

Gene position and relatedness confound genomic language model benchmarks

Tejasvi Singh Tomar

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

28 October 2026

Abstract

Benchmark scores of genomic language models are read as evidence of learned genome organisation and gene function; we tested, with baselines that use no sequence model, whether gene position and genome relatedness explain them instead. The answer depends on how a benchmark’s labels, inputs and splits were built: where these encode position or relatedness, such baselines can explain much of a model’s score. Position alone outperforms every language model we compared on bacterial operon benchmarks (as in a companion study), and distance matches or beats most published DNA language models on most eukaryotic long-range tasks. On two label sets we built without genome layout, no model we tested gains more than 0.002 AUROC over position on operons labelled from nanopore reads of expressed genes, and little model signal on experiment-only interaction labels survives strict controls. Relatedness explains much of the models’ performance on BacBench’s two genome-level tasks, and train–test overlap or negatives built from positives let a rule or lookup exceed the best published score we found on 24 of 86 tasks of widely used DNA benchmarks. Real signal remains on gene-level tasks, where it appears to come from the identity of genes and their neighbours; signal beyond position also remains on STRING interaction labels and in DNA language models fine-tuned on a causal eQTL task, but it is often not specifically genomic. Benchmarks therefore need these baselines, metrics with position held fixed, labels built without layout, and leak-free inputs and splits; we list a fix for each defect and release seqleak, which computes the position baseline and these metrics.

1 Introduction

Genomic language models learn from DNA, or from genomes written as ordered lists of proteins. For bacteria, gLM and gLM2 model metagenomic contigs [1, 2], Bacformer models whole genomes as ordered sets of proteins [3], and Evo and Evo 2 model DNA directly [4, 5]. Eukaryotic DNA language models such as the Nucleotide Transformer, DNABERT-2, HyenaDNA and Caduceus read windows of up to a megabase [6–9]. Their benchmarks ask about pairs of genes, pairs of elements, single genes and whole genomes [3, 10–13], and high scores are taken to show that the models capture genome organisation and gene function. Leaderboards rank models by these scores [10], and model papers report them as evidence of what pretraining on genomes captures [2, 3]. If simpler information explains the scores, these rankings reward that information rather than learned biology.

Two simpler explanations compete with that reading. The first is gene position. Genes of one operon lie on the same strand with short gaps [14, 15], and STRING scores protein pairs partly by the conserved proximity of their genes [16], so labels for operons and for STRING interactions both carry position. Dedicated operon predictors build on this with comparative genomics (MicrobesOnline) [17, 18], intergenic distance (UniOP) [19] or deep learning (the Operon Finder web server) [20].

In eukaryotes, enhancers near a transcription start site (TSS) are more likely to regulate its gene [21]. A model that encodes where genes and elements lie can therefore score well without learning what they do. The second is relatedness. Resistance and traits are shared within clades, so on a random split a model that recognises relatives of training genomes can score well. Some evaluations already control for distance: ENCODE-rE2G reports a distance baseline [22], and AlphaGenome and TraitGym balance or match controls on distance [23, 24]. An enhancer–gene benchmark found a distance method hard to beat [25], and elsewhere in genomics simple baselines rival foundation models [26–30]. In a companion paper we showed that a distance rule beats genomic language models at operon prediction and that STRING-derived interaction labels are built partly from genome organisation [31].

Here we ask, task by task, whether position or relatedness explains the models’ scores and what signal remains once they are held fixed. We audit six benchmarks [1, 10–13, 32], the KEGG task of the Bacformer preprint [3], two operon databases and two label sets built without genome layout, comparing models with baselines that use no sequence model and, where possible, evaluating them with position held fixed (Table 1, Figure 1). We also search 86 tasks of widely used DNA benchmarks for the same classes of defect. We then ask how the models come to encode position and whether scale or a cheaper model changes the picture, and we propose a position-aware evaluation standard (Table 8).

2 Results

2.1 Each model is reported with a position baseline and a share interval

For each task of the six benchmarks we use the split, metric and published numbers of the benchmark and, where a published value or model could be re-run, check our reproduction first (Table 9). For the Long-Range Benchmark (LRB) [12], values are recomputed from the released predictions. We add four quantities, computed with seqleak (Methods). The *no-model baseline* uses no sequence model. Its main form is the position model, fitted out of fold, which sees only position. For neighbouring genes it sees strand and gap, and for other pairs also how many genes apart they lie, the largest intervening gap and strand switches. For eukaryotic tasks it sees element–TSS distance. For single genes it sees strand, replication strand, distance to the origin, direction, local gene density, neighbouring gaps and the distance-rule score, but not gene length. The *within-position AUROC* compares positives with negatives only inside strata that share their position, so a score that is constant within strata yields 0.5. Position-only scores keep a little signal inside the strata, and the position model’s own within-position AUROC is 0.50–0.58. The *gain over position* is the change in out-of-fold AUROC when the score of a model is added to the position model. The *position share* is the fraction of the above-chance AUROC of a model that comes from position. It depends on whether position is a consequence of the label (in operons, co-transcription shapes spacing) or a cause of it (STRING scores links partly from neighbourhood), so seqleak reports an interval $[\text{share}, \text{share}_{\text{upper}}]$ whose lower end is exact in the first case and whose upper end is exact in the second. Genome-level baselines use relatedness (a nearest-relative lookup and a taxonomy prior), and genome-level models are also evaluated by AUROC within species or families.

2.2 Gene position outperforms most models on pairwise genome-organisation tasks

The companion paper covers operons and interactions in BacBench, operon databases and the gLM evaluation set [31]. The untrained rule $\mathbf{1}[\text{same strand}] \cdot \exp(-\text{gap}/500)$, zero above 500 bp, beats the operon score of every language model on these sets (Tables 1 and 2a). With position held fixed, em-

Table 1: The tasks audited with seqleak: the six benchmarks, two operon databases, Bacformer’s KEGG task and our two label sets built without genome layout (further benchmarks: Table 5). The no-model baseline is the strongest baseline we found without a sequence model, on the split and metric of the benchmark, except on two two-hybrid screens (with our folds) and on KEGG, where the distance rule and gene layout with gene length, respectively, score slightly higher; the DGEB cells give the position model trained on the other two genomes. “Best model” is the best published language model, or our reproduction where none exists (OperonDB windows, RegulonDB and SubtiWiki pairs, the two new label sets) or the released predictions (LRB), not counting our Evo 2 readouts (Table 7). Ranges cover genomes or data sets.

Task (benchmark)	Labels from	Metric	No-model baseline	Best model	Explained by
<i>Pairwise genome organisation, bacteria (Section 2.2)</i>					
Long-read operons (BacBench)	reads spanning genes	AUROC	position 0.833	0.776	position
Known operons, windows (BacBench)	OperonDB	AUROC	rule 0.755	0.648	position
Operon pairs (RegulonDB; SubtiWiki)	database operons	AUROC	position 0.958; rule 0.968	0.857; 0.861	position
Operon pairs, <i>E. coli</i> (gLM)	operon annotation	mean AP	position 0.905	0.775	position
Operonic pairs, 3 genomes (DGEB)	BioCyc; 2 of 3 predicted	AP	position 0.854–0.990	0.409–0.626	position (labels predicted)
Interactions, zero-shot (BacBench)	STRING score ≥ 600	AUROC	position 0.705	0.721	position, in part
<i>Labels not built from genome layout, this work (Section 2.3)</i>					
Experiment-only interactions, two-hybrid	two-hybrid screens	AUROC	position 0.519–0.535	0.519–0.592	neither
Operons from nanopore reads	both genes expressed	AUROC	position 0.987–0.991	0.742–0.753	position
<i>Eukaryotic long-range tasks (Section 2.4)</i>					
Enhancer–target (DNALongBench)	CRISPRi screen	AUROC; AUPRC	distance 0.821; 0.422	0.828; 0.380	distance, via input length
eQTL, 9 tissues (DNALongBench)	GTEEx fine-mapping	AUROC	distance 0.750	0.565	distance (models weak)
Causal eQTL, fine-tuned (LRB)	GTEEx fine-mapping	AUROC	distance + annotation 0.706	0.732	distance, in part
Causal eQTL, zero-shot (LRB)	GTEEx fine-mapping	AUROC	distance 0.600	0.500	models at chance
Enhancer annotation (BEND)	CRISPRi enhancers	AUPRC	distance 0.035	0.06	distance, for most models
Expression variants (BEND)	GRASP eQTLs	AUROC	distance 0.672	0.60	models weak
<i>Genome-level and gene-level tasks (Sections 2.6 and 2.7)</i>					
Antibiotic resistance (BacBench)	AST phenotypes	AUROC	baseline stack 0.929	0.9215	relatedness
Phenotypic traits (BacBench)	trait databases	AUROC	taxonomy prior 0.685	0.733	relatedness
Essential genes (BacBench)	DEG database	AUROC	gene layout 0.645	0.8245	protein identity
KEGG orthology (Bacformer)	KEGG	zero-shot F1	gene layout 0.010	0.901	protein identity

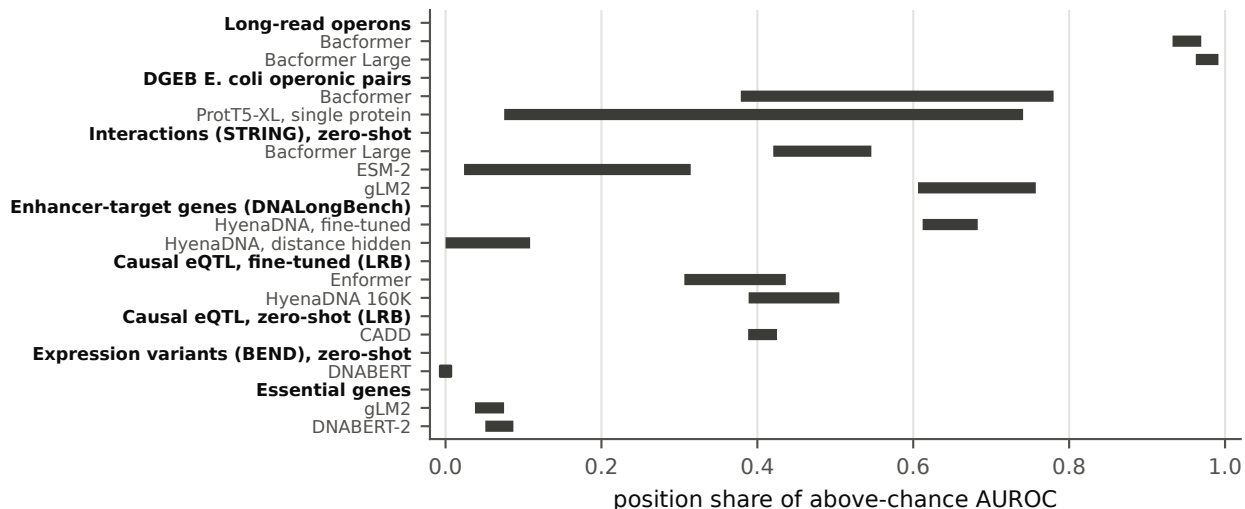


Figure 1: Position-share intervals across the audit. For representative models of eight tasks, the bar spans [share, share_{upper}], the interval for the share of the above-chance AUROC of the model that comes from position (squares: intervals narrower than 0.012). The lower end applies when position is a consequence of the label (operons), the upper end when it is a cause (STRING neighbourhood, enhancer and variant distance).

bedding similarity taken from the layers BacBench uses is at or below chance on the long-read operon labels of BacBench, pooled over the five strains. The exception is Evo 2 1B read at its final layer, which is just above chance (Section 2.9). Each of the 12 BacBench models has a position share of at least 0.93 (Figure 1 shows the two strongest). On the RegulonDB pairs, by contrast, both Bacformer models and the mid-depth layers of Evo 2 (the outputs of its middle block, standardised within each genome; Section 2.9) add a little beyond position with our folds, at most +0.006 AUROC (Table 2a). By embedding cosine, the mid-depth layer of Evo 2 1B scores a raw AUROC of 0.873 on these pairs.

For interactions, the six models that were re-run zero-shot on STRING labels differ in position share, which is lowest for ESM-2 and ESM-C and highest for gLM2 (Table 2c). Of the six, Bacformer Large keeps the most signal beyond position (gain +0.053), and the mid-depth readouts of Evo 2 1B and 7B keep more than Bacformer Large (Section 2.9).

DGE evaluates operonic pairs in three genomes, *Escherichia coli*, *Synechococcus elongatus* and *Vibrio cholerae*, by ranking gene pairs on the similarity of single-protein embeddings, with average precision (AP) as the metric [10]. The rule beats all 12 leaderboard models on all three tasks, with AP 0.847 against at most 0.626 on *E. coli* (Figure 2b, Table 2b). A position model trained on the other two genomes reaches AP 0.854 to 0.990.

In two of the three genomes, the rule’s lead comes largely from how the labels were made. DGE takes BioCyc transcription units, and BioCyc lists the *S. elongatus* and *V. cholerae* databases in its Tier 2, stating that “the 74 PGDBs in Tier 2 were generated by the PathoLogic program, which was used to predict their metabolic pathways, their operons (for bacteria only)” [33]. This predictor uses intergenic distance and functional relationships between adjacent genes [34, 35], and recent versions add GC content and codon-usage similarity. The Tier 2 labels follow a step at 40 bp (Figure 2a): “same strand and $-21 \leq \text{gap} \leq 40$ ” reproduces 98.7% of the *S. elongatus* labels, and “same strand and $\text{gap} \leq 40$ ” reproduces 95.1% of the *V. cholerae* labels. Two of the three DGE operon tasks therefore measure agreement with a predictor driven mainly by intergenic distance. The *E. coli* labels come from the intensively curated EcoCyc (Tier 1), and on experimentally supported RegulonDB units the rule still beats every model run on these units (AP 0.735 versus 0.521).

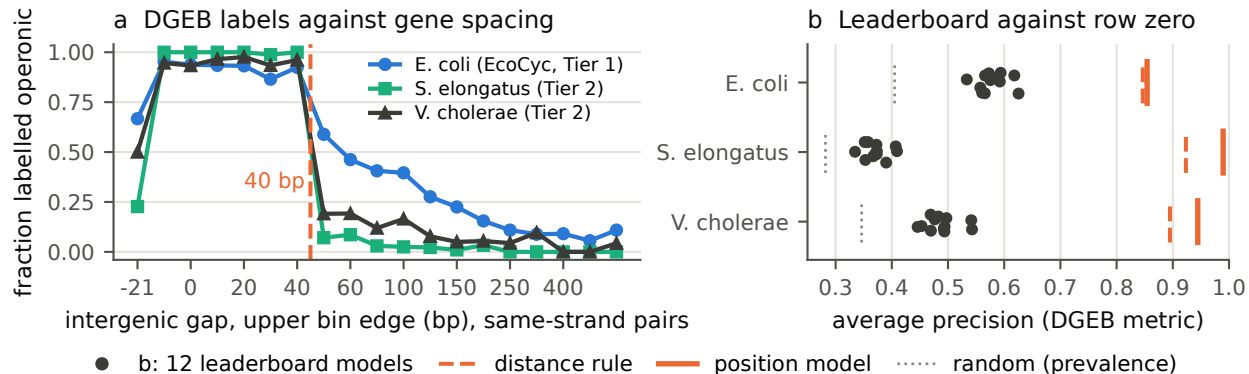


Figure 2: DGEb operonic pairs. (a) Fraction of same-strand pairs labelled operonic, by intergenic gap (gaps from -51 to 750 bp; every bin has at least 14 pairs), with the 40-bp step of the two Tier 2 label sets. (b) Average precision of the 12 leaderboard models (grey), the distance rule (dashed) and a position model trained on the other two genomes (solid), with the positive rate (dotted).

Table 2: Pairwise genome-organisation tasks in detail (Section 2.2): (a) operon label sets, (b) DGEb operonic pairs (average precision; last row: position-share interval), (c) zero-shot STRING interactions (ProtBERT, near chance, has no share interval). Brackets after a gain are 95% intervals; share intervals are [lower end, upper end].

(a) Operon label sets	Long-read	OperonDB	RegulonDB	SubtiWiki	gLM set
Rule (AUROC; gLM set: mean AP)	0.826	0.755	0.954	0.968	0.857
Evo 2 7B mid-depth: gain over position (within-position AUROC 0.68)			+0.006 [+0.003, +0.010]		
(b) DGEb operonic pairs (AP; last row: share interval)	<i>E. coli</i>	<i>S. elongatus</i>	<i>V. cholerae</i>		
Rule	0.847	0.923	0.895		
Best of the 12 leaderboard models	0.626	0.409	0.543		
Position model trained on the other two genomes	0.854	0.990	0.944		
Same-strand pairs: rule	0.848	0.923	0.895		
Same-strand pairs: best model (DGEb or BacBench protocol)	0.753	0.567	0.698		
Share interval of Bacformer (BacBench protocol)	[0.38, 0.78]				
(c) Zero-shot STRING interactions	ESM-2	ESM-C	Bacformer	Bacformer Large	gLM2
Share interval	[0.02, 0.31]	[0.16, 0.30]	[0.36, 0.51]	[0.42, 0.55]	[0.61, 0.76]

The DGEb models see one protein at a time and never see strand or gap, yet every opposite-strand pair is a negative. Restricted to same-strand pairs, the rule still leads in each genome (Table 2b). At fixed position the single-protein models keep within-position AUROCs close to their raw values, as expected for models that cannot see position, and on *E. coli* and *V. cholerae* they add a little AP to the position model trained on the other two genomes (+0.001 to +0.014). By contrast, the models that were run on these pairs under the BacBench protocol, with genome context for Bacformer, Bacformer Large, gLM2 and Evo, mostly have higher share intervals (for example Bacformer on *E. coli*, Table 2b) and add -0.004 to $+0.010$ AP. On *S. elongatus* only 0.35% of positive-negative comparisons fall within strata, so holding position fixed is uninformative there.

2.3 Little model signal survives strict controls on labels built without layout

A model that learned interaction or operon biology should keep its signal on labels built without genome layout, so we built two such label sets (Methods). The first, the experiment-only labels, holds 42,943 positive pairs in 17 data sets from experiments, curated complexes and one structure-

prediction screen in six genomes: two-hybrid screens [36–39], affinity-purification mass-spectrometry (AP-MS) screens [40–44], cross-linking (XL-MS) studies [45, 46], curated complexes [47], small-scale studies [48] and the structure screen [49], whose paired alignments can carry the gene order of other genomes. The second labels operons in three genomes from nanopore reads [50–55]. A same-strand pair is negative only when both genes are expressed and fewer than 10% of the reads reaching the downstream gene also extend into the upstream gene (the spanning fraction, with strict and lenient thresholds in Methods).

Few two-hybrid positives lie within 20 genes of each other (3.1–7.1%), so position is weak on the two-hybrid and AP-MS labels, though not on complexes, small-scale literature, the *B. subtilis* XL-MS set and the structure screen (Table 3a). The position model also scores lower on the two-hybrid and AP-MS screens than on STRING labels with negatives from the same screened spaces (Table 3a). On two-hybrid pairs more than 20 genes apart, with degree-preserving negatives, the best model per screen, counting Evo 2 readouts, scores 0.512–0.555 (Figure 3a; Table 3a gives the values on the strict subset). However, independent evidence built without the layout of the genome or the screen scores 0.607–0.633 on the same pairs (Methods), which shows that the labels are learnable.

On the four *E. coli* AP-MS screens, Bacformer Large is the best of the 12 models, and on each screen the better of the mid-depth layers of Evo 2 7B and 40B scores higher still. Both lose most of this signal on distant, degree-matched pairs without ribosomal proteins or neighbours in other genomes (Table 3a). By contrast, the other *E. coli* data sets (mainly the screen from the same group that shares about half of their positives) still predict two of these screens: those of Butland et al. [40] (AUROC 0.775) and Hu et al. [42] (0.702). Independent evidence also predicts these two screens (Table 3b).

A residual survives on curated complexes, on cross-linked pairs in *Mycoplasma pneumoniae* and on membrane AP-MS. On the cross-linked and membrane AP-MS pairs, single-protein models retain the residual signal about as well as the best of the 12 models, and the mid-depth layer of Evo 2 7B carries the most on membrane AP-MS (0.558, above every positive control: independent evidence, its single STRING channels and the other data sets of the same genome). The Nucleotide Transformer keeps an unexplained signal on the structure screen’s *P. aeruginosa* and *L. monocytogenes* sets while positive controls are at chance (Table 3b).

For the nanopore operon labels, the position model reaches an AUROC of 0.987–0.991 on the main set of each of the three genomes. Across these sets and two further *E. coli* direct-RNA sets (Figure 3b, Table 3c; a sixth, with 12 same-strand negatives, is too small to use), the largest gain of any model or Evo 2 readout is +0.0008 (at most +0.0017 under other fold orders), under both the strict and the lenient spanning-fraction thresholds. With our folds, the largest gain on SMRT-Cappable-seq units [56] has an interval that includes zero, and gains whose intervals lie above zero reappear only when the expression filter is removed and pairs without spanning reads are called negatives, as in the long-read labels of BacBench. Then some models gain up to +0.026 through similarity that, within strata, tends to rank negatives above positives (within-position AUROC 0.333). The BacBench labels of *P. aeruginosa* show the same signature, and those of *S. aureus* give a gain whose interval includes zero (Table 3c). We attribute the apparent operon information beyond position in BacBench to co-transcribed pairs labelled as negatives, as happens when pairs without spanning reads, including unexpressed ones, are called negatives.

2.4 Distance matches most DNA models on eukaryotic long-range tasks, with one exception

DNALongBench’s enhancer–target task asks whether a CRISPRi-tested element regulates a gene, given up to 450 kb of sequence [11, 21]. Distance alone reaches a test AUROC of 0.821 and an

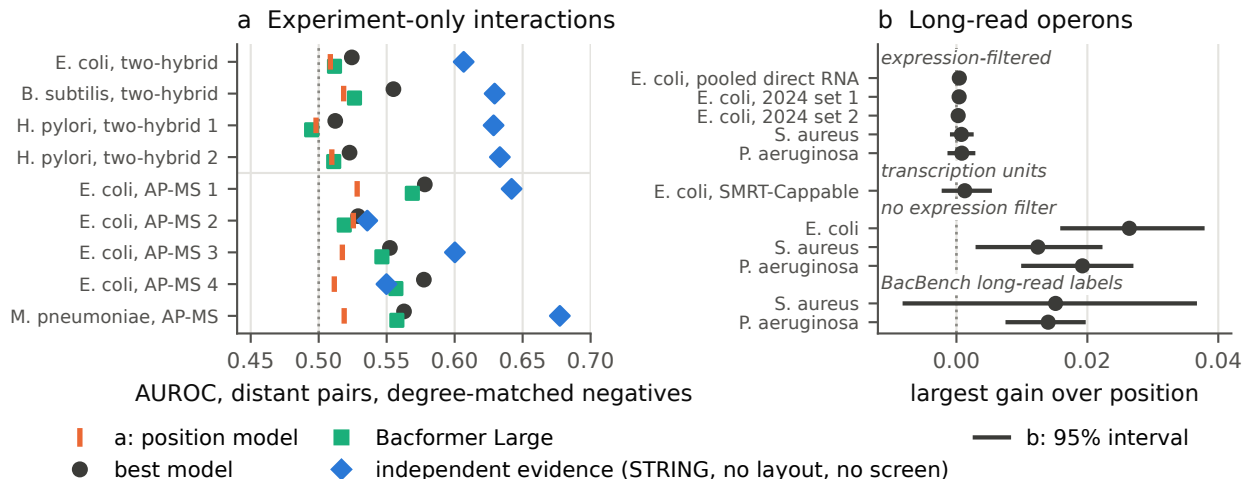


Figure 3: Labels not built from genome layout. (a) Two-hybrid and AP-MS screens, pairs more than 20 genes apart with degree-preserving negatives: the position model, the best model (Evo 2 readouts included), Bacformer Large and independent evidence (STRING’s transferred experiments and database entries, co-expression and text mining, built without the layout of the scored genome or the screen). (b) Largest gain of any model or Evo 2 readout over the position model on long-read operon labels (nanopore labels with and without the expression filter, SMRT-Cappable-seq units and the BacBench long-read labels), with 95% block-bootstrap intervals.

AUPRC of 0.422 (Figure 4a), matching the fine-tuned DNA language models on AUROC and exceeding all of them on AUPRC (Table 4a). The loader leaks distance because it pads the sequence between element and TSS with trailing N to 450,000 characters. The count of N therefore has a Spearman correlation of -0.9996 with distance, and a rule on that count alone matches distance and the published models (Table 4a). A HyenaDNA model we fine-tuned on these inputs exceeds its published value (Table 9) on the official test split at its best validation epoch but stays below distance plus a per-gene prior fitted on the training split (0.893 against 0.910). In cross-validation over all 2,602 pairs its gain over distance has an interval that includes zero, and with distance hidden from its input it falls to 0.602. The activity-by-contact (ABC) score, by contrast, keeps an AUROC of 0.740 within distance strata.

In the eQTL task of DNALongBench, distance averages 0.750 test AUROC over nine tissues and beats every DNA language model in all nine and the Enformer expert model [57] in seven (Figure 4b). An unpublished manuscript with code reports this distance result (Table 4b), the models’ indifference to the allele and a label leak through masked variants [58]. Its repository’s findings log notes distance encoded in the N padding of the eQTL inputs, but neither covers the enhancer–target task.

In BEND, distance to the nearest TSS scores 0.672 AUROC on expression variants, above all 13 published DNA language models, and on enhancer annotation its AUPRC matches or exceeds 11 of 16 models at the two decimals BEND reports [13]. Of the models BEND reports, only the supervised DeepSEA scores higher on expression variants (0.70). The zero-shot scores of DNABERT and two Nucleotide Transformers are weak but not distance-driven (their share intervals are $[0, 0]$), which rules out the position hypothesis for these models.

The Long-Range Benchmark’s causal eQTL task is the exception [12]. In the released predictions of 15 fine-tuned models, distance alone scores 0.650 AUROC, and every model gains over distance with an interval above zero, the 14 DNA language models by $+0.047$ to $+0.079$ (Figure 4c, Table 4c). Because distance likely helps cause these labels, the upper end of their share intervals applies, and distance accounts for roughly half or less of what the models achieve (Figure 1). However, two

Table 3: Labels not built from genome layout in detail (Section 2.3; AUROC; brackets: 95% block-bootstrap intervals): (a) experiment-only screens, ranges over the screens of each class, where the strict subset holds pairs more than 20 genes apart with degree-preserving negatives and without ribosomal proteins or neighbours in other genomes (four two-hybrid screens; AP-MS: the position rows cover five screens, four *E. coli* and one *M. pneumoniae*, the model rows and the strict AP-MS column the four *E. coli* screens); (b) single data sets of the strict subset; (c) long-read operon labels (the gain on the *S. aureus* labels has an interval that includes zero).

(a) <i>Experiment-only screens</i>	Two-hybrid	Two-hybrid, strict	AP-MS	AP-MS, strict
Position model, screened-space negatives	0.519–0.535		0.543–0.599	
Position model, STRING labels in the same spaces	0.607–0.665		0.575–0.642	
Best model per screen, Evo 2 readouts included		0.508–0.545		
Independent evidence		0.582–0.629		
Bacformer Large			0.552–0.649	0.502–0.536
Better Evo 2 mid-depth layer (7B or 40B)			0.565–0.680	0.514–0.558
Position model on complexes, literature, <i>B. subtilis</i> XL-MS and the structure screen	0.67–0.80 (these classes, not two-hybrid)			
(b) <i>Strict subset, single data sets</i>	Other <i>E. coli</i> data sets	Independent evidence	Model	
AP-MS, Butland et al.	0.775	0.629		
AP-MS, Hu et al.	0.702	0.588		
Membrane AP-MS, Babu et al. (Evo 2 7B mid-depth)			0.558 [0.541, 0.576]	
Structure screen, <i>P. aeruginosa</i> set (Nucleotide Transformer)			0.686	
Structure screen, <i>L. monocytogenes</i> set (Nucleotide Transformer)			0.745	
(c) <i>Long-read operon labels</i>	Position model	Largest gain over position	Within-position AUROC of that model	
Nanopore, two further <i>E. coli</i> direct-RNA sets	0.993			
SMRT-Cappable-seq units		+0.0013 [−0.002, 0.005]		
BacBench long-read labels, <i>P. aeruginosa</i>		+0.014 [+0.008, +0.019]	0.416	
BacBench long-read labels, <i>S. aureus</i>		+0.015	0.403	

results qualify this. First, distance and 18 genome-annotation flags without a sequence model score 0.706 with our folds (Supplementary note), around the weakest fine-tuned DNA language model (0.697). With distance and annotation class held fixed, the fine-tuned DNA language models fall to 0.59–0.62. Second, zero-shot over all SNPs, the DNA language models are at or slightly below chance, whereas distance scores 0.600. On the distance bins of the Caduceus evaluation of this task [9], our untrained distance rule matches every model below 30 kb. Beyond 30 kb, distance is uninformative, and the models reach 0.54–0.66 (Table 4b).

2.5 The same classes of defect recur in widely used DNA benchmarks

We also audited 86 tasks of widely used DNA benchmarks for the three classes of defect found in the six benchmarks: relatedness across splits, input construction and position. They comprise 81 tasks of GUE with its extension GUE+ [7], of the original [59] and revised [6] Nucleotide Transformer tasks and of Genomic Benchmarks [60], and five other BEND tasks [13]. Published scores below are those collected in our search, and the best published score is the highest of them. Each no-model rule or lookup is evaluated on the benchmark’s own split and metric and compared with every published score for the task (Table 5, Methods). The best rule or lookup exceeds them all on 24 of the 81 tasks,

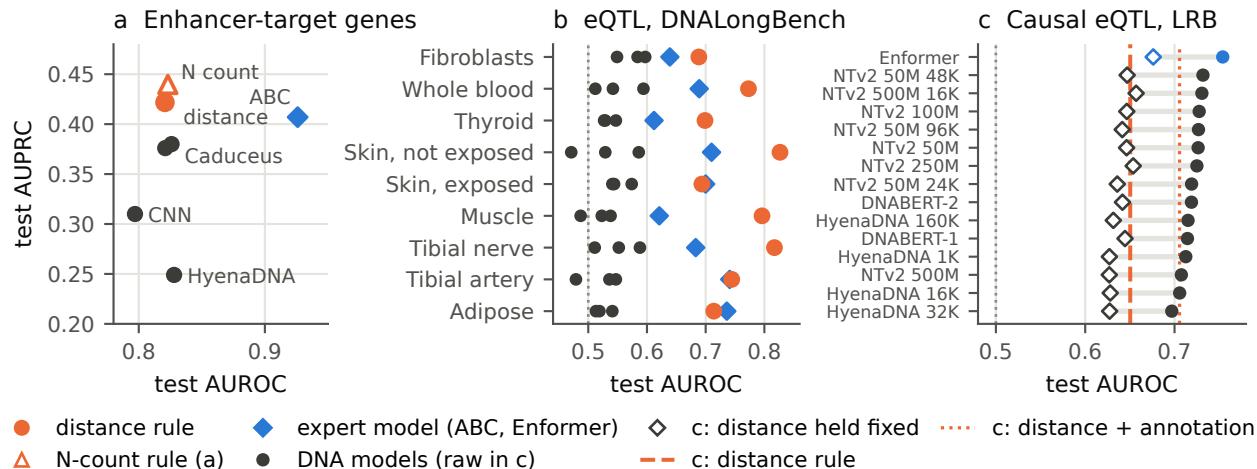


Figure 4: Eukaryotic long-range tasks. (a) DNALongBench enhancer-target genes, test split: published models (grey; blue: the ABC expert score), the distance rule and the N-count rule (number of N characters in the input). (b) DNALongBench eQTL, test AUROC by tissue: distance, Enformer and the three published DNA language models. (c) LRB causal eQTL: raw test AUROC of the 15 fine-tuned models and their AUROC with distance held fixed, with the distance rule and a model on distance and genome annotation.

every one of them in GUE, Genomic Benchmarks or the original Nucleotide Transformer tasks, and on none of the five other BEND tasks. These 24 tasks come from 18 distinct data sets, because six yeast histone marks count in both GUE and the original Nucleotide Transformer tasks. For 20 of the 24, the rule’s 95% interval lies entirely above the best published score, which is reported without a comparable interval. A leak is demonstrated on 49 of the 81 tasks (Methods).

The first class is relatedness across splits. In the yeast histone-mark tasks, which GUE and the original Nucleotide Transformer tasks take from the same ten data sets, 60–96% of test sequences overlap a training sequence, although none is a full-length copy. A lookup that copies the label of the most similar training sequence therefore exceeds the best published score on eight of the ten marks in GUE and on six in the original Nucleotide Transformer tasks. Test sequences also recur verbatim across benchmarks: 90–91% of each GUE yeast test set occurs in the original Nucleotide Transformer training set with identical labels. For splice sites, 60% of GUE’s test set occurs verbatim in the original Nucleotide Transformer training set, and 91% of that Nucleotide Transformer task’s test sequences occur in GUE’s training and development data, so a model tuned on one suite has already seen the other’s test sequences. For Genomic Benchmarks’ human non-TATA promoters, 99% of negative test sequences but 6% of positive ones share an identical stretch with a training sequence, and the lookup reaches an accuracy of 0.991. Exact duplicates occur as well: half of the positive test sequences of both original Nucleotide Transformer enhancer tasks (in enhancer types, all of one enhancer class) and 76% of those of Genomic Benchmarks’ human Ensembl enhancers appear in training, verbatim or reverse-complemented. In the enhancer-promoter interaction (EPI) tasks of GUE+, 45–49% of test pairs have a shifted copy in training, a training pair that matches both their elements (Methods). Predicting an interaction exactly for those pairs, without using the training pairs’ labels, reaches a Matthews correlation coefficient (MCC) of 0.942–0.978 in the six cell lines, above the highest EPI value DNABERT-2 reports (0.929) [7]. For enhancer-promoter pairs of these cell lines, random splits were reported to inflate scores because many pairs share a promoter, in K562 and GM12878 [61], and a baseline sequence CNN fell to near chance in all six when chromosomes were held out [62]. We add the share of test pairs with a shifted copy in training and a rule that ignores labels yet exceeds every published score. In BEND’s chromatin, histone and

Table 4: Eukaryotic, genome-level and gene-level tasks in detail, with 95% intervals. Section 2.4: (a) DNALongBench enhancer–target genes, test split, with cross-validation over all 2,602 pairs; (b) evaluations by others that account for distance (Caduceus rows: our untrained distance rule on the same bins); (c) LRB causal eQTL, released predictions of the 15 fine-tuned models (AUROC; last row: released zero-shot scores over all SNPs). Sections 2.6 and 2.7: (d) BacBench resistance, (e) essential genes and (f) phenotypic traits. Share intervals are [lower end, upper end].

<i>(a) DNALongBench enhancer–target</i>	AUROC	AUPRC	<i>(b) Other evaluations (AUROC unless stated)</i>	Models	Distance
Distance	0.821	0.422	ENCODE-rE2G enhancer–gene benchmark (AUPRC)	0.493 (Enformer)	0.436
Published fine-tuned DNA language models	0.821–0.828	0.249–0.380	Caduceus evaluation of LRB, below 30 kb	0.668–0.679	0.672
Count of N characters in the input	0.823	0.440	Caduceus evaluation of LRB, beyond 30 kb	0.54–0.66	0.50
Cross-validation: gain of HyenaDNA over distance	+0.005 [−0.003, 0.013]		DNALongBench eQTL, unpublished manuscript (mean of nine tissues)		0.751
Cross-validation: gain over distance plus a per-gene prior	−0.007				
<i>(c) LRB causal eQTL</i>			14 DNA language models	Enformer	Distance
AUROC			0.697–0.732	0.754	0.650
Within distance strata			0.627–0.657	0.676	
Gain over distance			+0.047 to +0.079	+0.098	
Share interval, all 15 models			lower ends 0.31–0.40, upper ends 0.44–0.51		
Zero-shot over all SNPs			0.489–0.500		
<i>(d) Antibiotic resistance (AUROC unless stated)</i>			<i>(e) Essential genes, 55 genomes out of fold (mean per-genome AUROC)</i>		
Nearest-relative lookup, AUPRC	0.813		Gene length alone		0.540
Gain over the baseline stack: ESM-2 and ESM-C	at most 0.0023		Evo 2 at BacBench’s layer: share interval with gene length in the position model		[0.76, 0.78]
Gain over the baseline stack: Bacformer	0.0031		Gain over position: Evo 2 mid-depth layer	+0.189 [+0.169, +0.208]	
Gain over the baseline stack: Bacformer Large	0.013 [0.010, 0.017]		Gain over position: the other models	+0.048 to +0.187	
<i>(f) Phenotypic traits (macro AUROC)</i>			Gain over position: Mistral-DNA	+0.001	
Taxonomy prior	0.685 [0.665, 0.706]		Gain over position: Evo 2 at BacBench’s layer	−0.0006	
Bacformer minus ESM-2, overall	+0.026 [+0.018, +0.034]				
Bacformer minus ESM-2, within families	+0.005 [−0.016, +0.024]				

CpG methylation tasks, by contrast, at most 0.4% of test regions have a near-duplicate in training, and the lookup scores AUROC 0.542–0.710, far below the best published scores.

Input construction is the second class. In the mouse transcription-factor tasks of GUE, negatives are dinucleotide shuffles of positives, as the GUE paper describes [7], and for 73–82% of test sequences a shuffle twin sits in training with the opposite label. The twin rule alone (Methods) beats the best published score on two of the five tasks, and an untrained shuffle detector for sequences without a twin adds at most 0.06 MCC without changing that count. The non-TATA promoter tasks of GUE and all three promoter tasks of the original Nucleotide Transformer set build each negative from a positive by replacing segments of it [63], whereas GUE’s TATA negatives are genomic windows that contain a TATA motif [7]. For GUE’s core TATA promoter task, a quality-control tool had already reported that the classes share sequences, that one class, GUE’s negatives, carries a TATA box at fixed positions and that 6% of test sequences have a training sequence at least 90% identical [64]. In the original Nucleotide Transformer task on all promoters, 81% of the 41,350 segment-twin pairs carry opposite labels, and 91% of test sequences have a segment twin in training. The twin rule is above the best published score on one GUE promoter task. In Genomic Benchmarks’ human

Table 5: Further benchmarks (Section 2.5): tasks audited (86 in total), tasks on which a no-model rule or lookup exceeds the best published score (as defined in Section 2.5; ties within half a unit of the published precision are not counted) and examples, each on the benchmark’s split and metric, with 95% bootstrap intervals for the rules. The yeast histone marks of GUE and of the original Nucleotide Transformer tasks are the same ten data sets, split differently, and are counted in both rows.

Benchmark	Tasks	Rule above all	Example: rule [95% interval] against the best published score (metric)
GUE [7], with the GUE+ species tasks	30	11	Yeast H4ac: lookup 0.794 [0.774, 0.816], DNABERT-2 (6-mer) + RandomMask 0.672 [68]; mouse TF 4: twin rule with shuffle detector 0.832 [0.809, 0.856], DNABERT-2 0.761, its Table 9, which states no metric (0.527 in its Table 6) [7] (MCC)
GUE+ enhancer–promoter interaction	6	6	Both-element rule 0.942 [0.928, 0.956] to 0.978 [0.969, 0.987] in the six cell lines; highest EPI value of DNABERT-2 0.929 [7] (MCC)
Nucleotide Transformer, original [59]	18	6	Yeast H3K4me1: lookup 0.707 [0.682, 0.730], Omni-DNA 0.617 [69] (MCC)
Nucleotide Transformer, revised [6]	18	0	None
Genomic Benchmarks [60]	9	1	Human non-TATA promoters: lookup 0.991 [0.989, 0.993], FinDNA 0.974 [70] (accuracy)
BEND [13], other tasks	5	0 (1 tie)	Disease variants: distance to the nearest exon–intron junction 0.981 [0.980, 0.983], PhyloGPN, PhyloGPN-X and GPN-MSA 0.98 [65], a tie (AUROC; best language model in the BEND paper: NT-MS, 0.77). Gene finding: 2.7% of test nucleotides in training regions, coordinate lookup MCC −0.04. Chromatin, histone and CpG methylation: near-duplicates 0.4%, 0.0% and 0.3%, homologs 6.5%, 7.4% and 31.7%, lookup AUROC 0.568, 0.542 and 0.710

Ensembl regulatory task, the three classes differ in mean length (565, 395 and 326 bases). Length alone classifies test sequences with an accuracy of 0.874, against 0.370 for the majority class.

Position is the third class. For BEND’s disease variants, the distance of a variant to the nearest GENCODE exon–intron junction scores AUROC 0.981 [0.980, 0.983] without using any labels. It ties the best published scores, 0.98 for PhyloGPN, PhyloGPN-X and GPN-MSA [65], and lies far above every language model in the BEND paper (at most 0.77) [13]. Of the pathogenic variants, 92% lie within 2 bases of a junction, against 0.2% of the benign ones. The no-model baseline for this task is the annotated consequence class of a variant, which scores 0.985 but is fitted on labels of the evaluated variants, each class’s pathogenic fraction being estimated on the other half of the chromosomes. It is not counted, because BEND evaluates this task zero-shot and only rules without labels are compared with the published scores. On ClinVar synonymous variants, the distance to the nearest exon junction was shown to exceed every neural representation tested [66], and TraitGym noted that expert-reviewed pathogenic ClinVar variants are skewed towards coding and splice-region variants and that BEND includes no conservation-based baseline [24]. Here the same rule is applied to BEND’s disease-variant task, where it reaches the best published scores.

The revised Nucleotide Transformer tasks hold out chromosomes 20 and 21 [6]. Their maintainers changed the tasks during peer review, replacing randomly generated negatives with genomic sequences and adding chromosome held-out test sets [67]. On them, no rule or lookup beats the best published score on any of the 18 tasks, and at most 61 of 3,000 test sequences in a task have a near-duplicate in training. Only the three splice tasks show a demonstrated leak: canonical dinucleotides at fixed positions reach MCC 0.852–0.893, below the best published scores.

2.6 Relatedness confounds genome-level tasks

BacBench’s antibiotic-resistance task splits genomes at random, and under this split 81% of test genomes have a training genome at Jaccard similarity of at least 0.9. A nearest-relative lookup with no model, which averages the labels of the most similar training genomes, beats all 12 published models on AUPRC and 11 of them on AUROC (AUROC 0.918; simple genome statistics alone score

0.859). A stack of genome statistics, a species prior and the nearest-relative lookup beats every published model (AUROC 0.929; Table 1). Added to it, ESM-2, ESM-C, Bacformer and Bacformer Large gain at most 0.013 AUROC (Bacformer Large; Table 4d).

The phenotypic-trait task holds out genera, and a prior that uses only family, order, class and phylum beats 8 of the 12 models (macro AUROC 0.685). Within families, the macro AUROC of Bacformer falls from 0.714 to 0.565, a loss of 70% of its above-chance AUROC, and its lead over its input model, ESM-2, shrinks to a difference whose interval includes zero (Table 4f). Genome-level benchmarks therefore need splits held out by clade and a nearest-relative baseline, as proposed for predicting resistance from genomes [71].

2.7 Gene-level tasks are not confounded by position, and models add real signal

A model on gene-layout features alone (strand, gaps and strands of the neighbours, operon context, distance to the replication origin, leading or lagging strand, local gene density and replicon) scores 0.645 median AUROC on the essential-gene test genomes of BacBench, above only three of the 12 published models. We evaluate all 55 genomes out of fold and read Evo 2 at BacBench’s layer, its final layer and its mid-depth layer (Table 9 lists the reproduction). Position, without gene length, accounts for at most 0.06 of the models’ above-chance AUROC at the lower end of the share interval and at most 0.10 at the upper end. Evo 2, read at BacBench’s layer, is barely above chance (AUROC 0.546), and its small excess comes mostly from gene length, which BacBench’s pooling builds into the embedding (Table 4e). The mid-depth layer of Evo 2 gains the most over the position model, +0.189 AUROC, and every other model gains at least +0.048 except Mistral-DNA and Evo 2 at BacBench’s layer. A Bacformer whose query gene is masked keeps a within-position AUROC of 0.739, which indicates that the identity of neighbouring genes, not their layout, carries this signal.

For KEGG orthology, a nearest-neighbour classifier on gene layout falls far below zero-shot ESM-2 and ESM-C (macro F1 0.010, Table 1; ESM-2 and ESM-C in Table 9).

2.8 Bacformer’s masked-gene predictions depend mainly on neighbouring genes in order

Bacformer is pretrained to predict the protein family of a masked gene, one of 50,000 clusters, from the other genes in order, with rotary position embeddings over gene index and no strand or coordinates [3]. We evaluated the released Bacformer and Bacformer Large under context ablations on the masked genes of the 53 test genomes of BacBench PPI (180,542 for Bacformer). These genomes were not used for label reconstruction or by the n -gram below. Scores are computed against labels reconstructed from each model’s own predictions (Methods), on which Bacformer’s top-1 accuracy is 0.311. Shuffling the genome while keeping its genes leaves 2% of the gain over a prediction without context. Keeping only the ± 1 neighbours in order and shuffling the rest leaves 65% of the gain, keeping ± 2 leaves 90% and keeping ± 4 leaves 98% (Figure 5a). Masking those neighbours instead leaves 39%, 14% and 3% for ± 1 , ± 2 and ± 4 , respectively.

A neighbour-copy n -gram, which has no transformer and is built from co-occurring families in training genomes disjoint from the test set, reaches 51% [47, 55] of the gain of Bacformer with 208 training genomes (Figure 5b). With 1,870 training genomes it reaches 80% [74, 87], and its top-1 accuracy is still rising by about 0.03 per doubling. Bacformer predicts genes with an operon-like neighbour with a top-1 accuracy of 0.337, compared with 0.149 for genes without one, although it sees neither strand nor gaps. Of Bacformer’s 96 attention heads, 62 attend most to the adjacent genes. Copying the target from neighbours is possible because operon partners and conserved neighbourhoods recur across genomes [17]. In our pretraining runs (Section 2.9) the baseline variant,

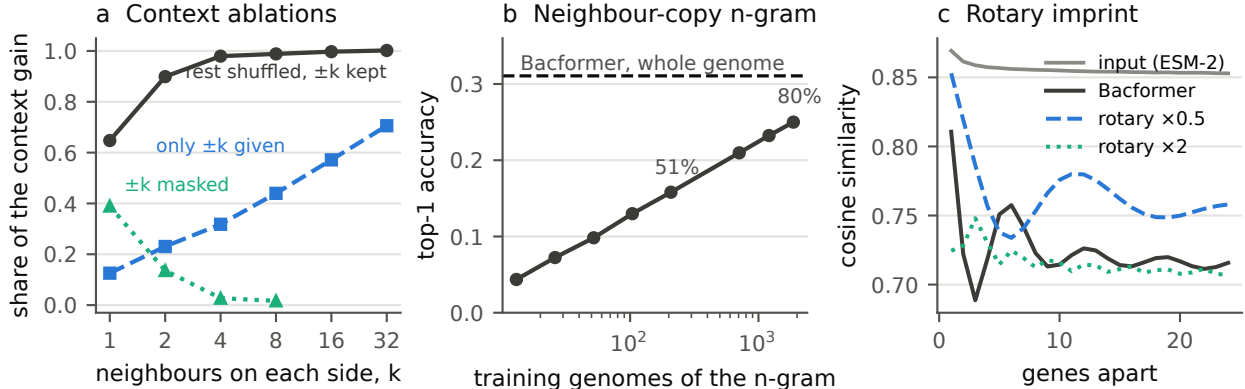


Figure 5: Bacformer’s masked-gene objective and its position signal. (a) Share of the masked-gene accuracy gain over a prediction without context that remains when the genome is shuffled except the $\pm k$ neighbours, when only the $\pm k$ neighbours are given, and when the $\pm k$ neighbours are masked (the 53 test genomes of BacBench PPI). (b) Top-1 accuracy of a neighbour-copy n -gram (± 4) against the number of training genomes, with its share of the gain of Bacformer at 208 and 1,870 genomes. (c) Cosine similarity of two genes’ embeddings against how many genes apart they are (1 = adjacent): input ESM-2 embeddings, Bacformer, and Bacformer with rotary positions scaled by 0.5 and 2.

which follows the Bacformer recipe, predicts 42% of masked genes when single genes are masked but 19% when whole operon-like units are, and the variant trained on shuffled gene order predicts 10%.

Our explanation is that position enters the embeddings as a by-product: the contextual embedding of a gene absorbs its neighbourhood, and the rotary encoding adds an imprint of relative position. The similarity of two genes’ embeddings oscillates with their distance in genes, and rescaling the rotary positions by s moves the peaks as predicted, to every $2\pi/s$ genes, the period of the fastest rotary frequency of one radian per gene (Figure 5c). The operon and interaction benchmarks reward this signal.

On zero-shot PPI, shuffling the genome lowers the raw AUROC of Bacformer Large from 0.721 to 0.627. Yet when each protein is embedded without any genome, the within-position AUROC of Bacformer Large stays at 0.625 against 0.628 with the genome, and its gain stays at +0.055 against +0.053. For Bacformer Large, we conclude that genomic context adds position and that the signal surviving position control comes from the proteins. For Bacformer, by contrast, a genome in any order raises the within-position AUROC from 0.535 for the protein alone to 0.595 (0.604 shuffled).

2.9 Larger models and tested pretraining changes do not remove the confound

We embedded the genes of the long-read operon, DGEb and zero-shot PPI tasks with models of increasing size from seven families. The families are ESM-2 (8M–3B) [72], the Nucleotide Transformer (v2 50M–500M and v1 2.5B) [6], gLM2 (150M and 650M) [2], GenomeOcean (100M–4B) [73], GENERator-v2 (1.2B and 3B) [74], Bacformer and Bacformer Large, and Evo 2 (1B–40B) [5]. On long-read operons no model is above chance with position held fixed, except Evo 2 1B at its final layer, and no larger model rises above chance (Table 6a). Larger ESM-2 and Evo 2 models score lower at fixed position (Table 6a, b), and the first-generation Nucleotide Transformer 2.5B scores above the second-generation 50M model only through position (Table 6b). For DGEb, no model gains more than +0.004 AUROC over position averaged over the three genomes, and the final layer of Evo 2 gains at most +0.0017. At 40B, the within-position AUROC of that layer lies above chance (Table 6a).

On zero-shot PPI, the final layers of the DNA models stay near or below chance within position strata, and ESM-2 loses within-position AUROC as it grows (Table 6b). Mid-depth layers of the causal DNA models (Evo 2, GenomeOcean, GENERator-v2), standardised within each genome, are

a positive exception. That of Evo 2 7B reaches a raw AUROC of 0.736, above the best BacBench model (0.721). In GenomeOcean their signal beyond position grows with size (Table 6b). The gains over position of the Evo 2 mid-depth layer at 1B and 7B are the largest of any zero-shot embedding tested, up to +0.092 AUROC, and at 40B the gain falls below that of 7B (Table 6a). This readout is the output of the fixed block $L/2$, where the 40B vectors are dominated by a few directions (Supplementary note). We ran no layer sweep, so the drop does not separate size from the choice of depth. Neither the mid-depth read nor the standardisation gives this signal alone, since for Evo 2 1B the unstandardised mid-depth layer scores 0.478 and the standardised final layer scores 0.558. Codon-usage similarity alone reaches a within-position AUROC of 0.542. Codon usage alone therefore does not account for this signal. Per-genome standardisation can still let shared composition drive similarity, and the STRING labels carry genome context of other genomes. The same Evo 2 layers, by contrast, stay at or below chance within position strata on long-read operons (Table 6a).

We then changed the pretraining itself. We trained six variants of a half-depth Bacformer-like model, three seeds each, on 10,286 genomes from genera disjoint from the evaluation genomes. The design was fixed before training (Methods), with these variants: the Bacformer recipe (the baseline), strand and gap inputs, masking of whole operon-like units, masking of random spans of matched length, both changes together, and shuffled gene order. The baseline passes a positive control, a raw PPI AUROC above that of its ESM-2 input (Table 6c). For long-read operons, adding the baseline’s similarity to the position model lowers the position model’s AUROC by 0.013. Operon-unit masking or shuffled order avoids that loss only by losing within-position signal. The position share of every variant on these operons is at least 0.86, and by the pre-specified decision rule (Methods) none of the variants learns operon structure beyond position. On STRING pairs far apart without genome-organisation evidence, every variant scores a lower AUROC than its ESM-2 input, significantly so for all but the strand-and-gap variant. By contrast, on experiment-only interactions the baseline reaches a within-position AUROC of 0.599, against 0.533 for ESM-2, mostly (about two thirds in an exploratory split of distant pairs) on pairs with STRING genome-context evidence from other genomes. Strand and gap inputs give the only improvement over the baseline that meets the decision rule (+0.013, Holm-corrected $p \leq 0.02$; Table 6c).

2.10 A protein language model with position matches or beats genomic models on operon benchmarks and, with transferred labels, on interaction benchmarks

If genomic models mostly learn position plus protein identity, a protein language model given position should match them. Our model uses one ESM-2 35M embedding per gene (the input stage of Bacformer), the gene coordinates, a few hand-crafted features, principal components of each gene’s embedding and, where a benchmark has training genomes, labels transferred from orthologs in those genomes (Methods). The hand-crafted features are strand and gaps, layout context (gene lengths, flanking gaps and the direction run), a label-free conserved-neighbourhood score from reciprocal best hits in genomes of other genera, and the cosine between the embeddings of a pair. The ESM-2 model with position matches or beats the best model on the long-read operon, database operon, DGEb and BacBench interaction benchmarks (Table 7). On essential genes it falls below the best AUROC but exceeds the best AUPRC. Per genome, its only neural step needs 55 to 1,323 times less compute than gLM2, Evo and Evo 2 1B. On long-read operons, its gain over position is consistent with the label artefact of Section 2.3, because its extra features lower AUROC on the nanopore labels.

Table 6: Size series and pretraining variants (Section 2.9); 95% intervals, size differences paired over resamples.

<i>(a) Evo 2 by size and readout</i>			
Long-read operons, within-position AUROC, final layer	1B 0.517 [0.502, 0.534]	7B	40B
Long-read operons, within-position difference from 1B, final layer		-0.089 [-0.107, -0.070]	-0.052 [-0.065, -0.039]
Long-read operons, within-position AUROC, mid-depth	0.456–0.460 (1B and 7B)		0.494
DGEB, final layer, gain over position (3-genome mean)	+0.0017	+0.0004	+0.0006 [+0.0002, +0.0011]
DGEB, largest gain of any model on one genome (Evo 2 1B mid-depth, <i>E. coli</i>)	+0.0075		
DGEB, final layer, within-position AUROC (3-genome mean)			0.584 [0.548, 0.626]
DGEB, final layer, average precision (3-genome mean)			0.482
Zero-shot PPI, within-position AUROC, mid-depth	0.681	0.686	0.601
Zero-shot PPI, gain over position, mid-depth	+0.088	+0.092 [+0.079, +0.104]	+0.034 [+0.028, +0.041]
Zero-shot PPI, gain difference from 7B, mid-depth			-0.057 [-0.066, -0.050]
<i>(b) Other size series (within-position AUROC)</i>		<i>(c) Pretraining variants</i>	
Long-read: ESM-2 3B minus 8M	-0.034 [-0.056, -0.014]	Raw PPI AUROC: baseline	0.725
Long-read: NT v1 2.5B minus v2 50M	-0.006 [-0.025, +0.009]	Raw PPI AUROC: its ESM-2 input	0.569
Long-read: the same, raw AUROC	+0.093	Long-read AUROC lost on adding the baseline’s similarity to the position model	0.013 [0.009, 0.017]
PPI: final layers of all DNA models (Evo 2 included)	0.42–0.53	Far STRING pairs without genome-organisation evidence: every variant	0.446–0.498
PPI: ESM-2 8M	0.607	The same pairs: ESM-2 input	0.512
PPI: ESM-2 3B	0.498	Strand and gap inputs: gain in experiment-only within-position AUROC	+0.013 [+0.004, +0.020]
PPI: GenomeOcean 100M mid-depth	0.615		
PPI: GenomeOcean 4B mid-depth	0.662		
PPI: GenomeOcean 4B minus 100M	+0.047 [+0.035, +0.060]		

On the BacBench PPI split, the 0.924 AUROC of the ESM-2 model with position comes from transferring training labels between orthologs. This transfer re-derives how STRING builds its transferred channels [16], and a leakage check found no test labels or same-genus genomes in the reference panels (the training genomes that supply orthologs and labels). Without transferred labels, position with the cosine or with conserved neighbourhood scores below the best model on AUROC but above it on AUPRC (AUROC 0.742–0.760 against 0.784).

The model with transferred labels does not carry over to experiment-only labels, with median AUROC 0.552 on two-hybrid and 0.512 on AP-MS screens. A version without transferred labels (features from position, cosine and conserved neighbourhood, also fitted on STRING labels) exceeds the cosine of the best genomic model on AP-MS and comes within 0.005 AUROC of it on XL-MS (Table 7). It falls below that cosine on two-hybrid screens (0.512 against 0.547).

2.11 A position-aware evaluation standard

seqleak, released with this paper, takes items with position features and per-item scores and returns the no-model baseline, the within-position AUROC, the gain over position with a Wilcoxon test

Table 7: Position alone, the best model and an ESM-2 35M model with position and hand-crafted features (Methods), each on the protocol of the benchmark (AUROC unless stated). “Best model” is the best published value, protein models included, or the best of our re-runs where none is published (RegulonDB, SubtiWiki, nanopore and experiment-only labels), counting Evo 2 readouts run on the same labels and protocol (not the raw-cosine RegulonDB readouts of Section 2.2). For nanopore operons, position is fitted on the long-read labels of other species. Experiment-only rows give position models as ranges over data sets and the other columns as medians, with screened-space negatives; the ESM-2 model with position there has no transferred labels.

Task	Position only	Best model	ESM-2 with position
Long-read operons, 5 strains	0.833	0.776 (Bacformer)	0.854
RegulonDB; SubtiWiki operons	0.958; 0.966	0.857; 0.861	0.961; 0.969
DGEB operonic pairs, AP (3 genomes)	0.854; 0.990; 0.944	0.626; 0.409; 0.543	0.919; 0.987; 0.951
Nanopore operons, 3 genomes	0.987–0.988	0.884–0.935	0.955–0.980
BacBench PPI split (AUROC; AUPRC)	0.713; 0.516	0.784; 0.508 (Bacformer Large)	0.924; 0.814
Experiment-only PPI: two-hybrid;	0.519–0.535;	0.547; 0.622; 0.767	0.512; 0.652; 0.763
AP-MS; XL-MS	0.543–0.599; 0.580–0.778		
Essential genes (median AUROC; AUPRC)	0.645; 0.263	0.8245 (ProtBERT); 0.641 (gLM2)	0.783; 0.764

over genomes and the share interval, each with bootstrap intervals (Methods). It reproduces Tables 2 and 5 of the companion paper exactly (Table 9). Table 8 lists, for each audited benchmark, the defect, the evidence and a fix. Together the fixes amount to three practices for benchmark maintainers: report the no-model baseline and position-held-fixed metrics next to every model, with a nearest-relative baseline for genome-level tasks; state how labels were made and include labels built without layout, with positive controls; and design inputs and splits so that position and relatedness cannot leak.

3 Discussion

How the labels of a benchmark were made largely decides how much of a model’s score reflects what it learned. When labels follow genome layout, position explains most of the models’ scores on long-read and predicted operon labels and part of their STRING interaction scores. When labels were made without layout, the models add almost nothing to position on the expression-filtered nanopore operon labels, and on interactions little survives once hub effects are controlled and ribosomal proteins and conserved neighbours are removed. On gene-level tasks the results indicate that the signal comes from the identity of the genes and their neighbours, not from their layout.

Nevertheless, the audit also finds real signal: in fine-tuned DNA models on the LRB causal eQTL task, in the standardised mid-depth layers of Evo 2 and in Bacformer Large on STRING labels. A residual co-complex signal also survives strict controls. This signal is often not specifically genomic (Sections 2.3, 2.4 and 2.8): on membrane AP-MS and cross-linked pairs, the residual is about as large for single-protein models as for the best of the 12 models, Bacformer Large keeps its gain over position when its proteins are embedded without a genome, and on LRB, distance plus genome annotation already comes close to the weakest DNA model.

Scaling up and the pretraining variants of Section 2.9 left the position confound in place. Because a masked-gene objective can be solved by copying from neighbours, our explanation (Section 2.8) is that the model comes to represent its neighbourhood, which benchmarks built on neighbourhood then reward. This would also explain why a protein language model with position and a few hand-crafted features does at least as well as the genomic models on operon benchmarks, and on interaction benchmarks once it is given labels transferred from orthologs (Section 2.10).

Table 8: Defects and fixes. For each audited benchmark: the defect, where the evidence is reported, and the fix. BacBench’s OperonDB windows and BEND’s enhancer annotation have the defects of the gLM and BEND expression-variant rows (Table 1). NT stands for Nucleotide Transformer. Fixes marked “recommended” were not tested here; the others were applied in this paper.

Benchmark (task)	Defect	Evidence in	Fix
DNALongBench (enhancer–target)	Trailing N padding to 450 kb encodes distance	Section 2.4, Table 4a	Input length independent of distance (HyenaDNA then falls to 0.602); distance and N-count baselines
DNALongBench (eQTL)	The same padding, noted in the findings log of an unpublished audit [58]	Section 2.4, Figure 4b	As for enhancer–target (recommended)
DGEB (operonic pairs)	Two of three label sets predicted mainly from distance	Section 2.2, Figure 2, Table 2b	Label provenance; experimentally supported units evaluated separately; position row, within-position AUROC
gLM operon set (<i>E. coli</i> pairs)	No position-only baseline (labels from operon annotation)	Section 2.2, Table 2a	Position row, as here; within-position metrics (recommended)
BacBench (long-read operons)	Pairs without spanning reads, unexpressed ones included, labelled negative	Section 2.3, Table 3c	Negatives only between expressed genes; position row
BacBench (zero-shot PPI)	STRING labels scored partly from conserved gene neighbourhood	Section 2.2, Table 2c	Share intervals; experiment-only labels; pairs with genome-context evidence reported separately
BacBench (resistance)	Random split; 81% of test genomes have a training genome at Jaccard ≥ 0.9	Section 2.6, Table 4d	Nearest-relative and baseline-stack rows; clade split (recommended)
BacBench (traits)	Genera held out, but higher taxonomic ranks predict traits	Section 2.6, Table 4f	Prior and within-family AUROC; family hold-out (recommended)
BEND (expression variants)	Distance is a strong baseline	Section 2.4, Table 1	Distance baselines and within-distance metrics
LRB (causal eQTL)	None found	Section 2.4, Table 4c	Distance row and within-distance metrics, as here
GUE; NT original (yeast histone marks)	Test windows overlap training windows	Section 2.5, Table 5	Hold out chromosomes, as the revised NT tasks do, or split by homology [75] (recommended)
GUE (mouse TF, non-TATA promoters); NT original (promoters)	Negatives are shuffled or segment-substituted copies of positives, with twins across splits	Section 2.5, Table 5	Remove twins or draw negatives from other loci, as the revised NT tasks do (recommended)
GUE+ (enhancer–promoter interaction)	Training pairs share both elements with test pairs	Section 2.5, Table 5	Split so that no element recurs across splits (recommended)
Genomic Benchmarks; NT original (enhancers)	Exact or overlapping copies across splits; class-dependent length (regulatory task)	Section 2.5, Table 5	Remove exact duplicates, near-duplicates and overlapping windows across splits; length-matched classes (recommended)
BEND (disease variants)	92% of pathogenic but 0.2% of benign variants lie within 2 bases of a junction (canonical splice sites)	Section 2.5, Table 5	Consequence-class and junction-distance baselines, as here; within-consequence AUROC or consequence-matched benign variants, as in TraitGym [24] (recommended)

The defects share one form. Something available without a model predicts the label (distance, gene layout, relatives or near-copies of test items in training, or the way negatives were built), and the benchmark does not separate it from what a model adds. The fixes in Table 8 follow from this, and the Nucleotide Transformer maintainers applied such fixes when they revised their tasks.

Some of these findings have been reported before. The Bacformer authors note a proximity bias in interaction data and train on pairs balanced across distance bins, but report no position-only baseline [3]. Population structure was known to confound resistance prediction from genomes [71], and sequence-based predictors of protein interactions learn node degree and sequence similarity on random splits [76]. Earlier reports of the DNALongBench eQTL result and of individual leaks in the further benchmarks are cited with each result (Sections 2.4 and 2.5). Besides the evaluations named in the Introduction, the Caduceus evaluation of the LRB causal eQTL task accounts for distance by reporting distance bins [9]. ProteomeLM learns interaction signal without positional encoding along the genome [77], consistent with our finding that the position-free part of the models’ interaction signal comes from proteins. What is new here is one protocol across task classes, the provenance of the DGEb labels, the padding leak in the enhancer–target task, label sets built without layout, and a share interval that is valid whatever the causal direction.

Our analysis has several limitations. First, the share is an interval, wide on some single-genome tasks, and within-position operon AUROCs often rest on less than 2% of positive–negative comparisons. Second, the experiment-only labels have approximate screened spaces and no tested negatives. Their three structure-screen sets are predictions from alignments paired across genomes, through which cross-genome neighbourhood can enter. STRING’s text-mining channel, one of our positive controls, partly draws on the same literature. The nanopore labels cover three genomes in few conditions. Eukaryotic test splits are small, although cross-validation over all pairs agrees with them. Most models are read at one layer, and the mechanism is shown for Bacformer with reconstructed labels. The n -gram saw 1,870 genomes, compared with about 1.3 million behind Bacformer, so its share is a lower bound. The pretraining variants test the objective at a small budget. Choosing the best of many models makes “best model” values optimistic, and we correct for multiple comparisons only in the pretraining experiment’s pre-specified decision rules. Published scores on the further benchmarks come from varied protocols, such as cross-validation or several seeds, and each rule there is the best of up to six families per task. We also did not check whether test genomes occur in pretraining corpora, an overlap that inflates interaction scores of models built on pretrained protein language models [78].

4 Methods

Benchmarks and models

BacBench tasks use the datasets and code of the benchmark [32], KEGG orthology is rebuilt from the Bacformer preprint [3], and operon, gLM and BacBench PPI analyses follow the companion paper [31]. DGEb uses its code and the dataset revisions pinned in its release [10]. DNALongBench uses its released data [79, 80] and its loader [11], LRB uses its released per-item predictions and zero-shot scores [12], and BEND uses its released data [13]. The 12 BacBench models are embedded as in the companion paper. Every Evo 2 1B result runs with its FP8 input projections on (Table 9). In the size series, Evo 2 embeds each gene from an 8,192-bp window that ends at its last base, GenomeOcean and GENERator-v2 embed it from the full window, and the Nucleotide Transformer embeds it as in BacBench. The causal models are also read after their middle block, standardised within each genome, and PPI size-series results use a fixed random half of the genes.

seqleak

The position model (row zero in seqleak’s output) is a logistic regression on cubic B-spline bases of the position features, separate for same- and opposite-strand pairs, fitted to convergence and evaluated out of fold (five genome-grouped folds for more than ten genomes, leave one genome out for two to ten, five contiguous chunks for one genome). The within-position AUROC compares positives only with negatives of the same genome and stratum, weighting strata by positives $\times$ negatives. Strata are strand state $\times$ 17 gap bins for neighbouring pairs, and 11 genes-apart bins $\times$ strand switches $\times$ 9 bins of the largest intervening gap $\times$ strand state for other pairs. For other tasks, strata are 20 quantile bins of the out-of-fold position model within each genome. Gene-level position features leave out gene length. With gene length added, the upper ends of the essential-gene shares reach at most 0.13. The exception is Evo 2 read at BacBench’s layer (0.78), whose pooling scales the embedding with gene length. Under seqleak’s coarser gene strata, the upper ends reach 0.13, or 0.18 with gene length, and 0.25 for that near-chance readout.

Metrics are computed per genome and averaged unless a benchmark pools them. The gain over position is the per-genome AUROC of a position-plus-score fit minus that of the position fit on the same folds, with a Wilcoxon signed-rank test over genomes. With raw AUROC r , within-position AUROC w and the AUROC e of the score replaced by its percentile within each (genome, stratum) cell, $s_{\text{within}} = 1 - (w - 1/2)/(r - 1/2)$ and $s_{\text{eq}} = 1 - (e - 1/2)/(r - 1/2)$. The share interval is $[\text{clip}(\min(s_{\text{within}}, s_{\text{eq}}), 0, 1), \text{clip}(\max(s_{\text{within}}, s_{\text{eq}}), 0, 1)]$, reported when $r - 1/2 > 0.02$. In simulated genomes with known shares (30 genomes per setting), the interval contained the true share to within 0.025 in every case except scores made only of position, whose upper end was 0.94–0.97 instead of 1 because strata leave some position signal inside them.

Intervals are 95% percentile intervals from one bootstrap plan per analysis, resampling genomes when there are ten or more and otherwise blocks of 50 neighbouring items (at most 1,000 blocks per genome). The experiment-only and nanopore analyses resample 50 contiguous gene blocks per genome (block pairs for gene pairs) and use five folds grouped by these blocks. The RegulonDB, SMRT-Cappable-seq and BacBench long-read gains use folds built the same way. All these folds, like those of the LRB distance-plus-annotation baseline below, were assigned by GroupKFold, and the values we report with our folds depend on this assignment (Supplementary note). Genome-level intervals are 95% percentile intervals over 10,000 resamples of drugs (antibiotic resistance) or traits (phenotypic traits). When folds split a genome, out-of-fold predictions are compared only within their fold.

New label sets

Experiment-only positives come from IntAct and MINT, BioGRID, journal supplements (including a computational structure screen), cross-linking files and Complex Portal [47, 48, 81]. Negatives are pairs in the screened space of a screen without a reported interaction, or degree-preserving rewirings of its positives. Degree-preserving negatives follow Bernett et al. [76], who found that interaction predictors trained on random splits learn node degree and sequence similarity. Distant pairs lie more than 20 genes apart and outside known operons. The strict subset also removes ribosomal proteins and pairs with any STRING neighbourhood, fusion or co-occurrence evidence. Positive controls score the same pairs with STRING’s transferred experiments, transferred database entries, co-expression, text mining and their combination, or with the positives of the other data sets of the same genome.

Nanopore reads were mapped with minimap2 [82]. For a same-strand adjacent pair, the spanning fraction is the share of reads covering the start of the downstream gene that also extend at least 30 nt into the upstream gene. Both genes need at least 10 reads, positives have a fraction of at least

0.5, and negatives a fraction below 0.1. The strict variant requires at least 0.8 for positives and less than 0.02 for negatives, the lenient variant requires 0.25 and 0.2, and the unfiltered variant keeps the positives and calls negative every other pair except partly spanned ones.

Task-specific analyses

DGEB entries were mapped to their genomes to obtain strand and gap. The label rules are fitted to the labels they reproduce (in sample), so they describe the labels rather than predict them. DGEB AP gains add a model’s standardised score to a strand $\times$ gap-spline model trained on the other two genomes. For DNALongBench, distance is the unpadded input length and inputs are rebuilt following the loader’s code. HyenaDNA (medium, 450 kb) was fine-tuned for five epochs on the official split, where it is read at its best validation epoch, and in five-fold cross-validation over all pairs, where its gains over distance compare logistic models with and without its score, with intervals from a bootstrap over genes. The distance-hidden input joins the element and TSS windows by a fixed 1-kb N spacer. For LRB, distance is the released distance to the nearest TSS, and the distance-plus-annotation baseline is a logistic regression on a spline of log distance and 18 released annotation flags, fitted out of fold on the test chromosomes (five groups of 1-Mb blocks), the only items with released annotations. For LRB the position model is linear in log distance and matches the untrained rule within folds (AUROC 0.647), whereas a spline position model underfits (0.630) and gives gains 0.008–0.015 larger.

BacBench’s drug filter on the released labels keeps 37 drugs, 6 of which differ from the 37 of the preprint’s Table S17, and trait results use BacBench’s three seeds, where Table S16 used five. Within-family analyses refit the BacBench heads in PyTorch to obtain per-genome predictions. The nearest-relative lookup, a form of genomic neighbour typing [83], sketches genomes with MinHash ($k = 21$) [84] and averages the labels of the most similar training genomes weighted by Jaccard similarity. Essential-gene layout features feed a gradient-boosted tree model, evaluated out of fold over 55 genomes like the language models, each with its own genus-grouped five-fold split.

Because the cluster centroids of Bacformer were not released, family labels were reconstructed from each model’s own predictions on 208 training genomes (24,550 recovered clusters for Bacformer). Context-ablation shares compare each condition with full and no context on the same masked genes. The Evo 2 readouts of Section 2.3 leave out the near-zero outputs of the 7B and 40B models at BacBench’s layer. The pretraining corpus has one complete RefSeq genome per species, at most 50 species per genus and no species or genus of the 69 genomes used to evaluate the variants, and the models have six layers and train for 15,000 steps of 32 genomes. A variant counts as improved only if it learns beyond position and beats the baseline, with a paired interval above zero after Holm correction over five variants and four endpoints. Analyses use PyTorch, Hugging Face Transformers, scikit-learn [85], NumPy, pandas and SciPy.

ESM-2 model with position

The ESM-2 model with position of Table 7 is a gradient-boosted model on position, the conserved-neighbourhood score and the cosine of a pair. For long-read operons it adds layout context (gene lengths, flanking gaps and strands, direction run) and 16 principal components of each gene’s embedding. For nanopore operons it adds 16 components, and for BacBench PPI it adds labels transferred from orthologs. For essential genes it uses the feature set selected on validation (the gene-layout features of the position model with gene length, transferred labels and 32 components; AUROC at most 0.802 over the eight sets), and for RegulonDB and SubtiWiki it is the additive position and conservation model. On experiment-only labels it uses position, cosine and conserved

neighbourhood without transferred labels. Orthologs are reciprocal best hits by embedding cosine, and labels are transferred only from training genomes of other genera.

Further benchmarks

GUE and GUE+ come from the authors’ archive, the original and revised Nucleotide Transformer tasks from their Hugging Face repositories, Genomic Benchmarks from the Hugging Face copies of its nine data sets and BEND from its archive (versions under Data and code availability). A test sequence is an exact duplicate if it or its reverse complement occurs in training. A test sequence occurs verbatim in another benchmark when it equals, ignoring case, a training sequence of that benchmark. Each test sequence is searched against all training sequences with BLASTN 2.17.0 [86] on both strands (E below 10^{-5}). Tasks whose median sequence length is at most 150 bases use blastn with word size 11, and the others use dc-megablast. GUE+ pairs are searched with megablast (E below 10^{-20}), and so are BEND regions (E below 10^{-10}). For BEND’s chromatin, histone and CpG tasks, the lookup scores each label by the mean label of the training hits within 90% of the best bit score, or by the training prevalence without a hit, and for gene finding it copies the label at the same genomic coordinate of an overlapping training region. A test sequence overlaps training if its best hit is fully identical over 5–99% of its length, and shares an identical stretch with training if its best hit is fully identical over part or all of its length. It is a near-duplicate at 95% identity or more over at least 80% of its length, and a homolog at 80% identity or more over at least half of it. The lookup copies the label of the best hit (highest bit score, majority vote among tied hits, the training majority class without a hit).

Shuffle twins have equal length, identical counts of all 16 dinucleotides and the same first and last base, the invariants of a dinucleotide shuffle. Segment twins, the pairs that segment substitution leaves, have equal length and share at least five identical 12-mers at the same offsets, counted over all offsets and ignoring 12-mers that occur in more than 20 sequences at one offset. The twin rule copies or flips the label of a test sequence’s training twin, following the label agreement of training pairs of the same twin type: it copies the label if more than three quarters of them agree, flips it if fewer than a quarter do, and ignores other twin types. Without a usable twin, it falls back to the training majority class when the twin rule is reported alone, and otherwise to the single feature that scores best on the training split among length, N count and share, leading and trailing N, lower-case count and share, count of other characters, GC content, CpG observed-to-expected ratio and density, the fifth-order Markov log-probability and an untrained shuffle detector. The detector scores the per-base log-probability under a fifth-order minus a first-order Markov chain, both estimated from 30 Mb of random GRCh38 sequence, so it sees no task sequence or label, and dinucleotide shuffling lowers it. The fixed-position motif rule chooses up to four presence indicators of a k -mer ($k \leq 3$) at a fixed position by training metric and predicts the majority training label of each combination. The combined rule uses the lookup where a close training sequence exists and a single-feature or positional k -mer rule otherwise. All thresholds are fitted on the training split. Table 5 counts, per task, the best of six pre-specified rule families (lookup, twin rule, best single feature, best artefact feature such as length or N count, fixed-position motif rule and combined rule), chosen on the test split, as is the best of the combined rule’s four variants. On every counted task, the best rule is the lookup, the twin rule or, for GUE+ pairs, the both-element rule. A GUE+ test pair has a shifted copy in training when one training pair matches both its enhancer and its promoter at 95% identity or more over at least half of each element (BLASTN megablast), and the both-element rule predicts an interaction exactly for such pairs. For BEND disease variants, the rule is the distance to the nearest GENCODE v44 [87] exon–intron junction. The consequence rule, the no-model baseline for this task, scores each annotated consequence class by its pathogenic

fraction on the other half of the chromosomes (odd against even chromosome numbers). BEND’s chromatin, histone and CpG methylation tasks are searched with 5,000 sampled test regions each.

Metrics follow each benchmark: MCC for GUE (F1 for the SARS-CoV-2 variant task) and the Nucleotide Transformer tasks, accuracy for Genomic Benchmarks, MCC for BEND gene finding and AUROC for its variant, chromatin, histone and CpG tasks. Published scores were collected from the benchmark papers and from later model papers, with a search date of 21 October 2026, and are listed with their sources and paper versions in the supplementary data. They follow each paper’s own protocol, for example 10-fold cross-validation on the training data or several seeds, whereas our rules are evaluated on the official test split. A task counts when a no-model rule or lookup exceeds every collected score for that task with the same metric, split and data version, by more than half a unit of the score’s last reported digit. Only results with a different metric or an explicitly different split are excluded, and where a paper reports two values for one model and task, the higher one is used. A leak is demonstrated when a rule tied to a detected leak reaches at least 75% of the best published score above chance (MCC 0 or majority-class accuracy), beats the best model of the benchmark’s own paper, or exceeds every published score. A lookup or combined rule is tied to a leak when at least 10% of test sequences have an exact duplicate or a near-duplicate in training, a quarter have a homolog, or, for GUE+, a quarter of test pairs have a shifted copy. A twin rule is tied to a leak when a quarter of test sequences have a training twin. A length, N or case rule is tied to a leak when it reaches half of the best published score above chance, and a fixed-position motif rule when it reaches three quarters of it. Composition rules (GC, CpG and Markov scores) are never tied to a leak. Intervals are 95% bootstrap intervals over 1,000 resamples of test items, stratified by class for BEND disease variants. Counts use point estimates under the strict rule. For four counted tasks, yeast H3K79me3 in GUE, mouse TF 0, yeast H3K4me2 in the original Nucleotide Transformer tasks and EPI in K562, the interval of the rule reaches the published score (source data of Table 5).

Data and code availability

Task tables, scores, results, logs, figure scripts, seqleak (called posbench in the supplement) and a table tracing every number to its result file are archived at doi:10.5281/zenodo.23605239. Input data are public, in these versions: BacBench at commit 64b7d45, DGEB at commit 1b187e6 with the dataset revisions pinned in its release, the DNALongBench loader at commit 6c43266, the GUE and GUE+ archive with SHA-256 581ba5a6, the original and revised Nucleotide Transformer tasks at Hugging Face commits 96d86d56 and 851f9946, and Genomic Benchmarks and BEND with the commits and checksums listed in the archive.

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] Hwang Y, Cornman AL, Kellogg EH, Ovchinnikov S, Girguis PR. Genomic language model predicts protein co-regulation and function. *Nat Commun.* 2024;15:2880. doi:10.1038/s41467-024-46947-9.
- [2] Cornman A, West-Roberts J, Camargo AP, et al. The OMG dataset: an Open MetaGenomic corpus for mixed-modality genomic language modeling. In: *International Conference on Learning Representations*. 2025. openreview.net/forum?id=jlzNb1iWs3.
- [3] 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.

- [4] Nguyen E, Poli M, Durrant MG, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*. 2024;386:eado9336. doi:10.1126/science.ado9336.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] West-Roberts J, Kravitz J, Jha N, Cornman A, Hwang Y. Diverse Genomic Embedding Benchmark for functional evaluation across the tree of life. *bioRxiv*. 2024. doi:10.1101/2024.07.10.602933.
- [11] 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.
- [12] 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=8O9HLDrmqtq>.
- [13] 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.
- [14] Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci USA*. 2000;97:6652–6657. doi:10.1073/pnas.110147297.
- [15] Moreno-Hagelsieb G, Collado-Vides J. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*. 2002;18 Suppl 1:S329–S336. doi:10.1093/bioinformatics/18.suppl_1.S329.
- [16] 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.
- [17] Price MN, Huang KH, Alm EJ, Arkin AP. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*. 2005;33:880–892. doi:10.1093/nar/gki232.
- [18] Dehal PS, Joachimiak MP, Price MN, et al. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res*. 2010;38:D396–D400. doi:10.1093/nar/gkp919.
- [19] Su H, Zhang R, Söding J. UniOP: a universal operon prediction for high-throughput prokaryotic (meta-)genomic data using intergenic distance. *bioRxiv*. 2024. doi:10.1101/2024.11.11.623000.
- [20] Tomar TS, Dasgupta P, Kanaujia SP. Operon Finder: a deep learning-based web server for accurate prediction of prokaryotic operons. *J Mol Biol*. 2023;435:167921. doi:10.1016/j.jmb.2022.167921.
- [21] Fulco CP, Nasser J, Jones TR, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat Genet*. 2019;51:1664–1669. doi:10.1038/s41588-019-0538-0.
- [22] Gschwind AR, Mualim KS, Karbalayghareh A, et al. An encyclopedia of human enhancer–gene regulatory interactions. *Nature*. 2026;657:179–190. doi:10.1038/s41586-026-10781-4.
- [23] Avsec Ž, Latysheva N, Cheng J, et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*. 2026;649:1206–1218. doi:10.1038/s41586-025-10014-0.
- [24] 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.
- [25] Moore JE, Pratt HE, Purcaro MJ, Weng Z. A curated benchmark of enhancer–gene interactions for evaluating enhancer–target gene prediction methods. *Genome Biol*. 2020;21:17. doi:10.1186/s13059-019-1924-8.
- [26] Ahlmann-Eltze C, Huber W, Anders S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nat Methods*. 2025;22:1657–1661. doi:10.1038/s41592-025-02772-6.
- [27] Kedzierska KZ, Crawford L, Amini AP, Lu AX. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biol*. 2025;26:101. doi:10.1186/s13059-025-03574-x.
- [28] 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.
- [29] Vishniakov K, Viswanathan K, Medvedev A, et al. Tokenization to transfer: do genomic foundation models learn good representations? *bioRxiv*. 2024 (version 3, 25 March 2026; versions 1 and 2 were titled “Genomic foundationless models: pretraining does not promise performance”). doi:10.1101/2024.12.18.628606.
- [30] 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.
- [31] Tomar TS. A simple distance rule beats genomic language models at operon prediction. Preprint. Zenodo. 2026. doi:10.5281/zenodo.23484461.
- [32] Wiatrak M, et al. BacBench. Software. 2025. <https://github.com/macwiatrak/BacBench>.

- [33] SRI International. BioCyc PGDB list, version 30.0. <https://biocyc.org/biocyc-pgdb-list.shtml>. Accessed 19 October 2026.
- [34] Romero PR, Karp PD. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics*. 2004;20:709–717. doi:10.1093/bioinformatics/btg471.
- [35] Karp PD, Midford PE, Billington R, et al. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform*. 2021;22:109–126. doi:10.1093/bib/bbz104.
- [36] Rajagopala SV, Sikorski P, Kumar A, et al. The binary protein–protein interaction landscape of *Escherichia coli*. *Nat Biotechnol*. 2014;32:285–290. doi:10.1038/nbt.2831.
- [37] Marchadier E, Carballido-López R, Brinster S, et al. An expanded protein–protein interaction network in *Bacillus subtilis* reveals a group of hubs: exploration by an integrative approach. *Proteomics*. 2011;11:2981–2991. doi:10.1002/pmic.201000791.
- [38] Rain JC, Selig L, De Reuse H, et al. The protein–protein interaction map of *Helicobacter pylori*. *Nature*. 2001;409:211–215. doi:10.1038/35051615.
- [39] Häuser R, Ceol A, Rajagopala SV, et al. A second-generation protein–protein interaction network of *Helicobacter pylori*. *Mol Cell Proteomics*. 2014;13:1318–1329. doi:10.1074/mcp.O113.033571.
- [40] Butland G, Peregrín-Alvarez JM, Li J, et al. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*. 2005;433:531–537. doi:10.1038/nature03239.
- [41] Arifuzzaman M, Maeda M, Itoh A, et al. Large-scale identification of protein–protein interaction of *Escherichia coli* K-12. *Genome Res*. 2006;16:686–691. doi:10.1101/gr.4527806.
- [42] Hu P, Janga SC, Babu M, et al. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol*. 2009;7:e1000096. doi:10.1371/journal.pbio.1000096.
- [43] Babu M, Bundalovic-Torma C, Calmettes C, et al. Global landscape of cell envelope protein complexes in *Escherichia coli*. *Nat Biotechnol*. 2018;36:103–112. doi:10.1038/nbt.4024.
- [44] Kühner S, van Noort V, Betts MJ, et al. Proteome organization in a genome-reduced bacterium. *Science*. 2009;326:1235–1240. doi:10.1126/science.1176343.
- [45] O'Reilly FJ, Xue L, Graziadei A, et al. In-cell architecture of an actively transcribing-translating expressome. *Science*. 2020;369:554–557. doi:10.1126/science.abb3758.
- [46] O'Reilly FJ, Graziadei A, Forbrig C, et al. Protein complexes in cells by AI-assisted structural proteomics. *Mol Syst Biol*. 2023;19:e11544. doi:10.15252/msb.202311544.
- [47] Balu S, Hugot S, Medina Reyes JJ, et al. Complex portal 2025: predicted human complexes and enhanced visualisation tools for the comparison of orthologous and paralogous complexes. *Nucleic Acids Res*. 2025;53:D644–D650. doi:10.1093/nar/gkae1085.
- [48] del Toro N, Shrivastava A, Ragueneau E, et al. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res*. 2022;50:D648–D653. doi:10.1093/nar/gkab1006.
- [49] Humphreys IR, Zhang J, Baek M, et al. Protein interactions in human pathogens revealed through deep learning. *Nat Microbiol*. 2024;9:2642–2652. doi:10.1038/s41564-024-01791-x.
- [50] Grünberger F, Ferreira-Cerca S, Grohmann D. Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli*. *RNA*. 2022;28:400–417. doi:10.1261/rna.078937.121.
- [51] Tan L, Guo Z, Shao Y, et al. Analysis of bacterial transcriptome and epitranscriptome using nanopore direct RNA sequencing. *Nucleic Acids Res*. 2024;52:8746–8762. doi:10.1093/nar/gkae601.
- [52] Mattick JSA, Bromley RE, Watson KJ, et al. Deciphering transcript architectural complexity in bacteria and archaea. *mbio*. 2024;15:e02359-24. doi:10.1128/mbio.02359-24.
- [53] Putzeys L, Wicke L, Boon M, van Noort V, Vogel J, Lavigne R. Refining the transcriptional landscapes for distinct clades of virulent phages infecting *Pseudomonas aeruginosa*. *microLife*. 2024;5:uqae002. doi:10.1093/femsml/uqae002.
- [54] Iturbe P, Martín AS, Hamamoto H, et al. Noncontiguous operon atlas for the *Staphylococcus aureus* genome. *microLife*. 2024;5:uqae007. doi:10.1093/femsml/uqae007.
- [55] Rostøl JT, Quiles-Puchalt N, Iturbe-Sanz P, Lasa Í, Penadés JR. Bacteriophages avoid autoimmunity from cognate immune systems as an intrinsic part of their life cycles. *Nat Microbiol*. 2024;9:1312–1324. doi:10.1038/s41564-024-01661-6.
- [56] Yan B, Boitano M, Clark TA, Ettwiller L. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat Commun*. 2018;9:3676. doi:10.1038/s41467-018-05997-6.
- [57] Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. 2021;18:1196–1203. doi:10.1038/s41592-021-01252-x.
- [58] Nanda K. Long-context DNA models do not use the allele: an audit of DNALongBench’s eQTL task. Unpublished manuscript with code. GitHub. 2026. <https://github.com/karan-nanda/Extending-DNALongBench>.
- [59] 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.

- [60] 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.
- [61] 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.
- [62] 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.
- [63] Oubounyt M, Louadi Z, Tayara H, Chong KT. DeePromoter: robust promoter predictor using deep learning. *Front Genet*. 2019;10:286. doi:10.3389/fgene.2019.00286.
- [64] Grešová K. genomic-benchmarks-qc: automated quality control for genomic machine learning datasets. Software. 2026 (version 1.0.0). github.com/genomic-benchmarks/genomic-benchmarks-qc. Accessed 21 October 2026.
- [65] Albors C, Li JC, Benegas G, Ye C, Song YS. A phylogenetic approach to genomic language modeling. *arXiv*. 2025 (version 2, 20 March 2026). arXiv:2503.03773v2.
- [66] Liang Y, Zhu W, Pan X, Li H. CodonBench: separating gene context, exon position, and variant-specific component in synonymous-variant classification. *bioRxiv*. 2026 (version 2, 11 October 2026). doi:10.64898/2026.08.12.744371.
- [67] Nucleotide Transformer authors. Nucleotide Transformer downstream tasks, revised (dataset card). Hugging Face. 2025 (commit 851f9946, 30 June 2025; key points added 23 September 2024). huggingface.co/datasets/InstaDeepAI/nucleotide_transformer_downstream_tasks_revised.
- [68] Liang C, Qiao L, Ye P, et al. Toward understanding BERT-like pre-training for DNA foundation models. *arXiv*. 2023 (version 3, 8 September 2024). arXiv:2310.07644v3.
- [69] Li Z, Subasri V, Shen Y, et al. Omni-DNA: a unified genomic foundation model for cross-modal and multi-task learning. *arXiv*. 2025 (version 1, 5 February 2025). arXiv:2502.03499v1.
- [70] Yu T, Cheng L, Khalitov R, Olsson EB, Yang Z. Self-distillation improves DNA sequence inference. *arXiv*. 2024 (version 1, 14 May 2024). arXiv:2405.08538v1.
- [71] Yu Y, Wheeler NE, Barquist L. Biased sampling driven by bacterial population structure confounds machine learning prediction of antimicrobial resistance. *PLoS Biol*. 2025;23:e3003539. doi:10.1371/journal.pbio.3003539.
- [72] 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.
- [73] 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.
- [74] 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.
- [75] 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.
- [76] Bennett 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.
- [77] Malbranke C, Zalaifi GP, Bitbol AF. ProteomeLM: a proteome-scale language model enables accurate and rapid prediction of protein–protein interactions and gene essentiality across taxa. *Proc Natl Acad Sci USA*. 2026;123:e2524201123. doi:10.1073/pnas.2524201123.
- [78] Szymborski J, Emad A. A flaw in using pretrained protein language models in protein–protein interaction inference models. *Nat Mach Intell*. 2026;8:197–208. doi:10.1038/s42256-025-01176-7.
- [79] DNALongBench authors. DNALongBench Enhancer-Target Gene Prediction [dataset]. Harvard Dataverse. 2025 (version 1.0). doi:10.7910/DVN/CTEQXX.
- [80] DNALongBench authors. DNALongBench eQTL data [dataset]. Harvard Dataverse. 2025 (version 1.0). doi:10.7910/DVN/YUP2G5.
- [81] Oughtred R, Rust J, Chang C, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021;30:187–200. doi:10.1002/pro.3978.
- [82] Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–3100. doi:10.1093/bioinformatics/bty191.
- [83] Brinda K, Callendrello A, Ma KC, et al. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nat Microbiol*. 2020;5:455–464. doi:10.1038/s41564-019-0656-6.
- [84] Irber L, Pierce-Ward NT, Abuelanin M, et al. sourmash v4: a multitool to quickly search, compare, and analyze genomic and metagenomic data sets. *J Open Source Softw*. 2024;9:6830. doi:10.21105/joss.06830.
- [85] 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>.
- [86] Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. doi:10.1186/1471-2105-10-421.
- [87] Mudge JM, Carbonell-Sala S, Diekhans M, et al. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res*. 2025;53:D966–D975. doi:10.1093/nar/gkae1078.

Supplementary note: reproducibility

Table 9 compares our reproductions with the published values. Nine of the 12 essential-gene reproductions lie within 0.012 of the published median AUROC, and the other three differ by +0.157 (Evo, on BacBench’s own windows), -0.082 (Bacformer Large) and -0.017 (Evo 2).

GroupKFold in scikit-learn 1.9.1, which assigned the genome-block folds (Methods), orders groups of equal size stably. Under earlier versions (1.1.0 to 1.8.0 checked) the order differs, and in our reruns the reported out-of-fold baseline AUROCs moved by up to 0.04 and gains over position by up to 0.003 (supplementary data). The LRB baseline on distance and genome annotation, for example, scores 0.69–0.70 under the other orders, against 0.706 with our folds.

The long-read intervals of the pretraining variants use 1,000 bootstrap replicates, where the pre-specified design called for 300. The mid-depth readout of Evo 2 40B (block $L/2$), measured on *E. coli* K-12, has a median vector norm of 7×10^{17} , and three directions carry 95% of its standardised variance.

Table 9: Reproductions of published values and checks of our pipelines.

Check	Published	This work
Long-read operons (12 models); zero-shot PPI (6); gLM probe, mean AP [31]	Tables S1, S13; 0.775	within 0.0024; match; 0.774
Evo 2 1B forward-test loss (repository reference), FP8 on (off); long-read AUROC $\times 100$, Evo 2 1B; Evo DGEb leaderboard AP, 10 models $\times$ 3 tasks	0.5020; 71.63; 74.77	0.5027 (1.364); 71.63; 74.77
BacBench essential genes, median AUROC of 12 models; Evo 2 at BacBench’s, final and mid-depth layer	leaderboard	$\max \Delta 3.6 \times 10^{-5}$
Antibiotic resistance AUROC: ESM-2; ESM-C; Bacformer; Bacformer Large	Table S14; 0.5605; $-$; $-$	9 within 0.012; 0.5437; 0.6195; 0.8189
Phenotypic traits AUROC: Bacformer; ESM-2	0.8505; 0.8568; 0.8761;	0.8497; 0.8698; 0.8777;
KEGG orthology, zero-shot macro F1: ESM-2; ESM-C	0.9215	0.9259
DNALongBench enhancer–target: ABC AUROC; AUPRC; fine-tuned HyenaDNA AUROC	0.7135; 0.6868	0.7125; 0.6876
BEND expression variants, zero-shot AUROC: DNABERT; NT-H; NT-MS	0.8785; 0.8758	0.8832; 0.8888
Bacformer masked-gene heads; seqleak on companion Tables 2 and 5	0.926; 0.407; 0.828	0.926; 0.407; 0.893 (best validation epoch)
Second transcription of Bacformer Tables S12, S14, S15	0.60; 0.55; 0.54	0.599; 0.546; 0.547
	released; [31]	load exactly; reproduced
	172 cells	172 agree