1, B.2, and B.4.

The benchmark includes five tasks spanning window-level and position-resolved prediction (Table 1). T1 classifies candidate splice sites—windows centered on splice-like sequence motifs—into background, donor, or acceptor classes. T2 performs the same three-class prediction at every posi-

Table 1: **Tasks, input lengths, and chromosome-based split sizes.** Counts denote YAO–GRCh38 window pairs. T2 contains multiple labeled positions within each paired window; the other tasks produce one prediction per window.

Task	Input (bp)	Train	Validation	Test	Total
T1 Splice classification	402	49,997	8,000	8,000	65,997
T2 Position-resolved splice prediction	2,004	14,077	2,162	2,544	18,783
T3 TSS classification	300	49,169	7,657	7,930	64,756
T4 Promoter classification	510	28,540	4,512	5,780	38,832
T5 Enhancer classification	510	49,998	8,000	8,000	65,998
Total	—	**191,781**	**30,331**	**32,254**	**254,366**

tion in a sequence window. T3–T5 perform binary classification of transcription start sites (TSSs), promoters, and enhancers against task-specific background windows.

Task labels come from YAO-coordinate gene and regulatory annotations derived from GRCh38-based RefSeq and ENCODE4 resources (Appendix A.1). We use these labels as shared benchmark targets for paired YAO and GRCh38 inputs. Gene, transcript, and exon annotations define T1–T3, while candidate promoter-like and enhancer-like annotations define T4 and T5. Background examples are selected outside the relevant target annotations. T1 uses splice-like sequence motifs outside annotated splice sites, T3 uses GC-matched windows that exclude annotated TSSs, and T2 assigns background to the remaining valid positions. Background construction for T4 and T5, together with detailed task construction and label provenance, is described in Appendices A.1 and A.6.

Correctness is defined relative to these shared benchmark labels, so annotation or transfer errors can affect both training and evaluation. Each pair is assigned to a split according to its YAO chromosome, and the paired GRCh38 region remains in the same split. Chromosomes 5, 9, and 22 are used for validation; chromosomes 4, 13, 19, and 21 for testing; and the remaining autosomes for training.

To evaluate a new predictor, users run it on the fixed paired sequences and provide the resulting probability arrays to the CrossRefEval scorer. The scorer checks sample order, sequence hashes, output shapes, probability ranges, and multiclass normalization before reporting task performance, cross-reference consistency, and paired correctness. Labels are provided for all splits. Output-position maps for position-resolved evaluation are provided for validation and testing (Appendix I.2).

3.2 Correspondence-aware evaluation

Predictions are generated independently for the two sequences in each pair. The output-position map is applied after prediction; it does not affect feature extraction or prediction. Window-level outputs are paired by sample identity, whereas position-resolved outputs are paired through the recorded position map.

For T2, let $\varphi_i(j)$ denote the GRCh38 output index corresponding to YAO index j in window i. We refer to this pairing as **mapped correspondence**. The YAO-to-GRCh38 map is built from aligned sequence blocks independently of the labels. A YAO position is included in the paired comparison only when its mapped partner lies within the GRCh38 window and both stored labels are valid. After fixing the map, we check label agreement at the retained pairs; no label conflicts occur (Appendix B.2). The same mapped positions are used across readout comparisons.

Task performance is evaluated separately on each reference representation. We use Matthews correlation coefficient (MCC) for the three-class T1 and T2 tasks and non-interpolated average precision (AP) for the binary T3–T5 tasks. As secondary T2 analyses, we also report donor-versus-rest and acceptor-versus-rest AP to examine class-specific score rankings (Appendix F.4).

Cross-reference consistency is defined from the predicted classes at paired outputs. We report its complement, the disagreement or flip rate F. For window-level tasks, F is the fraction of paired windows assigned different classes. For position-resolved prediction, disagreement is calculated over the retained mapped positions within each window and then averaged across windows:

$$F_{\mathrm{map}} = \frac{1}{W} \sum_{i=1}^{W} \frac{1}{|J_i|} \sum_{j \in J_i} \mathbf{1}\left[\hat{y}_i^Y(j) \neq \hat{y}_i^H(\varphi_i(j))\right]. \tag{1}$$

Here, W is the number of scored paired windows with $|J_i| > 0$; all 2,544 T2 test windows meet this condition. J_i contains the YAO indices in window i whose mapped GRCh38 partner lies within the paired window and whose stored task labels are valid on both sides. $\hat{y}_i^Y(j)$ and $\hat{y}_i^H(\varphi_i(j))$ are the predicted classes at the paired YAO and GRCh38 outputs. Each window receives equal weight, so windows with more retained mapped positions do not dominate the estimate. Windows may overlap, and a genomic position can therefore contribute to more than one window.

Primary T2 MCC is computed from all valid positions within each reference window, whereas cross-reference disagreement is computed on the mapped paired domain. Appendix G.1 also reports MCC on the same mapped positions used for cross-reference evaluation.

The shared task target assigns each paired prediction to one of four correctness states: both correct, only YAO correct, only GRCh38 correct, or both incorrect (Figure 1c). For window-level tasks, each paired window receives one state; for T2, each retained mapped position receives one. In binary tasks, a disagreement that disappears can become either correct agreement or a shared error. In multiclass tasks, two incorrect predictions may also disagree with each other, so we retain the predicted classes in addition to the correctness state. These records show how changes in consistency relate to changes in correctness.

3.3 Models and experimental protocol

We evaluate five frozen DNA encoders: DNABERT-2 (Zhou et al., 2024), GROVER (Sanabria et al., 2024), HyenaDNA (Nguyen et al., 2023), Caduceus-PH (Schiff et al., 2024), and NTv3-100M-pre (Boshar et al., 2025; InstaDeepAI, n.d.). Task-specific readouts operate on frozen final-layer representations. Where necessary, token representations are projected to DNA-base positions before prediction, with invalid positions masked. Model identities and feature-extraction procedures are detailed in Appendices C.1 and C.2.

The primary experiments compare two readout variants for each task. All readouts use two width-3 convolutional layers with 64 channels. For window-level classification, the resulting features are average-pooled either across the whole window (**pool1**) or across 16 ordered bins (**pool16**) before the linear classifier. Pool16 preserves coarse positional information and increases the classifier input dimension. For T2, the classifier operates at every sequence position; the two variants use dilation 1 or dilation 4 in both convolutional layers. The two dilation settings change the span and density of sampled positions without changing the parameter count. One-hot controls use the same readout designs but operate directly on one-hot encoded DNA bases.

Readouts are trained separately on YAO and GRCh38 for each task, variant, and seed, then evaluated unchanged on both references. In YY/YH/HY/HH, the first letter denotes the training reference and the second the test reference (Y: YAO; H: GRCh38). YY versus YH holds the YAO-trained readout fixed; HY versus HH holds the GRCh38-trained readout fixed. YY versus HH compares independently trained readouts. Readout variants are compared within each training-reference group.

All primary experiments use the fixed benchmark cohorts and seeds 14, 15, and 16. Paired readout variants share recorded initialization and batch-order controls. Readouts are trained with AdamW using batch size 32 for at most 20 epochs, with early stopping after five epochs without validation improvement. Checkpoints are selected by validation performance on the readout-training reference: MCC for T1/T2 and AP for T3–T5. The selected predictor is then applied unchanged to both test representations, using argmax for T1/T2 and a fixed threshold of 0.5 for T3–T5. Additional optimization details, compute costs, and comparator experiments are reported in Appendices C.3, C.4, H.1, and H.3.

Performance differences are summarized as means across the three fitted seeds. Uncertainty for cross-reference disagreement contrasts is estimated from 5,000 chromosome-stratified block-bootstrap replicates. Each replicate uses the same sampled block multiplicities for the paired readouts and fitted seeds so that they are resampled together. The primary block width is 1 Mb, with 2 Mb used as a sensitivity analysis. Pointwise 95% intervals describe genomic-sampling uncertainty con-

ditional on the fitted models. A finding is considered significant after correction only when its direction is preserved and Benjamini–Hochberg $q < 0.05$ at both block widths (Benjamini & Hochberg, 1995). Testing families separate T1, T2, and T3–T5 and also separate frozen representations from one-hot controls. The post-hoc correspondence diagnostic compares mapped correspondence with **equal-index matching**, which pairs output j in YAO with output j in GRCh38, and uses unadjusted intervals (Appendix B.3). The bootstrap protocol and testing families are detailed in Appendices D.1, E.3, and H.1.4.

4 RESULTS

4.1 CORRESPONDING POSITIONS CAN OCCUR AT DIFFERENT OUTPUT INDICES

In 908 of the 2,544 T2 test-window pairs (35.69%), at least one corresponding position maps to a different output index. Across all test pairs, the position maps retain 5,078,078 valid mapped observations, of which 563,787 (11.10%) have nonzero index displacement (Appendices A.6 and B.2). These displacements are defined by the paired sequences and position maps before model predictions are considered.

At a displaced position, mapped correspondence and equal-index matching select different GRCh38 outputs for the same YAO output (Figure 1b). If those outputs receive different predicted classes, the correspondence rule changes the measured disagreement even though the prediction arrays remain fixed.

4.2 OUTPUT CORRESPONDENCE CHANGES READOUT COMPARISONS

Changing only the correspondence rule changes the comparison between dilation 4 and dilation 1 across the ten encoder–training-reference combinations (Figure 2a,b). In nine of ten cases, the absolute correspondence-induced shift exceeds the magnitude of the readout difference under mapped correspondence. In six cases, equal-index matching reverses the sign of the three-seed mean comparison, changing which readout has lower measured cross-reference disagreement.

Prediction arrays, retained YAO positions, and window weights are held fixed across correspondence rules. For each correspondence rule, ΔF denotes the three-seed mean disagreement difference, dilation 4 minus dilation 1; positive values indicate greater disagreement for dilation 4. We define the correspondence-induced shift as

$$I = \Delta F_{\mathrm{equal}} - \Delta F_{\mathrm{map}}. \tag{2}$$

Across the ten settings, I ranges from -0.2642 to $+0.2516$ percentage points (pp), whereas the mapped difference ΔF_{map} ranges from -0.0230 to $+0.1272$ pp. Of the six mean sign reversals, three occur in all three fitted seeds; the remaining three occur in one or two seeds (Appendix B.3).

Setting-specific, unadjusted pointwise 95% intervals for I exclude zero at both block widths in nine of ten settings; GROVER under YAO training is the exception. Restricting the analysis to higher-support mapping subsets (Appendix B.4) preserves the direction of the mean shift in all ten settings.

The shift is larger in the one-hot controls, reaching $+2.9597$ and $+2.9361$ pp under YAO and GRCh38 training, respectively (Figure 2c). The identical-input control is reported in Appendix B.5.3.

4.3 BETTER TASK PERFORMANCE CAN ACCOMPANY GREATER CROSS-REFERENCE DISAGREEMENT

Changes in task performance and cross-reference disagreement do not follow a single pattern across the five tasks (Table 2; Figure 3a–e). The three-seed mean performance difference has the same sign on both test references in all 50 task–encoder–training-reference settings (five tasks $\times$ five encoders $\times$ two training references; Appendix E.5). The corresponding change in disagreement, however, varies across tasks and models.

T1 shows the clearest joint improvement: pool16 increases mean MCC on both test references and reduces cross-reference disagreement in all ten encoder–training-reference settings. All ten

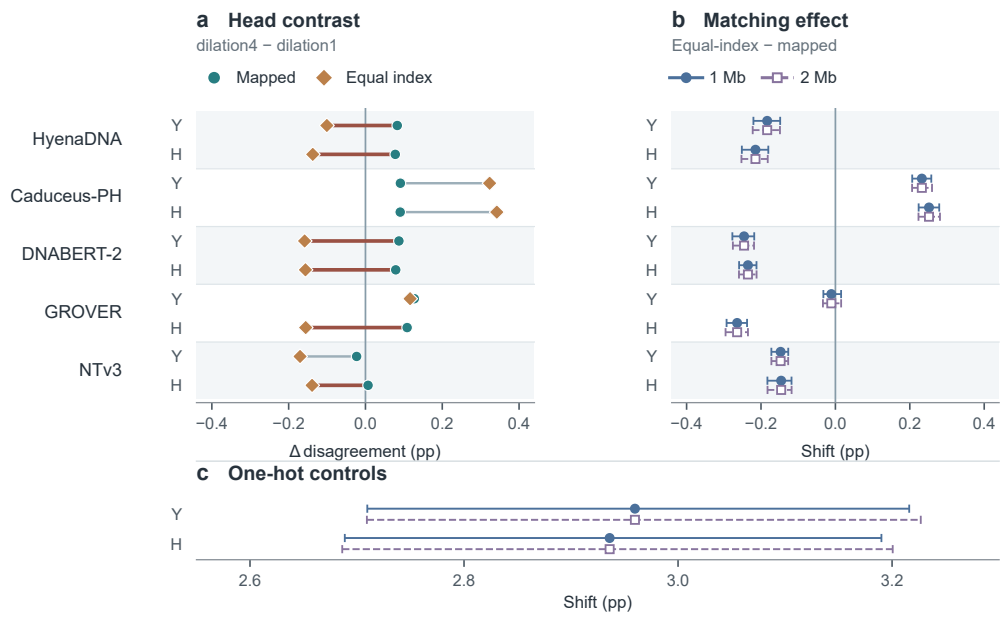

Figure 2: **Output correspondence changes readout comparisons with predictions fixed.** (a) Mean disagreement differences, dilation 4 minus dilation 1, under mapped and equal-index correspondence. (b) Correspondence-induced shift, $\Delta F_{\text{equal}} - \Delta F_{\text{map}}$. (c) One-hot controls shown on a separate scale. Means use three seeds. In (a), lines connect paired point estimates and six highlighted lines cross zero. Error bars in (b,c) show unadjusted 95% block-bootstrap intervals conditional on the fitted models at 1 Mb and 2 Mb block widths. Y/H denote YAO/GRCh38 readout training; pp denotes percentage points.

Table 2: **Task performance and cross-reference disagreement under readout changes.** Differences are pool16 minus pool1 for T1/T3–T5 and dilation 4 minus dilation 1 for T2. Ranges summarize three-seed means across 20 performance and ten disagreement settings per task. Significant disagreement changes require the same direction and Benjamini–Hochberg $q < 0.05$ at both block widths. pp, percentage points. See Appendix E for complete results.

Task	Metric	Performance difference range	Positive changes (of 20)	Disagreement difference (pp)	Significant decreases / increases (of 10)
T1	MCC	+0.1394 to +0.4659	20	−1.6042 to −0.4792	10 / 0
T3	AP	+0.0248 to +0.0670	20	−0.2396 to +0.1345	0 / 0
T4	AP	−0.0006 to +0.0166	18	−0.2191 to +0.1326	0 / 0
T5	AP	+0.0030 to +0.0426	20	−0.3833 to +0.2083	2 / 0
T2	MCC	−0.0228 to +0.0159	8	−0.0230 to +0.1272	1 / 8

disagreement decreases remain significant after correction. For T3–T5, mean AP increases on both test references in 29 of 30 settings, while disagreement can increase or decrease. Among the binary tasks, the only significant disagreement changes are the decreases for GROVER on T5 under both training references.

HyenaDNA on T2 shows a contrasting pattern (Figure 3e). Relative to dilation 1, dilation 4 increases mean full-window MCC by 0.0113 under YAO training and 0.0136 under GRCh38 training, averaged over both test references and three seeds. Cross-reference disagreement increases by 0.0829 and 0.0777 pp, respectively, and both increases remain significant after correction. Recomputing MCC on the same mapped positions used for cross-reference comparison preserves the direction of the performance change in every T2 setting, including HyenaDNA (Appendix G.1).

At the class level, donor-versus-rest and acceptor-versus-rest AP both improve for HyenaDNA on the mapped positions, but the fraction of acceptor positions classified correctly on both references decreases. The both-correct acceptor rate falls from 92.06% to 88.12% under YAO training and from 92.45% to 89.36% under GRCh38 training. Over the same readout change, the both-correct

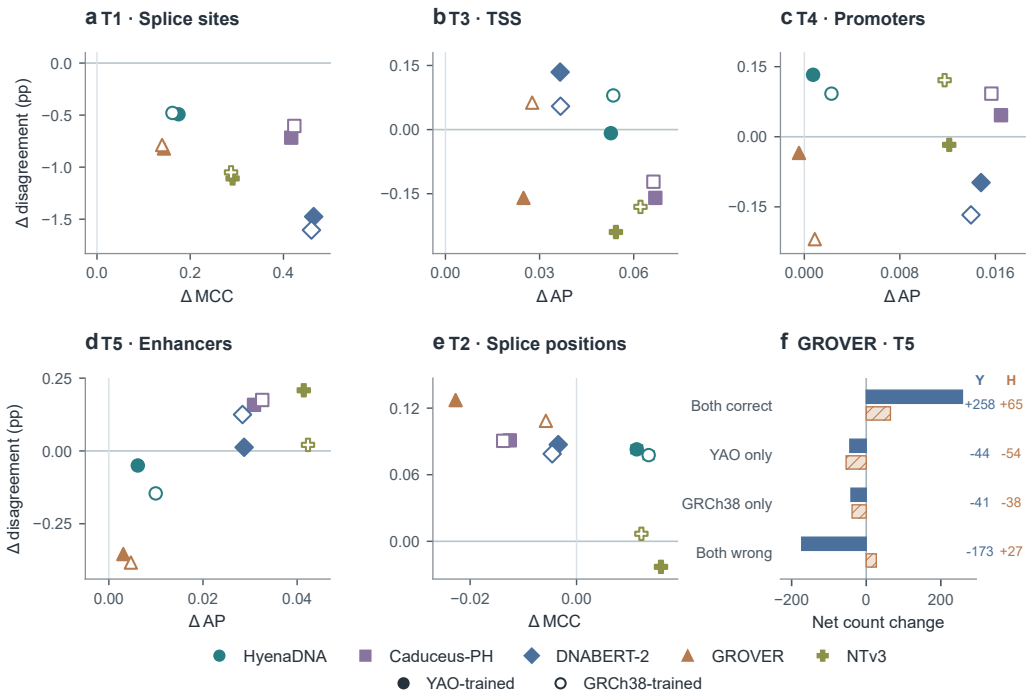

Figure 3: **Task performance and cross-reference consistency can change in different directions.** (a–e) Readout differences are pool16 minus pool1 for T1/T3–T5 and dilation 4 minus dilation 1 for T2, averaged over three seeds; performance differences are additionally averaged over the two test references. HyenaDNA MCC bars show unadjusted 95% block-bootstrap intervals conditional on the fitted models at the 1 Mb block width. (f) Net changes in GROVER T5 correctness-state counts across 8,000 window pairs and three seeds for each training reference. pp denotes percentage points.

background rate increases by 2.44 and 2.46 pp, respectively (Appendix F.3). Improved class-score rankings can therefore coexist with lower paired correctness for acceptor positions (Appendix F.4).

4.4 Newly agreeing predictions can share errors

Newly agreeing prediction pairs disagree across references under pool1 but agree under pool16. Shared task labels distinguish correct agreement from shared errors. Shared-error cases occur among these pairs in all 30 binary-task encoder–training-reference settings (three binary tasks $\times$ five encoders $\times$ two training references) when observations are pooled across the three fitted seeds (Appendix F.5).

GROVER on T5 illustrates both outcomes. Mean AP increases and cross-reference disagreement decreases under both training references, and both disagreement decreases remain significant after correction. Pooling the 8,000 test-window pairs across three fitted seeds gives 24,000 window–seed observations per training reference. Under YAO training, 105 of 280 eliminated disagreements become shared errors (37.5%); under GRCh38 training, 125 of 284 do so (44.0%). Pool16 also introduces 195 and 192 new disagreements, respectively, so the net change in disagreement reflects both resolved and newly introduced flips (Appendix F.6).

Full-cohort counts reveal changes that do not alter agreement status, such as transitions between both-correct and both-incorrect predictions. For GROVER, the total number of shared errors changes by -173 under YAO training and $+27$ under GRCh38 training (Figure 3f; Appendices F.6.1 and F.5). Per-seed changes differ in sign across seeds under both training references. Under YAO training, some eliminated disagreements become shared errors even though the total number of shared errors decreases. Under GRCh38 training, disagreement decreases while the total number of shared errors increases.

Across all 30 binary-task encoder–training-reference settings, pooling observations across the three fitted seeds, full-cohort shared errors decrease in 27 and increase in three (Appendix F.5). Transition counts identify shared errors among newly agreeing pairs; full-cohort counts give the net change in shared errors. Appendix G.4 and Figure S5 report the same quantities across common decision thresholds.

5 Discussion and Limitations

CrossRefEval makes reference representation part of genomic model evaluation. In the position-resolved splice task, changing only output correspondence alters readout comparisons with prediction arrays fixed. For the five frozen encoders, the absolute three-seed mean shifts remain below 0.27 percentage points but exceed the magnitude of the mapped readout difference in nine of ten settings. Equal-index matching reverses the sign of that mean difference in six. Output correspondence therefore belongs in the scoring protocol.

Better task performance can accompany greater cross-reference disagreement, while newly agreeing predictions can share errors. Consistency is not an objective by itself: genuine sequence differences may justify different predictions. It should be interpreted alongside task performance and paired correctness under the shared benchmark labels.

The benchmark covers accepted autosomal pairs between GRCh38 and the maternal haplotype assembly of T2T-YAO v2.0; the paternal haplotype is not evaluated. Unmapped and assembly-specific regions require a separate evaluation design. The experiments do not separate the contributions of sequence divergence, assembly-specific content and local coordinate structure to prediction changes. Shared benchmark labels do not provide independent functional validation on both references, and annotation or transfer errors can affect training and evaluation. The primary readout comparisons cover five frozen encoders, the tested readout variants and three fitted seeds; the correspondence analysis focuses on position-resolved splice prediction. Block-bootstrap intervals quantify genomic-sampling uncertainty conditional on the fitted models, not full retraining uncertainty.

CrossRefEval can be extended to other reference pairs given corresponding regions, compatible task labels and output maps for position-resolved tasks. These extensions would test whether the relationships observed here between task performance, cross-reference consistency and paired correctness persist across reference pairs, tasks and model classes. Independent functional measurements could further assess whether the prediction differences identified by paired evaluation correspond to functional differences.

6 Conclusion

CrossRefEval evaluates a fixed predictor across reference assemblies using paired inputs, shared task labels, and explicit output correspondence. In the YAO–GRCh38 benchmark, changing output correspondence alters readout comparisons even when predictions remain fixed. Better task performance can accompany greater cross-reference disagreement, while increased agreement can include shared errors. Cross-reference evaluation therefore requires more than separate task scores: consistency and paired correctness must be assessed at matched outputs. The protocol makes reference representation an explicit evaluation variable and can be applied to other reference pairs with compatible labels and output correspondence.

AI ASSISTANCE

AI assistance was used to develop and check training, data-processing, and analysis code; support literature searches and result interpretation; prepare figures; and draft, edit, and format the manuscript.

REFERENCES

Sergey Aganezov, Stephanie M. Yan, Daniela C. Soto, et al. A complete reference genome improves analysis of human genetic variation. *Science*, 376(6588):eabl3533, 2022. doi: 10.1126/science.abl3533. URL https://doi.org/10.1126/science.abl3533.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649:1206–1218, 2026. doi: 10.1038/s41586-025-10014-0. URL https://doi.org/10.1038/s41586-025-10014-0.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x.

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. URL https://doi.org/10.64898/2025.12.22.695963.

Yanan Chu, Zhuo Huang, Changjun Shao, Shuming Guo, Yiji Yang, Xinyao Yu, Yurong Luo, Jian Wang, Yabin Tian, Jing Chen, Ran Li, Yukun He, Stylianos E. Antonarakis, Jun Yu, Jie Huang, Zhancheng Gao, and Yu Kang. Approaching an error-free diploid human genome using a support-based validation framework. *The Innovation*, pp. 101543, 2026. doi: 10.1016/j.xinn.2026.101543. URL https://doi.org/10.1016/j.xinn.2026.101543. Journal pre-proof, available online 27 August 2026.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287–297, 2025. doi: 10.1038/s41592-024-02523-z. URL https://doi.org/10.1038/s41592-024-02523-z.

Nathan Dwarshuis, Divya Kalra, Jennifer McDaniel, Philippe Sanio, Pilar Alvarez Jerez, Bharati Jadhav, Wenyu Huang, Rajarshi Mondal, Ben Busby, Nathan D. Olson, Fritz J. Sedlazeck, Justin Wagner, Sina Majidian, and Justin M. Zook. The GIAB genomic stratifications resource for human reference genomes. *Nature Communications*, 15(1):9029, 2024. doi: 10.1038/s41467-024-53260-y. URL https://doi.org/10.1038/s41467-024-53260-y.

Haonan Feng, Lang Wu, Bingxin Zhao, Chad Huff, Jianjun Zhang, Jia Wu, Lifeng Lin, Peng Wei, and Chong Wu. Benchmarking DNA foundation models for genomic and genetic tasks. *Nature Communications*, 16(1):10780, 2025. doi: 10.1038/s41467-025-65823-8. URL https://doi.org/10.1038/s41467-025-65823-8.

Nancy F. Hansen, Nathan Dwarshuis, Hyun Joo Ji, Arang Rhie, et al. A complete diploid human genome benchmark for personalized genomics. *Cell*, 189(16):4857–4875.e31, 2026. doi: 10.1016/j.cell.2026.06.016. URL https://doi.org/10.1016/j.cell.2026.06.016.

Yukun He, Yanan Chu, Shuming Guo, Jiang Hu, Ran Li, Yali Zheng, Xinqian Ma, Zhenglin Du, Lili Zhao, Wenyi Yu, Jianbo Xue, Wenjie Bian, Feifei Yang, Xi Chen, Pingan Zhang, Rihan Wu, Yifan Ma, Changjun Shao, Jing Chen, Jian Wang, Jiwei Li, Jing Wu, Xiaoyi Hu, Qiuyue Long, Mingzheng Jiang, Hongli Ye, Shixu Song, Guangyao Li, Yue Wei, Yu Xu, Yanliang Ma, Yanwen Chen, Keqiang Wang, Jing Bao, Wen Xi, Fang Wang, Wentao Ni, Moqin Zhang, Yan Yu, Shengnan Li, Yu Kang, and Zhancheng Gao. T2T-YAO: A telomere-to-telomere assembled diploid reference genome for Han Chinese. *Genomics, Proteomics & Bioinformatics*, 21(6):1085–1100, 2023. doi: 10.1016/j.gpb.2023.08.001. URL https://doi.org/10.1016/j.gpb.2023.08.001.

C. Huang, R. W. Shuai, P. Baokar, et al. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, 55:2056–2059, 2023. doi: 10.1038/s41588-023-01574-w. URL https://doi.org/10.1038/s41588-023-01574-w.

InstaDeepAI. NTv3_100M_pre. Hugging Face model card, n.d. URL https://huggingface.co/InstaDeepAI/NTv3_100M_pre. Accessed 22 September 2026.

Wen-Wei Liao, Mobin Asri, Jana Ebler, et al. A draft human pangenome reference. *Nature*, 617:312–324, 2023. doi: 10.1038/s41586-023-05896-x. URL https://doi.org/10.1038/s41586-023-05896-x.

Phuc-Loi Luu, Phuc-Thinh Ong, Thanh-Phuoc Dinh, and Susan J. Clark. Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data. *NAR Genomics and Bioinformatics*, 2(3):lqaa054, 2020. doi: 10.1093/nargab/lqaa054. URL https://doi.org/10.1093/nargab/lqaa054.

Juan F. Macias-Velasco, Xiaoyu Zhuo, Chad Tomlinson, Eddie A. Belter, Milinn Kremitzki, Derek Albracht, Tina Lindsay, Xiaoyun Xing, Nina Tekkey, Wenjin Zhang, John E. Garza, Zheng Xu, Zilan Xin, Qichen Fu, Heather A. Lawson, Nathan O. Stitziel, Robert S. Fulton, Daofeng Li, Human Pangenome Reference Consortium, and Ting Wang. Benchmarking genome choice in functional genomics analyses. *Nature Communications*, 17:6837, 2026. doi: 10.1038/s41467-026-73663-3. URL https://www.nature.com/articles/s41467-026-73663-3. Published 26 May 2026.

Frederikke Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA language models on biologically meaningful tasks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 15246–15281, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/429e7b31625a8b7839f9e4d6e2aa9bb9-Paper-Conference.pdf.

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 Baccus. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201. Curran Associates, Inc., 2023. doi: 10.52202/075280-1872. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/86ab6927ee4ae9bde4247793c46797c7-Paper-Conference.pdf.

Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Nae-Chyun Chen, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G. S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson,

Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Ying Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, and Adam M. Phillippy. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022. doi: 10.1126/science.abj6987. URL https://doi.org/10.1126/science.abj6987.

Aman Patel, Arpita Singhal, Austin Wang, Anusri Pampari, Maya Kasowski, and Anshul Kundaje. DART-Eval: A Comprehensive DNA Language Model Evaluation Benchmark on Regulatory DNA. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62024–62061, 2024. doi: 10.52202/079017-1981. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/71998bfc3217ffe1cca1ee084dfadadd-Abstract-Datasets_and_Benchmarks_Track.html.

Melissa Sanabria, Jonas Hirsch, Pierre M. Joubert, and Anna R. Poetsch. DNA language model GROVER learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8): 911–923, 2024. doi: 10.1038/s42256-024-00872-0. URL https://doi.org/10.1038/s42256-024-00872-0.

Prasad Sarashetti, Josipa Lipovac, Qi Jia, Ling Wang, et al. A complete telomere-to-telomere diploid reference genome for south asian population. bioRxiv preprint, 2025. URL https://doi.org/10.1101/2025.07.12.664550.

Alexander Sasse, Bernard Ng, Anna E. Spiro, Shinya Tasaki, David A. Bennett, Christopher Gaiteri, Philip L. De Jager, Maria Chikina, and Sara Mostafavi. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nature Genetics*, 55:2060–2064, 2023. doi: 10.1038/s41588-023-01524-6. URL https://doi.org/10.1038/s41588-023-01524-6.

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. URL https://proceedings.mlr.press/v235/schiff24a.html.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 41642–41665, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b633e7052970b8f5aa1a69164d99e9e8-Paper-Conference.pdf.

The appendices follow the benchmark workflow: construct paired inputs, establish output correspondence, fit and score models, then interpret performance and correctness. Appendices A–I support the current main text. Appendix J retains earlier task-specific experiments under their recorded protocols.

For the main findings, read B.2–B.4 with Figure 2, E with Table 2 and Figure 3a–e, and F.3–F.6 with Figure 3f. Appendix I and the reproduction guide explain which results can be reconstructed from the retained material.

Appendix	Contents	Main-text connection
A	Benchmark construction and input provenance	3.1
B	Output correspondence and matching diagnostics	3.2; 4.1–4.2
C	Encoders, prediction heads and training protocol	3.3
D	Metrics, weighting and statistical inference	3.2–3.3
E	Complete primary performance and consistency results	4.3
F	Paired correctness and probability rankings	4.3–4.4
G	Sensitivity and annotation checks	4.2–4.4
H	Supervised comparators	3.3; 4.3
I	Reproducibility and verification	3.1–3.3
J	Earlier task-specific experiments	supplementary context

Literature references are collected in References. The current claim-to-file index links each main result to its appendix and recorded evidence.

A Benchmark construction and input provenance

A.1 Task definitions

The direct annotation inputs are the YAO-coordinate gene GFF3, `yao_matchrY_preciseannofix.gff3`, and cCRE BED, `YAO_cCREs_final.bed`. T1–T3 are extracted from the gene/transcript and exon annotations; T4 uses promoter-like cCREs (PLS), and T5 uses proximal and distal enhancer-like cCREs (pELS/dELS). Candidate counts enumerate task records after builder filters; windows may overlap. Their direct YAO-coordinate origin is distinct from their author-confirmed upstream RefSeq/ENCODE4 transfer provenance; the frozen input files have been hash-checked, whereas complete upstream processing records remain unresolved.

The GFF3 distributed in the official T2T-YAO v2.0 public resource release (He et al., 2023; Chu et al., 2026) decompresses to exactly the production annotation (1,281,335,154 bytes; SHA-256 `c8db1dcec3687ee801f793182964edb32fd4bd04101c0d7e28187976fa74a3f8`). This establishes the direct gene-annotation download binding. Appendix I.5 lists the remaining upstream transfer records. The public cCRE Registry V4 additionally supports a task-specific annotation check in Appendix G.3; it measures agreement with the recorded annotation category.

The archived freeze implementation records the following construction rules. T1 and T3 subsequently use the repaired original v2 cohort; their accepted counts and quarantine scope are in Appendix A.4.

Task	Positive target and window anchor	Background definition
T1	Annotated donor/acceptor; 402 bp window anchored at the second base of the strand-oriented GT/AG motif	Genomic GT/AG decoys outside annotated splice sites
T2	Base-level donor/acceptor labels in 2,004 bp gene-centered windows	Remaining valid positions; no separate negative-window dataset
T3	Unique transcript TSS boundary; 300 bp window	GC-matched background windows not containing an annotated TSS

Task	Positive target and window anchor	Background definition
T4	Promoter-like cCREs within the recorded common GC support; 510 bp window centered on the cCRE	Background not overlapping promoter-like cCREs
T5	Proximal or distal enhancer-like cCREs; 510 bp window centered on the cCRE	Background not overlapping enhancer annotations

Coordinates. Coordinates use zero-based, half-open intervals. Even-length windows start at center minus half the window length.

T1 anchoring. Each transcript contributes strand-relative intronic donor and acceptor dinucleotides; identical sites are deduplicated by contig, anchor, class, and strand before sampling. For a two-base site with $c = (\text{site_start}+\text{site_end})/2$, the recorded genomic-forward interval is $[c-201, c+201)$; the input is reverse-complemented for a negative-strand site. Thus the strand-oriented dinucleotide occupies oriented input indices 200–201 on both strands. On the positive strand, those indices map to genomic coordinates $c-1, c$; on the negative strand, after reverse complementation, they map to $c, c-1$. The second strand-oriented motif base is therefore input index 201 in both cases, while its genomic coordinate is c on the positive strand and $c-1$ on the negative strand. The frozen catalog and sequences verify this interval/orientation relation; Appendix G.7 separately reports central-motif compatibility exceptions.

T2 labels. T2 retains genomic-forward sequences and combines all transcript-derived sites in a window. The frozen three-class builder writes donor spans before acceptor spans, so a donor/acceptor overlap has deterministic acceptor-last precedence.

T3 boundaries and backgrounds. The positive-strand anchor is min(exon.start0), while the negative-strand anchor is max(exon.end0), an interbase boundary. For the negative strand, the boundary is end0 rather than the last included base, end0-1. T3 deduplicates positives by contig, boundary, and strand; negatives exclude every recorded TSS boundary from the sampled window and are GC-matched at a 1:1 target ratio. This distinction is retained in input preparation and matters when interpreting a one-position offset.

T4/T5 scope. T4/T5 labels represent the recorded cCRE category. Classification backgrounds and GC restrictions follow the frozen task-specific construction; Appendices A.4 and A.5 report the resulting class counts after repair.

A.2 Manuscript task names and original experimental keys

Paper ID	Task	Original key
T1	Splice-site classification	`A2`
T2	Position-resolved splice prediction	`A3`
T3	Transcription-start-site classification	`A4`
T4	Promoter classification	`A7`
T5	Enhancer classification	`A8`

The mapping changes presentation only. Original CSV task keys, paths, checkpoints and historical audit figures retain their original names. No data or results are relabeled at source.

A.3 Reference and annotation provenance

The YAO autosomal inputs come from the maternal (mat) assembly of T2T-YAO v2.0, as confirmed by the authors. The production file is `mat_chrY.fasta` (SHA-256 `31f4c4f51fb48fbb47158933f67726498a22f1ef9ddb4fd8bc5b6d027d75bdc6`). The production-reference identity record lists its 24 chromosome lengths. Paired evaluation uses the autosomes; the full-reference level-4 derivation also includes chrX and chrY. The source identity record associates the author-confirmed assembly and haplotype with this recorded file identity. Appendices I.2 and B.2 document checks of the fixed sequences and output maps, respectively.

Encoder-file identities, short weight-hash prefixes and replay coverage are summarized in Appendix C.2; complete file hashes are retained in the author-side inventory described in Appendix I.5.

The derived YAO GFF3 and cCRE BED were rehashed and match the frozen annotation audit. RefSeq release identification and the ENCODE4 family remain author-confirmed; exact upstream annotation checksums and transfer records are incomplete. The recorded FASTA hashes and GIAB source checks remain available. A public-file checksum match for the production FASTA and complete YAO-stratification production bindings remain outstanding.

Evaluating a new predictor uses the fixed paired sequences, shared targets and validation/test maps. Reconstructing benchmark construction from upstream assemblies also requires the production-file identities, public-file checksum bindings and complete annotation-transfer records. Reproducing the historical model results additionally requires the encoder execution identities and recorded training settings in Appendix C. Appendix I.5 records access and redistribution status for the fixed-input and analysis packages.

The author-side file inventory and detailed provenance audit are indexed in Appendix I.5.

A.4 Original-cohort repair flow

Task	Split	Before QC	Quarantined	Retained	Class 0 / 1 / 2
T1	train	50000	3	49997	24997 / 12500 / 12500
T1	valid	8000	0	8000	4000 / 2000 / 2000
T1	test	8000	0	8000	4000 / 2000 / 2000
T3	train	49808	639	49169	24544 / 24625
T3	valid	7904	247	7657	3814 / 3843
T3	test	7988	58	7930	3945 / 3985

For each original cohort, the count before QC equals the sum of retained and quarantined identities. Counts and quarantine files match their existing v2 validation hashes. Coordinate-only corrections preserve identities. T1 class order is background/donor/acceptor; T3 is background/positive. The original T1 repair record additionally reports 594/192/35 coordinate corrections in train/validation/test; these corrections precede filtering and are accounted for separately from removals.

A.5 Additional training counts and test composition

Task	Train windows	Validation windows	Training class counts	Validation class counts
T2	14077	2162	28116926 / 45425 / 47957	4319351 / 6405 / 6892
T4	28540	4512	14270 / 14270	2256 / 2256
T5	49998	8000	24999 / 24999	4000 / 4000

T2 counts valid position observations in background/donor/acceptor order; overlapping windows can count the same genomic base more than once. Binary order is background/positive. T1/T3 counts appear in Appendix A.4. New-experiment training/validation identity and label checks cover all model/reference settings.

The retained binary test cohorts have the following class composition.

Task	Test windows	Background	Positive	Positive fraction
T3	7930	3945	3985	50.2522%
T4	5780	2890	2890	50.0000%
T5	8000	4000	4000	50.0000%

T2 has 5,078,078 mapped valid position observations: 15,960 donor/acceptor observations and 5,062,118 background observations (0.3143% foreground). A genomic base appearing in overlapping windows contributes one observation per window. Binary counts come from the retained T3,

T4, and T5 test metadata. The dense denominator is checked across saved event summaries; the review derivation records the source hashes.

A.6 COHORT CONSTRUCTION, SELECTION AND ACCEPTED-TEST COVERAGE

The historical census begins with task-specific, builder-eligible positive identities; T2 instead counts informative gene-centered windows. Center-lift success is the earlier query status `ok`, before final exclusion of cross-chromosome/tied hits, overlap checks, T2 label-conservation checks, quota sampling and GC matching. Final inclusion also depends on the sampling and quality-control stages below.

The first four numerical columns count positive identities or informative T2 windows across all splits. The final column also includes sampled background windows for classification tasks. Sampling, background construction and refills connect these distinct selection stages.

Task	Builder-eligible	Center liftOK	Earlier selected	Paper-freeze positive/dense	Paper-freeze all windows
T1	541,418	538,870	33,000	33,000	66,000
T2	19,569	19,325	19,315	18,783	18,783
T3	73,585	72,813	33,000	32,850	65,700
T4	20,971	20,954	19,612	19,416	38,832
T5	1,643,361	1,642,768	33,000	32,999	65,998

Identity joins recover 5,251 removals and 4,025 refills between the earlier 256,539 selected windows and the 255,313-window paper freeze. All final positive/informative identities join to the historical eligible pool. The historical query snapshot records 4,153 `unmapped` queries and 21 sex-chromosome-target exclusions; these are recorded query outcomes. Available removal logs account for all 532 removed T2 windows: 63 cross-chromosome, 73 cross-split target overlaps, 273 within-split target overlaps, 70 unmapped labels and 53 out-of-bounds labels. Across the remaining tasks, 288 removed identities lack a recorded per-identity reason. The logged label failures identify unmapped and out-of-bounds labels, but do not recover every condition of the historical label-conservation check; a full-pool replay is unavailable.

The later geometry repair is a separate stage (Appendix A.4). In the original cohort it removes three T1 training windows and 639/247/58 T3 train/validation/test windows. The T1 test retains its 8,000 IDs; 35 GRCh38 test inputs have the recorded coordinate correction. T3 retains 7,930 of its 7,988 paper-freeze test windows. T2/T4/T5 retain their paper-freeze inputs. The cohort totals reconcile as $255{,}313-(3+639+247+58)=254{,}366$: 255,313 is the pre-repair paper freeze, whereas 254,366 is the accepted benchmark cohort. Mapping coverage and repair retention use their respective input cohorts as denominators.

Accepted-test composition is counted once per task-window identity, after these repairs, using the input hashes and saved full-window alignments. In particular, T1 now has 5,432 identical-input windows, compared with 5,423 before its coordinate correction. Sequence classes are assigned from the saved full-window alignments.

Task	Test windows	Identical	Substitution-only	Indel-containing	Identical fraction
T1	8,000	5,432	1,846	722	67.90%
T2	2,544	774	854	916	30.42%
T3	7,930	5,769	1,592	569	72.75%
T4	5,780	3,318	1,713	749	57.40%
T5	8,000	4,668	2,281	1,051	58.35%

The accepted test windows also retain support from all three recorded mapping levels. Levels identify the mapping procedure recorded for each window center.

The complete level counts are provided in the mapping-level table.

T2 contains 5,098,176 input window-position observations. Of these, 5,078,078 belong to the valid mapped comparison domain and 20,098 lie outside it. Within the mapped domain, 563,787 observations (11.10%) have nonzero output-index displacement, spanning 908 of 2,544 windows (35.69%). Thus correspondence affects a substantial subset of accepted windows even though most compared positions have zero displacement. The table partitions position observations by absolute displacement and, separately, by the minimum distance to either input-window edge. Foreground means donor or acceptor. The axes are separate marginal partitions; one window can contribute to several bins. The 15,960 foreground observations count each base once per containing window.

Axis	Bin (bp)	Position observations	Domain fraction	Foreground observations
Index shift	0	4,514,291	88.90%	14,500
Index shift	1	243,507	4.80%	732
Index shift	2-4	168,446	3.32%	393
Index shift	5+	151,834	2.99%	335
Edge distance	0-7	40,499	0.80%	108
Edge distance	8-31	121,423	2.39%	367
Edge distance	32-127	485,971	9.57%	1,431
Edge distance	128+	4,430,185	87.24%	14,054

These counts describe the cohort selected by quota sampling, GC matching, mapping restrictions and QC. Test IDs and input hashes match freeze/repair ancestors; T2 mapping and labels agree on IDs and mapped labels. Counts use no predictions or repetitions across models, seeds or readouts. The artifact includes source hashes, identity coverage, T2 window geometry and complete stage/repair tables. Sources: audit validation, source manifest, historical stages, later repair, current test identities, and T2 position coverage.

A.7 Resource organization and evaluation scope

We provide accepted paired windows with labels, correspondence maps, and regional evaluation records, alongside a separate set of level-4 intervals. We derive these intervals as the complement of the recorded level 1–3 YAO interval union on the production reference. Levels 1–3 support pairing subject to task-specific validity checks. Paired window centers lie within recorded level 1–3 intervals, while input context can extend beyond those intervals. Appendix A.9 documents this derivation, the exact input identities, and the difference from the source-reported level-4 totals.

The five-task candidate resource is extracted from YAO-coordinate gene and cCRE annotations. Gene/transcript and exon structures supply T1–T3, while cCRE classifications supply T4–T5. Across the five tasks, the historical builder-eligible pool contains 2,298,904 records. Of these, 2,294,730 have level1–3 center-mapping records: 1,561,545 at level1, 643,476 at level2, and 89,709 at level3. Counts sum task identities across the three splits. Center mapping is followed by window and label checks. Appendix A.1 identifies the direct annotation files and their upstream transfer provenance; Appendix A.6 reports construction stages.

After task-specific checks, sampling, background construction and subsequent repairs, the current benchmark retains 254,366 paired task windows across all splits, including 32,254 test pairs. The historical paper-freeze total before these repairs was 255,313 pairs. Each accepted task window contributes one record to these totals.

Of the 254,366 accepted paired windows, 61 have input intervals that partially overlap the derived level-4 regions, including 10 test windows. No paired window center lies in level 4, and no T2 test window overlaps level 4. Here, a window center is the integer midpoint of its zero-based, half-open YAO interval. Counts use task-window records across the frozen splits, including overlapping windows. Input-context overlap is distinct from the recorded output correspondence required for paired scoring. The interval and overlap audit records the input hashes and affected window identities.

Paired dense scoring is restricted to valid mapped outputs. The historical task-query audit records 4,153 unmapped candidates and 21 exclusions involving target sex chromosomes; these counts summarize the queried task candidates. Appendix A.9 reports the derived level-4 interval resource separately from the source-reported summary.

A.8 FILES, FIELDS, AND INTENDED USES

The resource catalog consolidates the accepted benchmark records and supporting regional annotations. The exported window manifest uses the original frozen T2/T4/T5 records and the accepted original-cohort T1/T3 repair records. The export preserves the accepted IDs and split assignments. Table A1 lists the files available in the local review materials; this catalog has not been publicly released.

Table A1. Resource components and their uses. The data dictionary specifies coordinate and array conventions. File hashes and source bindings are recorded in the manifest.

Component	Files and key fields	Intended use
Accepted paired windows	`paired_windows.tsv.gz`: task, split, sample ID, cohort version, label, YAO/GRCh38 intervals, annotation strands, input orientations, mapping level, and input sequence hashes	Identify and extract the same accepted inputs across references; preserve the frozen split and trace each pair
Dense test targets and correspondence	`T2_test_targets.npz`: sample IDs, labels and validity masks on both references; `T2_test_base_map.npz`: sample IDs and GRCh38 index for each YAO position	Compare mapped valid outputs and reconstruct paired correctness states; no equal-index substitution
Genomic context annotations	`giab_membership.csv.gz`: task/sample ID, membership definition, stratum, YAO/GRCh38 membership flags, and intervals	Select regional subsets while distinguishing annotation membership from mapping level
Composition controls	`giab_covariates.csv`, `giab_matched_pairs.csv`, and `giab_support.csv`: measured covariates, matched group-11/group-00 IDs, retention and block support	Reuse the same regional comparisons across fitted heads and audit common support
Analysis and provenance	`giab_primary_contrasts.csv`, `paired_window_counts_and_coverage.csv`, `source_manifest.json`, and `validation.json`	Inspect effect estimates, uncertainty, cohort coverage, source bindings, and completed checks
Separate level-4 intervals	`level4_derived/`: full-reference and autosomal BED files, chromosome lengths, per-chromosome counts, source chain, derivation script, and checksums	Reuse complement intervals as a separate resource; not interpreted as paired targets or functional copy counts
Fixed task inputs	`fixed_benchmark/data/T*_*.npz`: oriented sequences, labels and coordinates for T1–T5; T2 validation/test output maps	Direct model evaluation without production FASTAs

Table A2. Accepted window counts and nonoverlapping input coverage. Coordinates are zero-based, half-open. Coverage is the interval union within each reference, calculated separately by task across all splits; the final row merges across all five tasks as well. Overlapping or adjacent intervals are merged within chromosomes, without merging between references. Counts describe the selected benchmark inputs. The final row recomputes the cross-task interval union.

Task	Training windows	Validation windows	Test windows	Total windows	YAO union (bp)	GRCh38 union (bp)
T1	49,997	8,000	8,000	65,997	26,091,573	26,081,270
T2	14,077	2,162	2,544	18,783	37,356,289	37,324,838

Task	Training windows	Validation windows	Test windows	Total windows	YAO union (bp)	GRCh38 union (bp)
T3	49,169	7,657	7,930	64,756	19,224,777	19,218,237
T4	28,540	4,512	5,780	38,832	19,219,600	19,177,055
T5	49,998	8,000	8,000	65,998	33,570,501	33,542,187
All tasks	191,781	30,331	32,254	254,366	125,783,479	125,560,330

The coordinate manifest records hashes rather than nucleotide strings. The `fixed-input.zip` component of the current local release candidate supplies the exact oriented sequences and labels for all accepted T1–T5 train/validation/test windows, plus T2 validation and test output maps and a unified scorer. Using these fixed inputs no longer requires the production FASTAs. Rebuilding the package from upstream assemblies uses the production-file identities and construction records in Appendix A.3, including the author-confirmed T2T-YAO v2.0 maternal source. Gene and cCRE intersections with level-4 intervals have not been quantified. The candidate includes the fixed-input package and coordinate catalog; external reviewer access remains pending (Appendix I.5).

A.9 Derivation and source comparison of level-4 intervals

We operationally define level 4 as the complement of the chromosome-wise union of all level 1–3 YAO intervals in the supplied `fourlevel.chain`. Columns 1, 3, and 4 specify YAO chromosome, start, and end; column 13 specifies mapping level. Following the existing benchmark loader, starts and ends are interpreted as zero-based, half-open coordinates. All 22,410 level 1–3 records contribute, including multiple and cross-chromosome mappings. Overlapping and adjacent YAO intervals are merged before taking the complement over each complete chromosome, including terminal regions. No minimum-length or task-overlap filter is applied.

The derived interval resource includes the source chain, chromosome lengths, BED files, per-chromosome statistics, and a standalone derivation script. The chain SHA-256 begins `897092438d56`; the production FASTA SHA-256 begins `31f4c4f51fb48`, matching the frozen input identity in Appendix A.3. A fresh scan verifies all 24 chromosome lengths, totaling 3,074,790,210 bp. An independent endpoint-depth sweep reproduces every complement interval, and mapped-union length plus complement length equals each chromosome length.

Table A3. Level-4 intervals derived from the pinned chain and production reference. Counts enumerate disjoint complement intervals. The full reference includes chr1–chr22, chrX, and chrY; the autosomal subset matches the chromosome scope of the paired benchmark and is supplied as a separate interval resource.

Reference scope	Reference length (bp)	Level-4 intervals	Nonoverlapping level-4 length (bp)
Full reference	3,074,790,210	1,530	91,494,498
Autosomes	2,868,185,707	1,406	79,352,637

Interval length distribution. The 1,530 full-reference intervals range from 1 to 8,535,019 bp, with a median length of 2,407.5 bp and a 90th percentile of 119,873.8 bp, computed by linear interpolation between adjacent order statistics. The resource retains 140 single-base intervals because no minimum-length filter is applied. Lengths are measured on the complement of the recorded mapping intervals.

The archived YAOBridge documentation separately reports 1,591 level-4 blocks, a cumulative length of 94,184,462 bp, and a coverage fraction of 0.030631. Its maternal-haplotype-plus-chrY scope has the same total reference length. These historical totals differ from the present derivation by 61 intervals and 2,689,964 bp; their exact chain version and generation rules remain unresolved. The two totals are retained with their respective derivation records. The source audit records the distinction. Gene/cCRE overlap has not been quantified for this interval resource.

B OUTPUT CORRESPONDENCE AND MATCHING DIAGNOSTICS

B.1 REFERENCE PAIRING AND ACCEPTED EVALUATION UNITS

The archived YAOBridge documentation defines level 1 as a conserved framework based on single-copy orthologous genes and their flanks, preserving synteny and high identity/coverage across T2T human genomes. Level 2 contains uniquely aligned inter-framework regions retaining YAO–GRCh38 synteny and high identity/coverage. Level 3 contains inter-framework regions with lower identity/coverage or multiple valid alignments. These are the source procedure's definitions; the recovered README does not specify every numerical threshold needed to reconstruct their assignment. Level 4 in the local resource candidate is explicitly the complement of the recorded level 1–3 YAO interval union, as defined in Appendix A.7. The archived README and its hash accompany the input-binding audit.

The archived YAOBridge freeze procedure excludes gap hits, sex chromosomes, out-of-bounds windows, tied top mappings, and cross-chromosome mappings from the main cohort. Its mapping selection prioritizes the recorded lower levels within L1/L2 and then higher scores; L3 uses the highest score. These rules define the archived mapping selection. Task-specific repair and the accepted manifests determine the final cohort used in this paper.

Table B1. Recorded center-mapping levels of accepted paired windows. Counts come from the final catalog, after construction and repairs. Each cell gives all-split windows followed by test windows in parentheses. Each cell counts task-specific window records. Windows retain the level of their recorded center correspondence even when their context extends beyond that interval. The catalog stores this value in the `mapping_level` field, which defines the subsets in Appendix B.4.

Task	Level 1	Level 2	Level 3	Total
T1	47,366 (5,771)	14,316 (1,780)	4,315 (449)	65,997 (8,000)
T2	14,890 (2,054)	2,668 (400)	1,225 (90)	18,783 (2,544)
T3	45,417 (5,693)	15,385 (1,827)	3,954 (410)	64,756 (7,930)
T4	26,193 (3,785)	9,505 (1,545)	3,134 (450)	38,832 (5,780)
T5	37,423 (4,328)	23,696 (3,105)	4,879 (567)	65,998 (8,000)
All tasks	171,289 (21,631)	65,570 (8,657)	17,507 (1,966)	254,366 (32,254)

Frozen BED/H5 records preserve sample IDs, split assignments, coordinates, orientation, and labels. T2 uses genomic-forward sequences and labels; other tasks use their recorded strand orientation. For each YAO base, the frozen T2 map stores the corresponding GRCh38 index, using -1 when correspondence is missing. Evaluation retains positions with a map and valid labels on both sides, then verifies mapped label equality. Overlapping windows can contain repeated observations of a genomic base. The 5,078,078 mapped valid observations are therefore counted within their 2,544 containing windows.

Selected-chain orientation is applied before array indexing. For a reverse-oriented chain, source genomic position p is converted to the chain query coordinate $q = \text{source_start} + \text{source_end} - p - 1$, aligned through the selected blocks, and then converted to the GRCh38 index in its genomic-forward input window. Eight test windows and 79 validation windows use this rule; all other T2 maps use forward-oriented chains. The stored map therefore always indexes the genomic-forward GRCh38 prediction array.

The internal reconstruction entry retains 188,369 task-by-split paired identities for T2–T5 and binds 24 input files. Its full preparation check reproduced every array in those files. The check validates preparation from the frozen coordinates and labels. T1 retains a separate recorded execution path.

B.2 STRUCTURAL CORRESPONDENCE AND POSITION ACCOUNTING

T2 output correspondence follows the aligned blocks of the selected YAOBridge chain. The archived geometry audit recovers that chain from the source and target center coordinates, chromosome, direction, mapping level and score rounded to the recorded six decimal places. Exactly one chain satisfies these recorded criteria for each accepted T2 window; uniqueness is assessed within the

supplied records. The chain files and archived parser are content-pinned against the earlier audit. The source query is reflected within the chain interval for a reverse-direction chain; positions within an aligned block map by their offset from the block start. Source gaps and positions beyond the chain are unmapped. Absolute target coordinates are then converted to genomic-forward target-window indices, retaining -1 for positions without an in-window correspondence.

Reconstruction reproduces every entry of the existing test map. Applying the same rule without reading predictions supplies the validation map. A separate portable implementation reconstructs both maps from retained selected-chain blocks. Coordinates, IDs, range, injectivity and complete foreground-label retention are checked. Mapping is structural: label agreement is asserted after mapping rather than used to search for a corresponding position or silently remove disagreements. These mapping and agreement checks apply to the accepted cohort after the construction-stage exclusions in Appendix A.6. The position-accounting table below reports evaluation-domain exclusions within that cohort. There are no label conflicts in the retained valid domains.

Position-accounting stage	**Test**	**Validation**
Windows	2,544	2,162
All window-position observations	5,098,176	4,332,648
No aligned source block, including positions outside the selected chain	14,077	11,162
Aligned target position outside the paired window	6,021	3,020
Invalid Y label after those exclusions	0	0
Invalid mapped H label after those exclusions	0	0
Retained mapped observations	5,078,078	4,318,466
Label conflicts within retained observations (check, not a filter)	0	0
Retained donor/acceptor observations	15,960	13,297

All 2,544 T2 test windows retain at least one mapped valid position; the scorer excludes an empty window from window-averaged rates. The first four exclusion categories are sequential and mutually exclusive. Foreground counts and the conflict check describe retained positions. Counts use one observation per position within each window. Alignment blocks define correspondence; mapped-label agreement checks the shared annotation target. Mapping-level and motif analyses assess sensitivity to these recorded sources of support.

B.3 Effect of correspondence on the readout contrast

For each model, training reference, and seed, we calculate $I = (F_{\mathrm{equal},d4} - F_{\mathrm{equal},d1}) - (F_{\mathrm{map},d4} - F_{\mathrm{map},d1})$ using the same valid positions. F_{map} follows the recorded output correspondence; F_{equal} pairs equal output indices. Window rates receive equal weight, and the three fixed-seed means are averaged. The additional equal-index validity requirement removes no observations from the 5,078,078-position mapped domain. All 72 paired settings reproduce the previously archived rates before computing the new contrasts. All five frozen models and the one-hot control are retained.

This post-hoc analysis was specified before calculating the new intervals. For each block width, 5,000 chromosome-stratified bootstrap draws use source-window midpoints, with the same block multiplicities across rules, readouts, models and seeds (RNG seed 20260916). There are 340 occupied 1 Mb blocks and 193 occupied 2 Mb blocks. Each replicate computes the difference of differences using shared block multiplicities. Percentile intervals are pointwise, unadjusted and conditional on the frozen cohort and fitted seeds.

Table B2. Mean T2 readout contrasts under two correspondence rules. Values are three-seed mean disagreement differences (dilation 4 minus dilation 1), in percentage points; $I = \Delta F_{\text{equal}} - \Delta F_{\text{map}}$. Opposite signs mark mean sign reversals. The first ten rows are frozen-encoder settings: $|I| > |\Delta F_{\text{map}}|$ in nine, and six reverse sign. One-hot rows are controls.

Model	Training reference	Mapped ΔF	Equal-index ΔF	I	Sign reversal
HyenaDNA	Y	+0.0829	-0.1005	-0.1834	yes
HyenaDNA	H	+0.0777	-0.1372	-0.2149	yes
Caduceus-PH	Y	+0.0910	+0.3233	+0.2323	no
Caduceus-PH	H	+0.0905	+0.3421	+0.2516	no
DNABERT-2	Y	+0.0871	-0.1584	-0.2455	yes
DNABERT-2	H	+0.0788	-0.1563	-0.2351	yes
GROVER	Y	+0.1272	+0.1162	-0.0110	no
GROVER	H	+0.1085	-0.1556	-0.2642	yes
NTv3-100M-pre	Y	-0.0230	-0.1703	-0.1473	no
NTv3-100M-pre	H	+0.0068	-0.1390	-0.1458	yes
one-hot	Y	+0.0891	+3.0488	+2.9597	no
one-hot	H	+0.0912	+3.0273	+2.9361	no

Table B3. Pointwise, unadjusted 95% block-bootstrap intervals for the correspondence-induced shift. Intervals condition on the fitted seeds and use source-window midpoint blocks at the two stated widths. The one-hot rows are controls.

Model	Training reference	I (pp)	1 Mb 95% interval	2 Mb 95% interval
HyenaDNA	Y	-0.1834	[-0.2204, -0.1485]	[-0.2223, -0.1490]
HyenaDNA	H	-0.2149	[-0.2515, -0.1804]	[-0.2526, -0.1819]
Caduceus-PH	Y	+0.2323	[+0.2065, +0.2578]	[+0.2061, +0.2597]
Caduceus-PH	H	+0.2516	[+0.2238, +0.2791]	[+0.2233, +0.2811]
DNABERT-2	Y	-0.2455	[-0.2770, -0.2182]	[-0.2756, -0.2187]
DNABERT-2	H	-0.2351	[-0.2589, -0.2121]	[-0.2592, -0.2119]
GROVER	Y	-0.0110	[-0.0323, +0.0153]	[-0.0335, +0.0154]
GROVER	H	-0.2642	[-0.2925, -0.2375]	[-0.2953, -0.2351]
NTv3	Y	-0.1473	[-0.1721, -0.1267]	[-0.1721, -0.1270]
NTv3	H	-0.1458	[-0.1828, -0.1180]	[-0.1825, -0.1178]
one-hot	Y	+2.9597	[+2.7096, +3.2160]	[+2.7092, +3.2268]
one-hot	H	+2.9361	[+2.6886, +3.1900]	[+2.6862, +3.2006]

Nine of ten frozen settings have intervals excluding zero at both widths. GROVER Y has an interval spanning zero. The two matching rules give opposite signs for the mean readout contrast in six settings. Three reverse in all three fitted seeds, and the other three reverse in one or two seeds. The underlying contrasts are retained in the historical review-v30 source record, with model, training reference, seed, both readout effects and their difference. The per-seed records describe variation among fitted runs; the block intervals quantify conditional genomic sampling uncertainty.

B.4 Correspondence effects by recorded mapping level

The same per-window equal-index-minus-mapped readout-effect differences used in Figure 2 are grouped by the catalog's `mapping_level` field, which records the mapping level at each window center (Appendix B.1). Each subset retains identical observations under the two scoring rules, gives windows equal weight, and averages seeds 14/15/16. All full-cohort means are independently recovered from the retained window records before subsetting. These are descriptive estimates, with no new bootstrap intervals or hypothesis tests.

Subset	Windows	Mean effect range (percentage points), ten frozen settings	Mean directions agreeing with full cohort
L1	2,054	-0.2593 to +0.2508	10/10
L1_L2	2,454	-0.2536 to +0.2476	10/10
L3	90	-0.5739 to +0.3618	9/10

The mean-shift directions persist in the L1 and L1/L2 subsets. Seed variation and all one-hot results remain in the accompanying CSV tables.

B.5 Complete-panel correspondence and prediction diagnostics

These exploratory diagnostics were specified after inspecting the original results and before computing the new stratified summaries. They retain all 576 T2–T5 predictions and 288 fixed-head pairs, including the existing one-hot control. Task versions, labels, seeds and test splits remain fixed. T2 counts use window-by-position observations, with the same windows evaluated across model settings.

B.5.1 Correspondence on a common domain

The mapped valid domain is intersected with GRCh38 validity at the equal index. Here the comparison retains all 5,078,078 mapped valid observations in 2,544 windows. The equal-index diagnostic retains label mismatches and the four joint flip states in the accompanying CSV and reports disagreement. Task accuracy uses the valid labels for each input. Independent per-window recounting agrees on all 72 paired settings and 144 prediction files, with a maximum rate discrepancy of 1.39×10^{-17}. The three-seed mean contrasts and sign reversals are consolidated in Table B2; the author-side source CSV retains each seed and the sample SD.

Table B2 brings the matched-domain mapped and equal-index comparisons together with I; the independent per-window recounting above verifies these means.

B.5.2 Fixed-subset checks and geometry

Appendix B.5.3 reports the identical-input counts separately for zero-displacement and one-base-displaced observations.

For each subgroup, the flip rate is calculated within each window containing subgroup observations, then averaged equally across those windows. Each subgroup therefore has its own window denominator. The following differences use the same frozen membership across readouts. Foreground contains 15,960 position observations per setting. The interior criterion uses the minimum distance to either boundary of both corresponding outputs; contextual encoder features can extend beyond that distance. All models and both training references are retained.

Model	Training reference	All Δflip	Foreground Δflip	Interior $\geq$128 bp Δflip
HyenaDNA	Y	+0.0829 $\pm$ 0.0139	+0.1187 $\pm$ 0.0282	+0.0529 $\pm$ 0.0174
HyenaDNA	H	+0.0777 $\pm$ 0.0241	+0.1057 $\pm$ 0.0442	+0.0526 $\pm$ 0.0261
Caduceus-PH	Y	+0.0910 $\pm$ 0.0321	+0.0209 $\pm$ 0.0455	+0.0823 $\pm$ 0.0315
Caduceus-PH	H	+0.0905 $\pm$ 0.0335	-0.0045 $\pm$ 0.0765	+0.0823 $\pm$ 0.0317
DNABERT-2	Y	+0.0871 $\pm$ 0.1138	+0.0959 $\pm$ 0.1794	+0.0889 $\pm$ 0.1101
DNABERT-2	H	+0.0788 $\pm$ 0.0494	+0.0966 $\pm$ 0.0551	+0.0817 $\pm$ 0.0520
GROVER	Y	+0.1272 $\pm$ 0.0413	+0.0541 $\pm$ 0.0675	+0.1184 $\pm$ 0.0382
GROVER	H	+0.1085 $\pm$ 0.0115	+0.3080 $\pm$ 0.0669	+0.1063 $\pm$ 0.0149
NTv3-100M-pre	Y	-0.0230 $\pm$ 0.0469	+0.1071 $\pm$ 0.1009	-0.0253 $\pm$ 0.0452
NTv3-100M-pre	H	+0.0068 $\pm$ 0.0347	+0.0877 $\pm$ 0.1563	+0.0025 $\pm$ 0.0322
one-hot	Y	+0.0891 $\pm$ 0.0096	-0.0473 $\pm$ 0.0282	+0.0727 $\pm$ 0.0092
one-hot	H	+0.0912 $\pm$ 0.0143	-0.0629 $\pm$ 0.0162	+0.0751 $\pm$ 0.0142

The full geometry table crosses whole-input sequence category, absolute mapped index shift, distance to both output boundaries, and labels. Sequence categories are identical, substitution-only,

indel, and different-unclassified. Shift bins are 0, 1, 2–4, and ≥5 bp; boundary bins are 0–7, 8–31, 32–127, and ≥128 bp. T2 background, donor, and acceptor labels remain separate. Binary tasks have no position-shift or boundary strata. Zero-observation groups remain present with undefined rates.

B.5.3 IDENTICAL-INPUT SCOPE

For T2, equal-index probabilities and classes match exactly in all 774 identical-input windows across all 72 paired settings. Each setting holds the fitted predictor fixed across the two inputs. The output maps are derived from selected assembly-level alignment blocks and then converted to indices within the paired windows (Appendix B.2). They retain this recorded correspondence when the cropped input strings are identical.

In the audited background example in Figure S3c, the map links YAO index 2000 to GRCh38 index 1999 in two identical 2,004-bp inputs. Equal-index comparison checks the same entry in the two identical prediction arrays; mapped comparison checks adjacent entries, whose predicted classes can differ. This accounts for the disagreement computationally. Reconstruction in Appendix B.2 checks consistency with the recorded alignment blocks; biological homology at these exceptional positions has not been independently validated.

The audit contains 111,665,016 pooled zero-displacement position observations with zero flips and 3,240 one-base-displaced observations with 110 flips. All 110 identical-input flips therefore occur at displaced indices. They account for 0.00697% of the 1,577,548 flips across the complete T2 panel; the denominator is the complete panel's flip count.

Counts below pool both readouts, training references and all three seeds for all six models, including the one-hot control; they repeat window-position observations across settings. Identical-input binary windows have no class flips, although tiny numerical probability differences can occur.

Task	Identical/all windows	Identical-input flips	All flips	Share of all flips
T2	774/2544	110	1577548	0.00697%
T3	5769/7930	0	6430	0.00000%
T4	3318/5780	0	4150	0.00000%
T5	4668/8000	0	6481	0.00000%

B.5.4 DECISION MARGINS

For T2, the margin is the smaller top1−top2 probability gap across the paired outputs. For binary tasks, it is the smaller of twice each probability's absolute distance from 0.5. The fixed bins are [0,.01), [.01,.05), [.05,.10), [.10,.20), and [.20,1]. Margin membership changes with predictions and is excluded from fixed-subset readout contrasts. The table pools both readouts, both heads and three seeds within each model; all five bins remain in the source CSV. For T2, low-margin observations are enriched among flips, but no model has most flips in the first bin.

Model	Units in margin<.01 (%)	All flips in this bin (%)	Foreground flips in this bin (%)
HyenaDNA	0.325	21.684	17.130
Caduceus-PH	0.346	26.953	19.224
DNABERT-2	0.376	16.467	16.603
GROVER	0.246	12.847	12.422
NTv3-100M-pre	0.241	18.631	12.735
one-hot	1.101	5.725	6.000

The complete outputs retain both-correct, both-wrong, only-Y-correct, only-H-correct, flipping, and different-wrong-class flipping, with position-equal and window-equal rates. Sources: diagnostic protocol, complete tables, and independent correspondence audit.

C ENCODERS, PREDICTION HEADS AND TRAINING PROTOCOL

C.1 REPRESENTATION EXTRACTION AND REPRODUCIBILITY BOUNDARY

The executed readout consumes base-projected cached embeddings. Valid token offsets are clipped to the input window; each token vector is assigned to every base in its half-open interval, and overlapping contributions are averaged equally. Uncovered positions receive zero features and an invalid mask. This projection precedes the convolutional layers. Before cached representations are reused, the loader checks oriented sequence hashes and available configuration fields, along with IDs, labels and coordinates. The common-protocol T2–T5 run binds 120 manifests (24 per model) to its preflight and retains 160 feature-loading audit records. Independent sampled projection checks report zero missing projected positions and zero maximum error. The T1 extension reuses recorded cached representations; its protocol and verification are documented in Appendices C.4 and I.4.

Table C1. Recorded feature-extraction contract for the common-protocol T2–T5 panel. Each model has 24 bound cache manifests, uses final-layer features, and stores float32 embeddings with token offsets. Cache manifests record no truncation for the 300-, 510-, or 2,004-bp task inputs. "Not recorded" marks a missing historical revision field.

Model	Tokenization and offsets	Padding and biological-token handling	Historical revision
DNABERT-2	Non-overlapping BPE; Hugging Face native offsets	No recorded multiple; only valid offsets inside the input window are projected	Not recorded
GROVER	Non-overlapping BPE; Hugging Face native offsets	No recorded multiple; only valid offsets inside the input window are projected	Not recorded
HyenaDNA	One token per base	No recorded multiple; biological token mask is retained	Not recorded
Caduceus-PH	One token per base	No recorded multiple; biological token mask is retained	Not recorded
NTv3-100M-pre	One token per base	Batch padding to a multiple of 128; padding is removed by the token mask before projection	Not recorded

For every model, special or padding tokens without a valid biological offset do not contribute to the projected base features. Thus, NTv3's 2,004-bp T2 sequences are internally padded for encoder batching but retain exactly 2,004 biological base positions for the readout. The cached offset/mask contract specifies the features consumed by the fitted heads.

The readout implementation matches the frozen preflight hash. Archived worker and dispatch copies recover the preflight source hashes, while the later test-only replay manifest binds the corrected worker to the unchanged preflight and selected checkpoints. The manifests distinguish the archived extraction source from the corrected replay worker. The manifest metadata records last-layer extraction, offset strategy, context and precision. HyenaDNA additionally records the actual weight SHA-256 in its cache metadata. Most historical model/tokenizer revision fields and the actual Hyena source-commit/clean-worktree fields remain null. Appendix I.5 summarizes the remaining execution records. The input–extraction binding table separates verified cache consumption, recorded extraction settings, and unresolved encoder execution. Appendix I.3 adds content-pinned replays for HyenaDNA T2 and GROVER T5, testing numerical compatibility on the recorded samples.

The annotation file identified by the author is RefSeq `GCF_000001405.40_GRCh38.p14_genomic.gff`; ENCODE4 identifies the cCRE source family. The YAO assembly is T2T-YAO v2.0 maternal (mat); Appendix A.3 records the reference-file identities and retrieval-verification scope. Exact upstream annotation checksums, complete transfer records and production bindings for YAO strata remain incomplete. The benchmark therefore evaluates agreement with the frozen transferred labels and reports the provenance limits explicitly.

Construction sources: archived freeze documentation, freeze implementation, embedding entry, input and protocol lineage, and reconstruction entry. These links retain internal construction provenance. The historical review-v30 material supplies a portable subset for consuming fixed inputs

and checking central summaries; it does not contain the full historical construction environment or define the current delivery entry.

C.2 Model instance identity and validation coverage

All 34 files in the earlier five-model inventory were rehashed and match that inventory, which records agreement with the earlier preflight. The table separates current file identity from actual encoder replay. Full weight hashes, configuration/tokenizer files and local custom-code identities are retained in the linked inventory. Source snapshots and model-side configuration files identify the replay components. Model weights are not redistributed in the fixed-input package.

Encoder	Weight SHA-256 prefix	Inventoried files	Sequence-to-encoder replay coverage
HyenaDNA	`5ecc8a121feb9906`	2	sampled T2; 48 pairs per split; 24 selected-head test settings
Caduceus-PH	`1ab95562b7ee254d`	8	not replayed
DNABERT-2	`7ff39ec77a484dd0`	9	not replayed
GROVER	`81fa4ee056e65ff1`	6	sampled T5; 48 pairs per split; 24 selected-head test settings
NTv3	`8504dd659224628d`	9	not replayed

For HyenaDNA and GROVER, sampled replay covers 48 pairs per training/validation/test split and all 24 selected-head test prediction settings on the sampled inputs. Their test probabilities match the archived values exactly; the feature-tolerance results are reported in Appendix I.3. The other three encoders have file identities and cached-feature/readout audit records, but no new sequence-to-encoder replay. The historical execution fields are summarized in Appendix I.5. New predictors can use the fixed sequences and maps through the common scorer. In particular, NTv3 in Appendix H.3.2 refers to the file instance with full weight SHA-256 `8504dd659224628df906d6a876a8975e284fc9f75d194f2ca2bd9e98d4000ed2`.

C.3 Recorded training settings

The primary readout comparisons use the common-protocol results in Appendix E.2 for T2–T5 and Appendix E.1 for T1. The T1 extension includes 60 frozen-representation heads and 12 one-hot heads using seeds 14/15/16. Earlier task-specific results retain their own protocols in Appendix J.1.

The T2–T5 common-protocol readout experiment uses non-affine RMSNorm (epsilon 1e-6), two 64-channel width-3 convolutions, GELU and dropout 0.1. Training uses float32 without AMP or TF32, AdamW with betas (0.9, 0.999), epsilon 1e-8, learning rate 0.001, weight decay 0.0001, gradient-norm clipping at 1, and no scheduler or warmup. Batch size is 32 with the final incomplete batch retained. Training runs for at most 20 epochs, with early stopping after five epochs without improvement. We select the earliest epoch with the best validation score, using MCC for T2 and AP for binary tasks. Validation inputs use the head-training reference: YAO for YAO-trained heads and GRCh38 for GRCh38-trained heads. The binary decision threshold is fixed at 0.5. Binary pos_weight is the training negative/positive ratio; dense class weights are N/(3 n_c) using valid training positions. All 240 selections are frozen before testing.

For representation widths 256/768, binary pool1 heads contain 61,633/159,937 parameters, and pool16 heads contain 62,593/160,897. T2 heads contain 61,763/160,067 parameters under either dilation. Pool16 uses adaptive-pooling floor/ceil boundaries, so adjacent bins may overlap by one position. The binary pooling contrast adds 960 parameters; three-class T1 pooling adds 2,880 under both the historical and common protocols. Parameter counts depend on representation width.

The summed training and validation process times for successfully trained heads are 4.11 hours for T3, 2.60 for T4, 6.78 for T5 and 9.13 for T2. The timers cover successful training and validation; feature loading, testing, failed attempts and postprocessing are outside their scope. Completed heads ran for 7–20 epochs for T3/T4/T2 and 6–20 for T5.

Recorded compute and scope. The table consolidates existing training/validation timers; each row specifies its model and training scope.

Experiment	Summed recorded time (hours)	Record
T1 common-protocol extension, frozen and one-hot heads	7.04	C.4
T2 common-protocol frozen heads	9.13	C.3
T3 common-protocol frozen heads	4.11	C.3
T4 common-protocol frozen heads	2.60	C.3
T5 common-protocol frozen heads	6.78	C.3
T2 residual-CNN comparator	6.785	H.3.1

For the residual-CNN comparator, Appendix H.3.1 records six A800 GPUs and approximately 2.71 GiB peak allocated CUDA tensor memory per worker. Appendix I.3 documents the runnable CPU baseline. Appendix I.5 lists the remaining compute records.

C.4 T1 COMMON-PROTOCOL TRAINING

T1 was added to the common-protocol readout panel after the T2–T5 experiments. This extension uses the repaired v2 cohort: 49,997 training, 8,000 validation and 8,000 test window pairs, each 402 bp long, with background, donor and acceptor labels. Training class counts are 24,997, 12,500 and 12,500. The task-specific five-task panel and its regional predictions remain unchanged. The earlier T1 results in Appendix J.1 are retained as historical results under their own protocol.

The extension uses seeds 14, 15 and 16, both training references, pool1/pool16, five frozen representations and one-hot input. The common protocol in Section 3.3 and Appendix C.3 applies: batch size 32 with the final batch retained, at most 20 epochs, patience 5, float32, AdamW, and shared convolutional initialization and batch order within each paired readout comparison. Weighted cross-entropy uses N/(3 n_c); selection uses the earliest best validation MCC on inputs from the training reference. Ordered 16-bin pooling adds 2,880 classifier parameters. All 72 checkpoint selections were completed before test inference, yielding 144 test prediction sets. The heads train on the recorded encoder caches.

Completed runs lasted 8–20 epochs; 37 reached the 20-epoch cap and 13 selected their final epoch. Summed recorded training/validation process time was 7.04 hours. This sum is neither end-to-end wall time nor a GPU-utilization integral.

D METRICS, WEIGHTING AND STATISTICAL INFERENCE

D.1 PRIMARY METRICS AND BLOCK-BOOTSTRAP COMPARISONS

The primary readout contrast is the flip rate with the changed readout minus that with the baseline readout, averaged over three fixed seeds. The following resampling specification describes the T2–T5 common-protocol contrasts; Appendix E.3 documents the T1 extension. Binary tasks use windows; T2 first calculates the flip rate over valid mapped positions within each window, then averages these rates with equal weight per window. Position-weighted and positive-subset summaries are sensitivity analyses. Bootstrap replicates use shared within-chromosome block multiplicities, 5,000 replicates, and resampling seed 20260909. The approximate centered-bootstrap p-value is (1 + count(|delta_b - delta| >= |delta|))/(5001). The 30 frozen-model binary contrasts (T3–T5) and the 10 frozen-model dense contrasts (T2) form separate Benjamini–Hochberg families at each block width (Benjamini & Hochberg, 1995). One-hot controls use separate families (Appendix H.1.4). T1 adds a family of 10 frozen-model contrasts and a separate family of two one-hot contrasts (Appendix E.3). Performance means have no corresponding significance test.

The new binary experiment and one-hot controls use `sklearn.metrics.average_precision_score`: AP = Σ_k (R_k − R_{k−1})P_k, where P_k and R_k are precision and recall at the corresponding score thresholds. AP denotes this non-interpolated average precision. Historical source files label this quantity AUPRC. Their field names are retained in the evidence records, while common-protocol tables and figures use AP. The earlier task-specific table in Appendix J.1 retains its source label with a separate provenance note. T2 uses `matthews_corrcoef` on argmax classes, including background, donor, and acceptor. The AP and MCC API definitions specify the metrics. Binary flipping and the correctness states in Appendix F.6 use threshold 0.5. Appendix G.4

applies the preset common-threshold grid as a sensitivity analysis; primary binary decisions use 0.5 throughout.

For new readout contrasts, block identity is (source YAO chromosome, floor(YAO window start / block width)). At each width, only occupied blocks are included; within each chromosome, the observed number of blocks is sampled with replacement using uniform block probabilities. All observations from a window inherit its start block, and each replicate shares those block multiplicities across paired settings and seeds. A window crossing a block boundary inherits its start block in full. Reported intervals are the 2.5th and 97.5th percentiles of the bootstrap contrasts. They are pointwise intervals; multiplicity correction applies separately to the approximate p-values. The later correspondence diagnostic and HyenaDNA MCC sensitivity analysis use window-midpoint blocks, as specified in Appendices B.3 and G.5; their pointwise intervals are separate from these corrected readout tests.

D.2 Correctness identities and transitions

Joint incorrectness and correctness discordance determine mean two-reference accuracy. Using the same observations and weights, let A_Y and A_H denote accuracy on the two inputs, S the fraction incorrect on both, and D the fraction correct on exactly one. The mean accuracy is (A_Y + A_H)/2 = 1 − S − D/2. For binary predictions against a shared label, the flip rate F equals D, so $\bar{A} = 1 - S - F/2$. In multiclass tasks, F also counts disagreements between two incorrect classes; the accuracy identity therefore uses D. Here, S includes every pair for which both predictions are incorrect, whether or not the predictions agree on a class. In binary tasks with a shared label, two incorrect predictions necessarily agree on the incorrect class.

This decomposition uses accuracy on a common scoring domain; AP and MCC follow their definitions in Appendix D.1. Observation-level transition records identify which flips disappear or appear after a readout change and whether the resulting predictions are correct. Appendix F reports these transitions alongside the aggregate states.

D.3 Evaluation domains and aggregation

The reported quantities use the following observation domains and weights. Reproducing a joint plot requires reproducing both axes under their own definitions.

Quantity	Observation domain	Aggregation
T1 MCC; T3–T5 AP	All accepted test windows on each input reference	One window contributes one label and prediction; summarize the three fitted seeds
Primary T2 MCC	Full-window valid positions on each input reference	Pool the confusion matrix over positions within each run, then summarize seeds
Primary T2 cross-reference disagreement	Valid mapped positions in each paired window	Calculate a within-window fraction, then average windows equally and average the three seeds
T2 classwise performance and paired correctness	Shared mapped domain, with class-specific denominators	Give each retained position observation equal weight; overlapping windows can repeat loci
Binary eliminated-flip error fraction	Window–seed observations whose disagreement disappears	Shared errors among eliminated flips divided by all eliminated flips
Binary net shared-error change	All test window–seed observations	Difference in both-wrong counts between readouts

The primary corrected flip tests use source-window start blocks. The post-hoc correspondence diagnostic and conditional HyenaDNA MCC intervals use source-window midpoint blocks. Their intervals and correction families remain separate; the detailed procedures appear in D.1, B.3 and G.5. The historical regional comparisons in Appendix J.4 use a separate 200-contrast family under the earlier protocol.

E COMPLETE PRIMARY PERFORMANCE AND CONSISTENCY RESULTS

E.1 T1 TASK PERFORMANCE

MCC values are mean ± sample SD over the three seeds. Reference codes follow Section 3.3. Comparisons of means are descriptive.

Model	Readout	YY	YH	HY	HH
HyenaDNA	pool1	0.7376 ± 0.0303	0.7371 ± 0.0315	0.7489 ± 0.0338	0.7493 ± 0.0336
HyenaDNA	pool16	0.9124 ± 0.0022	0.9120 ± 0.0021	0.9108 ± 0.0061	0.9112 ± 0.0054
Caduceus-PH	pool1	0.5004 ± 0.0058	0.4997 ± 0.0055	0.4975 ± 0.0077	0.4967 ± 0.0080
Caduceus-PH	pool16	0.9171 ± 0.0060	0.9162 ± 0.0065	0.9202 ± 0.0021	0.9206 ± 0.0023
DNABERT-2	pool1	0.4090 ± 0.0252	0.4092 ± 0.0289	0.4123 ± 0.0212	0.4108 ± 0.0199
DNABERT-2	pool16	0.8750 ± 0.0037	0.8728 ± 0.0039	0.8720 ± 0.0075	0.8710 ± 0.0072
GROVER	pool1	0.7461 ± 0.0024	0.7458 ± 0.0020	0.7493 ± 0.0036	0.7478 ± 0.0044
GROVER	pool16	0.8901 ± 0.0033	0.8889 ± 0.0026	0.8887 ± 0.0052	0.8878 ± 0.0062
NTv3-100M-pre	pool1	0.6372 ± 0.0029	0.6368 ± 0.0027	0.6381 ± 0.0025	0.6374 ± 0.0039
NTv3-100M-pre	pool16	0.9279 ± 0.0012	0.9279 ± 0.0010	0.9253 ± 0.0042	0.9254 ± 0.0034
One-hot	pool1	0.2946 ± 0.0038	0.2926 ± 0.0048	0.2936 ± 0.0054	0.2922 ± 0.0064
One-hot	pool16	0.8975 ± 0.0024	0.8978 ± 0.0027	0.8997 ± 0.0003	0.8995 ± 0.0011

Pool16 increases mean MCC in all 20 frozen-model–reference-combination settings, by 0.1394–0.4659. Frozen representations exceed the matched one-hot mean in 32 of 40 task–readout–reference comparisons. One-hot itself increases mean MCC by 0.6029–0.6073. The pooling change jointly alters spatial aggregation and classifier capacity.

E.2 T2–T5 COMMON-PROTOCOL READOUT EXPERIMENT

Values are means ± sample standard deviations over training seeds. The first reference character denotes head training, and the second denotes test input. Paired bootstrap intervals are reported separately in E.3–E.4.

All rows use seeds 14/15/16. T2 uses original full-window MCC; other tasks use AP. The 40 rows account for the original 240 common-protocol frozen heads and 480 prediction sets; Appendix C.4 documents the 60 additional frozen-representation T1 heads and their 120 predictions. Contrasts are computed within this experiment.

Task	Model	Readout	YY	YH	HY	HH
T3	HyenaDNA	pool1	0.7760 ± 0.0067	0.7748 ± 0.0069	0.7753 ± 0.0021	0.7744 ± 0.0017
T3	HyenaDNA	pool16	0.8283 ± 0.0069	0.8280 ± 0.0066	0.8284 ± 0.0039	0.8282 ± 0.0040
T3	Caduceus-PH	pool1	0.7916 ± 0.0012	0.7914 ± 0.0008	0.7915 ± 0.0016	0.7912 ± 0.0015
T3	Caduceus-PH	pool16	0.8586 ± 0.0019	0.8582 ± 0.0020	0.8576 ± 0.0024	0.8576 ± 0.0023
T3	DNABERT-2	pool1	0.7759 ± 0.0025	0.7755 ± 0.0027	0.7751 ± 0.0018	0.7743 ± 0.0014
T3	DNABERT-2	pool16	0.8127 ± 0.0018	0.8118 ± 0.0026	0.8112 ± 0.0007	0.8116 ± 0.0009
T3	GROVER	pool1	0.7968 ± 0.0032	0.7964 ± 0.0032	0.7939 ± 0.0035	0.7935 ± 0.0040
T3	GROVER	pool16	0.8218 ± 0.0049	0.8212 ± 0.0050	0.8213 ± 0.0031	0.8213 ± 0.0033
T3	NTv3-100M-pre	pool1	0.8049 ± 0.0041	0.8045 ± 0.0044	0.8017 ± 0.0039	0.8016 ± 0.0038
T3	NTv3-100M-pre	pool16	0.8594 ± 0.0019	0.8587 ± 0.0017	0.8642 ± 0.0051	0.8634 ± 0.0051
T4	HyenaDNA	pool1	0.8844 ± 0.0011	0.8837 ± 0.0013	0.8833 ± 0.0011	0.8828 ± 0.0010

Task	Model	Readout	YY	YH	HY	HH
T4	HyenaDNA	pool16	0.8852 ± 0.0031	0.8844 ± 0.0035	0.8857 ± 0.0029	0.8849 ± 0.0030
T4	Caduceus-PH	pool1	0.8852 ± 0.0019	0.8853 ± 0.0019	0.8856 ± 0.0005	0.8858 ± 0.0006
T4	Caduceus-PH	pool16	0.9018 ± 0.0019	0.9017 ± 0.0018	0.9015 ± 0.0034	0.9013 ± 0.0036
T4	DNABERT-2	pool1	0.8933 ± 0.0017	0.8936 ± 0.0014	0.8925 ± 0.0019	0.8925 ± 0.0018
T4	DNABERT-2	pool16	0.9083 ± 0.0004	0.9082 ± 0.0003	0.9065 ± 0.0015	0.9065 ± 0.0015
T4	GROVER	pool1	0.8881 ± 0.0007	0.8875 ± 0.0010	0.8875 ± 0.0016	0.8869 ± 0.0014
T4	GROVER	pool16	0.8874 ± 0.0034	0.8872 ± 0.0035	0.8881 ± 0.0030	0.8880 ± 0.0032
T4	NTv3-100M-pre	pool1	0.8771 ± 0.0039	0.8767 ± 0.0045	0.8746 ± 0.0032	0.8738 ± 0.0031
T4	NTv3-100M-pre	pool16	0.8891 ± 0.0038	0.8889 ± 0.0031	0.8858 ± 0.0032	0.8861 ± 0.0027
T5	HyenaDNA	pool1	0.7801 ± 0.0018	0.7804 ± 0.0016	0.7765 ± 0.0027	0.7775 ± 0.0032
T5	HyenaDNA	pool16	0.7862 ± 0.0028	0.7867 ± 0.0026	0.7866 ± 0.0057	0.7873 ± 0.0058
T5	Caduceus-PH	pool1	0.7841 ± 0.0030	0.7842 ± 0.0030	0.7824 ± 0.0027	0.7829 ± 0.0027
T5	Caduceus-PH	pool16	0.8151 ± 0.0016	0.8149 ± 0.0013	0.8152 ± 0.0033	0.8153 ± 0.0030
T5	DNABERT-2	pool1	0.7813 ± 0.0019	0.7809 ± 0.0018	0.7812 ± 0.0022	0.7804 ± 0.0018
T5	DNABERT-2	pool16	0.8104 ± 0.0028	0.8092 ± 0.0032	0.8095 ± 0.0013	0.8089 ± 0.0011
T5	GROVER	pool1	0.7882 ± 0.0030	0.7884 ± 0.0027	0.7872 ± 0.0026	0.7871 ± 0.0029
T5	GROVER	pool16	0.7912 ± 0.0032	0.7916 ± 0.0036	0.7917 ± 0.0057	0.7920 ± 0.0060
T5	NTv3-100M-pre	pool1	0.7432 ± 0.0040	0.7412 ± 0.0052	0.7443 ± 0.0034	0.7429 ± 0.0028
T5	NTv3-100M-pre	pool16	0.7842 ± 0.0027	0.7831 ± 0.0029	0.7863 ± 0.0037	0.7855 ± 0.0030
T2	HyenaDNA	dilation1	0.1321 ± 0.0023	0.1321 ± 0.0023	0.1325 ± 0.0017	0.1326 ± 0.0017
T2	HyenaDNA	dilation4	0.1434 ± 0.0033	0.1434 ± 0.0033	0.1461 ± 0.0066	0.1461 ± 0.0066
T2	Caduceus-PH	dilation1	0.1616 ± 0.0068	0.1617 ± 0.0069	0.1623 ± 0.0061	0.1623 ± 0.0062
T2	Caduceus-PH	dilation4	0.1491 ± 0.0112	0.1491 ± 0.0112	0.1485 ± 0.0104	0.1485 ± 0.0104
T2	DNABERT-2	dilation1	0.1352 ± 0.0038	0.1352 ± 0.0039	0.1362 ± 0.0030	0.1364 ± 0.0029
T2	DNABERT-2	dilation4	0.1318 ± 0.0028	0.1318 ± 0.0028	0.1316 ± 0.0034	0.1319 ± 0.0035
T2	GROVER	dilation1	0.1692 ± 0.0051	0.1693 ± 0.0051	0.1566 ± 0.0044	0.1566 ± 0.0044
T2	GROVER	dilation4	0.1465 ± 0.0023	0.1465 ± 0.0021	0.1508 ± 0.0047	0.1509 ± 0.0047
T2	NTv3-100M-pre	dilation1	0.1793 ± 0.0087	0.1795 ± 0.0088	0.1778 ± 0.0118	0.1780 ± 0.0118
T2	NTv3-100M-pre	dilation4	0.1952 ± 0.0078	0.1953 ± 0.0079	0.1900 ± 0.0069	0.1901 ± 0.0069

E.3 T1 DISAGREEMENT CONTRASTS

Contrasts are pool16 minus pool1. All observations are window pairs; each setting uses 8,000 pairs per seed. At each width, 5,000 within-chromosome block replicates use source YAO window starts. The 10 frozen-model–training-reference contrasts form a new T1 family; the two one-hot contrasts form a separate family. Both families were specified before training. Original T2–T5 correction families are unchanged. Intervals are pointwise 95% percentile intervals conditional on the fitted seeds. Corrected decreases require the same direction and $q < 0.05$ at both 1 Mb and 2 Mb.

Model	Training reference	pool1 flip rate (%)	pool16 flip rate (%)	Difference (pp)	1 Mb interval (pp)	2 Mb interval (pp)	q at both widths
HyenaDNA	Y	0.7583	0.2667	-0.4917	[-0.6293, -0.3571]	[-0.6265, -0.3603]	0.0002
HyenaDNA	H	0.7625	0.2833	-0.4792	[-0.6355, -0.3273]	[-0.6411, -0.3245]	0.0002
Caduceus-PH	Y	0.9917	0.2750	-0.7167	[-0.8747, -0.5592]	[-0.8877, -0.5564]	0.0002
Caduceus-PH	H	0.8833	0.2792	-0.6042	[-0.7495, -0.4612]	[-0.7529, -0.4564]	0.0002
DNABERT-2	Y	2.0750	0.6000	-1.4750	[-1.7105, -1.2461]	[-1.7028, -1.2520]	0.0002
DNABERT-2	H	2.2333	0.6292	-1.6042	[-1.8255, -1.3860]	[-1.8234, -1.3840]	0.0002
GROVER	Y	1.3250	0.5042	-0.8208	[-0.9842, -0.6627]	[-0.9887, -0.6573]	0.0002
GROVER	H	1.2458	0.4583	-0.7875	[-0.9638, -0.6193]	[-0.9728, -0.6036]	0.0002
NTv3-100M-pre	Y	1.3917	0.2833	-1.1083	[-1.3419, -0.8889]	[-1.3571, -0.8784]	0.0002
NTv3-100M-pre	H	1.4000	0.3500	-1.0500	[-1.2821, -0.8267]	[-1.3097, -0.8066]	0.0002
One-hot	Y	1.4000	0.4208	-0.9792	[-1.2027, -0.7675]	[-1.2018, -0.7537]	0.0002
One-hot	H	1.3875	0.3417	-1.0458	[-1.2403, -0.8552]	[-1.2516, -0.8398]	0.0002

All 10 frozen contrasts and both one-hot contrasts decrease at both widths. The exact q-value is 1/5,001 for each contrast, the finite-resampling floor, rather than zero.

E.4 T2–T5 PERFORMANCE AND DISAGREEMENT

Figures S1–S2 cover the common-protocol T2–T5 comparisons in Appendix E.2. Appendix E.1 provides T1 task performance, and Appendix E.3 provides its disagreement contrasts and corrected results. Together they support the five-task summary in Table 2 and Figure 3 of the main text.

Tables E1–E4 report all 40 frozen-encoder contrasts underlying Figure S2, with intervals and BH-adjusted q values at both block widths. Each contrast is the mean over seeds 14, 15 and 16. T2 uses dilation4 minus dilation1 on 2,544 equally weighted windows; T3–T5 use pool16 minus pool1 on 7,930, 5,780 and 8,000 windows, respectively. Negative values indicate reduced disagreement. Means and interval endpoints are in percentage points (pp). The 95% pointwise intervals use 5,000 shared chromosome-stratified block-bootstrap replicates per width, conditional on the fitted seeds (Appendix D.1). BH correction is applied separately at each width to the ten T2 contrasts and, jointly, to the thirty T3–T5 contrasts.

A contrast passes the primary rule when its direction agrees and $q < 0.05$ at both widths. This yields one decrease and eight increases for T2, no passing contrasts for T3 or T4, and two decreases for T5 (GROVER under Y and H training). These counts reproduce Table 2. Unrounded values, per-seed inputs and file hashes are retained with the table data, original bootstrap summary, per-seed rates and source binding.

Table E1. T2 primary readout contrasts in cross-reference disagreement. Y/H denotes the training reference.

Model	Training	Δflip (pp)	1 Mb 95% interval (pp)	q, 1 Mb	2 Mb 95% interval (pp)	q, 2 Mb
HyenaDNA	Y	+0.0829	[+0.0647, +0.1006]	0.000222	[+0.0627, +0.1023]	0.000222
HyenaDNA	H	+0.0777	[+0.0593, +0.0941]	0.000222	[+0.0577, +0.0961]	0.000222
Caduceus-PH	Y	+0.0910	[+0.0749, +0.1108]	0.000222	[+0.0746, +0.1103]	0.000222
Caduceus-PH	H	+0.0905	[+0.0738, +0.1124]	0.000222	[+0.0731, +0.1126]	0.000222
DNABERT-2	Y	+0.0871	[+0.0668, +0.1142]	0.000222	[+0.0672, +0.1156]	0.000222
DNABERT-2	H	+0.0788	[+0.0603, +0.1022]	0.000222	[+0.0598, +0.1041]	0.000222
GROVER	Y	+0.1272	[+0.1038, +0.1566]	0.000222	[+0.1034, +0.1569]	0.000222
GROVER	H	+0.1085	[+0.0875, +0.1347]	0.000222	[+0.0869, +0.1344]	0.000222
NTv3-100M-pre	Y	-0.0230	[-0.0289, -0.0176]	0.000222	[-0.0294, -0.0174]	0.000222
NTv3-100M-pre	H	+0.0068	[-0.0011, +0.0137]	0.072585	[-0.0016, +0.0140]	0.083383

Table E2. T3 primary readout contrasts in cross-reference disagreement. Y/H denotes the training reference.

Model	Training	Δflip (pp)	1 Mb 95% interval (pp)	q, 1 Mb	2 Mb 95% interval (pp)	q, 2 Mb
HyenaDNA	Y	-0.0084	[-0.1908, +0.1657]	0.929814	[-0.1881, +0.1722]	0.928214
HyenaDNA	H	+0.0799	[-0.0954, +0.2509]	0.558520	[-0.0932, +0.2483]	0.554415
Caduceus-PH	Y	-0.1597	[-0.3303, +0.0083]	0.284515	[-0.3448, +0.0227]	0.345264
Caduceus-PH	H	-0.1219	[-0.2891, +0.0476]	0.370926	[-0.2876, +0.0460]	0.403556
DNABERT-2	Y	+0.1345	[-0.1055, +0.3776]	0.493666	[-0.0987, +0.3805]	0.478845
DNABERT-2	H	+0.0546	[-0.1706, +0.2762]	0.751290	[-0.1543, +0.2551]	0.731854
GROVER	Y	-0.1597	[-0.3560, +0.0314]	0.335933	[-0.3560, +0.0279]	0.345264
GROVER	H	+0.0631	[-0.1090, +0.2374]	0.655778	[-0.1073, +0.2459]	0.666412
NTv3-100M-pre	Y	-0.2396	[-0.4573, -0.0163]	0.205159	[-0.4615, -0.0168]	0.211158
NTv3-100M-pre	H	-0.1807	[-0.3998, +0.0262]	0.329934	[-0.4062, +0.0469]	0.369526

Table E3. T4 primary readout contrasts in cross-reference disagreement. Y/H denotes the training reference.

Model	Training	Δflip (pp)	1 Mb 95% interval (pp)	q, 1 Mb	2 Mb 95% interval (pp)	q, 2 Mb
HyenaDNA	Y	+0.1326	[-0.0596, +0.3200]	0.390351	[-0.0669, +0.3324]	0.411346
HyenaDNA	H	+0.0923	[-0.0684, +0.2610]	0.492651	[-0.0649, +0.2512]	0.474280
Caduceus-PH	Y	+0.0461	[-0.1210, +0.2231]	0.751290	[-0.1299, +0.2315]	0.731854
Caduceus-PH	H	+0.0923	[-0.1236, +0.3022]	0.592382	[-0.1306, +0.3134]	0.602451
DNABERT-2	Y	-0.0980	[-0.3412, +0.1371]	0.601880	[-0.3452, +0.1309]	0.602451
DNABERT-2	H	-0.1672	[-0.4343, +0.0832]	0.415117	[-0.4621, +0.0929]	0.474280
GROVER	Y	-0.0346	[-0.2243, +0.1636]	0.829219	[-0.2070, +0.1449]	0.805454
GROVER	H	-0.2191	[-0.4126, -0.0345]	0.205159	[-0.4180, -0.0135]	0.211158
NTv3-100M-pre	Y	-0.0173	[-0.2277, +0.1906]	0.928529	[-0.2390, +0.2051]	0.927952
NTv3-100M-pre	H	+0.1211	[-0.1214, +0.3542]	0.532893	[-0.1430, +0.3754]	0.554415

Table E4. T5 primary readout contrasts in cross-reference disagreement. Y/H denotes the training reference.

Model	Training	Δflip (pp)	1 Mb 95% interval (pp)	q, 1 Mb	2 Mb 95% interval (pp)	q, 2 Mb
HyenaDNA	Y	-0.0500	[-0.2345, +0.1415]	0.751290	[-0.2268, +0.1326]	0.731854
HyenaDNA	H	-0.1458	[-0.3421, +0.0429]	0.370926	[-0.3513, +0.0562]	0.411346
Caduceus-PH	Y	+0.1583	[-0.0207, +0.3446]	0.329934	[-0.0258, +0.3476]	0.345264
Caduceus-PH	H	+0.1750	[-0.0115, +0.3457]	0.277944	[-0.0083, +0.3582]	0.313937
DNABERT-2	Y	+0.0125	[-0.1707, +0.2008]	0.929814	[-0.1809, +0.1942]	0.927952
DNABERT-2	H	+0.1250	[-0.0536, +0.3011]	0.390351	[-0.0647, +0.3101]	0.411346
GROVER	Y	-0.3542	[-0.5298, -0.1857]	0.005999	[-0.5380, -0.1773]	0.002999
GROVER	H	-0.3833	[-0.5726, -0.1990]	0.005999	[-0.5804, -0.2001]	0.002999
NTv3-100M-pre	Y	+0.2083	[+0.0166, +0.3996]	0.205159	[+0.0214, +0.3889]	0.211158
NTv3-100M-pre	H	+0.0208	[-0.2078, +0.2415]	0.928529	[-0.1923, +0.2374]	0.927952

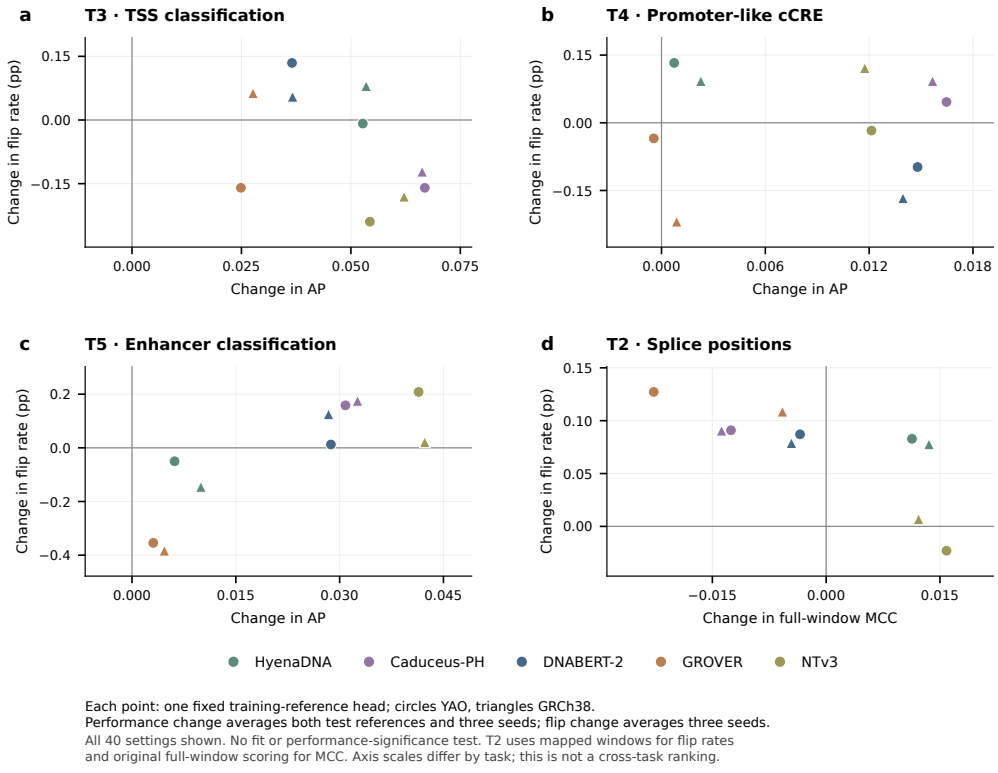

Figure S1: Descriptive performance–flipping comparison for T2–T5. Each point represents a task, model and fixed training reference; circles denote YAO training and triangles GRCh38. Horizontal differences average both test references and seeds 14/15/16; vertical differences average flipping over the same seeds. All 40 frozen-model–training-reference settings across T2–T5 are shown. Appendices E.1 and E.3 supply the ten T1 settings included in the main-text 50-setting panel. T2 MCC uses original full-window scoring while flipping uses window-equal mapped-position rates; Appendix G.1 checks the same-domain direction. Each task has its own axis scale. Figure S3 examines three diagnostic cases from these paired analyses.

E.5 Aggregate readout-effect direction audit

For each frozen representation, task and training reference, we compute alternative-minus-baseline task metrics on each test reference and average seeds 14, 15 and 16. Contrasts are pool16 minus pool1 (T1/T3–T5) or dilation4 minus dilation1 (T2); T2 MCC uses the primary full-window valid domain. Sources are the common-protocol task-performance results in Appendix E.1 for T1 and Appendix E.2 for T2–T5. The earlier task-specific panel is excluded.

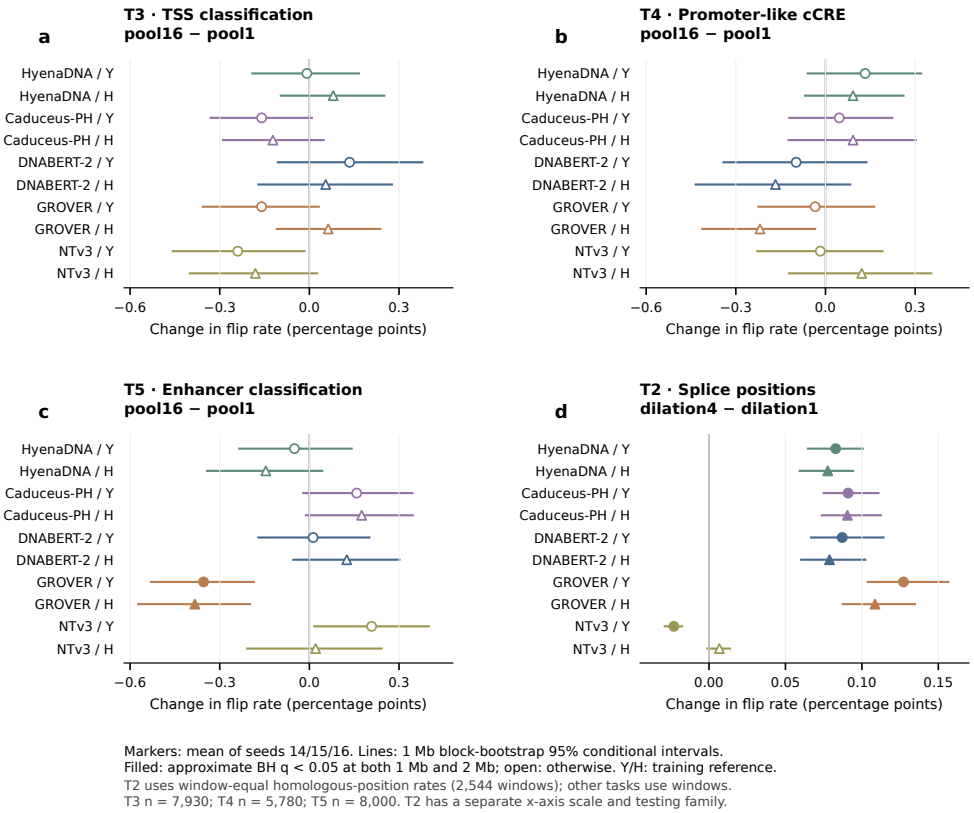

Figure S2: T2–T5 readout effects on cross-reference flipping. Points average seeds 14/15/16; intervals are pointwise 95% percentile intervals from 5,000 shared 1 Mb genomic-block replicates. Filled markers indicate $q < 0.05$ at both block widths within the stated task family. T2 uses a different horizontal scale and window-equal mapped-position flipping.

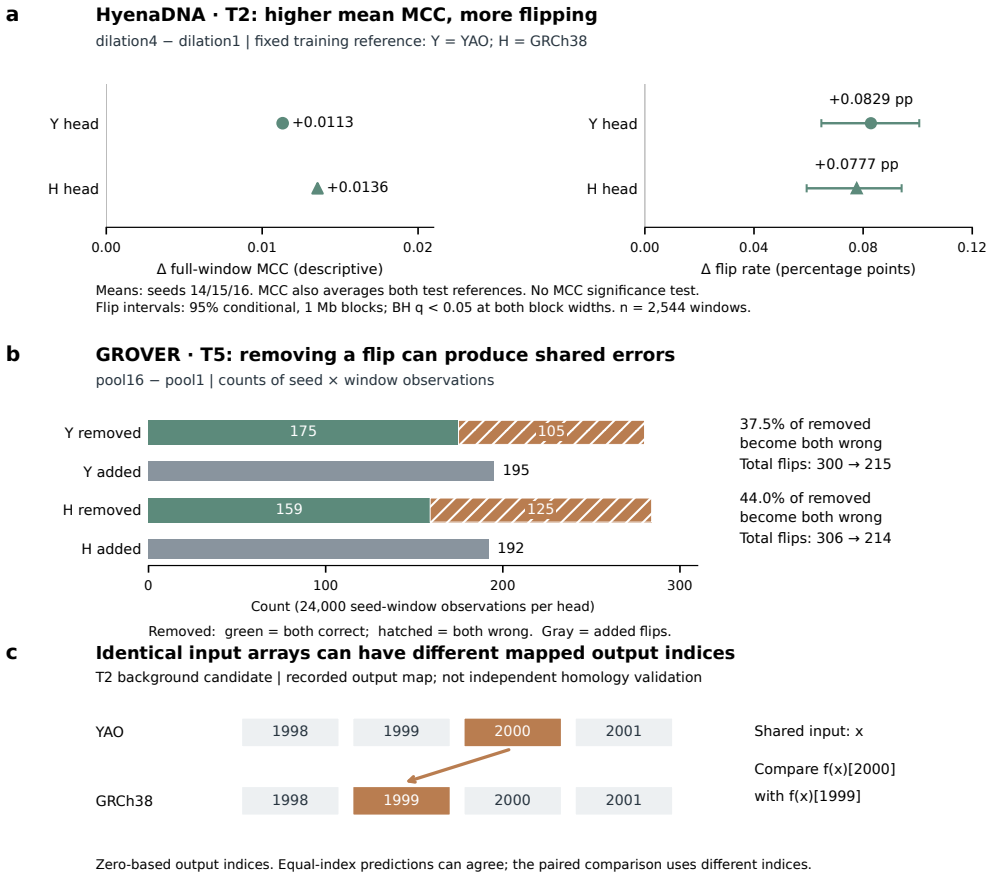

Figure S3: Paired diagnostics of readout changes. (a) HyenaDNA T2, dilation4−dilation1. MCC differences average both test references and three seeds; flipping uses 2,544 equally weighted paired windows. Bars are pointwise 95% conditional 1 Mb block intervals; both flipping contrasts pass correction at both widths. Y/H denotes training reference. (b) GROVER T5, pool16 versus pool1. Removed flips separate into both-correct and both-wrong agreement; added flips and totals are retained. For each readout and training reference, counts pool 8,000 test windows across three fitted seeds, yielding 24,000 seed-window observations. Binary decisions use threshold 0.5. Percentages condition on removed flips; Appendix F.6 also reports complete correctness states and overall accuracy. (c) One audited T2 background candidate has identical input arrays but mapped zero-based output indices 2000 and 1999. The schematic shows the recorded map indices, not independent validation of biological homology. Full settings and sources appear in Figure S2 and Appendices F.6/G.1/E.4.

All 50 reference-paired mean effects agree in sign, including ten of ten for each task. The audit compares signs of means across the fixed seeds. The retained numerical audit and source tables are listed in the claim-to-file index.

Task	Model	Training reference	ΔM(YAO)	ΔM(GRCh38)
T1	Caduceus-PH	Y	+0.416722	+0.416512
T1	Caduceus-PH	H	+0.422633	+0.423917
T1	DNABERT-2	Y	+0.465940	+0.463675
T1	DNABERT-2	H	+0.459755	+0.460247
T1	GROVER	Y	+0.144054	+0.143068
T1	GROVER	H	+0.139384	+0.140038
T1	HyenaDNA	Y	+0.174797	+0.174848
T1	HyenaDNA	H	+0.161857	+0.161901
T1	NTv3	Y	+0.290700	+0.291085
T1	NTv3	H	+0.287202	+0.287948
T2	Caduceus-PH	Y	-0.012528	-0.012567
T2	Caduceus-PH	H	-0.013773	-0.013801
T2	DNABERT-2	Y	-0.003423	-0.003466
T2	DNABERT-2	H	-0.004592	-0.004562
T2	GROVER	Y	-0.022698	-0.022759
T2	GROVER	H	-0.005760	-0.005772
T2	HyenaDNA	Y	+0.011316	+0.011310
T2	HyenaDNA	H	+0.013559	+0.013546
T2	NTv3	Y	+0.015902	+0.015830
T2	NTv3	H	+0.012229	+0.012139
T3	Caduceus-PH	Y	+0.066958	+0.066807
T3	Caduceus-PH	H	+0.066166	+0.066397
T3	DNABERT-2	Y	+0.036791	+0.036313
T3	DNABERT-2	H	+0.036097	+0.037256
T3	GROVER	Y	+0.024979	+0.024781
T3	GROVER	H	+0.027481	+0.027806
T3	HyenaDNA	Y	+0.052280	+0.053164
T3	HyenaDNA	H	+0.053164	+0.053808
T3	NTv3	Y	+0.054536	+0.054132
T3	NTv3	H	+0.062475	+0.061827
T4	Caduceus-PH	Y	+0.016580	+0.016364
T4	Caduceus-PH	H	+0.015839	+0.015493
T4	DNABERT-2	Y	+0.015017	+0.014581
T4	DNABERT-2	H	+0.013929	+0.013984
T4	GROVER	Y	-0.000642	-0.000254
T4	GROVER	H	+0.000596	+0.001156
T4	HyenaDNA	Y	+0.000752	+0.000713
T4	HyenaDNA	H	+0.002390	+0.002153
T4	NTv3	Y	+0.012092	+0.012178
T4	NTv3	H	+0.011204	+0.012293
T5	Caduceus-PH	Y	+0.030985	+0.030695
T5	Caduceus-PH	H	+0.032789	+0.032380
T5	DNABERT-2	Y	+0.029108	+0.028371
T5	DNABERT-2	H	+0.028354	+0.028438
T5	GROVER	Y	+0.003002	+0.003136
T5	GROVER	H	+0.004474	+0.004877
T5	HyenaDNA	Y	+0.006027	+0.006276
T5	HyenaDNA	H	+0.010101	+0.009809
T5	NTv3	Y	+0.040999	+0.041823
T5	NTv3	H	+0.042040	+0.042571

E.6 Absolute T2–T5 performance matrix

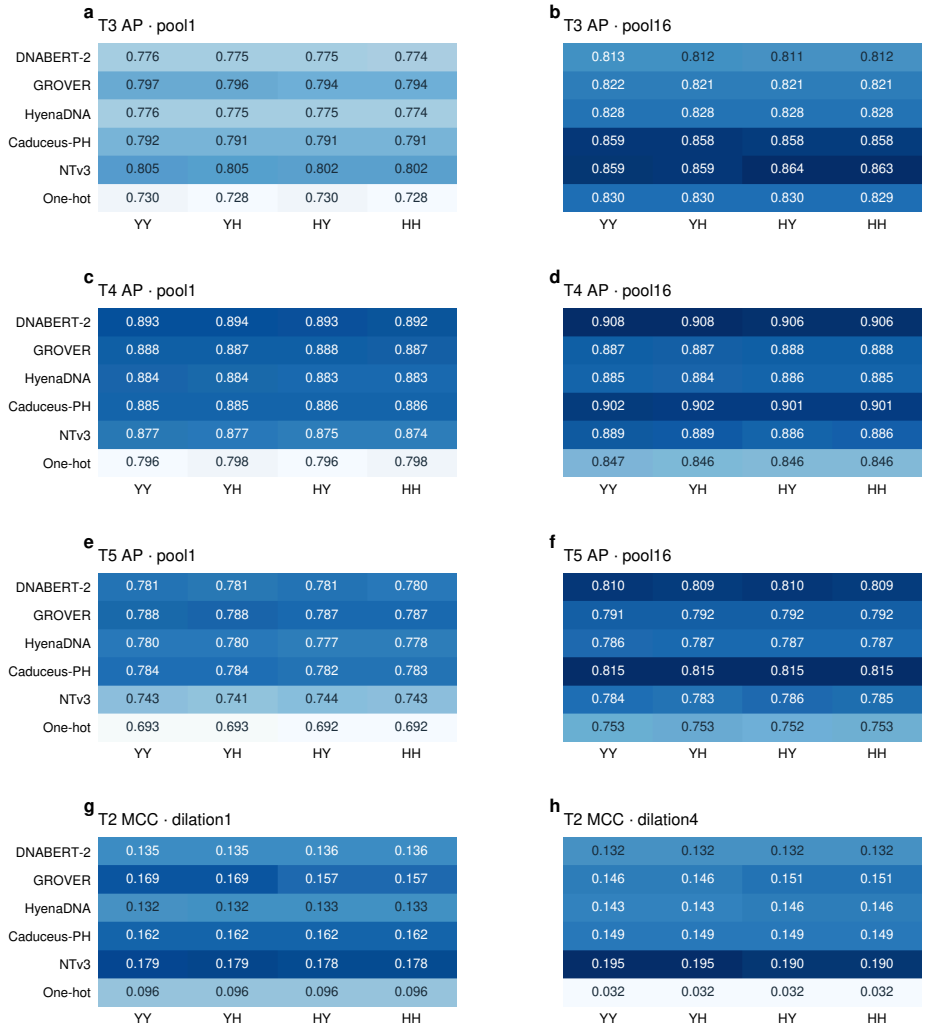

Figure S4: Absolute task performance for the T2–T5 common-protocol model settings. Rows retain five frozen models and the shallow one-hot control. Columns show YY/YH/HY/HH; the first letter denotes training reference. Each entry averages seeds 14/15/16. Both readout panels of a task share the color scale; scales differ across tasks. T2 uses full-window MCC, and binary tasks use average precision. Exact means, sample SDs and denominators appear in the source table.

F Paired correctness and probability rankings

F.1 Domains for classwise and binary correctness analyses

These descriptive analyses use the selected checkpoints and fixed decision rules for seeds 14, 15 and 16. The T2 panel includes 120 frozen-representation, 24 shallow one-hot and 12 residual-CNN prediction sets. The binary panel includes all five frozen representations on T3–T5, both readouts and both training references; the shallow binary one-hot controls are outside this state-transition summary.

For T2, predictions use the recorded three-class argmax rule. For class c, precision is TP_c/(TP_c+FP_c), and recall is TP_c/(TP_c+FN_c), using the complete evaluation domain for each denominator. The background false-positive rate is the fraction of true background observations predicted as donor or acceptor. Precision therefore counts false positives from every true-label class.

We compute each metric separately on full-window valid positions and on the shared mapped domain. The classwise comparison below uses the latter: 5,078,078 observations, comprising 5,062,118 background, 7,734 donor and 8,226 acceptor labels. Each observation receives equal weight. Overlapping windows can repeat genomic bases. Classwise paired correctness uses the same mapped domain, stratified by shared true label. These rates differ from primary window-equal flipping. Means and sample SDs summarize the three fitted seeds within each reference combination.

The archived construction labels both intronic bases of each annotated splice dinucleotide. Classwise counts follow this dinucleotide labeling convention, counting each labeled base within each containing window.

F.2 T2 CLASSWISE PERFORMANCE

All entries below are mean ± sample SD in percent on the mapped domain. All 52 model–readout–reference-combination settings are retained. Full-window results, individual-run counts and confusion matrices are supplied in the linked evidence.

Model	Readout	Pair	Donor precision	Donor recall	Acceptor precision	Acceptor recall	Background FPR
Caduceus-PH	dilation1	HH	3.578 ± 0.310	94.953 ± 1.225	2.807 ± 0.469	89.497 ± 0.637	9.079 ± 0.634
Caduceus-PH	dilation1	HY	3.575 ± 0.310	94.940 ± 1.220	2.805 ± 0.465	89.416 ± 0.734	9.080 ± 0.632
Caduceus-PH	dilation1	YH	3.579 ± 0.294	95.000 ± 1.286	2.758 ± 0.460	89.805 ± 0.796	9.192 ± 0.743
Caduceus-PH	dilation1	YY	3.576 ± 0.294	94.957 ± 1.262	2.755 ± 0.458	89.724 ± 0.838	9.194 ± 0.741
Caduceus-PH	dilation4	HH	3.264 ± 0.332	92.208 ± 1.564	2.282 ± 0.331	90.263 ± 1.344	10.570 ± 1.186
Caduceus-PH	dilation4	HY	3.265 ± 0.331	92.242 ± 1.541	2.280 ± 0.331	90.173 ± 1.376	10.570 ± 1.184
Caduceus-PH	dilation4	YH	3.157 ± 0.460	92.544 ± 1.360	2.359 ± 0.289	90.007 ± 1.324	10.511 ± 1.288
Caduceus-PH	dilation4	YY	3.156 ± 0.459	92.509 ± 1.337	2.358 ± 0.289	89.947 ± 1.329	10.509 ± 1.288
DNABERT-2	dilation1	HH	3.376 ± 0.388	90.488 ± 0.882	1.829 ± 0.091	84.014 ± 0.850	11.335 ± 0.594
DNABERT-2	dilation1	HY	3.375 ± 0.393	90.432 ± 0.838	1.826 ± 0.092	83.897 ± 0.771	11.335 ± 0.592
DNABERT-2	dilation1	YH	3.359 ± 0.462	90.290 ± 1.634	1.809 ± 0.156	83.811 ± 2.700	11.476 ± 1.029
DNABERT-2	dilation1	YY	3.360 ± 0.465	90.298 ± 1.603	1.809 ± 0.156	83.747 ± 2.757	11.469 ± 1.030
DNABERT-2	dilation4	HH	2.453 ± 0.266	87.070 ± 1.282	2.203 ± 0.223	79.966 ± 1.704	11.144 ± 0.903
DNABERT-2	dilation4	HY	2.449 ± 0.263	87.040 ± 1.251	2.199 ± 0.222	79.755 ± 1.748	11.145 ± 0.895
DNABERT-2	dilation4	YH	2.527 ± 0.301	86.230 ± 1.795	2.167 ± 0.283	80.574 ± 1.638	11.131 ± 0.370
DNABERT-2	dilation4	YY	2.529 ± 0.302	86.290 ± 1.810	2.166 ± 0.286	80.513 ± 1.523	11.128 ± 0.367
GROVER	dilation1	HH	3.616 ± 0.357	92.996 ± 0.711	2.455 ± 0.070	90.960 ± 0.335	9.689 ± 0.609
GROVER	dilation1	HY	3.620 ± 0.357	93.022 ± 0.723	2.451 ± 0.069	90.789 ± 0.292	9.686 ± 0.607

Model	Readout	Pair	Donor precision	Donor recall	Acceptor precision	Acceptor recall	Background FPR
GROVER	dilation1	YH	3.928 ± 0.211	92.630 ± 0.429	3.018 ± 0.266	89.006 ± 0.975	8.143 ± 0.585
GROVER	dilation1	YY	3.928 ± 0.211	92.669 ± 0.454	3.015 ± 0.266	88.836 ± 0.988	8.141 ± 0.585
GROVER	dilation4	HH	3.191 ± 0.360	89.260 ± 0.763	2.556 ± 0.235	85.967 ± 1.052	9.533 ± 0.705
GROVER	dilation4	HY	3.189 ± 0.360	89.260 ± 0.744	2.555 ± 0.238	85.886 ± 1.001	9.532 ± 0.703
GROVER	dilation4	YH	3.071 ± 0.120	89.260 ± 0.448	2.368 ± 0.103	86.356 ± 0.892	10.099 ± 0.284
GROVER	dilation4	YY	3.069 ± 0.122	89.255 ± 0.455	2.367 ± 0.107	86.259 ± 0.863	10.098 ± 0.292
HyenaDNA	dilation1	HH	3.038 ± 0.365	92.501 ± 1.305	1.687 ± 0.036	92.621 ± 0.791	13.334 ± 0.360
HyenaDNA	dilation1	HY	3.038 ± 0.364	92.475 ± 1.328	1.686 ± 0.036	92.556 ± 0.806	13.333 ± 0.358
HyenaDNA	dilation1	YH	2.804 ± 0.205	93.307 ± 0.756	1.733 ± 0.033	92.208 ± 0.414	13.460 ± 0.595
HyenaDNA	dilation1	YY	2.804 ± 0.206	93.272 ± 0.712	1.732 ± 0.033	92.159 ± 0.426	13.457 ± 0.593
HyenaDNA	dilation4	HH	3.215 ± 0.273	92.630 ± 0.396	2.194 ± 0.239	89.614 ± 1.591	10.836 ± 1.135
HyenaDNA	dilation4	HY	3.218 ± 0.271	92.656 ± 0.409	2.192 ± 0.239	89.505 ± 1.594	10.833 ± 1.135
HyenaDNA	dilation4	YH	3.081 ± 0.193	92.320 ± 0.392	2.157 ± 0.128	88.370 ± 0.452	10.978 ± 0.436
HyenaDNA	dilation4	YY	3.084 ± 0.194	92.371 ± 0.383	2.155 ± 0.128	88.253 ± 0.472	10.975 ± 0.437
NTv3-100 M-pre	dilation1	HH	5.611 ± 0.575	94.492 ± 0.318	2.766 ± 0.574	92.422 ± 1.853	7.897 ± 1.151
NTv3-100 M-pre	dilation1	HY	5.612 ± 0.574	94.500 ± 0.325	2.762 ± 0.573	92.264 ± 1.862	7.896 ± 1.151
NTv3-100 M-pre	dilation1	YH	5.651 ± 0.994	94.496 ± 0.856	2.865 ± 0.578	92.159 ± 1.652	7.715 ± 0.857
NTv3-100 M-pre	dilation1	YY	5.650 ± 1.000	94.492 ± 0.810	2.860 ± 0.576	92.001 ± 1.675	7.715 ± 0.855
NTv3-100 M-pre	dilation4	HH	6.492 ± 1.298	92.087 ± 2.147	3.168 ± 0.362	92.783 ± 0.596	6.747 ± 0.506
NTv3-100 M-pre	dilation4	HY	6.496 ± 1.298	92.121 ± 2.192	3.165 ± 0.362	92.661 ± 0.600	6.746 ± 0.506
NTv3-100 M-pre	dilation4	YH	6.219 ± 1.013	92.604 ± 1.818	3.483 ± 0.502	92.131 ± 1.286	6.394 ± 0.679
NTv3-100 M-pre	dilation4	YY	6.224 ± 1.020	92.669 ± 1.750	3.477 ± 0.502	91.969 ± 1.289	6.394 ± 0.680
One-hot	dilation1	HH	1.672 ± 0.073	93.298 ± 0.490	0.983 ± 0.024	93.816 ± 0.549	23.749 ± 0.526
One-hot	dilation1	HY	1.672 ± 0.073	93.272 ± 0.498	0.982 ± 0.024	93.699 ± 0.542	23.749 ± 0.526
One-hot	dilation1	YH	1.679 ± 0.067	93.195 ± 0.421	0.983 ± 0.024	93.816 ± 0.549	23.705 ± 0.521
One-hot	dilation1	YY	1.678 ± 0.067	93.169 ± 0.428	0.982 ± 0.024	93.699 ± 0.542	23.705 ± 0.521
One-hot	dilation4	HH	0.359 ± 0.005	59.870 ± 1.991	0.429 ± 0.011	61.468 ± 0.956	48.576 ± 1.871
One-hot	dilation4	HY	0.359 ± 0.006	59.922 ± 1.978	0.429 ± 0.011	61.496 ± 0.962	48.579 ± 1.871
One-hot	dilation4	YH	0.357 ± 0.010	60.236 ± 2.709	0.429 ± 0.009	61.435 ± 0.810	48.877 ± 2.394
One-hot	dilation4	YY	0.357 ± 0.010	60.292 ± 2.702	0.429 ± 0.009	61.460 ± 0.810	48.881 ± 2.393
Residual CNN	rescnn10 35	HH	8.404 ± 1.767	94.927 ± 0.792	6.806 ± 1.112	92.803 ± 0.856	3.754 ± 0.863

Model	Readout	Pair	Donor precision	Donor recall	Acceptor precision	Acceptor recall	Background FPR
Residual CNN	rescnn10 35	HY	8.403 ± 1.770	94.914 ± 0.771	6.806 ± 1.112	92.759 ± 0.862	3.753 ± 0.863
Residual CNN	rescnn10 35	YH	9.891 ± 0.184	94.233 ± 0.239	7.979 ± 0.825	92.406 ± 0.556	3.057 ± 0.224
Residual CNN	rescnn10 35	YY	9.891 ± 0.184	94.238 ± 0.235	7.977 ± 0.828	92.358 ± 0.556	3.057 ± 0.224

F.3 Correctness by T2 label

The evidence table `paired_class_states_pooled.csv` retains all 78 model–readout–training-reference–label settings. Counts pool three seeds. The corresponding per-seed table retains 234 rows. Both-correct, only-Y-correct, only-H-correct and both-wrong counts partition each stratum; flips additionally include different wrong predictions. The retained joint truth/Y-prediction/H-prediction histograms reconstruct both single-reference confusion matrices and all paired states.

Across the four reference combinations, residual-CNN mean donor precision is 8.40%–9.89% and acceptor precision is 6.81%–7.98%, despite recalls above 92%. Its background false-positive rate is 3.06%–3.75%. The residual CNN therefore combines high recall with frequent background false positives.

The classwise readout-change tables additionally retain 108 individual-run contrasts and 36 three-seed summaries for dilation4 minus dilation1 across the five frozen representations and shallow one-hot control. Every true-label stratum is retained. For HyenaDNA, the both-correct acceptor rate decreases from 92.0577% to 88.1190% under YAO training and from 92.4508% to 89.3630% under GRCh38 training, changes of -3.9387 and -3.0878 percentage points. The both-correct background rate increases by 2.4411 and 2.4610 percentage points, respectively. These position-equal changes locate the class-specific contributions within the mapped domain.

F.4 T2 probability-ranking diagnostics

We compute donor-versus-rest and acceptor-versus-rest average precision from all 48 existing test probability arrays for the residual CNN, NTv3 dilation4 and HyenaDNA dilation1/dilation4. The analysis retains seeds 14/15/16, both training references, both test references, and the same 5,078,078 mapped position observations. The positive counts are 7,734 donors and 8,226 acceptors. File/checkpoint hashes, IDs, labels and masks match the frozen records, and probability argmax matches the saved predicted classes. T2 checkpoint selection used validation MCC; AP evaluates those fixed checkpoints after selection. Average precision uses the existing implementation defined in Appendix D.1.

Model/readout	Donor AP	Acceptor AP
Residual CNN	0.545–0.562	0.470–0.498
NTv3 dilation4	0.643–0.645	0.579–0.580
HyenaDNA dilation1	0.180–0.185	0.108–0.112
HyenaDNA dilation4	0.225–0.230	0.123–0.130

Each entry spans the four three-seed means for YY/YH/HY/HH. NTv3 dilation4 exceeds the CNN in both classwise AP means in every reference combination, despite lower argmax MCC. Both HyenaDNA AP means improve with dilation4, while paired acceptor correctness declines as reported in Appendix F. AP summarizes probability rankings at the benchmark's foreground prevalence, while MCC uses the recorded argmax decisions. Validation prediction arrays were unavailable for choosing alternative class thresholds.

The MCC–AP comparison in this section uses exactly the same mapped observations and equal position weights. Recomputing MCC from the retained confusion matrices gives CNN means of 0.2828, 0.2829, 0.2604, and 0.2605 for YY/YH/HY/HH, respectively; NTv3 dilation4 gives 0.1956, 0.1957, 0.1904, and 0.1905. The CNN has higher mean MCC in every combination, while NTv3 has

higher mean donor and acceptor AP. The probability supplement retains all three class probabilities in the original precision and the mapped observation order, so these quantities can be reconstructed without private prediction paths. The supplement reconstructs these metrics from stored predictions.

F.5 Complete binary state transitions

The table retains every T3–T5 frozen-model–training-reference setting at the fixed threshold of 0.5. Counts pool seeds 14, 15 and 16. Removed-error percentage is conditional on eliminated flips; net shared-error changes use the full cohort. T3 has 23,790 seed–window observations per setting, T4 has 17,340 and T5 has 24,000. The individual-run source tables preserve all four correctness states. No case is excluded according to its direction or significance.

Task	Model	Training	Removed flips	Introduced flips	Removed → both wrong (%)	Net shared errors	seed 14	seed 15	seed 16
T3	Caduceus-PH	H	191	162	28.80	-1128	-363	-419	-346
T3	Caduceus-PH	Y	193	155	31.09	-1214	-424	-407	-383
T3	DNABERT-2	H	281	294	37.01	-532	-224	-146	-162
T3	DNABERT-2	Y	306	338	37.91	-645	-266	-141	-238
T3	GROVER	H	220	235	30.91	-482	-168	-164	-150
T3	GROVER	Y	272	234	31.62	-312	-159	-77	-76
T3	HyenaDNA	H	181	200	43.65	-883	-228	-325	-330
T3	HyenaDNA	Y	186	184	42.47	-781	-158	-225	-398
T3	NTv3-100M-pre	H	282	239	25.89	-1214	-406	-432	-376
T3	NTv3-100M-pre	Y	300	243	27.67	-1178	-440	-296	-442
T4	Caduceus-PH	H	109	125	38.53	-307	-79	-123	-105
T4	Caduceus-PH	Y	109	117	31.19	-253	-93	-69	-91
T4	DNABERT-2	H	188	159	39.89	-215	-22	-109	-84
T4	DNABERT-2	Y	174	157	45.40	-238	-71	-123	-44
T4	GROVER	H	161	123	45.34	+109	+69	+46	-6
T4	GROVER	Y	143	137	53.15	+56	+67	+6	-17
T4	HyenaDNA	H	125	141	50.40	-18	-12	+2	-8
T4	HyenaDNA	Y	121	144	45.45	-14	-2	-29	+17
T4	NTv3-100M-pre	H	176	197	42.05	-88	+4	-18	-74
T4	NTv3-100M-pre	Y	176	173	27.84	-287	-84	-42	-161
T5	Caduceus-PH	H	167	209	35.93	-854	-261	-290	-303
T5	Caduceus-PH	Y	173	211	41.62	-778	-202	-249	-327
T5	DNABERT-2	H	211	241	36.02	-495	-154	-131	-210
T5	DNABERT-2	Y	254	257	33.86	-741	-173	-355	-213
T5	GROVER	H	284	192	44.01	+27	-60	+127	-40
T5	GROVER	Y	280	195	37.50	-173	-240	+58	+9
T5	HyenaDNA	H	240	205	37.08	-233	-116	-21	-96

Task	Model	Training	Removed flips	Introduced flips	Removed → both wrong (%)	Net shared errors	seed 14	seed 15	seed 16
T5	HyenaDNA	Y	233	221	41.63	-138	-64	-37	-37
T5	NTv3-100M-pre	H	322	327	33.54	-686	-239	-216	-231
T5	NTv3-100M-pre	Y	241	291	40.25	-578	-136	-307	-135

All 30 pooled settings contain eliminated flips that end in shared errors. Full-cohort shared errors decrease in 27 settings and increase in three; the same windows and models contribute to multiple settings. This distinguishes the conditional destination of eliminated flips from the net error change across the cohort.

For GROVER T5, net shared-error changes are -240, +58 and +9 under YAO training and -60, +127 and -40 under GRCh38 training for seeds 14, 15 and 16, respectively. Their pooled sums, -173 and +27, correspond to -0.7208 and +0.1125 percentage points over 24,000 observations. The per-seed changes have mixed signs under both training references.

F.6 Absolute correctness-transition counts

The 37.5% and 44.0% values use observations with eliminated flips as their denominators (280 and 284, respectively), rather than all 24,000 seed-window observations. They quantify the fraction of those observations that are incorrect on both references after the readout change.

Training reference	Baseline flips	Changed-readout flips	Removed: both correct	Removed: both wrong	Added	Net flip change (pp)
Y	300	215	175	105	195	-0.3542
H	306	214	159	125	192	-0.3833

T5 GROVER, pool16 versus pool1. For each readout and training reference, counts pool 8,000 test windows across three fitted seeds, yielding 24,000 seed-window observations. Baseline flips = removed + retained; changed-readout flips = added + retained. Observations with eliminated flips are classified as either correct or incorrect on both references after the readout change. The same window contributes once per seed. Complete accuracy changes are reported in F.6.2. Values are derived from the previously validated transition summary, with net rates checked against the primary contrast table.

F.6.1 Complete correctness states

Training reference	Readout	Both correct	Both wrong	Only Y correct	Only H correct	Total observations	Flips
Y	pool1	17246	6454	143	157	24000	300
Y	pool16	17504	6281	99	116	24000	215
H	pool1	17443	6251	160	146	24000	306
H	pool16	17508	6278	106	108	24000	214

The four correctness states partition each row. For this binary task, flips equal only-Y-correct plus only-H-correct. Overall jointly wrong counts decrease by 173 under Y training and increase by 27 under H training; these net changes sum all cohort transitions. The individual-seed changes do not share a consistent sign under either training reference (Appendix F.5).

F.6.2 Overall accuracy at threshold 0.5

Reference combination	pool1 accuracy	pool16 accuracy	Change (percentage points)
YY	0.724542	0.733458	+0.8917
YH	0.725125	0.734167	+0.9042
HY	0.733458	0.733917	+0.0458
HH	0.732875	0.734000	+0.1125

Each value is the mean over seeds 14/15/16, equivalent here to aggregation across 24,000 equally sized seed-window observations. Accuracy is reconstructed as (both correct + correct on the specified input only) / 24,000. These values summarize accuracy over the fixed seeds. Complete-state counts and accuracy values are derived from the previously validated event source records and checked against the full transition summary. Their source hashes and reconstruction checks are recorded in review evidence.

F.7 T1 correctness states and transitions

Counts pool the three seeds, giving 24,000 seed–window observations per model, training reference and readout. Each window contributes once per seed. NTv3 abbreviates NTv3-100M-pre. In this three-class task, two wrong predictions can disagree, so flipping includes more than correctness discordance.

Model	Training reference	Readout	Both correct	Only Y correct	Only H correct	Both incorrect	Flips	Different incorrect predictions
HyenaDNA	Y	pool1	19964	93	87	3856	182	2
HyenaDNA	Y	pool16	22646	35	29	1290	64	0
HyenaDNA	H	pool1	20024	89	92	3795	183	2
HyenaDNA	H	pool16	22618	31	37	1314	68	0
Caduceus-PH	Y	pool1	16504	98	87	7311	238	53
Caduceus-PH	Y	pool16	22717	40	26	1217	66	0
Caduceus-PH	H	pool1	16472	94	80	7354	212	38
Caduceus-PH	H	pool16	22766	30	37	1167	67	0
DNABERT-2	Y	pool1	14920	197	200	8683	498	101
DNABERT-2	Y	pool16	22035	87	55	1823	144	2
DNABERT-2	H	pool1	14938	232	210	8620	536	94
DNABERT-2	H	pool16	21993	82	68	1857	151	1
GROVER	Y	pool1	20022	152	144	3682	318	22
GROVER	Y	pool16	22278	70	51	1601	121	0
GROVER	H	pool1	20083	155	128	3634	299	16
GROVER	H	pool16	22264	61	48	1627	110	1
NTv3	Y	pool1	18482	155	150	5213	334	29
NTv3	Y	pool16	22882	34	34	1050	68	0
NTv3	H	pool1	18495	160	151	5194	336	25
NTv3	H	pool16	22834	41	43	1082	84	0
One-hot	Y	pool1	13031	155	122	10692	336	59
One-hot	Y	pool16	22395	48	53	1504	101	0
One-hot	H	pool1	13052	143	120	10685	333	70
One-hot	H	pool16	22433	42	40	1485	82	0

Model	Training reference	Eliminated flips	Introduced flips	Eliminated → both correct	Eliminated → same incorrect class	Net both-incorrect change
HyenaDNA	Y	175	57	159	16	-2566
HyenaDNA	H	178	63	164	14	-2481
Caduceus-PH	Y	234	62	223	11	-6094
Caduceus-PH	H	207	62	200	7	-6187
DNABERT-2	Y	477	123	430	47	-6860
DNABERT-2	H	519	134	475	44	-6763
GROVER	Y	299	102	261	38	-2081
GROVER	H	276	87	245	31	-2007
NTv3	Y	329	63	306	23	-4163
NTv3	H	331	79	310	21	-4112
One-hot	Y	329	94	312	17	-9188
One-hot	H	325	74	308	17	-9200

Both-incorrect counts decrease across the full cohort in all model–training-reference settings, although some eliminated disagreements end in agreement on an incorrect class. The transition table identifies these new incorrect agreements; its net-change column counts all jointly incorrect predictions.

G Sensitivity and annotation checks

G.1 Same-domain dense-task sensitivity analysis

T2, dilation4 minus dilation1; means over seeds 14/15/16. All 120 frozen predictions are represented. The shared domain contains 2,544 windows and 5,078,078 mapped valid positions. MCC is computed from the pooled three-class confusion matrix across positions. GRCh38 outputs are mapped to YAO positions using the frozen correspondence; both validity masks and mapped label equality are checked. The primary tables and figures retain original full-window MCC. The table reports mean changes across the fitted seeds.

Model	Pair	Mapped MCC, dilation1	Mapped MCC, dilation4	Mapped ΔMCC	Full-window ΔMCC
HyenaDNA	YY	0.132347	0.143716	0.011369	0.011316
HyenaDNA	YH	0.132390	0.143760	0.011370	0.011310
HyenaDNA	HY	0.132762	0.146369	0.013607	0.013559
HyenaDNA	HH	0.132822	0.146420	0.013598	0.013546
Caduceus-PH	YY	0.161847	0.149293	-0.012554	-0.012528
Caduceus-PH	YH	0.161978	0.149362	-0.012616	-0.012567
Caduceus-PH	HY	0.162545	0.148738	-0.013807	-0.013773
Caduceus-PH	HH	0.162634	0.148785	-0.013849	-0.013801
DNABERT-2	YY	0.135444	0.132010	-0.003434	-0.003423
DNABERT-2	YH	0.135455	0.131986	-0.003469	-0.003466
DNABERT-2	HY	0.136434	0.131832	-0.004603	-0.004592
DNABERT-2	HH	0.136622	0.132051	-0.004571	-0.004562
GROVER	YY	0.169450	0.146671	-0.022779	-0.022698
GROVER	YH	0.169575	0.146744	-0.022831	-0.022759
GROVER	HY	0.156834	0.151043	-0.005791	-0.005760

Model	Pair	Mapped MCC, dilation1	Mapped MCC, dilation4	Mapped ΔMCC	Full-window ΔMCC
GROVER	HH	0.156929	0.151140	-0.005789	-0.005772
NTv3-100M-pre	YY	0.179651	0.195571	0.015920	0.015902
NTv3-100M-pre	YH	0.179816	0.195688	0.015871	0.015830
NTv3-100M-pre	HY	0.178185	0.190403	0.012218	0.012229
NTv3-100M-pre	HH	0.178324	0.190487	0.012163	0.012139

Model	Training reference	Position-equal Δflip (pp)	Window-equal Δflip (pp)
HyenaDNA	Y	+0.080083	+0.082859
HyenaDNA	H	+0.075324	+0.077651
Caduceus-PH	Y	+0.089706	+0.090965
Caduceus-PH	H	+0.089082	+0.090526
DNABERT-2	Y	+0.085616	+0.087126
DNABERT-2	H	+0.077037	+0.078802
GROVER	Y	+0.125618	+0.127204
GROVER	H	+0.107311	+0.108529
NTv3-100M-pre	Y	-0.022902	-0.022996
NTv3-100M-pre	H	+0.006590	+0.006791

Reproduction: the local analysis package contains all-seed aggregation and a 32-window prediction-to-metric example.

G.2 Mapping and sequence support

These analyses were specified after inspection of the original results. Subset rules were written before computing their prediction summaries; the registry analysis was specified after the direct public registry became available. The analyses use the original cohorts, fitted models and seeds 14/15/16. Subset membership is determined from mapping, sequence or annotation records before computing prediction summaries. All model settings are retained in the linked source tables. The subsets test sensitivity to mapping, sequence and annotation support.

The mapping/motif panel contains 252 fixed-head paired settings: 240 frozen-representation T2–T5 settings and 12 shallow one-hot T2 settings. The registry panel contains 120 frozen-representation T4/T5 settings. T1, the residual CNN and shallow binary-task controls are outside these new subset panels.

Mapping subsets retain recorded level-1 centers, or level-1/2 centers. T2 motif compatibility requires every foreground base in a window to participate in the annotated donor or acceptor dinucleotide in both inputs. The archived dense builder labels both intronic dinucleotide bases: donor GT/AC and acceptor AG/CT in genomic-forward sequence. For each mapped foreground pair, the strand-compatible motif types must intersect. Either adjacent dinucleotide containing the labeled base is checked. All positions, including background, remain in an accepted window. The rule tests local compatibility with the annotated splice dinucleotide. All 2,544 T2 windows contain foreground, so a foreground-bearing control retains the full cohort.

Table G1. Window retention and class composition. T2 counts are full-window background/donor/acceptor labels before restricting to mapped scoring positions. Binary counts are negative/positive. A window can belong to more than one subset.

task	subset	n_windows	background	class1	class2	unit
T2	all	2544	5082216	7734	8226	full_valid_positions
T2	mapping_L1	2054	4102901	6454	6861	full_valid_positions
T2	mapping_L1_L2	2454	4902382	7476	7958	full_valid_positions
T2	motif_supported	2417	4828622	7272	7774	full_valid_positions
T3	all	7930	3945	3985	0	windows

task	subset	n_windows	background	class1	class2	unit
T3	mapping_L1	5693	2593	3100	0	windows
T3	mapping_L1_L2	7520	3680	3840	0	windows
T4	all	5780	2890	2890	0	windows
T4	mapping_L1	3785	1491	2294	0	windows
T4	mapping_L1_L2	5330	2520	2810	0	windows
T5	all	8000	4000	4000	0	windows
T5	mapping_L1	4328	1822	2506	0	windows
T5	mapping_L1_L2	7433	3547	3886	0	windows

Table G2. HyenaDNA dilation4 minus dilation1 in each fixed subset. MCC here uses valid mapped positions, whereas primary task MCC and Table G4 use full-window labels. Flip differences weight windows equally. Values are means over the same three seeds; no new hypothesis tests are performed.

Training reference	Subset	Windows	ΔMCC Y input	ΔMCC H input	Δflip (pp)
H	all	2544	+0.01361	+0.01360	+0.07765
H	mapping_L1	2054	+0.01397	+0.01396	+0.08718
H	mapping_L1_L2	2454	+0.01379	+0.01376	+0.08569
H	motif_supported	2417	+0.01370	+0.01366	+0.08619
Y	all	2544	+0.01137	+0.01137	+0.08286
Y	mapping_L1	2054	+0.01163	+0.01164	+0.09206
Y	mapping_L1_L2	2454	+0.01158	+0.01158	+0.08963
Y	motif_supported	2417	+0.01147	+0.01145	+0.09086

The motif-compatible subset retains 2,417 windows. Both the mean MCC increase and the mean flip increase persist for HyenaDNA under either training reference. Mapping restrictions also preserve these directions. Because mapping restrictions change class prevalence, binary AP comparisons use the same subset for both readouts. The full tables report within-subset readout comparisons, task performance and all four correctness states for every setting.

G.3 Direct regulatory annotation support

We downloaded the public GRCh38 cCRE Registry V4 supplementary BED (2,348,854 records). A positive window is supported when its source identifier, after removing a terminal version suffix, has the relevant class in both the production YAO cCRE file and the public GRCh38 registry, and the recorded target-window midpoint falls within that same GRCh38 element. Relevant classes are PLS for T4 and pELS/dELS for T5. A compatible negative has no full-window overlap with the relevant GRCh38 class. The registry check tests annotation consistency. The registry and transferred labels can share upstream annotation sources; functional validation would require independent assay labels.

Task	All windows (negative/positive)	Retained windows (negative/positive)	Excluded
T4	5,780 (2,890/2,890)	5,767 (2,890/2,877)	13 positive ID/class/center failures
T5	8,000 (4,000/4,000)	7,988 (3,993/3,995)	5 positive failures; 7 negative overlaps

Table G3. GROVER T5 transitions at threshold 0.5. Counts pool the three seeds; each window contributes once per seed. Each supported analysis removes the same windows from both readouts and both inputs.

Subset	Training reference	Seed-windows	Flips before → after	Shared errors before → after	Removed flips ending wrong
all	Y	24000	300 → 215	6454 → 6281	105/280 (37.5%)
all	H	24000	306 → 214	6251 → 6278	125/284 (44.0%)
registry_supported	Y	23964	299 → 213	6441 → 6266	104/279 (37.3%)
registry_supported	H	23964	305 → 213	6240 → 6262	124/283 (43.8%)

On the registry-supported T5 subset, GROVER retains the mean flip reduction and the same opposing directions of total shared-error changes under Y and H training. The removed-flip error fractions are 37.3% and 43.8%, respectively. The shared-error pattern therefore persists in the registry-supported subset.

G.4 Common-threshold sensitivity

For GROVER T5, we apply each threshold in the fixed grid 0.1, 0.2, …, 0.9 to both reference inputs of each fitted head. The primary operating threshold remains 0.5. AP is unchanged because it ranks probabilities across thresholds. Counts below pool 8,000 windows and three seeds per training reference.

Training reference	Threshold	Δflips	Δshared errors	Removed flips ending wrong
Y	0.1	-23	-344	77/140
Y	0.2	-72	-203	88/211
Y	0.3	-91	-49	91/235
Y	0.4	-74	+52	98/242
Y	0.5	-85	-173	105/280
Y	0.6	-31	-701	84/227
Y	0.7	+19	-1080	73/202
Y	0.8	+59	-1114	42/93
Y	0.9	+13	-604	18/43
H	0.1	-66	+171	81/163
H	0.2	-88	+148	106/213
H	0.3	-90	+81	103/238
H	0.4	-88	+62	112/262
H	0.5	-92	+27	125/284
H	0.6	-86	-352	100/271
H	0.7	-31	-650	96/244
H	0.8	+10	-616	71/137
H	0.9	-13	-241	31/60

Both training references show fewer flips at thresholds 0.1–0.6. At 0.8, both show more flips; directions at other high thresholds differ. Every tested head–threshold setting includes removed flips that end in shared errors. The direction of the flip change depends on the operating threshold, and every tested threshold retains newly shared errors.

G.5 Uncertainty of the HyenaDNA MCC contrast

We resample source-midpoint genomic blocks within chromosomes, with shared block multiplicities across readouts, inputs and seeds. For each of 5,000 replicates (random seed 20260915), full-window confusion matrices are pooled within each seed, MCC differences are computed, and those differences are averaged over seeds 14/15/16. The 1 Mb and 2 Mb definitions contain 340 and 193 occupied blocks. These pointwise percentile intervals are unadjusted and condition on the three fitted seeds and frozen cohort.

Table G4. Full-window MCC: dilation4 minus dilation1.

Pair	Mean difference	1 Mb 95% interval	2 Mb 95% interval
YY	0.01132	[0.01027, 0.01235]	[0.01025, 0.01240]
YH	0.01131	[0.01028, 0.01235]	[0.01024, 0.01241]
HY	0.01356	[0.01260, 0.01449]	[0.01251, 0.01464]
HH	0.01355	[0.01257, 0.01448]	[0.01251, 0.01460]

All eight conditional intervals exclude zero in the positive direction. The MCC intervals form a separate conditional uncertainty analysis from the primary BH-corrected flip contrasts.

Figure 3e averages the two test-reference MCC differences within each stored 1 Mb replicate. The resulting mean and pointwise 95% interval are 0.01131 [0.01028, 0.01235] for Y training and 0.01355 [0.01258, 0.01449] for H training. The calculation uses the shared block weights and fitted seeds above. The figure source values retain the unrounded results. Appendix G.4 and Figure S5 report the common-threshold and shared-error sensitivity analyses.

G.6 Threshold and MCC sensitivity plots

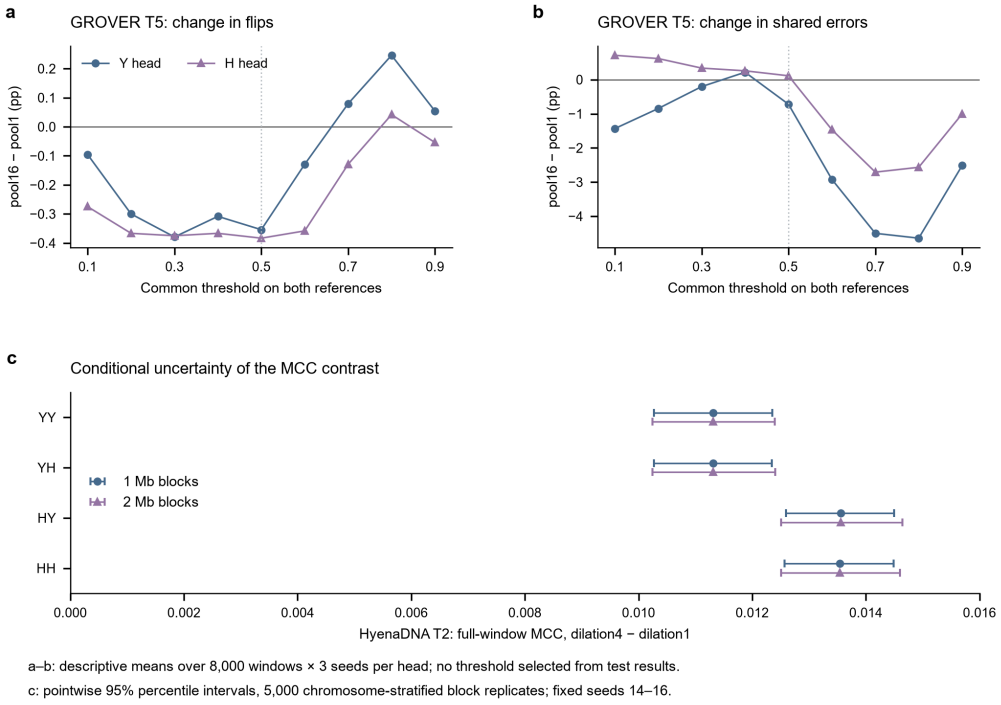

Figure S5: Supplementary sensitivity of the two key readout examples. (a–b) GROVER T5 pool16 minus pool1 changes in flips and total shared errors under a common threshold for both inputs. Counts are normalized by 8,000 windows × three seeds for each training reference; these are descriptive comparisons, and no operating point is selected. (c) HyenaDNA T2 full-window MCC differences with pointwise 95% percentile intervals from 5,000 chromosome-stratified block replicates at 1 Mb and 2 Mb. Block multiplicities are shared across paired inputs, readouts and the fixed seeds 14/15/16. These pointwise intervals are unadjusted and conditional on the fitted seeds.

G.7 T1 central-dinucleotide and catalog checks

This post-hoc data-quality check uses the unchanged fixed sequences and accepted catalog for all 65,997 T1 pairs. It checks 402-bp lengths, stored coordinates, input orientations, catalog labels and oriented sequence hashes. For the central dinucleotide at zero-based oriented indices 200–201, donor labels are checked for GT, acceptor labels for AG, and background decoys for either

motif. The check tests nominal anchor compatibility. Both positive and decoy examples can contain GT/AG. The historical construction and strand-specific anchoring conventions remain documented in Appendix A.1. All accepted examples and labels remain in the evaluation.

Split	Paired windows	Both-reference compatible	Background / donor / acceptor compatible	Exceptions
train	49,997	49,920	24,928 / 12,495 / 12,497	77
valid	8,000	7,983	3,983 / 2,000 / 2,000	17
test	8,000	7,989	3,990 / 1,999 / 2,000	11

All catalog coordinate/orientation/label checks and sequence hashes agree. In the test set, ten background pairs and one donor pair fail the central-motif rule on at least one reference. The donor has AG at the nominal central dinucleotide on both inputs. These exceptions are retained in primary evaluation and listed in `t1_central_motif_exceptions.csv`; the shared task labels remain the correctness targets.

G.8 Source-candidate context

New stable candidate unions by task are T2 28,855 observations (83 positive), T3 74 (28 positive), T4 42 (19 positive), and T5 47 (20 positive), totaling 29,018 (150 positive). T2 units are window-by-position observations; binary units are windows. The union combines candidates screened separately under two readouts, giving two selection opportunities per setting. This candidate screening retains its T2–T5 common-protocol scope. T1 retains its earlier candidate analysis; candidate screening for T1 follows its earlier protocol. Earlier 8/8 and newer 3/3 stability definitions remain separate. The 83 positive T2 observations correspond to 81 unique positive source positions and 57 source events; all 150 positive observations across tasks match their source annotations.

A stable S/S candidate flips in every run, without requiring the same correctness direction each time. We retain positive and background candidates and check their annotations, sequence identity, labels, and output mappings. Unions across readouts are recorded separately because they increase selection opportunities. Annotation agreement establishes the provenance of each candidate's task label.

One audited T2 background candidate had identical 2,004-bp input sequences but mapped zero-based output indices of 2000 and 1999, near the window boundary (Figure S3c). Comparing those outputs can yield a difference even when equal-index predictions agree. Earlier splice cases also included mapped motif bases that were nonadjacent on the target reference. Interpreting these candidates requires checking output correspondence and transferred-label geometry before attributing them to sequence-specific function. Across the full panel, 774 T2 windows have identical inputs and exactly matching probabilities at equal indices across all 72 paired settings, yet mapped correspondence contributes 110 of 1,577,548 pooled flips (0.00697%). These pooled counts repeat observations across settings. The example illustrates the computation behind this small subset of panel flips (Appendix B.5).

H Supervised comparators

H.1 Supervised one-hot controls for T2–T5

These controls add 48 heads and 96 predictions to the 240-head frozen-representation study. They retain the original frozen T2/T4/T5 data, repaired-original T3 v2 data, chromosome splits, seeds 14/15/16, labels, sequence orientation and validation-selection protocol. D10 and D50/enriched are excluded. A/C/G/T/N are five one-hot channels, with non-ACGT characters mapped to N; N remains a valid input, while T2 label -100 is excluded from loss and evaluation. The two-layer CNN and optimization settings follow Section 3.3 and Appendix C.3. Replacing the representation by five channels changes the first layer's parameter count and removes encoder context. The comparison measures the combined effects of representation, encoder context and first-layer capacity. All 48 heads were selected before test prediction. The T1 common-protocol controls are reported in Appendix E; earlier T1 controls remain in Appendices J.1 and J.2.

H.1.1 reports mean ± sample SD over seeds 14/15/16; H.1.2 links the source records. Columns are training reference followed by test reference: YY/YH share the Y-trained head and HY/HH share the H-trained head. T2 MCC uses full-window valid labels; binary tasks use AP. SDs summarize variation across the three fitted runs.

H.1.1 Primary metric: all reference combinations

Task	Metric	Readout	YY	YH	HY	HH
T2	MCC	dilation1	0.096315 ± 0.000932	0.096386 ± 0.000929	0.096265 ± 0.000959	0.096337 ± 0.000955
T2	MCC	dilation4	0.031978 ± 0.000479	0.031955 ± 0.000478	0.032005 ± 0.000354	0.031978 ± 0.000357
T3	AP	pool1	0.730018 ± 0.000303	0.728413 ± 0.000499	0.729831 ± 0.000653	0.728127 ± 0.000669
T3	AP	pool16	0.830222 ± 0.002749	0.829503 ± 0.003004	0.830015 ± 0.002186	0.829439 ± 0.002420
T4	AP	pool1	0.796381 ± 0.004428	0.797539 ± 0.003942	0.796151 ± 0.004273	0.797501 ± 0.003753
T4	AP	pool16	0.846877 ± 0.002509	0.846486 ± 0.002202	0.845968 ± 0.001835	0.845856 ± 0.001747
T5	AP	pool1	0.692685 ± 0.001231	0.692671 ± 0.001591	0.691995 ± 0.001447	0.692207 ± 0.001645
T5	AP	pool16	0.753092 ± 0.000868	0.753395 ± 0.000892	0.751746 ± 0.002187	0.752528 ± 0.002275

H.1.2 Source records

The complete run records retain the measurements underlying the three-seed summaries in H.1.1.

H.1.3 Frozen probe minus one-hot: descriptive mean differences

All 160 model × task × readout × reference comparisons are shown, including all 12 negative differences. Differences compare the complete frozen-probe and supervised one-hot systems. Raw comparison records also retain paired-seed difference SDs.

Task	Metric	Readout	Frozen representation	YY	YH	HY	HH
T2	MCC	dilation1	HyenaDNA	+0.035797	+0.035746	+0.036269	+0.036228
T2	MCC	dilation1	Caduceus-PH	+0.065293	+0.065305	+0.066031	+0.066003
T2	MCC	dilation1	DNABERT-2	+0.038933	+0.038860	+0.039972	+0.040075
T2	MCC	dilation1	GROVER	+0.072852	+0.072870	+0.060315	+0.060306
T2	MCC	dilation1	NTv3-100M-pre	+0.082981	+0.083070	+0.081553	+0.081633
T2	MCC	dilation4	HyenaDNA	+0.111451	+0.111488	+0.114087	+0.114133
T2	MCC	dilation4	Caduceus-PH	+0.117102	+0.117169	+0.116518	+0.116561
T2	MCC	dilation4	DNABERT-2	+0.099848	+0.099826	+0.099639	+0.099873
T2	MCC	dilation4	GROVER	+0.114491	+0.114543	+0.118814	+0.118894
T2	MCC	dilation4	NTv3-100M-pre	+0.163221	+0.163332	+0.158041	+0.158131
T3	AP	pool1	HyenaDNA	+0.046031	+0.046375	+0.045440	+0.046292
T3	AP	pool1	Caduceus-PH	+0.061609	+0.063000	+0.061622	+0.063117
T3	AP	pool1	DNABERT-2	+0.045897	+0.047067	+0.045304	+0.046167
T3	AP	pool1	GROVER	+0.066795	+0.067997	+0.064031	+0.065380
T3	AP	pool1	NTv3-100M-pre	+0.074847	+0.076136	+0.071865	+0.073488
T3	AP	pool16	HyenaDNA	-0.001892	-0.001551	-0.001581	-0.001212
T3	AP	pool16	Caduceus-PH	+0.028363	+0.028717	+0.027603	+0.028202
T3	AP	pool16	DNABERT-2	-0.017516	-0.017711	-0.018783	-0.017889
T3	AP	pool16	GROVER	-0.008430	-0.008312	-0.008672	-0.008126
T3	AP	pool16	NTv3-100M-pre	+0.029180	+0.029178	+0.034155	+0.034002
T4	AP	pool1	HyenaDNA	+0.088029	+0.086141	+0.087171	+0.085252
T4	AP	pool1	Caduceus-PH	+0.088817	+0.087774	+0.089482	+0.088260
T4	AP	pool1	DNABERT-2	+0.096902	+0.096055	+0.096391	+0.094994

Task	Metric	Readout	Frozen representation	YY	YH	HY	HH
T4	AP	pool1	GROVER	+0.091697	+0.089943	+0.091399	+0.089375
T4	AP	pool1	NTv3-100M-pre	+0.080672	+0.079166	+0.078402	+0.076280
T4	AP	pool16	HyenaDNA	+0.038285	+0.037906	+0.039743	+0.039051
T4	AP	pool16	Caduceus-PH	+0.054902	+0.055190	+0.055503	+0.055398
T4	AP	pool16	DNABERT-2	+0.061424	+0.061688	+0.060503	+0.060623
T4	AP	pool16	GROVER	+0.040560	+0.040742	+0.042178	+0.042176
T4	AP	pool16	NTv3-100M-pre	+0.042269	+0.042396	+0.039789	+0.040218
T5	AP	pool1	HyenaDNA	+0.087463	+0.087766	+0.084521	+0.085299
T5	AP	pool1	Caduceus-PH	+0.091446	+0.091550	+0.090449	+0.090671
T5	AP	pool1	DNABERT-2	+0.088576	+0.088192	+0.089199	+0.088221
T5	AP	pool1	GROVER	+0.095514	+0.095766	+0.095190	+0.094881
T5	AP	pool1	NTv3-100M-pre	+0.050511	+0.048563	+0.052278	+0.050681
T5	AP	pool16	HyenaDNA	+0.033084	+0.033317	+0.034871	+0.034787
T5	AP	pool16	Caduceus-PH	+0.062024	+0.061521	+0.063487	+0.062731
T5	AP	pool16	DNABERT-2	+0.057276	+0.055839	+0.057802	+0.056338
T5	AP	pool16	GROVER	+0.038109	+0.038178	+0.039913	+0.039438
T5	AP	pool16	NTv3-100M-pre	+0.031103	+0.029662	+0.034567	+0.032931

H.1.4 ONE-HOT READOUT CONTRASTS IN FLIPPING

T2 uses window-equal flipping at mapped valid positions; binary tasks use window flipping. The 5,000 within-chromosome block replicates condition on seeds 14/15/16. Within each width, baseline BH families contain six binary contrasts and two dense contrasts, separately from the frozen-probe families. Discoveries require the same direction and $q < 0.05$ at both widths. No binary contrast meets this rule; flipping increases under both T2 training references. The intervals quantify changes in cross-reference flipping.

Task	Training reference	Contrast	Block width	Δflip (pp)	95% interval (pp)	BH q
T3	Y	pool16-pool1	1 Mb	+0.008407	[-0.217087, +0.230931]	0.942412
T3	H	pool16-pool1	1 Mb	-0.033628	[-0.277178, +0.205676]	0.942412
T3	Y	pool16-pool1	2 Mb	+0.008407	[-0.203308, +0.217312]	0.929814
T3	H	pool16-pool1	2 Mb	-0.033628	[-0.265150, +0.182523]	0.919256
T4	Y	pool16-pool1	1 Mb	+0.034602	[-0.180930, +0.246620]	0.942412
T4	H	pool16-pool1	1 Mb	-0.063437	[-0.269549, +0.133136]	0.942412
T4	Y	pool16-pool1	2 Mb	+0.034602	[-0.173936, +0.248482]	0.919256
T4	H	pool16-pool1	2 Mb	-0.063437	[-0.276175, +0.138921]	0.919256
T5	Y	pool16-pool1	1 Mb	-0.045833	[-0.271538, +0.168461]	0.942412
T5	H	pool16-pool1	1 Mb	-0.150000	[-0.382976, +0.070574]	0.942412
T5	Y	pool16-pool1	2 Mb	-0.045833	[-0.261340, +0.166528]	0.919256
T5	H	pool16-pool1	2 Mb	-0.150000	[-0.384867, +0.079900]	0.919256
T2	Y	dilation4-dilation1	1 Mb	+0.089143	[+0.076929, +0.104163]	0.000200
T2	H	dilation4-dilation1	1 Mb	+0.091215	[+0.077050, +0.108706]	0.000200
T2	Y	dilation4-dilation1	2 Mb	+0.089143	[+0.077366, +0.103397]	0.000200

Task	Training reference	Contrast	Block width	Δflip (pp)	95% interval (pp)	BH q
T2	H	dilation4-dilation1	2 Mb	+0.091215	[+0.077176, +0.107845]	0.000200

Sources: protocol, all-seed predictions and checkpoint hashes, full comparisons, flipping records, analysis validation.

H.1.5 SUMMARY OF THE ONE-HOT COMPARISONS

Frozen probes exceeded the one-hot mean in 148 of 160 comparisons: 40/40 for T2, 28/40 for T3, and 40/40 each for T4 and T5. All 12 exceptions occurred in T3 pool16. HyenaDNA, DNABERT-2, and GROVER fell below one-hot in all four reference combinations by 0.001212–0.018783 AP. The count summarizes comparisons against the specified shallow supervised architecture, with shared test windows across settings.

One-hot performance and frozen-probe comparisons. Ranges span four reference combinations of three-seed means. Each task has five representations × two readouts × four combinations. Appendix H.1 gives means ± SD, model differences and links to source records.

Task	Metric	One-hot first readout: mean range	One-hot second readout: mean range	Probe mean above one-hot
T2	MCC	dilation1: 0.0963–0.0964	dilation4: 0.0320–0.0320	40/40
T3	AP	pool1: 0.7281–0.7300	pool16: 0.8294–0.8302	28/40
T4	AP	pool1: 0.7962–0.7975	pool16: 0.8459–0.8469	40/40
T5	AP	pool1: 0.6920–0.6927	pool16: 0.7517–0.7534	40/40

One-hot pool16 improved mean AP for T3–T5, while flipping intervals allowed increases and decreases. The direction of the binary flip change remains uncertain under the primary interval estimates. For T2, dilation4 lowered mean MCC from approximately 0.0963–0.0964 to 0.0320. Flipping increased by 0.089143 and 0.091215 percentage points under Y and H training, passing correction at both widths: both outcomes worsened.

Checks of all 12 T2 heads and 24 predictions detected no execution anomaly in the inspected records (Appendix H.2). The two dilation4 convolutions sample raw-input offsets $\{-8,-4,0,4,8\}$, excluding immediate neighbors. This restricted support offers a plausible explanation for the one-hot decrease and limits interpretation of the probe–baseline gap.

H.2 DIAGNOSTIC AUDIT OF THE T2 ONE-HOT DECREASE

An audit of all 12 heads and 24 predictions found no execution anomaly in the inspected records. Losses, validation scores and saved parameter tensors were finite; selection of the earliest best epoch and patience 5 matched the histories; class weights matched complete training counts. Each epoch covered 14,077 training windows in 440 batches, using batch size 32 except for the final batch of 29. Paired readouts shared initial convolutional parameters and the checked batch-order prefix. Each head had 13,571 parameters. All 24 predictions used all three classes, excluding complete majority-class collapse. Best validation MCC was 0.093342–0.094930 for dilation1 and 0.031476–0.032124 for dilation4: the decrease was already present during validation.

Both width-3 convolutions use the same dilation. At interior output positions, their raw-input support is $\{-2,-1,0,1,2\}$ for dilation1 and $\{-8,-4,0,4,8\}$ for dilation4. The five sampled offsets span 17 bp. The latter cannot access immediate offsets $\pm1/\pm2$ in raw one-hot input, whereas contextual encoder outputs can already carry information beyond their position. This restricted raw-input support provides a possible explanation for the decrease under the frozen protocol.

The training records retain checkpoint selection, epoch counts, validation scores and losses.

Audit sources: all epoch histories, 24 prediction diagnostics, validation and source hashes.

H.3 T2 SUPERVISED RESIDUAL CNN

The T2ResidualCNN architecture and training protocol were fixed after the preceding analyses and before training this comparator. We completed all six training runs and twelve test predictions. Appendices H.3.2–H.3.3 and Figure S6 report the results. It remains separate from the historical shallow one-hot counts and from the original significance families. The frozen protocol uses original T2 inputs of 2,004 bp, 14,077/2,162/2,544 training/validation/test windows per reference, the unchanged chromosome split, seeds 14/15/16, and two training references: six training runs and twelve paired test predictions.

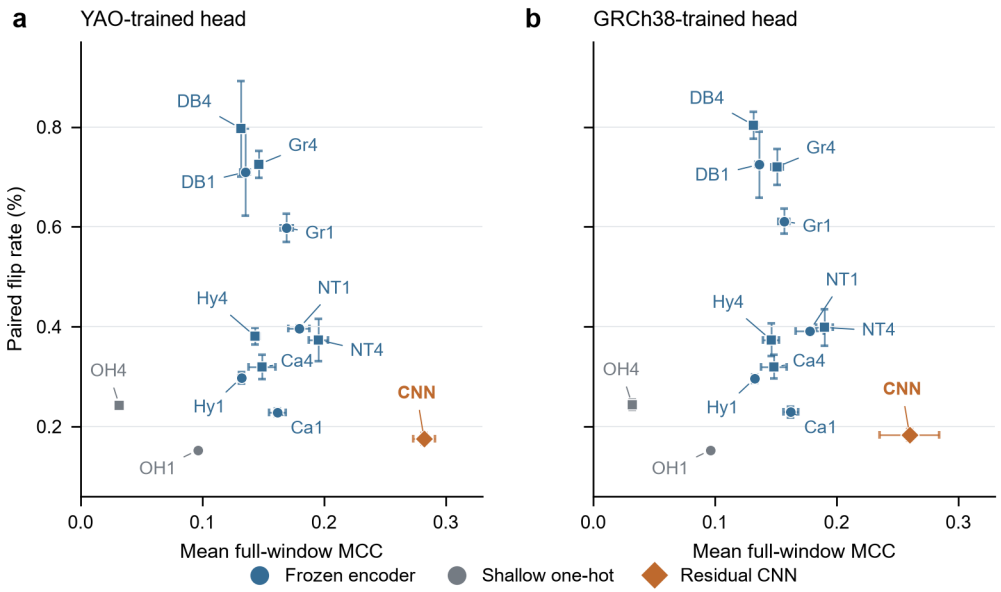

Figure S6: Joint task performance and reference consistency for all T2 pipelines. Panels separate YAO- and GRCh38-trained models. Horizontal values average full-window MCC over both test references; vertical values use mapped window-equal flipping. Markers show means over seeds 14/15/16; bars show sample SD across the three runs on each axis. All ten frozen model–readout settings, two shallow one-hot controls and the residual CNN are included in each panel. These descriptive comparisons evaluate complete pipelines with their respective capacities and input contexts.

T2ResidualCNN has 798,595 parameters: a width-15 input convolution, 128 channels, and eight two-convolution residual blocks with dilation 1, 2, 4, 8, 16, 32, 64, and 128. Position-wise non-affine channel RMS normalization, GELU, and dropout 0.1 preserve a continuous theoretical input support of 1,035 bp. Float32 AdamW uses learning rate 0.001, weight decay 0.0001, batch size 32 with the final incomplete batch retained, gradient clipping at 1, at most 20 epochs, and patience 5. Class weights are computed from training labels only. Each checkpoint is selected at the earliest epoch attaining the highest full-window validation MCC. Validation inputs use the head-training reference: YAO for YAO-trained heads and GRCh38 for GRCh38-trained heads. The protocol requires all six selections to be finalized before any test prediction.

Before training, implementation checks covered manual weighted-loss calculations, deterministic optimization, checkpoint reloading, input hashes, split membership, and frozen test-target identity. Evaluation retains all seeds, primary full-window MCC, mapped-domain MCC, complete correctness states, foreground and geometry diagnostics. The comparison measures performance and disagreement for each complete pipeline under its specified capacity and context. Execution records and source files: supervised extension.

H.3.1 CHECKPOINT SELECTION AND TRAINING RECORDS

We finalized all six checkpoint selections before testing and subsequently verified all six checkpoints, twelve predictions, nine result tables and the unchanged metadata, target and mapping files. No run failed or was discarded. Four runs reached the 20-epoch cap, and one selected the final epoch. No extension of the training budget was selected using test performance. The complete training records retain checkpoint selections and performance measurements; H.3.2 reports the three-seed means and sample SDs.

Training/validation worker times sum to 6.785 hours across six A800 GPUs; the total sums each worker's training and validation timer. Peak allocated CUDA tensor memory was approximately 2.71 GiB per worker. Input preparation, test inference and analysis are outside those training timers.

H.3.2 COMPLETE TASK-PERFORMANCE COMPARISON

All values are three-seed mean ± sample SD. The table includes every frozen model/readout, both shallow controls and the residual CNN. MCC includes all valid background, donor and acceptor observations in the full input windows. All 40 CNN-versus-frozen mean differences are positive; the settings share the same test cohort.

Model	Readout	YY	YH	HY	HH
HyenaDNA	dilation1	0.1321 ± 0.0023	0.1321 ± 0.0023	0.1325 ± 0.0017	0.1326 ± 0.0017
HyenaDNA	dilation4	0.1434 ± 0.0033	0.1434 ± 0.0033	0.1461 ± 0.0066	0.1461 ± 0.0066
Caduceus-PH	dilation1	0.1616 ± 0.0068	0.1617 ± 0.0069	0.1623 ± 0.0061	0.1623 ± 0.0062
Caduceus-PH	dilation4	0.1491 ± 0.0112	0.1491 ± 0.0112	0.1485 ± 0.0104	0.1485 ± 0.0104
DNABERT-2	dilation1	0.1352 ± 0.0038	0.1352 ± 0.0039	0.1362 ± 0.0030	0.1364 ± 0.0029
DNABERT-2	dilation4	0.1318 ± 0.0028	0.1318 ± 0.0028	0.1316 ± 0.0034	0.1319 ± 0.0035
GROVER	dilation1	0.1692 ± 0.0051	0.1693 ± 0.0051	0.1566 ± 0.0044	0.1566 ± 0.0044
GROVER	dilation4	0.1465 ± 0.0023	0.1465 ± 0.0021	0.1508 ± 0.0047	0.1509 ± 0.0047
NTv3	dilation1	0.1793 ± 0.0087	0.1795 ± 0.0088	0.1778 ± 0.0118	0.1780 ± 0.0118
NTv3	dilation4	0.1952 ± 0.0078	0.1953 ± 0.0079	0.1900 ± 0.0069	0.1901 ± 0.0069
One-hot	dilation1	0.0963 ± 0.0009	0.0964 ± 0.0009	0.0963 ± 0.0010	0.0963 ± 0.0010
One-hot	dilation4	0.0320 ± 0.0005	0.0320 ± 0.0005	0.0320 ± 0.0004	0.0320 ± 0.0004
Residual CNN	rescnn1035	0.2823 ± 0.0088	0.2823 ± 0.0087	0.2600 ± 0.0245	0.2599 ± 0.0244

H.3.3 COMPLETE PAIRED FLIPPING COMPARISON

Values below are mean ± sample SD of mapped window-equal flipping, in percent. Y/H identify training reference. Each seed uses 2,544 windows and 5,078,078 mapped valid position observations. The residual CNN has lower mean flipping than each corresponding frozen pipeline, while the shallow dilation1 control remains lower. The table reports each pipeline's absolute rate.

Model	Readout	Y-trained flip (%)	H-trained flip (%)
HyenaDNA	dilation1	0.2976 ± 0.0116	0.2957 ± 0.0087
HyenaDNA	dilation4	0.3805 ± 0.0166	0.3734 ± 0.0327
Caduceus-PH	dilation1	0.2284 ± 0.0078	0.2293 ± 0.0109
Caduceus-PH	dilation4	0.3194 ± 0.0248	0.3198 ± 0.0232
DNABERT-2	dilation1	0.7096 ± 0.0877	0.7247 ± 0.0661
DNABERT-2	dilation4	0.7967 ± 0.0956	0.8035 ± 0.0273
GROVER	dilation1	0.5978 ± 0.0282	0.6113 ± 0.0251
GROVER	dilation4	0.7250 ± 0.0273	0.7199 ± 0.0364
NTv3	dilation1	0.3964 ± 0.0061	0.3915 ± 0.0052
NTv3	dilation4	0.3734 ± 0.0423	0.3983 ± 0.0370
One-hot	dilation1	0.1524 ± 0.0035	0.1524 ± 0.0038
One-hot	dilation4	0.2415 ± 0.0061	0.2437 ± 0.0105
Residual CNN	rescnn1035	0.1756 ± 0.0113	0.1832 ± 0.0120

For T2ResidualCNN, foreground flipping is 0.3791%/0.3153% under Y/H training. The $\geq$128 bp interior group uses the same definition as the other diagnostics; the CNN's 517 bp support radius

can still reach window padding for some retained outputs. Foreground-only MCC uses the subset of donor/acceptor-labeled observations. Complete states, subgroup diagnostics, probabilities' recorded identities and all underlying measurements are retained in the results. The four-input numerical entry is described in the reproduction guide.

I Reproducibility and verification

I.1 Available materials and entry points

The compact local candidate contains fixed benchmark inputs, scoring code, numerical reconstruction records and a selected model replay. The larger author-side probability archive is not included in this candidate. The table distinguishes what reviewers can check from the compact package from analyses retained only in the author-side inventory. Appendix I.5 records source completeness and delivery status.

Reproduction route	Available material	Verification scope
Fixed-input evaluation	All five tasks and three splits; paired sequences, labels and T2 validation/test maps; common scorer	Data identities, correspondence and scoring on fixed inputs (I.2)
Numerical reconstruction	T1 compact predictions; T2–T5 summaries; correspondence replicates; T2 confusion matrices and joint histograms; binary transitions	Reported metrics, paired states and corrected contrasts (I.4)
Score-level AP diagnostics	Author-side 48 T2 probability arrays, omitted from the compact candidate	Reported results appear in Appendix F.4; independent AP recomputation from original scores is outside the compact delivery
Selected model replay	Full-test T5 one-hot checkpoint in the compact core; HyenaDNA T2 and GROVER T5 sampled execution records retained author-side	Core CPU baseline can be run; encoder sample checks are documented in I.3 but not part of the compact candidate
Rebuilding and retraining	Author-side historical scripts and records; no complete portable upstream builder or five-encoder retraining package	Scope and missing production/execution records are listed in I.5

The compact local candidate contains `fixed-input.zip`, a construction-and-workflow guide, a current claim-to-file index, a small compatibility diagnostic and the Appendix E.4 source table; external download access remains pending (Appendix I.5). The core ZIP contains `unpack.py` and `payload.tar.xz`. Running its unpacker produces `review_artifact/GenomeBench-T2T-review-v34/`; that directory name identifies the preserved packaged-data version. The separate 2.66 GB `analysis-supplement.zip` remains in the author-side inventory and is not part of the compact candidate. The reproduction guide maps this manuscript to the compact package entries.

The core `smoke_test.py` checks the manifest, numerical summaries, correspondence, data, maps, T1 predictions and classwise states, with an optional CPU baseline. The included small diagnostic reconstructs T1 compatibility and mapping-level summaries. Score-level AP reconstruction from the original 48 T2 probability arrays requires the separate author-side archive and is outside this compact candidate. The local source index records the author-side provenance paths.

I.2 Fixed inputs, correspondence and scoring

The fixed-input package contains 15 task/split files covering all 254,366 accepted pairs. All 508,732 oriented sequence hashes match the coordinate catalog. Labels match the accepted T1 catalog or the prepared T2–T5 targets, and train, validation and test IDs are disjoint within each task. Files provide both sequences, labels, validity masks, coordinates, orientation and center-mapping level. T2 validation and test files include `h_index_for_y_base`; the test map defines 5,078,078 valid

mapped observations. Portable chain-block reconstruction verifies both maps. Dense training maps remain outside this scoring package.

The common scorer accepts paired probability arrays for T1–T5 and supports `-split valid` and `-split test`, with test as the default. It checks exact ID order, frozen data hashes, probability shape and range, and multiclass normalization. Task performance uses the full valid domain; paired correctness and flipping use the mapped domain. Validation scoring permits users to assess paired behavior before choosing a model. The reported experiments select checkpoints using task performance on the training-reference validation inputs.

Real predictions from the Y-trained seed14 HyenaDNA T1 head and GROVER T2–T5 heads agree with independent observation-level state counts and scikit-learn metrics. T1 MCC also matches the retained original metric table; those predictions form an executable scoring fixture. A perfect-label software fixture checks the validation scoring path. Adding the validation map preserved all pre-existing arrays in that file and left the other 14 task/split files byte-identical. Real T1 and T5 test fixtures retain their recorded scores.

Source records: fixed-input quickstart, data verification, scorer checks, protocol and review guide, position accounting, and fixed-input package.

I.3 Selected model replays

Sequence baseline. All 24 prepared T2–T5 input arrays were reconstructed from the byte-pinned production FASTAs and frozen coordinates and targets. The selected T5 one-hot pool1 Y-trained seed14 checkpoint then evaluated 8,000 test windows on each reference. Its probabilities matched the archived values exactly in the tested GPU environment. Scoring recovered 5,272 both-correct, 2,631 both-wrong, 52 only-Y-correct and 45 only-H-correct observations, totaling 97 flips.

The packaged CPU wrapper reproduces the same classes and state counts. Maximum probability differences are 2.682209014892578e-07 for YY and 2.980232238769531e-07 for YH; small ranking changes affect the final digits of AP. The same-host relocation check ran from a separate directory with access to the original project source, configuration, reports, production data and audit-harness source blocked. The wrapper supplies the checkpoint, expected predictions and scoring inputs, and verifies all 24 prepared-array identities. Sources: baseline entry and relocation check.

Frozen encoders. HyenaDNA-large-1m on T2 and GROVER on T5 were replayed against historical cached features and selected-head predictions. A deterministic sample selected 48 paired IDs per split: the first 16, the middle 16 starting at floor(n/2)−8, and the last 16 in lexicographically sorted accepted IDs. This gives 144 pairs and 288 reference-specific sequences per model across training, validation and test. Sample selection preceded inspection of features and predictions, and the replay used the recorded weights and checkpoints.

The execution snapshot contains 174 content-hashed source files, model configurations and environment records. Checks cover model/tokenizer/configuration identities, oriented sequence hashes, offsets, valid masks, token features and base projections. Feature tolerance is `np.allclose(atol=1e-4,rtol=1e-4)`. Fresh test features pass through both readouts, all three seeds and both training-reference heads: 24 settings per model on 48 windows per input. Probability tolerance is a maximum absolute difference of 1e-5, with class changes recorded separately.

Model / task	Sequence inputs	Maximum token-feature difference	Maximum test-probability difference	Changed predicted classes
HyenaDNA / T2	288	0	0	0
GROVER / T5	288	1.00136×10^{-5}	0	0

All compared offsets and projected masks match. The 12 cache manifests match the archived preflight, and token and projected features pass their declared tolerances. The replay records hashes for 284 touched feature shards. The Hyena weight SHA-256 is `5ecc8a121feb9906ff3c11e671b8315156f0e595412ae8e5680781b0a0648104`; the GROVER weight SHA-256

is `81fa4ee056e65ff1c06c32512589f3ba3bcb9d054960f35f634b9cf6da7aadda`. The recorded environment uses PyTorch 2.7.1+cu126 and Transformers 4.33.3. These comparisons establish numerical compatibility on the predefined samples. Full file bindings and per-input comparisons accompany the execution evidence and model-instance files.

The feature-loading audit additionally binds 120 manifests to the original preflight and retains 160 loading records. Its independent sampled token-to-base projection checks report zero maximum error and zero missing projected positions (Appendix C.1). The binding audit records these checks separately from sequence-to-encoder replay.

I.4 Numerical reconstruction and verification

Primary comparisons. The T2–T5 frozen-encoder audit covers all 480 prediction sets and all 240 selected heads: aligned sample identities, labels, masks, prediction/checkpoint hashes, selected epochs and class weights agree with the frozen records. Per-seed flip rates reconstruct every primary mean contrast; independent BH calculations reproduce the adjusted values at both block widths. Tables E1–E4 expose all 40 settings, and the table validation records their source hashes and count checks.

The T1 evidence entry contains all performance and correctness tables, selection histories, checkpoint/prediction bindings and compact probabilities. Checks cover 72 selection histories and 144 prediction/checkpoint bindings. An independent confusion-matrix implementation reproduces all 144 MCC values exactly, and independent BH calculations agree for all four width–family combinations. T1 one-hot results are reported in Appendices E.1 and E.3, with the training protocol in Appendix C.4; Appendix H.1 covers T2–T5 one-hot controls.

Correctness and probability diagnostics. The classwise evidence entry binds all 156 T2 prediction files to fixed IDs, labels, masks and checkpoints. Reconstructed full-window MCC agrees with every recorded value to 1e-12. Joint-histogram marginals equal both mapped-domain confusion matrices. Binary state counts and transitions reconcile exactly; GROVER per-seed net changes also agree with the stored threshold analysis. The diagnostic records retain compatibility and mapping-level checks, while the author-side probability archive retains all 48 T2 arrays used in the prior numerical AP reconstruction. Those score arrays are not included in the compact candidate.

Figures and summaries. Retained numerical checks reconstruct the 50 aggregate direction comparisons, performance and flip ranges, Figure 2 mean contrasts, Figure 3 scatter coordinates and GROVER correctness counts. The claim-to-file index locates each source. Correspondence contrasts and intervals, together with the sensitivity protocol, regulatory protocol, subset results, regulatory results, all-model summaries and registry summaries, retain the diagnostic calculations.

The primary analyses use seeds 14, 15 and 16. Block-bootstrap intervals quantify genomic sampling uncertainty conditional on these fitted models. Reconstruction from stored probabilities, sufficient statistics or bootstrap replicates checks the reported calculations; selected execution checks are specified separately in I.3.

I.5 Source completeness and delivery status

The production YAO sequence is assigned to T2T-YAO v2.0 maternal by the authors; its local file identity is recorded in Appendix A.3. The direct public GFF3 matches the production annotation byte for byte (Appendix A.1). Remaining upstream records are the public FASTA checksum binding, the RefSeq source checksum and complete annotation-transfer history, including the cCRE processing chain. The recovered mapping documentation also omits some numerical assignment thresholds. Fixed-input evaluation uses the supplied sequences, labels and maps; reconstruction from upstream files depends on these additional records.

Historical encoder/tokenizer revision fields and some execution-environment fields remain incomplete. Per-shard hashes for the sampled caches were recorded during replay; historical per-shard hashes are unavailable. Sequence-to-encoder replay covers HyenaDNA T2 and GROVER T5 at the inputs specified in I.3. The other three encoders have model-file identities and feature-loading records. The compact package excludes frozen-encoder weights but includes the selected T5 one-

hot baseline checkpoint. Recorded compute covers the training/validation timers listed in Appendix C.3; a complete extraction, inference and scoring cost profile is unavailable.

The earlier two-archive local release candidate and its packaged claim index remain bound to an earlier manuscript. A separate compact local candidate now binds its manifest to the present PDF and has passed the core numerical/data/map checks plus the small compatibility check; the optional CPU baseline was not rerun for this candidate. The authors have confirmed anonymous-review distribution of the team-generated components, with source-specific terms recorded in the package notice. Final delivery still requires anonymous hosting and an external download-and-run test. The reproduction guide maps the working version to the compact package. The PDF does not embed these files, and the local checks do not establish reviewer access.

Upon publication, we will make the complete study-generated benchmark data and analysis materials publicly available, including paired windows and labels, position maps, scoring code and source files for the reported results. The underlying T2T-YAO v2.0 assembly and annotation are already available through the public T2T-YAO resources (He et al., 2023; Chu et al., 2026); the GRCh38.p14 FASTA and GFF are available from NCBI RefSeq assembly GCF_000001405.40. The public CrossRefEval release will link to these upstream resources and provide stable access to the derived materials.

J Earlier task-specific experiments

J.1 Earlier task-performance panel and T1 global-pooling control

The earlier T1 protocol uses a batch size of 128, weighted cross-entropy, learning rate 0.001, weight decay 0.0001 and patience five, selecting the first best validation MCC. Most representations have a 40-epoch ceiling; GROVER reuses early-stopped results under the earlier 20-epoch ceiling. Earlier T2/T4/T5 results retain their actual recorded settings rather than inheriting the new protocol. Among 240 earlier heads, 236 manifests record the monitor explicitly; four older manifests omit seed and monitor, with identity linked by validated prediction records and paths. The missing fields are retained as unrecorded.

The first table uses the earlier task protocols: T1 segmented pooling and T3 use three seeds 14/15/16; T2/T4/T5 use their full eight seeds 12–19. The second table is the T1 global-pooling control. These tables summarize 684 earlier prediction sets under their recorded task-specific budgets. Metric provenance: AUPRC below is the historical source label. Metric implementation was verified for the common-protocol experiment and one-hot controls (Appendix D.1). The earlier runs retain their recorded AUPRC label because their metric implementation history is incomplete.

Task (metric)	Model	YY	YH	HY	HH
T1 (MCC)	DNABERT-2	0.8662 ± 0.0029	0.8649 ± 0.0030	0.8636 ± 0.0076	0.8622 ± 0.0051
T1 (MCC)	GROVER	0.8862 ± 0.0064	0.8863 ± 0.0075	0.8794 ± 0.0032	0.8782 ± 0.0029
T1 (MCC)	HyenaDNA	0.9056 ± 0.0083	0.9060 ± 0.0085	0.9135 ± 0.0068	0.9123 ± 0.0061
T1 (MCC)	Caduceus-PH	0.9171 ± 0.0029	0.9169 ± 0.0037	0.9155 ± 0.0013	0.9159 ± 0.0015
T1 (MCC)	NTv3-100M-pre	0.9240 ± 0.0058	0.9234 ± 0.0051	0.9257 ± 0.0016	0.9254 ± 0.0019
T1 (MCC)	one-hot	0.8967 ± 0.0029	0.8965 ± 0.0028	0.8982 ± 0.0040	0.8985 ± 0.0032
T2 (MCC)	DNABERT-2	0.1352 ± 0.0062	0.1353 ± 0.0062	0.1375 ± 0.0083	0.1376 ± 0.0083
T2 (MCC)	GROVER	0.1627 ± 0.0076	0.1628 ± 0.0076	0.1612 ± 0.0112	0.1612 ± 0.0112
T2 (MCC)	HyenaDNA	0.1325 ± 0.0035	0.1325 ± 0.0035	0.1309 ± 0.0037	0.1309 ± 0.0037
T2 (MCC)	Caduceus-PH	0.1630 ± 0.0068	0.1631 ± 0.0068	0.1620 ± 0.0052	0.1621 ± 0.0053
T2 (MCC)	NTv3-100M-pre	0.1899 ± 0.0086	0.1900 ± 0.0087	0.1866 ± 0.0088	0.1867 ± 0.0088
T3 (AUPRC)	DNABERT-2	0.7759 ± 0.0057	0.7751 ± 0.0049	0.7703 ± 0.0040	0.7699 ± 0.0040
T3 (AUPRC)	GROVER	0.7920 ± 0.0025	0.7910 ± 0.0023	0.7929 ± 0.0035	0.7923 ± 0.0036
T3 (AUPRC)	HyenaDNA	0.7699 ± 0.0074	0.7682 ± 0.0075	0.7648 ± 0.0083	0.7633 ± 0.0086
T3 (AUPRC)	Caduceus-PH	0.7839 ± 0.0059	0.7836 ± 0.0059	0.7873 ± 0.0041	0.7872 ± 0.0042
T3 (AUPRC)	NTv3-100M-pre	0.8029 ± 0.0042	0.8026 ± 0.0039	0.8035 ± 0.0027	0.8031 ± 0.0020
T4 (AUPRC)	DNABERT-2	0.8890 ± 0.0027	0.8895 ± 0.0026	0.8874 ± 0.0019	0.8878 ± 0.0018
T4 (AUPRC)	GROVER	0.8787 ± 0.0038	0.8781 ± 0.0042	0.8796 ± 0.0025	0.8791 ± 0.0024

Task (metric)	Model	YY	YH	HY	HH
T4 (AUPRC)	HyenaDNA	0.8670 ± 0.0037	0.8655 ± 0.0037	0.8668 ± 0.0038	0.8655 ± 0.0038
T4 (AUPRC)	Caduceus-PH	0.8744 ± 0.0031	0.8747 ± 0.0032	0.8743 ± 0.0036	0.8746 ± 0.0037
T4 (AUPRC)	NTv3-100M-pre	0.8564 ± 0.0035	0.8561 ± 0.0037	0.8580 ± 0.0055	0.8574 ± 0.0053
T5 (AUPRC)	DNABERT-2	0.7758 ± 0.0014	0.7757 ± 0.0015	0.7743 ± 0.0028	0.7741 ± 0.0029
T5 (AUPRC)	GROVER	0.7778 ± 0.0031	0.7785 ± 0.0029	0.7775 ± 0.0040	0.7782 ± 0.0043
T5 (AUPRC)	HyenaDNA	0.7411 ± 0.0156	0.7410 ± 0.0156	0.7482 ± 0.0117	0.7479 ± 0.0116
T5 (AUPRC)	Caduceus-PH	0.7680 ± 0.0051	0.7682 ± 0.0050	0.7697 ± 0.0037	0.7700 ± 0.0038
T5 (AUPRC)	NTv3-100M-pre	0.7325 ± 0.0033	0.7310 ± 0.0036	0.7321 ± 0.0035	0.7310 ± 0.0031

T1 representation	YY	YH	HY	HH
DNABERT-2	0.4316 ± 0.0127	0.4325 ± 0.0152	0.4317 ± 0.0270	0.4315 ± 0.0266
GROVER	0.7256 ± 0.0085	0.7264 ± 0.0079	0.7363 ± 0.0068	0.7369 ± 0.0044
HyenaDNA	0.7583 ± 0.0151	0.7576 ± 0.0140	0.6951 ± 0.0294	0.6949 ± 0.0293
Caduceus-PH	0.5107 ± 0.0066	0.5109 ± 0.0036	0.4960 ± 0.0191	0.4945 ± 0.0208
NTv3-100M-pre	0.6242 ± 0.0071	0.6243 ± 0.0070	0.6319 ± 0.0029	0.6315 ± 0.0039
one-hot	0.2915 ± 0.0069	0.2890 ± 0.0078	0.2874 ± 0.0024	0.2837 ± 0.0009

J.2 Existing T1 supervised baselines

Baseline	YY	YH	HY	HH
k-mer TF-IDF + RF	0.2453 ± 0.0026	0.2452 ± 0.0042	0.2461 ± 0.0047	0.2464 ± 0.0051
one-hot CNN, pool1	0.2915 ± 0.0069	0.2890 ± 0.0078	0.2874 ± 0.0024	0.2837 ± 0.0009
one-hot CNN, pool16	0.8967 ± 0.0029	0.8965 ± 0.0028	0.8982 ± 0.0040	0.8985 ± 0.0032

These historical one-hot rows reproduce the controls in Appendix J.1. The primary common-protocol T1 controls appear in Appendix E. The k-mer row reports 12 existing prediction sets. All use repaired original T1, seeds 14/15/16 and 8,000 paired test windows. The k-mer protocol fits overlapping oriented TF-IDF features and 256-tree random forests. Candidate k values are 3–6; each reference selects k using mean validation MCC across three seeds, choosing k=3 on both references. Vocabulary and IDF use training data only. Selection is frozen before testing, without re-fitting on combined training and validation data. Its feature construction and model-selection budget follow the k-mer protocol above.

J.3 Recorded representation metadata

Model	Dimension	Layer	Tokenization	Model/tokenizer revision
caduceus_ph	256	-1	single nucleotide	None / None
dnabert2	768	-1	non-overlapping BPE	None / None
grover	768	-1	non-overlapping BPE	None / None
hyenadna_large_1m	256	-1	single nucleotide	None / None
ntv3_100m_pre	768	-1	single nucleotide	None / None

The table summarizes 90 recorded embedding manifests for earlier T2/T4/T5 runs. "None" indicates that the field was not recorded. These manifest fields apply to the earlier T2/T4/T5 caches; Appendix C documents the current input and model bindings.

The author identified the upstream gene annotation as RefSeq GCF_000001405.40_GRCh38.p14_genomic.gff; the source checksum and full transfer history remain unresolved. The actual GRCh38 stratification inputs match GIAB v3.1 release files from the genomic stratifications resource (Dwarshuis et al., 2024). Appendix J.4.1 uses these strata for supplementary regional comparisons.

J.4 REGIONAL COMPARISON DESIGN

The regional analysis uses segmental duplications, low-mappability regions, nominal tandem repeats longer than 200 bp, and their three-way union. Membership is defined by window centers; coverage of at least 50% is a sensitivity analysis. The tandem-repeat group combines the recorded 201–10,000 bp and >10,000 bp strata, separately from older >100 bp analyses. One-to-one matching uses 5-percentage-point GC bins on each reference, sequence-change classes (identical, substitution-only, or indel), nonidentity bins (≤1%, 1–5%, 5–10%, 10–20%, and >20%), source chromosome, and task-specific label summaries. For T1 and T3–T5, the label summary is the window class. For T2, it is the ordered pair of donor and acceptor counts across the full YAO label window, encoded as `donor_count:acceptor_count`. Counts are taken before restriction to the mapped scoring domain; the signature contains no positional arrangement or model predictions. Within each joint cell, the smaller group determines the matched count, with fixed salted sample-ID hashes selecting the pairs without reading predictions. The main outcome is window-equal flipping in group 11 minus group 00. Inference requires at least 30 windows and 10 supported blocks per side and at least 4,950 valid replicates; unsupported settings remain in the correction family with p=1. These thresholds define the minimum support for computing a contrast. Residual imbalance is flagged at absolute standardized mean difference >0.1. Full settings are in the regional protocol.

At each block width, BH correction is applied to the 200 post-matching, window-center regional flip-rate contrasts, separately from the common-protocol readout families. Unmatched contrasts are not included in this primary correction family. The regional analysis uses the earlier task-specific five-task head panel (Appendix J.1) with its recorded readouts. Matched-set intervals quantify sampling uncertainty conditional on the fixed matched sets; regional comparisons estimate adjusted associations.

J.4.1 REGIONAL CONTROLS

Regional analyses compare membership on both references (11) with membership on neither (00). Matching uses the task-specific label summaries defined above, both GC bins, sequence-change type, nonidentity bins, and chromosome. Deterministic identity hashes define shared one-to-one matched sets. The 200 post-matching contrasts cover four strata, five tasks, five models, and two training references, with a separate BH correction at each block width. They use the earlier task-specific five-task head panel in Appendix J.1. All regional means use seeds 14/15/16. Internal task identifiers A2/A3/A4/A7/A8 correspond to T1/T2/T3/T4/T5; the primary T1 regional analysis uses its earlier 16-bin readout. The global-pooling T1 and one-hot comparisons remain supplementary to this family.

Table J1. Composition-matched regional comparisons for the three-stratum union. Each range spans five frozen models and two fixed training references, averaging three seeds within each setting. Differences are flipping in group 11 minus group 00, in percentage points. Ranges span the ten setting-specific means. Corrected counts require $q < 0.05$ at both block widths and are assessed in the full 200-contrast post-matching family.

Task	Matched windows per group	Retained fraction of group 11	Unadjusted difference range (pp)	Matched difference range (pp)	Corrected contrasts / 10
T1	578	81.2%	+0.307 to +1.202	−0.058 to +0.519	0 / 10
T2	152	65.2%	+0.517 to +1.403	+0.060 to +0.160	0 / 10
T3	610	78.3%	+0.323 to +1.007	−0.765 to +0.219	0 / 10
T4	616	73.7%	+0.302 to +1.618	−0.649 to +0.433	0 / 10
T5	665	72.7%	+0.136 to +1.793	−0.752 to +0.451	0 / 10

After composition matching, none of the 200 regional contrasts survives correction at either block width; minimum $q = 0.608$ at 1 Mb and $q = 0.619$ at 2 Mb. The union comparisons meet the recorded support criteria, with maximum absolute standardized mean differences of 0.035–0.078. The separate tandem-repeat comparisons retain nonidentity imbalance in T1 and T2 (absolute standardized mean differences 0.113 and 0.174). Matching retains 65.2%–81.2% of the original jointly annotated union group, so inference concerns that common-support subset. The matched contrast

measures the residual regional association after conditioning on the specified sequence differences and other covariates.

The task summary, complete primary contrasts, matched identities, and support table provide the numerical basis for Table J1. These regional analyses use the existing training runs and predictions.

The union group has higher mean flipping than the jointly unannotated group before matching in all 50 task–model–training-reference settings. For T4, the differences range from +0.302 to +1.618 percentage points before matching and from -0.649 to +0.433 afterward. The table reports the corresponding ranges for all five tasks. Pairs annotated on only one reference (01 or 10) remain in the benchmark but are outside this 11-versus-00 comparison. Regional annotations describe genomic context; mapping levels describe the recorded correspondence procedure.

J.5 Supplementary task-specific panel

In the supplementary task-specific panel (Appendix J.5; Figure S7), net accuracy difference and correctness discordance pool valid mapped position observations across T2 window pairs within each fitted model, then average rates over seeds. Each base contributes once per containing window. Classification tasks use window pairs. Supplementary diagnostics retain position-equal and window-equal correctness rates separately; primary T2 readout flipping remains window-equal.

The five-task panel compares each fixed head on paired YAO and GRCh38 inputs. Across all 50 task–model–training-reference settings, mean net accuracy differences had magnitudes below 0.12 percentage points, while correctness discordance ranged from 0.228% to 1.311% (Figure S7). Gains and losses in correctness therefore largely offset one another in aggregate accuracy. Correctness discordance counts observations with exactly one correct prediction; in the multiclass splice tasks, label flipping can also include two incorrect predictions that disagree. This panel uses seeds 14/15/16 under the earlier task-specific protocols in Appendix J.1. Appendix J.1 reports task performance and all four reference combinations, retaining the seed sets of each historical performance table. The primary common-protocol readout comparisons are reported in Appendix E.

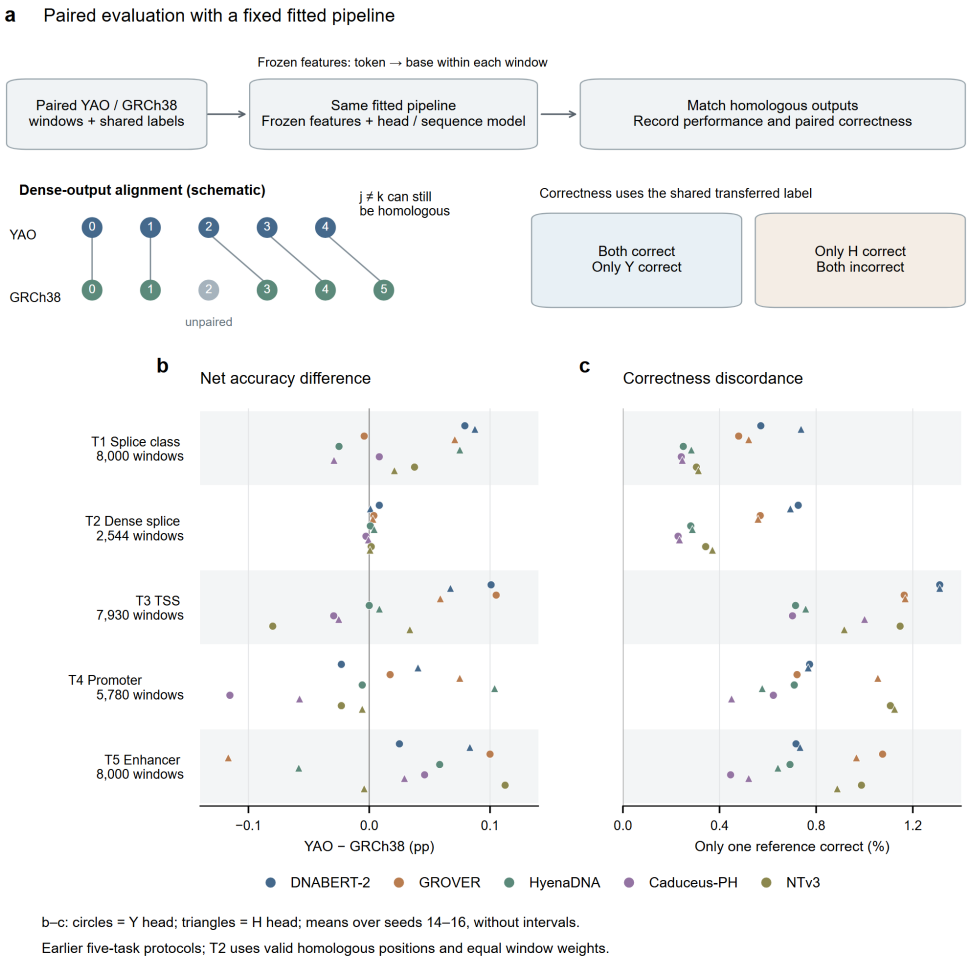

Figure S7: Paired evaluation and correctness across five tasks. (a) Keep the fitted pipeline fixed, match outputs by the recorded correspondence and score against shared benchmark labels. The schematic distinguishes within-window feature projection from between-reference output mapping; an insertion can shift corresponding output indices. (b) Net accuracy difference, YAO minus GRCh38, in percentage points. (c) Correctness discordance, the percentage of observations with exactly one correct prediction. Panels (b–c) show all 50 task–model–training-reference settings, averaging seeds 14/15/16. Colors identify models; circles and triangles denote YAO and GRCh38 training. Task labels give paired test-window counts. For T2, both rates pool valid mapped position observations across window pairs, weighting each observation equally. Other tasks use one observation per window pair. Comparisons follow the earlier task-specific protocols in Appendix J.1 and summarize the three fitted seeds.