# Latent Prediction Beats Scale
## A World Model for Held-Out Microbial Traits

Miyu Horiuchi

Replicater Labs

Abstract

Foundation models (FMs) for biological sequence are increasingly proposed for microbial phenotype prediction, yet recent benchmarks find they rarely beat simple baselines, and it is unclear what they add. Rather than ask "does the FM win," we ask what actually governs generalization, and whether a small group can build a competitive model by getting the objective and the data diversity right rather than by scaling. We report three results. (1) The objective works. A Joint-Embedding Predictive Architecture (JEPA) trained from scratch on the amino-acid sequences of just 5,675 bacterial genomes (no ESM, no Evo, no pretraining) matches ESM-2 [Lin et al., 2023], pretrained on ∼250M proteins, on held-out-family machinery-trait transfer (AUROC 0.748 vs 0.747); a matched-data ablation shows the latent-predictive objective is the driver. (2) The right coordinate is mechanistic, not phylogenetic. In a resampled matched-split analysis, proximity in machinery space (functional-ortholog content) predicts held-out generalization while phylogenetic distance carries no independent signal, and the effect grows with extrapolation distance. Grounding in functional-ortholog machinery (secretion/defense/virulence), not core metabolism, is what moves the hard traits. (3) More data is not the lever; the right data is. Scaling the JEPA from 5,675 to 14,949 genomes under a matched protocol decreases machinery-trait AUROC (0.748 → 0.654), because the added genomes contribute phylogenetic, not mechanistic, diversity. Competitive microbial world models come from a latent-predictive objective grounded in mechanistic diversity, not from scale.

## 1 Introduction

Neural scaling laws promise loss falling as a power law in data and model size [Kaplan et al., 2020, Hoffmann et al., 2022], and biological FMs such as ESM-2 [Lin et al., 2023] and Evo 2 [Brixi et al., 2025] were expected to inherit it. On function and phenotype prediction they largely do not: added pretraining data gives diminishing returns [Cheng et al., 2025, Serrano et al., 2025], simple baselines match far larger models [Ahlmann-Eltze et al., 2025, Laine et al., 2019], and single-cell FMs neither beat simple baselines [Kędzierska et al., 2023] nor improve when pretraining data is made more diverse [DenAdel et al., 2026]. Prior work in this program named this a compositional ceiling and argued it is a statement about which axis one samples [Kanehisa and Goto, 2000].

We ask whether that diagnosis points to a constructive path: a model whose inductive bias and training data match it. We adopt the JEPA framing [LeCun, 2022, Assran et al., 2023]: predict masked content in a learned latent space rather than reconstructing tokens, so the target encoder discards unpredictable low-level detail and concentrates capacity on functional structure [Baevski et al., 2022]. We build such a model for bacterial genomes from scratch and evaluate under held-out-taxon phenotype transfer. Our contributions:

1. A from-scratch genome-JEPA that, on 5,675 genomes with no pretrained input, matches ESM-2 on held-out-family machinery-trait transfer; a matched-data ablation isolates the latent-predictive objective as the driver and shows a genome-level supervision head is harmful (Section 4).
2. A resampled matched-split analysis establishing that mechanistic, not phylogenetic, diversity is the coordinate governing generalization, with the effect growing under deeper extrapolation (Section 3).
3. A clean scaling test showing that naive data scaling hurts (because it adds phylogenetic, not mechanistic, diversity), reframing the "more data" intuition (Section 5).

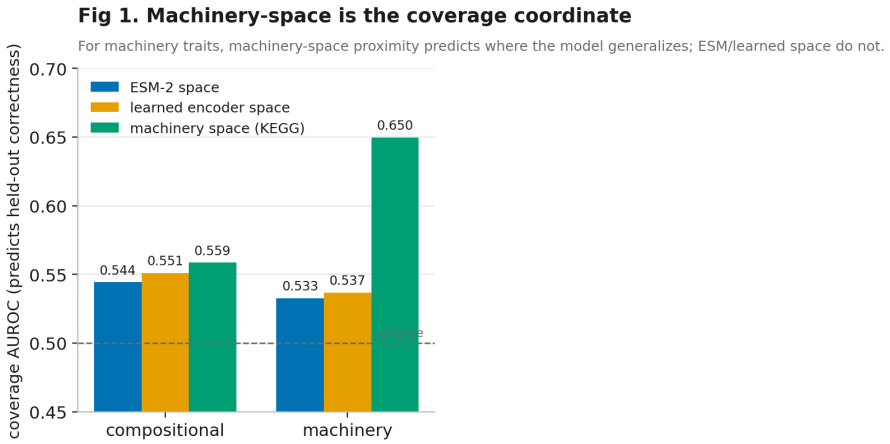


Figure 1: Machinery-space proximity predicts where the model generalizes; ESM/learned space are near chance. Coverage-AUROC by coordinate and trait class.

Table 1: Matched-split effect by axis (mean accuracy $\Delta$, machinery traits). Machinery distance is varied with phylogeny held fixed, and vice versa.

| Held-out level | Machinery effect (phylogeny matched) | Phylogeny effect (machinery matched) |
|---|---|---|
| species | +0.015 | −0.072 |
| genus | +0.086 | −0.042 |
| family | +0.149 | +0.036 |

## 2 Related Work

FMs versus simple baselines. Ahlmann-Eltze et al. [2025] show deep perturbation models fail to beat linear baselines; Kędzierska et al. [2023] show zero-shot single-cell FMs lose to classical pipelines. We do not merely reproduce the null; we identify the axis that governs generalization, which these do not isolate. Latent-predictive learning. JEPA [LeCun, 2022, Assran et al., 2023] and data2vec [Baevski et al., 2022] predict latent targets rather than reconstruct; ProteinJEPA [Anonymous, 2026] shows a latent loss helps out-of-distribution protein tasks but collapses without an anchor. Genome-context FMs. Bacformer [Wiatrak et al., 2025] models proteins-as-tokens in genome context with a generative objective; we hold the substrate fixed and change the objective to latent-predictive.

## 3 The coordinate that governs generalization is mechanistic

Before building a model, we ask which coordinate predicts where a trait model generalizes. For each held-out-family test genome we compute its proximity (max cosine similarity) to the training set in three spaces (ESM-2 embedding, a learned encoder, and KEGG machinery space [Kanehisa and Goto, 2000]) and measure how well each predicts per-genome correctness. Machinery space is the coverage coordinate (Figure 1): for machinery traits it predicts held-out correctness at AUROC 0.65, versus 0.53 for embedding or learned-encoder proximity.

We de-confound machinery from phylogeny by construction: we build matched test groups with an identical phylogenetic-distance distribution but differing machinery distance (and the symmetric control), and contrast accuracy across species/genus/family held-out splits (Table 1, Figure 2). Machinery-space distance drives generalization; phylogenetic distance carries no independent signal at any depth, and the machinery effect grows monotonically with extrapolation distance.

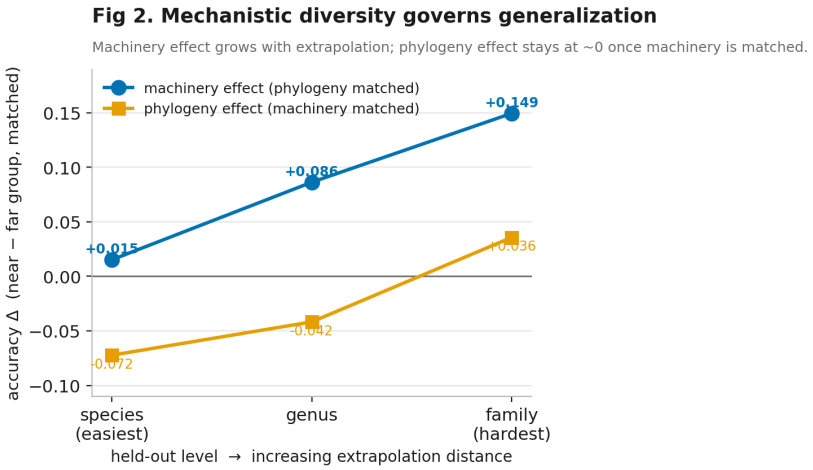


Figure 2: The machinery effect grows with extrapolation distance while the phylogeny effect stays near zero once machinery is matched.

Table 2: Matched-data, matched-compute ablation (held-out-family machinery AUROC). One component changed at a time; ESM-2 reference 0.747.

| Configuration | Machinery AUROC |
|---|---|
| Pure JEPA | **0.752** |
| Mean-pool stem (no context transformer) | 0.716 |
| Machinery-supervised head only (no JEPA) | 0.641 |
| JEPA + genome-level machinery head | 0.619 |

# 4 A from-scratch genome-JEPA matches ESM-2 at small scale

Architecture (Figure 3). A learnable residue stem (amino-acid embedding + 1-D convolutions) maps each protein to a gene embedding, replacing ESM entirely. A transformer context encoder over the gene sequence, with a masked contiguous gene span, predicts the masked genes' latents produced by an EMA target encoder (stop-gradient, LayerNorm). There is no tokenizer and no pretrained input.

Result. Trained on 5,675 genomes (3,005 held-out-family train) for 40 epochs on a single A100, the frozen genome embedding matches ESM-2 on machinery traits: pure JEPA 0.748 versus ESM-2 0.747 AUROC, parity from a model trained on ∼40,000× less data.

Ablation. Holding data and compute fixed and varying one component (Table 2, Figure 4), the latent-predictive objective is the driver. A genome-level machinery-prediction head (a plausible-sounding form of supervision) is actively harmful (−0.13), over-constraining the pooled representation.

# 5 The right data, not more data

Which machinery (Figure 5). Grounding a representation to predict functional-ortholog occupancy (eggNOG: secretion/defense/virulence orthologs) lifts machinery-trait AUROC from 0.807 (ESM) to 0.846, whereas grounding in core-metabolic modules (KEGG) leaves it flat (0.811). The hard traits live in specialized machinery, not core metabolism.

What kind of diversity. We scaled the from-scratch JEPA from 5,675 to 14,949 genomes under a strictly matched protocol (same A100, batch size, epochs, and model), varying only the data (Table 3, Figure 6). More data decreased performance (0.748 → 0.654). This is not an artifact: the 5,675 run reproduced Section 4 exactly, and only the data differed. The added ∼9,000 genomes contributed phylogenetic breadth, not mechanistic diversity, diluting a fixed-capacity representation. One cannot raise the machinery ceiling by

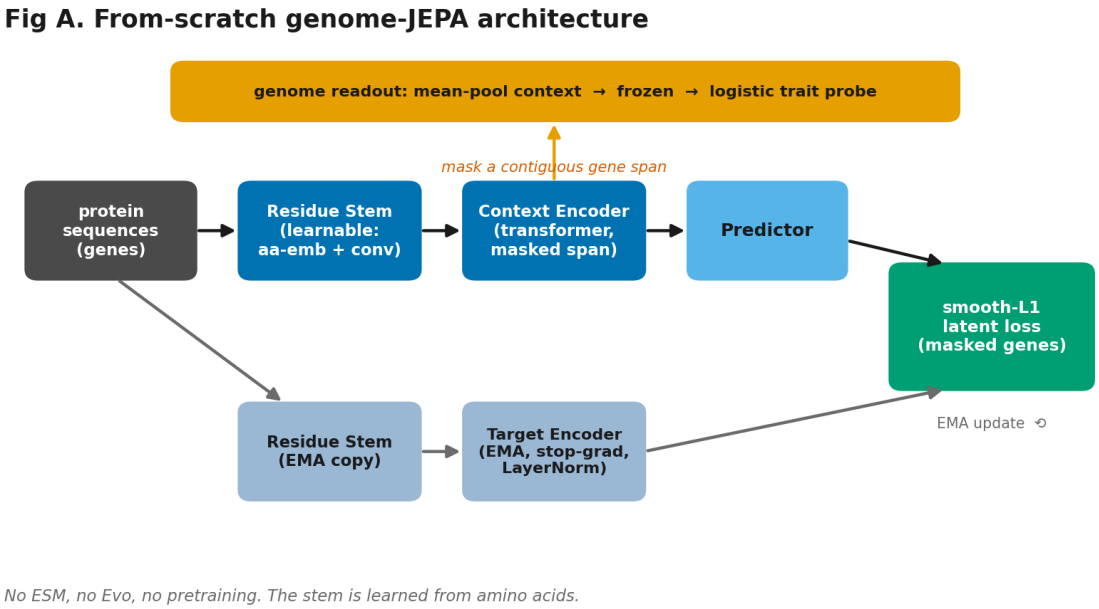


Figure 3: From-scratch genome-JEPA. A learnable residue stem feeds a context encoder whose masked-gene latents are predicted against an EMA target (smooth-L1); the pooled context is frozen and read out with a logistic trait probe. No ESM, no Evo, no pretraining.

Table 3: Clean scaling test (matched batch, epochs, GPU, model; only data varies). Held-out-family AUROC.

| Genomes | Machinery | Compositional |
|---|---|---|
| 5,675 | **0.748** | 0.751 |
| 14,949 | 0.654 | 0.720 |

adding phylogenetically-diverse genomes; the scaling experiment is a third, independent confirmation that mechanistic, not phylogenetic, diversity is the lever.

## 6 Discussion

Three independent lines (a coverage diagnostic, a from-scratch model, and a scaling test) point to one conclusion: a latent-predictive objective is sufficient to reach FM parity at tiny scale, and the binding constraint is the mechanistic diversity of the data, which neither scale nor naive supervision supplies. This reframes the "FMs don't beat baselines" critique constructively: the failure is one of sampling, and the remedy is mechanistic-diversity-aware curation with a self-supervised objective, not larger models. The natural next step is a mechanistic-diversity sampler that de-duplicates genomes in machinery space rather than sequence space, which our results predict is the way scale finally helps.

## Limitations

Machinery-trait evaluation uses $n$=3 traits and a single seed, so 0.748-vs-0.747 is parity, not a statistically clean win; the large effects (machinery head −0.13; scaling −0.09) are robust to this. "Scaling hurts" is demonstrated at fixed model capacity; a larger model on more data is untested, though the mechanistic-versus-phylogenetic diagnosis is capacity-independent. The mechanistic-diversity sampler that the discussion motivates is proposed, not yet built.

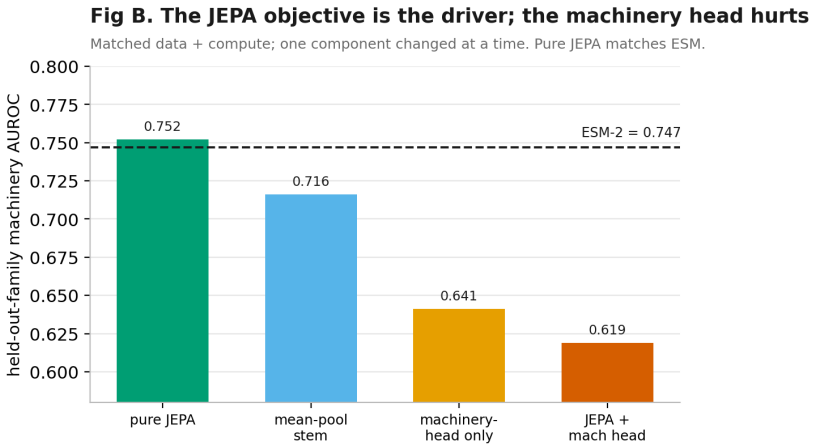


Figure 4: The self-supervised JEPA objective alone matches ESM-2; adding a genome-level machinery head hurts.

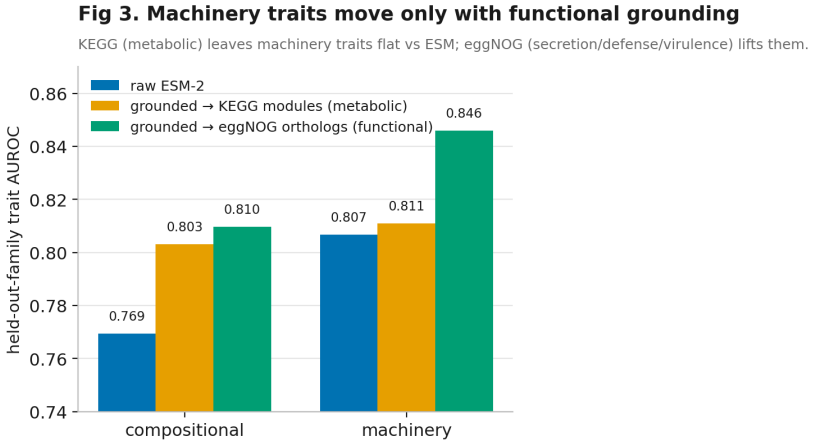


Figure 5: Machinery traits move only with functional-ortholog grounding, not core-metabolic KEGG.

# References

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods, 2025.

Anonymous. ProteinJEPA: latent-predictive pretraining improves out-of-distribution protein property prediction. arXiv preprint arXiv:2605.07554, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In CVPR, 2023.

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. ICML, 2022.

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.

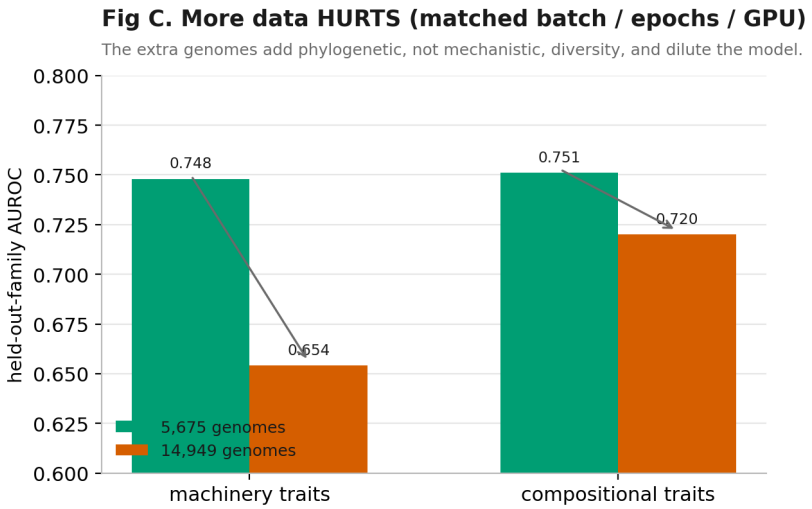


Figure 6: Under a matched protocol, more genomes decrease machinery-trait AUROC: the added diversity is phylogenetic, not mechanistic.


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.

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

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.

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.

Yann LeCun. A path towards autonomous machine intelligence. Open Review, 62(1), 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.

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

Maciej Wiatrak et al. Bacformer: a genome-context foundation model for bacteria. bioRxiv, 2025.