Machinery Space is the Right Coordinate for Biological Coverage: Diagnosing the Compositional Ceiling in Genomic Foundation Models

Miyu Horiuchi

Replicater Labs

Abstract

Foundation models of protein and genome sequence plateau on function and phenotype prediction: added data yields vanishing returns and simple linear or evolutionary baselines match billion-parameter encoders. A natural diagnosis, advanced by recent work on this exact setting, is that generalization is coverage-limited: held-out lineages fall where no labelled training neighbour exists. We agree coverage binds, but show the field has been measuring coverage in the wrong coordinate system. We ask, for held-out bacterial families, which notion of "closeness to the training set" actually predicts whether a model generalizes: closeness in the foundation model's own embedding space, in taxonomy, or in machinery space, the space of metabolic pathway content (KEGG modules). On 15,002 genomes we find a sharp trait-class×coordinate interaction. For machinery-localized traits (pathogenicity, biosafety level), machinery-space proximity predicts per-genome generalization at AUROC ≈ 0.652, whereas embedding-space (≈ 0.531) and taxonomy (≈ 0.548) proximity are near chance; for compositional traits no coordinate is strongly predictive. A pre-registered control rules out that machinery space is merely a taxonomy proxy: machinery distance is nearly orthogonal to taxonomy (η^2=0.042) and taxonomy coverage is itself near chance. The coordinate is not an artifact of a hand-built ontology: a learned genome-context foundation model (Bacformer) recovers it (coverage-AUROC 0.68 on machinery traits) while mean-pooled ESM-2 is near chance (0.52), localizing the failure to gene-content-destroying pooling rather than to foundation models as such. We further show, via multi-seed learning curves, that (i) on compositional traits every diversity operator saturates at the same plateau, reproducing the published single-cell diversity null in genomes, and (ii) maximizing taxonomic diversity by dereplication specifically harms machinery traits (-0.033 AUROC) while leaving compositional traits untouched ($+0.004$). The compositional ceiling is thus a statement about which axis one samples. The corrective is not more taxa but coverage measured in machinery space.

1 Introduction

Neural scaling laws promise that loss falls as a power law in data and model size [Kaplan et al., 2020, Hoffmann et al., 2022], and biological foundation models, protein language models such as ESM-2 [Lin et al., 2023] and genome models such as Evo 2 [Brixi et al., 2025] and the Nucleotide Transformer [Dalla-Torre et al., 2024], were expected to inherit it. On function and phenotype prediction they largely do not: added pre-training data gives diminishing or absent returns [Cheng et al., 2025, Serrano et al., 2025], simple evolutionary or linear baselines match far larger models [Laine et al., 2019, Frazer et al., 2021, Notin et al., 2023], and single-cell foundation models neither beat simple baselines [Kędzierska et al., 2023] nor improve when their pre-training data is made more diverse [DenAdel et al., 2026]. We call this saturation a compositional ceiling.

A prominent diagnosis is that the bottleneck is coverage in the model's representation space: held-out lineages land in regions sparsely populated by labelled training examples, so there is no reliable neighbour to transfer from. This diagnosis is appealing but, as stated, non-actionable: it does not say in what space coverage should be measured or maximized. The diversity operators used to attack the problem quietly assume an answer. Genome dereplication [Olm et al., 2017] and taxonomy-aware cross-validation [Roberts

et al., 2017] work in taxonomy space; the single-cell diversity study of DenAdel et al. [2026] works in transcriptomic space via geometric sketching [Hie et al., 2019]. All are compositional coordinates.

We argue the correct coordinate is machinery space: the presence and completeness of metabolic pathway modules [Kanehisa and Goto, 2000]. The intuition is that phylogenetic redundancy, the same machinery re-sampled across related lineages, is low-rank and linearly recoverable, so it does not stress a nonlinear model, whereas the residual signal that governs hard phenotypes is organized by mechanism, not by taxonomy. Our contributions:

1. We turn the "coverage-limited" diagnosis into a testable, coordinate-system question and answer it: for machinery-localized traits, generalization to held-out families is predicted by proximity in machinery space but not in embedding space or taxonomy (Section 5.3). This is the paper's main result.
2. We pre-register and pass a collinearity gate showing machinery- and taxonomy-space are distinct operators, and a taxonomy-coverage control showing the effect is not taxonomy in disguise (Sections 5.2, 5.3).
3. Via multi-seed learning curves we reproduce the compositional ceiling in genomes (operator-invariant saturation on compositional traits) and show that maximizing taxonomic diversity specifically harms machinery traits (Section 5.4).
4. We report honestly that a strong training-recipe effect, machinery-space dereplication dramatically out-climbing other operators, does not survive multi-seed evaluation: re-sampling data is not the lever (Section 5.4).
5. We show the machinery coordinate's robust value is as a coverage/confidence signal, not as readout features or a sampling recipe: it makes a strong out-of-distribution selective-prediction signal (Section 5.7), while swapping input features gives only a weak, non-significant trend (Section 5.6). We report both the trend and its collapse under more data.

2 Related Work

Scaling laws and redundancy. Power-law scaling [Kaplan et al., 2020, Hoffmann et al., 2022] presumes each example reduces irreducible entropy along the task axis; redundancy breaks this. Deduplication improves LM scaling [Lee et al., 2022], principled data pruning beats power-law scaling [Sorscher et al., 2022], and the exponent is formally governed by redundancy in the data covariance [Bi, 2025]. These prune in generic embedding/loss geometry; we ask what geometry is right for biology.

Biological foundation models. Scale helps structure and perplexity [Lin et al., 2023, Brixi et al., 2025] but stalls on function/variant effect [Cheng et al., 2025, Serrano et al., 2025], where likelihood conflates fitness with phylogeny [Yin et al., 2025] and alignment-based methods stay competitive [Laine et al., 2019, Frazer et al., 2021, Hopf et al., 2017, Notin et al., 2023].

Diversity as a controlled operator. DenAdel et al. [2026] manipulate single-cell pre-training diversity (random, cell-type reweighting, geometric sketching) and find none restores scaling. This is the closest experiment to ours and, we argue, the control condition: those axes are compositional. Dereplication [Olm et al., 2017] and k-center coresets [Sener and Savarese, 2018, Sorscher et al., 2022] supply operator machinery; taxonomy-aware evaluation [Roberts et al., 2017, Kapoor and Narayanan, 2023] supplies the split. What is missing, and what we test, is coverage in machinery space.

3 Setup and Hypothesis

Each genome g has a frozen foundation-model representation $\phi(g) \in \mathbb{R}^{640}$ (mean-pooled ESM-2), a machinery vector $\mu(g) \in [0,1]^{570}$ (KEGG module completeness), a taxonomy, and binary phenotype labels. We evaluate under held-out-family splits: no family in test appears in training, the "microbial dark matter" regime. Following prior work we partition traits, before analysis, into compositional (signal expected diffuse; e.g. Gram stain, motility, oxygen tolerance) and machinery (signal expected gene/pathway-localized; e.g. pathogenicity, biosafety level), and choose machinery traits non-circular with the KEGG space (pathogenicity is virulence-driven, not a metabolic-module readout).

Table 1: Pre-registered collinearity gate. Machinery-space dereplication covers fewer families than taxonomy-space at every budget, ruling out taxonomic coverage as an explanation for any machinery effect. Machinery vs. taxonomy distance: Spearman $\rho = 0.20$, $\eta^2 = 0.042$; mean selection Jaccard 0.12.

budget k	random	taxonomy-derep	machinery-derep
100	64	100	75
250	94	250	129
500	126	407	194
1000	173	407	272
2000	219	407	355
4000	296	407	397

Hypothesis. Coverage binds generalization, and machinery space is the coordinate in which coverage is generalization-relevant. Concretely, for a held-out genome let $\mathrm{cov}_Z(g) = \max_{g' \in \mathrm{train}} \cos\big(z(g), z(g')\big)$ be its coverage in coordinate $Z \in \{\phi, \mathrm{taxonomy}, \mu\}$. We predict that for machinery traits cov_μ predicts whether g is classified correctly, while cov_ϕ and $\mathrm{cov}_{\mathrm{taxonomy}}$ do not; for compositional traits no coordinate is strongly predictive.

4 Method

Coverage-coordinate analysis (main). For each trait we train an ℓ_2-regularized logistic readout on standardized ESM-2 features of all labelled training genomes, score the labelled held-out-family test genomes, and record per-genome correctness. For each coordinate Z we compute cov_Z and report the AUROC of cov_Z as a predictor of correctness, with 95% bootstrap CIs. Taxonomy coverage uses the depth of the deepest shared rank (order/class/phylum) with the nearest training genome, since family and genus are held out by construction.

Dereplication operators (learning curves). Three operators each order the training pool; a curve at budget k trains on the first k: random (uniform), taxonomy (round-robin over families then genera, one per family first; the dRep/DenAdel analogue), and machinery (greedy k-center [Sener and Savarese, 2018] in KEGG cosine space). We report mean±95% CI over five seeds. The readout is an ℓ_2-regularized logistic regression, the "linear readout of aggregate features" central to the compositional-ceiling argument; a nonlinear-readout robustness replication is left to future work.

5 Experiments

5.1 Data

We use 15,002 bacterial genomes for which frozen ESM-2 embeddings, KEGG module completeness, BacDive phenotype labels [Schober et al., 2025], and a family split all exist, spanning 596 families (407 in the training pool; 20.3% of genomes in held-out test families).

5.2 Are the operators distinct? A pre-registered collinearity gate

Because KEGG content is partly vertically inherited, taxonomy- and machinery-space dereplication could coincide. They do not. Machinery (cosine) distance is only weakly related to taxonomic rank distance (Spearman ρ=0.20); a taxonomy bucket explains just η^2=0.042 of machinery-distance variance; the two orderings select largely different genomes (mean Jaccard 0.12). Crucially, machinery-space dereplication covers fewer families than taxonomy-space at every budget (Table 1), so no downstream machinery effect can be attributed to broader taxonomic coverage.

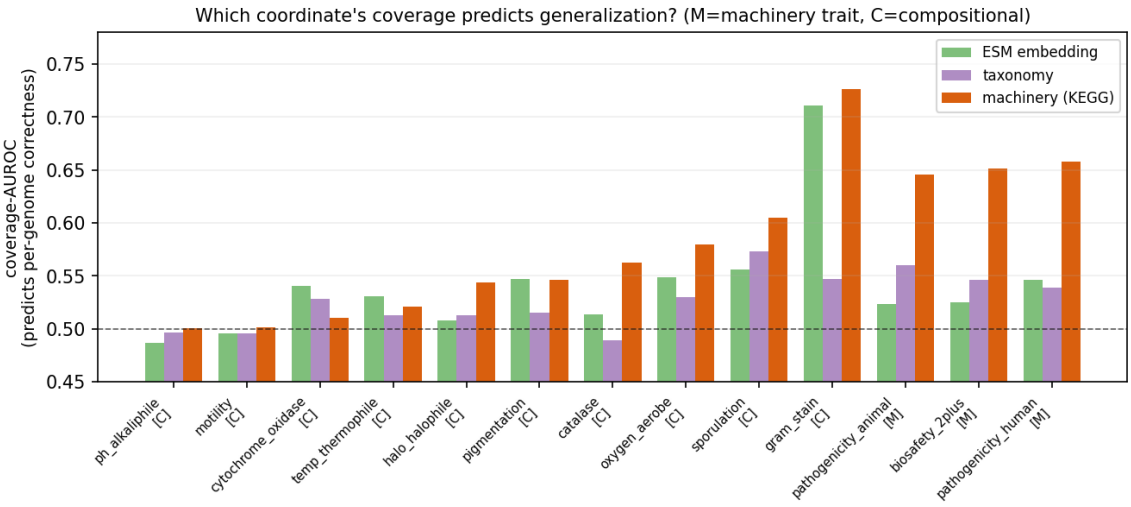

Figure 1: Coverage-AUROC by coordinate. For each held-out genome, coverage = max cosine similarity to any training genome, in ESM embedding space (green), taxonomy (purple), or KEGG machinery space (orange); bars show how well each predicts whether the readout classifies that genome correctly. Traits marked [M]=machinery, [C]=compositional. On the machinery traits (right) machinery-space coverage rises far above chance while embedding and taxonomy coverage do not.

5.3 Main result: machinery space is the coordinate of coverage

Table 2 and Figure 1 give the central finding. For machinery traits, machinery-space coverage predicts per-genome generalization at AUROC 0.652, while embedding-space (0.531) and taxonomy (0.548) coverage sit at chance; per-trait bootstrap CIs for machinery coverage exclude both 0.5 and the embedding/taxonomy values. For compositional traits the three coordinates are similar and only weakly predictive (0.560 vs. 0.544). The taxonomy-coverage control is decisive: were machinery space a taxonomy proxy, taxonomy coverage would carry the same signal; it does not. This localizes why the "coverage-limited" diagnosis resisted fixes: coverage was measured in embedding space, which is blind to machinery-trait generalization.

5.4 Learning curves: the compositional ceiling and a taxonomy penalty

Figure 2 and Table 3 show the sampling view. On compositional traits all three operators reach the same plateau at the same small N, the compositional ceiling, and a reproduction of the single-cell diversity null [DenAdel et al., 2026] in genomes. On machinery traits, maximizing taxonomic diversity hurts: taxonomy-space dereplication costs -0.033 AUROC relative to random selection, versus $+0.004$ on compositional traits. Machinery-space dereplication avoids this penalty and reaches the full-data ceiling. We note plainly that machinery-space dereplication does not dramatically out-climb random selection under multi-seed evaluation, an earlier few-seed signal did not survive, so we frame the operator result as a taxonomy penalty on mechanism rather than a machinery-diversity windfall; the robust, actionable claim is the coordinate diagnosis of Section 5.3.

5.5 A genome-context foundation model recovers the machinery coordinate

The main result uses mean-pooled ESM-2 as the embedding coordinate, and finds it near chance for machinery traits. Is that a property of foundation models in general, or of mean-pooling a protein language model, an operation that averages away gene content, the very thing machinery depends on? We repeat the coverage analysis with a fourth coordinate: Bacformer, a foundation model that represents proteins in genomic context (gene order and content), on the 9,062 genomes with valid Bacformer features (readout held fixed to ESM-2 logistic, so only the coverage coordinate changes). Table 4 and Figure 3 give the answer: for machinery traits, Bacformer-space proximity predicts generalization at 0.683, as well as or better than

Table 2: Coverage-coordinate analysis (main result). AUROC of a held-out genome's coverage (max cosine similarity to training) as a predictor of whether the linear readout classifies it correctly, measured in ESM embedding space, taxonomy, and KEGG machinery space. For machinery traits only machinery-space coverage predicts generalization; embedding and taxonomy coverage are near chance (0.5). Per-trait machinery 95% bootstrap CIs shown.

trait	class	ESM	taxonomy	machinery (95% CI)
ph_alkaliphile	comp	0.487	0.497	0.500 [0.464,0.536]
motility	comp	0.495	0.495	0.501 [0.472,0.533]
cytochrome_oxidase	comp	0.541	0.528	0.510 [0.469,0.552]
temp_thermophile	comp	0.531	0.513	0.521 [0.454,0.583]
halo_halophile	comp	0.508	0.513	0.543 [0.499,0.591]
pigmentation	comp	0.547	0.515	0.546 [0.502,0.588]
catalase	comp	0.513	0.489	0.563 [0.525,0.602]
oxygen_aerobe	comp	0.548	0.530	0.580 [0.551,0.608]
sporulation	comp	0.556	0.573	0.605 [0.547,0.660]
gram_stain	comp	0.711	0.547	0.726 [0.666,0.783]
pathogenicity_animal	mach	0.523	0.560	0.646 [0.604,0.688]
biosafety_2plus	mach	0.525	0.547	0.651 [0.615,0.686]
pathogenicity_human	mach	0.546	0.539	0.658 [0.620,0.697]
mean compositional		0.544	0.520	0.560
mean machinery		0.531	0.548	0.652

explicit KEGG pathways (0.633) and far above near-chance mean-pooled ESM-2 (0.523) and taxonomy (0.563). The machinery coordinate is thus not specific to a hand-built pathway ontology: a learned genome-context representation recovers it, provided it does not discard gene content by pooling. This localizes the earlier ESM-2 failure to the representation, not to foundation models as such, and predicts that larger genome language models (e.g. Evo 2) should pattern with Bacformer rather than with mean-pooled ESM-2, a scaling test we leave to future work (we had Evo 2 features for only 364 genomes, too few for held-out-family analysis).

5.6 Is the coordinate also a better readout? A weak, honest trend

The coverage analysis is diagnostic; can the coordinate also improve the model directly, as readout features? We test this: for each trait we compare a logistic readout on a gene-content representation against one on mean-pooled ESM-2, on matched genomes under held-out-family evaluation, and ask whether any gain is confined to machinery traits (a representation×trait-class interaction). Table 5 and Figure 4 give a deliberately unflattering three-part answer. (i) Bacformer, a genome-context foundation model, trends toward helping machinery traits more than compositional (+0.033 vs +0.012; interaction +0.021), consistent with the coverage result—but at n=9,062 genomes no per-trait gain is individually significant (pathogenicity human +0.036, animal +0.044, biosafety +0.020; all 95% CIs include 0). An earlier estimate on a smaller Bacformer subset (n=5,506) looked significant (+0.10); it did not survive more data, and we report the weaker, larger-sample result. (ii) Raw orthogroup content (eggNOG) improves readout broadly (+0.03 to +0.04 on both classes) with no interaction. (iii) KEGG module completeness is a poor readout (−0.053 on machinery traits) despite being a good coverage coordinate. The through-line is a dissociation: the machinery coordinate is a robust coverage/confidence signal (Sections 5.3, 5.7) but only a weak readout representation. Measuring coverage in machinery space is the reliable lever; swapping input features is not.

5.7 A practical payoff: machinery-coverage as an abstention signal

If machinery-space coverage predicts whether a held-out genome is classified correctly (Section 5.3), it should be usable as a selective-prediction signal: a deployed pipeline could abstain on low-coverage genomes and trust the rest. The sharp question is whether it beats the model's own predicted-probability margin, the standard confidence signal, which is known to be poorly calibrated out-of-distribution. Figure 5 and Table 6 show the risk-coverage behaviour. On machinery traits, ranking genomes by machinery-coverage and keeping

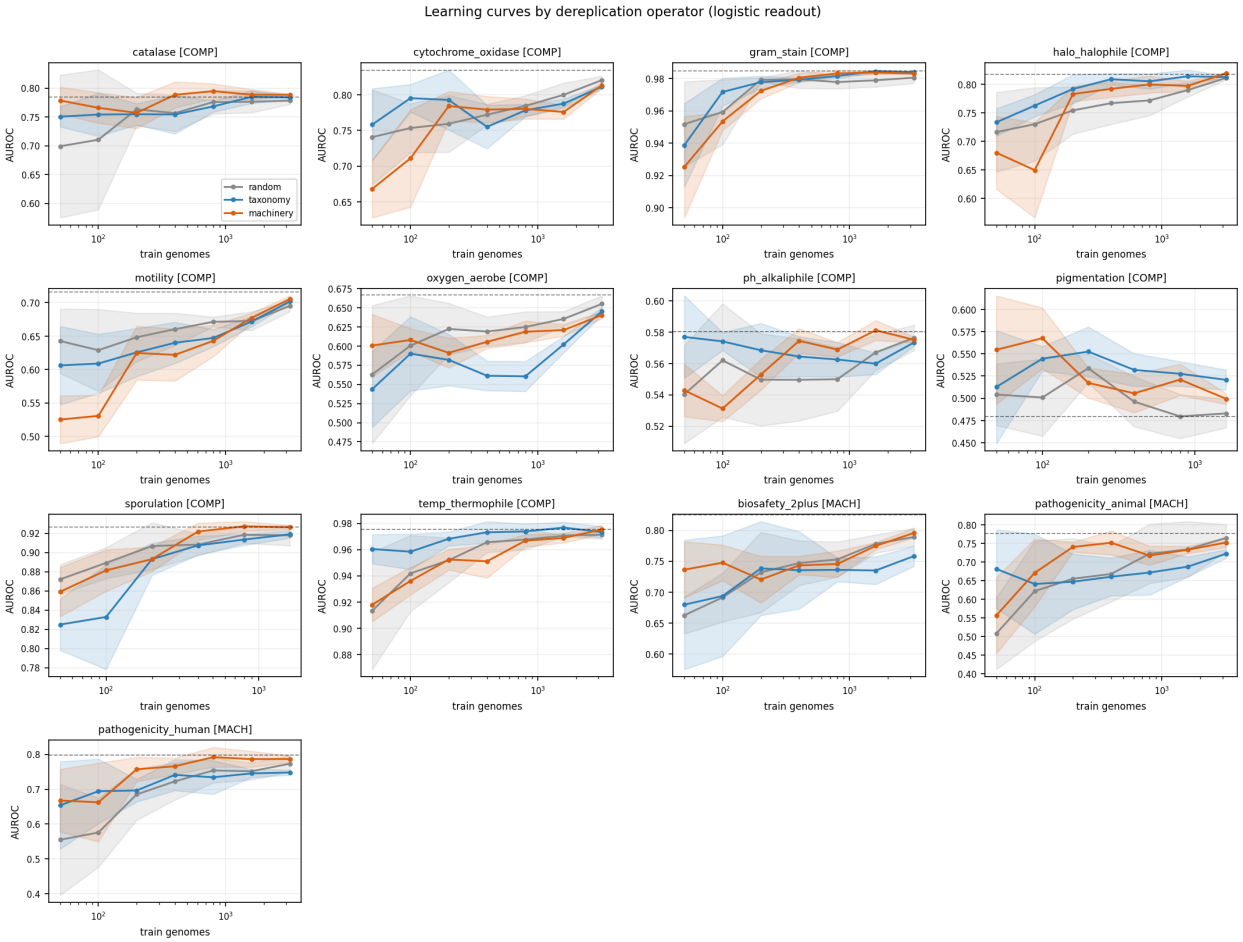

Figure 2: Learning curves by operator (linear readout, 5 seeds; band = ±std). Orange=machinery-derep, blue=taxonomy-derep, grey=random; dashed=full-data AUROC. Compositional [COMP] curves collapse together; on machinery [MACH] traits taxonomy-derep (blue) trails while machinery-derep (orange) reaches the ceiling.

the confident half raises balanced accuracy far more than ranking by the model's own probability margin (+0.054 vs +0.014); the two signals are complementary, and combining them (rank-averaging) is best of all (+0.079), winning on every machinery trait. On compositional traits, where the model is well-calibrated, its own confidence suffices and machinery-coverage adds little. The practical rule for mechanism-driven function prediction: trust predictions on genomes that are both confidently scored and machinery-covered. We note the machinery-trait effect is concentrated on pathogenicity; a broader machinery-trait battery would test its generality.

6 Discussion

Our results recast a widely reported failure. "Biology breaks foundation models" is a statement about which axis one samples. Compositional diversity, more taxa, cells, or metadata strata, cannot restore a scaling law, because the redundancy it fails to remove is mechanistic; and coverage, the correct diagnosis, is only actionable once measured in the right coordinate. Machinery space is that coordinate: it is where held-out generalization on hard, gene-localized phenotypes is predictable, and where taxonomy-driven diversity operators go wrong by discarding the within-clade replication those phenotypes rely on. This unifies three observations: absent data-scaling laws (measured along the wrong axis), the failure of single-cell diversity operators (compositional axes), and foundation models losing to linear baselines (low-rank compositional

Table 3: Learning curves at the largest shared budget (linear readout, 5 seeds). full: AUROC using all training genomes. tax. penalty: AUROC(taxonomy-derep)−AUROC(random): taxonomy-space dereplication specifically harms machinery traits (negative) while leaving compositional traits unchanged. On compositional traits all operators reach the same plateau (the compositional ceiling).

trait	class	full	random	taxon.	mach.	tax. penalty
oxygen_aerobe	comp	0.667	0.655	0.645	0.640	-0.010
cytochrome_oxidase	comp	0.834	0.820	0.811	0.813	-0.009
ph_alkaliphile	comp	0.580	0.576	0.573	0.575	-0.003
sporulation	comp	0.926	0.918	0.919	0.926	+0.001
halo_halophile	comp	0.818	0.812	0.813	0.820	+0.002
temp_thermophile	comp	0.975	0.971	0.973	0.975	+0.002
gram_stain	comp	0.985	0.980	0.984	0.983	+0.003
catalase	comp	0.784	0.778	0.784	0.789	+0.006
motility	comp	0.716	0.695	0.702	0.705	+0.007
pigmentation	comp	0.479	0.483	0.521	0.499	+0.038
pathogenicity_animal	mach	0.777	0.765	0.723	0.753	-0.043
biosafety_2plus	mach	0.825	0.789	0.758	0.796	-0.031
pathogenicity_human	mach	0.797	0.773	0.748	0.787	-0.025
mean compositional						+0.004
mean machinery						-0.033

Table 4: A genome-context FM recovers the machinery coordinate. Mean coverage-AUROC by coordinate and trait class, on the n=9,062 genomes with valid Bacformer features (readout fixed to ESM-2 logistic; only the coverage coordinate varies). For machinery traits, proximity in Bacformer space, a genome-context foundation model that sees gene content and order, predicts generalization (0.683) as well as or better than explicit KEGG pathways (0.633) and far above near-chance ESM-2 (0.523) and taxonomy (0.563). Compositional traits show no such coordinate.

trait class	ESM-2	taxonomy	Bacformer	KEGG machinery
compositional	0.541	0.522	0.545	0.568
machinery	0.523	0.563	0.683	0.633

signal a linear head already captures).

Toward mechanism-varying data. Our operators re-weight observational genomes, which can only remove redundancy, not add mechanism. The natural next step is data that varies mechanism while holding composition fixed, exactly what our coordinate analysis says is scarce. In microbiology this is perturbation data: transposon-insertion and knockout screens (Tn-seq), CRISPRi libraries, and growth-condition/media challenge panels, which measure the causal effect of disabling or stressing a specific piece of machinery. Such data decouples mechanism from phylogeny by construction, the perturbed and reference genomes are identical in composition and differ only in the targeted machinery, and it supplies the independent mechanistic instances that convergent evolution provides in the wild. It also connects directly to function-discovery pipelines: cultivation-medium and growth-phenotype prediction are, in effect, coarse perturbation responses, and pathway-completion tools (gapseq) are mechanistic simulators of them. The tradeoff is coverage: perturbation atlases exist for hundreds of organisms, versus tens of thousands of observational genomes. The design our results suggest is therefore a hybrid: broad observational genomes curated and weighted in machinery space, anchored by sparse perturbation data where mechanism is measured directly, and testing whether perturbation data breaks the compositional ceiling is the key experiment this diagnosis motivates.

Limitations. (i) The strongest intervention claim, that machinery-space dereplication as a training recipe substantially beats random selection, did not survive multi-seed evaluation; our robust result is diagnostic. (ii) Pathogenicity is partly clade-confounded, which is why taxonomy-dereplication can hurt it; we control the comparison with a fixed held-out-family test set and a taxonomy-coverage baseline, but a non-confounded machinery trait battery would strengthen the claim. (iii) Machinery space is instantiated by KEGG metabolic modules; eggNOG orthogroups and gapseq reactions may localize other traits and are natural robustness axes. (iv) Labels are from BacDive and carry annotation noise; we restrict to traits with adequate coverage

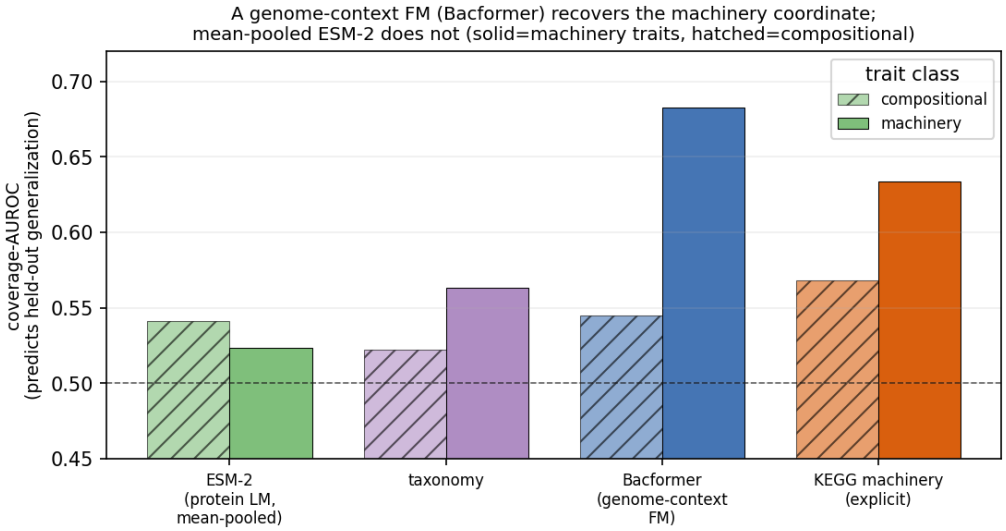

Figure 3: Coverage-AUROC by coordinate and trait class (Bacformer subset). Solid bars=machinery traits, hatched=compositional. For machinery traits, the genome-context FM (Bacformer, blue) and explicit KEGG machinery (orange) predict generalization; mean-pooled ESM-2 (green) and taxonomy (purple) do not.

Table 5: Representation as readout (H1): a weak, non-significant trend. AUROC gain of a gene-content readout over mean-pooled ESM-2 under held-out-family evaluation (n=9,062 for Bacformer). Interaction is machinery-minus-compositional mean gain. Bacformer trends toward helping machinery traits more (interaction +0.021) but no per-trait gain is individually significant at this sample size; eggNOG helps broadly; KEGG completeness is a poor readout. The machinery coordinate's robust value is as a coverage/confidence signal (Sec. 5.3, 5.7), not as readout features.

representation	Δ compositional	Δ machinery	interaction
Bacformer	+0.012	+0.033	+0.021
eggNOG	+0.039	+0.027	-0.012
KEGG_machinery	-0.001	-0.053	-0.052

and weight conclusions on rising, not flat, signals. (v) We manipulate the downstream training set on frozen features, not the encoder's pre-training corpus; whether the same coordinate governs pre-training is the key next test.

7 Conclusion

The compositional ceiling is a sampling artifact, and coverage, long fingered as the bottleneck, becomes actionable only in the right coordinate system. For machinery-localized phenotypes that coordinate is machinery space, not taxonomy or embedding geometry. Measuring biological coverage in machinery space is a concrete, deployable change to how these models are evaluated and their data curated.

References

Wei Bi. Scaling laws are redundancy laws. arXiv preprint arXiv:2509.20721, 2025.

Garyk Brixi, Matthew G Durrant, Jerome Ku, et al. Genome modeling and design across all domains of life with evo 2. bioRxiv, 2025. 2025.02.18.638918.

Cheng et al. Understanding protein language model scaling on mutation effect prediction. bioRxiv, 2025. 2025.04.25.650688.

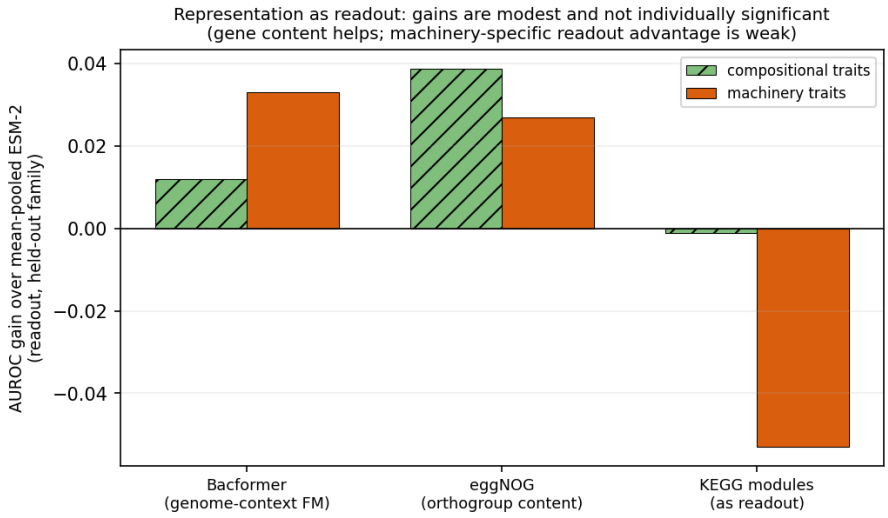

Figure 4: Representation as readout: AUROC gain over mean-pooled ESM-2 under held-out-family evaluation. Bacformer only weakly (non-significantly) trends toward helping machinery traits more (orange); eggNOG helps both classes; KEGG completeness is a coverage coordinate, not a readout.

Table 6: H2: machinery-coverage is a complementary out-of-distribution confidence signal. Mean gain in balanced accuracy from keeping the top-50% most-confident held-out genomes (retained minus all), by abstention signal and trait class. For machinery traits, combining the model's own probability margin with machinery-space coverage yields the largest gain (+0.079), beating the probability margin alone (+0.014); for compositional traits the model's own confidence suffices.

abstention signal	compositional	machinery
model probability margin	+0.068	+0.014
machinery coverage	-0.018	+0.054
combined (prob + machinery)	+0.039	+0.079

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 2024.

Alan DenAdel et al. Evaluating the role of pre-training dataset size and diversity on single-cell foundation model performance. Nature Methods, 2026. bioRxiv 2024.12.13.628448.

Jonathan Frazer, Pascal Notin, Mafalda Dias, et al. Disease variant prediction with deep generative models of evolutionary data. Nature, 599:91–95, 2021.

Brian Hie, Hyunghoon Cho, Benjamin DeMeo, Bryan Bryson, and Bonnie Berger. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. Cell Systems, 8(6):483–493, 2019.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, et al. Mutation effects predicted from sequence co-variation. Nature Biotechnology, 35:128–135, 2017.

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 28 (1):27–30, 2000.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

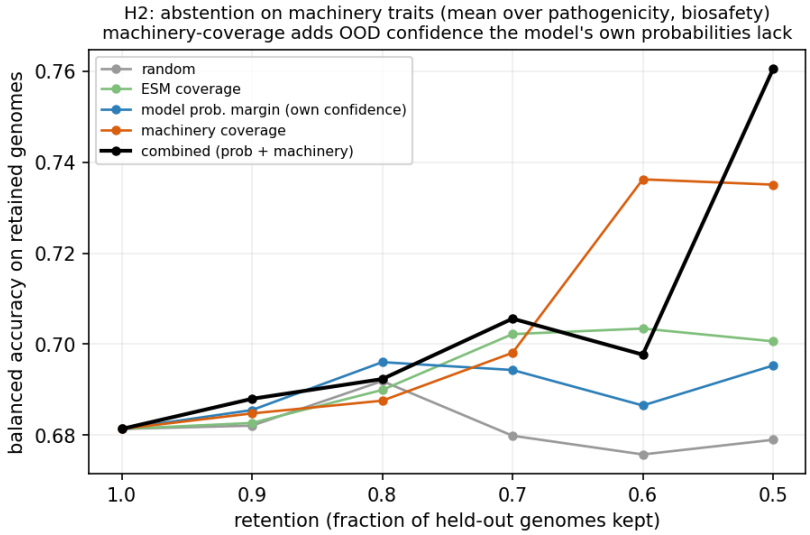

Figure 5: Selective prediction on machinery traits (mean over pathogenicity and biosafety). As the retained fraction shrinks (abstain on the least-confident), balanced accuracy rises fastest when genomes are ranked by machinery coverage or by the combined signal; the model's own probability margin (blue) is nearly flat, i.e. poorly calibrated out-of-distribution.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 2023.

Kasia Z Kędzierska, Lorin Crawford, Ava P Amini, and Alex X Lu. Assessing the limits of zero-shot foundation models in single-cell biology. bioRxiv, 2023. 2023.10.16.561085.

Elodie Laine, Yasaman Karami, and Alessandra Carbone. GEMME: A simple and fast global epistatic model predicting mutational effects. Molecular Biology and Evolution, 36(11):2604–2619, 2019.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, 2023.

Pascal Notin, Aaron W Kollasch, Daniel Ritter, et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks, 2023.

Matthew R Olm, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. The ISME Journal, 11:2864–2868, 2017.

David R Roberts, Volker Bahn, Simone Ciuti, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8):913–929, 2017.

Isabel Schober et al. BacDive in 2025: the bacterial diversity metadatabase. Nucleic Acids Research, 2025.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In International Conference on Learning Representations (ICLR), 2018.

Yara Serrano et al. Scaling and data saturation in protein language models. arXiv preprint arXiv:2507.22210, 2025.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Yin et al. From likelihood to fitness: Improving variant effect prediction by marginalizing over selective pressure. bioRxiv, 2025. 2025.05.20.655154.