PREDICTION TARGETS AND VOCABULARY COVERAGE IN BACTERIAL PANGENOME MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Genome language models over bacterial pangenomes predict each gene as one of a fixed set of protein families; we compare nine prediction targets for such models under a matched corpus, backbone, masking scheme, token budget and evaluation. Contrary to our pre-registered primary hypothesis, a head tied to a frozen protein encoder's centroids has higher held-out cross-entropy than a free classifier, by 0.225 nats. The gap depends on distribution shift. It is 0.435 nats for unseen species of a seen genus and indistinguishable from zero at unseen phylum, where resampling phyla rather than species widens the bound to 0.051 nats and the species-level reversal seen with a 302M-parameter backbone does not survive. Part of the gap reflects head size: at matched trainable parameters, the tied head is ahead. The target further sets what a model can represent. Out-of-vocabulary genes are 21.25% of held-out gene mass, and a closed-vocabulary head cannot emit their labels at all; a tied head can, but its exact-family accuracy on them is near zero and its net advantage depends on how finely the vocabulary is extended. An auxiliary target predicting cross-taxon gene persistence works on either head at small cost. Finally, frozen-probe conclusions under the widely used protocol depend on the layer probed: at the final layer no target we tested beats a protein-only baseline, while at the first block every arm's mean does on essential genes.

1 INTRODUCTION

Bacterial pangenome models predict a gene as one of a fixed set of protein families. This choice is rarely examined, although it limits predictions for genes absent from reference genomes.

A large and growing share of bacterial genomics studies metagenome-assembled genomes rather than cultured isolates, and these genomes routinely contain genes absent from reference databases. A fixed 50,000-way softmax cannot emit labels for such genes because its output matrix has no corresponding rows. Across 1,088 held-out species, these genes account for 21.25% of held-out gene mass, rising monotonically to 29.56% at unseen phylum. The model thus has no label of its own for a fifth of a novel genome's genes; §5 measures whether an existing label is an adequate substitute. These genes are rare within their own species, falling below the corpus's frequency filter; this does not imply biological novelty.

We make the following contributions.

1. A measurement of what a closed protein-family vocabulary excludes (§3): coverage and discriminability decouple under taxonomic shift, and a fifth of held-out gene mass falls outside the vocabulary entirely.
2. A comparison of two capabilities unavailable from a closed vocabulary, under matched corpus, backbone, masking, token budget and evaluation mask, with their respective baselines stated: tying the head to encoder centroids increases cross-entropy by 0.2249 nats relative to the free classifier, a gap that shrinks with taxonomic distance until it is indistinguishable from zero at unseen phylum (§4); a population-statistical head adds 0.0125 nats to the tied head, or +0.0176 to the free classifier while retaining nearly full predictive strength (§7). Neither capability is expressible by the 50,000-way output alone.
3. An evaluation of open-vocabulary coverage (§5), on two million real out-of-vocabulary families: the architecture supports vocabulary extension, exact-family accuracy is near

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

zero, and the metric head's advantage reverses between two pre-registered extension rules while a continuous target's does not.

4. A geometry-based soft target and its null result (§6): a soft target over prototypes shaped by the centroid geometry, at a temperature derived from §3's near-tie rate rather than tuned, does not resolve the cross-entropy gap.
5. A trainable-parameter-matched control and a permutation control (§4.1): at the metric head's own trainable-parameter count a free classifier is 1.22 nats worse than the metric head and a permuted geometry worse again by 1.33, so the true geometry improves performance and the permuted geometry degrades it at matched trainable size. The metric head's frozen centroid matrix is supplied rather than learned. Permuting increases cross-entropy by 2.5517 nats relative to the metric head; this deficit grows with distance while the true geometry's deficit shrinks.
6. A depth-dependent frozen-probe result: standalone final-layer frozen probes show no transfer advantage for any target we tested, but at block 0 every arm's mean exceeds the protein-encoder baseline on essential genes (19 of 20 checkpoints); on operon the result is mixed and within seed noise. The widely used protocol evaluates genome context at a layer whose performance is close to the lowest observed (§8).

2 Related work

The architecture we study — a transformer over a genome's proteins, each supplied as a protein-language-model embedding — is standard: gLM (Hwang et al., 2024) and Bacformer (Wiatrak et al., 2025) are the closest comparators; our backbone follows this architectural family.

This paper isolates the prediction target, which differs across these models. gLM's masked objective is continuous, regressing a whitened protein embedding; Bacformer's causal variant takes and predicts a protein-family token, a closed vocabulary in the sense used here. Existing systems vary this choice alongside their corpora, encoders, parameter counts and evaluations, so their reported results do not isolate its performance effect.

The metric head uses an established metric classifier over frozen centroids. Label-embedding classifiers, prototypical networks (Frome et al., 2013; Snell et al., 2017), tied input–output embeddings (Press & Wolf, 2017) and the image–text contrastive family (Radford et al., 2021) replace a free output matrix with similarities to fixed vectors. Our contribution is the measurement, not the head. Appendix G provides further discussion, including the masked-image analogue (Bao et al., 2022; He et al., 2022; Baevski et al., 2022).

3 Closed-vocabulary coverage and ambiguity

The corpus is a bacterial pangenome release over GTDB r214 species (Parks et al., 2022): 290,981 genomes of 27,669 species, each gene assigned to a protein cluster, and labelled core or accessory by mOTUpan (Buck et al., 2022). Every cluster has one representative protein, embedded with ESM-C (EvolutionaryScale Team, 2024), and the protein families the models predict are prototypes over these embeddings. The split is by species: 1,088 test and 338 validation species are held out and assigned to six bands by taxonomic distance to the nearest training species, from unseen species, whose genus keeps a training species, to unseen phylum, whose whole phylum is withheld; training uses 282,312 genomes of the remaining 26,243 species. Appendix A gives the full construction.

We use $K = 50{,}000$ spherical-k-means prototypes over mean-centred ESM-C embeddings, fit on 12,800,000 rows of training species, 29 of which (0.0002%) occur only in withheld in-corpus genomes (Appendix A). We characterise this label space through three measurements.

The label space is near-balanced: $\ln K - H = 0.464$ nats, with a median of 10,560 training tokens per prototype and only 7 of 50,000 prototypes below 100. The motivation for a different target is label ambiguity rather than an undertrained tail.

The taxonomic bands are degrees of distribution shift, and coverage and discriminability decouple across that range. Mean assignment cosine to the nearest centroid barely moves — 0.908 in-corpus,

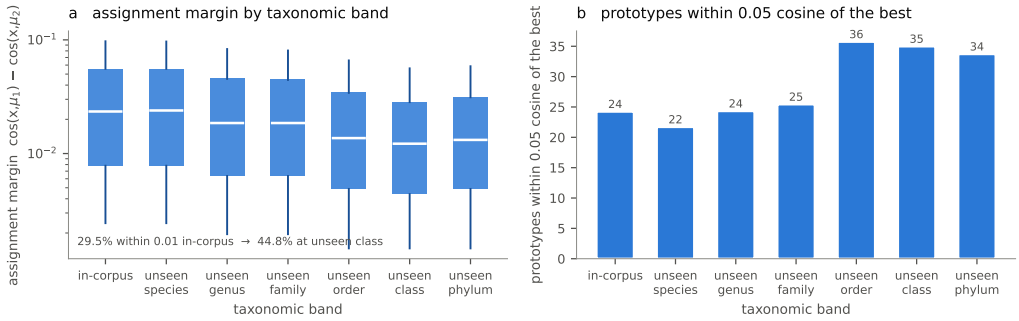


Figure 1: Coverage and discriminability under taxonomic shift. (a) Assignment margin, cosine to the nearest centroid minus cosine to the second, by taxonomic band, with the share of proteins whose margin is below 0.01. (b) A different and looser statistic: the mean number of prototypes within 0.05 cosine of the best. Persistence by band is in Appendix D.

0.909 at unseen species, 0.872 at unseen phylum — while the share of proteins whose second-best prototype lies within 0.01 cosine of the best rises from 29.5% in-corpus to 44.8% at unseen class (Figure 1a, all rows). A looser statistic, the mean number of prototypes within 0.05 cosine of the best, rises from 22–25 through unseen family to 34–36 from unseen order on (Figure 1b). Held-out proteins are well covered but poorly distinguished, and hard cross-entropy penalises a near neighbour as severely as an unrelated one.

A fifth of held-out gene mass is absent from the label space (Table 1). The corpus filter $t = \min(5, n)$, which we chose to generalise the fixed threshold of 20 in the corpus construction we started from to shallow species (Appendix A), drops clusters seen in fewer than t genomes of their species; those clusters were never in the codebook fit and never in training.

Table 1: Out-of-vocabulary genes in three of the six held-out bands and overall, on the 1,088 held-out species. Share is out-of-vocabulary over source genes. Exact counts; no uncertainty applies. All six bands: Table 4.

band	genes	out of vocabulary	share
d0 unseen species	5,340,349	984,770	18.44%
d4 unseen class	754,063	215,422	28.57%
d5 unseen phylum	620,405	183,404	29.56%
all	12,399,979	2,634,603	21.25%

The share rises monotonically with distance (20.73%, 23.51%, 23.64% at unseen genus, family and order; Table 4). This share is defined by observed corpus membership: a gene is out of vocabulary if it is in its source genome and absent from the corpus tokens, counted by subtracting the one from the other (Appendix B). An earlier count detected these genes by aligning contigs, which cannot see the 137,173 contigs the filter emptied entirely — they retain no token to align against — and reported 16.25% (Appendix E).

4 Closed-vocabulary performance comparisons

We compare nine arms under an identical corpus, backbone (12 layers, width 768, 85M parameters), masking scheme, 4B-gene-token budget and evaluation mask. Only the prediction target differs (Figure 2); Table 2 names each arm by a letter, and we use the letters from here on. Inputs and targets live in ESM-C's 960-dimensional space; $q = \text{normalize}(Wh)$ maps the 768-wide backbone output into it.

Table 2: The nine arms. Head parameters are instantiated counts including the class bias b_c; every arm also shares the 84,955,392-parameter backbone and a 443,136-parameter input projection.

arm	target	head params
A free classifier	learned $50{,}000 \times 768$ matrix $+ b_c$, hard cross-entropy	38,450,000
A′ factorised	$768 \to 256 \to 50{,}000 + b_c$	13,046,864
A″ matched to B	$768 \to 15 \to 50{,}000 + b_c$, B's trainable count (§4.1)	811,535
B metric head	$\text{logit}_c = (q \cdot \mu_c)/\tau + b_c$, μ frozen centroids	787,281
C + lineage	B plus a 3-d regression on gene-persistence residuals	789,588
R permuted	B with the label→centroid map permuted, b_c intact	787,281
D continuous	regression onto the masked protein's ESM-C embedding	737,280
E soft target	B's head, target a distribution over prototypes (§6)	787,281
F free + lineage	A's classifier plus arm C's identical lineage head (§7)	38,452,307

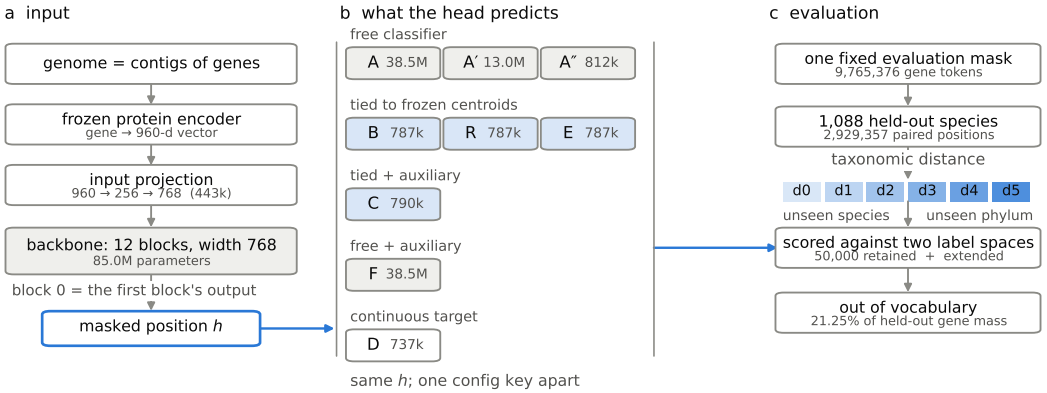


Figure 2: Model pipeline and evaluation: input dimensions, nine heads selected by a single configuration key, instantiated parameter counts, and evaluation design.

Each genome is presented as its contigs, each contig as its genes' prototype tokens in the order the upstream pangenome construction gives. That order is not genomic position (Appendix A), so we make no positional or syntenic claim anywhere.

We evaluate species-macro paired cross-entropy on `test`, using 1,088 held-out species, 2,929,357 paired positions and 5 seeds per arm. We bootstrap over species and evaluate every arm on the same fixed mask.

Arm B, the metric head, has higher cross-entropy than arm A, the free classifier, on the closed task: Δ(B−A) = +0.2249 nats, CI [+0.2112, +0.2393]. P1, the paper's primary hypothesis, failed, and we apply the interpretation registered for that outcome.

The cross-entropy gap decreases with taxonomic distance, falling from +0.4350 at unseen species to +0.0055 [−0.0076, +0.0188] at unseen phylum — an interval straddling zero, so we report an upper bound: at the greatest distance the gap is at most 0.019 nats, an at least twenty-three-fold decline, on a slope of −0.0887 per band, CI [−0.0950, −0.0835]. Species inside a phylum are not independent, so we recompute every band with the phylum as the unit (Table 9). The decline survives — +0.264 to +0.013 — but the bound widens to at most 0.051, and at the two most distant bands the phyla disagree in sign (9 of 15 and 7 of 13 positive). `d5` is 117 species in 13 phyla, 80 of them in one.

The cross-entropy gap persists under changes to the corpus's presentation order. With training repeated on a within-contig shuffle of the corpus (Appendix A), Δ(B−A) = +0.2188 [+0.2064, +0.2316] across two matched pairs 0.0006 apart — a difference from +0.2249 equal to an eighth of the 0.05 material threshold. Both arms improve by roughly 0.09 nats: changing the order affects both models similarly and has little effect on their comparison.

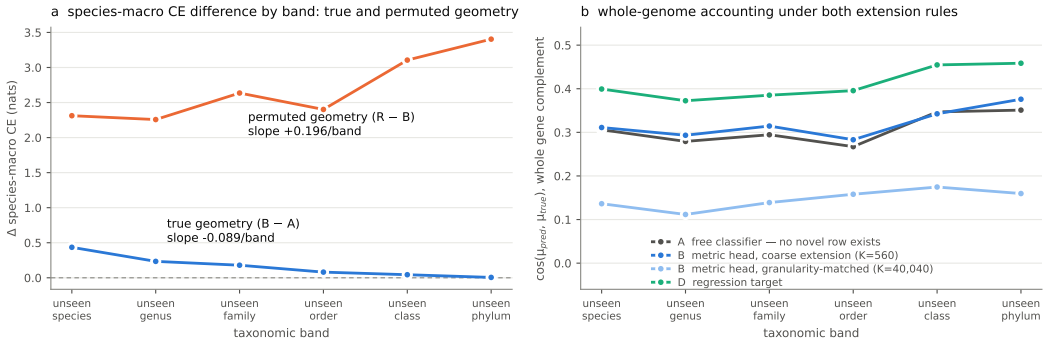


Figure 3: (a) The metric head's deficit decreases with taxonomic distance, whereas the permuted-geometry control's increases, despite identical head architecture and parameter count. (b) The whole-genome accounting under both pre-registered extension rules.

4.1 Trainable-parameter-matched and permuted-geometry controls

Arm R is arm B with the label→centroid map permuted and the class bias intact: the same architecture, parameter count and code path, with the label–centroid correspondence removed.

Permuting the map increases cross-entropy by +2.5517 nats, CI [+2.5115, +2.5926] — and its deficit grows with taxonomic distance, at +0.1959 per band (+2.31 at unseen species to +3.40 at unseen phylum), where the true geometry's shrinks at −0.0887. The two slopes have opposite signs (Figure 3a).

Architecture alone does not explain the B–R difference. Matching trainable parameter counts reverses the comparison with the free classifier. Arm A″, arm A's free head factorised to 811,535 head parameters, within 3.1% of arm B's 787,281, has +1.45 nats higher cross-entropy than arm A, and arm B improves on it by 1.22 [1.20, 1.24] — by more the further from training (−0.090 per band). Its rank is forced by the budget, not chosen: the 50,000 class biases cost 50,000 parameters and each rank dimension 50,769, so a dense free head of arm B's size has rank 14.5; arm A′, at rank 256 and +0.106 over arm A, is the intermediate point.

This matches trainable parameters, not capacity: arm B also reads a frozen 50,000 × 960 centroid matrix, information a free head of its size must learn instead. Nor does arm B win by receiving 960-dimensional class vectors: arm R receives a matrix of the same size and dimensionality at the same trainable count and is 1.33 nats worse than A″ (one seed, unpaired): the benefit is the geometry's correctness. Arm B's remaining gap reflects recovery of 84% of the cross-entropy increase associated with reducing the head size. This is the interpretation registered for this outcome, reported as directional under the registered spread rule (Table 11): A″ varies 6× as much as arm B, although the difference is 57 seed-level standard errors.

4.2 Backbone scale and taxonomic distance

Model sizes are quoted as backbone parameters, 85M for the main runs and 302M here (exact counts: Table 2, Appendix D); totals differ by head. We compare arms A and B with a 302M backbone (24 × 1024), using two matched pairs, the same 4B budget and the same effective batch size as the 85M runs. The arms do not scale symmetrically: arm A's output table falls from 0.45 of the backbone to 0.17, while arm B's tied head is negligible at both.

Δ(B−A) falls from +0.2249 at 85M to +0.1916 and +0.1912 in the two 302M pairs, with slopes of −0.0849 and −0.0866 per band against −0.0887 (every cell and interval: Table 8).

Increasing backbone capacity reduces the cross-entropy gap by about 15%, with little change in its taxonomic trend; the metric head benefits more. At unseen phylum the species-level sign flips in both pairs — −0.0290 [−0.0441, −0.0145] and −0.0430 [−0.0567, −0.0298] — but the reversal is driven by one clade: with the phylum as the unit it is +0.0022 [−0.0289, +0.0356] and −0.0144 [−0.0584, +0.0255], carried by Omnitrophota, 80 of the band's 117 species. We therefore do not

claim a reversal at 302M; what both sizes support is that at the greatest distance the gap is indistinguishable from zero.

This scale comparison is based on two seeds; the 85M five-seed result remains the primary result.

5 OPEN-VOCABULARY PREDICTION AND EXTENSION GRANULARITY

We extend the vocabulary to evaluate predictions for the genes excluded in §3. The extension uses 1,996,474 excluded representatives embedded with the encoder that built the corpus store; Appendix B describes construction and verifies pipeline identity.

Arm D, the continuous target, was included before evaluation to test whether "the capability follows from the target rather than from our particular head" (Appendix E). The results in §5.2 confirm this registered hypothesis and show that open-vocabulary evaluation depends on extension granularity.

The metric head's net advantage occurs with poor novel-cell discrimination. At $K_{\text{new}} = 560$, a decoder given the out-of-vocabulary protein's true embedding selects its assigned novel cell 7.7% of the time and a retained centroid 92.3% of the time: retained centroids are generally closer to the new cells' members than the new cells are. Refining the extension sufficiently to distinguish its cells (Figure 4b) removes the metric head's advantage.

5.1 NOVEL-FAMILY COVERAGE AND EXACT-FAMILY ACCURACY

Under the granularity-matched extension, where the true-embedding decoder scores comparably on novel (0.6201) and retained (0.6739) families, arm A, the free classifier, has accuracy 0 on novel families because no corresponding output row exists, while arms B–E, each with under 0.8M head parameters, produce a prediction. However, exact-family accuracy is near zero (Table 7). Relative to the true-embedding decoder, arm B, the metric head, reaches 0.17% on families absent from its vocabulary against 0.59% on families present in it; arm E, the soft target, reaches 0.19%, the best of the four. The decoder is a reference, not a bound: at coarse granularity its nearest centroid is usually a retained one, so a model predicting the assigned novel cell could beat it. We therefore make no claim of practical utility. Appendix C gives both extension variants and the per-arm ranking rules.

5.2 WHOLE-GENOME EVALUATION AND METRIC DEPENDENCE

The sign convention reverses here: in §4 a positive Δ(B−A) in cross-entropy means arm B is worse; on this cosine endpoint a positive Δ(B−A) means arm B is better. The primary endpoint was registered before the evaluation: $\cos(\mu_{\text{pred}}, \mu_{\text{true}})$ over every gene of held-out genomes, each kind reweighted to its band's out-of-vocabulary share — a weight only, since genes on contigs the filter emptied supply no evaluation position. Arm A, which has no novel row, is scored at the centroid of its argmax over the 50,000 classes it has rather than zeroed; the metric arms rank by their own logit, a novel prototype taking the mean learned bias (Appendix C).

Δ(B−A), Δ(C−A) (arm C, the metric head with a lineage head), Δ(D−A) are +0.0120, +0.0151, +0.1006 under the ratio rule (registered primary) and −0.1613, −0.1551, +0.1028 under the matched rule (registered sensitivity).

Arm B's advantage survives one of the two pre-registered rules; arm D's survives both, at every band (Figure 3b), and is 8× arm B's even where arm B wins — consistent with the registered hypothesis. The continuous reference target exhibits this capability; it is not specific to the metric head.

The two metrics favour different training objectives. $\cos(\mu_{\text{pred}}, \mu_{\text{true}})$ rewards landing near the right region of protein space — arm D's training objective — while top-1 over the extended vocabulary rewards committing to one cell, which is what A, B, C and E are trained to do. We report both metrics and their respective biases without selecting one as preferable.

Ranking every arm by cosine alone, as the registration specified, gives Δ(B−A) = −0.0312 at $K_{\text{new}} = 560$ on identical predictions, against +0.0120 under each arm's own rule: the ranking convention alone reverses the sign (Appendix C).

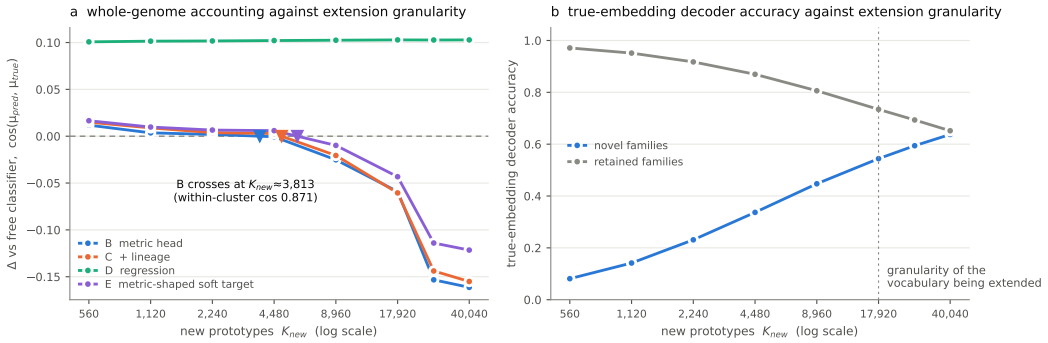


Figure 4: (a) Whole-genome performance as a function of extension granularity: the sign of the comparison depends on this parameter, and arm B crosses zero at a granularity coarser than the vocabulary it extends. (b) The true-embedding decoder – nearest centroid to the protein's own embedding – on novel and retained families: original prototypes are selected more often than a coarse extension's novel cells.

5.3 Performance across extension granularities

To locate the sign reversal between the two registered extension rules, we swept K_{new} across eight codebooks, each fit by the same spherical Lloyd on the same support-side rows with the same frozen centring vector (Figure 4a; Appendix C).

Δ(B−A) crosses zero at $K_{\text{new}} \approx 3{,}813$, at a within-cluster cosine of 0.871 — coarser than the 0.894 of the vocabulary being extended. The metric head is ahead only where the new vocabulary is coarser than the one it extends; at the existing granularity it is already 0.03–0.06 behind.

Arm D remains positive, at +0.1007 to +0.1028 across a 72-fold range of K_{new}. Its continuous target's advantage does not depend on extension granularity over this range. The metric head's advantage at coarse granularity does: cells that are easy to predict correspond to regions already covered by the retained vocabulary.

6 Geometry-based soft targets

§3 measured two things about the label space: it is metric and it is ambiguous. Arm B, the metric head, incorporates the metric geometry but uses a one-hot target that imposes distinctions the label space does not support; arm D, the continuous target, accommodates ambiguity by abandoning categorical structure. Arm E, the soft target, incorporates both properties through a distribution over prototypes, shaped by the geometry the head scores in. This is knowledge distillation in form (Hinton et al., 2015), with the teacher replaced by the label space's own geometry.

$$t_i(c) = \text{softmax}_c\left(\frac{\langle z_i, \mu_c \rangle}{\tau_t}\right), \qquad L = \text{KL}\Big(t_i \,\|\, \hat{P}(\cdot \mid h_i)\Big),$$

where z_i is the masked protein's own mean-centred, L2-normalised ESM-C vector — the input the model never saw at that position — and $\hat{P}(\cdot \mid h_i)$ is the head's softmax over prototypes, arm B's form.

Arm E inherits arm B's head and `score_prototypes`: evaluation code is identical, with one configuration-key difference. The target temperature τ_t is derived rather than tuned: it solves mean top-1 target mass = 1 − near-tie rate, giving 0.010951, fixed before training.

Species-macro CE is 7.2844 ± 0.0062 for arm B over five seeds and 7.2770 ± 0.0263 for arm E over three, giving Δ(E−B) = −0.0074 and arm E's deficit to arm A, the free classifier, is +0.2175.

Arm E is 0.0074 nats better — 3.3% of arm B's cross-entropy gap — and the species-bootstrap interval excludes zero, but its seed spread is 4.2 times arm B's (3.7 times that of arm C, the metric

head with a lineage head): at the seed level the arms are not distinguishable in either direction, and the one-seed version had the opposite sign (Appendix E).

The two registered experiments constrain the explanation for the gap. Arm R, the permuted-geometry control, differs substantially from arm B and its deficit increases with distance despite identical architecture, parameter count and code path, excluding architecture alone as the explanation. Arm E shows that the geometry-based target reshaping motivated by our ambiguity analysis does not resolve the gap. We tested one reshaping at one derived temperature and do not claim that no reshaping could resolve it. Nor is the gap attributable to discreteness, since the free classifier also predicts a discrete family without this deficit. Both results differ from our preregistered expectations.

7 CROSS-TAXON GENE-PERSISTENCE PREDICTION

Arm C adds to the metric head a 3-d regression onto gene-persistence residuals at species, genus and family rank. This is not a vocabulary item but a population statistic about how persistent a gene family is within a clade, which a 50,000-way softmax cannot express. The target is a smoothed, leave-one-lineage-out residual $y = \mathrm{logit}(\tilde{p}) - \mathrm{logit}(\pi)$ (Appendix D); for held-out species its counts come from held-out genomes and species, and only the prior $\pi_{c,r}$ is frozen from training.

Against $\hat{y} = 0$, on `test`, three seeds, every one of the 21 band $\times$ rank cells is positive, R^2 0.3116–0.5200 out to unseen phylum; overall 0.3424 ± 0.0026 at species rank, 0.4503 ± 0.0118 at genus, 0.4879 ± 0.0156 at family. For species whose entire genus was withheld, genus-rank persistence is predicted at $R^2 = 0.3952 \pm 0.0089$, from gene content alone (every cell: Table 10).

Persistence prediction and its cross-entropy increase are not specific to the metric head. Arm F is the free classifier plus the identical lineage head — verified to use identical code, differing from arm A, the free classifier alone, by one configuration key and evaluated over three seeds. Δ(C$-$B) = +0.0125 [+0.0111, +0.0139] on the metric head; Δ(F$-$A) = +0.0176 [+0.0162, +0.0191] on the free classifier; both flat across taxonomic distance (slopes -0.0013 and -0.0003 per band), unlike the metric head's cross-entropy gap, which decreases steeply.

They differ by 0.0051 — non-overlapping intervals, but a tenth of the 0.05 threshold. Arm F's R^2 is positive in all 21 cells, 0.3016–0.5241; it is slightly weaker, mean gap -0.0107, lower in 16 of 21 cells (reported descriptively: three cells aggregate the others and the ranks share observations). The metric geometry contributes a small systematic amount; it does not create the capability. Persistence prediction is therefore a separate result: a prediction outside the label space, at nearly full strength, for an additional +0.0176 nats on the better-performing arm in the P1 comparison. At 302M the additional cross-entropy on the metric head is +0.0199 on one seed; unlike the metric head's gap, this increase does not shrink with capacity, although the result is directional only (Appendix D).

Predicted persistence could separate genes truly absent from a draft genome from assembly gaps: a family predicted highly persistent but not observed is a candidate for reassembly.

Two limitations apply. Since y is a residual against the family's prior, $\hat{y} = 0$ already predicts each family's corpus-wide persistence, so the R^2 is what context adds beyond it; a stronger baseline — persistence in the nearest training clade — was not run. The two heads' differences (0.0051 nats, 0.0107 R^2) are of the order of the seed spreads, with no third base head tested.

8 FROZEN-PROBE PERFORMANCE ACROSS LAYERS

We evaluate frozen representations with linear probes and one hyperparameter-selection rule on three tasks: essential genes and operon membership are gene-level, and phenotype, genome-level, is in Appendix F. Operon is scored by leave-one-genome-out (LOGO) cross-validation over eight genomes. The comparator is a protein-only baseline without a genome model: ESM-C's protein embedding, mean-pooled. Block 0 is the output of the first transformer block — the ESM-C vectors after the input projection and one attention layer over the genome, not the raw input the floor is built from — and was fixed as the first block before the sweep rather than selected by performance.

Better frozen representations at early layers have been reported for masked language models generally (Alain & Bengio, 2016; Tenney et al., 2019) and for genomic embeddings specifically (West-

Table 3: Frozen-probe AUPRC for the arms of Table 2, mean±sd over seeds (A, B: 5; C, E, F: 3; D, Bacformer: 1). † denotes the registered primary layer. Above-floor means are described in the text.

task, layer	floor	A	B	C	D	E	F	Bacf.
essential, block 0	0.5045	0.5228±0.012	0.5314±0.010	0.5388±0.004	0.5583	0.5309±0.018	0.5425±0.006	0.5901
essential, final†	0.5045	0.4384±0.016	0.4601±0.009	0.4638±0.009	0.4903	0.4860±0.005	0.4510±0.011	0.4325
operon, block 0	0.4129	0.4214±0.003	0.4098±0.006	0.4118±0.009	0.4222	0.4130±0.003	0.4238±0.005	0.4312
operon, final†	0.4129	0.3913±0.006	0.3893±0.006	0.3909±0.004	0.4106	0.3962±0.003	0.3914±0.004	0.3868

Roberts et al., 2024). Our contribution is the comparison across prediction targets under matched training, including the difference between the final-layer and early-layer conclusions.

At block 0, genome context improves on the protein encoder on essential genes. Every arm's mean exceeds the floor, as do 19 of 20 checkpoints (one arm A seed misses by 0.0004); the improvement satisfies the registered test — mean minus two seed standard deviations above the floor — for arms B, C and F, not for A or E.

The operon result remains unresolved: means exceed the floor for A, E (by 0.0001), F and single-seed D and fall short for B and C, and only A and F satisfy the test. The earlier seed-1 result, in which three of four arms exceeded the operon baseline, was one observation within this spread.

The profile declines with depth in every arm on both tasks, though not monotonically (Appendix F). At the final layer — the registered primary — the result is a null: none of our tested representations or Bacformer's exceeds the floor. We keep that as the registered result, because selecting a layer by score is excluded elsewhere in this work. Block 0 was fixed independently of score, so we report it alongside the primary result. The ordering among our arms is inconsistent across tasks, so Table 3 supports no ranking.

The final-layer null holds for every target and seed. Concatenation also exceeds the floor, and the ordering control bounds only gene order (Appendix F); BacBench's reported figures were unavailable, so no claim relies on them.

9 Limitations and discussion

The primary comparison covers 78.75% of held-out gene mass — every gene that reaches the corpus text; the other 21.25% is out of vocabulary. The corpus token count and the evaluation split agree exactly, so nothing is lost between them; only the 21.25% enters §5's accounting.

The scale check covers one additional size rather than a scaling law; it rests on two matched pairs and arm C, the metric head with a lineage head, at 302M on one seed, and its unseen-phylum reversal is carried by a single phylum (§4.2). Our corpus is 282,312 genomes against Bacformer's roughly 1.3M, so no cross-model comparison supports a superiority claim either way.

Two findings concern evaluation beyond the bacterial setting. §5.2 and §5.3 concern how open-vocabulary prediction is evaluated: no metric is neutral between a categorical and a continuous target, and the sign of the comparison depends on extension granularity. Both can arise in any label space built by clustering an encoder; we show they do in one, not how often.

The effect of output geometry depends on distribution shift. Near training, the free classifier performs better and the constraint enabling open-vocabulary prediction increases cross-entropy; far from training, the gap is indistinguishable from zero at this resolution. Its performance therefore depends on distance from training, the regime most applications occupy and a random split cannot assess. The target also determines what the model can represent: a closed vocabulary has no label of its own for a fifth of a novel genome's genes, although the resulting coverage advantage survives only one of two defensible extension rules. Part of the cross-entropy gap reflects head size: at matched trainable parameters, the true geometry is 1.22 nats ahead. The population statistic in §7 is predicted with consistently positive results for an additional 0.012 to 0.018 nats on either head. Frozen-probe conclusions also depend on the layer evaluated (§8). Appendix E documents the preregistration and the verification record.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Maxwell L. Bileschi, David Belanger, Drew H. Bryant, Theo Sanderson, Brandon Carter, D. Sculley, Alex Bateman, Mark A. DePristo, and Lucy J. Colwell. Using deep learning to annotate the protein universe. *Nature Biotechnology*, 40:932–937, 2022.

Moritz Buck, Maliheh Mehrshad, and Stefan Bertilsson. motupan: a robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. *NAR Genomics and Bioinformatics*, 4(3):lqac060, 2022.

Andre Cornman, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin Beracochea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. The OMG dataset: An open metagenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, 2024. doi: 10.1101/2024.08.14.607850.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 2025. doi: 10.1038/s41592-024-02523-z.

EvolutionaryScale Team. ESM Cambrian: Revealing the mysteries of proteins with unconditional protein language models. Model release `esmc-300m-2024-12`, 2024.

Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, 2013.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NIPS 2014 Deep Learning Workshop.

Yunha Hwang, Andre L. Cornman, Elizabeth H. Kellogg, Sergey Ovchinnikov, and Peter R. Girguis. Genomic language model predicts protein co-regulation and function. *Nature Communications*, 15:2880, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336.

Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 2015.

Donovan H. Parks, Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, 2022.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Gerry Tonkin-Hill, Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh, John A. Lees, Rebecca A. Gladstone, Stephanie Lo, Christopher Beaudoin, R. Andres Floto, Simon D. W. Frost, Jukka Corander, Stephen D. Bentley, and Julian Parkhill. Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome Biology*, 21(1):180, 2020.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan L. Yuille. NormFace: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM International Conference on Multimedia*, 2017.

Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse genomic embedding benchmark for functional evaluation across the tree of life. *bioRxiv*, 2024. doi: 10.1101/2024.07.10.602933.

Maciej Wiatrak et al. A contextualised protein language model reveals the functional syntax of bacterial evolution. *bioRxiv*, 2025. 2025.07.20.665723.

Maciej Wiatrak et al. Bacbench: Evaluating genomic language models for bacteria. OpenReview submission 2wu96xvRIG, 2026. Unpublished; no number reported by it is used in this paper.

Xu Zhang, Felix X. Yu, Svebor Karaman, Wei Zhang, and Shih-Fu Chang. Heated-up softmax embedding. *arXiv preprint arXiv:1809.04157*, 2018.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. In *International Conference on Learning Representations*, 2024.

USE OF AI ASSISTANTS

The authors conceived the research question, formulated the hypotheses, designed the nine-arm comparison and the matched-training protocol, chose the prediction targets and training objectives, designed the evaluation and its taxonomic banding, and set the pre-registration discipline under which the work was run. Some experimental controls, analyses and corpus choices were first proposed by AI tools or AI-assisted review, including the trainable-parameter-matched control, the clade-level robustness analysis and the depth-aware corpus filter, and the authors decided whether to adopt each.

Within that design, we used generative AI tools for implementation, execution and drafting: to write and refactor pipeline, training, evaluation and analysis code; draft preregistration amendments; run registered experiments and generate figures, tables and the number registry; retrieve literature; and draft and revise this manuscript.

We did not use generative AI tools to generate, synthesise or impute data, produce numbers not computed by the released code from released artifacts, or decide endpoints, thresholds or interpretations. Formal proofs and theoretical derivations are not applicable because this work is entirely empirical.

The authors reviewed all AI-assisted work and set, reviewed and approved every experimental design decision, preregistered endpoint, threshold, interpretation and manuscript claim. Verification included the following procedures:

- Endpoints, thresholds and both interpretations of each comparison were registered before the relevant runs, with each amendment committed separately (Appendix E). Four preregistered adverse outcomes occurred, including failure of the primary hypothesis; each was interpreted as specified.
- A split fingerprint is checked at every training start and a `config_core_sha256` at every matched run. Artifacts pass 28 leakage checks and 30 corpus-integrity checks.
- Each manuscript number is generated into a registry of 1,130 values linked to source artifacts and generating scripts, without manual transcription. The anonymised release reproduces 1,053 values from a clean checkout with zero mismatches.
- Cross-checks reconciling derived artifacts with their sources identified eleven defects. Three had affected results reported in an internal draft and were corrected. Appendix E documents all eleven and states whether each was identified before or after reporting.

Several defects in Appendix E originated in AI-generated analysis code and were identified through AI-executed cross-checks. Preregistration was adopted specifically to constrain post-hoc analysis given the speed at which these tools can generate analyses.

The authors take responsibility for the final content, including all claims, numerical results and code.

A Corpus construction, ordering and shuffle controls

Corpus and split. The corpus is a pangenome release over GTDB r214 species (Parks et al., 2022) in which every gene is assigned to a protein cluster with one representative protein, and labelled core or accessory by mOTUpan over its species' genomes (Buck et al., 2022). A cluster is retained in a species if it occurs in at least $t = \min(5, n)$ of the species' n genomes; core clusters are always retained. This rule is ours: it replaced the fixed threshold of 20 in the corpus construction we started from, keeps every cluster that threshold kept, and was chosen among bounded alternatives for the accessory clusters it adds in species with at least 20 genomes, where occupancy can be estimated; the corpus was rebuilt under it. The corpus has 27,669 species and 290,981 genomes. The split is by species and is built before any other artifact: 26,243 training species (284,652 genomes), 338 validation species (1,602) and 1,088 test species (4,727), holding out 1.51% of gene tokens. A held-out species' band is its taxonomic distance to the nearest training species: `d0`, its genus retains a training species; `d1`–`d4`, its genus, family, order or class is fully held out while the next rank up retains a training member; `d5`, its phylum is fully held out. Because a band requires removing a whole clade, bands are carved coarse to fine, and each species' band is then recomputed from the final training set and asserted to match. Species with at least 300 genomes, and any clade containing one, are kept in training. A seventh evaluation band, in-corpus, withholds two genomes from each training species with at least 20 genomes; training uses the remaining 282,312 genomes. Every downstream artifact is built through a guard that reads only the training-species list, and training asserts the split fingerprint at start (Appendix E). The codebook's fit sample, 12,800,000 rows, is drawn from training species only; 29 of those rows (0.0002%) occur only in the in-corpus band's withheld genomes, so they are held-out representatives that entered the fit, and we count them as leakage.

The corpus presents each contig as a sequence of prototype tokens in the order the upstream pangenome construction gives. That order is not genomic position. It is monotone in the true gene coordinate for 3.26% of contigs, and no candidate sort key explains it either — representative protein position (3.26%), cluster identifier lexically (2.97%) or numerically (2.69%), cluster prevalence (never testable, fewer than three distinct values within a contig). Genomic position is recoverable from the source proteomes — FASTA header order reproduces the definitional protein index for 100.00% of 5,598 contigs — but it is not what the corpus encodes. We therefore make no positional, syntenic or neighbourhood claim anywhere in the paper. This has three consequences: the contig-bigram baseline is a bigram over the corpus ordering, not genomic adjacency; $R^2 - R^2_{\text{null}}$

measures the part attributable to the corpus's ordering, so the permutation null removes less than its registration assumed; and the operon task is unaffected because it uses BacBench's own coordinates.

Evaluation-time shuffling increases arm A's cross-entropy by +0.4852 nats. This increase of almost half a nat is ten times the 0.05 threshold preregistered in `12j` for requiring a training ablation. It is an upper bound, not an estimate: a model trained on one ordering degrades on any other whether or not the order carries information.

In the training ablation, the shuffled-order model is 0.0620 nats better. We compare arm A trained on the corpus order with arm A trained on a within-contig shuffle, matched at 1B gene tokens and each evaluated on its own corpus: 7.5330 against 7.4710, CI [+0.0558, +0.0679] over 1,088 species. The corpus ordering degrades performance. However, the evaluation has an asymmetry: the fixed eval mask selects flat positions and the shuffle changes which gene sits at each, so the two evaluations do not mask an identical set of genes and the 0.0620 is a direction and rough magnitude, not a precise effect size.

At the full 4B budget the effect is larger — seed 1 of arm A improves from 7.0590 to 6.9727 and seed 1 of arm B from 7.2856 to 7.1918 — while Δ(B−A) moves from +0.2249 to +0.2188 (§4). The two Δ are paired over every seed (five as-is, two shuffled) and every evaluated position; they are not the difference of these single-seed means, which would give 0.2266 and 0.2191. These results support shuffling within contigs before training on a pangenome corpus of this form, without additional training expenditure.

B BUILDING THE OUT-OF-VOCABULARY EVALUATION SET

The 1,996,474 excluded representatives were resolved from the source proteomes (1,996,474/1,996,474, zero length mismatches against FAA headers) and embedded with the encoder that built the corpus store. We verified pipeline identity by re-embedding retained representatives: 2,048 retained representatives were processed through the same code path and compared against their stored vectors — max absolute deviation 4.3×10^{-4}, minimum cosine 0.9999996.

Table 4 reports out-of-vocabulary counts for all six bands over source genes from the corpus's 1,088 held-out species (§3 gives four rows). The count is a subtraction: every gene in the source proteomes of those genomes, minus every token in the corpus text. It uses no alignment, so a contig whose genes were all filtered — which leaves no corpus token to align against — is counted like any other. The 127,857 genes of 35 genomes on the benchmark-exclusion list are outside the denominator; they are not in the corpus at all.

Table 4: Out-of-vocabulary genes in all six held-out bands, on the 1,088 held-out species, counted by subtraction (source genes minus corpus tokens). Exact counts; no uncertainty applies.

band	genes	out of vocabulary	share
d0 unseen species	5,340,349	984,770	18.44%
d1 unseen genus	3,125,252	647,952	20.73%
d2 unseen family	1,579,707	371,376	23.51%
d3 unseen order	980,203	231,679	23.64%
d4 unseen class	754,063	215,422	28.57%
d5 unseen phylum	620,405	183,404	29.56%
all	12,399,979	2,634,603	21.25%

The counts reconcile by construction. As an independent check, the corpus-token count from the text equals the 9,765,376 gene tokens in the pre-tokenised evaluation split exactly, confirming that no genes are lost between the corpus and evaluation.

The decomposition by cause uses the alignment pass (`scripts/recount_excluded.py`) and the orphan-contig pass (`scripts/recount_orphans.py`): 2,036,341 genes on contigs that survive — 2,035,552 the frequency threshold, 633 unattributable, 156 without an embedding — and 599,301 on 137,173 contigs the filter emptied entirely, 599,051 of them the frequency threshold. Those sum to 2,635,642, which is 1,039 more than the subtraction: the genes of the 186 corpus

contigs the alignment never placed, counted on both sides. We therefore report the cause-specific decomposition separately and use subtraction for the total.

Genes on emptied contigs have no surviving context to be inserted into, so §5's evaluation positions are drawn from aligned contigs; the reweighting in §5.2 uses the full share. The representatives embedded for the extension (1,996,474 clusters) come from a count that already included the emptied contigs, so the extended vocabulary itself was never incomplete — only the share was.

Evaluation sequences follow `12k`'s option (b): the context is exactly what the model was trained on — retained genes, corpus order, contig structure intact — and an out-of-vocabulary gene is inserted at one position, masked. Its position is not drawn: a contig's corpus tokens are an order-preserving subsequence of exactly one source contig's row order, so aligning the two recovers where the frequency filter took the gene from. Alignment succeeds for 505,149 of 505,335 contigs (99.96%) with zero genes unplaceable, yielding 762,609 leakage-free novel positions, of which we evaluate 30,000 per band against 30,000 matched in-vocabulary controls from the same genomes — 360,000 single-mask sequences per arm.

Label leakage is prevented at the cluster level rather than the genome level (`12l` §1). Every occurrence of a cluster shares one representative embedding, so a 50/50 genome split alone would leave a query gene's own vector inside its own label's centroid. A cluster enters the clustering only if every occurrence lies in a support genome, and is evaluated only if every occurrence lies in a query genome; 340,966 clusters (17.08%, 33.8% of occurrences) span both partitions and enter neither. The disjointness is asserted in code.

Table 5 reports granularity before accuracy is examined (`12i`).

Table 5: Granularity of the retained vocabulary and of the two registered extensions: codebook size, mean within-cluster cosine and representatives per cell. Each is a single codebook fit; no uncertainty is estimated.

codebook	K	mean within-cluster cosine	representatives per cell
original (retained vocabulary)	50,000	0.8942	192.2
ratio (registered primary)	560	0.8392	1,287.6
matched (registered sensitivity)	40,040	0.9093	18.0

The registration's arithmetic gave N_{new} = round(1,996,474 / 1,286.5) = 1,552; the same ratio over the 721,032 rows the leakage boundary permits into the fit gives 560 (`12l` §2). This applies the same rule to the eligible support rows. As anticipated in `12i`, the ratio rule produces cells 6.7× coarser than the vocabulary it extends.

C Evaluation across K_{NEW} values

We fit eight codebooks with the same spherical Lloyd algorithm, seed, support-side rows and frozen centring vector (Table 6). Each arm is re-scored against all eight from one saved query vector per masked position, using its own scoring rule, as in §5.2. The curve agrees with the two registered points: exactly at 40,040 cells (-0.1613), and to 0.0002 at 560 (+0.0118 against +0.0120), because the ladder's 560-cell rung is a refit rather than the registered codebook itself. Arm A is K-invariant by construction and is not re-scored: its output layer has no novel row, so its net accounting cannot depend on how the novel vocabulary is built (0.3022 at every rung).

Crossings: arm B at $K_{\text{new}} \approx 3{,}813$ (cosine 0.871), arm C at 4,874 (0.875), arm E at 5,815 (0.878). Arm D does not cross zero. The vocabulary being extended has a within-cluster cosine of 0.8942, between the 8,960- and 17,920-cell rungs. Its cells are tighter, and so finer, than those at every crossing: the metric head is ahead only where the extension is coarser than the vocabulary it extends (§5.3).

The ladder's decoder row is pooled over positions on eight refit codebooks, so its 560-cell rung reads 0.081. Table 7 is species-macro on the registered codebooks, where the same quantity reads 0.0773

Table 6: Extension-granularity ladder over eight refit codebooks. Δ is the difference from arm A in the whole-genome endpoint $\cos(\mu_{\text{pred}}, \mu_{\text{true}})$, positive where the arm is better; the decoder row is the true-embedding decoder's accuracy on novel families, pooled over positions. Seed 1 of each arm; point estimates without intervals.

K_new	560	1,120	2,240	4,480	8,960	17,920	26,880	40,040
within-cluster cosine	0.8394	0.8518	0.8632	0.8739	0.8845	0.8953	0.9021	0.9093
reps per cell	1,288	644	322	161	80	40	27	18
Δ(B−A)	+0.0118	+0.0036	+0.0019	−0.0006	−0.0252	−0.0591	−0.1533	−0.1613
Δ(E−A)	+0.0165	+0.0098	+0.0066	+0.0059	−0.0098	−0.0432	−0.1140	−0.1216
Δ(D−A)	+0.1007	+0.1015	+0.1017	+0.1021	+0.1025	+0.1028	+0.1027	+0.1028
true-embedding decoder, novel families	0.081	0.142	0.231	0.337	0.447	0.544	0.594	0.637

(pooled on the registered codebook: 0.0825). Both the averaging rule and codebook differ, but the interpretation is unchanged: at 560 cells the decoder lands on a retained centroid 92.3% of the time (species-macro; 91.7% pooled) and on some other novel cell 0.01%.

Scoring on the extended vocabulary. Every arm's prediction is a class over the 50,000 retained plus K_{new} novel prototypes, and μ_{pred} is that class's centroid. The net endpoint is $\cos(\mu_{\text{pred}}, \mu_{\text{true}})$, μ_{true} the gene's own mean-centred, L2-normalised ESM-C vector.

- Arms B, C, E rank by their own logit $(q \cdot \mu_c)/\tau + b_c$. A novel prototype has no learned b_c; it is given the mean of the 50,000 learned biases. Zero is not neutral — the learned biases are log-frequencies with mean -10.9, so assigning zero would give novel cells a much larger prior. Cosine-only ranking reverses the sign of the primary comparison in the sensitivity analysis (§5.2).
- Arm D ranks by cosine; it has no bias.
- Arm A has no row for a novel prototype. Its μ_{pred} is the centroid of its argmax over the 50,000 retained classes — its available fallback — so on exact-family accuracy it scores zero on novel families by construction, and on the net endpoint it is credited for landing near the right region of protein space, never zeroed.

Table 7: Exact-family accuracy on novel families at the two registered extension rules, species-macro, each arm scored by its own rule; the decoder column gives novel / retained families. Point estimates without intervals. Every arm is below 1% on novel families, the outcome `12i` and `12k` register as a nominal capability.

variant	true-embedding decoder (novel / retained)	A	B	C	D	E
ratio	0.0773 / 0.9747	0	0.0098	0.0095	0.0065	0.0094
matched	0.6201 / 0.6739	0	0.0011	0.0009	0.0008	0.0012

Arm E's rows were computed after the other four using a routine verified to reproduce all previously reported rows above to within 6×10^{-4}.

D BACKBONE SCALE, SEED VARIABILITY AND LINEAGE TARGETS

The 302M pairs (Table 8). Arms A and B at 24 × 1024 — 301,993,984 backbone parameters, against the main runs' 84,955,392 at 12 × 768 — two matched pairs, 4B tokens, effective batch held identical to the 85M runs (microbatch halved, gradient accumulation doubled), sharing a config-core hash, batch plan and backbone fingerprint.

Table 8: Δ(B−A) in species-macro cross-entropy (nats; positive: arm B worse) at 85M over five seeds and in the two 302M matched pairs, overall, by band and as a per-band slope. Brackets: 95% bootstrap intervals over species.

	85M (5 seeds)	302M pair 1	302M pair 2
Δ(B−A) overall	+0.2249 [+0.2112, +0.2393]	+0.1916 [+0.1780, +0.2057]	+0.1912 [+0.1781, +0.2047]
d0	+0.4350	+0.3919 [+0.3643, +0.4193]	+0.3923 [+0.3652, +0.4184]
d1	+0.2342	+0.1987 [+0.1796, +0.2167]	+0.2019 [+0.1809, +0.2224]
d2	+0.1794	+0.1561 [+0.1408, +0.1706]	+0.1534 [+0.1370, +0.1696]
d3	+0.0809	+0.0422 [+0.0274, +0.0575]	+0.0490 [+0.0330, +0.0645]
d4	+0.0439	+0.0413 [+0.0238, +0.0578]	+0.0393 [+0.0208, +0.0569]
d5	+0.0055 [−0.0076, +0.0188]	−0.0290 [−0.0441, −0.0145]	−0.0430 [−0.0567, −0.0298]
slope	−0.0887	−0.0849	−0.0866

`12o` registered, while the replication was running, that two concordant pairs would be reported as a scale-dependent reversal with two seeds stated as the entire evidence base, that discordant pairs would be reported as not surviving replication, and that no third pair would be run either way. The pairs are concordant and nothing further was run.

The lineage head at 302M. Arm C at 24 × 1024, one seed, against the 302M arm B of pair 1: Δ(C−B) = +0.0199 [+0.0161, +0.0235], against +0.0125 at 85M on three seeds. The lineage head's additional cross-entropy does not shrink with capacity as the metric head's gap does; with one seed, this result is directional only.

Clade-level dependence (`12s`). The band intervals in §4 resample species, and species inside a phylum are not independent. Table 9 recomputes every band with the phylum as the unit, from the same cached per-position losses.

Table 9: Δ(B−A) by band with the species and with the phylum as the resampling unit (nats; positive: arm B worse). sp.: held-out species; phyla: phyla represented; in largest: species in the band's largest phylum; +/−: phyla with positive / negative Δ. Brackets: 95% bootstrap intervals over the stated unit.

band	sp.	phyla	in largest	Δ, species unit	Δ, phylum unit	+/−
d0 unseen species	335	32	84	+0.4350 [+0.408, +0.463]	+0.2637 [+0.204, +0.322]	31/1
d1 unseen genus	222	26	87	+0.2342 [+0.213, +0.255]	+0.1213 [+0.063, +0.180]	20/6
d2 unseen family	167	19	69	+0.1794 [+0.163, +0.195]	+0.0952 [+0.058, +0.133]	17/2
d3 unseen order	149	9	62	+0.0809 [+0.066, +0.097]	+0.0728 [+0.020, +0.124]	8/1
d4 unseen class	98	15	26	+0.0439 [+0.020, +0.066]	-0.0004 [-0.063, +0.056]	9/6
d5 unseen phylum	117	13	80	+0.0055 [-0.007, +0.019]	+0.0125 [-0.019, +0.051]	7/6

The species-unit column is recomputed inside the same script so that both units are bootstrapped identically; it reproduces the registered intervals to 0.0002 (d5: [−0.007, +0.019] here against the registered [−0.0076, +0.0188]). The cross-entropy gap and its decline remain evident: at the four nearer bands the phyla agree almost unanimously and the clade-level interval excludes zero. The two most distant bands do not resolve at clade level — the phyla split 9/6 and 7/6 — and `d5`'s bound widens from 0.019 to 0.051. `d5` is dominated by one phylum: Omnitrophota is 80 of its 117 species, and the 302M species-level reversal is driven by that phylum (pair 1 −0.0428, pair 2 −0.0558 within Omnitrophota; +0.0022 [−0.0289, +0.0356] and −0.0144 [−0.0584, +0.0255] with the phylum as the unit). The species-level bootstrap remains the registered endpoint and is what §4 reports first;

this is reported beside it, as `12s` registered, with corresponding qualifications on the claims based on `d5`.

Measured seed spreads, species-macro CE on `test`: arm B 0.0062 (5 seeds), arm C 0.0071 (3), arm E 0.0263 (3). Arm E is 4.2× the other two, indicating higher variability without a resolved performance improvement for the soft target.

The lineage target is defined as follows. For a gene token carrying prototype c in a genome of species s, at rank $r \in$ species, genus, family, the support units are the other genomes of s, the other species of its genus, and the other genera of its family — leave-one-lineage-out, so the token's own lineage never counts toward its own target. With n the number of support units, k the number containing c, and $\pi_{c,r}$ the corpus-wide prior computed on training species,

$$\tilde{p} = (k + \alpha\,\pi_{c,r})/(n + \alpha),\ y = \mathrm{logit}(\tilde{p}) - \mathrm{logit}(\pi_{c,r}),\ w = n/(n + \alpha),\ \alpha = 5,$$

with y clipped to ± 6. A rank with no support units has $w = 0$ and enters neither the loss nor the R^2. y and w depend on (species, prototype) alone. For a held-out species the same construction is applied with the support units drawn from the held-out genomes and species themselves; only π is frozen from training, so no held-out occupancy enters the prior. For a species whose whole genus was withheld, genus-rank support is the other held-out species of that genus: the target is defined from held-out data alone, and a genus with a single held-out species has $n = 0$ and contributes nothing at that rank.

Table 10 reports every band-by-rank cell for arm C, using three seeds and R^2 against $\hat{y} = 0$ on `test`; arm F is the same auxiliary head on the free classifier, three seeds.

Table 10: Gene-persistence R^2 against $\hat{y} = 0$ on `test`, every band × rank cell, arm C / arm F. Arm C: mean ± sd over three seeds; arm F: mean over three seeds.

band	species C / F	genus C / F	family C / F
ALL	0.3424 ± 0.0026 / 0.3366	0.4503 ± 0.0118 / 0.4383	0.4879 ± 0.0156 / 0.4942
d0 unseen species	0.3475 ± 0.0023 / 0.3453	0.4970 ± 0.0137 / 0.4878	0.5200 ± 0.0135 / 0.5241
d1 unseen genus	0.3388 ± 0.0048 / 0.3325	0.3952 ± 0.0089 / 0.3836	0.4840 ± 0.0146 / 0.4956
d2 unseen family	0.3291 ± 0.0060 / 0.3135	0.3959 ± 0.0106 / 0.3656	0.4504 ± 0.0201 / 0.4522
d3 unseen order	0.3374 ± 0.0020 / 0.3212	0.3453 ± 0.0291 / 0.3188	0.3382 ± 0.0336 / 0.3477
d4 unseen class	0.3251 ± 0.0063 / 0.3171	0.4725 ± 0.0130 / 0.4332	0.4676 ± 0.0316 / 0.4644
d5 unseen phylum	0.3514 ± 0.0144 / 0.3233	0.3570 ± 0.0264 / 0.3236	0.3116 ± 0.0180 / 0.3016

All 21 arm F cells are positive. The mean gap is −0.0107, the size of arm C's own across-seed spread, but consistent in sign: arm F is lower in 16 of 21 cells, reported descriptively — the ALL row aggregates the other eighteen cells and the three ranks share observations, so a sign test over the 21 would treat dependent cells as independent.

Table 11 assesses seed variability in light of the sign reversal in §E.2. Ratios are given against arm E's spread (a conservative reference because it is the noisiest arm measured) and against the ∼0.0065 typical of arms B and C.

Two rows are not resolved at the seed count available, and both are written in the main text as directions rather than results.

E PREREGISTRATION AND ERROR AUDIT

The analysis plan `12` and nineteen amendments, `12a`–`12s`, were each committed separately before the analysis they governed. Endpoints moved from `val` to `test` before any S2 result was examined. The plan specified interpretations for adverse outcomes: failure of the metric or lineage hypothesis, null downstream results, a dimensional explanation for the prototype bag's advantage, nominal (near-zero) prediction capability, discordant replication, or mixed depth results across arms. Four such outcomes occurred, and each was interpreted as registered. P1, the primary hypothesis, failed and is reported first. The metric-aware secondary endpoint $\cos(\mu_{\text{top1}}, \mu_{\text{true}})$ could have been designated primary, since the paper's argument predicts that cross-entropy is unsuitable for a target

Table 11: Seed-variability audit. Seeds are runs per side; effect is in the comparison's own units; ×E and ×typ express it as a multiple of arm E's seed sd (0.0263) and of the ∼0.0065 typical of arms B and C.

comparison	seeds	effect	×E	×typ	reading
Δ(R−B), permuted geometry	1 v 5	+2.5517	97	393	supported — larger than any seed spread we measured, on an arm we did not replicate
Δ(B−A) 85M, CE gap	5 v 5	+0.2249	8.6	35	supported — seed is a variance component
Δ(A′−A), factorised	1 v 5	+0.1056	4.0	16	probably supported; this qualification is explicit
Δ(B−A″), matched trainable params	5 v 3	−1.2233	47	188	supported in magnitude; A″'s spread is 6× arm B's, so `12r` states it as a direction
Δ(C−B) 85M, lineage	3 v 5	+0.0125	0.5	1.9	supported — 3 seeds, sd 0.0071, seed-aware CI
Δ(E−B), soft target	3 v 5	−0.0074	0.3	1.1	not resolved, reported as a direction
shuffled − as-is	2 v 2	−0.0061	0.2	0.9	resolved — two pairs 0.0006 apart
302M − 85M	2 pairs	−0.0333	1.3	5.1	acceptable — the pairs land 0.0004 apart
302M d5 reversal	2 pairs	−0.0290	1.1	4.5	replicated; two seeds is the whole basis
Δ(C−B) 302M − 85M	1 v 1	+0.0074	0.3	1.1	not resolved — directional only

designed around metric structure. It was not substituted: the plan specified that "no leg is rescued by switching endpoints" before results were available.

Integrity checks comprise a split fingerprint over five sorted lists at every training start, 30 corpus-integrity checks, 28 leakage checks, a per-shard checksum manifest over the 120 GiB embedding store, and a `config_core_sha256` verifying that every run of a matched stage shares one configuration. The configuration check was added because the run overlays lived in `/tmp`, were referenced by absolute path, and did not survive a node migration: a resume would have trained at the wrong learning rate and temperature while all existing fingerprints would still have matched.

E.1 Eleven defects and their reporting status

Eight of the eleven defects in Table 12 produced no error, warning or implausible number. Each was identified by reconciling derived totals with source artifacts rather than by a test. Three affected results reported in an internal draft; all are documented here to make the verification record explicit.

Defect 9: omission of emptied contigs. The out-of-vocabulary set is defined by observation: align each surviving corpus contig to its source contig and count the genes the corpus dropped. A source contig whose every gene fell below the frequency threshold leaves no corpus tokens, so nothing aligns to it, and its genes were in the denominator but in neither the out-of-vocabulary count nor the evaluation — 599,051 genes on 137,173 contigs, 4.78 points of held-out gene mass. The reported share was 16.25%; the complete accounting gives 21.25%, rising monotonically with distance where the undercount had not. The defect was identified when reconciling an unexplained 5.8-point gap between "survives the filter" and "reaches the evaluation": checking the pretokenised evaluation split against the recount showed that no gene is lost between the corpus text and the evaluation, so the gap had to be out-of-vocabulary genes the recount never saw. With the corrected weights §5.2 moves by at most 0.006 in any cell and no sign or reading changes — the newly counted genes sit on contigs with no retained genes, so they alter the weights and can supply no evaluation position — and the superseded weights are kept in the release and reproduce the earlier table exactly. An independent computation sharing no alignment code confirmed the share: source genes minus corpus tokens, giving 2,634,603 of 12,399,979 = 21.25%, monotone across bands, with the corpus-token side reproducing the evaluation split exactly. That subtraction is what §3 reports; the by-cause decomposition sums to 1,039 genes more, which are the 186 contigs the alignment never placed.

Table 12: The eleven defects identified by cross-checks: the defect, its observable symptom, and whether it was identified before or after a result was reported.

#	defect	symptom	identified before or after reporting
1	`searchsorted` off-by-one left one store row unembedded	none	before
2	`row_index` base offset skipped an entire accessory parquet — 4,239,621 representatives	none	before
3	class-bias weights computed over withheld genomes	none	before
4	`torch.clamp` gradient ratchet froze τ permanently after one overshoot	a run logging 0.00% movement	before — but only after a wrong interpretation had been drawn from it (§E.4)
5	a genome-level probe bypassed the `StandardScaler` the gene-level probes applied	none	before
6	four training ranks read-modify-wrote one JSON integrity file at startup	none until file truncation terminated a 4B launch at step 0	before
7	the paired-comparison cache was keyed on the run directory basename	none	after — see §E.3
8	a single-seed interval read as a training-recipe result	none	after — see §E.2
9	the out-of-vocabulary recount omitted genes on contigs the filter had emptied	none — 16.25% was plausible	after — in the abstract, §3 and §5.2's weights; see below
10	the leakage gate's default pointed at a superseded split and died on a missing file; at the live split, three artifacts carried a stamp from an earlier definition of the fingerprint	a crash, unnoticed because the gate was not being run	before — no artifact's inputs had changed
11	the anonymised release shipped a tracked symlink whose target named the account; the scrub greps file contents and a symlink has none	none — the scrub reported clean	before — nothing was released

Defect 11. `data` is a symlink to the data root, tracked in git, so `git archive` reproduced it in every export; its target is an absolute path carrying the account name. The scrub could not see it, because it reads file contents and a symlink has none. The builder now replaces every link pointing outside the tree with a note, and the leak check reads symlink targets as well as file contents. The defect was identified by inspecting the release contents during packaging.

Defect 10. The split fingerprint was widened twice — to cover genome-level lists, then the val/test assignment — and the codebook, prior table and persistence table still carry the first definition's stamp. All three definitions reproduce exactly from the current split, and the build inputs are identical: the gate now matches a build-time artifact against every historical definition and reports which one it matched, without waiving the check.

E.2 Sampling uncertainty and training-seed variability

With one seed per arm, we measured Δ(E−B) = +0.0107, with bootstrap interval [+0.0076, +0.0138] over 1,088 species and 2.9M paired positions. The interval excluded zero and indicated higher cross-entropy for arm E; an internal draft of §6 used this result. Two further seeds gave Δ(E−B) = −0.0074, [−0.0094, −0.0052], indicating lower cross-entropy for arm E, again with an interval excluding zero. The two analyses therefore give opposite signs despite their narrow intervals.

Neither bootstrap is incorrect. Both resample species within the fixed held-out set, so the interval measures variability under a different draw of held-out species, rather than under a different initialisation. Only the latter supports a conclusion about the training recipe. The difference reflects seed variance: arm E's is 0.0263, compared with arm B's 0.0062 and arm C's 0.0071. The seed-level standard error of the difference is 0.0154, twice the effect.

We therefore report the seed count and seed-level spread alongside conclusions based on single-seed comparisons. This applies to the 302M reversal in §4.2, the lineage head at scale in §7, and every comparison in Appendix D. The initial one-seed interval appeared precise but did not capture variability across initialisations.

E.3 Cache-key collision in the ordering comparison

The paired comparison caches per-position NLL to avoid repeated scoring. The key was the run directory's basename, and `runs/s2/A_s1` and `runs/s2sh/A_s1` are different runs — the same arm and seed on the as-is and the shuffled corpus — with the same basename. The shuffled comparison silently read the as-is corpus's cached tensors, and an internal draft reported a "shuffled" Δ(B$-$A) of +0.2634, which is `B_s1 - A_s1` on the as-is corpus to four decimal places. The erroneous value was used to support an ordering-invariance conclusion, with an incorrect direction of change and an accompanying mechanistic explanation. The correct value is +0.2188, which preserves the conclusion despite the invalid original check.

The collision was detected in the species-level table: the supposedly shuffled arm A scored 7.0590, the value already reported for the original corpus. Summary statistics had not revealed the mismatch, but this repeated value identified it. Keys are now stage-qualified, every entry records its absolute source run directory, and the loader raises an error on mismatch, so cache keys identify their contents.

E.4 Temperature-clamping defect and re-evaluation

τ was clamped to [0.01, 0.2] by `torch.clamp` in the forward pass, which zeroes the gradient strictly outside the range: one step past a bound froze the parameter for the rest of the run. A margin diagnostic showed the clamp left the logit margin invariant up to a constant factor, which was initially interpreted as showing no performance effect from clamping. That conclusion was withdrawn: the registered criterion, run afterwards on unclamped arms at the full budget, showed 0.0267 nats. The diagnostic was correct, but it did not establish the performance effect attributed to it. The arm comparisons in §4 and §7 were rebuilt on the nine unclamped runs. We make no novelty claim for the temperature finding — scale and temperature in normalised-embedding classifiers are studied in (Wang et al., 2017; Zhang et al., 2018) — and it is reported here rather than as a contribution.

E.5 Run-queue completion reporting

The runner queueing arm E's extra seeds waited for the previous run's `run_summary.json` to appear. `rank 0` writes that file before the process group tears down, so both seeds launched into GPUs still holding 49–56 GiB, failed with CUDA OOM within two minutes, and the runner printed `COMPLETE`. The failure left eighteen GPU-hours idle before inspection. The runner now waits for available device memory rather than a file and retries once, addressing the incorrect success report.

F Downstream baselines, probe protocols and reproducibility

Concatenation also exceeds the floor. When concatenated with ESM-C, arm B reaches 0.5419 on essential genes against the 0.5045 floor, with C and D at 0.5410 and 0.5352 — but $R^2 - R^2_{\text{null}}$, the same concatenation with gene order permuted at matched dimension, is within 0.008 of zero for every arm on both gene-level tasks, and never above +0.0015 (+0.0007 to -0.0073 on essential genes; +0.0015 to -0.0038 on operon) and unresolvable on phenotype. That bounds what is attributable to ordering and nothing more: permuting order preserves gene content.

Two additional baselines constrain interpretation of the downstream results. A per-prototype constant $\pi_{c,r}$ — three numbers per protein family, no genome context — reaches AUPRC 0.2190 on

essential genes against a 0.1119 base rate: 43% of the 0.5045 an ESM-C mean-pool probe reaches, or 27% of that probe's margin above the base rate, and all it encodes is whether a family is present across bacteria generally. The 50,000-dimensional TF-IDF prototype bag beats ESM-C mean-pool on phenotype (0.5095 against 0.4910), but the same bag randomly projected to 960 dimensions scores 0.4880, below ESM-C — so the supported conclusion is limited to a high-dimensional sparse gene-identity representation outperforming low-dimensional dense representations under a linear probe.

$R^2 - R^2_{\text{null}}$ on phenotype ranges -0.0012 to +0.0036 for our arms against a variation across the grid of 0.0178: not resolvable, by a rule registered before it was computed.

We probed every block and the final layer on both gene-level tasks. The decline is not monotone in any arm: on essential genes each has two to four single-step rises, the largest 0.0171 (arm C, blocks 1–2), and Bacformer's sampled profile has two of its own. Spearman ρ against block index, essential genes: -0.901 (A), -0.912 (B), -0.951 (C), -0.687 (D), -0.951 (Bacformer). Operon: -0.813, -0.841, -0.951, -0.703 and -0.762 (Bacformer, sampled every third block, whose final layer reproduces §8's 0.3868 exactly), with three to six rises per arm. Operon final-layer values are 0.3913, 0.3828, 0.3869, 0.4106 and the best layers 0.4243 (A, block 1), 0.4019 (B, block 1), 0.4214 (C, block 0), 0.4246 (D, block 3) — so seed 1 of arm B is below the 0.4129 floor at every depth. These profiles are seed 1; §8's block-0 and final-layer values are means over every seed (five for A and B, three for C, E and F), from `scripts/probe_depth.py`, which reproduces each seed-1 value here to within 3×10^{-8}.

Node equivalence. S2 spans two compute nodes, 15 runs on one and 1 on the other. Step-0 cross-entropy is bit-identical, the batch plan hash is identical, and cross-node divergence is below $3\times$ the intra-node band.

Code availability. The anonymised release described below accompanies this submission as supplementary material, and will be released publicly on acceptance. It contains the full pipeline, the pre-registration amendments, the number registry and the scripts that produce every table and figure in this paper.

Reproduction. Every number in the paper is generated into a registry, `paper/numbers.md`: 1,130 numbers, each with its source artifact and the script that produced it, none transcribed by hand. The anonymised release is built from `git archive HEAD` into a tree with no `.git`, with identifying paths rewritten and the scrub asserted. Running the number registry inside that release reproduces 1,053 of 1,130 registered numbers exactly, with zero mismatches; 71 require run directories or the 120 GiB embedding store, both excluded from the code release, and 6 are machine facts (host, driver, commit) expected to differ. The keys of all 77 are listed to identify which values can be checked from the release alone.

G Extended related work

This section expands the discussion in §2.

Genome-level language models over protein tokens. The architecture this paper studies — a transformer over a genome's proteins, each supplied as a protein-language-model embedding rather than as residues — belongs to a small, established family. gLM (Hwang et al., 2024) trains $\sim$1B parameters on seven million metagenomic contig fragments of 15–30 genes, with ESM-2 (Lin et al., 2023) embeddings as input. Bacformer (Wiatrak et al., 2025) scales the same idea to $\sim$1.3M genomes and $\sim$3B proteins at 26M parameters (ESM-2 35M input) and 300M (ESM-C input), ordering each genome's proteins by genomic coordinate. Our backbone follows this architectural family — 12 layers, 768 hidden, 85M parameters, flash-attention varlen — because the contribution is not the encoder. Our comparison differs from both studies in that we hold the encoder fixed across nine arms and vary only the prediction target, which is the comparison neither of those papers is designed to make. Nucleotide-level models (Dalla-Torre et al., 2025; Zhou et al., 2024; Nguyen et al., 2024) and the mixed-modality gLM2 (Cornman et al., 2024) use different modalities and are not direct comparators; our label space is a quantised encoder (van den Oord et al., 2017).

Prediction targets in existing models. gLM's masked objective is continuous: a linear head emits four 100-dimensional vectors with associated likelihoods, and the loss is an MSE against the closest

prediction plus a cross-entropy over the likelihoods, with the label a PCA-whitened ESM-2 embedding. This belongs to the same target family as arm D rather than arm A. Bacformer's masked variants likewise operate on protein embeddings, while its causal variant takes and predicts a protein-family token — a closed vocabulary in exactly the sense this paper uses. Existing systems therefore use different discrete or continuous targets, but also differ in corpus, protein encoder, parameter count and evaluation. Their reported results do not isolate the performance effect of the target choice; §4–§6 provide that comparison.

Prediction targets outside genomics. Masked image modelling made the discrete/continuous target an explicit design question: BEiT (Bao et al., 2022) predicts tokens from a learned visual vocabulary, MAE (He et al., 2022) regresses pixels, and data2vec (Baevski et al., 2022) regresses latent representations of the input. The finding there, as here, is that the target changes which tasks the representation supports rather than simply how accurate it is. We adopt this framing without adopting those methods; our §5.2 observation — that no evaluation metric is neutral between a categorical and a continuous target — makes explicit a difficulty that this literature addresses implicitly through downstream-transfer evaluation.

Tying a classifier to an encoder's geometry. Arm B uses an established metric classifier over frozen centroids. Label-embedding classifiers (Frome et al., 2013) replace a free output matrix with a similarity against fixed semantic vectors; prototypical networks (Snell et al., 2017) classify by distance to class means and support few-shot classification; tied input–output embeddings (Press & Wolf, 2017) share one table between input and output; and the image–text contrastive family (Radford et al., 2021) uses an encoder-defined output space, allowing classification over categories named at test time. Open-vocabulary recognition in vision rests on exactly that substitution. We claim no novelty in the head. Our contribution is a controlled measurement of the cross-entropy difference, in nats, relative to a free classifier on identical data and budget, its dependence on distribution shift, and the predictions enabled by vocabulary extension. These controlled comparisons are not reported in the cited studies.

Protein families as a closed set. The out-of-vocabulary problem is long familiar in protein annotation: a large share of environmental protein space matches no characterised family, and family classifiers must decide what to do with it. ProtCNN (Bileschi et al., 2022) is the closest reference point, extending Pfam coverage with a learned classifier and confronting the same question of what a fixed label set excludes. Our §3 is the pangenome-model version of that count, with a specific distinction: it is measured against the vocabulary a model was actually trained with, not against a reference database, and by observation of what the corpus contains rather than by reimplementing the filter that built it.

Pangenome construction. The corpus we train on inherits its gene clusters and its core/accessory split from standard pangenome tooling. The frequency filter $t = \min(5, n)$ that produces §3's out-of-vocabulary set is ours: a depth-aware generalisation of the fixed threshold in the corpus construction we started from (Appendix A). Core-genome estimation from incomplete genomes is itself an active problem (Buck et al., 2022) — corpora of this shape are built with pangenome clustering pipelines of the kind Roary and Panaroo provide (Page et al., 2015; Tonkin-Hill et al., 2020) — and the taxonomy our held-out bands are defined over is GTDB (Parks et al., 2022). We take the clusters and the taxonomy as given; the filter is a modelling choice, and §3's out-of-vocabulary share is measured under it. The paper's claims are about what a model does with such a corpus, not about how the corpus should be built.

Benchmarks. We report Bacformer as a reference row on BacBench's tasks (Wiatrak et al., 2026). BacBench is unpublished, so no claim in this paper rests on a number it reports; §8 and Appendix F state what its tasks can and cannot demonstrate under a frozen-probe protocol, including a floor that no final-layer representation in our comparison exceeds and that block-0 representations of our arms exceed on essential genes in 19 of 20 checkpoints (§8).