

A simple distance rule beats genomic language models at operon prediction

Tejasvi Singh Tomar

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

28 October 2026

Abstract

Language models trained on bacterial genomes are evaluated on operon identification and on protein–protein interaction (PPI) prediction, and good scores are read as evidence that they learn genome organisation. Yet genes in one operon lie on the same strand with short gaps between them, and interacting proteins are often encoded by neighbouring genes. We re-ran the 12 protein and genomic language models of the BacBench benchmark and asked how much of their performance gene position (strand, intergenic distance and adjacency) explains. An untuned rule that uses only strand and intergenic distance exceeds every model’s BacBench score on long-read operon maps (0.826 against at most 0.776 AUROC), and it outperforms the operon predictor of the genomic language model gLM on gLM’s own evaluation set. At fixed position, the models’ embedding similarity carries little or no operon signal, so scores on these operon benchmarks leave open whether the models have learned operon structure. For PPI, the labels from STRING are themselves built partly from genome organisation, and position alone outperforms most models zero-shot, although five of the six models we re-ran for this task keep significant AUROC signal beyond it. In supervised probes, the advantage of the genomic language model Bacformer Large over the protein language model it builds on disappears once position features are added. Benchmarks of genomic language models should report position baselines and what the models add to position.

Keywords: genomic language models; operon prediction; protein–protein interactions; gene position; benchmarking; bacterial genomics

1 Introduction

Language models trained on genomes are proposed as general tools for biology. For bacteria, gLM represents metagenomic contigs as sequences of protein embeddings [1], gLM2 mixes protein and intergenic DNA tokens [2], Bacformer models whole genomes as ordered sets of proteins, each given by its embedding from a protein language model (ESM-2 for Bacformer and ESM-C for Bacformer Large) [3], and Evo and Evo 2 model DNA directly [4, 5]. Operon structure and protein–protein interactions are among the main lines of evidence that these models capture how genes are organised. gLM reports an operon predictor built on its attention weights, with a mean average precision of 0.775 on *Escherichia coli* K-12 [1]. Wiatrak et al. [3], who introduced Bacformer, score operons on long-read RNA-sequencing maps of five strains and interactions on STRING-derived labels, and compare Bacformer with 11 other protein and genomic language models in their BacBench benchmark.

Gene position has been used to predict operons for more than 25 years. Adjacent genes on the same strand with short intergenic distances tend to be co-transcribed [6, 7], and conserved gene order across genomes adds further evidence [8]. STRING, too, scores protein pairs partly by the conserved proximity of their genes [9]. Yet neither gLM’s evaluation nor BacBench reports a baseline built on gene position alone. On the long-read operon maps, BacBench scores a gene pair by embedding similarity times a same-strand indicator, and Bacformer’s own operon formula multiplies the similarity by that indicator and a distance term [3], so part of what each score credits to the model belongs to the strand indicator or the distance term. Wiatrak et al. do point to a proximity bias in interaction data: they fine-tune Bacformer Large on training pairs balanced across bins of gene distance and report performance within those bins [3], but they compare the models with no baseline that uses position alone.

Elsewhere in genomics, simple baselines rival deep models and foundation models in predicting gene perturbation effects [10] and in zero-shot use of single-cell models [11], probes on pretrained DNA language models gain little over one-hot sequence [12], and deep protein–protein interaction predictors learn shortcuts such as sequence similarity and node degree [13]. In enhancer–gene prediction, a curated benchmark found the best supervised method only modestly better than a distance baseline, and often no better across cell types [14].

Here we re-ran gLM’s and BacBench’s evaluations and asked whether the models’ predictions go beyond gene position. By gene position we mean strand, the intergenic gap between genes and adjacency, together with quantities derived from these for a pair of genes: the number of genes between the two, the largest intervening gap and the number of strand switches. We re-embedded the benchmark genomes with the public checkpoints of the BacBench models, all 12 for operons and six for interactions (ESM-2, ESM-C, ProtBERT, gLM2, Bacformer and Bacformer Large). We then compared the models with an untuned rule that uses only strand and gap: Bacformer’s operon formula with its similarity factor removed; for interactions, the rule’s ties are broken by the number of genes between the two proteins. We made this comparison on the benchmark data and on curated operon sets, measured the models’ signal with position held fixed, rebuilt gLM’s evaluation from its released code and checkpoint, broke the PPI labels down by STRING evidence channel, and trained supervised probes with and without position features.

2 Results

We embedded the genomes of each benchmark as BacBench does (Methods), scored them with BacBench’s protocols and compared the results with the values published in Tables S1 and S13 of Wiatrak et al. [3] (Appendix, Table A1). All 12 models reproduce the published long-read operon AUROC to within 0.0024, and the six models we re-ran on the PPI test split reproduce the published zero-shot AUROC and AUPRC and their spread over genomes. With BacBench’s probe head, frozen embeddings score 0.004 to 0.007 AUROC above the published fine-tuned values, so the probes we use in Section 2.5 are at least as strong as the published ones.

2.1 Operon scores against a distance rule

Bacformer’s long-read set has 21,639 adjacent gene pairs in five strains (*Escherichia coli*, *Mycobacterium tuberculosis*, *Mycobacteroides abscessus*, *Pseudomonas aeruginosa* and *Staphylococcus aureus*). The rule $1[\text{same strand}] \cdot \exp(-\text{gap}/500)$, set to zero for gaps above 500 bp, takes its constants from Bacformer’s operon formula, and we did not tune it. Averaged over strains, it reaches 0.826 AUROC (Table 1, Figure 1a). Scored as in the published benchmark, where the embedding cosine is multiplied by a strand indicator, no model reaches the rule: the best is Bacformer at 0.776, and

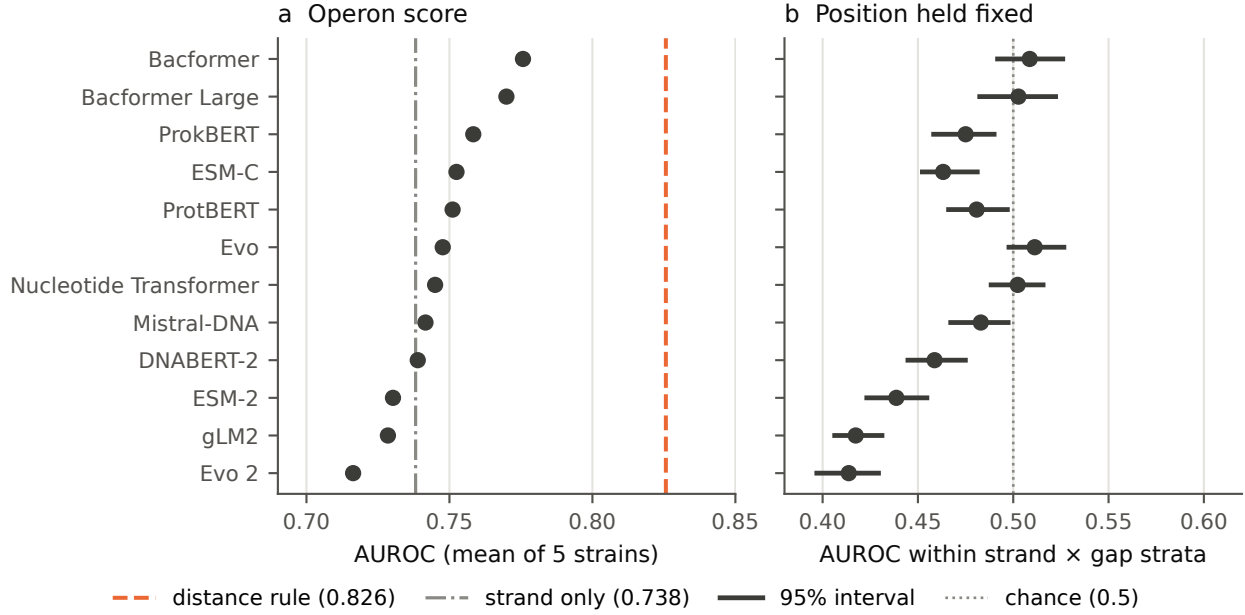


Figure 1: Long-read operons. (a) BacBench operon score of the 12 models against the distance rule and strand alone. (b) AUROC of their embedding similarity within strand $\times$ gap strata, with 95% block-bootstrap intervals.

ESM-2, gLM2 and Evo 2 fall below strand alone. When the cosine is instead multiplied into the rule, as in Bacformer’s operon formula, Bacformer and Bacformer Large reach 0.828, Evo leaves it at 0.826, and the other nine models lower it. A logistic model on a spline of gap by strand, trained on four strains and tested on the fifth, does better than every product of rule and cosine (position model, Table 1).

The rule also scores above every model on three curated operon sets (Table 2). It reaches 0.954 on RegulonDB operons of *E. coli* K-12 [15] and 0.968 on SubtiWiki operons of *Bacillus subtilis* 168 [16], about 0.1 above the best model. On the RegulonDB pairs, the comparative-genomics predictions of MicrobesOnline [8, 17] score 0.955, level with the rule. BacBench’s known-operon set, built from OperonDB [18], is scored differently: its protocol compares each operon with ten random windows of the same length, and the models see neither strand nor distance. Under that protocol the rule scores 0.755, and the 12 models score 0.500–0.648. The protocol’s negatives are also impure, since 25.9% of the negative windows lie entirely inside a known operon.

2.2 Operon signal at fixed position

To ask whether the models know anything about operons beyond position, we compared pairs only within strata that share strand state and one of 17 gap bins. On the long-read set, every model’s within-strata AUROC lies between 0.414 and 0.511, averaged over strains, and every 95% interval includes 0.5 or lies below it (Table 1, Figure 1b). No single strain exceeds 0.551. Pooled over strains, Bacformer’s AUROC lies between 0.48 and 0.52 in each of the ten same-strand gap bins up to 200 bp, which hold 12,878 of the 14,948 same-strand pairs. In those bins the mean cosines of co-transcribed and non-co-transcribed pairs differ by at most 0.004.

Nor does similarity help a fitted position model appreciably: adding a model’s cosine changes its leave-one-strain-out AUROC by less than 0.01 (Table 1). The four increases whose intervals exclude zero belong to ESM-C, DNABERT-2, ProkBERT and Evo 2, whose similarities all fall below chance

Table 1: Long-read operons (AUROC, mean over five strains). For the 12 models, “AUROC” is that of the BacBench score, $(\cos + 1)/2$ times a same-strand indicator. “Rule $\times$ cosine” multiplies the rule by $(\cos + 1)/2$, as in Bacformer’s operon formula. “Within strata” compares only pairs that share strand state and one of 17 gap bins, with 95% intervals from a block bootstrap. “Gain over position” is the change in leave-one-strain-out AUROC when cosine is added to the position model.

Scorer	AUROC	Rule $\times$ cosine	Within strata [95% CI]	Gain over position [95% CI]
strand only	0.738			
distance rule	0.826			
position model	0.833			
Bacformer	0.776	0.828	0.509 [0.491, 0.527]	+0.0002 [−0.0005, +0.0010]
Bacformer Large	0.770	0.828	0.503 [0.481, 0.523]	−0.0082 [−0.0125, −0.0032]
ProkBERT	0.758	0.824	0.475 [0.457, 0.491]	+0.0016 [+0.0001, +0.0032]
ESM-C	0.752	0.818	0.463 [0.451, 0.482]	+0.0025 [+0.0015, +0.0034]
ProtBERT	0.751	0.824	0.481 [0.465, 0.498]	−0.0061 [−0.0101, −0.0026]
Evo	0.748	0.826	0.511 [0.496, 0.528]	−0.0054 [−0.0079, −0.0031]
Nucleotide Transformer	0.745	0.823	0.502 [0.487, 0.517]	+0.0004 [−0.0011, +0.0020]
Mistral-DNA	0.742	0.825	0.483 [0.466, 0.499]	+0.0002 [−0.0005, +0.0010]
DNABERT-2	0.739	0.822	0.459 [0.443, 0.476]	+0.0013 [+0.0004, +0.0022]
ESM-2	0.730	0.818	0.439 [0.422, 0.456]	−0.0027 [−0.0040, −0.0015]
gLM2	0.728	0.810	0.417 [0.405, 0.432]	−0.0059 [−0.0077, −0.0041]
Evo 2	0.716	0.825	0.414 [0.396, 0.431]	+0.0038 [+0.0022, +0.0056]

Table 2: Curated operon sets (AUROC). Models use the BacBench score: cosine $\times$ strand for pairs, mean pairwise cosine for windows; the rule scores a window by its mean over the window’s adjacent pairs. Ranges cover all 12 models. The OperonDB window row counts operons, the other rows adjacent pairs; Methods defines the OperonDB pair rows and the 8-genome subset.

Set	Size	Rule	Position	Models	Within strata	Gain over position
OperonDB, BacBench windows	4,073	0.755		0.500–0.648		
OperonDB pairs, 11 genomes	12,423	0.793	0.792	0.664–0.716	0.505–0.588	−0.004 to +0.005
OperonDB pairs, 8 genomes	11,003	0.919	0.919	0.739–0.818	0.483–0.621	+0.001 to +0.010
RegulonDB, <i>E. coli</i>	4,114	0.954	0.957	0.778–0.857	0.503–0.635	−0.0002 to +0.005
SubtiWiki, <i>B. subtilis</i>	2,657	0.968	0.967	0.792–0.861	0.493–0.584	−0.001 to +0.002

at fixed position. A supervised probe on the full embedding pair (principal components of $e_i \odot e_j$ and $|e_i - e_j|$, trained on four strains and tested on the fifth) fares no better. It reaches 0.48–0.73 AUROC alone and changes the position model’s AUROC by −0.023 to +0.002 for all 12 models. We read each model at the layer BacBench uses, and for Evo 2 the layer matters: at its final layer, Evo 2 scores 0.767 rather than 0.716 under BacBench’s protocol, still below the rule, and 0.517 [0.502, 0.534] within strata, the only long-read interval above 0.5; even so, its cosine lowers the position model’s AUROC by 0.002.

ESM-2, gLM2 and Evo 2, which fall below strand alone, are also the furthest below chance at fixed position (0.414–0.439). The long-read maps call a pair co-transcribed only when reads span it, so operons without expression in the sequenced conditions count as negatives. Some of these negatives are among the most similar pairs. Among same-strand neighbours with gaps of −50 to 150 bp, pooled over strains, the pairs in the top decile of ESM-2 similarity that are labelled negative in *E. coli* include genes of well-characterised operons (*dppABCDF*, *nikABCDE*, *livKHMGF*), for

Table 3: gLM operon evaluation: mean average precision over the published five folds, and AUROC.

Scorer	Mean AP	AUROC
gLM attention probe (190 heads, published protocol)	0.774 ± 0.020	0.813
gLM attention probe, input with correct strand	0.786 ± 0.022	0.825
distance rule	0.857 ± 0.030	
position logistic regression (gap spline $\times$ strand)	0.905 ± 0.020	0.950
attention + position	0.918 ± 0.017	0.947
UniOP predictions [19]	0.914 ± 0.015	
MicrobesOnline predictions [8]	0.945 ± 0.018	

example *dppD/dppF*, *nikE/nikD*, *nikC/nikB* and *livG/livF*. Below-chance similarity at fixed position is therefore weak evidence on its own, which is why we also tested the curated sets.

On the curated sets the models do keep a small residual at fixed position (within-strata AUROC up to 0.64), but added to position models that score 0.79–0.97 it raises AUROC by at most 0.010 (Table 2).

2.3 gLM’s operon evaluation

We rebuilt gLM’s evaluation from its released code and data. It has 2,900 adjacent gene pairs inside 30-gene windows of *E. coli* K-12 and 1,147 positives, identical to the source data of Figure 4c of the gLM paper [1]. Running the released checkpoint on the released input reproduces the attention probe (Table 3).

Position alone outperforms the attention probe (Table 3). Because the published protocol splits folds over pairs, pairs from one window can fall in both training and test folds. We therefore also grouped folds by window; attention, position and attention plus position then scored 0.770, 0.904 and 0.917, respectively. A bootstrap over windows puts attention’s gain over position at +0.013 AP (95% CI -0.005 to $+0.032$), and adding attention lowers AUROC from 0.950 to 0.946 with window-grouped folds (0.947 with the published folds, Table 3).

gLM’s released input carries almost no strand information. Its orientation feature is +0.5 for 4,281 genes and -0.5 for only 34 (*araB*, *lacZ*, *galK* and others), whereas the NCBI annotation places 2,214 of these genes on the minus strand. Rewriting the feature from the annotation raises the attention probe from 0.774 to 0.786, still below the rule’s 0.857 (Table 3).

The published ablation baseline, a probe on a one-hot version of gLM, scores 0.426 [1], barely above the 0.396 expected from a random ranking (the positive rate). The evaluation’s labels also vary in support: of the 1,147 positives, 559 lie in operons with experimental evidence in RegulonDB v14.5, 313 in operons with computational evidence only, and 275 in operons that v14.5 lists without evidence or not at all.

2.4 Zero-shot interactions

BacBench’s PPI test split has 53 genomes and 16,410,869 pairs listed in STRING v12. A pair is positive when its STRING combined score is at least 600 (9.6% of pairs); the negatives are STRING-listed pairs that score 150–599, not random pairs. Zero-shot, each model scores a pair by the cosine of its two protein embeddings, and metrics are averaged over genomes.

The rule, with ties broken by the number of genes between the two proteins, reaches 0.669 AUROC and 0.402 AUPRC (Table 4). That puts it above 11 of the 12 models on AUROC and above all 12 on AUPRC, counting the six models we did not re-run at their published values

Table 4: Zero-shot PPI (mean over 53 genomes). Position strata combine 11 genes-apart bins, strand switches (0, 1, 2, 3 or more), 9 bins of the largest intervening gap and strand state. “Gain over position” is the change in AUROC when the cosine is added to the fitted position model, with 95% intervals over genomes, and “genomes up” counts the genomes where it is positive.

Scorer	AUROC	AUPRC	Within strata [95% CI]	Gain over position [95% CI]	Genomes up
genes apart	0.666	0.342			
distance rule	0.669	0.402			
position model (fitted)	0.705	0.508			
Bacformer Large	0.721	0.375	0.628 [0.614, 0.641]	+0.053 [+0.045, +0.062]	51/53
Bacformer	0.649	0.314	0.595 [0.583, 0.607]	+0.030 [+0.025, +0.036]	52/53
ESM-C	0.574	0.208	0.562 [0.551, 0.573]	+0.020 [+0.016, +0.024]	51/53
ESM-2	0.569	0.220	0.568 [0.561, 0.575]	+0.024 [+0.019, +0.029]	51/53
gLM2	0.540	0.197	0.510 [0.497, 0.522]	−0.0002 [−0.0015, +0.0010]	33/53
ProtBERT	0.507	0.183	0.466 [0.455, 0.479]	+0.003 [+0.001, +0.005]	39/53

(0.508–0.562 AUROC, 0.168–0.211 AUPRC). A logistic regression on genes apart, strand, gap, the largest intervening gap and the number of strand switches, fitted on other test genomes, does better still (Table 4).

Holding position fixed separates the models (Table 4, Figure 2a). Position explains gLM2’s AUROC signal: within strata gLM2 is at chance, and its cosine adds nothing to the position model’s AUROC. The other five models, by contrast, keep significant signal beyond position (Wilcoxon over genomes, $p < 10^{-3}$), although ProtBERT’s comes from similarity that is below chance at fixed position (0.466). ESM-2 keeps almost all of its raw AUROC within strata, and Bacformer and Bacformer Large add the most to the position model.

The labels themselves are built partly from genome organisation. Removing STRING’s neighbourhood and fusion channels from the combined score drops 19.1% of positives below 600, removing co-occurrence drops 18.2%, and removing all three drops 38.6% (606,049 positives). On these positives the position model outscores Bacformer Large, the best of the five models scored on them (gLM2 was not): 0.804 vs 0.768 AUROC and 0.576 vs 0.257 AUPRC. On the remaining positives, Bacformer Large is ahead on AUROC (0.688 vs 0.648), whereas the position model is ahead on AUPRC (0.303 vs 0.256). Gene distance alone, with the channel binarised at 600, scores 0.956 AUROC against STRING’s neighbourhood channel (median over genomes), so agreement with that channel is largely attainable from position alone; Wiatrak et al. report a median of 0.99 ± 0.03 for attention weights against that channel on other genomes [3].

For pairs more than 20 genes apart we split the positives by the STRING evidence behind them, keeping the same negatives (Table 5, Figure 2b); genome-organisation evidence means a non-zero STRING neighbourhood (direct or transferred), fusion or co-occurrence score. On these groups the fitted position model scores only 0.52–0.56. Bacformer’s gain over the protein model it builds on, ESM-2, is +0.057 on pairs that are neighbours in other genomes, but −0.067 on pairs supported only by co-occurrence and −0.009 on pairs without genome-organisation evidence. Bacformer Large, by contrast, gains over ESM-C on every row of Table 5 (+0.042 to +0.099), though not on far positives with transferred database evidence but no genome-organisation evidence, a subgroup not shown in Table 5 (0.476 for Bacformer Large vs 0.512 for ESM-C). For positives without neighbourhood, fusion or co-occurrence evidence, no model exceeds 0.52 AUROC, and gLM2 and ProtBERT fall well below chance (0.364 and 0.331; ProtBERT is not shown in Table 5). Across all 1,570,225 positives in these genomes, none has direct co-expression evidence and 12 have direct experimental evidence,

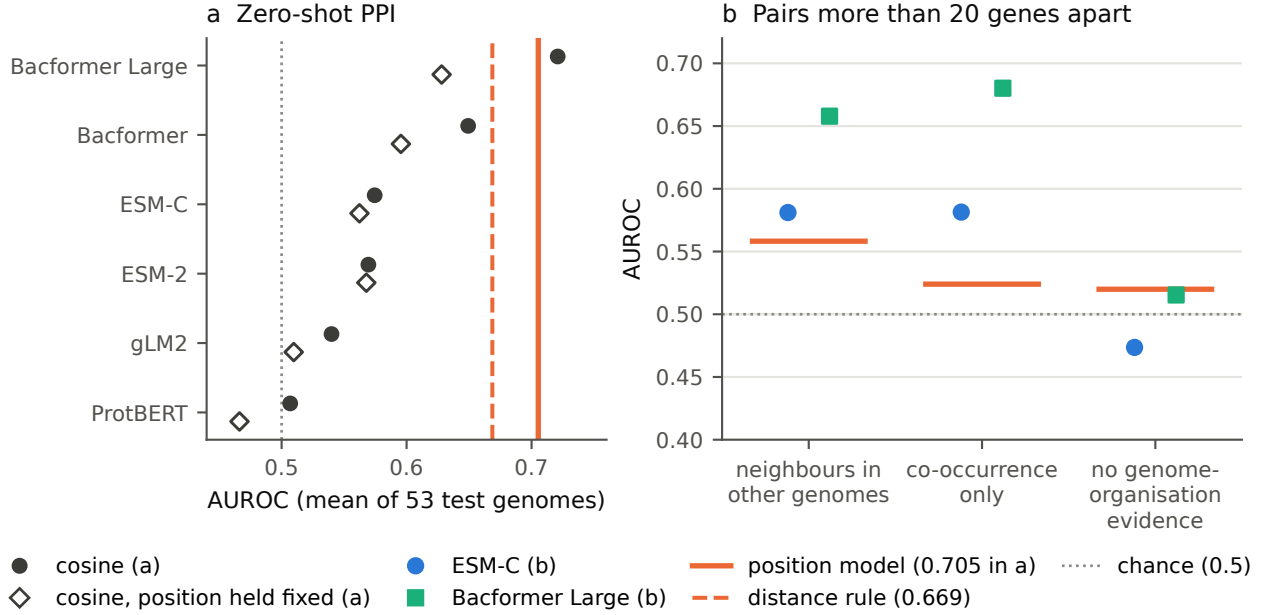


Figure 2: Zero-shot PPI. (a) AUROC of the cosine of six models, raw and with position held fixed (within position strata), with the fitted position model and the distance rule for reference. (b) For pairs more than 20 genes apart, AUROC of ESM-C and Bacformer Large by the STRING evidence behind each positive, with the fitted position model as a reference.

Table 5: Zero-shot AUROC for pairs more than 20 genes apart, by the STRING evidence behind each positive (same negatives in every row). “Position” is the fitted position model.

Positives	<i>n</i>	Genomes	Position	ESM-2	Bacformer	ESM-C	Bacformer Large	gLM2
neighbours in other genomes	626,351	53	0.558	0.574	0.631	0.581	0.658	0.520
co-occurrence only	161,476	50	0.524	0.658	0.590	0.581	0.680	0.600
no genome-organisation evidence	252,579	52	0.520	0.512	0.503	0.474	0.515	0.364
with transferred experiments	168,362	51	0.527	0.488	0.477	0.435	0.506	0.315

while direct database entries cover 4.7% and direct text mining 0.3%, so at least 95% of positives carry no direct evidence from their own organism outside the genome-context channels.

2.5 Supervised interactions

For the supervised probes we trained BacBench’s head on frozen embeddings of the 159 training genomes and tested it on the 53 test genomes (Methods). The position-only probe, a small network on the position features, scores 0.696 AUROC and 0.496 AUPRC (Table 6). On AUPRC it exceeds 11 of the 12 published fine-tuned models; only fine-tuned Bacformer Large is higher, at 0.508. On AUROC it exceeds 5 of the 11 non-degenerate ones; the twelfth, Evo 2, scores a constant 0.500 there.

Genomic pretraining raises the probe by +0.006 AUROC over ESM-2 (Bacformer) and by +0.051 over ESM-C (Bacformer Large). With position features, however, the two differences are −0.002 and −0.003. The advantage of Bacformer Large sits mainly in pairs one to five genes apart, where it scores 0.760 compared with 0.668 for ESM-C, while the position probe scores 0.880 there (Figure 3).

Table 6: Supervised PPI with BacBench’s probe head on frozen embeddings (test, mean $\pm$ SD over 53 genomes). “With position” adds a small network on the position features, trained jointly with the head; the position-only row is that network alone.

Probe	AUROC	AUPRC	AUROC with position	AUPRC with position
position only (no sequence)	0.696 ± 0.104	0.496 ± 0.204		
ESM-2 (35M)	0.735 ± 0.036	0.451 ± 0.092	0.862 ± 0.041	0.672 ± 0.149
Bacformer (on ESM-2)	0.742 ± 0.043	0.462 ± 0.099	0.860 ± 0.043	0.673 ± 0.153
ESM-C (300M)	0.738 ± 0.035	0.452 ± 0.097	0.865 ± 0.041	0.676 ± 0.155
Bacformer Large (on ESM-C)	0.789 ± 0.027	0.520 ± 0.111	0.862 ± 0.046	0.679 ± 0.153

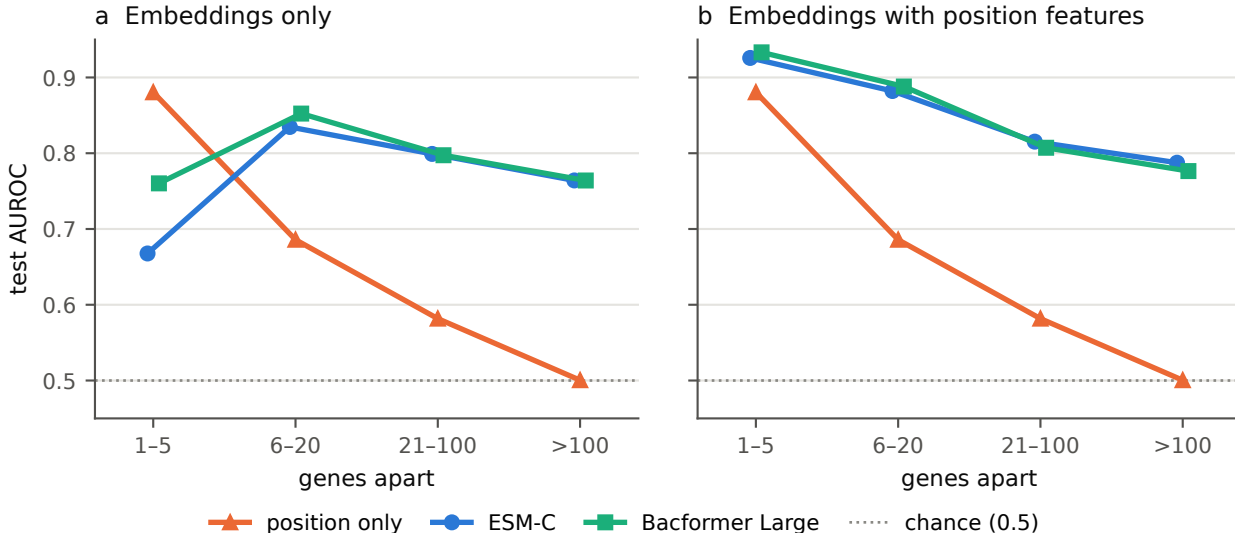


Figure 3: Supervised PPI test AUROC by gene distance for ESM-C, Bacformer Large and the position-only probe. (a) Embeddings only. (b) Embeddings with position features.

Beyond 20 genes ESM-C matches it (0.799 vs 0.797 at 21–100 genes; 0.764 for both beyond 100). With position features, the four models agree to within 0.02 AUROC in every distance bin. The probes with position features exceed the best published fine-tuned result (0.784 and 0.508), and the position features account for most of the excess.

Supervised probes also rank positives that lack genome-organisation evidence above the negatives: for pairs more than 20 genes apart, the probes on embeddings alone score 0.834 (ESM-2), 0.831 (ESM-C) and 0.808 (Bacformer Large) AUROC, whereas the position probe scores 0.483 . The head, however, adds a per-protein term to the pair’s cosine. We ran no per-protein baseline, and BacBench’s split of the genomes by genus does not keep homologous proteins out of the training genomes. This signal is therefore consistent with a per-protein interaction propensity transferred by homology.

3 Discussion

On operons, the models’ predictions add almost nothing to strand and intergenic distance, and gLM’s attention-based predictor stays below a rule built on those two quantities on its own evaluation

pairs even when given the correct strand. A benchmark on which a heuristic more than 25 years old outperforms the models does not show that they have learned operon structure.

On interactions, part of the evidence behind the STRING labels is genome organisation itself (Section 2.4), so a model that represents gene order can score well by recovering position. Zero-shot, however, every re-run model except gLM2 retains AUROC signal that position does not explain, and the two Bacformer models retain the most. For Bacformer, the gain over ESM-2 among pairs more than 20 genes apart is confined to protein pairs that are neighbours in other genomes. That is a form of conserved gene order learned across genomes, consistent with the report of Wiatrak et al. that shuffling gene order lowers AUROC by 0.13 [3]. The supervised probes on embeddings show a second source of signal, at least as strong in the protein models as in Bacformer Large; as Section 2.5 notes, it may be a per-protein propensity carried over between homologous proteins. Once the probes see position features, a protein language model that sees one protein at a time performs as well as the genomic model.

Our analysis has limits. The long-read labels miss operons that are not expressed in the sequenced conditions, and RegulonDB and SubtiWiki include computationally inferred operons, which favour distance-based scores. The long-read intervals come from resampling pairs within five fixed strains, so they ignore strain-to-strain variation, and we applied no correction for multiple comparisons. Each supervised probe is a single run, so differences of a few thousandths, including Bacformer’s +0.006, may be noise. All models are read at BacBench’s layers, a choice that matters for Evo 2 (Section 2.2). Because the PPI negatives are STRING-listed pairs, these AUROCs are not comparable with benchmarks that use random negatives. Finally, we did not check whether the test genomes and long-read strains occur in the models’ pretraining corpora; such overlap could favour the models.

To be informative, operon and interaction benchmarks of genomic language models should report a position baseline next to every model score, together with position-conditional metrics such as AUROC within strata of strand and distance or the gain over a position model. They should also evaluate interactions on labels that do not derive from genome context, and report separately the pairs that STRING supports through neighbourhood, fusion or co-occurrence.

4 Methods

Data. We used three of BacBench’s datasets on Hugging Face, each also in a DNA version:

- long-read operons:
macwiatrak/operon-identification-long-read-rna-sequencing-protein-sequences;
- known operons:
macwiatrak/bacbench-operon-identification-protein-sequences;
- PPI, with 159 training, 49 validation and 53 test genomes:
macwiatrak/bacbench-ppi-stringdb-protein-sequences-small.

We joined each PPI test pair to its record in STRING v12 (`protein.links.full`) by locus tag [9]. RegulonDB v14.5 operons and gene products [15], SubtiWiki operons [16], MicrobesOnline operon predictions [17] and UniOP predictions [19] are taken from their public downloads. gLM’s code, data and checkpoint are taken from its repository and Zenodo record [1, 20].

Embeddings. Protein models and Bacformer follow BacBench’s `embed_prot_seqs.py` [21]: the mean of last-layer states over each protein’s tokens including special tokens, truncation at 1,024

tokens including special tokens, and the `masked-complete-genomes` checkpoints. Bacformer takes ESM-2 t12 35M embeddings [22] and Bacformer Large takes ESM-C 300M embeddings [23], through the ESM++ implementation pinned by BacBench. As in BacBench’s script, genomes are capped at 9,000 proteins for Bacformer. ProtBERT is Rostlab/prot_bert [24]. gLM2 (650M) [2] reads each contig as protein and intergenic DNA tokens with strand tokens. For operons it uses 4,096-token windows centred on each gene, and for PPI windows tiled at a stride of 2,048 tokens. The gene embedding is the mean over the gene’s amino-acid tokens. DNA models (Nucleotide Transformer v2 250M [25], DNABERT-2 [26], Mistral-DNA [27], ProkBERT-mini-long [28]) embed each gene with 128 bp of promoter in overlapping chunks of 2,048 bp (Nucleotide Transformer, ProkBERT) or 512 bp (DNABERT-2, Mistral-DNA) and average the chunks. ProkBERT’s 2,048-bp chunk reproduces the published long-read operon AUROC; with 512 bp it scores 0.752. Evo (`evo-1-8k-base`, final-layer output) [4] and Evo 2 (`evo2_1b_base`, layer `blocks.24.mlp.13`) [5] follow BacBench’s protocol: an 8,192-bp window centred on each gene, averaged over the gene’s tokens. For Evo 2 we cut each window at the gene’s last base; because the model is causal, this leaves the gene’s states unchanged. Evo reads the full window, as in BacBench. Evo 2 runs with its FP8 input projections (Transformer Engine) switched on, which its authors require for the numerical accuracy of the 1B model. Its embeddings are stored in float32.

Operon scores and strata. The rule is $\mathbf{1}[\text{same strand}] \cdot \exp(-\text{gap}/500)$ for gaps up to 500 bp and 0 otherwise. The BacBench score is $(\cos + 1)/2$ times a same-strand indicator, with AUROC computed per strain and averaged. Within-strata AUROC compares positive and negative pairs only inside the same stratum (strand state $\times$ 17 gap bins) and weights each stratum by its number of positive-negative pairs. Its 95% intervals come from 300 bootstrap resamples of blocks of 50 consecutive pairs within each strain. The position model is a logistic regression on a cubic B-spline of gap, separately for same-strand and opposite-strand pairs, fitted on four strains and tested on the fifth. The gain of a model is the change in this AUROC when its cosine and cosine $\times$ same strand are added as features. The supervised operon probe uses the leading 32 or 128 principal components of $e_i \odot e_j$ and $|e_i - e_j|$.

Curated operon sets. For RegulonDB and SubtiWiki an adjacent pair is positive when its genes share any operon and negative when both genes are assigned to operons but share none; pairs with an unassigned gene are excluded. 654 SubtiWiki genes belong to more than one operon. For OperonDB, BacBench’s protocol compares each operon’s mean pairwise cosine with ten random contiguous windows of the same size for five seeds, and the pair protocol compares pairs inside a known operon with the pairs at its two ends. In three OperonDB genomes 24–67% of operons span both strands, against 0.5% in *E. coli* and none in the other genomes, so the 8-genome rows exclude them. Position models are fitted leave-one-genome-out on the OperonDB pairs and with five-fold cross-validation over pairs on RegulonDB and SubtiWiki.

gLM. We rebuilt the evaluation from `operon_logreg.py` and `operon.annot`: genes in file order, 30-gene windows, the first 100 windows, adjacent pairs inside windows, positives sharing an operon name, and five folds from `KFold(shuffle=True, random_state=16)`. Attention features are the 190 symmetrised heads of the released checkpoint run on the released input. Gene coordinates come from the NCBI annotation of *E. coli* K-12 MG1655, matched by name and synonym. Mean AP is averaged over folds.

PPI. A pair is positive at a STRING combined score of at least 600. BacBench’s caps apply: the first 2,000,000 pairs and the first 6,000 proteins of each contig. Metrics are computed per genome and averaged. For a pair, gap is the distance in base pairs between the two genes, clipped at zero, and strand state is whether the two genes lie on the same strand. The fitted position model is a logistic regression on splines of genes apart, gap and the largest intervening gap, each separately by strand state, with indicators for zero to four or more strand switches, fitted to convergence by Newton’s method with genome-grouped five-fold cross-validation. Gains over it are tested with Wilcoxon signed-rank tests over genomes, with 95% intervals from bootstrapping genomes. For the channel ablations we rebuilt the combined score from STRING’s channels with the prior 0.041 and without STRING’s homology correction. We assigned far pairs to evidence groups by their direct or transferred neighbourhood, fusion and co-occurrence scores, and for each group the mean AUROC was taken over genomes with at least 20 positives in that group.

Supervised probes. The head is BacBench’s `run_train_mlp.py` head: a shared dense layer with GELU, layer normalisation and dropout applied to each protein, the mean over the pair, the pair’s cosine, and a linear output. Position features (genes apart, strand, gap and largest intervening gap, the three distances log-transformed, and strand switches as a scaled count and a zero-switch indicator) feed a network with two hidden layers of 64 units, whose output is added to the head’s. Embedding probes use AdamW with learning rate 10^{-3} , batch size 1,024 and at most ten epochs with patience three on validation AUROC; four of them choose between learning rates 10^{-3} and 3×10^{-4} on validation AUROC. The position-only probe is the position network alone, trained with Adam, batch size 8,192 and at most 20 epochs with patience five, choosing among learning rates 10^{-3} , 3×10^{-4} and 10^{-4} . Both optimisers use weight decay 0.01. All probes use at most 200,000 training and 100,000 validation pairs per genome, and each probe is one run.

Software. Analyses use PyTorch, Hugging Face Transformers and scikit-learn [29]. The README of the archived code (Data and code availability) gives the commands and the run settings of the embeddings.

A Published values and this re-run

Table A1 compares the published values with our re-run. The join of BacBench’s PPI test pairs to STRING v12 covers all pairs, and STRING’s combined score equals the benchmark’s label score for every pair. Our rebuild of the combined score from STRING’s channels matches it to within five points for every pair.

Table A1: Published values and this re-run. Operon AUROC $\times 100$ on the long-read set; PPI AUROC / AUPRC $\times 100$, mean $\pm$ SD over test genomes (53 in the released split; Wiatrak et al. state 54); gLM mean average precision over its five folds, as published by Hwang et al. [1]. The fine-tuned PPI re-run trains BacBench’s probe head on frozen embeddings.

Benchmark	Model	Published	This re-run
Long-read operons	Bacformer	77.59	77.57
	Bacformer Large	77.00	76.99
	ESM-2 (35M)	73.02	73.03
	ESM-C (300M)	75.25	75.25
	ProtBERT	75.11	75.11
	gLM2	72.85	72.85
	Nucleotide Transformer	74.50	74.50
	Mistral-DNA	74.16	74.16
	DNABERT-2	74.13	73.89
	ProkBERT	75.77	75.83
	Evo	74.77	74.77
	Evo 2	71.63	71.63
Zero-shot PPI	Bacformer Large	$72.09 \pm 6.44 / 37.51 \pm 15.20$	$72.09 \pm 6.45 / 37.50 \pm 15.20$
	Bacformer	$64.93 \pm 4.96 / 31.43 \pm 12.61$	$64.93 \pm 4.96 / 31.37 \pm 12.61$
	ESM-C	$57.44 \pm 3.13 / 20.81 \pm 9.24$	$57.45 \pm 3.13 / 20.80 \pm 9.24$
	ESM-2	$56.94 \pm 2.44 / 22.08 \pm 9.15$	$56.94 \pm 2.44 / 22.05 \pm 9.14$
	gLM2	$54.02 \pm 3.72 / 19.76 \pm 9.34$	$53.99 \pm 3.71 / 19.71 \pm 9.27$
	ProtBERT	$50.69 \pm 4.11 / 18.38 \pm 8.86$	$50.69 \pm 4.11 / 18.34 \pm 8.85$
Fine-tuned PPI	Bacformer Large	$78.35 \pm 2.40 / 50.84 \pm 10.34$	$78.88 \pm 2.66 / 51.97 \pm 11.06$
	Bacformer	$73.79 \pm 4.69 / 45.73 \pm 9.29$	$74.16 \pm 4.31 / 46.21 \pm 9.87$
	ESM-C	$73.20 \pm 4.29 / 44.75 \pm 9.08$	$73.77 \pm 3.46 / 45.21 \pm 9.75$
	ESM-2	$72.83 \pm 4.53 / 44.24 \pm 8.64$	$73.55 \pm 3.56 / 45.11 \pm 9.19$
gLM operons	gLM attention probe	0.775 ± 0.028	0.774 ± 0.020

Data and code availability

Code, result tables, logs and figures are archived at <https://doi.org/10.5281/zenodo.23484463>. The input data are public and are listed in Methods.

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, Roux S, Beracochea M, Mirdita M, Ovchinnikov S, Hwang Y. The OMG dataset: an Open MetaGenomic corpus for mixed-modality genomic language modeling. *bioRxiv.* 2024. doi:10.1101/2024.08.14.607850.
- [3] Wiatrak M, Abelson DC, Mikhaylenko R, Viñas Torné R, Ntemourtsidou M, Dinan A, Arora D, Horsfield ST, Lees J, Brbić M, Weimann A, Floto RA. 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 modeling and design across all domains of life with Evo 2. *bioRxiv.* 2025. doi:10.1101/2025.02.18.638918.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] Salgado H, Gama-Castro S, Lara P, et al. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in *E. coli* K-12. *Nucleic Acids Res.* 2024;52:D255–D264. doi:10.1093/nar/gkad1072.
- [16] Pedreira T, Elfmann C, Stülke J. The current state of SubtiWiki, the database for the model organism *Bacillus subtilis*. *Nucleic Acids Res.* 2022;50:D875–D882. doi:10.1093/nar/gkab943.
- [17] 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.
- [18] Okuda S, Yoshizawa AC. ODB: a database for operon organizations, 2011 update. *Nucleic Acids Res.* 2011;39:D552–D555. doi:10.1093/nar/gkq1090.
- [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] Hwang Y, Cornman A, Kellogg E, Ovchinnikov S, Girguis P. Genomic language model predicts protein co-regulation and function [dataset]. Zenodo. 2023. doi:10.5281/zenodo.7855545.
- [21] Wiatrak M, et al. BacBench. Software. 2025. <https://github.com/macwiatrak/BacBench>.
- [22] 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.

- [23] EvolutionaryScale. ESM Cambrian: revealing the mysteries of proteins with unsupervised learning. 2024. <https://www.evolutionaryscale.ai/blog/esm-cambrian>.
- [24] Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell.* 2022;44:7112–7127. doi:10.1109/TPAMI.2021.3095381.
- [25] 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.
- [26] 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.
- [27] Mourad R. Mistral-DNA: Mistral large language model for DNA. Software. 2024. <https://github.com/raphaelmourad/Mistral-DNA>.
- [28] Ligeti B, Szepesi-Nagy I, Bodnár B, Ligeti-Nagy N, Juhász J. ProkBERT family: genomic language models for microbiome applications. *Front Microbiol.* 2024;14:1331233. doi:10.3389/fmicb.2023.1331233.
- [29] 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>.