FUNCTIONAL COVERAGE FOR SELECTIVE MICROBIAL TRAIT PREDICTION UNDER TAXONOMIC SHIFT

ABSTRACT

Functional-pathway similarity may help identify reliable predictions for microbial lineages absent from a model's labelled training set. We evaluate this hypothesis on 15,002 genome accessions using taxonomic holdouts, a fixed protein-embedding classifier, and three complementary coverage coordinates. In the initial family-held-out analysis, KEGG proximity ranks correct predictions above errors on three pathogenicity/biosafety targets with mean AUROC 0.652, compared with 0.531 for protein-embedding proximity. We then examine whether this diagnostic signal improves selective prediction, regenerating accession-level predictions and quantifying uncertainty across families. Combining pathway coverage with classifier confidence yields a mean balanced-accuracy advantage of 0.069 over confidence alone at 50% retention, but the paired family-bootstrap 95% interval [-0.055, 0.202] includes zero. Data-provenance checks further show that all negative pathogenicity labels are biosafety-derived proxies and that a small number of records have inconsistent or placeholder taxonomy. Across 468 refitted taxonomic-holdout models, aggregate phenotype AUROCs reproduce to within 0.00074. Our results distinguish informative coverage geometry from established selection utility: functional proximity is a useful exploratory diagnostic in this setting, but a robust reliability improvement is not demonstrated. We provide prediction-level supplementary results and reproducible statistical checks.

1 INTRODUCTION

A microbial genome can suggest an organism's physiology, but labelled organisms are an uneven sample of microbial diversity. A predictor intended for unfamiliar lineages needs more than high average accuracy: users need to know which predictions warrant follow-up. An additional representation of functional content may help identify where a fixed phenotype model is reliable even when that representation is not itself the best predictive input.

Protein language models such as ESM-2 supply sequence-derived features (Lin et al., 2023). KEGG modules describe annotated pathway completeness (Kanehisa & Goto, 2000), while Bacformer contextualizes proteins within genomes (Wiatrak et al., 2025). These representations induce different notions of proximity. We ask whether similarity in pathway space complements a classifier's own confidence for selecting predictions under taxonomic shift. This is a question about error ranking and abstention, not a claim that pathway annotations reveal causal mechanisms.

Our evaluation has two components. A *taxonomic-shift analysis* measures phenotype discrimination under progressively coarser holdouts. A *coverage analysis* holds an ESM-2 logistic predictor fixed and compares fitting-reference similarity in embedding, taxonomic, and pathway spaces, followed by a selective-prediction comparison. We connect these components through an explicitly verified common dataset, while preserving their different evaluation partitions. A reproduction phase reconstructs the cached inputs, regenerates individual predictions, inspects label provenance, and estimates selection uncertainty with families as sampling units.

Three distinctions organize the contribution. First, phenotype-prediction AUROC is not AUROC for predicting whether an existing classifier is correct. Second, a positive retained-subset accuracy gain is not necessarily a reproducible or statistically clear improvement over ordinary model confidence.

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

Third, family-label separation is not equivalent to complete biological independence. Our findings support functional coverage as an exploratory diagnostic while identifying concrete limits to the original reliability interpretation.

2 Related work

Representations and labels. BacDive provides heterogeneous strain-level phenotype and cultivation records (Schober et al., 2025). ESM-2 (Lin et al., 2023), genome-context models such as Bacformer (Wiatrak et al., 2025), and functional inventories such as KEGG (Kanehisa & Goto, 2000) offer complementary inputs for phenotype analysis. We compare their proximity scores within specified evaluation populations rather than rank their universal ability to represent biology. Annotation-derived proxy labels must be distinguished from direct measurements.

Structured evaluation and selective prediction. Grouped cross-validation is appropriate when related examples violate independent-sample assumptions (Roberts et al., 2017), but it does not remove every form of leakage or dependence (Kapoor & Narayanan, 2023). Selective classification trades prediction coverage for lower conditional risk (Geifman & El-Yaniv, 2017). Trust Score already uses neighborhood structure to assess classifier reliability (Jiang et al., 2018); this work does not introduce distance-based confidence itself. Its empirical focus is the coverage coordinate and the validity of its evaluation in microbial data. Class-conditioned Trust Score and other stronger error-ranking baselines remain untested here.

3 Data, provenance, and audit scope

Shared base, different experiments. Both evaluation components use the same cached dataset. Reconstruction from the upstream ESM-2 features, KEGG module table, BacDive labels, and split table confirms exact agreement of row order, accessions, BacDive IDs, both feature matrices, and all 13 binarized label arrays. It contains 15,002 distinct accessions, including after version-suffix removal, 640-dimensional ESM-2 features, and 570-dimensional pathway-completeness vectors. The original family partition has 8,845 training, 3,105 validation, and 3,052 test rows across 596 stored family labels.

The coverage analysis fits on all labelled *non-test* rows, pooling the original training and validation partitions; test rows are excluded from fitting and standardization. Thus no unused validation partition remains for selecting the combined confidence policy. The taxonomic ladder independently redraws groups at each rank and does not reuse the coverage analysis's family-test partition. The Bacformer comparison restricts the shared base to 9,062 aligned, nonzero feature vectors. Each trait uses only rows with an eligible binary label, so its fitting and test counts differ from the total cohort size.

Label provenance. The focal targets are human pathogenicity, animal pathogenicity, and biosafety level at least two. Inspection of all 15,002 matching raw BacDive records found that **all 7,769 human and 8,008 animal negative labels** lack an explicit negative pathogenicity annotation and instead arise from the parser's BSL-1 fallback. Test negatives number 1,663 and 1,701, respectively; test positives number 105 and 76. Thus these targets operationalize annotated positives versus a biosafety-derived negative proxy. BSL-1 is not experimental proof of non-pathogenicity, and the three focal outcomes are coupled by construction. Missing values outside these rules remain excluded, not silently converted to negatives. The remaining ten traits are listed in Appendix A.

Identity and taxonomy checks. No duplicate accession or BacDive ID was found in the cache, and no accession crossed the stored partitions. Each stored family label is confined to one partition. However, ten genus labels appear in both fitting and test populations, affecting 11 test records. Ten of these records use the placeholder family label "Not assigned to family"; one reflects discordant family assignments for the same genus. Therefore, string-level family separation alone does not establish strict unseen-genus or fully harmonized unseen-family evaluation. There are no nonfinite or all-zero ESM/KEGG rows. ESM vectors are unique, whereas KEGG contains 206 repeated rows; one identical KEGG vector occurs across partitions for two distinct accessions. Equal pathway profiles are not by themselves duplicate genomes.

Audit boundaries. Checks are at accession, cached-feature, metadata, and downstream-prediction levels. They do not verify assembly-sequence identity across distinct accession systems, near-homology separation, pretrained-model corpus overlap, every genome download, or the biological truth of annotations. KEGG generation code uses module definitions and KOfam hits, not phenotype labels. The audit did not repeat upstream annotation or foundation-model embedding generation. Exact historical dependency locks and all original checkpoint revisions are unavailable.

4 PREDICTORS, COVERAGE SCORES, AND UNCERTAINTY

4.1 FIXED PHENOTYPE PREDICTOR AND REFERENCE GEOMETRY

For each binary trait t, the coverage study fits a standardized, ℓ_2-regularized logistic model to the eligible non-test ESM-2 rows. Verified settings are `C=1`, `class_weight=balanced`, the default L-BFGS solver, tolerance 10^{-4}, and at most 2,000 iterations. Standardization is fitted on eligible non-test rows only. The probability $p_t(g)$ is thresholded at 0.5; class weighting is not a probability-calibration procedure. Correctness is $a_t(g) = \mathbf{1}\{\widehat{y}_t(g) = y_t(g)\}$.

For a representation z and the same trait-specific fitting reference set $\mathcal{R}_t$, the coverage score is

$$c_{z,t}(g) = \max_{h \in \mathcal{R}_t} \langle u(z(g)), u(z(h)) \rangle, \qquad u(x) = x / \max(\|x\|_2, \mathbf{1}\{\|x\|_2 = 0\}). \tag{1}$$

Equivalently, the implementation replaces zero normalization denominators by one. No zero ESM/KEGG vectors were present in this cache. Proximity uses unstandardized, unit-normalized representations, distinct from the standardized predictive inputs. Taxonomic coverage scores 3, 2, 1, or 0 for shared order, class, phylum, or no shared rank with a fitting record. No null values were found for these metadata ranks, although placeholder strings require separate scrutiny.

We evaluate $\mathrm{AUROC}(c_{z,t}, a_t)$, which measures ranking of correct versus incorrect predictions. This differs from phenotype AUROC, $\mathrm{AUROC}(p_t, y_t)$. Correctness-ranking scores can still reflect class prevalence, trait difficulty, or clade-specific annotation patterns.

4.2 RETROSPECTIVE SELECTION

The model-confidence baseline is $m_t(g) = |p_t(g) - 0.5|$. The combined score is

$$s_t(g) = R(m_t(g)) + R(c_{\mu,t}(g)), \tag{2}$$

where $R(x) = \mathrm{argsort}(\mathrm{argsort}(x))/(n-1)$ within a trait's test population. Ties are not averaged in the original code. Multiplying this sum by one half does not change its ordering. The procedure uses unlabelled test-score distributions for ranking, not test labels to calculate the score, but is a batch-relative, retrospective policy rather than an independently calibrated deployment threshold.

For retention $q \in \{1, .9, .8, .7, .6, .5\}$, the implementation keeps the first $\max(\mathrm{round}(qn), 10)$ rows after descending score sort and evaluates balanced accuracy if both classes remain. Define $\Delta\,\mathrm{BA}_t(q)$ as retained-subset balanced accuracy minus full-test balanced accuracy. These gains do not improve rejected predictions and do not imply full-population accuracy improvements. The audit also checks an average-rank tie convention as a sensitivity analysis.

Exploratory evaluation. The combined-score analysis has no documented validation-only selection protocol or verifiable pre-test commitment to its rule and operating point. We therefore treat the 50% retention comparison as exploratory. Regenerating predictions and adding bootstrap intervals tests numerical reproducibility and sampling sensitivity; it does not provide an untouched confirmatory evaluation or correct for adaptive choices made during development.

4.3 REGENERATION AND FAMILY-AWARE UNCERTAINTY

We regenerated predictions with pinned NumPy 1.26.4, pandas 2.3.3, scikit-learn 1.6.1, SciPy 1.13.1, and PyArrow 21.0.0. These versions match the identified local analysis packages but are not a recovered lock of every historical run. We save accession-keyed probabilities, labels, classifications, and coverage scores. All 13 main logistic fits converged within the configured limit without

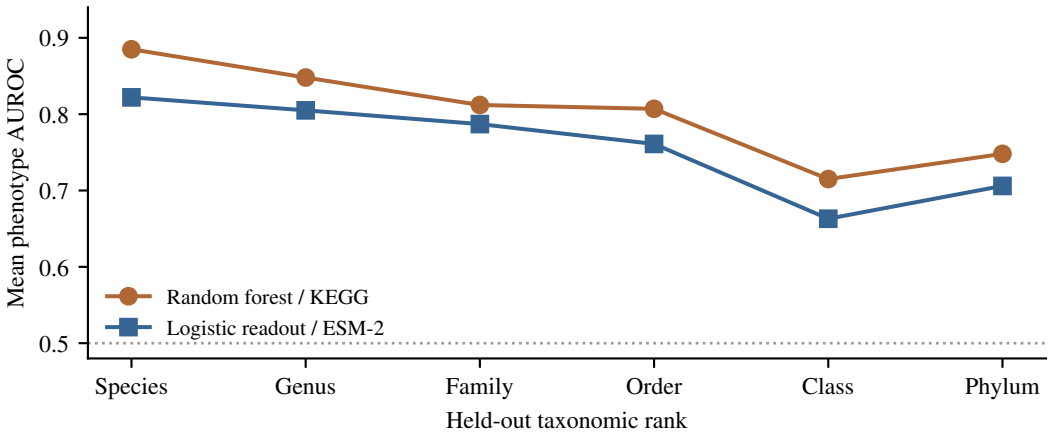

Figure 1: Historical phenotype AUROCs under taxonomic-rank holdout. Both models lose discrimination from species to class holdout and remain above chance. Phylum scores rise again, so the trend is not globally monotonic. All 468 underlying configurations were rerun during the audit; the maximum aggregate discrepancy is 0.00074.

convergence warnings. Repeating the three focal fits in the same environment produced identical probabilities; agreement with historical aggregate tables is assessed separately.

The ladder regeneration uses two methods, six taxonomic ranks, three resamples, and 13 eligible traits: 468 fits. It uses 120 depth-10 random-forest trees with seed zero, or standardized unweighted logistic regression with `C=1` and 1,000 iterations. Group selection uses seeds 0–2 and a 25% group holdout, preserving the original group order. Held-group manifests and individual probabilities are retained; no exception is silently discarded in this rerun.

For the focal selection comparison, we draw 2,000 bootstrap samples of the 165 represented test families, using the same family draws across the three traits. All labelled rows within a sampled family are retained with its multiplicity. We recompute score ranks and top-half selection within each resample and calculate the mean paired difference between combined and confidence-only balanced accuracy. Models remain fixed: these intervals describe test-population sampling uncertainty, not training uncertainty or prospective generalization. Resamples lacking a required class are undefined; their number is reported instead of substituting a score. The historical coverage intervals instead resample individual test rows 1,000 times and are not family-aware.

5 Results

5.1 The taxonomic ladder is approximately reproducible

Figure 1 shows the historical ladder retained for provenance. KEGG random-forest AUROC decreases from 0.885 at species holdout to 0.715 at class holdout, and the ESM-2 readout decreases from 0.822 to 0.663. The audit regenerates every fit and reproduces the aggregate ladder to within 0.00074 AUROC. This is approximate numerical reproduction, not bitwise identity or independent confirmation on a new dataset.

Coarser holdouts alter the test population, label distribution, and fitting-set size; they do not manipulate evolutionary distance alone. Placeholder taxonomy can also become a sampled group. No accession overlaps the fitting and test populations of the regenerated ladder partitions, but this does not establish sequence-level independence. The ladder motivates reliability assessment, not a claim that every model fails to generalize.

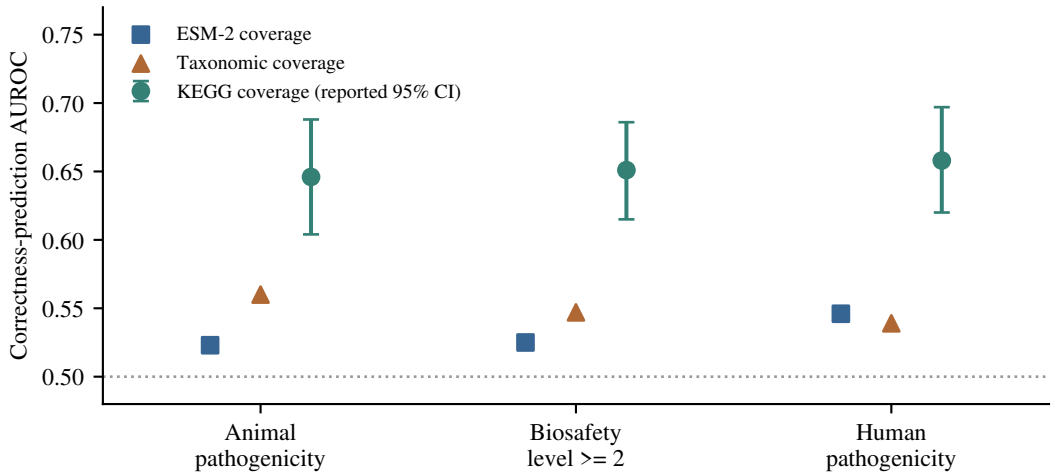

Figure 2: Historical correctness-ranking results for the fixed ESM-2 readout. Error bars are the source study's individual-row bootstrap intervals, not family-bootstrap intervals or paired differences. "Pathogenicity" targets contrast annotated positives with BSL-1-derived negative proxies. These qualifications limit a mechanistic interpretation.

5.2 Functional coverage ranks correctness, with qualifications

The historical coverage summary gives mean focal-label correctness AUROC 0.652 for KEGG, 0.531 for ESM-2, and 0.548 for taxonomy. Our rerun gives 0.651, 0.532, and 0.549, respectively. Figure 2 retains the historical per-trait values and row-bootstrap intervals. Cached summaries and regenerated metrics are kept separate, rather than replacing discrepancies silently.

A matched 9,062-row Bacformer subset historically yields focal-label coverage AUROC 0.683, versus 0.633 for KEGG, 0.523 for ESM-2, and 0.563 for taxonomy. The audit verifies the subset by accession and nonzero-vector filtering and regenerates all 13 comparison fits. Regenerated focal-label Bacformer coverage AUROC is 0.683; per-genome probabilities and coordinate scores are saved separately. Architectural, pretraining, and extraction differences prevent attributing this result solely to pooling. The ten other traits have heterogeneous coverage results; functional proximity is not universally useful.

5.3 Selection gains are exploratory and uncertain

The historical top-half gains are 0.079 for combined ranks, 0.014 for model confidence, and 0.054 for pathway coverage on the focal group (Figure 3). Regenerated combined and confidence gains are 0.088 and 0.019, giving a paired advantage of 0.069. Different numerical predictions near thresholds and different rank orderings can affect top-half metrics; repeatability within one environment does not establish exact reproduction of an older aggregate. The historical per-genome probabilities were not saved, so a direct probability-by-probability comparison to the original run is unavailable.

The group 95% interval is [-0.055, 0.202]. It includes zero, as do all three per-trait intervals. A second complete 13-trait rerun in the existing Anaconda environment gives a mean advantage of 0.074 with interval [-0.056, 0.198], again including zero. Only 1885 joint draws have defined metrics for all traits. This substantial uncertainty and the class-support failures argue against claiming a statistically established benefit across families. Excluding shared-genus and placeholder-family test records is a post-hoc sensitivity analysis, not a new validation set; its results are reported in Appendix B.

5.4 Interpretation of the negative controls

The original data also distinguish a useful coverage coordinate from a useful predictive input or sampling rule. KEGG readout substitution reportedly reduces focal-trait phenotype AUROC, and

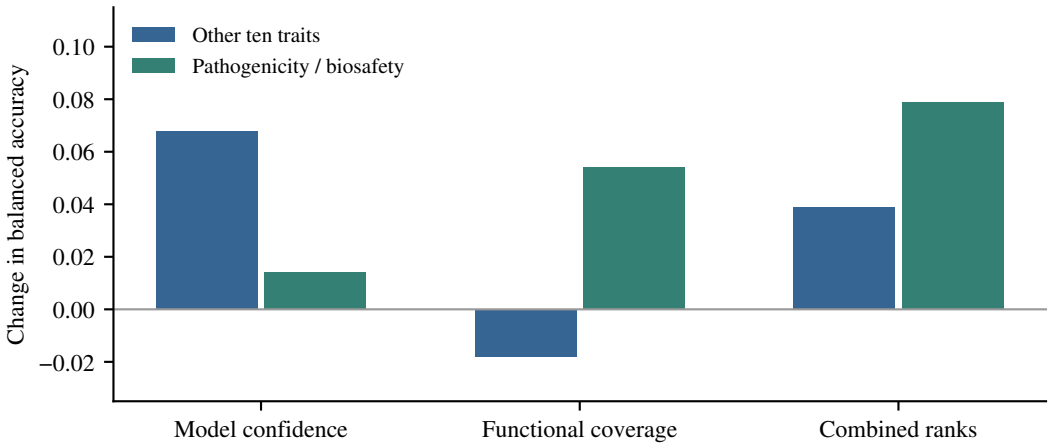

Figure 3: Historical balanced-accuracy changes at 50% retention. Functional coverage is not uniformly beneficial: ordinary confidence performs better for the other ten traits. These original point estimates are not statistical evidence of superiority. Figure 4 supplies uncertainty for the regenerated paired comparison.

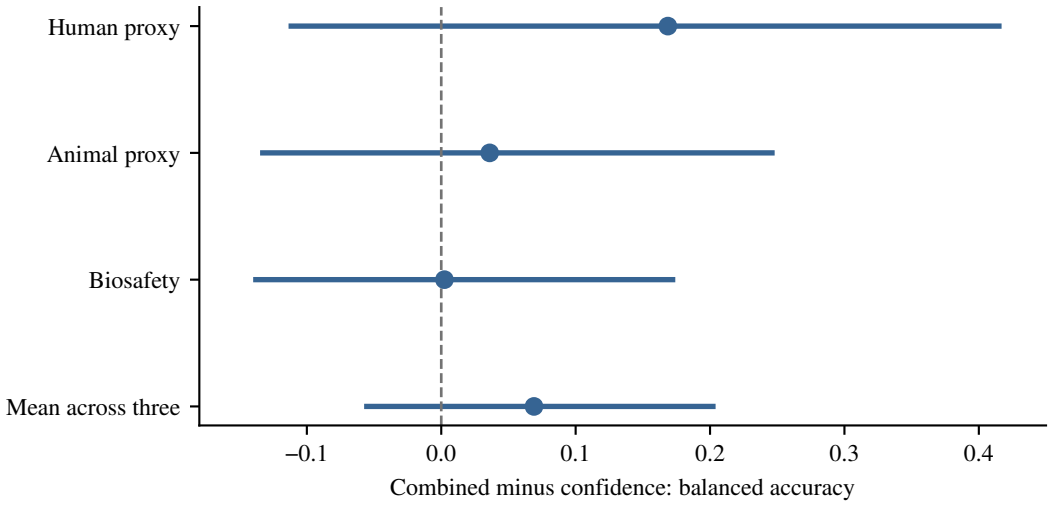

Figure 4: Regenerated combined-minus-confidence balanced accuracy at 50% retention, with paired family-bootstrap percentile intervals. Every interval includes zero. Resampling is joint across focal traits, with fixed fitted models and ranks recomputed within each draw. The group interval uses 1885 valid draws out of 2,000. Undefined draws lack a required class after selection.

a large advantage of machinery-based sampling over random sampling did not survive the source study's multi-seed analysis. Those secondary experiments are traceable to saved tables but were not independently rerun in this audit. They do not justify claims about pretraining scaling laws. The current contribution is a narrower diagnostic comparison and its reproducibility limits.

6 Limitations and conclusion

Functional coverage provides descriptive information about correctness on the evaluated labels. However, the proxy construction of every negative pathogenicity label couples the focal outcomes to biosafety, the apparent selection advantage has wide family-level uncertainty, and the development history does not establish an untouched test. No deployment guarantee, confirmed causal mechanism, or prospectively validated cultivation result follows.

The data and reproduction checks establish a common cached base, exact feature and label reconstruction, accession-disjoint partitions, and traceable individual predictions. Residual limitations include placeholder taxonomy, unverified sequence-level homology, incomplete metadata for the initial runs, and numerical sensitivity of selective metrics. The strongest supported conclusion is therefore a distinction between informative coverage geometry and demonstrated reliability improvement.

A stronger confirmatory study would use harmonized taxonomy, independently supported labels across additional biological functions, an untouched external evaluation population, and a fixed comparison against strong confidence baselines. It would jointly report retention, class support, family-balanced effects, and uncertainty. These controls are necessary to determine whether the observed coverage signal supports dependable selection beyond this benchmark.

AI USE STATEMENT

Writing and polishing assistance. Generative AI tools assisted drafting, revising, organizing, and polishing the original studies and this combined manuscript, including clarity, grammar, and terminology.

Retrieval and discovery. AI-assisted search supported finding related work and locating publication records for checking selected titles, authors, and bibliographic details. Retrieved sources were consulted rather than relying solely on generated summaries.

Implementation and research assistance. Generative AI assisted experiment code and data processing in the initial analyses. It also assisted framing and narrowing claims, methodological criticism, implementation of prediction regeneration and statistical checks, interpretation of audit findings, figures, and LaTeX preparation. AI-assisted implementations were exercised through feature/label reconstruction assertions, repeated fits, saved-prediction consistency checks, and comparisons with initial result tables. These checks do not establish the biological validity of annotations or independent generalization. New downstream computations used existing data; no synthetic biological observations or wet-lab measurements were generated. The author is responsible for the final text, claims, and artifacts, including material produced with AI assistance.

ETHICS STATEMENT

The analyses use existing microbial features and annotations and report no new laboratory experiments. Biosafety-derived negative labels must not be interpreted as proof of non-pathogenicity. Predictions and abstention scores are not organism-safety certifications and do not replace institutional biosafety review. Selective reporting can disadvantage understudied clades; subgroup retention and error reporting are important. Data, model, and code redistribution require compliance with applicable licenses and access terms.

REPRODUCIBILITY STATEMENT

The accompanying anonymous supplementary archive contains an accession-level cohort manifest, regenerated probabilities for the 13-trait coverage analysis, the Bacformer-subset comparisons and all 468 ladder configurations, recorded split groups, initial-versus-regenerated metric summaries, and paired family-bootstrap draws. A standalone script reproduces the key summary comparisons and family-bootstrap calculation from the supplied predictions. Appendix B describes the verification procedure and its boundaries. Raw genome sequences, upstream ESM/KEGG/Bacformer feature matrices, and pretrained checkpoints are not redistributed in this archive; the statistical checks can be reproduced from it, but model retraining and upstream sequence processing require those separate inputs. No external repository link is needed to inspect the supplied numerical evidence.

REFERENCES

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, 2018.

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

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

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.

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, Julia Koblitz, Joaquim Sardà Carbasse, et al. BacDive in 2025: The core database for prokaryotic strain data. *Nucleic Acids Research*, 53(D1):D748–D756, 2025. doi: 10.1093/nar/gkae959.

Maciej Wiatrak, Ramón Viñas, Maria Ntemourtsidou, Adam M. Dinan, David C. Abelson, Divya Arora, Maria Brbić, Aaron Weimann, and R. Andrés Floto. A contextualised protein language model reveals the functional syntax of bacterial evolution. *bioRxiv*, 2025. doi: 10.1101/2025.07.20.665723.

A HISTORICAL PER-TRAIT COVERAGE RESULTS

These values reproduce the original coverage table for provenance; the intervals use individual-row resampling. They are not the new family-bootstrap selection intervals. The descriptive grouping must not be interpreted as a verified mechanistic taxonomy.

Table 1: Historical correctness-prediction AUROC for the fixed ESM-2 readout.

Trait	ESM-2	Taxonomy	KEGG (row-bootstrap 95% CI)
Alkaliphile	.487	.497	.500 [.464, .536]
Motility	.495	.495	.501 [.472, .533]
Cytochrome oxidase	.541	.528	.510 [.469, .552]
Thermophile	.531	.513	.521 [.454, .583]
Halophile	.508	.513	.543 [.499, .591]
Pigmentation	.547	.515	.546 [.502, .588]
Catalase	.513	.489	.563 [.525, .602]
Aerobe	.548	.530	.580 [.551, .608]
Sporulation	.556	.573	.605 [.547, .660]
Gram stain	.711	.547	.726 [.666, .783]
Animal pathogenicity proxy	.523	.560	.646 [.604, .688]
Biosafety level ≥ 2	.525	.547	.651 [.615, .686]
Human pathogenicity proxy	.546	.539	.658 [.620, .697]

B AUDIT DETAILS AND UNRESOLVED VALIDATION

Dataset lineage. The common dataset is the accession intersection of the upstream ESM-2 and KEGG tables, linked to traits and partitions by BacDive ID. The upstream tables have unique join keys, and every cached label matches reconstruction under the original binarization rules. Input hashes and reconstruction checks are supplied under anonymous artifact names. The ladder uses this full base with resampled taxonomic groups; the coverage analysis uses the stored family partition; the Bacformer comparison further restricts eligibility to nonzero aligned features. These evaluations are not paired across components simply because they share inputs.

Taxonomy and sensitivity. Stored family labels do not span partitions. Nonetheless, 11 held-out rows share one of ten genus labels with fitting rows. Ten use an unassigned-family placeholder and one uses discordant family classification. After excluding shared-genus and placeholder-family test records without refitting the models, combined-versus-confidence balanced-accuracy differences are 0.142 for the human proxy, 0.036 for the animal proxy, and -0.010 for biosafety in the pinned-environment rerun. This post-hoc sensitivity does not establish significance or replace taxonomy harmonization. Genome-sequence near-duplicates and cross-accession assembly equivalence were not checked.

Labels and uncertainty. All negative pathogenicity labels in this cohort are BSL-1 fallbacks. A sensitivity restricted to explicit pathogenicity annotations would have no explicit negatives and therefore cannot support ordinary binary AUROC on these stored annotations. The paired bootstrap samples the union of 165 test families jointly across the focal traits with seed 20260916. Models remain fixed, while reference-score ranks and retention are recalculated within each draw. Per-trait valid draws and intervals, and the joint-valid subset, are retained. Invalid draws are reported, not imputed. The fraction of nonpositive bootstrap draws is not presented as a formal hypothesis-test p-value.

Reproduction is not historical identity. Numerical differences occur relative to saved aggregates despite matching the currently identified package versions. The source studies did not save per-genome probabilities, so the exact origin of historical differences cannot be proven by comparing predictions. The new three focal fits repeat identically within the pinned environment; this does not

establish that historical libraries, BLAS implementations, solver stopping points, or row-ranking ties were identical. The audit retains both old and regenerated numbers. The ladder's 468 fits use the original rank/replicate/trait grid; the audit records all eligible configurations rather than silently ignoring errors.

Scope of statistical inference. The combined rule and the 50% operating point do not have a verifiable pre-test commitment, so the reported intervals should not be interpreted as correcting for model or analysis selection. The bootstrap is conditional on the fitted models and the available test families; it does not incorporate uncertainty in biological annotations, pretraining exposure, or deployment populations. A confirmatory experiment must lock the policy independently of its final evaluation data and use outcomes whose provenance supports the intended biological interpretation.

C Complete source-study visual archive

This appendix preserves all five original figures from the coverage study and all three from the taxonomic-shift study, including every panel, alongside the four figures in the main text. All source panels and plotted quantities are retained, with typography redrawn in a consistent Times-style serif family; the captions below supply the audited interpretation. They are historical visualizations, not additional independent experiments. Original plot titles and the labels "machinery"/"compositional" describe the source analyses and are not evidence of a verified mechanism. In particular, claims of universal collapse, causal representation effects, or established selective-prediction superiority in original titles are not conclusions of this manuscript. The three focal targets remain coupled by biosafety-derived pathogenicity negatives.

C.1 Coverage coordinates across all 13 traits

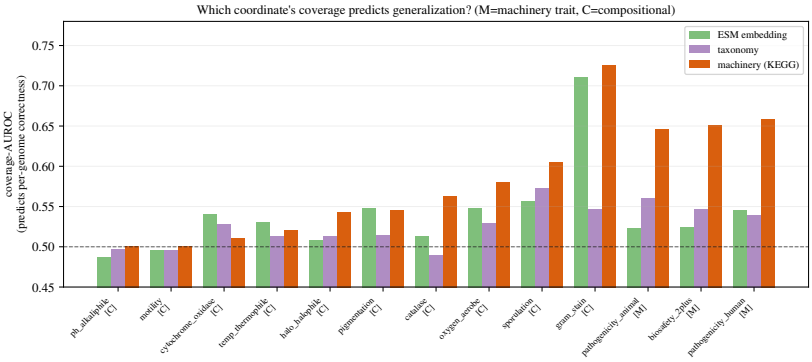

Figure 5: Original coverage-study Figure 1, preserved in full. Bars show correctness-prediction AUROC for proximity in ESM-2, taxonomy, and KEGG spaces across all 13 traits, rather than phenotype-prediction AUROC. The original [M] group comprises the pathogenicity proxies and biosafety target; [C] denotes the other ten traits. This full panel complements the focal-trait summary in Figure 2 and the values in Appendix A. It does not establish that functional coverage is uniformly useful or free of phylogenetic confounding.

C.2 Genome-context coverage comparison

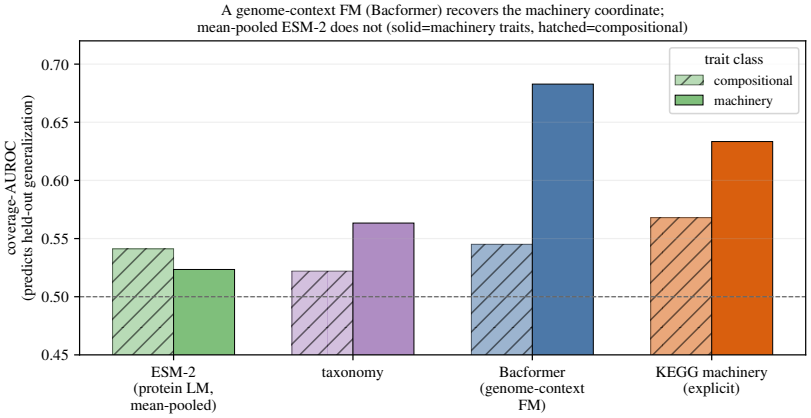

Figure 6: Original coverage-study Figure 3. Historical mean correctness-ranking AUROC on the matched 9,062-genome Bacformer subset, with the ESM-2 phenotype readout held fixed. Solid bars average the focal proxy/biosafety group; hatched bars average the other traits. The original title's claim about recovering a machinery coordinate is a descriptive source interpretation, not a causal attribution to genomic context or pooling. The audit regenerated all 13 matched-subset comparisons; the plotted historical values are retained rather than replaced by regenerated estimates.

C.3 COMPLETE MULTI-SEED LEARNING CURVES

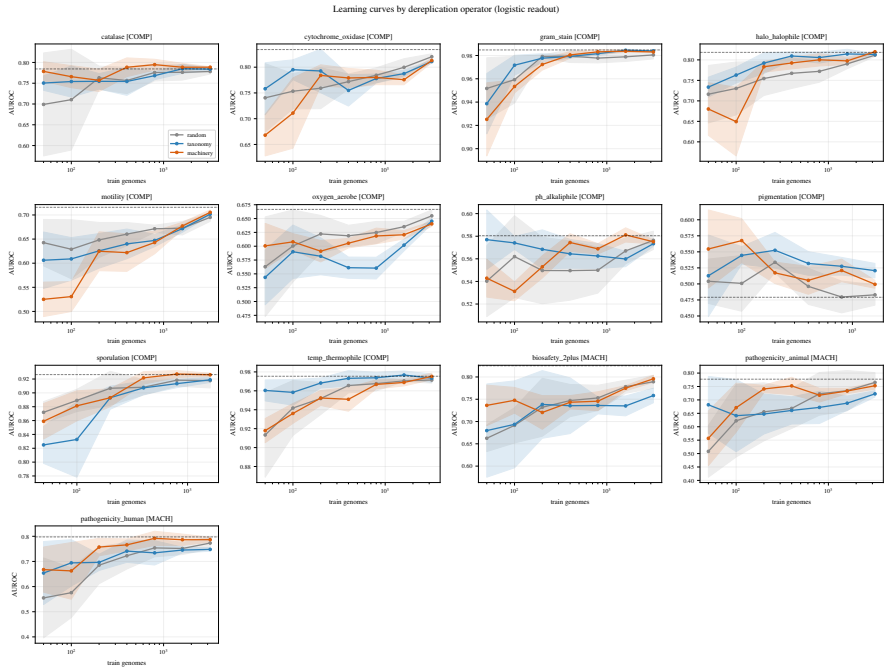

Figure 7: Original coverage-study Figure 2, redrawn as vector graphics with all 13 learning-curve panels and their original plotted values. The horizontal axis is labelled fitting-genome budget and the vertical axis is phenotype AUROC. Grey, blue, and orange denote random, taxonomy-based, and pathway-based sampling; dashed lines mark full-data AUROC. Bands are mean ± one sample standard deviation over five seeds, not 95% confidence intervals, as verified in the plotting code. These secondary experiments were not rerun in the downstream audit. They do not establish a general scaling law or a consistent advantage of pathway-based sampling over random sampling.

C.4 SPECIES-TO-FAMILY COMPARISONS ACROSS FOUR MODELS

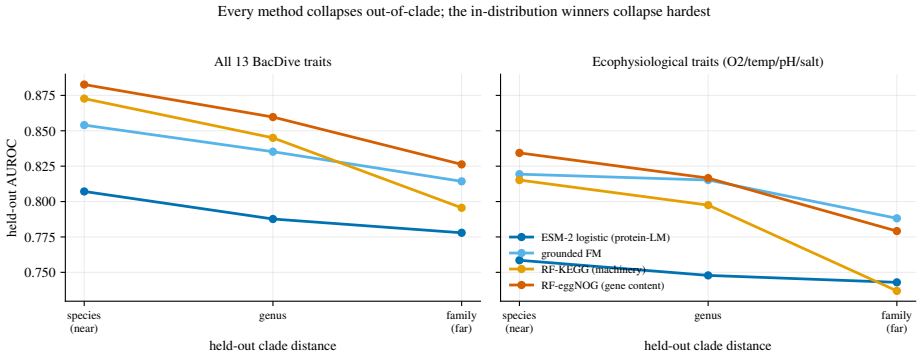

Figure 8: Original taxonomic-shift-study Figure 1, preserving both the all-trait panel and the ecophysiological-subset panel. Historical phenotype AUROCs compare ESM-2 logistic regression, the source grounded encoder, RF-KEGG, and RF-eggNOG on the source benchmark's species/genus/family partitions. These are distinct from the resampled six-rank ladder and are not paired with the coverage study's test population. These four-method results were not regenerated by the 468-fit ladder audit. The original title's "every method collapses" wording means a decline among the displayed configurations, not chance-level performance or a claim about all microbial models. No bootstrap intervals are shown.

C.5 Representation substitution as predictive input

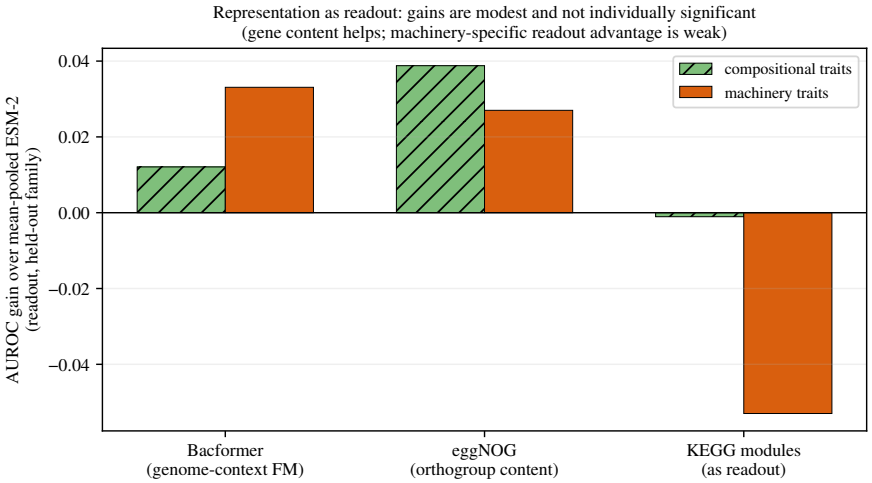

Figure 9: Original coverage-study Figure 4. Historical phenotype-AUROC differences when substituting Bacformer, eggNOG, or KEGG predictive inputs for mean-pooled ESM-2, grouped by the source trait categories. These are changes in the phenotype predictor, not correctness-ranking AUROCs or selection gains. Comparisons use the eligible matched populations for each representation, not necessarily one common cohort across all three input types. No intervals are drawn, and these secondary readout experiments were not independently rerun in the audit. Their point estimates should not be interpreted as established representation superiority.

C.6 Full selective-prediction curves

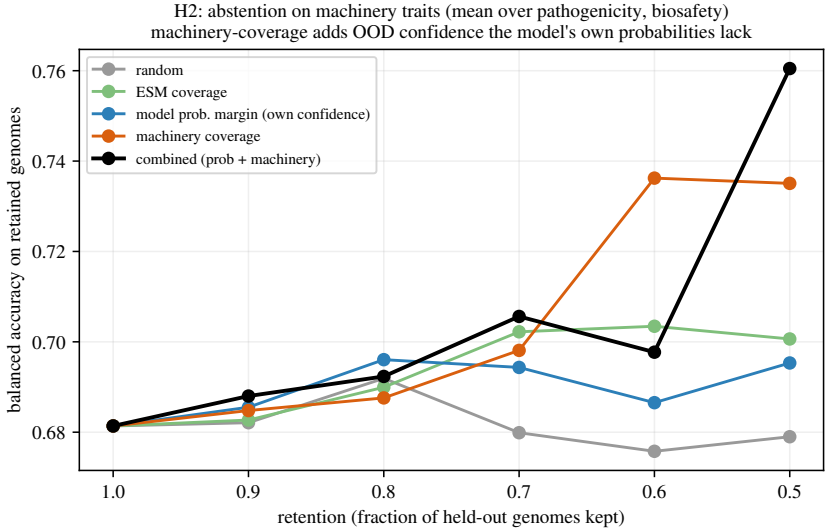

Figure 10: Original coverage-study Figure 5. Historical mean balanced accuracy on retained predictions for the focal pathogenicity-proxy/biosafety group, from full retention to one-half retention. All five original policies are retained: random selection, ESM proximity, probability margin, pathway proximity, and combined ranks. Unlike Figure 3, this plot shows the full retention sweep and absolute retained-subset accuracy rather than the top-half change. The source title's confidence claim is exploratory: the regenerated combined-minus-confidence family-bootstrap interval includes zero (Figure 4). These curves neither establish probability calibration nor guarantee prospective reliability.

C.7 Original six-rank taxonomic ladder

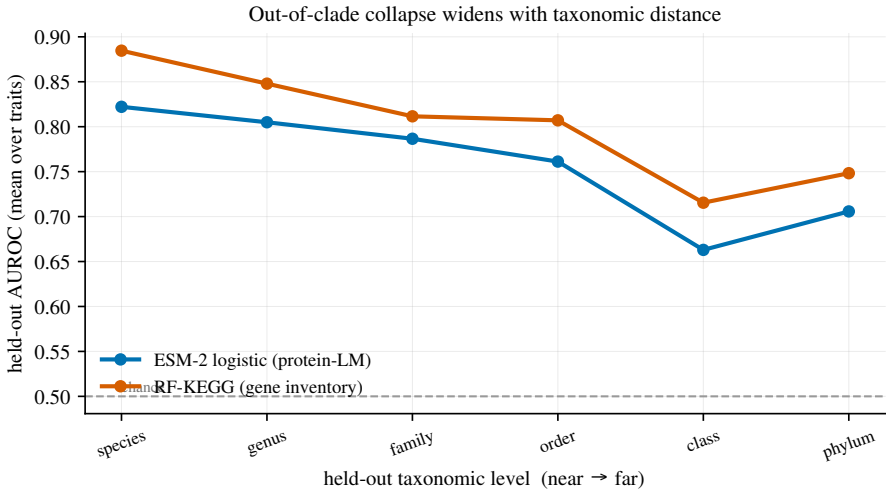

Figure 11: Original taxonomic-shift-study Figure 2. This is the source visualization of the ladder summarized in Figure 1, not an additional experiment. Historical mean phenotype AUROCs are averaged over the source traits and three group-holdout resamples at each rank. Both displayed models remain above chance; performance rises again at phylum holdout, so the trend is not globally monotonic. The audit reran all 468 configurations and found maximum aggregate discrepancy 0.00074. Changes in population, label composition, and fitting-set size limit a causal interpretation in terms of evolutionary distance alone.

C.8 Matched-subset Evo2 comparison

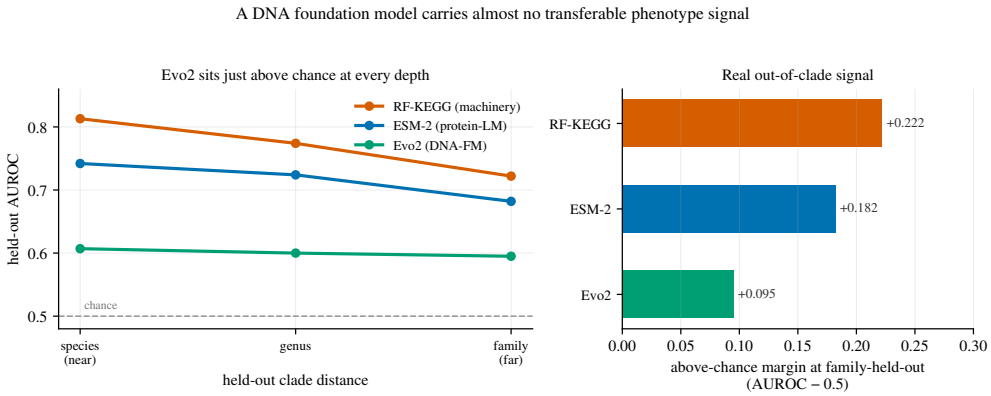

Figure 12: Original taxonomic-shift-study Figure 3, preserving both panels on the shared 675-genome comparison subset. Left: historical phenotype AUROC at species, genus, and family holdout. Right: the family-held-out AUROC margin above 0.5. The Evo2 result applies to the source study's mean-pooled DNA-window extraction and fitted readout, not to all DNA foundation models or all ways of using Evo2. Its family AUROC of about 0.595 is above chance; the original title's broad transferable-signal claim is not established. This subset experiment was not rerun in the audit, has no bootstrap intervals here, and must not be conflated with the full-cohort ladder or the 9,062-genome Bacformer comparison.