

Copy and relatedness leaks in genomic language model benchmarks are common and can be found and largely removed automatically

Tejasvi Singh Tomar

ORCID: <https://orcid.org/0000-0003-4668-3639>

Abstract

Genomic language model benchmarks rank models, but on some tasks a no-model rule, such as a nearest-training lookup that copies the label of the most similar training item, matches or beats the best published score. Such rules exploit leaks: test items that are copies or relatives of training items, negatives constructed from the positives, and genomic positions that reveal the label. Copies and relatives occur in 8 of the 11 DNA and RNA benchmarks published from 2024 to 2026 that we audited, although on most tasks no leak rule comes close to the published scores. The frozen rules of our linter, `seqleak`, recover all 24 copy and relatedness flags of that audit on 41 tasks not used to set them. In pre-registered fine-tuning experiments on GUE’s yeast histone-mark tasks, however, relatives in training raise test scores only slightly on average. Constructed negatives are rarer, mostly draw only a warning from the frozen rules and survive re-splitting. Yet each of six models fine-tuned on GUE’s constructed negatives instead of matched genome windows scores lower on matched genomic negatives, by 0.15 and 0.41 AUROC on average for shuffled and segment-substituted negatives, net of the same comparison on control tasks. On independent G+C-matched test sets the losses are 0.14 and 0.31, whereas an earlier independent test without G+C matching did not meet its registered criterion. About two thirds of the improvement reported on GUE and the original Nucleotide Transformer tasks since their release does not exceed a no-model leak rule’s score, so published scores alone cannot tell it apart from leak exploitation; on Genomic Benchmarks, most of it lies above. We release the linter, leak-reduced versions of 60 tasks, of which the re-split ones remove most of the lookup’s signal, and PairCtrl, a leaderboard that controls for genomic position.

Keywords: genomic language models; benchmarks; data leakage; homology; negative sampling; pre-registration

1 Introduction

Genomic language models are compared on suites of sequence tasks: promoters, enhancers, splice sites, chromatin marks, transcription-factor binding, variant effects and gene pairs [1–7]. A high score is read as evidence that a model has learned the biology the task asks about. Some tasks can be won, or nearly won, another way. In DART-Eval task 1 [8], each positive is paired with a dinucleotide shuffle of itself, and an order-10 Markov rule ranks the positive above its shuffle in 0.917 of pairs, compared with 0.976 for the best published model. In an earlier audit we found tasks in widely used DNA benchmarks on which a no-model rule, such as a lookup that predicts the label of the nearest training item, matches or beats every published score [9]. The rules exploit three classes of leak: relatedness between training and test items, negatives constructed from positives, and genomic position, such as the distance to a transcription start site. That audit was done by hand and did not cover most benchmarks published from 2024 to 2026; it also left open whether the leaks change what benchmarks say about models.

Here we ask whether such leaks persist in these newer benchmarks, whether they can be found and removed automatically, and whether they alter the conclusions drawn about models. Section 2

Table 1: Census verdicts per benchmark (the earlier audit’s strict comparison; Appendix B); DNA benchmarks first, then RNA. For PGB and LRB, D3 counts defects, with mechanical D3 verdicts in brackets; the 5 mechanical D3 verdicts of PGB at the labelled site count as biological signal. Other: no comparable published score, relatedness by design, or not audited.

Benchmark	Tasks	D1	D2	D3	C	N	Other
Plant Genomic Benchmark [17]	31	0	0	1 (6)	0	25	0
LRB, tasks not audited before [6]	8	0	0	1 (1)	0	7	0
GENEB, classic tasks [18]	17	4	1	4	5	3	0
GENERator tasks, eukaryote [19]	2	0	0	0	1	1	0
GENERanno tasks, prokaryote [20]	4	0	0	1	0	3	0
DART-Eval, tasks 1 and 3–5 [8]	19	0	0	1	0	18	0
TraitGym [21]	2	0	0	1*	0	1	0
PlantCaduceus, cross-species [22]	4	0	0	0	0	4	0
NTv3, genome annotation [23]	13	0	0	0	0	12	1
Tang and Koo [24]	22	0	0	0	0	0	22
BEACON [25]	13	0	0	2	0	6	5
RGB [26]	7	0	0	0	0	1	6
mRNABench [27]	59	2	0	11	0	46	0

audits 11 benchmarks published from 2024 to 2026 by hand and finds copies or relatives in 8 of them. Section 3 introduces seqleak, a linter whose rules, frozen before testing, recover all relatedness and copy flags of that audit on 41 held-out tasks and fail 1 of the 13 tasks it did not flag, but mostly only warn on construction and motif flags. Section 4 releases leak-reduced drop-in versions of 60 tasks of GUE, the original Nucleotide Transformer tasks and Genomic Benchmarks. Section 5 tests what the leaks do to models, in pre-registered fine-tuning experiments and a pre-registered analysis of published scores. Section 6 presents PairCtrl, a leaderboard of four pair tasks that controls for genomic position, and two pre-registered attempts at position-controlled training, neither of which, on labels without gene layout or transferred evidence, outperforms an untrained cosine of the embeddings it starts from.

Related work. Related training and test items inflate scores across biological machine learning: sequence-based predictors of protein interactions learn node degree and sequence similarity under random splits [10], homology between chromosomes leaks into genome-trained sequence models even under a chromosome hold-out [11], and random splits of enhancer–promoter pairs inflate scores when pairs share a promoter [12, 13]. For DNA benchmarks, genomic-benchmarks-qc [14] flags features that separate the classes on their own and test sequences similar to training sequences, and Civale et al. find normalised k -mer frequencies competitive with the frozen embeddings of most of five DNA foundation models on 57 data sets without checking overlap between splits [15, 16]. Section 3 compares our linter with genomic-benchmarks-qc, and Appendix A gives details and relates this work to our earlier audit.

2 A census of benchmarks

We applied the earlier audit’s verdict rules and thresholds unchanged, with benchmark-specific rule families added (two of them post hoc; Appendix B), to genomic language model benchmarks published from 2024 to 2026 and to benchmarks it had not covered in full: the Plant Genomic Benchmark (PGB) [17], the eight tasks of the Genomics Long-Range Benchmark (LRB) [6] that the earlier audit had left out and BEND’s remaining defect classes [5]. We score each rule or lookup on the benchmark’s own split and metric against the best published score we found. A task is D1 when a rule tied to a detected leak exceeds every published score, D2 when it beats the best model of the benchmark’s own paper, and D3 when it reaches 0.75 of the best published score above chance. It is C when only a composition rule does so, which reflects biology rather than a leak, and N otherwise (Table 1). For 9 of the 13 benchmarks we first reproduced one published number (Appendix Table 3).

Benchmarks published from 2024 to 2026. Of the 26 benchmarks listed, we audited 11: eight DNA benchmarks and three RNA-sequence benchmarks (BEACON [25], RGB [26] and mRNABench [27]). Six task rows in three of them were not audited, and a twelfth benchmark was audited in part

(Appendix B). Most of the 162 task rows show no defect: 95 are assigned verdict N and 6 verdict C, which marks composition and thus a biological signal. A leak rule assigns a defect verdict to 27 rows (Table 1), mostly D3; the verdict for TraitGym (marked *) is caveated. For four of the six D1 verdicts, the only comparable published scores come from frozen-embedding probes, and for two the per-seed interval includes zero. Relatedness recurs across benchmarks: it is flagged in 65 rows across 8 of the 11 benchmarks and contributes to 17 defect verdicts in 4 benchmarks. mRNABench accounts for 45 of these rows and 13 of these verdicts. Construction defects, by contrast, are rarer; DART-Eval task 1 of the introduction is one. No defect appears in NTv3’s 12 genome-annotation tasks [23], in PlantCaduceus’ cross-species tasks [22] or in DART-Eval tasks 3–5.

PGB, LRB and BEND. On PGB, one of 31 tasks has a counted defect: in the *Arabidopsis* splice-donor task, 37.4% of test negatives lack the central GT, so a fixed-position motif table reaches 0.963 AUROC (D3). In LRB, the promoter task’s positives are defined by their distance to an annotated TSS, and TSS distance alone reaches AUPRC 0.811, compared with the best published 0.890 (D3); in BEND’s remaining tasks, TSS distance stays at AUROC 0.560–0.630 (Appendix B).

Protein and RNA tasks. A census of 204 public protein and RNA tasks flags relatedness in 148 (0.725; 95% interval 0.661–0.782, or 0.512–0.872 when source families are resampled) and gives a defect verdict to 8 of the 42 tasks with a comparable published score (one of them a family-level test that shares families by design). Of the splits matched to documentation of a homology-aware split or of whole families or folds held out, only 9 of 82 are consistent with it under the frozen matching rule, and 8 of 67 under its post hoc correction (Appendix Table 4)¹.

3 A leak linter for genomic benchmark tasks

Design. The command `seqleak audit` reads a task’s splits, optionally with genomic coordinates and distance fields, and runs 17 checks, 16 in five families and a reference lookup (Appendix Table 5). *Overlap* checks measure exact, reverse-complement and near copies, overlapping or shifted windows, copied pairs and BLASTN homologs between the test split and the splits a model can be fitted on, and a *reference* check finds copies of another benchmark’s items. *Shortcut* checks fit one-feature rules, such as length, composition, an order-3 Markov classifier and fixed-position motifs, on the training split and score them on the test split. *Construction* checks find the composition twins of dinucleotide shuffles and the positional twins of segment substitution, and the *position* check fits `seqleak`’s position-only model to distance fields. Each check returns PASS, WARN, FAIL or SKIP, and a task takes its worst verdict. FAIL marks a leak at the level that the earlier manual audit called a defect; WARN marks weaker signal of the same kind, or a composition baseline, which never fails. The companion command `seqleak clean` writes a filtered or re-split version that cuts every relation the overlap and twin checks found. The tool `genomic-benchmarks-qc` [14] shares the length, composition, position-specific sequence and cross-split duplicate checks; `seqleak audit` adds checks for constructed negatives, overlapping genome windows, pair items, copies of another benchmark and distance confounds, sets its FAIL levels at a manual audit’s defect levels and is validated against manual audits with rules frozen before a held-out test.

Rule history. We registered the checks and thresholds before the linter ran on any audited task. However, we set the thresholds after reading the earlier audit’s tables, so its 80 tasks in scope are an in-sample check of the implementation. Two post hoc amendments (Appendix C) then moved the FAIL levels of copy shares to the manual audit’s operating points, let twins fail only for constructed negatives, because the registered rules had failed 21 of 27 tasks without a documented leak, and added an AUROC criterion for shortcut rules when the minority class is below 20%. We then froze the rules as v1.0; every later v1.0 number uses them unchanged.

¹The frozen rule also applied a family’s documented split to third-party re-splits with other test sets; the correction (amendment C6) is a disclosed post hoc fix, made after partial results were read. The relatedness flags and all verdicts are the same under both rules, and verdicts exist only for the 42 tasks whose published scores could be parsed. Ten tasks first left out for their size were audited later with the same frozen tool (amendment C7).

Validation. On the 80 tasks of the earlier audit, every rule version recovers all 87 documented (task, class) pairs at FAIL. The frozen v1.0 rules fail all 48 tasks with a demonstrated defect and none of the 27 without a documented leak. A synthetic suite fixed before any run met all 32 expectations under each rule version, and no control drew a leak FAIL; it comprised 9 planted leaks at 4 doses and 3 controls with 20 seeds each. Appendix C gives the validation by rule version and checks the first amendment out of sample on PGB.

After the freeze, we scored two task sets with the rules unchanged (Figure 1). On the 42 tasks of newer benchmarks (Section 2), v1.0 recovers 28 of 39 manual (task, class) pairs at FAIL and 38 of 39 at FAIL or WARN, against the manual flags as they stood before the comparison (Appendix C). Relatedness and copy flags transfer: all 26 such pairs fail. Only 8 of them lie on tasks with a defect verdict. One of the 42 tasks is a re-hosted copy of an in-sample GUE task. Without it, recall is 26 of 36 at FAIL and 35 of 36 at FAIL or WARN, and all 24 relatedness and copy flags still fail. Construction and motif classes mostly stay at WARN (10 of 13 such pairs), mainly because the motif FAIL bar is absolute (MCC 0.7), unlike the manual audit’s comparison with the best published score; twins that never cross the split (DART-Eval task 1) are a WARN by design. Of the 13 tasks that the manual audit did not flag, one fails: a rice splice-donor task whose motif table captures the splice consensus (AUROC 0.853 with 16.7% positives). On the second set of 6 LRB tasks, the linter misses the one defect counted in Table 1, a TSS-distance shortcut that the linter cannot detect without distance fields.

At scale. Under a pre-registered protocol, we then ran the frozen rules on every genomic sequence data set with a published train and test split that we found on Hugging Face and GitHub beyond the audited benchmarks: 790 tasks from 92 data sets (22 tasks excluded by scope). In 58 of the 92 data sets (0.630 [0.528, 0.722]) at least one task fails, and 134 of the 790 tasks fail. One repository contributes 507 similar tasks but only 5 failures; without it, 129 of 283 tasks fail (0.456 [0.399, 0.514]). Copies and relatedness dominate (an overlap FAIL on 117 tasks), followed by fixed-position motifs (14) and artefacts (10), whereas constructed negatives inside a split reach only WARN. Checks of the raw files confirm the clearest cases. Three pre-registered controls and a separate recount of exact copies passed, whereas a registered external check met only one of three expectations (Appendix D).

A second held-out test. Version v1.1 adds four checks for construction defects; we froze it on the in-sample sets before a fresh sample of the census pool, one task per data set, was audited blind by hand (58 tasks with 40 task–class pairs; 3 data sets with human variant-pathogenicity labels were outside the auditor’s scope). Both versions find 32 of the 40 pairs at FAIL and 36 of the 40 at FAIL or WARN, and both fail 4 of the 32 tasks that the audit did not flag. The new checks fail no task; on a secondary sample, mostly from one repository, v1.1 finds 2 more pairs. The misses are constructed negatives that are detectable only through genome membership, motifs anchored in sequences of variable length, shuffled negatives only in the test split and borderline relatedness (Appendix C).

4 Leak-reduced drop-in versions

We built leak-reduced versions of 60 tasks of GUE (with the GUE+ enhancer–promoter tasks) [1], of the first version of the Nucleotide Transformer tasks [2] and of Genomic Benchmarks [4], 309 versions in total, plus two DNALongBench drop-ins [7] and cleaned versions of 17 failing tasks from the census of public data sets (Section 3; 34 versions). Each version keeps its suite’s format and can be rebuilt from the original download (Appendix E). The default drop-in, *resplit*, holds out chromosomes chosen by a label-free rule and cuts every leak relation that crosses splits, with the original split sizes and class balance. Other versions also cut the linter’s relations (*resplit_union*), replace constructed negatives by matched genome windows (*fixneg*, which changes the task), do both (*fixneg_resplit*), keep the original test set (*purged*) or re-split on other chromosomes (*placebo*). Every version passes through the negative control and the frozen linter, and each of the 60 tasks through the positive control (Appendix Table 9); census versions, which lack coordinates, assign each group of related items as a whole to one split instead of holding out chromosomes. The negative control (C1) re-runs the audit’s leak rules, and the version passes when their scores are within 0.05 of chance, or

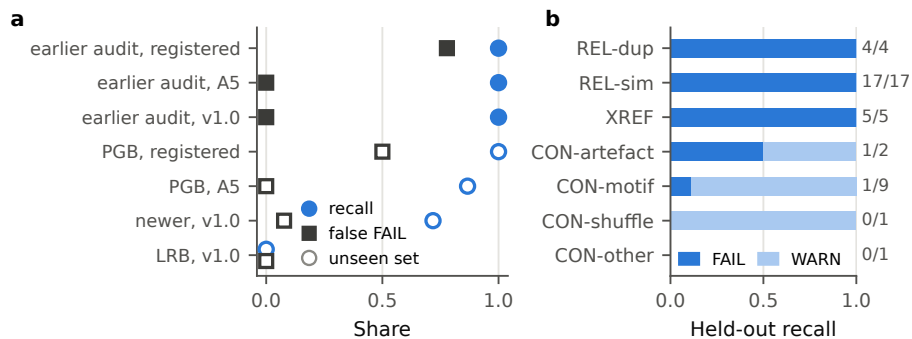


Figure 1: Validation of the leak linter. (a) Recall of documented leak classes at FAIL and false-FAIL rate on tasks without a documented leak, by task set and rule version (earlier audit: 80 tasks; PGB: 31 tasks, relatedness classes; newer benchmarks: 42 tasks; LRB: 6 tasks, recall at the mechanism level). (b) Held-out recall of the frozen v1.0 rules per manual flag class on 42 tasks of newer benchmarks, at FAIL and at FAIL or WARN. REL-dup: exact or reverse-complement copies; REL-sim: near copies, homologs and overlapping windows; XREF: copies of another benchmark; CON: construction classes (artefact, motif, shuffled negatives, other).

their interval contains chance, and no test item has a training twin. The positive control (C2) adds shifted copies of 10% of the test items to the training split, which must raise the lookup’s score.

Re-splitting removes a median 91.1% of the nearest-training lookup’s score above chance on the 46 tasks where it originally exceeds chance by more than 0.1 (Appendix Figure 3a). The positive control’s injected copies raise the lookup’s score in 57 of 60 tasks. The frozen linter separates originals from drop-ins: it fails 47 of the 60 original tasks but no resplit_union, placebo or fixneg version, and only 3 resplit versions, where it finds segment twins that the audit’s definition misses.

However, passing its checks does not make a version free of leaks. On the 12 tasks with constructed negatives (Appendix Table 8), the genome-membership rule of Section 5 still scores MCC 0.8925–1.000 on every version that keeps those negatives, including the re-split ones. With this rule added to the negative control (a post hoc amendment), no resplit_union version passes (Appendix Table 9). The fixneg versions remove the construction (genome rule MCC 0.000–0.013) but keep a documented residual of the relatedness rules in 25 of their 36 draws. Combining both repairs (*fixneg_resplit*, three draws per task) gives 36 versions that all score 0.000 on the genome rule and get no linter FAIL. A recommended version passes every check in all three draws for 7 of the 12 tasks: the combined version for six and fixneg for one (Appendix Table 8). No relative remains in the combined versions under either tool’s definitions, yet on 4 tasks they keep a nearest-training lookup score above chance (0.053–0.323); the cause is sub-threshold homology. Strict variants that also cut every BLAST hit, registered for these 4 tasks and extended post hoc to GUE prom_300_all, bring the lookup’s score to 0.000–0.050. With them, every construction task now has a version whose three draws pass every check (Appendix Table 10).

As released, 28 resplit versions fail the negative control, and the README of each documents its residual (Appendix E).

5 Do leaks change what benchmarks say about models?

A no-model rule that matches published scores leaves open whether models trained on shortcut data learn less of what transfers, and whether the leaks inflate the scores that rank models. We test both with a design registered before any test evaluation. Appendix F describes its amendments, and Appendix Table 17 lists those entered after the results were first reported.

Design. Six models from 0.4 M to 500 M parameters form the registered panel: the Genomic Benchmarks CNN [4], HyenaDNA [28], Caduceus [29], DNABERT-2 [1] and the Nucleotide Transformer v2 models with 50 M and 500 M parameters [3]. Three further models compared in the GUE paper are added to the ten-mark test of Q2 (defined below). We fine-tune all of them with the protocol

Table 2: Pre-registered tests of whether leaks change what benchmarks say about models, each with its registered criterion, result and verdict. Codes are those of the registered design (Appendix F; in F5bx and F3bx, b marks the cost statistic and x the extension from two to all five transcription-factor tasks). Frozen probes: at the registered layer.

Question and test	Registered criterion	Result	Verdict
Q1, promoter test (E4)	both metrics Holm-significant in ≥ 4 of 6 models	2 of 6; AUROC +0.043, GC-stratified +0.266	not met
Q1, replacement test	Holm-significant in ≥ 4 of 6 models per class	6 of 6 per class: +0.142 (binding sites), +0.306 (promoters); controls pass	met
Q1, GUE construction tasks (F5bx)	Holm-significant in ≥ 4 of 6 models per class	6 of 6 per class: +0.151 (shuffled), +0.408 (segment)	met
Q1, controlled constructions (F3bx)	mechanism, Holm over 12 tests	12 of 12: +0.249, +0.414; genomic control near zero	shown
Q1, frozen probes	per model and class: lower bound > 0 and Holm-significant	12 of 12: +0.192, +0.397 (genomic negatives); +0.186, +0.273 (independent sets); controls pass	met
Q2, four marks (F1)	Holm-significant in ≥ 4 of 6 models	4 of 6; panel +0.016 [+0.011, +0.022]; controls: negative passes, positive 2 of 6 (one more borderline)	met, no clear null
Q2, ten marks (F1x)	Holm-significant in ≥ 4 of 9 models; clear null if the panel’s upper bound is below 0.02	6 of 9; panel +0.013 [+0.011, +0.016]	met, clear null
Published margins	inside when below the inflation’s lower bound	GUE: 3 of 11 inside (7 at the point estimate); NT – original: 1 of 10; published order kept: Kendall’s τ 0.728 (cleaned), 0.787 (original)	–
Published wins	“most survive” if $\geq 50\%$ of reproduced wins survive	19 of 31 reproduce; with position-matched negatives 11 of 19 survive, aligned-input 6 of 6 vanish or reverse; control passes	most survive

of the DNABERT-2 GUE scripts (Appendix F), and each family of tests has its own Holm control (Table 2).

- *Q1, external validity (primary).* Do models trained on a shortcut version do worse on independent leak-free data than those trained on the cleaned version? An independent promoter test (E4) scores promoter models on revised Nucleotide Transformer rows, and two construction-cost tests, on GUE tasks (F5bx) and on controlled constructions (F3bx), measure what constructed negatives cost on genomic negatives. After the promoter test’s result, a replacement test was registered before its data existed: independent binding sites of the GUE mouse factors from other experiments and independent promoters, each negative being a genome window with the same G+C count as its positive; on the 300-bp promoter set, 97.9% of the negatives match exactly. Q1 needs a supported effect in at least 4 of 6 models.
- *Q2, inflation of the benchmark’s own scores.* On the GUE yeast histone-mark tasks, arm P removes every training row that overlaps a test row in the genome or shares an exact 50-mer with it, and arm R removes as many rows, with the same class counts, as whole loci around random unrelated anchors. Training relatives inflate scores when $S(R) - S(P) > 0$ on the original test set (a four-mark test, F1, and a ten-mark test with nine models, F1x). Two random arms and injected shifted copies serve as negative and positive controls. A family has a clear null when the upper 95% bound of its panel-mean effect is below 0.02.
- *Published margins.* For each task with comparable published scores, the margin between the two best published model families is compared with the panel-mean inflation, measured on the original test sets ($S(R) - S(P)$) and at the benchmark level (original minus re-split). A margin falls inside the inflation when it is below the inflation’s lower 95% bound.

Results. Q1 meets its registered criterion for both construction classes (Table 2). Training on the benchmark’s constructed negatives instead of matched genomic negatives lowers the AUROC on test sets with genomic negatives by 0.151 [0.136, 0.165] for dinucleotide-shuffled negatives (GUE mouse tasks) and by 0.408 [0.395, 0.422] for segment substitution (GUE promoter tasks); both drops are net of the same contrast on five transcription-factor tasks. The drop is Holm-significant for all 6 models in each class, and the control contrast lies between -0.003 and $+0.008$. By contrast, the

independent promoter test, on 13,843 rows (1,650 promoters), does not meet the registered criterion: only 2 of 6 models are supported in both metrics, with a plain AUROC difference of $+0.043$ [$+0.019$, $+0.068$]. Within GC strata, the difference is $+0.266$ [$+0.233$, $+0.299$] and significant for 6 of 6, mostly because the models trained on constructed negatives fall from 0.543–0.664 to 0.390–0.457, below chance. GC composition thus appears to mask part of the deficit. The replacement test meets the criterion for both classes: models trained on genomic negatives score higher than those trained on constructed negatives, beyond the same contrast on the five human transcription-factor tasks, by 0.142 [0.135, 0.149] on the independent binding sites and by 0.306 [0.295, 0.317] on the independent promoters, Holm-significant for all 6 models in each class. In a further pre-registered test, linear probes on frozen embeddings of three large models from Section 6 (Evo 2 7B, GenomeOcean 4B and GENERator 3B) show the same cost in every model: $+0.156$ to $+0.225$ (shuffled) and $+0.359$ to $+0.426$ (segment) on the genomic-negative test sets, and $+0.165$ to $+0.220$ and $+0.252$ to $+0.297$ on the independent sets, 12 of 12 tests Holm-significant, with both controls passing (Appendix Table 15).

Q2 asks whether training relatives inflate the benchmark’s own scores. On the four GUE yeast histone-mark tasks, the registered four-mark test meets its criterion: training relatives inflate fine-tuned scores on this benchmark. At the last checkpoint, $S(R) - S(P)$ in test MCC is Holm-significant for 4 of the 6 models, with the largest effect for DNABERT-2 ($+0.052$ [$+0.041$, $+0.063$]; Appendix Table 11). The panel mean, $+0.016$ [$+0.011$, $+0.022$], does not meet the clear-null criterion. The negative control passes. At the checkpoint chosen on development data, the effect is smaller. The ten-mark test, with nine models, also meets its criterion, with 6 of 9 models Holm-significant, but its panel mean of $+0.013$ [$+0.011$, $+0.016$] has an upper bound below 0.02 and therefore also meets the registered clear-null criterion at that margin. The inflation is real but small, except for DNABERT-2 ($+0.041$). The inflation replicates for 4 of 6 models in a secondary test on the ten marks’ original Nucleotide Transformer test sets, and its panel mean is not a clear null (Appendix F). The positive control detects the strong injected dose for Caduceus and DNABERT-2, and NT-v2 500M is borderline. Because the dose goes undetected for the CNN and HyenaDNA, their null four-mark results show only that no exploitation of training relatives is detectable (Appendix F).

By contrast, the same leaks inflate the no-model rule’s score far more: on the histone-mark tasks, original minus re-split MCC is 0.45 to 0.77 for the nearest-training lookup but -0.03 to $+0.09$ for the models, with the largest change for DNABERT-2 (descriptive). The six models rank differently on the size-matched original and the re-split histone-mark tasks in the registered ranking test, which compares the six models’ order on the two versions (Kendall’s τ 0.867, beyond replicate noise), with a significant reversal of DNABERT-2 and NT-v2 50M.

We then compare the published margins with the inflation of fine-tuned scores. The published lead of the best model family over the next falls inside the inflation measured on the original test sets on 3 of the 11 GUE tasks where it was measured (ten histone marks and splice sites), and below its point estimate on 7 of the 11. The lead also lies below the lower bound of the benchmark-level inflation on 3 of 18 tasks, or on 5 of 18 when DNABERT-2’s Table 9 rows, set aside in Appendix G, are excluded. On these five tasks, however, a reference 3-mer classifier without relatives also loses 0.082 to 0.604, which indicates a shift between the splits. Two of the three Genomic Benchmarks tasks with complete cells have leads inside the inflation, as does 1 of 10 original Nucleotide Transformer tasks (Appendix Table 19). After cleaning, seven models with published values keep most of the published order (mean τ 0.728, compared with 0.787 for our runs on the original splits and 0.819 between our original and cleaned runs). Most of the difference from the published order therefore comes from reproduction, not cleaning (Appendix F).

Reported progress and the leak ceiling. A separate pre-registered analysis asks how much of the progress reported on GUE, the Nucleotide Transformer tasks and Genomic Benchmarks lies above what a no-model leak rule scores on the same test sets (Figure 2). For each task, the reported improvement is the best published score minus the best score of the benchmark’s own authors, and the part above the ceiling is the improvement beyond the larger of that baseline and the best documented leak rule. We froze the 5,553 published rows from 71 sources before computing any primary number; an independent audit reproduced every share, and its row-validity corrections, which raised the shares above the ceiling, define the primary row set P' , with the registered set P alongside it (Appendix G).

In P' , the share of the summed improvement that lies above the ceiling is 0.356 [0.317, 0.402] on GUE (27 tasks with an improvement), 0.365 [0.310, 0.395] on the original Nucleotide Transformer tasks (18 tasks) and 0.862 [0.858, 0.866] on Genomic Benchmarks (9 tasks); in P , it is 0.247, 0.318 and 0.863. By the pre-registered wording rule, most of the reported improvement on GUE and on the original Nucleotide Transformer tasks, about two thirds, lies at or below what a no-model leak rule scores on the same test sets, whereas most of it on Genomic Benchmarks lies above. These verdicts hold in P and in every comparable sensitivity variant. We do not treat the pooled share over the four suites as a finding, because its verdict changes with the row set. None of this shows that the reported progress is spurious. From published scores alone, the part at or below the ceiling cannot be told apart from leak exploitation. The leak rules were selected on the test splits, so the ceiling is optimistic, which tends to raise the share at or below it. Because the intervals become wide under task resampling, the result describes these task sets, not genomic tasks in general (Appendix G).

Genome membership and published wins. Because constructed negatives do not occur in the reference genome, a no-model rule that calls a sequence positive if it aligns full-length to the genome scores MCC 1.000 on seven construction tasks, above every published sequence-only score (post hoc, descriptive; Appendix G). In a pre-registered test, we re-trained recent models that have public code and claimed wins on the 12 construction tasks, and the models they beat, on the original and construction-free versions (Appendix Table 20). Of 31 claims, 19 reproduce. Five of the 6 reproduced wins of a model whose inputs are genome alignments reverse or vanish on the construction-free versions, all Holm-significant, while all 13 reproduced wins of four sequence-only models survive. A registered negative control failed for the aligned-input model on the transcription-factor tasks, so a second registered test drew every negative from the genome, 2–10 kb from its positive. Its registered verdict is that most survive (11 of 19): all 6 wins of the aligned-input model vanish or reverse with a Holm-significant change, whereas 11 of the 13 sequence-only wins survive. Its control passes; at the last checkpoint, however, the verdict is mixed, and these negatives are only loosely matched on G+C (Appendix G).

6 A position-controlled leaderboard and two negative results

Position is the third leak route of the census (Section 2), and PairCtrl lets a reader measure and remove it: it scores genomic language models against row zero, the out-of-fold position-only model, with gene-block bootstrap intervals, on four pair tasks built to control position and relatedness (Appendix H). We scored Evo 2 7B and 40B [30], GenomeOcean 100M, 500M and 4B [31] and GENERator 1.2B and 3B [32] on every pair-task genome, at their final and middle layers and at the layer that a pre-registered rule chose on development genomes (R1). On strict experiment-only interactions, the best genomic readouts cannot be separated from row zero: Evo 2 7B at that layer scores AUROC 0.544 [0.523, 0.567], compared with 0.519 [0.501, 0.533] for row zero, whereas a text-mining control that uses no genome layout reaches 0.597 [0.580, 0.611], which shows that the labels can be learned. For Evo 2 40B, the registered rule applied to all its blocks picks block 12, while its middle layer, block 25, lies past a blow-up of the residual stream. The lower middle-layer gain of Evo 2 40B, which our earlier work could not separate from the choice of depth [9], is therefore due to depth (Appendix H).

Position-controlled training. Two pre-registered attempts to train models whose signal survives strict position controls fail (Appendix I): a classifier on layout-free interaction labels and a pair head with a position adversary. On labels without layout or transferred evidence, neither beats an untrained cosine of the same embeddings, which PairCtrl reports alongside any trained readout.

7 Discussion

Why do constructed negatives cost so much more than relatives? A constructed negative differs from every genomic sequence in a way that needs no biology to detect: a full-length match to the reference genome separates positives from negatives on seven construction tasks (Section 5), and

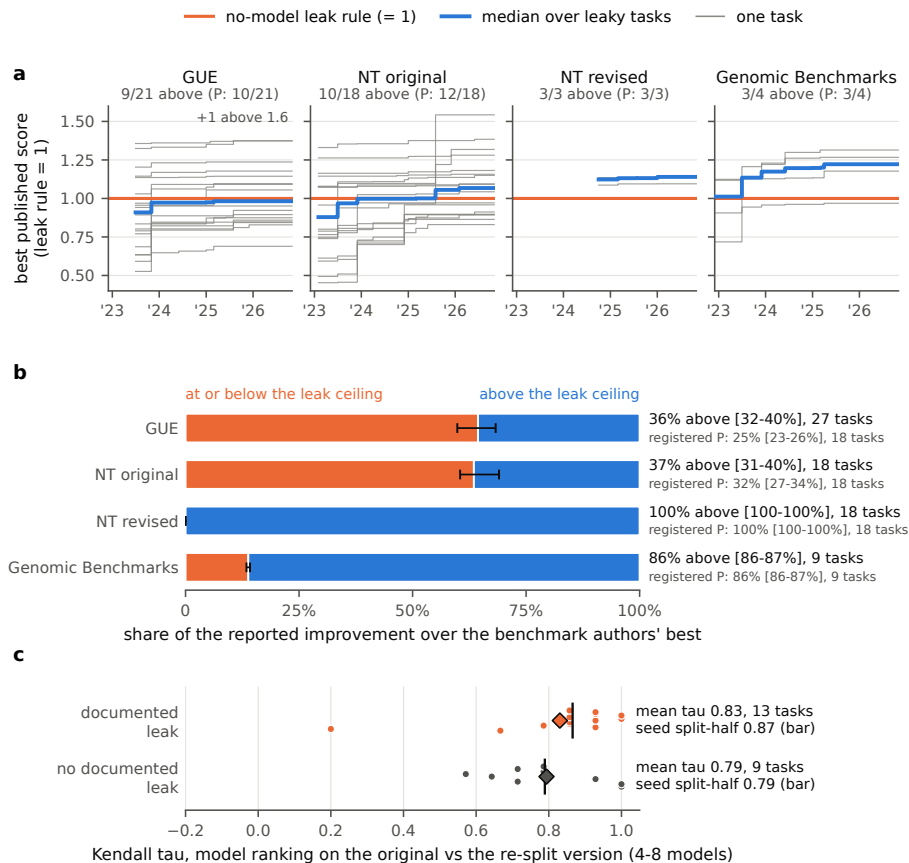


Figure 2: Reported progress and the no-model leak ceiling, for the primary row set P' with the registered set P alongside each label. (a) Best published score over time for tasks with a documented leak, in units of the leak rule's above-chance score on the same test set ($1 =$ the rule): one grey line per task and the median in blue. (b) Share of the reported improvement over the benchmark authors' best that lies above the leak ceiling (blue) or at or below it (orange), with ceiling-bootstrap intervals. The revised Nucleotide Transformer bar is 100% largely by construction. (c) Kendall's τ between the rankings of four to eight fine-tuned models on the original and the re-split version of each task, for tasks with and without a documented leak: one point per task, the mean as a diamond and the seed split-half reliability as a bar. 7 of the 13 re-split versions with a documented leak keep their constructed negatives; with fixneg versions the mean is 0.752 compared with 0.878, and at the last checkpoint 0.767 compared with 0.881 (Appendix G).

within G+C strata the models trained on constructed negatives fall below chance on independent promoters. A model can thus learn the construction instead of the signal, and that cue is useless when every negative is a matched genome window. A relative helps only on the test items that have one, and fine-tuned models gain far less from relatives than a nearest-training lookup does. For benchmark designers, this makes the source of the negatives matter more than deduplication, and it means that the checks that catch copies do not catch constructions (Section 3).

Small as it is, the inflation from relatives exceeds the best model family's published lead on 3 of 11 GUE tasks. Most of the progress reported on the GUE and original Nucleotide Transformer suites cannot be separated from leak exploitation by published scores, so progress there is better measured on the leak-reduced versions. For maintainers we suggest running the linter on every release, drawing negatives from the genome and reporting untrained references alongside every model.

Limitations. The linter's thresholds were set with the earlier audit's tables in view; its held-out tests contain few construction and motif cases (13 pairs in the first). Manual verdicts depend on the published scores we could read (Appendix B), and repackaged tasks count more than once. Drop-in scores are not comparable with published numbers. The leaderboard evaluates zero-shot readouts at a few layers; the fine-tuning results hold for their panel, tasks and recipes.

Acknowledgements

Generative AI tools were used to write code, run analyses and draft text; the author directed the work and takes full responsibility for its content.

References

- [1] Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: efficient foundation model and benchmark for multi-species genome. In: *International Conference on Learning Representations*. 2024. arXiv:2306.15006.
- [2] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, et al. The Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *bioRxiv*. 2023 (version 3, 19 September 2023). doi:10.1101/2023.01.11.523679.
- [3] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat Methods*. 2025;22:287–297. doi:10.1038/s41592-024-02523-z.
- [4] Grešová K, Martinek V, Čechák D, Šimeček P, Alexiou P. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genom Data*. 2023;24:25. doi:10.1186/s12863-023-01123-8.
- [5] Marin FI, Teufel F, Horlacher M, et al. BEND: benchmarking DNA language models on biologically meaningful tasks. In: *International Conference on Learning Representations*. 2024. arXiv:2311.12570.
- [6] Trop E, Schiff Y, Marroquin EM, et al. The Genomics Long-Range Benchmark: advancing DNA language models. OpenReview (submitted to ICLR 2025). 2024. <https://openreview.net/forum?id=8O9HLDrmTq>.
- [7] Cheng W, Song Z, Zhang Y, et al. DNALONGBENCH: a benchmark suite for long-range DNA prediction tasks. *Nat Commun*. 2025;16:10108. doi:10.1038/s41467-025-65077-4.
- [8] Patel A, Singhal A, Wang A, Pampari A, Kasowski M, Kundaje A. DART-Eval: a comprehensive DNA language model evaluation benchmark on regulatory DNA. *Adv Neural Inf Process Syst* (NeurIPS 2024, Datasets and Benchmarks Track). 2024;37:62024–62061. doi:10.52202/079017-1981.
- [9] Tomar TS. Gene position and relatedness confound genomic language model benchmarks. Preprint. Zenodo. 2026. doi:10.5281/zenodo.23605241.
- [10] Burnett J, Blumenthal DB, List M. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Brief Bioinform*. 2024;25:bbae076. doi:10.1093/bib/bbae076.
- [11] Rafi AM, Kiyota B, Yachie N, de Boer CG. Characterizing homology-induced data leakage and memorization in genome-trained sequence models. *bioRxiv*. 2025 (version 2, 25 May 2026). doi:10.1101/2025.01.22.634321.
- [12] Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Comput Biol*. 2018;14:e1006625. doi:10.1371/journal.pcbi.1006625.
- [13] Tahir M, Khan SS, Davie J, Yamanaka S, Ashraf A. LOCO-EPI: leave-one-chromosome-out (LOCO) as a benchmarking paradigm for deep learning based prediction of enhancer–promoter interactions. *Appl Intell*. 2025;55:71. doi:10.1007/s10489-024-05848-6.
- [14] Grešová K. genomic-benchmarks-qc: automated quality control for genomic machine learning datasets. Software. 2026 (version 1.0.0, released 26 August 2026). github.com/genomic-benchmarks/genomic-benchmarks-qc. Accessed 22 October 2026.
- [15] Civale VY, Bagdanov AD, Magi A. What DNA foundation models learn beyond sequence composition. In: *Advances in Neural Information Processing Systems* (NeurIPS 2026, Track on Evaluations and Datasets). 2026. arXiv:2610.15711.
- [16] Feng H, Wu L, Zhao B, et al. Benchmarking DNA foundation models for genomic and genetic tasks. *Nat Commun*. 2025;16:10780. doi:10.1038/s41467-025-65823-8.

- [17] Mendoza-Revilla J, Trop E, Gonzalez L, et al. A foundational large language model for edible plant genomes. *Commun Biol.* 2024;7:835. doi:10.1038/s42003-024-06465-2.
- [18] Ledneva D, Nuridinov M, Kuznetsov D. GENEb: why genomic models are hard to compare. In: *International Conference on Machine Learning (ICML 2026)*. 2026. arXiv:2606.04525.
- [19] Wu W, Li Q, Zhang Y, et al. GENERator: a long-context generative genomic foundation model. *arXiv.* 2025 (version 5, 31 January 2026). arXiv:2502.07272v5.
- [20] Li Q, Wu W, Zhu Y, Feng F, Ye J, Wang Z. Generanno: a genomic foundation model for metagenomic annotation. *bioRxiv.* 2025 (version 3, 14 August 2025). doi:10.1101/2025.06.04.656517.
- [21] Benegas G, Eraslan G, Song YS. Benchmarking DNA sequence models for causal regulatory variant prediction in human genetics. *bioRxiv.* 2025 (version 2, 4 March 2025). doi:10.1101/2025.02.11.637758.
- [22] Zhai J, Gokaslan A, Schiff Y, et al. Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model. *Proc Natl Acad Sci USA.* 2025;122:e2421738122. doi:10.1073/pnas.2421738122.
- [23] Boshar S, Evans B, Tang Z, et al. 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.
- [24] Tang Z, Somia N, Yu Y, Koo PK. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. *Genome Biol.* 2025;26:203. doi:10.1186/s13059-025-03674-8.
- [25] Ren Y, Chen Z, Qiao L, et al. BEACON: benchmark for comprehensive RNA tasks and language models. In: *Advances in Neural Information Processing Systems (NeurIPS 2024, Datasets and Benchmarks Track)*. 2024. arXiv:2406.10391.
- [26] Yang H, Chen R, Li K. Bridging sequence-structure alignment in RNA foundation models. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025)*. 2025. arXiv:2407.11242.
- [27] Shi R, Dalal T, Fradkin P, et al. mRNABench: a curated benchmark for mature mRNA property and function prediction. *bioRxiv.* 2025. doi:10.1101/2025.07.05.662870.
- [28] Nguyen E, Poli M, Faizi M, et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *Adv Neural Inf Process Syst (NeurIPS 2023)*. 2023;36:43177–43201. arXiv:2306.15794.
- [29] Schiff Y, Kao CH, Gokaslan A, Dao T, Gu A, Kuleshov V. Caduceus: bi-directional equivariant long-range DNA sequence modeling. *Proc Mach Learn Res (ICML 2024)*. 2024;235:43632–43648. arXiv:2403.03234.
- [30] Brixi G, Durrant MG, Ku J, et al. Genome modelling and design across all domains of life with Evo 2. *Nature.* 2026;652:1349–1361. doi:10.1038/s41586-026-10176-5.
- [31] Zhou Z, Riley R, Kautsar S, et al. GenomeOcean: an efficient genome foundation model trained on large-scale metagenomic assemblies. *bioRxiv.* 2025 (version 1, 5 February 2025). doi:10.1101/2025.01.30.635558.
- [32] Li Q, Zhan Z, Feng S, et al. GENERator-v2: reconciling coarse tokenization with single-nucleotide resolution in genomic language modeling. *bioRxiv.* 2026 (version 2, 4 May 2026). doi:10.64898/2026.01.27.702015.
- [33] Omri Y, Sagan N, Min E, Choi H, Moon T, Weissman T. Genomic data classification via universal compression. *Research Square.* 2026 (version 2, 3 March 2026). doi:10.21203/rs.3.rs-6363017/v2.
- [34] Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–3100. doi:10.1093/bioinformatics/bty191.
- [35] Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10:421. doi:10.1186/1471-2105-10-421.
- [36] Yang X, Qin C, Wang F, et al. Alignment-aware genomic coordinate encoding for long-range DNA modeling. *bioRxiv.* 2026 (version 1, 12 October 2026). doi:10.64898/2026.10.10.757278.
- [37] Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 2023;51:D638–D646. doi:10.1093/nar/gkac1000.

- [38] Wiatrak M, Abelson DC, Mikhaylenko R, et al. A contextualised protein language model reveals the functional syntax of bacterial evolution. *bioRxiv*. 2025 (version 3, 17 August 2026). doi:10.1101/2025.07.20.665723.
- [39] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379:1123–1130. doi:10.1126/science.ade2574.
- [40] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>.

A Related work in detail

The first author of Genomic Benchmarks released a quality-control tool, `genomic-benchmarks-qc` [14]. It flags single features that separate the classes (sequence length, GC content, nucleotide and dinucleotide composition, per-position nucleotide composition and exact duplicates between classes) by the AUROC of a one-feature classifier, with a warning above 0.6 and a failure above 0.7, and it searches test sequences against training sequences with `MMseqs2`, flagging similarity above 90% by default. Variant-effect benchmarks need matched controls [21], and a compression classifier solves many genomic classification tasks from sequence statistics alone [33]. Closest in aim, Civalé et al. compare frozen embeddings of five DNA foundation models with normalised k -mer frequencies on 57 classification data sets of the benchmark of Feng et al., which repackages tasks of GUE and Genomic Benchmarks among others [15, 16]. The k -mer features remain competitive with most of the models. That study keeps the source splits without measuring overlap between them, and it evaluates frozen representations only, not fine-tuned models. We measure the leaks in the splits themselves, build leak-reduced splits and test fine-tuned models on them, and composition rules enter our audit as a baseline that marks biology, not as a leak. Our earlier audit [9] measured relatedness, construction and position defects across benchmarks of genome organisation and the widely used DNA suites, and introduced `seqleak` (then named `posbench`) for position-controlled metrics. The present work turns that audit into a tool, extends it to newer benchmarks, repairs the leaky tasks and tests whether the leaks change fine-tuned results.

B Census details

Verdict rules. The verdict rules, thresholds and leak definitions are those of our earlier audit [9], unchanged. For the newer benchmarks, benchmark-specific rule families were added: structure-transfer lookups, 9-mer motif rules for NTv3, regression lookups and a reporter-library rule, fixed before the test evaluation of their benchmarks (the reporter-library rule was amended after a first run, so that it reads the library from the sequence alone), and two analyses added after looking at results and marked post hoc, the `TraitGym` population-variation model and a k -mer sketch lookup for `GENERator`’s long taxonomic inputs. A D1 verdict needs the rule to beat the best published score by more than half a unit of that score’s last reported digit. Published scores follow each paper’s own protocol, whereas the rules are scored on the official test split. Each rule is the best of several pre-specified families, chosen on the test split, which makes the rule scores optimistic; per task the choice is among at most three families and, counting the variants inside them, at most nine rules (six on tasks with a defect verdict). Whether a mechanical D3 verdict counts as a leak was decided per task and recorded with its reason. Motifs at the labelled site count as biological signal. For 9 of the 13 benchmarks, one published number was reproduced first (Table 3).

Plant Genomic Benchmark. In the poly-A tasks, every negative lies within 50 bp of a positive, as the benchmark’s authors describe [17]. A rule that combines the arrangement of the test windows with a motif table therefore exceeds every published score on 5 of the 6 poly-A tasks. It needs the whole test set (it is transductive), so we do not count it; when windows are scored one at a time, no rule reaches the published scores. Two pairs of data sets share sequences across tasks. The promoter-strength tasks of the two assays (leaf and protoplast) share most test sequences with the other assay’s training split (79.2% and 76.0% of test items), and a lookup that copies the other assay’s measurement explains little of the variance (r^2 0.197 and 0.204). Similarly, in the two rice chromatin data sets 85.5% and 43.1% of the test windows are near-identical to the other cultivar’s training windows. Whether published models were trained on both members of a pair is not stated, and rules built on these copies stay below the published scores. In the gene-expression tasks, 7.0–33.4% of test windows have a near-duplicate training window from a neighbouring gene, but the nearest-training lookup explains at most 0.026 of the variance, so expression does not transfer through these copies. The chromatin tasks hold out chromosomes, and no test window overlaps a training window. The two rice chromatin data sets come from two cultivars. The per-tissue AUROCs reported for the two

Table 3: Reproduced published numbers: one per audited benchmark where possible (four were not reproduced), and the models of the fine-tuning panel.

Benchmark	Reproduced number	Ours	Published
Plant Genomic Benchmark	AgroNT AUROCs recomputed from the released ROC points	14 of 14	–
LRB	released predictions that reproduce the latest paper version at its precision	50 of 102	–
GENEB	SpliceBERT MCC, iDHS task	0.7080	0.7080
DART-Eval	HyenaDNA zero-shot paired accuracy, task 1	0.8931	0.891
TraitGym	CADD logistic regression AUPRC, Mendelian traits	0.8760	0.8747
GENERator	DNABERT-2 fine-tuned, weighted F1 (one run; published: mean of 10)	0.6573	0.660
GENERanno	DNABERT-2 fine-tuned, weighted F1 (one run; published: mean of 10)	0.9012	0.902
RGB	ViennaRNA token macro F1, ArchiveII	0.7476	0.7399
mRNABench	naïve-4 linear probe, one seed	0.504605	0.504604
PlantCaduceus, BEACON, NTv3, Tang and Koo	not reproduced: model inference or fine-tuning not run, no released baseline code, or relatedness checks only	–	–
Fine-tuning panel (Section 5): DNABERT-2	mean MCC, five human transcription-factor tasks of GUE (published by its authors)	0.6846	0.7011
Fine-tuning panel (Section 5): HyenaDNA	same tasks and metric (published run by the HybriDNA authors)	0.6369	0.6174
Fine-tuning panel (Section 5): Caduceus	same tasks and metric (published run by the HybriDNA authors)	0.6521	0.6517
Fine-tuning panel (Section 5): NT-v2 50M	mean accuracy, same tasks (published run by the dnaGrinder authors)	0.7969	0.7680
Fine-tuning panel (Section 5): CNN	accuracy, Genomic Benchmarks human enhancers of Cohn et al.	0.6621	0.6950

cultivars are identical to 15 digits in the published source data, in all 15 tissues, while the baseline’s values differ [17].

LRB. On the zero-shot OMIM task, label-free TSS distance scores AUPRC 0.079, above every single-sequence DNA language model in the LRB paper (at most 0.003) and below the alignment-based GPN-MSA (0.350). Windows made entirely of N (assembly gaps) are 5.67% of the promoter task’s test items and all are negatives, but an N-count rule scores only AUROC 0.530. No CAGE test window overlaps a training or validation window (0 of 1,922). On the ClinVar task, the label-free distance to the nearest exon junction scores AUROC 0.637 (verdict N). Its metadata carries a consequence annotation, and a prior fitted on it would score 0.955, so we treat that field as a documentation issue and not as a rule.

Benchmarks published from 2024 to 2026. We first give details of the flagged rows and verdicts, then of the benchmarks audited in part or not at all. mRNABench’s flagged rows include 40 eCLIP targets that share one split per cell line. Examples are an iPro promoter set of GENEB [18] (19% near-duplicates, D1), GENERanno’s SSU taxonomic task (81% near copies, D3) and mRNABench’s designed-UTR task (62% near copies, D3). DART-Eval task 1 [8] pairs each positive with a dinucleotide shuffle; an order-10 Markov rule trained on its training sequences reaches a paired accuracy of 0.917, against the best published 0.976 (D3). Of the six D1 verdicts, three come from motif tables on 4mC sets of GENEB, one from a near-duplicate lookup on a promoter set of GENEB, and two concern eCLIP targets of mRNABench’s homology split. For these two, the lookup’s score exceeds the best published score on the mean of 10 seeds, but the paired per-seed difference has an interval that includes zero. The 4mC motif is the sequence context of the modified base, which is partly biology, and the only comparable published scores on these splits come from frozen-embedding probes. mRNABench’s homology split assigns rows of genes missing from its HomoloGene map one at a time, and the map covers only 41–67% of rows; nine further eCLIP targets are assigned D3. The remaining findings each concern a single benchmark or data set. TraitGym’s Mendelian controls are common variants that are not matched on allele frequency [21]. On TraitGym, 10 population-variation annotations alone reach AUPRC 0.689, 0.760 of the best published score above chance (D3, caveated). GENEB labels one of its sets 5mC, whereas its sequences match the

5hmC set of the source benchmark. Two of GENEanno’s taxonomic tasks have exact copies of training items for 39.6% and 55.2% of their test items, with the same label, but the lookup scores stay below the published scores. In the Tang and Koo eCLIP data of one RNA-binding protein, 27.7% of test sequences are exact copies of training sequences. Its published values appear only in figures, so no strict comparison was made. A twelfth benchmark, that of Feng et al. [16], shares 15 of its 17 classic test sets with GENE, which were audited once through GENE, and its two enhancer tasks were not audited. The other 14 need controlled access or a submission for test labels, have no training split, are out of scope or remain listed for later audits.

Protein and RNA tasks. Table 4 summarises the census of public protein and RNA tasks by category. A later amendment (C7) added the ten VenusX residue tasks, which the census had excluded for their size. In the eight domain tasks among them, 0.42–0.78 of the test proteins are exact copies of training proteins, but the copied proteins carry another domain’s annotation, and the residue lookup stays below the majority baseline.

Table 4: Census of 204 public protein and RNA tasks in 50 source families: tasks with a relatedness flag (copies in at least 0.10, or relatives at 30% identity in at least 0.25, of the test items; RNA tasks: the linter’s copy or homolog checks), tasks with a comparable published score, and defect verdicts (D1–D3) among them.

Category	Tasks	Relatedness flag	With a published score	Defect verdict
Function	96	92 of 96	11	6
Localisation	26	6 of 26	10	0
RNA structure	8	6 of 8	2	0
Solubility	14	13 of 14	5	0
Stability	23	20 of 23	10	1
Structure	37	11 of 37	4	1
All audited tasks	204	148 of 204	42	8

Sources not read. Some full texts could not be retrieved: 5 plant-genomics preprints, 3 journal articles behind access controls and the review threads of the LRB submissions. No D1 verdict depends on them, and a higher unread score could only lower a rule’s ratio to the best published score.

C Linter details

Checks and thresholds. Table 5 lists the checks of the frozen rules v1.0. Shares count test items with a relative in the fit splits (training and development). Shortcut rules use one feature, with threshold and direction chosen on the training split, and the feature of a check is chosen by grouped cross-validation on the training split. When the minority class of a binary task is below 20%, shortcut rules are also scored by AUROC, with AUROC thresholds $(1 + \text{MCC threshold})/2$, and a check takes the more severe of its two statuses. Twins fail only when their label relation, chance-corrected, lets a twin rule predict test labels and the task’s negatives were constructed. The position check needs distance fields, or a seqleak gene-pair task. Near copies are found by containment of canonical 21-mers, and a nearest-training lookup is reported as a reference. seqleak clean writes a manifest of its changes. Without the optional homology search, one thread audits 0.50 s per million bases on eight benchmark tasks.

Rule history. Table 6 gives the validation by rule version. Four amendments preceded the first run. Two later amendments were post hoc. A5 sets the FAIL levels of near-duplicate shares (0.10) and of window and pair copies (0.25) at the manual audit’s operating points, and it lets twins fail only for constructed negatives. It was chosen from a sensitivity grid after the registered rules had failed 21 of the 27 tasks that lack a documented leak. A6 scores shortcut rules by AUROC when the minority class is below 20%, after two splicing tasks of the Plant Genomic Benchmark (2.3% and 3.4% positives) had passed the motif check. In a first form that covered every binary task, A6 failed one task without a documented leak, so we restricted it to imbalanced tasks before use.

Table 5: Checks of the frozen linter rules v1.0, with WARN and FAIL thresholds on their statistics.

Check	Statistic	WARN	FAIL
overlap.exact	share of test items with an exact or reverse-complement copy in the fit splits	0.005	0.05
overlap.near	share with a near copy (containment of sampled canonical 21-mers)	0.01	0.1
overlap.windows	share overlapping a fit-split window (coordinates) or with a collinear shifted copy	0.01	0.25
overlap.pair	share of pair items whose elements all match one fit-split pair	0.01	0.25
overlap.homology	share with a BLASTN homolog (optional search)	0.05	0.25
reference.copies	share copied from or into another benchmark’s fit splits	0.01	0.05
shortcut.length	test MCC of a length rule fitted on the training split	0.1	0.3
shortcut.ambiguous	same, ambiguous-base features	0.1	0.3
shortcut.case	same, lowercase share	0.1	0.3
shortcut.composition	same, GC, CpG and Markov naturalness (never FAIL)	0.5	never
shortcut.kmer	order-3 Markov classifier (never FAIL)	0.5	never
shortcut.position_motif	table of fixed-position motifs	0.4	0.7
construction.shuffle	share with a composition twin across the split	0.01	0.05
construction.segment	share with a positional twin across the split	0.01	0.05
position.distance	AUROC of the position-only model on distance fields	0.6	0.75
position.row_zero	same, seqleak row zero for gene-pair tasks	0.6	0.75
baseline.lookup	nearest-training lookup with interval (reference only)	–	–

Table 6: Validation of the linter by rule version. IN: the task set informed the rules. OUT: it did not. Recall counts documented (task, leak class) pairs with a FAIL in a check of that class. False FAIL counts tasks without a documented leak that fail.

Rule version	Earlier audit (80 tasks): recall; false FAIL	Synthetic: cells met	PGB (31): false FAIL; relatedness FAIL	Held out: newer benchmarks (42) recall at FAIL; at FAIL or WARN; false FAIL. LRB (6): recall at FAIL; false FAIL
Registered (A1–A4 before any run)	IN: 87/87; 21/27	OUT: 32/32	OUT: 4/8; 15/15	–
A5 (post hoc)	IN: 87/87; 0/27	OUT: 32/32	OUT: 0/8; 13/15	–
v1.0 = A6 (post hoc), frozen	IN: 87/87; 0/27	OUT: 32/32	IN	OUT: 28/39; 38/39; 1/13. LRB: 0/1; 0/5

Validation details. After the held-out comparison on the newer benchmarks, the manual audit added two exact-copy flags that the linter’s reports had pointed to; against the revised flags, recall at FAIL is 30 of 41. We report 28 of 39 because the revised reference is no longer independent of the linter. Only 8 of the 26 relatedness and copy flags lie on tasks with a defect verdict. The others lie on tasks with verdict N or C (14), on a task that is relatedness by design (2) and on the re-hosted in-sample task (2). The synthetic suite has 780 runs; its cells without an expectation, the weakest doses of some planted leaks, did not fail. The Plant Genomic Benchmark was audited by hand in parallel (Section 2) and is out of sample for A5. There, A5 fails none of the 8 tasks to which the audit assigned no leak class (the registered rules failed 4) and 13 of the 15 relatedness tasks. The two misses are copies of another task’s data, which only the reference check detects. Of the 6 LRB tasks that were scored, only the promoter task has a defect counted in Table 1 (TSS distance), and the linter cannot detect it without distance fields. That task fails instead for exact copies of all-N assembly-gap windows among its negatives. None of the other five fails.

Synthetic planted leaks. The base task has random sequences of length 300 from an order-3 Markov chain fitted to the human genome, 4,000 training and 1,000 test items and balanced labels that do not depend on sequence. Nine kinds of leak are planted at doses of 2, 10, 30 and 90% of the test items: exact copies, near copies, shifted windows, dinucleotide-shuffled and segment-substituted negatives with their sources in training, a length shortcut, an N-padding shortcut, a fixed-position motif and a position field. The three controls are the base task, the base task with a genuine composition signal and the base task with a motif at random positions. A cell with an expectation passes when at least 19 of 20 seeds meet it.

`seqleak clean`. Run on the earlier audit's 80 tasks with all links and re-audited, the filter mode leaves the overlap and twin checks at PASS on 68 tasks and WARN on 12, and no task without a documented leak fails. Twins inside a split remain a WARN, because re-splitting cannot remove constructed negatives. Filtering drops a median 96.1% of the test items of leaky tasks and 16.1% of the others. The drop-in versions of Section 4 re-split instead, which keeps the split sizes.

Agreement with the manual audit. The linter's near-duplicate share correlates with the earlier audit's BLAST near-duplicate share (Spearman 0.775 over 74 tasks), and its alignment-free window share with the BLAST homolog share (Spearman 0.898). All 10 tasks with a BLAST near-duplicate share of at least 0.10 have a linter share of at least 0.10. The linter's *k*-mer lookup agrees with the earlier audit's BLAST lookup (Pearson 0.969 over the 80 tasks). On the 74 single-sequence tasks, its best no-model rule exceeds every published score we found on 24 tasks, against 18 for the earlier audit's rules, whose count of 24 includes six pair tasks. The six new tasks are three GUE mouse tasks, so that all five now count, and two promoter tasks, where a twin rule with an order-3 Markov fallback wins. One of the promoter tasks is a near-tie (0.9507 against 0.9503). The sixth is GUE's 20-species classification, where the winning fallback is a composition model, which captures biological signal rather than a leak. As in the earlier audit, these rules are selected on the test split.

Limitations found on the held-out sets. Each comes with the evidence that showed it.

1. The motif FAIL bar is absolute. Six 4mC methylation tasks reach test MCC 0.287–0.437 and stay at WARN, and so does an *Arabidopsis* TATA promoter set (MCC 0.672, minority class 0.342, so the AUROC criterion does not apply). The manual audit judges these tables against the best published score.
2. Constructed negatives inside a split are a WARN. In DART-Eval task 1, 80.4% of items are in composition-twin pairs with different labels, and no twin crosses the split.
3. Tasks without a training split (LRB OMIM) cannot be scored.
4. No check flags windows made entirely of N. The LRB promoter task fails `overlap.exact` at 0.0552 because of such windows, all of them negatives.
5. No guard exists for splice consensus. A rice splice-donor task fails on its motif table (MCC 0.507, AUROC 0.853, minority 0.167), the one held-out false FAIL.
6. No check covers pooled reporter libraries (BEACON APA), which the manual audit finds predictable from sequence.
7. The position check needs distance fields. Without them it skips (LRB promoter), and with them it runs (LRB ClinVar, WARN at AUROC 0.644).
8. Rule-based checks skip regression labels, and overlap checks still run.
9. WARN is common on tasks without a documented leak: 26 of 27 in sample and 9 of the 13 held out. FAIL is the verdict, and WARN marks a task for manual review.
10. The lookup uses exact 21-mers. It is lower than the earlier audit's BLAST lookup by more than 0.1 on 11 of the 80 tasks.
11. Copies of another benchmark are found only with `-reference`: the PGB promoter-strength tasks fail `reference.copies` at 0.80 and 0.77 with it. Leaks through the arrangement of test windows, as in the PGB poly-A tasks, have no check.

Second held-out test (v1.1 against v1.0). Rule version v1.1 addresses known limitations of v1.0 with four new checks (twin pairs with different labels anywhere in the data, windows with less than half A, C, G or T, negatives that lack a site carried by at least 95% of a class, and constant library segments at a fixed offset), a published-relative motif criterion and the scoring of tasks without a training split; it was designed and frozen on the in-sample sets only. The primary held-out sample contains one task per data set of the census pool: 58 tasks audited blind by hand. Of these, 26 together contain 40 task–class pairs. Both versions miss the same 8 pairs: 2 constructed-negative pairs detectable only through genome membership, 2 anchored motifs in sequences of variable length (the motif check skips them, and both tasks fail the length check), shuffled negatives that occur only

within the test split (a WARN), a borderline segment-substitution pair (a WARN) and two relatedness pairs, one borderline and one of same-label positional twins, which the linter reports as a WARN by design. Both raise the same 4 false flags: exact copies for 5.5% of the test items of one task, near copies in a task for which the audit’s measures are not available, and two tasks that the audit marks as borderline relatedness. In a paired comparison, the two versions differ on no pair (exact McNemar $p = 1$). The secondary sample (the audit’s first draw) contains 59 tasks, 39 of which come from one repository. Of its 92 pairs, v1.1 finds 16 at FAIL and v1.0 finds 14; the 2 extra pairs are shuffled negatives that the new twin check catches ($p = 0.5$). Both versions miss its 39 genome-membership pairs. Rule version v1.2 adds a genome-membership check, frozen before it was run on two fresh pools of data sets whose negatives were labelled blind from their documentation. In each pool, the one applicable data set with documented synthetic negatives fails the check, and no data set with genomic negatives does (0 of 8 and 0 of 5). All of these counts are below the registered minimums, so these results are descriptive, and we make no claim from the pooled secondary line (2 of 2 and 0 of 13, upper bound 0.247).

Spike-in test of v1.2 (registered). In a simulation frozen before any spike-in data were generated, we replaced the negatives of the 13 data sets with documented genomic negatives from parts B and B2 and ran the frozen genome-membership check on them. With whole negatives made by dinucleotide shuffling, by GC-matched random sampling or by an order-5 Markov model, 11 of the 13 fail the check in at least two of three replicates, for each of the three recipes (a share of 0.846; 95% interval 0.546–0.981). Replacing a single contiguous 10% segment of each negative already makes 12 of the 13 fail. None of the 13 unmodified data sets fails (upper bound on the share failing, 0.247). With genome windows as negatives, 1 of the 13 fails; its positives are transcript-derived. For this control the upper bound on the share failing, 0.360, is above the threshold of 0.30, so the control’s criterion is not met. One miss is structural: when the training items come from other species, as in one data set, the rule, which is fitted on the training split, has little cross-validation signal (MCC 0.160; the check requires 0.50). The other came from a defect of v1.2: in multi-class tasks the binning can collapse to one bin when membership takes only two values, as for a seven-class data set of part B2, whose prediction was then constant (MCC 0). The fixed rules (v1.2.1, released as seqleak 0.4.1) remove this defect: with them the seven-class data set fails (MCC 0.795), 12 of the 13 fail under each recipe (a share of 0.923; 95% interval 0.640–0.998). The fix changes no other number in this paper. The simulation reuses the fresh pools’ data sets and does not make the descriptive result on those pools confirmatory.

D The linter on public data sets

Protocol. The census followed a design registered before any screening of search results and before any linter run, and its amendments were made before the runs they affect. Fixed query lists on Hugging Face and GitHub gave the candidates: every classification or regression data set of genomic sequences whose train and test split is published, beyond the benchmarks audited in Section 2 and their identical re-hostings. Each task was converted to unified splits and audited with the frozen rules v1.0 unchanged. Tasks whose test items lie mostly inside an audited task stay in the census and are marked as derived. The data sets that no audit has seen form the held-out pool of a later rule version.

Controls. None of the three registered controls failed. The pipeline reproduces the frozen reports of 10 re-hosted validation tasks of Section 3 on all 112 compared checks. A synthetic task without a planted leak passes, and the same task with exact copies planted in 10% of its test items fails the exact-copy check, end to end through the converter. A task re-run on a second machine reproduces its report exactly. A separate recount of exact copies, which was not registered, agrees exactly with the linter’s exact-copy shares. An external check registered before any linter run ran the linter on the five example data sets of genomic-benchmarks-qc, with expectations for three of them, and met one: the hidden-motif example is caught as expected, but the tool’s clean example, a re-hosted subsample of GUE mouse1, fails the shuffle-twin check, and its composition-bias example, a subsample of GUE prom_core_all, stays below WARN on the composition checks.

Mechanisms and limits. In scope, 117 tasks fail overlap checks, 14 fail the fixed-position motif check and 10 fail artefact checks. By type, the overlap failures are near copies (97 tasks), homologs (95), overlapping windows (87), exact copies (42) and pair copies (2), and the artefact failures are length (9) and letter case (1). No task fails a construction check, and 610 tasks warn on one, 475 of them in one data set; v1.0 cannot fail constructed negatives that stay inside a split, the second limitation in Appendix C. The largest data set holds 507 tasks in one format, and 5 of them fail. For two very long tasks, the reports are provisional. As a pre-registered contingency allows, they were made with the homology search off, before the main runs with the search on had finished. Both tasks already fail other overlap checks, so the homology search cannot change their verdicts.

Examples checked on the raw files. Independent of the linter, the test sequences of three promoter data sets all occur in their raw training files (10,363 of 10,363, 1,168 of 1,168 and 4,221 of 4,221). In another data set the training, validation and test files are byte-identical, and in a third the test file equals the validation file while the labels differ. A methylation benchmark re-split in 2026 puts 81% of the test items of one task in its training split, against 0.02% in the original split.

Table 7: The frozen linter rules v1.0 on public data sets, per data set (tasks in scope only). FAIL: tasks with a FAIL in a leak family. Failing checks: in order of the number of tasks failing them.

Data set	Tasks	FAIL	Failing checks
GitHub: aaroshdas/crispr-cas12a-on-target-prediction	1	0	–
GitHub: AnonymousISCBConf/Win-Win-Privacy-Utility-Analysis	1	0	–
GitHub: BionicsAI/model_4mc	1	0	–
GitHub: Cpillar/MOSAIC	18	16	exact, homology, near, windows, motif
GitHub: FakeEnd/iDNA_ABF	13	4	motif, near
GitHub: Jackie-Suv/iPro-MP	11	9	homology, near, windows
GitHub: jun-csbio/Co6mA	4	0	–
GitHub: katarinagresova/benchmarks	18	0	–
GitHub: Kshitizk623/Promoter_prediction	1	0	–
GitHub: leirufeng/iEnhancer-TTMM	1	1	exact, homology, near, windows
GitHub: lftxd1/iEnhancer-DLRA	1	1	length
GitHub: LIU-CT/MSNet-4mC	6	0	–
GitHub: liu-h-y/5hmC_model	1	0	–
GitHub: Lyannas94/HybProm	5	5	exact, homology, near, windows, motif, case
GitHub: mldlproject/2022-i4mC-GRU	1	1	length
GitHub: mldlproject/2023-i6mA-CNN	1	1	length
GitHub: narissu/4mC-prediction-methods	6	1	exact, homology, near, windows
GitHub: Nisasoylu/BERT2OME	1	0	–
GitHub: NWAUFU-LiuLab/m7Gpredictor	1	0	–
GitHub: OluwadareLab/EnsembleSplice	6	5	near, motif, homology
GitHub: Rafid013/CRISPRpredSEQ	3	3	exact, homology, near, windows
GitHub: Rehman1995/DCNN-4mC	12	2	exact, near, homology, windows
GitHub: Rhys-sg/e.-coli-promoter-prediction	1	1	homology, near, windows
GitHub: sadmansakibmahi2/DNA-Promoter-Gene-Prediction-with-ml-model	1	1	exact, near, windows, length
GitHub: ShengWang-XTU/VoxLncLoc	3	2	homology, near, windows, exact
GitHub: Stegobully/merged_ncRNAclassifier	1	1	homology, length
GitHub: tanvirehsan/prediction-of-atoirna-editing-sites	1	1	exact
GitHub: UdeS-CoBIUS/LSC-ncRNA	9	5	homology, windows, near
GitHub: XueqiangF/Ense-i6mA	2	1	exact
GitHub: yangyn533/HCRNet	31	0	–
GitHub: yanzhu22/TIMER	7	0	–
GitHub: yjccm/BF-m7GPred	1	1	near
GitHub: yq-geng/PL-m5C	1	0	–
GitHub: yygen89/riceMirP	1	1	length
GitHub: Z-Abbas/m5C-pred	1	0	–
HF: addykan/genome-classification-with-random	1	1	homology, near, windows
HF: AI4Biology/Cross-Molecular-Benchmark	3	3	windows, exact, homology, near, pair
HF: anikethjr/promoter_design	8	8	homology, near, windows
HF: anudaw/genome-classification	1	1	homology, near, windows
HF: BatatasLM/Batatas_GUE	1	1	homology, near, windows
HF: BGI-HangzhouAI/Benchmark_Dataset-Genomic_element_classification	1	0	–
HF: bioinfothb/Fish_GUE	1	1	exact, homology, near, windows
HF: build-small-hackathon/clinvar-cls	1	1	near, pair, windows
HF: cerebras/exome_bench	1	1	exact, homology, near, windows
HF: chejames/nucleosome-condensability-shuffled-split	2	2	homology
HF: dimostzim/siRBench-data	2	0	–
HF: dnagpt/laya-bio	1	0	–
HF: edbeeching/malinois-mpra-regression	3	3	homology
HF: genbio-ai/rna-downstream-tasks	10	5	motif, homology, near, windows

Data set	Tasks	FAIL	Failing checks
HF: GenerTeam/DeepSTARR-enhancer-activity	2	0	–
HF: genomic-benchmarks/miRBench	1	1	near, windows
HF: GleghornLab/mrna_half_life	1	0	–
HF: GleghornLab/mrna_stability_other	1	1	exact, homology, near, windows
HF: GleghornLab/tc_riboswitches_fungal	1	1	homology, near, windows
HF: InstaDeepAI/true-cds-protein-tasks	3	2	homology, windows, near
HF: Lab-Rasool/ENCODE-TFBS	507	5	homology
HF: LiukG/mus_promoter	1	1	homology, near, windows
HF: lsalsi/5utr_pat_ben_class	1	1	homology, near, windows
HF: lsalsi/genomic_regions_annotated	1	1	near, windows
HF: lsalsi/mrl_prediction	3	0	–
HF: mahdibaghbanzadeh/asm_name	1	1	homology, near, windows
HF: mahdibaghbanzadeh/dna_sample_30000	1	1	near, windows
HF: mahdibaghbanzadeh/genome_origin_datasets	2	1	near
HF: mogam-ai/CDS-BART-Ecoli-proteins	1	0	–
HF: mogam-ai/CDS-BART-fungal-expression	1	1	near
HF: mogam-ai/CDS-BART-mRFP-expression	1	1	homology
HF: mogam-ai/CDS-BART-mRNA-stability	1	1	exact, homology, near, windows
HF: neuralbioinfo/bacterial_promoters	2	2	near, homology, windows
HF: Nhoodie/hgt-bootstrap-v2-synthetic	1	1	near, windows
HF: roa7n/G_quad_DNA_tokenized_K6	1	0	–
HF: saiden89/xiang2021-spcas9	1	1	homology, near, windows
HF: SaumyaGupta-99/spliceformer-data	1	1	exact, homology, near, windows
HF: sheryylli/utr_total_reads	1	0	–
HF: suzuki-2001/athaliana_rmaseq_splicing	1	0	–
HF: tattabio/ec_classification_dna	1	0	–
HF: tattabio/mibig_classification_dna	1	0	–
HF: Taykhoom/functional-multiclass-gamba	3	0	–
HF: Taykhoom/functional-random-gamba	2	0	–
HF: Taykhoom/functional-random-noannot-gamba	2	0	–
HF: Taykhoom/functional-upstream-gamba	2	0	–
HF: TerminatorJ/localization_multiRNA	1	0	–
HF: tydmy/150bp_human_vs_microbial_dna	1	1	exact, homology, near, windows
HF: wanglab/variant_effect_non_snv	1	1	near
HF: zhangtaolab/fungi-histone-modifications	1	1	exact, homology, near, windows
HF: zhangtaolab/human_lentiMPRA	1	1	exact, homology, near, windows
HF: zhangtaolab/plant-multi-species-core-promoters	1	0	–
HF: zhangtaolab/plant-multi-species-histone-modifications	1	1	homology, near
HF: zhangtaolab/plant-multi-species-lncRNAs	1	1	homology, near, windows
HF: zhangtaolab/plant-multi-species-promoter-strength	1	1	exact, homology, near, windows
HF: zhangtaolab/plant-multi-species-sequence-conservation	1	0	–
HF: ZhejiangLab/RiceBenchmark	19	13	near, windows, homology, exact, length
HF: ZJdog/hf_tutorial_dna	3	0	–

E Release details

Contents. The release holds 33 GUE tasks (ten yeast histone marks, five human and five mouse transcription-factor tasks, six promoter tasks, six enhancer–promoter tasks and splice sites), the 18 original Nucleotide Transformer tasks and the 9 Genomic Benchmarks tasks, plus the two DNALongBench drop-ins. Each version folder holds the data in its suite’s format (for GUE and the Nucleotide Transformer tasks, rebuilt from the original download by the build script), an item table with the original split and row of each item, the leak rules of the version and a READY file with the sha256 of each file. Released versions are immutable: a rebuild rewrites a version only if its content changes.

Versions. Each version stores the row ids of every construction round and the genomic coordinates of every item. *resplit_union*, recommended for the 7 tasks that have it, cuts the union of the audit’s and the linter’s relations, so that neither tool finds a relative across splits. *fixneg*, in three independent draws for the 12 construction tasks and, as controls, the five GUE human transcription-factor tasks, replaces each negative by a reference-genome window matched on length, GC count and repeat content and drawn on the chromosomes of its split, which changes the task. *purged* and *purged_tight* keep the original test set and remove every related training row, for comparison with published numbers, and *placebo*, a second chromosome hold-out for 6 tasks, separates the effect of the chromosome choice from that of leak removal. *fixneg_resplit*, in three draws for the 12 construction tasks, re-splits each fixneg draw with the relations of resplit_union. Its README recommends it where all three draws pass every check, and fixneg where only the fixneg draws all pass (Table 8).

Table 8: Combined versions (fixneg_resplit) of the 12 construction tasks, three draws each. Passing: draws that pass C1 with every rule, including the genome-membership rule, and draw no linter FAIL. Lookup: nearest-training lookup (MCC) of the combined draws, chance 0. Survival: share of the fixneg draw’s items kept. The genome rule scores 0.000 and the linter finds no leak in all 36 combined versions. Recommended: the version the README recommends; –, none passes in all draws, and the README lists what remains.

Task	fixneg_resplit passing	lookup	survival	fixneg passing	recommended
GUE mouse0	3 of 3	0.039	0.994–0.995	3 of 3	fixneg_resplit
GUE mouse1	0 of 3	0.322–0.323	0.893–0.894	0 of 3	–
GUE mouse2	0 of 3	0.105–0.137	0.980–0.991	0 of 3	–
GUE mouse3	3 of 3	0.034–0.127	0.982–0.983	1 of 3	fixneg_resplit
GUE mouse4	3 of 3	0.013–0.023	0.989–0.991	3 of 3	fixneg_resplit
GUE prom_300_all	1 of 3	0.048–0.069	0.972	1 of 3	–
GUE prom_300_notata	0 of 3	0.053–0.079	0.969–0.972	0 of 3	–
GUE prom_core_all	3 of 3	0.000	0.971–0.972	0 of 3	fixneg_resplit
GUE prom_core_notata	3 of 3	0.000	0.971	0 of 3	fixneg_resplit
NT promoter_all	0 of 3	0.074–0.088	0.972–0.975	0 of 3	–
NT promoter_no_tata	3 of 3	0.011–0.035	0.971–0.976	0 of 3	fixneg_resplit
NT promoter_tata	2 of 3	0.069–0.100	0.979–0.984	3 of 3	fixneg

Table 9: Leak-reduced versions and their checks. C1: the audit’s leak rules at chance; with genome rule: C1 with the genome-membership rule added on the 12 construction tasks. Linter: versions with a FAIL in an overlap or construction check of the frozen v1.0 rules. Survival: median share of the original rows kept; fixneg versions replace the constructed negatives with genome windows (some splits have up to 0.5% fewer), so for them it counts the original positives only, and re-splitting a fixneg draw keeps a median 97.3% of its items. The 51 fixneg draws are 36 on the construction tasks and 15 on the transcription-factor controls; with the genome rule, 11 and 9 of them pass C1; the 27 strict versions are listed in Table 10.

Version	Purpose	C1 pass	with genome rule	Linter FAIL	Survival
original	the benchmark as released	–	–	47 of 60	–
purged	original test set kept	14 of 60	13 of 60	1 of 60	0.802
purged_tight	as purged, tighter relations	26 of 59	22 of 59	1 of 59	0.619
resplit	default drop-in	32 of 60	28 of 60	3 of 60	0.938
resplit_union	recommended drop-in	2 of 7	0 of 7	0 of 7	0.927
resplit_twin	linter’s own re-split	0 of 3	0 of 3	0 of 3	0.994
placebo	second chromosome hold-out	4 of 6	4 of 6	0 of 6	0.691
fixneg (3 draws)	genomic negatives	20 of 51	20 of 51	0 of 51	0.464
fixneg_resplit (3 draws)	genomic negatives, re-split	21 of 36	21 of 36	0 of 36	0.461

Checks and residuals. Figure 3 shows the lookup’s score before and after re-splitting and the causes of the residuals. Of the 28 resplit versions that fail the negative control as released, five retain segment twins. The resplit_union versions remove these twins, but the lookup’s score barely changes. All five still contain constructed negatives. In the other 23, the score above chance comes from sub-threshold homology, which is removed only by a looser homology relation (as in the strict variants), or from constructed negatives, composition or fixed-position motifs, which no choice of split removes. In the three exceptions to the positive control, all promoter tasks, the lookup’s score falls after injection, for a reason we did not test. Two shortcuts survive any re-split: sequence length in the Ensembl regulatory task of Genomic Benchmarks, and fixed-position motifs in the three Nucleotide Transformer splice tasks, for which no fixneg version was built. In the fixneg versions of the 300-bp promoter tasks, exact GC matching covers only 87.2–93.3% of negatives, which leaves a residual: a GC rule still reaches MCC 0.082–0.102.

The *eqtlp_corrected* drop-in is built with the strand-aware variant–TSS distance, a chromosome hold-out, distance-matched negatives and fixed-length inputs. The test AUROC of its distance rule falls from a mean of 0.751 on the released split to 0.499–0.506 on the test split of every tissue. In *etgp_nodist*, which joins the element and TSS windows by a fixed spacer, input length correlates with distance at $r = 0.021, 0.137$ and -0.058 in the three data sets, against 1.000 for the released inputs.

Construction. We map items to their reference genome with minimap2 [34] (query coverage at least 0.9 and identity at least 0.95). Leak relations are overlapping genome windows, BLASTN homologs [35] (identity at least 80% over at least 50% of a sequence, and in the tightened round any hit over at least 30%), composition and positional twins and exact or reverse-complement duplicates. Held-out chromosomes are chosen from item counts only, without labels or model results. A fixneg

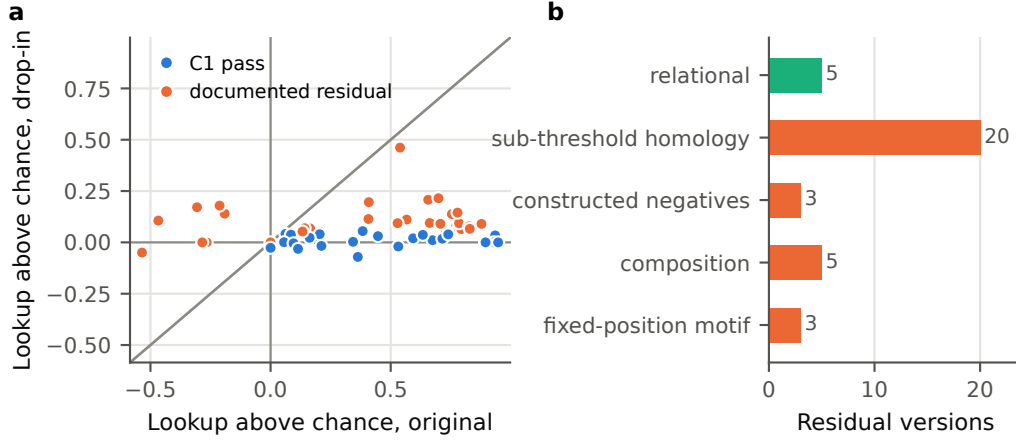


Figure 3: Leak removal by re-splitting. (a) Nearest-training lookup above chance on the original test set and on the recommended drop-in (resplit_union where it exists, resplit otherwise) of 60 tasks, coloured by the negative control C1 with the genome-membership rule for the construction tasks (32 fail). (b) Causes of the 28 residuals of resplit versions as released, which fail C1 without the genome rule. Relational residuals are cut by resplit_union, but their lookup barely changes (all five keep constructed negatives), and a non-relational residual can carry several labels.

negative is a reference-genome window matched to its parent positive on length, GC count and repeat-content bin, more than 500 bp from any positive of the task family (for promoters, more than 1 kb from any annotated TSS), and redrawn while it is related to an item of another split. One generator version, with its sha256 recorded in each READY file, builds every fixneg version.

Licences and redistribution. GUE states no separate data licence. Its versions therefore ship as row ids into the original files, the generated negatives and the build script, which rebuilds each version from the original archive and checks the sha256 values. The Nucleotide Transformer versions store no original sequences, and a content check of the rebuilt files finds no mismatching row. The DNALongBench tables are released under CC0 1.0. Each suite has a README and a datasheet. For the GUE, Nucleotide Transformer, Genomic Benchmarks and census versions, the README gives the status of the negative control C1 (labelled E1 in the release files), the residual and the survival of every version and the rebuild instructions; the first three also list the shortcuts that re-splitting cannot remove.

Table 10 gives the strict variants and the census suite; the re-split version passes the negative control and the linter on 16 of the 17 census tasks.

Table 10: Strict construction-free variants, in which every BLAST hit is a relation (hcut cuts each cross-split pair; hclust holds out homology clusters by chromosome), registered for the four tasks with a lookup residual after the combined versions and extended to GUE prom_300_all after their results, and the census suite. Passing: negative control with every rule (census: the audit’s rules; the genome rule needs coordinates) and no linter FAIL; strict draws must also keep enough original test items, which hclust on mouse1 does not (0.34 of the lowest class). Lookup: nearest-training lookup (MCC). Survival: share of the draw’s items kept.

Task	Version	Draws passing	Lookup (MCC)	Survival
GUE mouse1	fixneg_hcut	3 of 3	0.000	0.862–0.863
GUE mouse1	fixneg_hclust	0 of 3	0.000	0.907–0.908
GUE mouse2	fixneg_hcut	3 of 3	0.000	0.976–0.984
GUE mouse2	fixneg_hclust	3 of 3	0.000	0.987–0.996
GUE prom_300_notata	fixneg_hcut	3 of 3	0.039–0.050	0.954–0.958
GUE prom_300_notata	fixneg_hclust	3 of 3	0.029–0.036	0.997–0.998
NT promoter_all	fixneg_hcut	3 of 3	0.000	0.958–0.959
NT promoter_all	fixneg_hclust	3 of 3	0.000	0.994–0.996
GUE prom_300_all	fixneg_hcut	3 of 3	0.042–0.044	0.956–0.958
GUE prom_300_all	fixneg_hclust	3 of 3	0.000	0.995–0.997
Census suite (17 tasks)	resplit	16 of 17		
Census suite (17 tasks)	purged	14 of 17		

F Fine-tuning experiment details

Registration and amendments. The design was registered on 21 October 2026, before any test-set evaluation, and every amendment is archived with the amended design. The amendments fix a transcription error in the training steps of the mouse tasks, subsample original training sets to the clean versions’ sizes, pin compared runs to one accelerator type, replace the bootstrap unit by 5-kb genomic blocks, move the confirmatory relatedness test to the purge arms (F1) with class-stratified random arms, make the construction effect its own family with a negative control, and reinstate the GUE construction tasks and the external-validity test once their clean versions were released. Twelve of them (A10–A21) respond to an independent review of the design; Table 17 lists the amendments entered after the results were first reported. The first of these (A45) corrected two implementation errors that an independent audit of the analysis found after the first results: the promoter test had filtered its test rows against one of the three fixneg draws instead of all three, and the two relatedness tests and their negative control had resampled the seeds of the two arms independently instead of in pairs; no registered decision changed. A later one (A48) made the same pairing correction for the positive control, which moves NT-v2 500M to the border of detection, and another (A47) computed the registered AUROC sensitivity analysis of the four-mark test after the MCC results were known. The same audit found that two amendments, which add the Q2 framing, the clear-null criterion and the panel-mean interval, were recorded after the four-mark runs had finished but before the test was first computed; neither changes the test, its Holm family or the criterion for its conclusion.

Protocol. Training follows the DNABERT-2 GUE scripts: batch 32, AdamW with weight decay 0.01, linear warm-up and decay, gradient clipping and bf16 autocast, with 1,000 optimiser steps on the mouse tasks. Learning rates were chosen on development data of five transcription-factor tasks only: 3×10^{-4} for Caduceus and 3×10^{-5} for the other language models. Numbers of seeds were sized from three pilot seeds per model: 3 for DNABERT-2, 4 for the CNN, 5 for NT-v2 50M and 8 for HyenaDNA, Caduceus and NT-v2 500M, with at least 5 in the purge arms. Intervals come from a two-level bootstrap over seeds and test units (5-kb genomic blocks of test items; single rows for the promoter test), with the seeds of the two arms paired in the two relatedness tests and their controls, and 10,000 replicates for confirmatory p -values. Before the tests, the panel reproduced published scores to within 0.05 for five of its six models, each on one task set with a published score (Appendix Table 3). From the pilot runs, the effects detectable with 80% power at the first to last Holm step are 0.018–0.025 AUROC for the promoter test, 0.019–0.029 for the GUE construction tasks, 0.022–0.032 for the controlled constructions and 0.015–0.025 and 0.009–0.016 MCC for the four- and ten-mark tests; for the replacement test, they are 0.016–0.023 and 0.027–0.065 at the first Holm step for the two classes.

Q1 families. In the promoter test (E4), models trained with segment-substituted negatives (resplit) or genomic negatives (fixneg) are tested on Nucleotide Transformer revised promoter rows that share no 32-mer with either training set, by AUROC and GC-stratified AUROC. For the GUE construction tasks (F5bx), the cost is the AUROC on genomic negatives of fixneg-trained minus resplit-trained models, relative to the same quantity on five transcription-factor tasks. The controlled-construction test (F3bx) measures the same cost for controlled constructions built from the same positives, which show the mechanism: shuffled and segment-substituted negatives cost +0.249 [+0.234, +0.264] and +0.414 [+0.394, +0.432] (12 of 12 tests significant), against –0.008 to +0.002 for matched genomic negatives. An independent audit recomputed the replacement test of Section 5 from the per-run predictions with its own code and reproduced every point estimate and Holm decision; with the 300-bp and the 70-bp core promoter test sets, which share their positives, resampled jointly, the interval of the promoter difference is [+0.293, +0.318], and no decision changes. The replacement test’s control contrast is –0.007 [–0.011, –0.003], inside the registered ± 0.02 margin, and on 300-bp promoter sets the difference falls from +0.336 with composition-matched negatives to +0.106 without that matching, as in the promoter test (Appendix Table 14).

Data. The clean training sets are smaller than the originals (62.9–73.7% of the original size for the transcription-factor tasks and 93.3–95.3% for the yeast tasks), hence the size-matched original versions. Purge arm P removes 6,903, 5,922, 6,121 and 2,003 training rows from the four yeast tasks

(H3K4me3, H3K14ac, H4ac and H3), and the residual 25-mer lookup on their test sets falls from 0.784, 0.837, 0.790 and 0.653 to 0.052, 0.055, 0.051 and 0.238. The external-validity test set keeps 13,843 of the 31,584 revised promoter rows (1,650 positives) after removing every row that shares a 32-mer with a training or development sequence of either arm, over all three fixneg draws. Synthetic self-tests of each family’s analysis recover planted effects and do not reject planted nulls.

Results. Table 11 gives the four-mark test (F1) per model at both checkpoints, with the negative control, and Table 12 gives the Q1 families per model. At the checkpoint chosen on development data, which is a secondary readout, the four-mark effect is smaller: 2 of 6 models are Holm-significant, and the panel mean of $+0.008$ [$+0.003$, $+0.013$] is above zero but meets the clear-null criterion (upper bound below 0.02). The CNN and HyenaDNA show no significant effect at either checkpoint.

The positive control, paired by seed as registered, detects the strong dose at the last checkpoint for Caduceus ($+0.078$ [$+0.044$, $+0.113$]) and DNABERT-2 ($+0.049$ [$+0.017$, $+0.082$]). NT-v2 500M ($+0.027$ [$+0.0003$, $+0.054$]) is borderline: its lower bound is zero within Monte Carlo error. NT-v2 50M ($+0.025$ [-0.011 , $+0.060$]) and HyenaDNA ($+0.011$ [-0.009 , $+0.031$]) are positive but not significant, and the CNN is at -0.001 . No model detects the weak dose ($+0.015$ to $+0.046$, every interval including zero), and at the checkpoint chosen on development data only Caduceus detects the strong dose. Read alongside the four-mark test, a missed positive control does not exclude inflation: NT-v2 50M shows significant inflation in that test although it misses its positive control, and NT-v2 500M shows it with a borderline positive control; the CNN and HyenaDNA are null in both tests and show no detectable exploitation of training relatives, although this does not rule such exploitation out.

In the promoter test, the CNN goes the other way in plain AUROC (-0.081 [-0.129 , -0.037]). The secondary ranking test on the purge arms finds no change (τ 1.000). No model has a significantly positive difference in differences on the segment-substituted promoter tasks, measured against the clean TATA tasks (-0.042 to $+0.029$), and on the shuffled-negative mouse tasks, which have no admitted control, original minus re-split scores are $+0.037$ to $+0.086$ for all six models (descriptive).

Table 16 gives the secondary test on the original Nucleotide Transformer test sets per model: 4 of 6 models are Holm-significant (DNABERT-2 $+0.041$ [$+0.030$, $+0.051$]), and the panel mean of $+0.015$ [$+0.009$, $+0.0203$] is not a clear null; at the development checkpoint, only 2 of 6 models reach Holm significance and the panel mean is a clear null. Table 13 gives the ten-mark test (F1x) per model; with 20-kb blocks, the relatedness tests (F1, F1x) and the construction-cost tests (F3bx, F5bx) keep their conclusions. Only DNABERT-2 scores lower at the benchmark level on the re-split than on the size-matched original histone-mark tasks (by 0.053 [0.027, 0.081]); the other five models change by -0.013 to $+0.008$. Original minus re-split scores average $+0.005$ [-0.035 , $+0.042$] on the clean TATA control and $+0.119$ on the transcription-factor tasks. A reference 3-mer classifier already loses 0.090 of this, which indicates a distribution shift between the splits. A memorisation signature, agreement with the lookup rule on test items where the rule is wrong, is as frequent in one purge arm as in the other. For H3K36me3, H3K9ac and splice sites, the published lead lies inside the inflation on the original test set. For tf1, tf4 and mouse2, the published lead lies inside the benchmark-level inflation, as it does for tf0 and tf3 once DNABERT-2’s Table 9 rows are excluded, and a reference 3-mer classifier loses 0.140, 0.116, 0.604, 0.110 and 0.082 on these tasks from the original to the re-split version.

On the original Nucleotide Transformer tasks, four leads of 0.097–0.161, all set by a multi-task model whose training includes these test sequences (Appendix G), lie far above the inflation. The Genomic Benchmarks leads on human non-TATA promoters ($+0.006$) and the enhancers of Cohn et al. ($+0.003$) lie below the lower bounds of the inflation ($+0.030$ and $+0.016$); the Ensembl enhancer task was refused because 3.2% of its test rows lack coordinates (Table 19). Table 18 gives the re-ranking by task group. Repeated descriptively with eight models, the ranking test gives τ 0.929.

The registered AUROC sensitivity analysis of the four-mark test, computed late, after the MCC results were known, finds the same 4 of 6 models Holm-significant at the last checkpoint (panel mean $+0.009$ [$+0.007$, $+0.011$] AUROC; 2 of 6 at the checkpoint chosen on development data), so the inflation does not depend on the fixed threshold of MCC; the clear-null margin of 0.02 is defined for MCC and is not applied here.

Linear probes on frozen large models. A further pre-registered test asks whether the cost of constructed negatives arises without fine-tuning. Three large models from the leaderboard, GenomeOcean 4B, GENERator 3B and Evo 2 7B, embed each sequence once, as the mean over its tokens at the leaderboard’s registered layer, chosen on bacterial gene-pair data (layer 9, layer 14 and block 10; Appendix H). On each train split, an L2-penalised logistic regression is fitted to the standardised embeddings, with its penalty chosen by 5-fold inner cross-validation on folds that are fixed in advance and shared by every model. The cost and its control are those of the GUE construction test (F5bx) on the benchmarks’ genomic-negative test sets and those of the replacement test on the independent G+C-matched sets. The cost is present for a model and class if the lower 95% bound of its DiD is above zero and the test is Holm-significant among the 6 tests of its set. It is present in all 12 tests (Table 15), with panel means of +0.192 and +0.397 on the genomic-negative test sets and +0.186 and +0.273 on the independent sets. The negative control lies inside ± 0.02 for every model and set (-0.011 to $+0.008$); three of the six intervals exclude zero (Evo 2 on both test sets, GENERator on the genomic-negative sets), much like the replacement test’s -0.007 . Both planted constructions are detected everywhere, and probes refitted on folds drawn with seeds 1 and 2 give the same decisions. At the final layer, the secondary analysis, the cost is again present in all 12 tests, and every control passes except Evo 2’s negative control on the independent sets (-0.025 [-0.031 , -0.019]); as planned, Table 15 shows Evo 2’s plain costs on those sets alongside the DiD (+0.196 and +0.178, against DiDs of +0.221 and +0.204). On 300-bp promoter sets whose negatives are not matched on composition, the cost is only +0.017 to +0.035, against +0.292 to +0.300 with matched negatives, the composition masking seen in the promoter test (Section 5). The run logs of this appendix, among them the bootstrap checks of the positive control and the fitting diagnostics of every probe, are in the archive at <https://doi.org/10.5281/zenodo.23717097>.

Table 11: Four-mark test (F1) per model: $S(R) - S(P)$ in test MCC, mean over the four GUE yeast histone-mark tasks, with 95% intervals and Holm-corrected p -values, at the last checkpoint (registered) and at the checkpoint chosen on development data (secondary), and the negative control $S(R_a) - S(R_b)$. A Holm p of zero means that no bootstrap replicate fell at or below zero; it is shown as below m/B for m models and B replicates: <0.0006 at the last checkpoint ($B = 10,000$) and <0.003 at the development checkpoint ($B = 2,000$).

Model	last checkpoint	Holm p	development checkpoint	Holm p	negative control
CNN	-0.0027 [-0.0231 , $+0.0176$]	1.0000	$+0.0015$ [-0.0135 , $+0.0165$]	0.9070	$+0.0055$ [-0.0242 , $+0.0351$]
HyenaDNA	-0.0014 [-0.0122 , $+0.0101$]	1.0000	-0.0011 [-0.0125 , $+0.0103$]	0.9070	-0.0137 [-0.0450 , $+0.0131$]
Caduceus	$+0.0152$ [$+0.0052$, $+0.0251$]	0.0060	$+0.0130$ [$+0.0031$, $+0.0229$]	0.0325	-0.0069 [-0.0325 , $+0.0168$]
DNABERT-2	$+0.0523$ [$+0.0411$, $+0.0634$]	<0.0006	$+0.0288$ [$+0.0146$, $+0.0437$]	<0.003	-0.0055 [-0.0314 , $+0.0196$]
NT-v2 50M	$+0.0162$ [$+0.0050$, $+0.0276$]	0.0069	$+0.0036$ [-0.0078 , $+0.0154$]	0.8220	-0.0007 [-0.0310 , $+0.0283$]
NT-v2 500M	$+0.0179$ [$+0.0087$, $+0.0274$]	0.0005	$+0.0036$ [-0.0051 , $+0.0124$]	0.8220	-0.0228 [-0.0484 , $+0.0028$]
Panel mean	$+0.0162$ [$+0.0109$, $+0.0215$]	–	$+0.0082$ [$+0.0031$, $+0.0133$]	–	-0.0073 [-0.0192 , $+0.0038$]

Table 12: Q1 families per model. F5bx (GUE construction tasks) and F3bx (controlled constructions): AUROC lost on test sets with genomic negatives by training on shuffled or segment-substituted negatives. TF ctrl.: the same contrast on five transcription-factor tasks; genomic: training on matched genomic negatives instead of the task’s own. The registered statistic is the cost minus its control, Holm-significant in all 12 tests of each family. E4: fixneg-trained minus resplit-trained AUROC on the independent promoter test set, plain and within GC strata; *, Holm-significant.

Model	F5bx:			F3bx:			E4:	
	shuffled	segment	TF ctrl.	shuffled	segment	genomic	AUROC	GC strata
CNN	+0.186	+0.383	–0.003	+0.197	+0.343	–0.005	–0.081	+0.146*
HyenaDNA	+0.149	+0.404	+0.008	+0.194	+0.398	–0.008	+0.015	+0.277*
Caduceus	+0.123	+0.418	+0.001	+0.294	+0.439	+0.001	+0.022	+0.291*
DNABERT-2	+0.152	+0.392	+0.006	+0.270	+0.432	+0.002	+0.049	+0.269*
NT-v2 50M	+0.177	+0.455	+0.005	+0.251	+0.405	–0.001	+0.114*	+0.284*
NT-v2 500M	+0.140	+0.420	+0.005	+0.275	+0.455	+0.000	+0.141*	+0.329*

Table 13: Ten-mark test (F1x) per model: $S(R) - S(P)$ in test MCC, mean over the ten GUE yeast histone-mark tasks, with 95% intervals and Holm-corrected p -values, at the last checkpoint (registered) and at the checkpoint chosen on development data (secondary); a Holm p of zero is shown as below m/B , as in Table 11 (<0.0009 and <0.0045 for the nine models).

Model	last checkpoint	Holm p	development checkpoint	Holm p
CNN	+0.0013 [−0.0127, +0.0164]	0.4424	+0.0040 [−0.0045, +0.0121]	0.1815
HyenaDNA	+0.0048 [−0.0011, +0.0106]	0.1116	+0.0047 [−0.0015, +0.0104]	0.1310
Caduceus	+0.0154 [+0.0100, +0.0208]	<0.0009	+0.0140 [+0.0089, +0.0195]	<0.0045
DNABERT-2	+0.0409 [+0.0335, +0.0484]	<0.0009	+0.0238 [+0.0160, +0.0317]	<0.0045
NT-v2 50M	+0.0152 [+0.0084, +0.0223]	<0.0009	+0.0070 [−0.0002, +0.0142]	0.1220
NT-v2 500M	+0.0181 [+0.0121, +0.0241]	<0.0009	+0.0079 [+0.0023, +0.0137]	0.0210
DNABERT-6mer	+0.0106 [+0.0046, +0.0168]	0.0010	+0.0097 [+0.0037, +0.0155]	<0.0045
NT-500M-human	+0.0056 [−0.0004, +0.0112]	0.0993	+0.0053 [−0.0005, +0.0112]	0.1220
NT-2500M-multi	+0.0067 [+0.0017, +0.0116]	0.0160	+0.0061 [+0.0011, +0.0115]	0.0400
Panel mean	+0.0132 [+0.0106, +0.0159]	–	+0.0092 [+0.0069, +0.0115]	–

Table 14: Replacement external-validity test. AUROC of models trained on genomic negatives minus models trained on constructed negatives, on independent test sets with G+C-matched negatives, minus the same difference on five human transcription-factor tasks (control column); best checkpoint unless stated; 95% intervals.

Model	Binding sites	Promoters	Control
CNN	+0.125 [+0.112, +0.137]	+0.245 [+0.229, +0.260]	−0.003 [−0.011, +0.004]
HyenaDNA	+0.113 [+0.103, +0.124]	+0.288 [+0.275, +0.301]	+0.003 [−0.003, +0.009]
Caduceus	+0.148 [+0.133, +0.163]	+0.312 [+0.294, +0.330]	−0.013 [−0.023, −0.003]
DNABERT-2	+0.168 [+0.151, +0.186]	+0.295 [+0.278, +0.313]	−0.017 [−0.026, −0.008]
NT-v2 50M	+0.131 [+0.119, +0.143]	+0.332 [+0.313, +0.350]	+0.000 [−0.008, +0.009]
NT-v2 500M	+0.166 [+0.153, +0.179]	+0.364 [+0.336, +0.387]	−0.012 [−0.021, −0.004]
Panel mean	+0.142 [+0.135, +0.149]	+0.306 [+0.295, +0.317]	−0.007 [−0.011, −0.003]
Panel mean, last checkpoint	+0.132 [+0.125, +0.138]	+0.288 [+0.277, +0.298]	−0.0003 [−0.004, +0.003]
Panel mean, within GC strata	+0.143 [+0.136, +0.150]	+0.303 [+0.293, +0.313]	−0.008 [−0.012, −0.004]
Panel mean, within composition strata	+0.120 [+0.113, +0.128]	+0.308 [+0.297, +0.319]	−0.001 [−0.004, +0.003]
Planted constructions, shuffled / segment	+0.172 [+0.164, +0.179] / +0.242 [+0.233, +0.251]		
Promoter cost (300-bp sets), matched / unmatched negatives	+0.336 [+0.320, +0.351] / +0.106 [+0.091, +0.120]		

Table 15: Linear probes on frozen embeddings of three large models (registered test): DiD of the cost of constructed negatives (AUROC, 95% intervals) on the benchmarks’ genomic-negative test sets and on the independent G+C-matched sets, for dinucleotide-shuffled negatives (GUE mouse tasks) and segment substitution (GUE promoter tasks). Negative control: the difference between two realistic negative sets on the five transcription-factor tasks (pass: interval inside ± 0.02). Planted constructions: shuffled and segment-substituted negatives built from the same positives (pass: lower bound above zero; all pass). Layer: the leaderboard’s registered layer (R1; a block for Evo 2) and, as the secondary analysis, the final layer. Every DiD test is Holm-significant, with no bootstrap replicate at or below zero. [†]Negative control fails; as planned, the row below gives the plain costs, without the control.

Model	Layer	Genomic-negative test sets		Independent sets		Negative control		Planted constructions	
		Shuffled	Segment	Shuffled	Segment	Genomic-negative	Independent	Genomic-negative	Independent
<i>Registered layer (primary)</i>									
Evo 2 7B	10	+0.156 [+0.136, +0.176]	+0.359 [+0.345, +0.372]	+0.220 [+0.212, +0.228]	+0.297 [+0.285, +0.309]	+0.008 [+0.003, +0.013]	−0.011 [−0.014, −0.009]	+0.346, +0.471	+0.266, +0.330
GenomeOcean 4B	9	+0.196 [+0.170, +0.222]	+0.426 [+0.411, +0.441]	+0.174 [+0.165, +0.183]	+0.269 [+0.256, +0.282]	+0.004 [−0.001, +0.008]	+0.001 [−0.002, +0.004]	+0.384, +0.506	+0.263, +0.324
GENERator 3B	14	+0.225 [+0.196, +0.253]	+0.407 [+0.392, +0.422]	+0.165 [+0.156, +0.174]	+0.252 [+0.239, +0.265]	+0.006 [+0.001, +0.012]	+0.003 [−0.001, +0.006]	+0.265, +0.390	+0.144, +0.206
Panel mean		+0.192 [+0.174, +0.210]	+0.397 [+0.385, +0.410]	+0.186 [+0.179, +0.193]	+0.273 [+0.261, +0.284]	+0.006 [+0.003, +0.009]	−0.003 [−0.005, −0.0005]	+0.332, +0.456	+0.224, +0.286
<i>Final layer (registered secondary)</i>									
Evo 2 7B	final	+0.149 [+0.120, +0.179]	+0.339 [+0.323, +0.355]	+0.221 [+0.209, +0.232]	+0.204 [+0.189, +0.218]	+0.007 [+0.002, +0.013]	−0.025 [−0.031, −0.019] [†]	+0.382, +0.429	+0.257, +0.273
plain cost [†]				+0.196 [+0.186, +0.205]	+0.178 [+0.165, +0.191]				
GenomeOcean 4B	final	+0.208 [+0.182, +0.234]	+0.409 [+0.395, +0.424]	+0.169 [+0.160, +0.177]	+0.276 [+0.263, +0.288]	+0.004 [−0.001, +0.009]	−0.002 [−0.005, +0.002]	+0.382, +0.533	+0.261, +0.348
GENERator 3B	final	+0.238 [+0.209, +0.265]	+0.392 [+0.376, +0.407]	+0.174 [+0.166, +0.182]	+0.243 [+0.230, +0.256]	+0.008 [+0.002, +0.014]	−0.0003 [−0.004, +0.003]	+0.265, +0.375	+0.139, +0.205
Panel mean		+0.198 [+0.181, +0.217]	+0.380 [+0.367, +0.393]	+0.188 [+0.181, +0.195]	+0.241 [+0.229, +0.252]	+0.006 [+0.003, +0.010]	−0.009 [−0.012, −0.006]	+0.343, +0.446	+0.219, +0.275

G Reported progress: details

Row sets and the audit. The audit recomputed the shares, intervals, counts and wordings of the registered set P with its own code and found no mismatches. It also found rows whose validity fails on evidence about the rows themselves: multi-task models whose training includes the test inputs (28,732 of 28,741 test sequences of the original Nucleotide Transformer histone tasks are in the training split of another histone task), the transcription-factor and mouse rows of DNABERT-2’s Table 9, which contradict its Table 6 for the same model, protocol and split by 7.9–24.8 MCC points, and maxima over runs scored on the test set. P' is P without these rows, and the audit’s own computation of P' gives the same shares. Each correction raises the shares above the ceiling.

Table 16: NT1 per model: $S(R) - S(P)$ in test MCC on the original Nucleotide Transformer test sets of the ten histone marks, with 95% intervals and Holm-corrected p (a p of zero shown as below m/B), at the last and at the development checkpoint.

Model	last checkpoint	Holm p	development checkpoint	Holm p
CNN	+0.0017 [−0.0272, +0.0255]	0.4319	+0.0054 [−0.0102, +0.0194]	0.7200
HyenaDNA	+0.0046 [−0.0034, +0.0125]	0.2442	+0.0047 [−0.0031, +0.0126]	0.5000
Caduceus	+0.0124 [+0.0044, +0.0204]	0.0032	+0.0119 [+0.0039, +0.0203]	0.0050
DNABERT-2	+0.0405 [+0.0299, +0.0515]	<0.0006	+0.0197 [+0.0096, +0.0310]	<0.003
NT-v2 50M	+0.0153 [+0.0075, +0.0234]	<0.0006	−0.0006 [−0.0093, +0.0083]	1.0000
NT-v2 500M	+0.0143 [+0.0034, +0.0249]	0.0177	−0.0002 [−0.0109, +0.0100]	1.0000
Panel mean	+0.0148 [+0.0089, +0.0203]	–	+0.0068 [+0.0023, +0.0109]	–

Table 17: Amendments entered after the results were first reported, in the order they were entered; A38–A43 followed the first four-mark result but each preceded every result it could affect.

Amendment	Subject, as logged
A45	implementation corrections from the independent audit
A44	NT-original arm
A46	NT-2500M-multi GUE re-scoring restored
A47	A23’s F1 AUROC sensitivity computed late
A48	implementation correction: PC1 paired by seed

Table 18: Kendall’s τ between published, original-split and cleaned rankings by GUE task group, with NT-2500M-multi re-scored.

Group	Models	τ published, original	τ published, cleaned	τ original, cleaned	Pairs that swap
Mean over 18 tasks	–	+0.787	+0.728	+0.819	–
Histone marks (four)	5	+1.000	+1.000	+1.000	none
Human transcription factors	7	+0.905	+0.905	+0.810	Caduceus/DNABERT-6mer
Mouse transcription factors	6	+0.867	+0.867	+1.000	Caduceus/HyenaDNA
Promoters	6	+0.200	+0.200	+0.733	6

Sensitivity variants and controls. The verdict wording was fixed in the pre-registration: most of a suite’s improvement lies at or below the ceiling when the upper bound of its share’s interval is below 0.5, above it when the lower bound is above 0.5, and the verdict is undetermined otherwise. In GUE and in the original Nucleotide Transformer tasks, the shares above the ceiling are 0.247 [0.228, 0.263] and 0.318 [0.270, 0.345] in P, 0.356 [0.317, 0.402] and 0.365 [0.310, 0.395] in P’, 0.365 [0.327, 0.410] and 0.365 [0.310, 0.395] in P’ with the preprints that missed the freeze, and 0.331 [0.295, 0.376] and 0.357 [0.301, 0.388] in P’ with stricter row criteria. Most of the improvement lies at or below the ceiling in each. Adding the rows of USM-AA, whose inputs include genome coordinates, makes GUE undetermined (0.490 [0.459, 0.527] with these rows added to P’ and the late preprints), and the registered rules do not count such rows as comparable. Genomic Benchmarks keeps most of its improvement above the ceiling (0.862 [0.858, 0.866] in P’). For the four suites pooled, the share is 0.462 [0.441, 0.475] in P, 0.503 [0.478, 0.523] in P’ and 0.529 [0.506, 0.548] with the late preprints, so it moves from at or below the ceiling, through undetermined, to above it. When tasks are resampled, the intervals for GUE and the original Nucleotide Transformer tasks reach 0.593 and 0.601. Only the splice tasks of the revised Nucleotide Transformer set have a documented leak, and the ceiling of the others is chance, so the contrast of the revised with the original tasks shows what removing the leaks does to the bar rather than better models. On 20 tasks without a documented leak the share is 1.000 (negative control), and replacing each ceiling by the same rule on the leak-reduced versions raises the share to 1.000 on GUE, on Genomic Benchmarks and on the original Nucleotide Transformer tasks. Rows of models with coordinate inputs are kept out of the analysis.

Scores over time. At the last date, the best published score exceeds the leak rule on 9 of 21 GUE and 10 of 18 original Nucleotide Transformer tasks with a documented leak (medians 0.982 and 1.068 in units of the rule; Figure 2a).

Ranking survival. The ranking analysis uses the 930 admitted complete cells of the fine-tuning experiment at the checkpoint chosen on development data, with the last checkpoint as a sensitivity

Table 19: Published margins of the Genomic Benchmarks and original Nucleotide Transformer tasks against the panel-mean inflation (95% intervals); inside: margin below the lower bound; [†], lead set by a multi-task model whose training includes these test sequences (Appendix G).

Suite	Task	Lead (best family over the next)	Margin	Inflation	Inside
Genomic Benchmarks	human_nontata_promoters	FinDNA over Omni-DNA	+0.0059	+0.0506 [+0.0301, +0.0717]	yes
Genomic Benchmarks	human_enhancers_cohn	Life-Code over MergeDNA	+0.0031	+0.0273 [+0.0157, +0.0396]	yes
Genomic Benchmarks	human_enhancers_ensembl	Life-Code over FinDNA	+0.0019	refused	—
Genomic Benchmarks	human_ocr_ensembl	Caduceus over DNA-xLSTM	+0.0020	−0.0060 [−0.0139, +0.0015]	no
NT original	H3	JanusDNA over AIDO.DNA	+0.0029	+0.0174 [+0.0048, +0.0294]	yes
NT original	H3K14ac	Omni-DNA [†] over JanusDNA	+0.1610	+0.0116 [+0.0007, +0.0227]	no
NT original	H3K36me3	JanusDNA over Omni-DNA	+0.0130	+0.0061 [−0.0047, +0.0174]	no
NT original	H3K4me1	Omni-DNA over JanusDNA	+0.0010	+0.0022 [−0.0113, +0.0154]	no
NT original	H3K4me2	Omni-DNA [†] over JanusDNA	+0.1020	−0.0031 [−0.0182, +0.0113]	no
NT original	H3K4me3	Omni-DNA [†] over JanusDNA	+0.0970	−0.0001 [−0.0235, +0.0219]	no
NT original	H3K79me3	Omni-DNA over JanusDNA	+0.0050	+0.0106 [+0.0017, +0.0191]	no
NT original	H3K9ac	Omni-DNA over ConvNova	+0.0200	+0.0071 [−0.0046, +0.0197]	no
NT original	H4	AIDO.DNA over NT	+0.0135	+0.0094 [−0.0003, +0.0192]	no
NT original	H4ac	Omni-DNA [†] over JanusDNA	+0.1480	+0.0068 [−0.0084, +0.0212]	no

readout. For the re-split versions, the mean Kendall’s τ is 0.830 (median 0.857) on 13 tasks with a documented leak and 0.794 on 9 tasks without one, against seed split-half reliabilities of 0.866 and 0.789, and the top model is kept on 10 of 13 and 7 of 9; at the last checkpoint the means are 0.767 and 0.810, against 0.881 and 0.860, with the top model kept on 7 of 13 and 8 of 9. On 7 of the 13 tasks with a documented leak, the re-split version keeps the constructed negatives. Where fixneg replaces them, the means are 0.752 on 7 tasks with a documented leak and 0.787 on 5 without, against split-half reliabilities of 0.878 and 0.797, with the top model kept on 5 of 7 and 2 of 5. Three tasks fall more than 0.3 below their split-half reliability: the re-split human non-TATA promoters of Genomic Benchmarks (τ 0.200 against 0.839), where the nearest-training lookup scores 0.991 on the original split, and the fixneg versions of GUE mouse2 (0.467 against 0.777) and prom_core_all (0.333 against 0.844). The pre-registered convergent check, a positive Spearman correlation over the 13 tasks with a documented leak between the gap to the leak ceiling and the measured inflation, is not supported: 0.099 [−0.518, 0.771] in P' and 0.220 [−0.367, 0.779] in P .

Genome membership as a shortcut (post hoc, descriptive). Constructed negatives, such as dinucleotide shuffles and segment-substituted copies, do not occur in the reference genome, whereas the positives do. A no-model rule that calls a test sequence positive when it aligns over its full length to the reference genome reaches MCC 1.000 on the three promoter tasks of the original Nucleotide Transformer set and on four GUE tasks (mouse0, mouse3, mouse4 and the 300-bp non-TATA promoters), above every published score of a sequence-only model on these tasks. On the versions of the 12 construction tasks with genomic negatives (fixneg, 36 draws) the rule falls to chance (MCC at most 0.0127), whereas every released version that keeps the constructed negatives scores 0.8925–1.000. USM-AA, a published model whose inputs include each token’s chromosome, strand and comparison against the reference [36], reports 1.000 and 0.9974 on GUE’s 300-bp and 70-bp non-TATA promoter tasks, where the alignment rule scores 1.000 and 0.9985. Its scores there match what the rule reaches without a model; on the 300-bp and 70-bp promoter tasks with TATA and non-TATA promoters together it reports 0.9593 and 0.9603, above the rule’s 0.8986 and 0.9008.

Table 20 lists the primary construction-task claims of the published-wins test. Because the negative control of this test failed for the aligned-input model, a second test, registered before its data were built, repeated the 19 reproduced claims with new negatives: for each positive, a genome window of the same length 2–10 kb away on the same chromosome, the closest of eight candidates in G+C count (Table 21). By the decision rule, most of the 19 wins survive (11). For the aligned-input model, all 6 wins vanish or reverse with a Holm-significant change. Of the 13 sequence-only wins, 11 survive and 2 vanish, and the change is Holm-significant for one of the two. The procedural control, which compares the claimants with DNABERT-2 on the five human transcription-factor tasks, passes (pooled change +0.004 [−0.011, +0.020]), and so does the planted positive control. At the last checkpoint, 8 of the 19 vanish or reverse with a Holm-significant change and 7 survive, which the rule classifies as mixed. An independent recomputation from the per-run outputs reproduced the point estimates exactly, the same seven Holm-significant vanishings and reversals, and the passing

control. In the recomputation, drawing seed picks jointly across the control pairs widens the control interval to $[-0.018, +0.025]$, which passes only because it contains zero. With a registered extension to 6 seeds per arm, the three thinnest survivors all still survive: the advantage of C24 on the original version is $+0.029 [+0.005, +0.054]$, the change of C17 is $-0.046 [-0.115, +0.020]$ (Holm p 0.94), and the advantage of C28 on the original version is $+0.021 [+0.006, +0.037]$. The count stays at 11 of 19, so the verdict stands. At the last checkpoint, survivors rise from 7 to 8 of the 19, still a mixed result. G+C matching is loose for these negatives: the mean absolute difference from the positive is 0.06–0.10 of the sequence length on the promoter tasks and 0.02–0.09 on the mouse tasks. The negatives follow the recipe of a post hoc check on the five human transcription-factor tasks, in which the aligned-input model’s advantage over DNABERT-2 fell from $+0.152 [+0.130, +0.174]$ to $-0.014 [-0.032, +0.003]$ MCC. The second test drew new negatives, its 19 claims were not part of that check, and its control includes the aligned-input model on the same tasks.

Table 20: Published wins on construction tasks: claimant minus comparator (MCC, mean over seeds, 95% intervals) on the original and the construction-free versions, with the registered category; H1: Holm-significant change.

Claim	Task	Claimant vs comparator	Published	Original version	Construction-free version	Category
C01	GUE prom_300_all	USM-AA vs NT-2500M-multi	+0.047	+0.031 [+0.020, +0.043]	−0.057 [−0.074, −0.039]	reverses (H1)
C02	GUE prom_300_notata	USM-AA vs DNABERT-2	+0.057	+0.062 [+0.050, +0.074]	+0.055 [+0.031, +0.078]	survives
C03	GUE prom_core_all	USM-AA vs NT-2500M-multi	+0.257	+0.246 [+0.226, +0.267]	−0.070 [−0.088, −0.051]	reverses (H1)
C04	GUE prom_core_notata	USM-AA vs NT-2500M-multi	+0.282	+0.280 [+0.257, +0.304]	−0.086 [−0.108, −0.066]	reverses (H1)
C05	GUE mouse1	USM-AA vs DNABERT-2	+0.013	−0.002 [−0.013, +0.008]	+0.007 [−0.003, +0.018]	not reproduced
C06	GUE prom_300_notata	USM-AA vs GenoJEPa-B	+0.040	+0.051 [+0.042, +0.060]	+0.006 [−0.011, +0.023]	vanishes (H1)
C07	GUE prom_core_notata	USM-AA vs GenoJEPa-B	+0.272	+0.264 [+0.245, +0.283]	−0.077 [−0.113, −0.047]	reverses (H1)
C08	GUE prom_300_notata	GenoJEPa-B vs DNABERT-2	+0.040	+0.011 [−0.003, +0.026]	+0.049 [+0.027, +0.070]	not reproduced
C09	GUE prom_core_notata	GenoJEPa-B vs DNABERT-2	+0.013	+0.035 [+0.020, +0.050]	+0.140 [+0.119, +0.162]	survives
C10	GUE mouse1	GenoJEPa-B vs NT-500M	+0.011	−0.008 [−0.017, +0.002]	−0.003 [−0.016, +0.010]	not reproduced
C11	GUE mouse2	GenoJEPa-B vs DNABERT-2	+0.007	−0.038 [−0.103, +0.026]	−0.085 [−0.189, +0.020]	not reproduced
C12	GUE mouse4	GenoJEPa-B vs DNABERT-2	+0.024	+0.010 [−0.024, +0.044]	+0.018 [−0.014, +0.050]	not reproduced
C13	GUE prom_300_all	HybridNA-300M vs Caduceus-Ph	+0.050	+0.004 [−0.014, +0.021]	+0.032 [+0.014, +0.050]	not reproduced
C14	GUE prom_300_notata	HybridNA-300M vs Caduceus-Ph	+0.023	+0.025 [+0.016, +0.036]	+0.033 [+0.016, +0.049]	survives
C15	GUE prom_core_all	HybridNA-300M vs HyenaDNA	+0.022	+0.014 [−0.011, +0.040]	+0.033 [−0.001, +0.064]	not reproduced
C16	GUE prom_core_notata	HybridNA-300M vs Caduceus-Ph	+0.008	+0.014 [−0.006, +0.032]	+0.060 [+0.028, +0.090]	not reproduced
C17	GUE mouse0	HybridNA-300M vs HyenaDNA	+0.123	+0.159 [+0.096, +0.221]	+0.087 [+0.033, +0.139]	survives
C18	GUE mouse1	HybridNA-300M vs Caduceus-Ph	+0.008	+0.038 [+0.023, +0.055]	+0.029 [+0.016, +0.042]	survives
C19	GUE mouse2	HybridNA-300M vs HyenaDNA	+0.079	+0.051 [−0.030, +0.135]	−0.050 [−0.209, +0.091]	not reproduced
C20	GUE mouse3	HybridNA-300M vs Caduceus-Ph	+0.268	−0.011 [−0.083, +0.064]	−0.076 [−0.198, +0.038]	not reproduced
C21	GUE mouse4	HybridNA-300M vs HyenaDNA	+0.043	+0.115 [+0.076, +0.155]	+0.043 [+0.010, +0.078]	survives
C22	NT promoter_all	GENERator vs Caduceus-PS	+0.021	+0.017 [+0.009, +0.024]	+0.077 [+0.056, +0.097]	survives
C23	NT promoter_no_tata	GENERator vs Caduceus-PS	+0.022	+0.013 [−0.001, +0.026]	+0.054 [+0.032, +0.076]	not reproduced
C24	NT promoter_tata	GENERator vs Caduceus-PS	+0.045	+0.029 [+0.000, +0.058]	+0.131 [+0.032, +0.220]	survives
C25	GUE prom_300_all	GROVER vs TF-IDF 3-mer	+0.102	+0.093 [+0.076, +0.111]	+0.150 [+0.129, +0.171]	survives
C26	GUE prom_300_notata	GROVER vs TF-IDF 3-mer	+0.072	+0.070 [+0.053, +0.087]	+0.118 [+0.091, +0.147]	survives
C27	GUE prom_core_all	GROVER vs TF-IDF 2-mer	+0.040	+0.043 [+0.020, +0.065]	+0.149 [+0.120, +0.178]	survives
C28	GUE prom_core_notata	GROVER vs TF-IDF 2-mer	+0.024	+0.023 [+0.004, +0.041]	+0.102 [+0.075, +0.129]	survives
C29	NT promoter_all	GROVER vs TF-IDF 3-mer	+0.065	+0.066 [+0.053, +0.080]	+0.142 [+0.116, +0.168]	survives
C30	NT promoter_no_tata	GROVER vs TF-IDF 3-mer	+0.066	+0.062 [+0.047, +0.077]	+0.115 [+0.091, +0.140]	survives
C31	NT promoter_tata	GROVER vs TF-IDF 3-mer	+0.016	+0.028 [−0.013, +0.070]	+0.362 [+0.299, +0.429]	not reproduced

Table 21: Second published-wins test (registered): claimant minus comparator (MCC, mean over seeds, 95% intervals) on the original version and with position-matched genomic negatives, the change, and its Holm-corrected p over the 19 primary claims (0.0000: no bootstrap replicate reached zero); the three secondary claims (b) are outside the Holm family. H1: vanishes or reverses with a Holm-significant change.

Claim	Task	Claimant vs comparator	Original version	Position-matched negatives	Change	Holm p	Category
C01	GUE prom_300_all	USM-AA vs NT-2500M-multi	+0.031 [+0.020, +0.043]	−0.092 [−0.124, −0.064]	−0.123 [−0.158, −0.094]	0.0000	reverses (H1)
C02	GUE prom_300_notata	USM-AA vs DNABERT-2	+0.062 [+0.050, +0.074]	−0.021 [−0.045, +0.007]	−0.083 [−0.111, −0.052]	0.0000	vanishes (H1)
C03	GUE prom_core_all	USM-AA vs NT-2500M-multi	+0.246 [+0.226, +0.267]	−0.090 [−0.107, −0.072]	−0.336 [−0.363, −0.309]	0.0000	reverses (H1)
C04	GUE prom_core_notata	USM-AA vs NT-2500M-multi	+0.280 [+0.257, +0.304]	−0.098 [−0.120, −0.076]	−0.378 [−0.411, −0.347]	0.0000	reverses (H1)
C01b	GUE prom_300_all	USM-AA vs DNABERT-2	+0.082 [+0.066, +0.097]	−0.025 [−0.047, −0.003]	−0.107 [−0.133, −0.081]	–	reverses (secondary)
C03b	GUE prom_core_all	USM-AA vs DNABERT-2	+0.272 [+0.244, +0.301]	+0.007 [−0.019, +0.035]	−0.265 [−0.304, −0.225]	–	vanishes (secondary)
C04b	GUE prom_core_notata	USM-AA vs DNABERT-2	+0.299 [+0.280, +0.318]	+0.020 [−0.004, +0.046]	−0.279 [−0.310, −0.248]	–	vanishes (secondary)
C06	GUE prom_300_notata	USM-AA vs GenoJEPa-B	+0.051 [+0.042, +0.060]	−0.044 [−0.065, −0.022]	−0.094 [−0.118, −0.071]	0.0000	reverses (H1)
C07	GUE prom_core_notata	USM-AA vs GenoJEPa-B	+0.264 [+0.245, +0.283]	−0.075 [−0.095, −0.055]	−0.339 [−0.366, −0.312]	0.0000	reverses (H1)
C09	GUE prom_core_notata	GenoJEPa-B vs DNABERT-2	+0.035 [+0.020, +0.050]	+0.095 [+0.075, +0.114]	+0.060 [+0.036, +0.084]	1.0000	survives
C14	GUE prom_300_notata	HybridNA-300M vs Caduceus-Ph	+0.025 [+0.016, +0.036]	+0.019 [+0.005, +0.033]	−0.006 [−0.023, +0.010]	1.0000	survives
C17	GUE mouse0	HybridNA-300M vs HyenaDNA	+0.159 [+0.096, +0.221]	+0.066 [+0.026, +0.106]	−0.093 [−0.169, −0.019]	0.0979	survives
C18	GUE mouse1	HybridNA-300M vs Caduceus-Ph	+0.038 [+0.023, +0.055]	−0.013 [−0.045, +0.024]	−0.051 [−0.088, −0.010]	0.0912	vanishes
C21	GUE mouse4	HybridNA-300M vs HyenaDNA	+0.115 [+0.076, +0.155]	+0.019 [−0.008, +0.046]	−0.097 [−0.144, −0.049]	0.0013	vanishes (H1)
C22	NT promoter_all	GENERator vs Caduceus-PS	+0.017 [+0.009, +0.024]	+0.043 [+0.028, +0.057]	+0.026 [+0.010, +0.043]	1.0000	survives
C24	NT promoter_tata	GENERator vs Caduceus-PS	+0.029 [+0.000, +0.058]	+0.139 [+0.084, +0.194]	+0.110 [+0.049, +0.170]	1.0000	survives
C25	GUE prom_300_all	GROVER vs TF-IDF 3-mer	+0.093 [+0.076, +0.111]	+0.082 [+0.060, +0.103]	−0.011 [−0.040, +0.016]	1.0000	survives
C26	GUE prom_300_notata	GROVER vs TF-IDF 3-mer	+0.070 [+0.053, +0.087]	+0.067 [+0.050, +0.085]	−0.003 [−0.027, +0.021]	1.0000	survives
C27	GUE prom_core_all	GROVER vs TF-IDF 2-mer	+0.043 [+0.020, +0.065]	+0.127 [+0.104, +0.152]	+0.085 [+0.052, +0.117]	1.0000	survives
C28	GUE prom_core_notata	GROVER vs TF-IDF 2-mer	+0.023 [+0.004, +0.041]	+0.090 [+0.062, +0.117]	+0.067 [+0.034, +0.101]	1.0000	survives
C29	NT promoter_all	GROVER vs TF-IDF 3-mer	+0.066 [+0.053, +0.080]	+0.100 [+0.081, +0.119]	+0.034 [+0.011, +0.057]	1.0000	survives
C30	NT promoter_no_tata	GROVER vs TF-IDF 3-mer	+0.062 [+0.047, +0.077]	+0.069 [+0.052, +0.087]	+0.007 [−0.016, +0.030]	1.0000	survives

H PairCtrl details

Tasks. *ppi_exp* has 1,971,014 gene pairs with labels from 17 experimental, curated and structure-screen data sets of six genomes. The strict subset keeps pairs more than 20 genes apart and outside known operons, with degree-preserving negatives, no ribosomal proteins and no STRING neighbourhood, fusion or co-occurrence evidence [37]. The primary metric is the strict-subset AUROC, averaged over the data sets of a class with at least 25 strict positives, and the overall mean leaves out the structure screen. *operon_expr* has 12,636 pairs from nanopore direct RNA and Cappable-seq data, with negatives only between expressed genes, and its primary metric is the gain over the position model. *operon_regulondb_exp* has 4,314 pairs labelled by experimentally supported RegulonDB units, through a join script. *etgp_nodist* has 5,534 enhancer-gene pairs from CRISPRi screens with distance-free inputs and folds that hold out loci. One engine scores every task: AUROC, AUROC within position strata, the gain of a position-plus-score model over the out-of-fold position model on frozen folds, the share interval and AUPRC. Intervals come from 1,000 bootstrap replicates shared by all scorers of a task, two-way over gene blocks for pair tasks and over loci for the enhancer task. The engine reproduces the earlier strict-subset AUROCs and intervals exactly (102 comparisons). The genomes of the pair tasks are unrelated to each other (Mash distance at least 0.30). The licence audit and the datasheet are summarised in Appendix J.

The R1 readout. A pre-registered rule chose, for each model, the layer and normalisation with the best within-position AUROC on 17 STRING development genomes and evaluated it once on 36 held-out genomes. Seven models have a registered R1 choice (Evo 2 1B and the models added to the leaderboard except Evo 2 40B); for each, the choice is a layer standardised within the genome, at 0.34–0.67 of the model’s depth (block 10 for Evo 2 1B and 7B). On the held-out genomes, the R1 readouts gain +0.0433 (GENERator 3B) to +0.0910 (Evo 2 7B) over position, whereas raw cosines and final layers carry little signal beyond position. Evo 2 40B was swept over all 50 blocks on the pairs whose genes both lie in a fixed tenth of the genes (167,090 items: 17 development and 36 held-out genomes). The rule picks block 12, standardised (development within-position AUROC 0.731 [0.709, 0.755]); on the held-out genomes, its within-position AUROC is 0.714 [0.683, 0.745], and its gain over position has a Holm-corrected p of 2.3×10^{-10} over the eight models. On 36 held-out STRING genomes, block 12 gains 0.0924 [0.0746, 0.1124] over the position-only model, and the middle layer, block 25, gains only 0.0244. The median norm of the 40B’s gene vectors rises from about 150 at block 20 to about 1.5×10^{18} at block 24, and the development within-position AUROC of the standardised readout falls from 0.731 at block 12 to 0.623–0.630 from block 23 on. The block-12 gain is positive in all 36 held-out genomes, against +0.0013 at the final layer. On the leaderboard, the block-12 readout gains +0.0005 on expression-filtered operons, with an interval that includes zero, and on strict interactions the 40B’s final layer and block 25 score 0.492 and 0.504. In the sweep’s own evaluation, the block-12 readout gains +0.0052 [+0.0014, +0.0090] on RegulonDB units and clears chance on 4 of 7 strict experiment-only *E. coli* interaction data sets, against none for block 25 and the final layer, where clearing chance means that the intervals of its within-position AUROC and of its gain lie above 0.5 and 0.

Results in detail. Rows that did not change reproduce the earlier leaderboard exactly. Final layers score 0.482–0.506. Where position carries signal, on all pairs, the middle and R1 readouts gain +0.022 to +0.048 over row zero, except the middle layer of Evo 2 40B (+0.0133). No readout has a gain interval above 0 on expression-filtered operons (largest +0.0007), and the largest gain on experimentally supported RegulonDB units is +0.0085 [+0.0045, +0.0132] (Evo 2 7B, R1). A HyenaDNA [28] fine-tuned on the distance-free inputs of *etgp_nodist* scores 0.682 [0.634, 0.730] with loci held out, below distance alone (0.836 [0.812, 0.863]). Figure 4 shows the strict and gain readouts.

Full leaderboard. Table 22 lists each scorer, with 95% intervals.

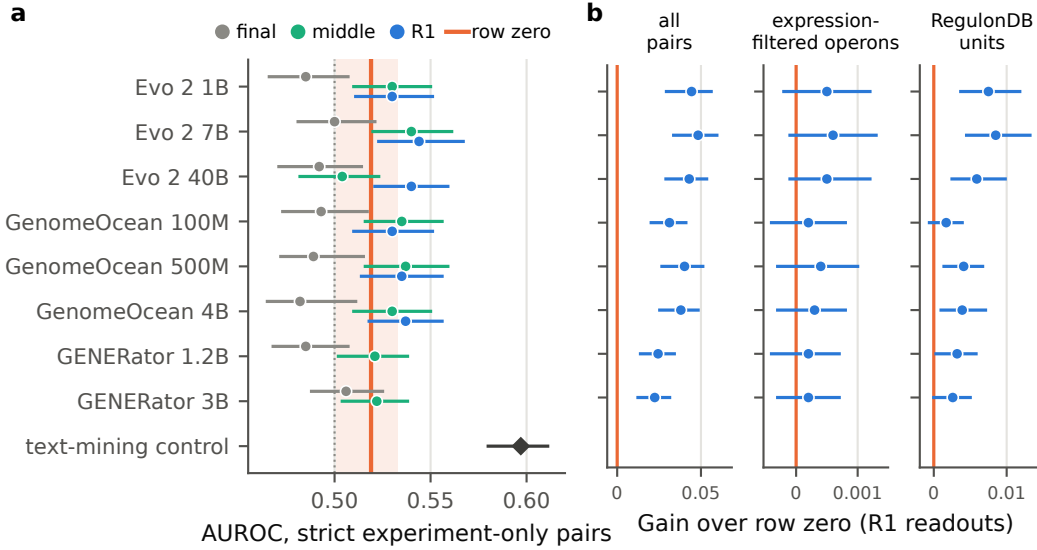


Figure 4: PairCtrl leaderboard with the added models. (a) Strict experiment-only interactions (14 data sets, six genomes): AUROC with 95% intervals for each model’s final, middle and R1 readouts, with row zero (orange line, interval shaded) and a text-mining positive control. (b) Gain over row zero on all interaction pairs, expression-filtered operons and RegulonDB units for the R1 readouts.

Table 22: PairCtrl with the added models: strict experiment-only interactions (AUROC) and gains over row zero on all interaction pairs, expression-filtered operons and RegulonDB units, with 95% intervals.

Scorer	ppi_exp strict: AUROC	ppi_exp all pairs: gain	operon_expr: gain	RegulonDB units: gain
row zero (position model)	0.519 [0.501, 0.533]	–	–	–
distance rule	0.512 [0.492, 0.531]	–	–	–
Evo 2 1B, final	0.485 [0.466, 0.507]	+0.0034 [+0.0001, +0.0063]	+0.0002 [–0.0002, +0.0006]	+0.0044 [+0.0002, +0.0084]
Evo 2 1B, middle	0.530 [0.510, 0.550]	+0.0460 [+0.0311, +0.0576]	+0.0004 [–0.0003, +0.0011]	+0.0075 [+0.0036, +0.0122]
Evo 2 1B, R1	0.530 [0.511, 0.551]	+0.0444 [+0.0294, +0.0561]	+0.0005 [–0.0002, +0.0012]	+0.0075 [+0.0037, +0.0118]
Evo 2 7B, final	0.500 [0.481, 0.521]	+0.0047 [+0.0017, +0.0080]	+0.0001 [–0.0001, +0.0003]	+0.0015 [–0.0002, +0.0033]
Evo 2 7B, middle	0.540 [0.520, 0.561]	+0.0446 [+0.0313, +0.0547]	+0.0007 [–0.0000, +0.0014]	+0.0073 [+0.0036, +0.0116]
Evo 2 7B, R1	0.544 [0.523, 0.567]	+0.0483 [+0.0337, +0.0595]	+0.0006 [–0.0001, +0.0013]	+0.0085 [+0.0045, +0.0132]
GenomeOcean 100M, final	0.493 [0.473, 0.517]	+0.0087 [+0.0037, +0.0135]	–0.0000 [–0.0002, +0.0001]	+0.0023 [–0.0005, +0.0050]
GenomeOcean 100M, middle	0.535 [0.516, 0.556]	+0.0369 [+0.0236, +0.0491]	+0.0002 [–0.0003, +0.0006]	+0.0026 [+0.0001, +0.0051]
GenomeOcean 100M, R1	0.530 [0.510, 0.551]	+0.0312 [+0.0203, +0.0410]	+0.0002 [–0.0004, +0.0008]	+0.0017 [–0.0006, +0.0039]
GenomeOcean 500M, final	0.489 [0.472, 0.515]	+0.0094 [+0.0043, +0.0137]	+0.0001 [–0.0001, +0.0003]	+0.0015 [–0.0006, +0.0035]
GenomeOcean 500M, middle	0.537 [0.516, 0.559]	+0.0389 [+0.0254, +0.0503]	+0.0005 [–0.0002, +0.0011]	+0.0038 [+0.0012, +0.0066]
GenomeOcean 500M, R1	0.535 [0.514, 0.556]	+0.0402 [+0.0267, +0.0511]	+0.0004 [–0.0003, +0.0010]	+0.0041 [+0.0014, +0.0067]
GenomeOcean 4B, final	0.482 [0.465, 0.511]	+0.0089 [+0.0041, +0.0131]	–0.0000 [–0.0002, +0.0002]	+0.0011 [–0.0009, +0.0031]
GenomeOcean 4B, middle	0.530 [0.510, 0.550]	+0.0374 [+0.0248, +0.0472]	+0.0003 [–0.0003, +0.0009]	+0.0044 [+0.0012, +0.0079]
GenomeOcean 4B, R1	0.537 [0.518, 0.556]	+0.0380 [+0.0254, +0.0483]	+0.0003 [–0.0003, +0.0008]	+0.0039 [+0.0010, +0.0071]
GENERator 1.2B, final	0.485 [0.468, 0.507]	+0.0026 [–0.0008, +0.0058]	–0.0002 [–0.0004, –0.0000]	+0.0001 [–0.0008, +0.0010]
GENERator 1.2B, middle = R1	0.521 [0.502, 0.538]	+0.0244 [+0.0139, +0.0341]	+0.0002 [–0.0004, +0.0007]	+0.0032 [+0.0003, +0.0058]
GENERator 3B, final	0.506 [0.488, 0.525]	+0.0025 [–0.0001, +0.0054]	+0.0001 [–0.0002, +0.0004]	+0.0001 [–0.0007, +0.0010]
GENERator 3B, middle = R1	0.522 [0.504, 0.538]	+0.0224 [+0.0124, +0.0313]	+0.0002 [–0.0003, +0.0007]	+0.0026 [+0.0000, +0.0050]
Evo 2 40B, final	0.492 [0.471, 0.514]	+0.0003 [–0.0027, +0.0025]	–0.0000 [–0.0003, +0.0002]	+0.0018 [–0.0000, +0.0038]
Evo 2 40B, middle	0.504 [0.482, 0.523]	+0.0133 [+0.0065, +0.0187]	–0.0001 [–0.0004, +0.0002]	+0.0067 [+0.0036, +0.0106]
Evo 2 40B, R1	0.540 [0.521, 0.559]	+0.0430 [+0.0291, +0.0534]	+0.0005 [–0.0001, +0.0012]	+0.0059 [+0.0025, +0.0098]
Bacformer Large	0.526 [0.508, 0.545]	+0.0291 [+0.0168, +0.0393]	–0.0000 [–0.0003, +0.0003]	+0.0061 [+0.0024, +0.0098]
protein LM + position alternative	0.529 [0.509, 0.546]	+0.0210 [+0.0118, +0.0286]	–	–
text-mining control	0.597 [0.580, 0.611]	+0.0698 [+0.0554, +0.0796]	–	–

I Constructive experiments: details

Training on layout-free labels. The labels are experiment-only interactions of four bacterial genomes, on the strict subset of Section 6. Each fold trains on three genomes and tests on the fourth, and a test pair is kept only if neither protein has an MMseqs2 hit to a protein of the fold’s training pairs. The 11 test sets with at least 20 strict positives after this filter were fixed from label counts before scoring. Eleven candidates were pre-registered: classifiers on frozen protein and DNA embeddings, Bacformer Large fine-tuned end to end and a pretraining variant with position input. A candidate beats position when its mean AUROC minus that of the position model and its mean gain over the position model are both above zero, with a Holm correction over candidates. Folds are saved

as files and the decisions were rechecked under several tie orders of the fold assignment. Untrained per-genome standardised cosines of the same embeddings serve as references. Classifiers on frozen embeddings, and Bacformer Large fine-tuned end to end [38], were trained on experiment-only interactions of three genomes and tested on the fourth, on the strict subset and after a homology filter. Position alone is at chance on this subset (mean AUROC 0.4872). With the fold order of the original run, one protein-only classifier, a small network on frozen ESM-2 35M embeddings [39], passes the registered rule (Holm p 0.0221). The pass depends on the order in which cross-validation places gene blocks of equal size: with the stable order of a later scikit-learn release [40], which is equally arbitrary, nothing passes (lowest Holm p 0.0993). The classifier’s AUROC (0.5479) is no higher than that of the untrained cosine of the same embeddings (0.5756) or of 22 generic sequence properties (0.5610) (point estimates). In the primary analysis, no trained candidate whose features carry genomic context passes under either order. In the pre-registered secondary setting with all-pairs training, where the rule is applied descriptively, a logistic classifier on ESM-2 35M and a small network on gene-centred Evo 2 1B DNA embeddings pass under both orders (Holm p 0.0221 in the logged order and 0.0441 in the stable one), at AUROC 0.5690 and 0.5489, not above the untrained ESM-2 cosine (0.5756; point estimates).

Position-controlled readouts and heads. The registered candidates are a residualised readout, which projects out of each embedding the directions that predict gene layout, and a pair head trained on STRING development genomes with a position adversary. The controls are the same head without the adversary, a head trained on permuted labels, a random subspace of the residualised readout’s rank and a head trained on labels recomputed without the transferred channels of STRING. Table 23 gives the held-out results. On 36 held-out STRING genomes, a pair head trained with a position adversary improves on the plain R1 readout for 7 of 7 genomic models under each of three fold seeds, with within-position AUROC 0.7867–0.8271 against 0.6224–0.6863. Residualising the embeddings against position improves none of them. The controls show where this gain comes from. The same head without the adversary reaches 0.7780–0.8238, so the adversary adds at most +0.0098. A head on a protein-only embedding (ESM-2 35M) reaches 0.8428, the highest of the 11 embeddings. Heads trained on labels without STRING’s transferred evidence channels reach only 0.6126–0.6298, no better than the untrained cosines. No candidate and tie-order cell, 0 of 110, beats the untrained cosine of the same embedding on layout-free experimental labels. Adding STRING’s four context channels to the position model removes a ceiling-adjusted share of 0.274–0.370 of the mid-depth gain over position, so the gain is mostly not conserved neighbourhood. Re-embedding each gene without its neighbouring DNA keeps a share of 0.802 [0.770, 0.840] of the gain of Evo 2 1B. On held-out pairs without any STRING context evidence, the Evo 2 readouts keep gains of +0.0564 (1B) and +0.0523 (7B).

Table 23: Readouts and heads on 36 held-out STRING genomes: AUROC within position strata / gain over the position model (fold seed 0).

Embedding	untrained cosine	residualised readout	adversarial head	head without adversary
Evo 2 1B, R1	0.6797 / +0.0853	0.6689 / +0.0764	0.8103 / +0.1658	0.8019 / +0.1612
Evo 2 7B, R1	0.6863 / +0.0910	0.6744 / +0.0816	0.8271 / +0.1780	0.8238 / +0.1763
GenomeOcean 100M, R1	0.6224 / +0.0478	0.5782 / +0.0273	0.8105 / +0.1671	0.8006 / +0.1625
GenomeOcean 500M, R1	0.6489 / +0.0605	0.6221 / +0.0440	0.8161 / +0.1726	0.8104 / +0.1704
GenomeOcean 4B, R1	0.6557 / +0.0652	0.6166 / +0.0446	0.8171 / +0.1713	0.8097 / +0.1656
GENERator 1.2B, R1	0.6370 / +0.0473	0.5616 / +0.0202	0.7867 / +0.1525	0.7780 / +0.1468
GENERator 3B, R1	0.6315 / +0.0433	0.5491 / +0.0158	0.7937 / +0.1554	0.7879 / +0.1532
ESM-2 35M	0.6209 / +0.0527	0.6042 / +0.0408	0.8428 / +0.1853	0.8429 / +0.1847
ESM-C 300M	0.6366 / +0.0611	0.6002 / +0.0388	0.8167 / +0.1701	0.8135 / +0.1679
Bacformer Large	0.6434 / +0.0613	0.6349 / +0.0545	0.7594 / +0.1375	0.7626 / +0.1421
Evo 2 1B, block 12	0.6779 / +0.0843	0.6661 / +0.0748	0.8062 / +0.1640	0.8003 / +0.1606

J Datasheets, licences, maintenance and availability

Datasheets. Each released suite of drop-in versions and PairCtrl come with a datasheet that states motivation, composition, intended uses, distribution and maintenance. The checks and the residual of every drop-in version are listed in the READMEs (Appendix E).

Licences. The licence audit of PairCtrl covers 51 sources. Of the 17 interaction data sets, 14 can be shipped with attribution. The structure screen (CC BY-NC-ND 4.0), STRING's database-channel controls and DGEB's own labels are not shipped, and RegulonDB labels and the enhancer input sequences are rebuilt by join scripts. GUE versions ship as row ids and a build script, as described in Appendix E.

Maintenance. PairCtrl's data and code share semantic versions, a released version never changes, and corrections produce a new version with a changelog and the old tables kept. Leaderboard submissions are per-item scores with a short model card, evaluated by the maintainer with the same engine. A task found to be solvable by an uncontrolled confound is flagged and kept for comparison. The linter's rules are versioned: v1.0 to v1.2.1 are frozen and archived, and a later version will be validated on a new held-out set.

Release checks. Before release, the frozen linter, the released versions and PairCtrl's evaluation engine passed their integrity checks and tests; the logs are in the archive.

Availability. The archive at <https://doi.org/10.5281/zenodo.23717097> holds the linter (seqleak 0.4.1, with seqleak audit and seqleak clean) and its frozen rule versions, the build scripts of the drop-in versions, PairCtrl's evaluation package and leaderboard, and the scripts and logs behind the paper's numbers, with tables that link each number to its source. The drop-in versions themselves are at <https://huggingface.co/datasets/tstomar/leak-reduced-dna-benchmarks> and PairCtrl's task tables at <https://huggingface.co/datasets/tstomar/pairctrl>; the code of the linter and of PairCtrl is also at <https://github.com/tejasvi/seqleak> and <https://github.com/tejasvi/pairctrl>. All input data are public.