# The Taxonomy Ceiling: Why Protein-Language-Model Adaptation Fails to Generalize Across Bacterial Families

Miyu Horiuchi
m@replicater.xyz

2026-06-17

**Abstract** A bacterial genome is a *set* of proteins, and recent phenotype predictors aggregate frozen protein language-model embeddings into a genome representation. A natural next lever, once the pooling step is fixed, is to adapt the encoder itself: rather than treat ESM-2 as a frozen feature extractor, fine-tune it to the genomic task. We ask whether this lever survives the regime that matters for uncultured organisms: prediction on held-out taxonomic families. Holding data, splits, pooling (a Set Transformer), task heads, and training recipe fixed, we compare a frozen ESM-2-150M encoder against the same encoder adapted with Low-Rank Adaptation (LoRA), on a family-held-out split of 19,278 bacterial genomes and 21 traits. On the aggregate primary metric, LoRA *appears* to win by a small margin (0.580 vs 0.568). We show this gain is a measurement artifact: it is produced entirely by two imbalanced binary pathogenicity heads on which LoRA collapses to majority-class prediction, raising accuracy (0.70⟶0.94; 0.75⟶0.96) while driving F1 to *exactly zero*. With those two heads removed, encoder adaptation is net negative (−0.011), and on F1 it is neutral-to-harmful across the imbalanced traits. The conclusion mirrors and extends our earlier pooling study: under cross-family shift, neither the pooling operator nor encoder adaptation is the binding constraint. We document the artifact as a reusable warning, that accuracy is the wrong primary metric for imbalanced genomic traits, and offer this as a focused, single-seed pilot rather than a finished benchmark.

## Introduction

Foundation models for bacterial genomes increasingly follow a common template: a protein language model such as ESM-2 (Lin et al. 2023) embeds each open reading frame, and a genome-level model pools the resulting variable-size protein set into one vector for phenotype prediction (Wiatrak et al. 2025; MicroGenomer authors 2025; BacPT authors 2026). In prior work we isolated the **pooling** step and showed that adaptive (attention / Set-Transformer) pooling helps gene-localized "machinery" traits more than diffuse "compositional" traits, but that the entire advantage collapses under family-level distribution shift. That study deliberately *froze* the encoder so that the result could be read as a pooling effect. It left open the obvious follow-up question:

If the pooling operator is not the binding constraint under cross-family shift, is the **encoder**?

The standard hope is that fine-tuning the protein encoder, rather than using it frozen, lets the representation specialize to the genomic task and recover generalization that a fixed encoder cannot. This is the most expensive lever available short of retraining a foundation model, and it is the lever most practitioners reach for next. We test it in the cleanest possible way: hold the pooling, heads, loss, data, and split fixed, and change *only* whether the ESM-2 encoder is frozen or adapted with Low-Rank Adaptation (LoRA; (Hu et al. 2022)). LoRA is the natural parameter-efficient choice because it adapts the encoder while adding only ~1.8M trainable parameters, keeping the comparison close to a controlled ablation rather than a wholesale change of model capacity.

Our headline result is a **cautionary null**. On the aggregate primary metric the adapted encoder edges out the frozen one (0.580 vs 0.568, +0.011). But the gain is not real: it is generated wholly by two imbalanced binary pathogenicity heads on which the adapted model degenerates to predicting the majority (negative) class. There, accuracy rises sharply (human 0.70⟶0.94, animal 0.75⟶0.96) while F1 falls to *exactly zero*, the signature of a classifier that has stopped making positive calls. Once these two heads are excluded, encoder adaptation is net negative, and measured by F1 rather than accuracy it is neutral-to-harmful on the imbalanced traits it most needed to improve. The encoder, like the pooling operator before it, is not what cross-family generalization is waiting on.

**Contributions.**

1. **A controlled frozen-vs-LoRA comparison under family shift.** With encoder adaptation as the only experimental variable, on 19,278 genomes / 21 traits at family-held-out evaluation, we find no genuine benefit from fine-tuning ESM-2-150M (§4).
2. **A concrete metric-artifact diagnosis.** We show that the apparent aggregate improvement is an *accuracy illusion* driven by majority-class collapse on imbalanced binary heads, with F1 going to zero where accuracy goes up (§4.2). This is a reusable warning for any genomic-trait benchmark that aggregates accuracy across imbalanced labels.
3. **A second, independent confirmation of the cross-family bottleneck.** Combined with our pooling study, this gives two separate architectural levers, set aggregation and encoder adaptation, that both fail to move family-level generalization, sharpening the claim that distribution shift, not architecture, is the open problem (§5).

We are explicit about scope. This is a **single-seed pilot** with one encoder scale and three training epochs; it is intended as a focused, honest measurement and a methodological warning, not a benchmark sweep or a claim that fine-tuning can never help.

# Related Work

**Encoder fine-tuning vs frozen features.** Using a pretrained encoder frozen and training only a lightweight head is the cheaper alternative to fine-tuning the encoder end-to-end; parameter-efficient methods, of which LoRA (Hu et al. 2022) is the most widely used, sit between these poles by adapting the encoder through small low-rank updates. In protein modeling, ESM-2 (Lin et

al. 2023) is routinely used both frozen and fine-tuned, but the genome-level trait-prediction literature (Wiatrak et al. 2025; MicroGenomer authors 2025; BacPT authors 2026) reports benchmark numbers without isolating the frozen-vs-adapted decision under controlled distribution shift. Our contribution is to run exactly that ablation and report a null.

**Distribution shift in genomic prediction.** Taxonomy-aware (species / genus / family held-out) evaluation is the relevant stress test for predicting traits of clades unlike anything seen in training, the "microbial dark matter" regime. Our prior work established that the pooling advantage decays to zero at family holdout and that traits such as pathogenicity are heavily clade-confounded. The present note inherits that family-held-out protocol and asks the encoder-adaptation question within it.

**Metrics for imbalanced labels.** That accuracy is misleading under class imbalance is textbook, but it remains common to aggregate per-head accuracy into a single benchmark score. We provide a clean, mechanistic example from genomic trait prediction in which this aggregation *inverts* the qualitative conclusion: a model that gets strictly worse at the positive class is scored as better.

# Setup

## Data and splits

Labels are drawn from BacDive (Schober et al. 2025), giving **21 prediction heads** across seven biological blocks. For each strain with an NCBI genome we predict open reading frames and embed each protein independently with ESM-2-150M (`facebook/esm2_t30_150M_UR50D`), producing a ragged protein set per genome. We use the **family-held-out split**: no taxonomic family appears in more than one fold, the regime in which our earlier pooling advantage vanished. After restricting to genomes with retrievable sequences (19,278 genomes), the family split yields 11,600 train / 3,871 validation / 3,807 test genomes.

## Model and the single variable

The genome representation is produced by a Set Transformer pool (Lee et al. 2019) over the protein embeddings, feeding 21 task heads under a masked multi-task loss that contributes zero gradient for missing labels. Everything in this pipeline (pooling, heads, loss, optimizer family, data, split, and seed) is held fixed. The **only** experimental variable is the encoder treatment:

- **Frozen:** ESM-2 weights are fixed; only the Set-Transformer pool and heads train (5.54M trainable of 153.68M total). The encoder is a fixed feature extractor and is run under `no_grad`.
- **LoRA:** low-rank adapters (rank 16, $\alpha = 32$) are injected into the ESM-2 encoder and trained jointly with the pool and heads (7.39M trainable of 155.52M total). The encoder is adapted to the task.

This deliberately narrow contrast is what licenses reading any difference as an *encoder-adaptation* effect rather than a change in pooling capacity or data.

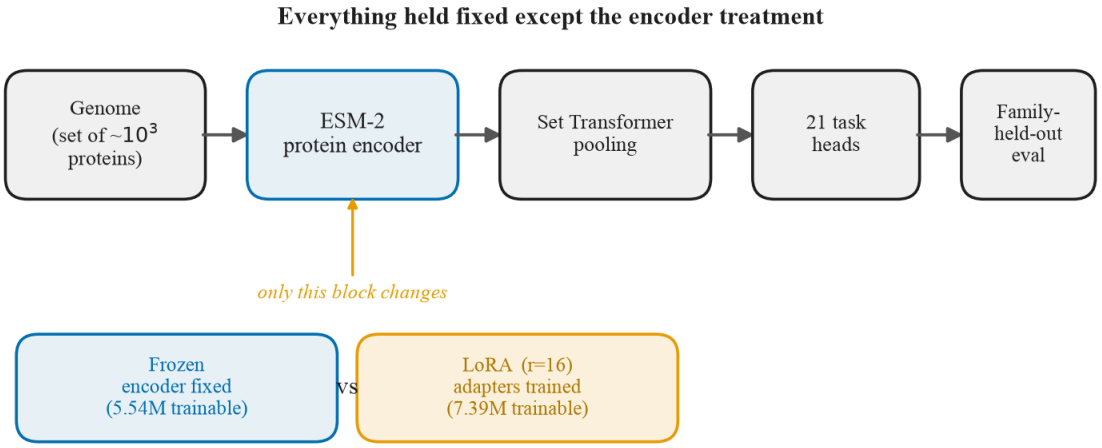


Figure 1: **Controlled comparison.** Data, pooling, task heads, loss, optimizer, split, and seed are held fixed; the *only* experimental variable is whether the ESM-2 encoder is frozen (a fixed feature extractor) or adapted with LoRA. Any difference in family-held-out performance is therefore attributable to encoder adaptation.

### Training

Both arms train for 3 epochs with balanced-family sampling and class-weighted losses, identical protein caps (128 proteins/genome, encoder micro-batch 8), and the same base learning rate ($5 \times 10^{-4}$). Both runs shard across 8 GPUs via distributed data parallelism; following standard large-batch practice we scale the learning rate by the square root of the world size ($\sqrt{8} \approx 2.83\times$, to $1.41 \times 10^{-3}$) with a short linear warmup. We report each head's primary metric (accuracy for single-label classification heads, micro-F1 for multilabel heads, RMSE for the regression-vector head) and the dataset-level average of primary metrics (`avg_primary`). Crucially, because some primary metrics are accuracy on imbalanced binary labels, we also inspect **F1** on those heads directly.

## Results

### Aggregate: a small apparent win

Table 1: Family-held-out aggregate (single seed, seed 0).

| Encoder treatment | Trainable params | `avg_primary` (test) |
|---|---|---|
| Frozen | 5.54M | 0.568 |
| LoRA | 7.39M | **0.580** (+0.011) |

Read at face value, encoder adaptation gives a +0.011 improvement in the aggregate primary metric. In-sample the adapted model also fits better (LoRA train loss 0.88 vs frozen 0.95 by epoch 2) and shows a slightly higher validation average (0.621 vs 0.604). A naive reading would conclude that fine-tuning the encoder modestly helps cross-family generalization. **The rest of this section shows that reading is wrong.**

## The gain is a majority-class artifact

Decomposing the aggregate by head reveals that the *entire* improvement is carried by two heads, and in a way that signals failure rather than success.

Table 2: The two heads that produce the aggregate gain. Accuracy is the primary metric; F1 is shown alongside. The *no-skill* column is the accuracy of a trivial classifier that always predicts the majority (negative) class.

| Head | No-skill acc | Acc (frozen ⟶ LoRA) | F1 (frozen ⟶ LoRA) |
|---|---|---|---|
| pathogenicity_human | 0.939 | 0.699 ⟶ 0.939 | 0.176 ⟶ **0.000** |
| pathogenicity_animal | 0.956 | 0.753 ⟶ 0.956 | 0.152 ⟶ **0.000** |

The decisive observation is that **LoRA's accuracy equals the no-skill baseline to three decimals** (0.939 and 0.956): the adapted model has converged exactly onto the trivial majority-class predictor. The frozen encoder scores *lower* accuracy precisely because it still spends predictions on the rare positive class (nonzero F1). On both pathogenicity heads the adapted model's accuracy jumps by ~0.20 to 0.24 while its F1 drops to *exactly zero*. F1 = 0 with high accuracy is the unambiguous signature of **majority-class collapse**: the adapted model has stopped predicting the (rare) positive class entirely, scoring well on accuracy precisely because ~94% of test genomes are negative. The frozen encoder, by contrast, still recovers some positives (F1 0.18 and 0.15). In other words, on the two traits that drive the headline number, encoder adaptation makes the classifier *strictly worse* at the thing the trait is about (Figure 2).

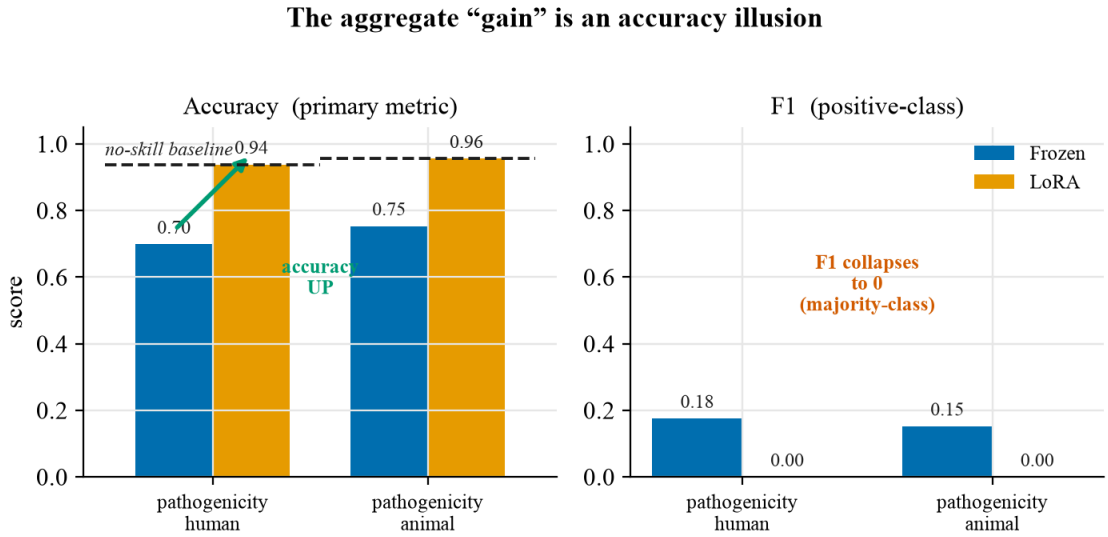


Figure 2: **The aggregate gain is an accuracy illusion.** On the two pathogenicity heads that drive the headline number, encoder adaptation *raises* accuracy (left) while *collapsing* F1 to exactly zero (right). High accuracy with zero F1 is the signature of a classifier that has stopped predicting the rare positive class; it is rewarded by accuracy precisely for getting worse at the trait.

These two heads contribute +0.021 to `avg_primary` on their own, more than the entire +0.011 aggregate gain. Removing them flips the conclusion:

Table 3: Aggregate with and without the two pathogenicity heads.

| Subset | `avg_primary` Δ (LoRA − frozen) |
|---|---|
| All 21 heads | +0.011 |
| Excluding 2 pathogenicity heads | **−0.011** |

Excluding only the two artifact heads, encoder adaptation is **net negative** across the remaining 19 heads (Figure 3).

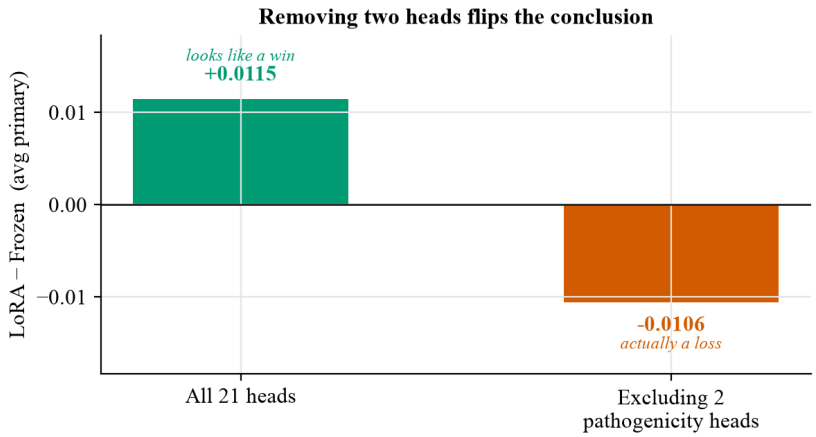


Figure 3: **Removing two heads flips the sign.** The aggregate primary metric favors LoRA by +0.0115 over all 21 heads, but once the two majority-class-collapse pathogenicity heads are excluded, the same comparison becomes −0.0106. The "win" lives entirely in the artifact.

## On F1, adaptation is neutral-to-harmful

The artifact is specific to accuracy on imbalanced labels, so we look at F1 where it is informative (Figure 4). The largest F1 movements under adaptation are **losses**, not gains: temperature_class −0.144, pathogenicity_human −0.176, pathogenicity_animal −0.152, halophily −0.047, ph_class −0.021, isolation_source −0.026. The few F1 gains are small and concentrated in multilabel/regression heads: carbon_utilization +0.034 (micro-F1), motility +0.023, fatty-acid RMSE improves by 0.026. The balance is firmly toward *no benefit or harm* from adaptation on exactly the imbalanced traits a better encoder was supposed to help. Across all heads the tally is roughly even (a handful up, a handful down, many unchanged), with the average dragged positive only by the accuracy artifact of §4.2.

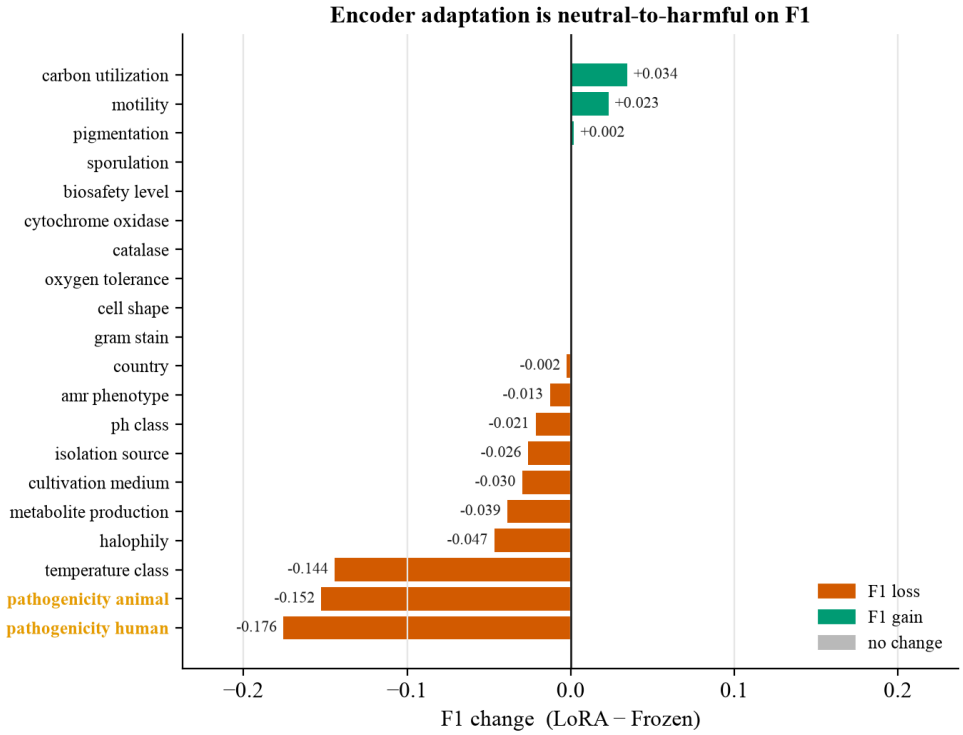


Figure 4: **Per-head F1 change under encoder adaptation (LoRA – frozen).** Heads are sorted by effect. The large movements are losses, including both pathogenicity heads (highlighted labels) that collapse to F1 = 0; gains are few and small. On the metric that actually tracks positive-class performance, adaptation is neutral-to-harmful.

The artifact also has a clear signature across heads: the traits where adaptation most *raises* accuracy are exactly the traits where it most *lowers* F1 (Figure 5). The two pathogenicity heads sit alone in the bottom-right "accuracy-up, F1-down" quadrant, precisely the corner an accuracy-only benchmark rewards and an F1-aware one penalizes.

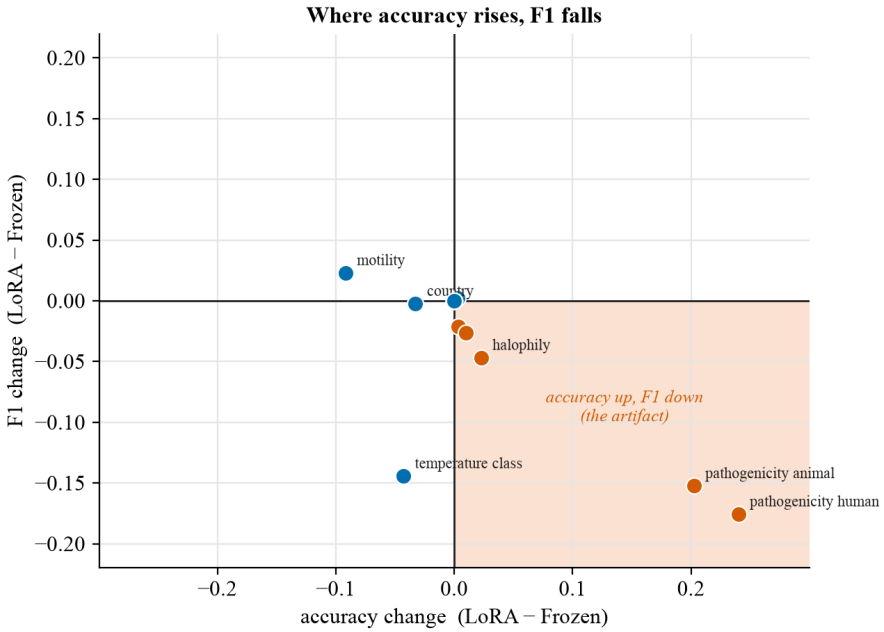


Figure 5: **Where accuracy rises, F1 falls.** Each point is a classification head, plotted by its change in accuracy (x) against its change in F1 (y) under encoder adaptation. The two pathogenicity heads occupy the shaded bottom-right quadrant, where accuracy increases while F1 collapses, isolating the metric artifact that drives the aggregate number.

## Discussion

**The encoder is not the binding constraint either.** Our earlier study showed that swapping the pooling operator (mean vs attention) does not survive family-level shift. This note shows that the far more expensive move, adapting the encoder weights themselves with LoRA, *also* fails to move family-level generalization once the accuracy artifact is removed. Two independent architectural levers, set aggregation and encoder adaptation, both come up empty in the same regime. This is consistent with the interpretation that the limiting factor in cross-family genomic trait prediction is not how we represent or pool a genome's proteins, but whether *any* protein-set representation transfers across evolutionary distance at all. Architecture is not the lever; distribution shift is the problem.

**A warning about aggregated accuracy.** The cleanest practical takeaway is methodological. A genomic-trait benchmark that aggregates per-head accuracy will *reward* a model for collapsing to the majority class on imbalanced labels, exactly the failure mode that matters most for rare, high-stakes traits like pathogenicity. Had we reported only `avg_primary`, we would have published a (false) positive result for encoder fine-tuning. Reporting F1 alongside accuracy on every imbalanced head, and being suspicious of accuracy gains that coincide with F1 collapse, is a cheap and necessary guard.

**Why might adaptation collapse here?** We avoid over-interpreting a single-seed run, but note two plausible, testable contributors. First, with class-weighted losses and a learning rate scaled up

$\sqrt{8}\times$ for 8-way distributed training, the adapted encoder may overfit the abundant negative class on the rare pathogenicity labels under family shift, where the positive examples in the test families are unlike anything in training. Second, three epochs of LoRA at the scaled learning rate may simply move the encoder toward a representation that fits validation families while transferring worse to held-out families on the imbalanced heads. Both are checkable with a learning-rate / epoch sweep and multiple seeds, which we have not yet run.

**Limitations, stated plainly.**

- **Single seed.** Each arm is one run (seed 0). We report no error bars; the small aggregate differences are well within plausible seed variance, which is itself part of why the +0.011 "win" should not be trusted.
- **One encoder scale, short schedule.** All results use ESM-2-150M and 3 epochs. Whether a larger encoder, a longer schedule, or an unscaled learning rate changes the picture is untested.
- **Recipe coupling in the LoRA arm.** The adapted arm inherits $\sqrt{\text{world-size}}$ LR scaling and warmup; disentangling adaptation from the LR/large-batch interaction requires a dedicated ablation.
- **Not an end-to-end fine-tune.** We test LoRA, the parameter-efficient lever, not full-encoder fine-tuning, which has more capacity but also more overfitting risk under family shift.
- **Family-split only.** We did not re-run species/genus holdout here; our claim is specifically about the cross-family regime that motivates the question.

**What would change the conclusion.** A genuinely positive result for encoder adaptation would need to show, *on F1 and across seeds*, gains on the imbalanced heads under family holdout, not an accuracy bump that coincides with F1 collapse. That experiment, with multiple seeds, an LR/epoch sweep, and matched within-family pathogenic / non-pathogenic controls, is the natural next step.

# Conclusion

Asked whether fine-tuning the protein encoder, rather than freezing it, helps bacterial trait prediction generalize to unseen families, the honest answer in this pilot is **no**. The aggregate metric says yes by a small margin, but that margin is manufactured by majority-class collapse on two imbalanced pathogenicity heads, where accuracy rises and F1 falls to zero. Net of that artifact, LoRA adaptation of ESM-2-150M does not improve, and slightly degrades, family-held-out performance. Together with our pooling study, the result points the same direction twice: reading and representing a genome's proteins is achievable, but making those representations *transfer across clades*, by better pooling or by adapting the encoder, remains the open problem, and aggregated accuracy will hide that fact if you let it.

# Data and Code Availability

The repository includes the trait schema, split definitions, model and fine-tuning code, and the frozen and LoRA result records (`runs/lora/frozen_family_s0.json` and the corresponding LoRA metrics) used for every number above. The comparison is reproducible from these two run

records; large protein embedding stores are not bundled but the scripts and identifiers needed to regenerate them are included.

# References


BacPT authors. 2026. "A Bacterial Proteome Foundation Model." *bioRxiv*, ahead of print. https://doi.org/10.64898/2026.03.07.710335.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. 2022. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations (ICLR)*.

Lee, Juho, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. "Set Transformer: A Framework for Attention-Based Permutation-Invariant Neural Networks." *International Conference on Machine Learning (ICML)*, 3744–53.

Lin, Zeming, Halil Akin, Roshan Rao, et al. 2023. "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model." *Science* 379 (6637): 1123–30. https://doi.org/10.1126/science.ade2574.

MicroGenomer authors. 2025. "MicroGenomer: A Genome Language Model for Microbial Phenotype Prediction." *bioRxiv*, ahead of print. https://doi.org/10.64898/2025.12.28.696777.

Schober, Isabel, Carola Söhngen, Lorenz Christian Reimer, et al. 2025. "BacDive in 2025: The Expanding Bacterial and Archaeal Diversity Knowledge Resource." *Nucleic Acids Research*, ahead of print. https://doi.org/10.1093/nar/gkae959.

Wiatrak, Maciej, Aaron Weimann, R. Andres Floto, Maria Brbić, et al. 2025. "Bacformer: A Foundation Model for Bacterial Genomes." *bioRxiv*, ahead of print. https://doi.org/10.1101/2025.07.20.665723.