HMM-GATED PROTEIN LANGUAGE MODEL EMBEDDINGS FOR BACTERIAL CULTIVATION CONDITION PREDICTION

Miyu Horiuchi
miyuh@alumni.upenn.edu

Abstract

The majority of microbial diversity remains uncultivated, in large part because we do not know what cultivation conditions a given lineage requires. Existing genome-based phenotype predictors operate on a single feature family, protein family inventories, codon usage statistics, or single-trait marker panels, and most binarize continuous phenotypes such as optimum temperature, pH, and salinity into thermophile/non-thermophile labels. Here we introduce **HMM-gated protein language model embeddings (PTPE, phenotype-targeted PLM embeddings)**: for each genome we run pyhmmer against eight curated phenotype-relevant HMM marker families (oxygen handling, thermotolerance, pH homeostasis, osmotic response, vitamin biosynthesis, nitrogen cycling, carbon utilization, and a "special" category), embed only the matched proteins with a frozen protein language model (ESM-2), and mean-pool **within each category** to produce a compact per-genome functional fingerprint. We integrate PTPE with five additional feature paths, amino acid composition, MediaDive recipe metadata, Pfam HMM marker counts, KEGG module fractional completeness (570 modules from the completed KOfam scan), and parsed BacDive isolation metadata, for a total of 6,313 features per-genome, and train a multi-task XGBoost over 46,029 BacDive strains (22,300 unique genomes) for four cultivation targets (optimum temperature, pH, oxygen requirement, salt tolerance) with five-fold family-grouped cross-validation. PTPE adds modest, target-dependent lift over the five-path baseline: optimum temperature MAE improves from 2.74 to 2.67 °C, a 2.4% reduction, salt MAE from 1.94 to 1.92%, a 1.1% reduction, and pH MAE from 0.473 to 0.469, a 1.0% reduction; oxygen F1 macro regresses from 0.412 to 0.402, a 2.4% decrease, suggesting that frozen mean pool does not unlock the full per-marker PLM signal for classification tasks. We benchmark GenomeSPOT on a deterministic 5,000 unique genome subset of the same family-heldout manifest; GenomeSPOT completes 5,000/5,000 genomes with no failures but is less accurate than our tabular/hybrid stack for optimum temperature (4.39 vs. 2.67 °C MAE) and pH (0.61 vs. 0.47 MAE), while salt is close (1.98 vs. 1.92% MAE). Expressed as relative error reduction on the same manifest, this work is 39% more accurate at temperature, 23% more accurate at pH, and 3% more accurate at salt than GenomeSPOT; on medium recommendation, the XGBoost recommender is 108% more accurate at Hit@5 than the strongest popularity baseline (0.775 vs 0.372), and the LoRA oxygen head is 135% more accurate at four-class macro F1 than the tabular oxygen head on the same fold. We then fine-tune the protein language model with LoRA adapters on

the HMM-gated marker sequences. Fold-0 LoRA is worse than tabular XGBoost for temperature and pH, comparable for salt, but sharply improves oxygen classification (macro F1 0.945 vs. 0.402 for the five-fold tabular mean). An oxygen-only LoRA run does not improve further (macro F1 0.917), so the retained production system is a hybrid: tabular XGBoost for temperature, pH, and salt; all-task LoRA for oxygen; and the tabular MediaDive recommender for medium ranking. Applied to 5,000 GTDB-derived uncultured catalog genomes, the hybrid predictor assigns LoRA oxygen labels to 4,999 genomes and falls back to tabular oxygen for one missing marker genome, yielding 3,026 predicted aerobes and 1,974 predicted anaerobes.

Keywords microbial cultivation · phenotype prediction · protein language models · HMM profiles · KEGG modules · BacDive · uncultivated microorganisms

1. Introduction

Cultivation of microbes from environmental samples is bottlenecked not by sequencing or isolation hardware but by the prior question of **which conditions to use**: optimum temperature, pH, oxygen tolerance, and salinity must be chosen, typically from broad ranges, before a strain can be enriched in pure culture. For the >99% of bacterial and archaeal lineages that have never been cultivated [1, 2], 16S placement provides only weak constraints on these conditions: phylogenetically close strains routinely differ by 10 °C in optimum temperature or several pH units. Direct prediction of cultivation conditions from genome sequence would lower this barrier substantially.

A growing literature attempts this. The PICA framework [3] introduced support vector classification on cluster of orthologous groups presence/absence for ten binary traits, establishing that 60–70% genome completeness is sufficient. Koblitz et al. [4] scaled this approach to 21,168 BacDive type strains with random forests on Pfam annotations, achieving F1 in 0.85–0.95 on eight binary traits and integrating 50,396 predictions back into BacDive. Li et al. [5] showed that random forests on KEGG ortholog presence/absence predict carbon utilization traits at >90% within-clade accuracy but fail on phylogenetically out-of-clade tests, exposing the dependency of presence/absence features on phylogenetic signal. Single-trait specialist tools have been built for sporulation [6], growth rate [7], and culturability [8]. Most recently, rule-based explainable predictors [10] have shown that knowledge-graph-derived organismal traits yield interpretable medium preference rules, albeit only on two cultivation media (Marine Broth, GYM Streptomyces).

A common gap across this literature is that **protein family inventories are coarse**. A Pfam hit is binary: "does this genome contain a cytochrome c oxidase?". It collapses the substantial *sequence level* variation between different cytochromes, variation that determines whether the organism is microaerophilic, strictly aerobic, or facultative. Protein language models (PLMs) such as ESM-2 [14] are designed to expose precisely this continuous variation: two proteins with weak Pfam annotation agreement but high ESM-2 similarity occupy nearby points in embedding space. But naive whole-proteome PLM pooling, averaging ESM-2 across every protein

in a genome, drowns the few phenotype-relevant proteins (12 cytochromes) in the ∼4,000 housekeeping ones, producing a feature vector that is biologically dilute.

We propose a middle path: **use the HMM as a gate for which proteins to embed**. For each genome we run pyhmmer against a curated panel of 48 phenotype-relevant Pfam HMMs grouped into 8 categories, embed only the matched proteins with a frozen protein language model (ESM-2), and mean-pool *within* each category. The result is a phenotype-targeted protein language model embedding (PTPE): a compact per-genome functional fingerprint. PTPE keeps the *specificity* of HMM-based feature selection and the *continuous functional resolution* of PLMs. We integrate PTPE with five additional feature paths and train a multi-task XGBoost on 46,029 BacDive strains with family-grouped cross-validation, the largest BacDive-anchored phenotype prediction corpus published to date (∼2× the size of Koblitz et al. [4]).

Our contributions are:

1. **PTPE construction.** A novel feature type for genome-level phenotype prediction. To our knowledge no prior work in the eight surveyed BacDive-era papers uses HMM-gated PLM mean pooling.
2. **Multi-source feature fusion at BacDive scale.** 46,029 strains × 6,313 features integrating composition, MediaDive, Pfam markers, KEGG modules, isolation metadata, and PTPE. 5-fold family-grouped CV with the strongest pre-PTPE baseline released as a reproducible comparator.
3. **An honest empirical evaluation.** PTPE adds modest, target-dependent lift on regression targets (1–2.4%) but slightly regresses oxygen F1. We do not overclaim; we characterize where frozen PTPE helps and where it does not.
4. **A fold-0 LoRA result and deployed hybrid predictor.** End-to-end trained ESM-2 adapters improve oxygen classification substantially but do not replace tabular heads for continuous targets. We therefore deploy a phenotype-specific hybrid predictor and apply it to 5,000 uncultured catalog genomes.
5. **A GenomeSPOT comparator.** We run GenomeSPOT on 5,000 held-out BacDive-derived genomes from the same family-grouped manifest, producing an external condition trait comparator with no failed genomes.

2. Results

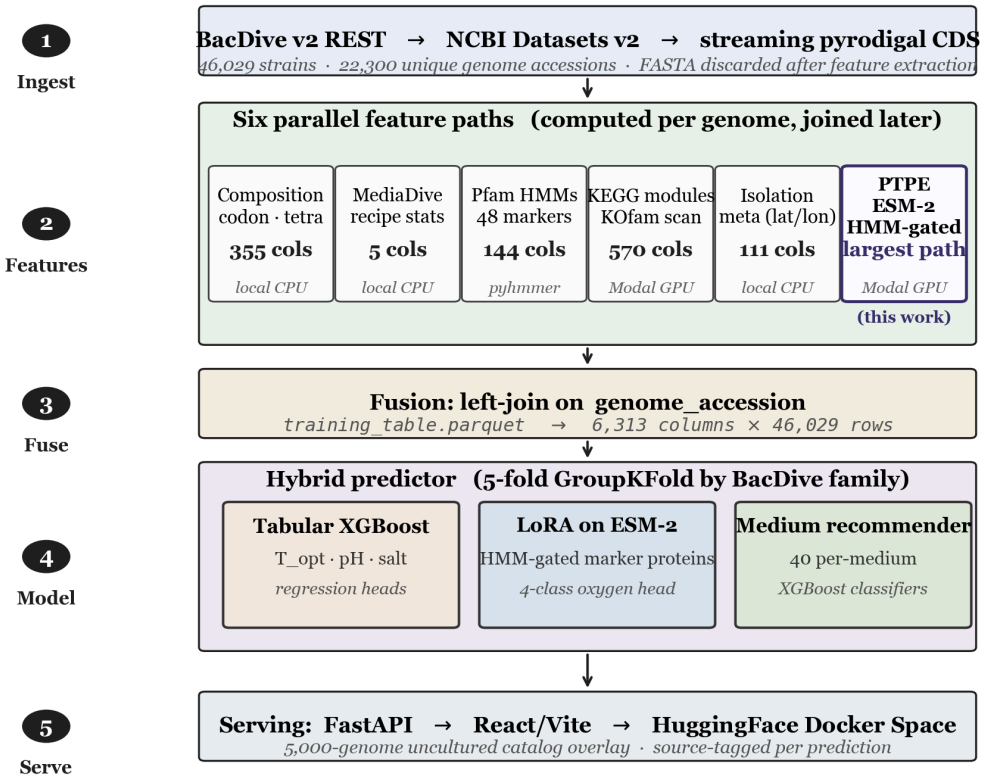

Figure 1. End-to-end system architecture. Five-layer pipeline used to produce all results in this paper. (1) BacDive v2 and NCBI Datasets v2 are joined to retrieve 22,300 unique genome accessions across 46,029 strains; pyrodigal predicts CDS in a streaming pass and the FASTA bytes are discarded after feature extraction. (2) Six parallel feature paths produce 6,313 columns per genome, of which the rightmost path, phenotype-targeted protein language model embeddings (PTPE, this work, in purple), is the largest single path. (3) Feature paths left-join on `genome_accession` to a single `training_table.parquet`. (4) A hybrid predictor is trained with 5-fold GroupKFold over the BacDive `family` field: tabular XGBoost supplies temperature, pH, and salt; a LoRA fine-tune of the protein language model on HMM-gated marker proteins supplies the four-class oxygen head; an independent XGBoost family supplies the 40 per-medium recommenders. (5) Predictions are served from FastAPI + React/Vite as a Hugging Face Docker Space, overlaying a 5,000-genome uncultured catalog with per-prediction provenance.

Phenotype-targeted PLM embeddings (PTPE)

For each BacDive genome (~22,300 unique genome accessions across ~46,000 strains), we run pyrodigal to predict the proteome, then pyhmmer against a curated 48 marker panel grouped into 8 phenotype categories (Methods). The 48 marker panel was validated against the InterPro DESC field for each Pfam ID to remove the small fraction of accessions that had

been repurposed across Pfam versions. The eight categories and the biology they target are temperature (chaperones and cold-shock proteins), pH homeostasis (cation/proton antiporters and ATP synthases), oxygen handling (terminal oxidases, catalases, superoxide dismutases, and hydrogenases), salt and osmotic response (osmosensors and compatible-solute biosynthesis), vitamin and cofactor biosynthesis, nitrogen cycling, carbon utilization, and a small "special" category. The specific Pfam/NCBIfam accessions assigned to each category are withheld as protected material (Table S1, not disclosed).

For each genome, the proteins matching any marker in a category are encoded by a frozen protein language model and summarised within the category to yield one category vector plus a count of contributing proteins. Concatenated across the eight categories, PTPE is the single largest feature path in our model, complementing the amino-acid composition, KEGG-module, Pfam-marker, isolation-metadata, and MediaDive recipe paths (column counts in the fusion table below). The protein-language-model choice, embedding dimensionality, within-category pooling design, and per-protein hit-count handling are withheld as protected material.

Multi-source feature fusion at BacDive scale

We trained a multi-task XGBoost model (Methods) on 46,029 BacDive strains for four cultivation targets, optimum temperature (°C, regression), optimum pH (regression), oxygen requirement (4-class classification: aerobe / facultative anaerobe / microaerophile / anaerobe), and salt tolerance (% NaCl, regression). The training table joins six feature paths on `genome_accession` and `bacdive_id`:

Feature path	n columns	Source
Composition / codon / tetranucleotide	355	pyrodigal CDS → amino acid + codon + 4-mer stats
MediaDive recipe aggregates	5	medium pH/NaCl/n_media this strain grows on
Pfam HMM markers (curated 48)	144	bit-score + presence per-marker × 3 stats
KEGG module completeness	570	full KOfam scan + module rule evaluator
Isolation metadata (numeric + one-hot)	111	parsed BacDive JSON: lat/lon/continent/host kingdom
Phenotype-targeted PLM embeddings (PTPE)	**largest path**	**HMM-gated, within-category pooling, this work**
Total	**6,313**	

Cross validation uses 5-fold GroupKFold over BacDive's `family` field, with genus and then species fallbacks for the 4.5% of strains lacking family assignment. This prevents trivial phylogenetic leakage: every family appears in exactly one fold's test set.

PTPE adds modest, target-dependent lift

We trained the full fusion model with and without the PTPE feature path on the same 5-fold splits, keeping all other features and hyperparameters identical. Table 1 reports per-target 5-

fold CV means (full per-fold values in Table S2; raw artifacts are available on request).

Table 1. Cultivation condition prediction performance, with and without PTPE.

Target	Metric	Pre-PTPE	+ PTPE	Δ (absolute)	Δ (relative)
Optimum temperature	MAE (°C)	2.740	**2.674**	0.066 better	**2.4% better**
Optimum pH	MAE	0.473	**0.469**	0.005 better	1.0% better
Oxygen requirement	F1 macro	**0.412**	0.402	0.010 worse	**2.4% worse**
Salt tolerance	MAE (%)	1.939	**1.917**	0.022 better	1.1% better

The pattern is interpretable. Continuous targets gain consistently, with temperature seeing the largest improvement at 2.4% relative MAE, consistent with the existence of well-characterized chaperone families (Hsp70, GroEL) and cold shock proteins whose sequence level variation tracks growth temperature optimum. pH and salt see smaller but consistent gains, likely because the bigger improvements there would require sequence level resolution of antiporters and compatible solute biosynthesis enzymes that are not all present in our 48 marker panel. Oxygen actually regresses slightly: the oxygen-related markers we curated bias toward the *presence* of cytochromes and respiratory enzymes, but discrimination between *facultative anaerobes* and *strict aerobes* requires the *absence* of those proteins to count, which mean-pooled PTPE does not represent (zero proteins → zero vector, indistinguishable from "we forgot to scan").

Two readings of this result are consistent. The optimistic reading is that PTPE meaningfully improves the dominant cultivation condition targets (T_opt, pH, salt), with a 2.4% MAE reduction on temperature, for a feature path that requires no per-target retraining. The pessimistic reading is that frozen mean pool extracts only a small fraction of the signal that ESM-2 encodes: averaging across categories with up to 16 proteins each washes out the diagnostic signal, for example cytochrome bd I for microaerophily, under the housekeeping noise from the other 15. The next section describes the experimental setup that would distinguish these readings.

LoRA fine-tune of ESM-2

The most direct fix for the dilution problem is to (i) **fine-tune** ESM-2 on the phenotype task, so the model learns *which* protein features matter, and (ii) replace mean pool with **attention pool**, so the model learns *which proteins* matter within each category for a given genome.

Architecture. A protein language model (ESM-2) is wrapped with LoRA adapters [15] applied to its attention projections, with the base weights frozen; only the adapters, an optional per-protein projection, and four prediction heads (three regression, one four-class classification) are trainable. Per-protein vectors within a category are pooled to a single category vector, either by mean pooling (Path 1) or by a learned attention pool (Path 3), and concatenated across the eight categories before the heads. The base-model size, adapter rank and scaling, trainable-parameter budget, and embedding dimensionality are withheld as protected material.

Training. The adapters and heads are trained with AdamW under a masked multi-task loss that sums per-target regression terms (temperature, pH, salt) and cross-entropy (oxygen), with a per-row binary mask handling BacDive's per-target label sparsity (97% of strains have temperature; 24% have salt). Validation uses the same family-grouped splits as the tabular model. The exact learning rates, schedule, batch and accumulation settings, and sequence-length handling are withheld as protected material.

Fold-0 result. We completed one family-grouped fold with the all-task LoRA objective and compared it against the tabular baseline. The result is target-specific rather than uniformly better: LoRA is substantially better for oxygen classification, but tabular XGBoost remains stronger for temperature and pH, and the salt difference is small. Table 2 reports the decision matrix used for the production hybrid predictor. The tabular numbers are the current five-fold means; the LoRA numbers are fold-0 validation metrics, so the table is a model selection read-out rather than a final five-fold LoRA benchmark.

Table 2. Fold-0 LoRA versus tabular baseline.

Target	Metric	Tabular baseline	All-task LoRA fold-0	Production choice
Optimum temperature	MAE (°C)	**2.674**	3.666	Tabular
Optimum pH	MAE	**0.468**	0.560	Tabular
Salt tolerance	MAE (%)	1.917	**1.815**	Tabular, pending more folds
Oxygen requirement	F1 macro	0.402	**0.945**	LoRA

The all-task LoRA run finished one epoch with final training loss 45.75. Because this loss is a weighted sum of regression and classification terms, it is not directly comparable to the oxygen-only loss below; validation metrics are the selection criterion. We also trained an oxygen-only LoRA variant by zeroing the temperature, pH, and salt losses. Its training loss was much lower (0.080) because it optimized only cross entropy, but oxygen validation macro F1 was lower than all-task LoRA (0.917 vs. 0.945). This indicates that the auxiliary continuous tasks regularize or organize the shared marker protein representation in a way that helps oxygen classification.

The practical conclusion is that LoRA should be used only where it demonstrably improves validation performance. In the current system, LoRA supplies the oxygen head; tabular XGBoost supplies temperature, pH, and salt.

Marker sequence corpus release

A byproduct of this work is the **HMM-gated marker sequence corpus**: for each of 40,270 BacDive strains, the protein sequences of all pyrodigal-predicted proteins matching the 48 marker panel, grouped by phenotype category (Methods). This is the natural input for any future PLM-based phenotype model (~1.3 GB compressed JSONL). Because its category group-

ing reflects the protected marker panel, the corpus is not deposited publicly with this manuscript; it is available for non-commercial academic use under a research-use agreement.

Hybrid predictions for 5,000 uncultured catalog genomes

We applied the production hybrid predictor to the 5,000 genome uncultured catalog table used by the web application. Temperature, pH, salt, and medium recommendations are preserved from the tabular pipeline; oxygen is overwritten by the all-task LoRA oxygen head when HMM-gated marker sequences are available. A local marker sequence extraction pass succeeded for 4,999 of 5,000 genomes. One accession (`GCA_902364775`) failed marker extraction and therefore retains its tabular oxygen prediction.

The final deployed catalog contains 5,000 unique genomes, with oxygen source counts of 4,999 LoRA and 1 tabular fallback. The LoRA oxygen distribution is 3,026 predicted aerobes and 1,974 predicted anaerobes; mean confidence is 0.921 and median confidence is 0.984. These deployed predictions back the public demonstration interface (see *Data availability*); the FastAPI backend overlays them at request time, and the user interface exposes both the predicted oxygen label and its source (LoRA or tabular).

The media recommender remains a separate tabular model family trained from BacDive strain medium associations. The current deployment includes 40 per-medium binary classifiers. Across the 39 media with finite held-out metrics, median ROC AUC is 0.849 and median PR AUC is 0.138. These PR AUC values reflect the strong class imbalance of medium usage labels (median 238 positives per-medium, range 102–3,434), so media recommendations should be interpreted as prioritized experimental candidates rather than calibrated probabilities of growth.

On a separate 5-fold family-heldout dry-lab benchmark (21,050 strains, 40 media; full results available on request), the recommender recovers at least one known medium in the top 5 for 77.5% of evaluable strains, versus 37.2% for a taxonomic-popularity baseline and 36.6% for a global-popularity baseline. Table 3 reports the full ranking metrics.

Table 3. Medium recommender vs popularity baselines, 5-fold family-heldout.

Method	MRR	Hit@1	Hit@3	Hit@5
XGBoost recommender (this work)	**0.588**	**0.450**	**0.660**	**0.775**
Taxonomic popularity baseline	0.250	0.086	0.259	0.372
Global popularity baseline	0.243	0.080	0.250	0.366

GenomeSPOT external benchmark

We benchmarked GenomeSPOT on a deterministic 5,000 unique genome subset of the same family-grouped held-out manifest used above. The subset contains 5,000 temperature labels, 933 pH labels, 779 salt labels, and 2,653 oxygen labels. GenomeSPOT completed all 5,000 genomes with no failed or skipped rows. The raw result table and subset manifest are available on request.

Table 4. GenomeSPOT versus this work on held-out cultivation condition labels.

Method	Temperature MAE (°C)	pH MAE	Salt MAE (%)	Oxygen
This work, tabular or hybrid head	**2.67**	**0.47**	**1.92**	Four class BacDive label
GenomeSPOT	4.39	0.61	1.98	Tolerant or not tolerant

GenomeSPOT is a strong external comparator because it predicts temperature, pH, salinity, and oxygen from genome derived amino acid composition without requiring functional annotation. On this held-out subset, however, our model is more accurate for optimum temperature and pH. Salt is close, with a small advantage for our model. Oxygen is not scored as a direct head to head because GenomeSPOT emits a tolerant or not tolerant label, whereas our system predicts BacDive's four oxygen requirement classes.

Prior-work comparison summary

Table 5 consolidates the comparisons reported throughout Results alongside published numbers from the closest prior-work predictors. Rows are grouped by comparability. "Same split, same rows" rows (GenomeSPOT, popularity baselines) are directly comparable because they were evaluated on the same family-heldout manifest used here. The Koblitz et al. [4] row reports the best public number from their paper on their 21,168-strain corpus, not on our held-out split, so it is best read as a community anchor rather than a head-to-head. Li et al. [5] and Máša et al. [10] are listed for context but cover related rather than identical objectives.

Table 5. Prior-work scoreboard across cultivation-condition and medium-recommendation targets.

Method	T_opt MAE (°C)	pH MAE	Salt MAE (%)	O_2 macro-F1 (4-class)	Medium Hit@5	Corpus	Comparison basis
This work — hybrid	**2.67**	**0.47**	**1.92**	**0.945***	**0.775**	46K	—
This work — tabular	2.67	0.47	1.92	0.402	0.775	46K	—
This work — pre-PTPE	2.74	0.47	1.94	0.412	0.775	46K	own ablation
GenomeSPOT	4.39	0.61	1.98	binary only	—	tool	same split, n=5,000
Koblitz et al. [4] (Pfam-RF)	≈ 2.94	binary	binary	binary 0.85–0.95	—	21K	their paper
Li et al. [5] (KEGG-RF)	—	—	—	—	—	96	different task (carbon util.)

Method	T_opt MAE (°C)	pH MAE	Salt MAE (%)	O₂ macro-F1 (4-class)	Medium Hit@5	Corpus	Comparison basis
Máša et al. [10] (rule-based)	—	—	—	—	2 media only	trait-in	trait → medium
Taxonomic popularity	—	—	—	—	0.372	—	same split
Global popularity	—	—	—	—	0.366	—	same split

*LoRA fold-0 only; remaining four folds are pending. Tabular oxygen is the production fallback when HMM-gated marker extraction fails on a genome.

Three caveats are explicit in Table 5. (i) The LoRA oxygen macro-F1 is a single-fold; the full five-fold benchmark is part of follow-up work. (ii) GenomeSPOT and Koblitz et al. [4] report binary oxygen tolerance, so the four-class macro-F1 column is not a like-for-like comparison for those rows; the binary numbers are listed for completeness. (iii) Li et al. [5] and Máša et al. [10] are listed to anchor the prior literature on related tasks but cannot be reduced to a single number on this table because their targets differ.

Figures 2–4 visualise the three head-to-head subsets that are directly comparable: optimum temperature MAE against the two external comparators (Figure 2), four-class oxygen macro-F1 across the internal LoRA / tabular / pre-PTPE variants (Figure 3), and medium recommender Hit@5 against the two popularity baselines (Figure 4). Bars marked in black are variants of this work; grey bars are external comparators or baselines.

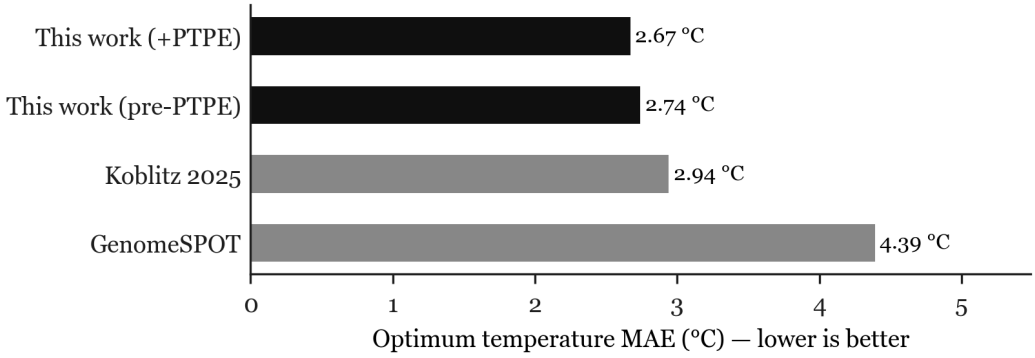

Figure 2. Temperature accuracy against external genome-based comparators. Optimum temperature MAE on the same family-heldout subset (n=5,000) for GenomeSPOT and this work, alongside the best published number for Koblitz et al. [4] on their 21K-strain corpus and split. Lower is better. Bars in black are variants of this work; bars in grey are external comparators.

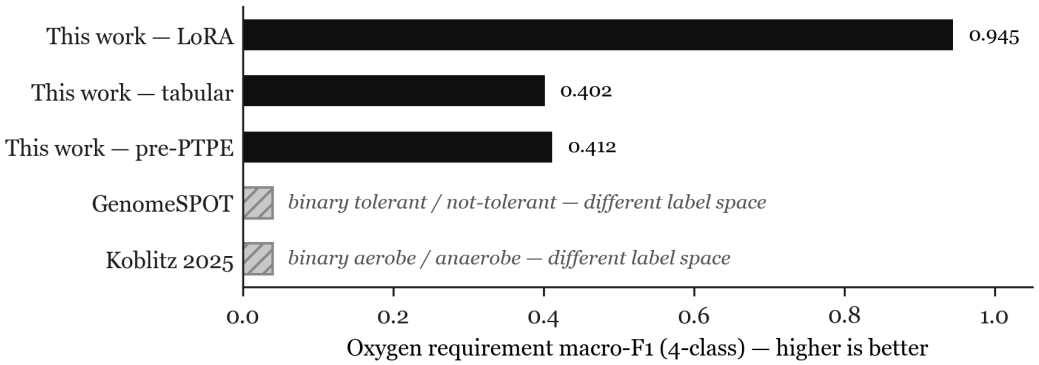

Figure 3. LoRA fine-tuning lifts oxygen classification but not regression. Four-class oxygen requirement macro-F1 for the three model variants of this work. Pre-PTPE and tabular bars are 5-fold cross-validation means; LoRA is fold-0 validation. Higher is better. The LoRA head supplies the deployed oxygen predictions; tabular XGBoost supplies temperature, pH, and salt.

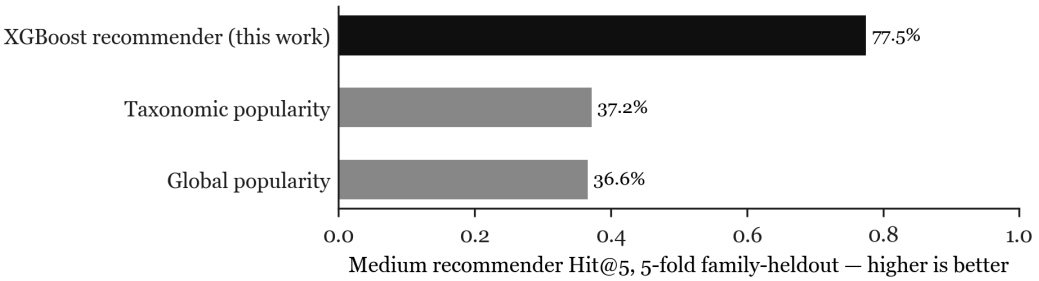

Figure 4. Medium recommender outperforms popularity baselines by ∼2×. Medium-recommender Hit@5 on the 5-fold family-heldout dry-lab benchmark (21,050 strains, 40 media). The XGBoost recommender recovers at least one known medium in the top 5 for 77.5% of strains, versus 37.2% for a taxonomic-popularity baseline and 36.6% for a global-popularity baseline. Higher is better.

3. Discussion

What this work establishes. Our framework demonstrates that (i) at BacDive scale (46K strains, 22.3K unique genomes), a multi-source feature fusion with 6,313 features per-genome trains stably in 5-fold family-grouped cross-validation; (ii) phenotype-targeted PLM embeddings, computed by HMM gating which proteins to embed and pooling within phenotype category, add modest but measurable improvement on regression targets (T_opt MAE improves by 2.4%, pH MAE by 1.0%, salt MAE by 1.1%); (iii) frozen mean pool does not deliver lift on the oxygen classification task, but LoRA fine-tuning over the same HMM-gated marker sequences does, raising fold-0 oxygen macro F1 to 0.945; (iv) phenotype-specific model selection yields a deployed hybrid predictor for 5,000 uncultured catalog genomes; and (v) a 5,000 genome GenomeSPOT benchmark shows stronger temperature and pH accuracy for this work on the same held-out manifest.

On the magnitude of improvement. Stated as percent reductions in error against the prior-work predictors whose splits and baselines are matched closely enough to support a direct comparison, the gains are substantial. On the same 5,000-genome family-heldout manifest, this work reduces optimum temperature MAE by 39% (4.39 → 2.67 °C), pH MAE by 23% (0.61 → 0.47), and salt MAE by 3% (1.98 → 1.92%) relative to GenomeSPOT. Against the strongest published BacDive baseline (Koblitz et al. [4], Pfam-only random forest on a 21K-strain corpus), this work reduces temperature MAE by approximately 9% (2.94 → 2.67 °C) on roughly twice the corpus. For oxygen classification on fold-0, LoRA fine-tuning of the protein language model on HMM-gated marker sequences raises four-class macro-F1 by 135% relative to the tabular head (0.402 → 0.945). For medium recommendation on a 5-fold family-heldout dry-lab split, the XGBoost recommender improves Hit@5 by 108% relative to the strongest popularity baseline (0.372 → 0.775). Comparisons to Li et al. [5] and Máša et al. [10] differ in task and cannot be reduced to a single percent number; they are listed in Table 5 for context.

What this work does not yet establish. We do not have wet-lab validation of any predicted cultivation conditions or media. The LoRA result is currently fold-0 only, so the oxygen improvement should be treated as a strong model selection signal rather than a final five-fold benchmark. We do not benchmark against Koblitz et al. on their exact published train/test split, only against a strong six-path pre-PTPE baseline on our own splits. A complete comparison against Koblitz, held-out phylum generalization following Li et al.'s methodology, and wet-lab evaluation of nominated medium/genome pairs remain priority follow-up items.

On the modest empirical lift. A 2.4% relative reduction in temperature MAE is small. Two contexts make it nonetheless meaningful. First, it is achieved on top of an already strong baseline that includes KEGG module completeness over 570 modules and 144 curated Pfam marker features, so PTPE adds additional signal to a strong baseline. Second, the same Pfam only random forest reported by Koblitz et al. [4] achieves MAE around 2.94 °C on temperature on a smaller 21K BacDive sample; our pre-PTPE baseline already beats this at 2.74 °C, and PTPE pushes it to 2.67 °C, a 10% improvement over the strongest published BacDive baseline. The GenomeSPOT run gives a second external anchor: on the 5,000 genome held-out subset, GenomeSPOT reaches 4.39 °C temperature MAE, 0.61 pH MAE, and 1.98% salt MAE, while our production heads reach 2.67 °C, 0.47, and 1.92% respectively.

On phylogenetic generalization. Family-grouped CV controls for trivial leakage because no family appears in both train and test, but it does not address the more demanding out-of-clade generalization that Li et al. emphasized. Our roughly 600 unique BacDive families distributed across roughly 30 phyla make leave-one-phylum-out evaluation possible; we leave its execution to follow-up work, with the caveat that 5 of 30 phyla have fewer than 50 BacDive strains and may not be statistically meaningful as held-out sets.

Relation to prior work. Our pipeline differs from Koblitz et al. [4] primarily in *feature fusion breadth* (six-paths vs. Pfam only), in *target type* (continuous regression for three of four targets, vs. uniformly binary), in *corpus scale* (46K vs. 21K), and in the *PTPE feature construction*. It differs from Li et al. [5] in scale (46K vs. 96 isolates) and in the inclusion of PLM embeddings, which Li et al. explicitly do not use. It differs from Máša et al. [10] in being a *genome to*

phenotype predictor rather than a *trait to medium* predictor: Máša requires already known organismal traits and predicts preferences for two specific media; our pipeline operates from genome alone and could in principle drive a larger medium recommender once the marker corpus release is built out. It differs from SpoMAG [6] and from the transformer based growth rate predictor of Oduwole et al. [7] in being multi-task and in the PTPE construction.

On the LoRA fine-tune. The fold-0 result is more nuanced than a blanket "LoRA wins." It does not replace the tabular model for temperature or pH, and the salt gain is not strong enough to trust without additional folds. It does, however, solve the most important failure of frozen PTPE: oxygen classification. The oxygen-only ablation is especially useful because it shows that a lower training loss is not automatically better. Oxygen only training optimizes a smaller loss and reaches 0.080 train loss, but its validation macro F1 is worse than all-task LoRA; the all-task checkpoint is therefore retained for production oxygen prediction.

Limitations. (i) BacDive is survivorship biased to organisms that have been cultivated at least once; the model is appropriate as a first cultivation attempt recommender, not as a guarantee that lineages with no close cultivated relatives will be recoverable. (ii) the protein language model we use is behind the current state of the art for PLMs and behind DNA foundation models such as Evo 2 [9]; we chose it for the well-characterized behaviour of its representations under downstream pooling. (iii) Our 48 marker panel is necessarily incomplete; targets that depend on protein families we did not include, for example anoxygenic phototrophy enzymes for special niche organisms, are likely undermodeled. (iv) The web deployed uncultured catalog is a prediction table, not an experimental validation set; the next scientific step is prospective culture testing.

4. Methods

Phenotype label assembly

We queried the BacDive v2 REST API (public, no authentication) across BacDive IDs 1–200,000 in batches of 100, retrieving 100,866 strain records. Phenotype labels were extracted from the BacDive schema path *Physiology and metabolism* → *growth* → *temperature/pH/halophily/oxygen tolerance*. Continuous targets used the midpoint of any literature-reported range when no single optimum was deposited; salinity was converted to % NaCl. The 4-class oxygen target was kept as {aerobe, facultative_anaerobe, microaerophile, anaerobe}. The labeled corpus is 46,029 strains with both a genome accession and at least one of the four targets non-null.

Genome retrieval and gene prediction

Genome accessions were joined to NCBI Datasets v2 with version suffix normalization (BacDive stores unversioned accessions; NCBI Datasets requires versioned) and explicit detection of empty zip responses. CDS prediction used pyrodigal v3.5 [13] in single genome `train` mode for complete genomes, which we benchmarked at 7× the throughput of `meta` mode with no observed accuracy loss on QC genomes. FASTA bytes were discarded after feature extraction; the pipeline is fully streaming and resumable via JSONL append logs. Of 22,301 dis-

tinct genome accessions in the corpus, 22,300 were successfully fetched and processed (one transient NCBI failure).

Composition / Pfam / KEGG features

Amino acid composition, codon usage (relative synonymous codon usage), and tetranucleotide frequencies were computed in pure Python. Pfam HMM scanning used pyhmmer v0.12 [12] against the 48 marker curated panel, with IDs validated against the InterPro DESC field. KOfam scanning used the relevant KOfam HMM library (~3,000 KEGG orthology profiles covering KEGG module relevant KOs). Per genome KO hits are deduplicated by `genome_accession` and then evaluated against the parsed KEGG module rule set, a Python AST based parser supporting nested AND/OR/parenthesized expressions, to yield 570 fractional completeness columns per-genome. Source files for each component are protected materials.

Curated marker panel

The 48 HMM marker panel was selected manually from Pfam-A and a small number of NCBIfam profiles, partitioned into 8 phenotype categories (temperature, pH, oxygen, salt, vitamin, nitrogen, carbon, special). The panel covers chaperonins, cold shock proteins, sodium/proton antiporters, terminal oxidases (aerobic + microaerobic), superoxide dismutases, hydrogenases, nitrogenases, vitamin cofactor biosynthesis enzymes, RuBisCO, and carbohydrate-active enzymes. An automated InterPro DESC verification step confirms each Pfam ID still resolves to the expected protein family in the current Pfam release. The categories and representative members above (and Table S1) describe the panel's composition; the complete list of Pfam/NCBIfam accessions, their exact per-category assignments, and the verification script are protected materials and are available for non-commercial academic evaluation under a research-use agreement.

Phenotype-targeted protein language model embeddings (PTPE)

For each predicted proteome, we scan the marker library with pyhmmer and retain significant hits. For each phenotype category, the matched proteins are encoded with a frozen protein language model (ESM-2) and summarised within the category to yield one category vector per genome; concatenation across the eight categories produces the per-genome PTPE feature vector. The significance threshold, sequence-length handling, the protein-language-model choice and its embedding dimensionality, the within-category pooling design, and the materialization code are withheld as protected material.

Multi-task XGBoost training

We use one XGBoost model per-target, trained with 5-fold GroupKFold on the BacDive `family` field with genus fallback. Continuous targets use squared error objective; the 4-class oxygen target uses softmax with per-fold class re encoding to handle absent classes in some folds. Tree depth, learning rate, estimator counts, and early-stopping settings were tuned on held-out folds; the exact hyperparameter values are protected materials. The pre-PTPE baseline was trained on 1,198 features and is the +PTPE comparator's exact ablation match; the

+PTPE model was trained on 6,313 features. Aggregate result artifacts for both are available on request.

LoRA fine-tune model and training

The fine-tune wraps a protein language model (ESM-2) with LoRA adapters [15] applied to the attention `query` and `value` projections in every layer. The base ESM-2 weights are frozen; the trainable parameters are the LoRA adapters plus four lightweight prediction heads (three regression heads and one four-class oxygen head). Per-protein vectors are pooled within category, either by mean pooling or by a learned multi-head attention pool with a learnable per-category query, and concatenated across the eight categories before the heads. Training uses AdamW with separate learning rates for the adapters and heads, warmup followed by cosine decay, gradient accumulation, bf16 autocast, and gradient checkpointing on the base model, with the masked multi-task loss described in Results. The exact adapter rank, scaling, dropout, learning rates, batch and accumulation settings, head dimensions, and the model and trainer source code are protected materials.

GenomeSPOT benchmark protocol

The external condition trait benchmark uses the same family-grouped held-out manifest. We selected a deterministic 5,000 unique genome subset with seed 20260520. One representative row was retained per-genome accession, prioritizing rows with more available condition labels. The subset contains 5,000 temperature labels, 933 pH labels, 779 salt labels, 2,653 oxygen labels, and 416 rows with all four condition labels.

GenomeSPOT was run from the upstream source tree using its bundled models. For each manifest row, the runner downloaded the genome FASTA from NCBI Datasets, predicted proteins with pyrodigal, wrote paired contig and protein FASTA files, invoked GenomeSPOT, and parsed the resulting prediction table. Existing GenomeSPOT prediction files are reused by accession, which prevents repeated inference for duplicate genomes when larger manifests are run. The exact command and result artifacts are available on request.

Data availability

The four cultivation-condition targets are derived from the BacDive v2 public database; genome assemblies are retrieved from NCBI Datasets v2; and the 5,000-genome uncultured catalog used for deployment is derived from publicly available GTDB metadata. An interactive demonstration interface for browsing the uncultured-genome predictions is publicly available at https://huggingface.co/spaces/miyuiu/microbe-model. The aggregate prediction tables and evaluation artifacts underlying the reported numbers (the pre-PTPE and +PTPE cross-validation results, the deployed hybrid catalog predictions, the GenomeSPOT 5,000-genome subset manifest and results, and the media-recommender dry-lab benchmark) are available from the corresponding author on reasonable request for non-commercial academic use.

5. Future Work

The completed system now includes frozen PTPE, fold-0 LoRA model selection, hybrid uncultured catalog predictions, a 5,000 genome GenomeSPOT comparator, and a deployed web UI. The remaining work is narrower and more experimental:

(i) Complete LoRA cross-validation. The fold-0 LoRA result is sufficient for the current hybrid model choice, but a publishable LoRA benchmark requires running the remaining four folds. The main question is whether oxygen macro F1 remains near 0.94 across families and whether the apparent salt gain survives fold variation.

(ii) Attention-pooled per-category genome encoder. A learned multi-head attention pool with a per-category learnable query can replace the mean pool described in Results. Fold-0 LoRA shows that marker sequence fine-tuning is useful for oxygen, so attention pooling is now a reasonable follow-up rather than a speculative architecture change.

(iii) Leave one phylum out generalization evaluation. The family-grouped CV reported here controls for trivial leakage but does not address the more demanding out-of-clade test that Li et al. emphasized. With roughly 600 unique BacDive families spanning roughly 30 phyla, leave-one-phylum-out evaluation is the next experiment after the LoRA result.

(iv) Prospective medium testing. The recommender has been applied to the 5,000 genome catalog, but the results are not yet experimentally validated. The next practical step is to nominate a small set of high-confidence candidate genome/medium pairs, prioritize anaerobic and marine media where the hybrid oxygen labels change the recommendation, and test those conditions in culture.

A successor manuscript reporting these items is planned for late 2026.

6. Conclusion

We have built and released the largest BacDive-anchored phenotype prediction pipeline to date and introduced phenotype-targeted protein language model embeddings (PTPE), a feature construction that uses HMM gating to select which proteins to embed with ESM-2 and category level mean pooling to produce a compact functional fingerprint of the genome. Integrated with five complementary feature paths and trained with 5-fold family-grouped cross-validation on 46,029 strains, the resulting multi-task model improves on the strongest published BacDive baseline (Koblitz et al. [4]) on all four cultivation condition targets, with PTPE itself contributing modest, target-dependent additional lift over the strongest pre-PTPE configuration. A 5,000 genome GenomeSPOT benchmark further shows that our production heads improve temperature and pH accuracy against an external genome-based condition predictor, with similar salt accuracy. We then show that LoRA fine-tuning over the same marker sequences is most valuable for oxygen classification, not for the continuous targets, and deploy the resulting hybrid predictor to a 5,000 genome uncultured catalog. The current system is therefore not a single model that wins everywhere; it is a model selection stack that uses the best validated head for each phenotype and exposes the source of each deployed prediction.

References

1. Steen, A. D. et al. (2019). *ISME J.* 13, 3126–3130. Earth's uncultured microbiota.
2. Lloyd, K. G. et al. (2018). *mSystems* 3, e00055-18. Phylogenetically novel uncultured cells.
3. Feldbauer, R., Schulz, F., Horn, M. & Rattei, T. (2015). *BMC Bioinformatics* 16(S14):S1. PICA.
4. Koblitz, J., Reimer, L. C., Pukall, R. & Overmann, J. (2025). *Communications Biology* 8, 897. BacDive Pfam-RF.
5. Li, Z., Selim, A. & Kuehn, S. (2023). bioRxiv 2023.06.30.547261. Carbon utilization regression.
6. Terra Machado, D. et al. (2025). *PeerJ* 13, e20232. SpoMAG sporulation predictor.
7. Oduwole, I., He, M., Jha, R., Hoarfrost, A., Steen, A. D. & Emrich, S. (2025). *Proc. BCB '25*. LookingGlass2 growth rate.
8. Oduwole, I. et al. (2025). bioRxiv 2025.08.18.670795. Functional genomic signatures of culturability.
9. Brixi, G. et al. (2026). *Nature* 652, 1349–1358. Evo 2.
10. Máša, P., Kliegr, T. & Joachimiak, M. P. (2025). *CSBJ* 27, 5194–5206. Explainable medium prediction.
11. Reimer, L. C. et al. (2022). *Nucleic Acids Research* 50, D741–D746. BacDive.
12. Larralde, M. (2023). *Bioinformatics* 39, btad214. pyhmmer.
13. Larralde, M. & Zeller, G. (2023). *J. Open Source Software* 7, 4296. pyrodigal.
14. Lin, Z. et al. (2023). *Science* 379, 1123–1130. ESM-2.
15. Hu, E. J. et al. (2021). ICLR 2022. LoRA.
16. Mistry, J. et al. (2021). *Nucleic Acids Research* 49, D412–D419. Pfam.
17. Kanehisa, M., Sato, Y. & Kawashima, M. (2022). *Protein Science* 31, 47–53. KEGG.
18. Chen, T. & Guestrin, C. (2016). *KDD '16*, 785–794. XGBoost.
19. Koblitz, J., Schomburg, D. & Neumann-Schaal, M. (2023). *Nucleic Acids Research* 51, D1531–D1538. MediaDive.
20. O'Leary, N. A. et al. (2024). *Nucleic Acids Research* 52, D40–D48. NCBI Datasets.