Seven Silent Failure Modes in Genome-Mining Pipelines

Ivan Meić[1], Buğra Alptekin Sarı[1], Steve Muller[1], Fedor Chishchin[1], Olamide Isreal[1], Gabriel Richman[1,*]

[1] Omic Inc., Seattle, Washington, USA

* Correspondence: gabe@omic.ai

ORCIDs: F.C. 0009-0008-5679-7617; O.I. 0009-0008-6590-5384. [ORCIDs NEEDED for I.M., B.A.S., S.M. and G.R.]

Abstract

Background: Genome-mining pipelines chain annotation, function prediction, cluster detection, substrate prediction and structure assembly, each stage reporting a confidence score. A benchmark does not measure whether a stage's output reveals when that stage is wrong on new data — the property deciding whether a pipeline can be trusted without ground truth, which genome mining never has.

Results: Re-deriving every reported number in a working pipeline from the files that produced it, we found seven stages where wrong output was indistinguishable from correct, five detectable only by recomputation. Three patterns recur. Metrics measure something adjacent to what they claim: an annotation-quality rule counting "hypothetical protein" scored one genome 0.0% where 54.1% of products were empty; a combined CLEAN/DeepFRI recovery figure built across mismatched denominators was 14.5% over the 2,455 proteins both scored, not 15-20% as first reported. Evaluations cannot see their own confound: a classifier scoring one protein 0.988 then 0.000 at unchanged ROC-AUC 0.995 reproduced neither value once rebuilt with documented classes and seeds, and the same confound cut DeepBGC's classification accuracy from 70.4% to 49.5% on unseen genera. Provenance records describe the invocation, not the experiment: a seven-method benchmark spanning BLAST/Pfam, Foldseek, CLEAN, ESM-2, ProTrek and DeepFRI graded all seven wrong against a ground truth from an unverified code comment, when four were correct.

Conclusions: A benchmark score reported without a companion test separating "the model is correct" from "the evaluation cannot tell" leaves a pipeline unfalsifiable in use. We release the checks so the audit can be repeated elsewhere.

Background

A genome-mining pipeline is a chain of inferences. A genome is annotated; coding sequences are assigned functions; clusters of biosynthetic genes are detected; substrates are predicted for the relevant domains; a structure is assembled; and the product is scored for novelty or activity. Each link has been benchmarked, often carefully, and most emit a confidence score.

Benchmarks measure whether a stage is right on the data it was tested on. They do not measure whether, when that stage is wrong on new data, anything in its output reveals it. These are different properties, and the second is what determines whether a pipeline can be trusted without ground truth — which is the situation genome mining is always in, since the compounds it is used to find are by construction uncharacterised.

The distinction is sharpest for what we will call *errors of commission*. A method that returns "unknown" for a protein it cannot handle is behaving safely: the failure is legible and a downstream step can route around it. A method that returns a specific, confidently scored, wrong function is behaving unsafely, because its output is formally identical to a correct one. The same distinction applies above the level of individual predictions. A metric that silently measures something adjacent to what it claims, a pipeline that degrades instead of crashing when misconfigured, two tools whose combined recovery is reported but whose overlap is not — all produce plausible numbers that no inspection of those numbers can invalidate.

Individual failures of this kind are documented. Shortcut learning — a model reaching high measured performance by keying on a property that correlates with the label rather than causes it — is well characterised in machine learning [1], and its canonical example, a pneumonia classifier responding to a hospital-specific scanner token, is structurally identical to one of the cases we report. Reviews of machine-learning-assisted enzyme and protein engineering [18, 19] treat the same generalisation failure as a known hazard of the field rather than a novel observation. Spurious and propagated genome annotation is likewise a known problem, and OMArk was built because "no existing methods estimate the extent of

spurious annotation, which is common in publicly available genomes" [2]. Reviews of computational BGC discovery [22] and of how genomic background shapes measured cluster diversity [23] survey the methods and their scope. What has not been assembled is an account of how these failures compose along a multi-stage discovery pipeline: which stage each sits in, which are visible in that stage's own output, and what it costs to find the ones that are not.

This paper reports seven such failures, found while auditing a working natural-product genome-mining system of our own. We did not design experiments to elicit them. All seven surfaced from a single discipline: re-deriving every number in a manuscript from the files that produced it, rather than from the summary that reported it. Five were invisible in the artefact and were caught only by recomputation — the last of them invisible to an earlier audit of ours as well, which read the run's own record of its arguments and drew the opposite conclusion.

We demonstrate them on biosynthetic gene cluster (BGC) mining in archaea, a setting chosen because it maximises the distance between the training distributions of the available tools and the data they are applied to. The tools concerned are actively maintained and the reference data actively curated — antiSMASH is now at 8.0 [16] and MIBiG at 4.0 [17] — so the versions audited here are a snapshot, and the failure modes rather than the version numbers are the subject. That choice makes failures larger and easier to characterise; we make no claim here that archaeal biology is the source of any of them, and where we can test that question, as in failure mode 1, the answer is that it is not.

Relationship to a companion manuscript. Several of the measurements below — the MA_4560 seven-method benchmark, whose ground truth is corrected in both manuscripts, the CLEAN/DeepFRI recovery comparison, the eight-genome annotation survey, the biosynthetic-classifier confound and the DeepSEMS reimplementation — also appear in a companion manuscript by the same authors, which reads them as evidence about archaeal biology and organises them into a model of why archaeal natural products evade detection. The experiments were performed once and the numbers are the same in both. What differs is what is asked of them: that manuscript asks what archaea contain, this one asks what the tools report when they are wrong. The population-scale generalisation studies in failure modes 1, 2 and 4, and the probes of GECCO and DeepBGC, appear only here, and no argument in this paper depends on the archaeal interpretation.

Results

Failure 1 — An annotation-quality metric that counts vocabulary, not annotation

The fraction of a proteome annotated as "hypothetical" is a standard proxy for how much of an organism's biology is unknown. The usual implementation, ours included, tests whether a coding sequence's `/product` field contains one of a set of keywords (*hypothetical*, *predicted protein*, *uncharacterized*, *unknown function*). Applied to eight archaeal genomes this gives a mean of 40.2% (range 13.1-56.4%; Extended Data Table 1). Applying the identical rule to eight bacterial genomes spanning curation intensity gives a bacterial mean of 25.0% — a comparison that is wrong, for the reason this failure mode describes.

Synechocystis sp. PCC 6803 scored 0.0%, but 1,712 of its 3,167 coding sequences (54.1%) have an **empty** `/product` field: the genome is not well annotated, it is unannotated in a different dialect, and the keyword rule measures which words a submitter chose for their unknowns rather than how many unknowns exist. Counting a blank product as unannotated, the bacterial mean rises to 32.1% and the archaea–bacteria gap falls from 15.2 points to 8.1. No archaeal genome in our set has any blank products, so only the comparison group was distorted, in the direction that would have flattered an archaeal claim.

Even the corrected comparison does not separate the domains: ranges overlap almost entirely (archaea 13.1–56.4%, bacteria 2.5–59.0%), Mann-Whitney gives $U = 24$ against a critical value of 13 at $n = 8$ per group, and *S. aureus* NCTC 8325 (59.0%) is less completely annotated than every archaeon we examined. Grouping bacterial genomes by curation intensity instead of domain is more informative: intensively curated model organisms average 12.0% hypothetical, while well-studied and environmental tiers average 44.2% and 44.0% — closer to the archaeal figure than to the curated bacterial one. An annotation gap of this size looks like a property of under-curated genomes, not of any lineage, and only a control reveals that.

How often does this happen, and is our own control representative? A single genome shows the metric *can* fail, not how often. Over 3,570 complete genomes sampled with a recorded seed from the 66,714 RefSeq and 90,043 GenBank assemblies available (Extended Data Table 2), the failure tracks annotation provenance almost exactly: 1 of 3,163 carries any blank product (all PGAP-annotated, which always writes one, by construction), against 34 of 393 do (8.65%) among genomes retaining submitter annotation — a two-orders-of-magnitude difference. Yet across all 3,570 genomes, **none** reproduces the pattern that makes *Synechocystis* striking (keyword score under 1% with true unannotated fraction over 20%); the largest disagreement anywhere is 22.1 percentage points, and disagreement exceeded 1 percentage point for 0.89% of GenBank genomes and none of RefSeq. The metric is unsafe for submitter-annotated records and sound for PGAP-annotated ones, not generally wrong — and this bears on our own eight-genome bacterial control, which is one outlier (Synechocystis, roughly 1-in-3,570) carrying most of its 25.0%->32.1% correction. The corrected comparison is still the right one to report, since it moves against our own archaeal claim, but it rests on a small, one-outlier-dominated control. Separately, 429 of the 2,000 sam-

pled GenBank assemblies (21.5%) carry no CDS features at all, and "percent hypothetical" is undefined for them; our own script silently excluded them from its denominator until we checked why the parsed count was short.

Dedicated annotation-quality tools address a related but different problem: OMArk was built because "no existing methods estimate the extent of spurious annotation" [2] — completeness of a gene repertoire, not whether the genes present are real. This failure is a third thing: a proxy in routine use that measures the vocabulary a submitter chose for their unknowns, and disagrees with itself across annotation dialects by more than the effect it is used to detect.

Failure 2 — A combined recovery figure that was never measured, and a Jaccard index that could not mean what we said

We applied two enzyme-function predictors to the 2,476 proteins annotated as hypothetical in *M. acetivorans*: CLEAN, which predicts EC numbers from sequence by contrastive learning, and DeepFRI, run here in sequence-only mode. At a 0.5 confidence threshold, CLEAN returns high-confidence predictions for 93 proteins (3.8%); DeepFRI returns high-confidence EC assignments for 54 and molecular-function GO terms for 166 of the 1,081 hypothetical proteins in its own input set (5.0% and 15.4%).

We first summarised this as 15–20% of the unknown proteome illuminated. That cannot be computed from those three percentages: they have two different denominators, and the overlap between the two sets was never measured. Cross-tabulated over the 2,455 proteins both tools actually scored (Fig. 1b), CLEAN recovers 93, DeepFRI recovers 274, the union is 356 (14.5%), and the **intersection is 11 proteins** — a Jaccard index of 0.03.

A Jaccard index of 0.03 does not on its own indicate that the two predictors disagree; that reading requires the observed overlap to fall *below* what the marginal coverage rates predict. It does not: independent coverage predicts 10.4 shared proteins against 93 and 274 calls; 11 were observed (Fisher exact odds ratio 1.07, 95% CI 0.59–2.09, $p = 0.87$), and the Jaccard expected under independence, 0.029 against the 0.031 observed, is close to what was seen. A Jaccard of 0.03 is simply what two tools calling 3.8% and 11.2% of the same proteins produce whether or not they agree, and we withdraw the reading we placed on it. What survives is the original arithmetic error — a combined figure cannot be built from percentages with different denominators, and a union needs its intersection measured, not assumed — and the corrected union of 14.5% is lower than our first 15–20%.

Does this generalise? Repeating the measurement on 58 independently sampled genomes (20 bacteria, 20 archaea, 18 fungi; Extended Data Table 3) with the same independence test per genome, the prokaryotic result looks emphatic read naively — zero intersection in all 20 bacterial genomes and 19 of 20 archaeal ones — but coverage there is small enough that independence predicts fewer than one shared protein in **every one of the 44 genomes** where none was observed (bacteria: 0 observed against 1.50 expected; archaea 1 against 4.11, Mantel-Haenszel odds ratio 0.23). Fungi are the only domain with enough coverage to test agreement at all, and there the tools agree substantially more than chance: 136 shared proteins against 51.3 expected, odds ratio of 4.46 (95% CI 3.54–5.63) — the opposite direction from what we first reported, for reasons we did not design this study to identify.

The defensible claim is narrow: two tools each covering a few percent of a prokaryotic proteome will produce a near-zero Jaccard regardless of agreement, so that statistic is not evidence of corroboration or its absence; where coverage permits a test, these two tools agree more than chance. The failure this section documents is upstream of any of that — combining percentages across denominators, and reporting a union without its intersection.

Failure 3 — Seven methods graded against a ground truth that was wrong

Seven methods applied to a single protein returned what read as unanimous failure, and we first took four of them to be confidently and identically wrong in a way that would indicate a systematic limit of annotation transfer. The benchmark's ground truth was itself wrong, and correcting it inverts that reading.

The protein was misidentified. We had taken MA_4560 to be TfuA, the *M. acetivorans* thioamidation enzyme Nayak and colleagues characterised in 2017 — an assignment that came from a comment in our own analysis code, never checked against a primary source. It is wrong. The real TfuA is MA_0164 (UniProt Q8TUA9), Pfam PF07812 (TfuA core), the protein crystallised as PDB 6XP8, paired with YcaO at MA_0165 — matching Nayak and colleagues' own numbering. MA_4560 (UniProt Q8THF7) is an unrelated, uncharacterised protein: its own Pfam domains are PF08350 (FilR1_middle) and PF31122 (FilR1_N), the winged helix-turn-helix FilR1 family, with no TfuA-family domain in UniProt, InterPro, or the GenBank record.

Every measurement in Table 1 stands; only the grading inverts. Graded against what MA_4560 actually is, four of the seven were right and the other three correctly declined to call it. Three claims are withdrawn in full: that four methods returned a wrong, confidently-supported assignment (they were right); that the case shows a structure/function dissociation (there is none — the fold is a regulator and so, on all evidence, is the function); and that agreement between the four is uninformative (here it agreed with the truth). We no longer have a verified instance of fold-convergent annotation-transfer failure in this paper.

What survives is the evaluation failure. A locus identity taken from an internal code comment, never traced to a primary source, became the reference against which seven methods were scored, producing a clean, publishable-looking re-

sult in which every method failed. Nothing in the pipeline — not a confidence score, an ensemble, agreement between methods — was positioned to catch a wrong ground truth; the methods' agreement read as corroboration of a shared blind spot rather than as a signal the key was wrong. The check that resolved it took one query each to UniProt and InterPro — the same check Failure 4 turns on: verify a small number of independently-known answers *at the source*, not from the pipeline's own record.

This is one case, and a failure of our own evaluation rather than of any tool; it does not establish how often misattributed ground truth survives into published benchmarks.

Table 1. Seven methods on MA_4560 (UniProt Q8THF7), regraded.

Approach	Method	Prediction returned	Correct
Sequence homology	BLAST / Pfam	PF08350 (FilR1_middle), transcriptional regulator	Yes
Structural search	Foldseek	963 AFDB50 hits, all transcriptional regulators	Yes
Enzyme prediction	CLEAN	score 0.000, no enzyme function detected	Yes (abstains)
Protein language model	ESM-2 (650M)	rank 1903 / 4540, not discriminative	Yes (abstains)
Tri-modal contrastive	ProTrek	"winged helix-turn-helix transcription factor" (−0.035) ranked first; "thioamide bond formation" (−0.054) third	Yes
Structure–function	DeepFRI	DNA binding (GO:0003677), confidence 0.669	Yes
Purpose-trained classifier	this work (ESM-2 + logistic regression)	score 0.000, rank 3718 / 4540	Yes (abstains)

Failure 4 — A classifier whose headline score was a property of its training set, not of the protein

Believing off-the-shelf methods had failed (Failure 3), we trained a purpose-built classifier — logistic regression on mean-pooled ESM-2 embeddings separating biosynthetic from non-biosynthetic proteins — and reported two versions. The first, with 17,056 MIBiG proteins as positives and archaeal housekeeping proteins as negatives, scored MA_4560 at 0.988. The second added 7,205 bacterial non-biosynthetic negatives and reported the score collapsing to 0.000 at rank 3718/4540, with cross-validated performance unchanged (ROC-AUC 0.9946, F1 0.9844) — read as shortcut learning, positives bacterial and negatives archaeal, so the model separated domains rather than function.

Neither reported number reproduces, and the diagnosis does not survive either. The first version's output was never retained; no script trains the second as described; the positive downsampling called an unseeded `np.random.choice`, so 0.988 was never exactly reproducible even when the original code ran. We rebuilt the experiment with documented rules and explicit seeds across three arms, 25 seeds each (Extended Data Table 4) — a reconstruction, not a reproduction, since no documented rule recovers the published class counts.

The reported collapse is not there. The arm that adds bacterial non-biosynthetic negatives — the change credited with exposing the confound — scores MA_4560 between 0.793 and 0.9997 across all 25 seeds and never approaches zero. Score collapse appears only in a third arm, where what changes is the *positive* class: restricting it to proteins MIBiG annotates as biosynthetic (`gene_kind`), which excludes 14,079 BGC-resident coding sequences carrying no `gene_kind` annotation at all. Every coding sequence in a MIBiG record had been labelled biosynthetic with no functional filter, silently absorbing 2,780 transport, 1,945 regulatory, 874 other and 55 resistance proteins alongside the 14,079 unannotated ones — the class meant "located in a BGC," not "biosynthetic," and was reported as though it meant the latter.

The score was never a statement about the protein. Scoring all 4,540 *M. acetivorans* proteins per seed, MA_4560 sits between the 50th and 91st percentile in every arm, while its raw score ranges from 0.0016 to 0.9996 — because the whole proteome's median score moves with it, from 0.0001 in one arm and up to 0.614 in another. Its rank tells a more consistent story: across all 75 fits its best rank is 398 of 4,540, never near the top, consistent with Failure 3 — it was not a missed enzyme to begin with, so the near-zero score in the label-corrected arm is the right answer, for a different reason than originally claimed. (One shared, uncorrected contamination: the archaeal negatives come from the same eight genomes the ranking proteome is drawn from, so 196 of those 4,540 proteins were also trained on as negatives; excluding them moves MA_4560's rank by at most three places.) The published 0.988 is reached by 3 of 25 seeds in the first arm;

the published 0.000 is reached by none of the 25 fits in the second arm — the one that makes the change the published collapse was attributed to — and by all 25 fits in the third, where the change is to the positive class instead.

What survives. All three arms reach ROC-AUC 0.988–0.999 while disagreeing about one protein by three orders of magnitude — cross-validated performance measured on the training population is uninformative here, which is the finding that does hold up. What does not survive is our specific diagnosis: the artefact sits in the definition of the positive class and in probability calibration on an out-of-distribution proteome, not in a taxonomic confound in the negatives. Shortcut learning [1] remains the right frame; we no longer claim to know which shortcut this model took.

Does the same confound appear in classifiers we did not build? We probed two published BGC detectors, GECCO and DeepBGC, on their own shipped models against populations their own published training data excluded (GECCO's 1,880 MIBiG v2.0 training BGCs, prokaryotic only; DeepBGC's 1,814 training BGCs, including 267 fungal entries). Cluster detection held up regardless of population for GECCO, including on phyla absent from training (73.3–80.0% across four groups, all pairwise comparisons non-significant), DeepBGC's detection recall was flat for its novel fungal genera (98.2% covered against 98.5% novel) and showed only a borderline dip for novel bacterial genera (95.5% vs 88.0%, Fisher $p = 0.051$). Classification did not hold up: GECCO's accuracy declines with taxonomic distance from training in a significant trend (84.9% on training BGCs to 67.6% on eukaryotic, Cochran-Armitage $z = -2.41$, $p = 0.016$), and DeepBGC's classification collapses on novel bacterial genera (70.4% to 49.5%, $p = 0.0035$) while its deliberately-covered novel fungal genera show no decline (67.7% against 63.2% on covered fungal genera, $p = 0.62$). Neither tool's own confidence score would have caught this: GECCO's cluster probability is higher on correct calls (0.998 against 0.992, Mann-Whitney $p = 0.0004$) but by less than the score's own resolution — the minimum confidence among its wrong calls is 0.92 — and DeepBGC's classifier probability discriminates more usefully (0.873 correct against 0.722 incorrect) but most of its misses return no class prediction at all rather than a confident wrong one ($n = 26$ confident-and-wrong). The pattern tracks what each tool's authors actually trained on, not a fixed rule such as "prokaryotic tools fail on eukaryotes": DeepBGC held up on the population it deliberately covered and collapsed on the one it did not, for the same structural reason as our own classifier above — the training population, not the target property, is what a model learns to key on, and neither cross-validation nor a tool's own reported score reveals it. 420 BGCs (GECCO) and 472 BGCs (DeepBGC) were run for this probe (Methods).

Failure 5 — A correlation and a retrieval result that disagree, both correct

The premise behind embedding-based structure prediction is that similar sequences make similar molecules. Measured directly on 399 MIBiG BGCs having both a mean-pooled ESM-C embedding and a reference structure (79,401 pairs, embedding cosine against Morgan fingerprint Tanimoto; Fig. 3a), the relationship is essentially absent across all pairs: Pearson $r = 0.037$, $r^2 = 0.001$, Spearman $\rho = 0.005$. Restricting to the 47,143 pairs at cosine ≥ 0.9 does not rescue it — mean Tanimoto there is 0.126, with only 0.7% above 0.4.

The confidence interval on that r is itself an instance of this paper's subject. We first reported 95% CI 0.029–0.044 from a bootstrap resampling pairs — but the 79,401 pairs are every unordered pair among 399 BGCs, so each BGC appears in 398 of them and they are not independent observations. Resampling BGCs instead (a cluster bootstrap, 2,000 replicates) gives **95% CI −0.005 to 0.081**, 5.7 times wider and no longer excluding zero. Nothing about the finding changes — $r = 0.037$ was never large enough to support the premise — but the published interval claimed a precision the design could not deliver, in the direction that made the number look more definite.

On the same embeddings, nearest-neighbour retrieval works: each BGC's closest neighbour has mean Tanimoto 0.278 against 0.124 for a random pair, a 2.24-fold lift, with 19.8% of BGCs above 0.4 (Fig. 3b). Part of that lift is database redundancy: for 21 of the 399 BGCs (5.3%) the nearest neighbour is a different MIBiG entry encoding the same compound (Tanimoto 1.0); excluding them, mean nearest-neighbour Tanimoto falls to 0.237 and the lift to 1.92-fold. We found this only because the figure made the spike visible — the argument of this paper, applied to our own result.

Both results are correct and support opposite summaries: $r \approx 0.04$ implies embeddings are useless for structure, which would not predict that template matching on the same embeddings reaches 0.376 mean Tanimoto; the retrieval lift implies a usable relationship, which would not predict $r^2 = 0.001$. The relationship is real but confined to the extreme tail of the similarity distribution, where mean nearest-neighbour cosine of 0.993 and 47,143 of 79,401 pairs above 0.9 show that mean-pooled embeddings compress into a narrow cone with little dynamic range — so a value-based statistic like Pearson r has almost no sensitivity even where rank-based retrieval still works. We flag, without being able to reconcile, that our own internal records once carried a different figure ($r = 0.229$) for this relationship, attributed to an experiment whose output no longer exists.

Failure 6 — A benchmark that does not reproduce

We reimplemented DeepSEMS, a published structure predictor, following the authors' protocol, and evaluated it on our MIBiG validation set: mean Tanimoto 0.266 across 647 BGCs, against the "up to 0.6" the preprint reported. We

are not in a position to adjudicate this — our reimplementation may differ from the authors' in ways we cannot see, we did not contact them before writing this, and the work has since been peer-reviewed [9] under a version we have not rechecked against our reimplementation.

What the observation supports is narrower: the distance between a reported figure and an independent reimplementation, 0.266 against 0.6, is larger here than the distance between any two methods in Table 2, and a comparison assembled from reported numbers cannot resolve differences smaller than that — which includes most comparisons in this area, ours included. We release the reimplementation so the discrepancy can be located rather than argued about.

Table 2. Substrate and structure prediction in context.

Metric	This work	antiSMASH 7.0	PRISM 4	DeepSEMS (reimplemented)
Substrate top-1 accuracy	74.7% ± 0.3	~85%	—	—
Substrate top-3 accuracy	84.1% ± 0.4	—	—	—
Structure Tanimoto (mean)	0.374	~0.5	0.5–0.7	0.266
Structure Tanimoto (cyclic)	0.489	—	—	—
Open-vocabulary prediction	Yes	No	No	No

Our own substrate predictor reaches 74.7% ± 0.3 top-1 and 84.1% ± 0.4 top-3 over 54 substrate classes, on a held-out split of 1,191 domains from 8,744 bacterial adenylation domains, as the mean of five seeded reruns; it does not outperform antiSMASH and its structure prediction is below PRISM 4; dedicated substrate predictors such as PARAS [20] and annotation platforms such as Gaia [21] report stronger performance on their own bacterial benchmarks. We include it as the system in which the other six failures were found, not as a competitive claim — and these are not the numbers we first reported, for the reason given in Failure 7. Its confidence scores are well calibrated on a separate 1,107-domain comparison set (overall accuracy 72.2%): 86.7% accurate above 0.9 confidence (n = 360), 71.7% between 0.7 and 0.9 (n = 590), 40.8% between 0.5 and 0.7 (n = 152) — a different evaluation from the 74.7% headline, and the two should not be read as one experiment.

Failure 7 — A provenance record that was accurate about the invocation and wrong about the experiment

We found this one last, by accident, while re-running an unrelated ablation. Our substrate predictor was reported at 76.4% top-1 and 85.5% top-3 over 60 substrate classes. Re-running the recorded configuration — temperature 0.07, minimum class count 10 — on this repository's own inputs returns 54 classes, 5,701 training domains and 74.7% top-1 over five seeds. A gap of that shape is not a seed effect.

It was 1,993 self-generated pseudo-labels, accepted at a 0.95 confidence floor and added to the 8,744 labelled domains. Three lines of evidence agree: the original training log records loading 10,737 rows, exactly 8,744 plus 1,993; the file at the path the training script reads today holds 8,744 rows with a modification time three days *before* the pseudo-labels were generated, an older copy restored with its timestamp preserved; and of the stored model's 60 classes, thirteen have fewer than the minimum ten labelled examples and **six have none at all**, carrying 99, 100, 100, 17, 100 and 99 pseudo-labelled examples instead — exactly the gap between the 54 classes the labelled data supports and the 60 reported.

The failure is what happened when we audited this the first time. An earlier audit of ours asked exactly this question and answered it wrongly, concluding no pseudo-labels were used, on the evidence that the run's own recorded arguments show `pseudo_labels: null`. That field is accurate — it records the `--pseudo-labels` command-line flag, which was not set — but the data entered through a hard-coded embeddings path that bypasses the flag entirely. The record was correct about the invocation and wrong about the experiment, and an audit that read the record rather than the data inverted a true claim into a false one, which then propagated into a manuscript and a status document. A run's self-description is therefore a claim to be checked against the data it names, not evidence of what the run did.

Two further claims did not survive the same rerun. The stored model's 76.4% / 85.5% / 60 classes figures are unreproducible, since the file that produced them is gone; we report the reproducible 74.7% ± 0.3 / 84.1% ± 0.4 over 54 classes throughout. And an ablation cited for choosing 34 binding-pocket positions over the 10 Stachelhaus positions (+1.1 percentage points) does not reproduce either: over five seeds, 34 positions give 74.74% ± 0.32 and the 10 give 74.41% ± 0.46, a paired difference of +0.34 points with a standard de-

viation of 0.50 and an inconsistent sign across seeds — not distinguishable at this sample size.

What each failure cost to find

Five of the seven (1, 2, 4, 5, 7) produced output that was internally consistent and indistinguishable from correct; none was found by inspecting a result, only by recomputing a number from upstream files — failure 7 only after an earlier audit had trusted a stored argument instead of doing that. Failure 3 was found by returning to a primary source for an answer the pipeline had only held second-hand, and it exposed our own key rather than any tool's. Failure 6 required an independent reimplementation. Neither an authoritative external record nor a reimplementation is available in the normal case, which is exactly the case genome mining addresses.

Discussion

Confidence scores do not report the failures that matter

Every method in Table 1 emits a confidence score, and none carried the information that mattered: DeepFRI's 0.669 for DNA binding is an ordinary value returned for what we now know was a correct call, not a failure — the score cannot see that the problem was our label, not its prediction. The Failure 4 classifier reported ROC-AUC 0.995 while measuring the wrong thing, and all three rebuilt arms report 0.988–0.999 while disagreeing about one protein by three orders of magnitude. A confidence score is a statement about the distribution a model was fitted to; it cannot express that an input is unlike its training data or that its label is confounded — exactly the conditions genome mining operates under. The fix is a reporting convention, not a new algorithm: alongside a benchmark score, state what would distinguish "the model is right" from "the evaluation cannot tell." For Failure 4 that check — verifying a small number of independently-known answers — was cheap and decisive where cross-validation on thousands of proteins was not; GECCO's and DeepBGC's own confidence scores were similarly uninformative about their generalisation collapse.

Aggregation hides the structure of a result

Failures 1, 2 and 5 share a mechanism: a summary statistic that is correct arithmetic and misleading description. A mean over eight genomes hid that one was measured by a rule that did not apply to it; a combined recovery figure hid that its components had different denominators and an unmeasured overlap; a correlation over 79,401 pairs hid that the relationship exists only among the closest few hundred of them. In each case the disaggregated view was available in data we already had and cost minutes to produce — the aggregate simply looked reasonable, so nothing prompted a second look. This argues for disaggregating pipeline-evaluation statistics by default, particularly across heterogeneous inputs, rather than on suspicion.

Fail-open stages turn misconfiguration into quiet accuracy loss

One defect does not belong on the list above, since it is a bug rather than a property of a method, but it illustrates the same principle in plumbing. Our structure pipeline reaches mean Tanimoto 0.374 on 453 MIBiG NRPS BGCs against an oracle ceiling of 0.416; re-running it during this audit returned 0.339, because a valid but differently-keyed embedding cache matched nothing, silently emptied the BGC embedding dictionary, and routed the 269 BGCs that should have used the highest-performing template path to weaker ones — no error, no warning, a plausible number. The difference (0.035 Tanimoto) is smaller than the spread between published methods in this area, so neither a reader nor we could detect it from output alone; we found it by comparing method-selection counts between two runs. The general form is cheap to prevent: a stage that proceeds without a missing input converts a configuration error into silent accuracy loss, and the fix is a precondition that refuses to run on zero matches rather than documentation.

Errors of commission deserve separate accounting

Benchmarks conventionally report accuracy, treating a wrong answer and an abstention as equivalent losses. For a pipeline stage feeding downstream inference they are not: an abstention propagates as missing data, which later stages can detect, while a confident wrong answer propagates as data, which they cannot. We would report abstention rates alongside error rates for any method run inside a pipeline. Failure 3 is a caution about that accounting as much as an argument for it: the same partition reads as four unsafe and three safe failures under one key and as four correct calls and three abstentions under the right one — separating commission from omission inherits whatever is wrong with the ground truth underneath it.

Limitations

The seven are not a survey: they are what one audit of our own pipeline turned up while working for a different purpose, and five of the seven come from our own analyses. We cannot state their prevalence — that needs auditing pipelines we did not build, and the checks are released for that purpose. Failures 5 and 6 concern tools we did not write (ESM-C, DeepSeMS); failures 1, 2, 3, 4 and 7 concern how we measured, labelled and reported ourselves, and in Failure 3's case the tool outputs were right and our own grading was wrong. We think failures 1 and 2 are latent in any study using those conventions, failure 3 in any benchmark whose key is not traced to a primary source, failure 4 in any cross-domain classifier, and failure 7 in any pipeline where a training input can be written to a fixed path — but we cannot demonstrate that from one pipeline.

Several findings are narrower than they might read. Failure 3 rests on a single protein and misattribution, and no longer

supports any claim about the tools it tested; Failure 4 on a single classifier design, and its rebuild is a reconstruction (30,780 positives against the published 17,056, 7,149 negatives against 8,528) that shows the published values do not reproduce, not that they are wrong. The Failure 1 bacterial control uses eight hand-picked genomes per domain — enough to show the archaeal figure does not separate from the bacterial one, a negative result robust to the sampling, but not enough for a positive claim, and it cannot fully disentangle domain from curation intensity. Failure 2's cross-domain result (58 genomes) was not pre-registered; 18 fungal genomes is not enough to characterise the fungal case beyond "it differs," and a study designed around that question would sample more deeply. The GECCO/DeepBGC probe (Failure 4) tests shipped models rather than a retrained reproduction, since neither tool's training data was fully recoverable; GECCO's detection recall was measured on isolated, pre-curated records without a matched genomic negative background, an easier task than genome-scale scanning, which may explain why recall held up where classification did not.

A few results are contingent in ways worth stating plainly. Failure 6 reproduces to every reported digit given the correct embedding file and degrades silently given a different valid one — a property of the pipeline, not a resolved issue. We could not compare Failure 5's embedding–chemistry correlation against our own prior figure ($r = 0.229$), because that experiment's output does not survive; the measurement reported here does. Failure 5's 2.24-fold retrieval lift falls to 1.92 once same-compound neighbours are excluded, a correction we only made because a figure made the redundancy visible. Substrate-prediction accuracy (74.7%) is bacterial throughout: no archaeal adenylation domain with an experimentally known substrate exists to test against, which makes archaeal accuracy not merely unmeasured but currently unmeasurable. Finally, some of this is ordinary bugs rather than ideas — Failure 1's metric, the DeepFRI mode confusion in Failure 2, and the fail-open embedding store above are all fixable defects, included because none was fixed for as long as nobody re-derived the number that would have caught it.

What we changed

All seven failures, and the fail-open embedding store, are corrected in the pipeline and its outputs: the annotation metric counts blank products; the recovery figure is a measured union; the classifier carries balanced negatives and documented seeds; the structure pipeline's embedding requirement is stated explicitly; the substrate model is reported from a rerun rather than a stored checkpoint; and every number here is re-derived from its inputs by a checking script rather than copied from a summary file. That script found thirteen numerical discrepancies between this manuscript and our own results directory before any of the analyses above were run, two of them errors we introduced while correcting others — a standing check, not a one-time audit, because the failure mode it defends against is the ordinary reasonableness of a plausible number.

Methods

Annotation-quality measurement and bacterial control

Coding sequences were parsed from GenBank flat files (Biopython 1.85): keyword-flagged if `/product` (lowercased) contained *hypothetical*, *predicted protein*, *uncharacterized*, or *unknown function*; blank if `/product` was absent or empty; unannotated if either. Archaeal genomes: *H. salinarum* NRC-1, *H. mediterranei* ATCC 33500, *H. volcanii* DS2 (CP001956.1), *M. acetivorans* C2A (AE010299.1), *Natrinema* sp. J7-2, *P. furiosus* COM1, *S. acidocaldarius* DSM 639 (CP000077.1), *T. kodakarensis* KOD1. [ACCESSIONS NEEDED for five of the eight.] Bacterial genomes (NCBI: U00096.3, AL009126.3, AL123456.3, AL645882.2, AE004091.2, CP000253.1, BA000022.2, AP008226.1) span curation intensity from models to environmental isolates. The implementation reproduces all eight published archaeal counts before touching any bacterial genome and aborts on disagreement. Two-sided Mann-Whitney ($n = 8$/group, 5% critical value $U \leq 13$); no multiple-comparison correction, one pre-specified comparison. `bin/bacterial_annotation_control.py`.

Annotation-metric failure rate at scale

2,000 complete assemblies per population, recorded seed, drawn from 66,714 RefSeq and 90,043 GenBank summaries (archaea a quarter of each sample). Each `.gbff.gz` was downloaded, parsed and deleted in turn; provider/pipeline read from each record's structured comment. The implementation aborts unless it reproduces all eight published archaeal counts, confirmed under both Biopython 1.85 (published counts) and 1.86 (this run). No-CDS assemblies reported separately, not dropped. `bin/stageA_annotation_metric_at_scale.py`.

Enzyme function prediction at scale and cross-tabulation

CLEAN (Yu et al. 2023; native GPU install, pretrained split100 weights/GMM ensemble) and MetagenomicDeepFRI (Gligorijević et al. 2021; v1.1.8, v1.1 weights, sequence-only CNN/MF mode, no structures) were each validated against the single-proteome result before scaling: CLEAN reproduced 92 of 93 published calls, agreeing top EC on all overlapping; DeepFRI-MF reproduced 97% of the published count. For the 2,476 *M. acetivorans* hypothetical proteins, CLEAN ran in max-separation mode and DeepFRI in CNN sequence-only mode over the whole proteome; high confidence is score ≥ 0.5 for both. Counts are per protein, not per prediction — two loci carry two high-confidence CLEAN calls each, so 95 high-confidence predictions correspond to 93 proteins — computed over the 2,455 proteins in CLEAN's prediction file, the only comparable-output set. `bin/reconcile_function_prediction_recovery.py`.

At scale, 58 complete RefSeq assemblies were sampled (recorded seed, even bacteria/archaea/fungi representation; RefSeq chosen since Stage A showed PGAP annotation carries no blank-product confound), keyword-hypothetical proteins (60–1000 aa) capped at 400/genome. A CLEAN embedding-step defect — a failed, unchecked `subprocess.run()` letting a stale cached embedding substitute silently — was found live and is now guarded: each genome's cache is cleared beforehand and verified complete after, or the genome is discarded. `bin/stageB_predictor_overlap_at_scale.py`.

MA_4560 benchmark

Table 1's seven entries are not seven instances of one task — a domain match, a structural search, two fixed-label classifiers, a centroid ranking, a text ranking, a purpose-trained classifier — tabulated because a practitioner faces this menu, not because scores are commensurable; an abstention means something different for each. BLAST/Pfam from InterPro (PF08350). Foldseek (van Kempen et al. 2023) queried the AlphaFold model of Q8THF7 against AFDB50/PDB100 (3Di+AA, E<0.01; all 963 AFDB50 hits inspected). ESM-2 650M (Lin et al. 2023) mean-pooled embeddings for all 4,540 *M. acetivorans* proteins, ranked by cosine to a MIBiG-derived biosynthetic centroid. ProTrek (Su et al. 2025) queried with fifteen phrasings in protein-to-text mode (phrase-sensitive; full ranking in Supplementary Table 1). Identity resolved from the UniProtKB API, not a stored copy, by `bin/verify_ma4560_identity.py`; Table 1's grading follows its result.

Biosynthetic-protein classifier

As published: logistic regression on mean-pooled ESM-2 650M embeddings. Positives: 17,056 proteins from MIBiG BGCs. Negatives, first version: 1,323 archaeal housekeeping proteins; corrected version: the same 1,323 plus 7,205 non-biosynthetic proteins from the same MIBiG BGCs (8,528 total; ROC-AUC 0.9946, F1 0.9844). None of these counts is recoverable from the repository, so Extended Data Table 4 is a rebuild, not a rerun: `bin/recreate_failure4_classifier.py` extracts every 30–1000-residue CDS from 5,272 MIBiG v4 records (capped 50/record, deduplicated, `gene_kind` retained), paired with 1,495 archaeal primary-metabolism proteins from the eight reference genomes above. Three arms differ only in class cuts (Extended Data Table 4); each is fitted 25 times (seeds 0–24, controlling positive downsampling and the 5-fold split). Reported ROC-AUC/F1 are cross-validated; the probe score refits on the full arm and scores all 4,540 *M. acetivorans* proteins. `bin/audit_failure4_seed_sensitivity.py`, `bin/audit_failure4_score_distribution.py`.

Generalisation to published BGC classifiers (GECCO, DeepBGC)

GECCO v0.9.10 and DeepBGC v0.1.31 ran their own shipped, pretrained models (GECCO's CRF; DeepBGC's LSTM detector + Random Forest classifier). Each tool's actual training population was reconstructed from its authors' own published data: GECCO's `mibig-2.0.proG2.clusters.tsv` (1,880 MIBiG v2.0 training BGCs, prokaryotic; the matching domain-feature/negative-background tables have no public remote, so GECCO was not retrained), DeepBGC's v0.1.0 release assets (1,814 training BGCs, including 267 fungal). Taxonomy was resolved for every training taxid and every post-cutoff MIBiG v4.0 BGC; held-out BGCs were grouped by whether their genus (GECCO: phylum) occurs in that population. 420 BGCs (GECCO) and 472 BGCs (DeepBGC) ran through each tool's own pipeline on curated GenBank annotation, bypassing each tool's gene caller identically across groups. Detection recall: fraction receiving ≥1 predicted cluster. Type-match accuracy: restricted to BGCs with an unambiguous MIBiG class. Two-sided Fisher's exact tests (plus Cochran-Armitage trend for GECCO's four ordered groups); confidence calibration by Mann-Whitney U on each tool's own score, correct vs incorrect calls. `bin/run_stageC.py`, `bin/run_stageC_deepbgc.py`, `bin/analyze_stageC.py`, `bin/extract_class_probs.py`, `bin/stageC_confound_summary.py`.

Embedding–chemistry correlation, substrate prediction, and structure prediction

BGC-level embeddings: mean-pooled, L2-normalised ESM-C per-protein embeddings, matching the structure pipeline's pooling. Of 453 BGCs with a reference compound (MIBiG modules), 399 had both a parseable structure and an embedding (79,401 pairs). Chemical similarity: Tanimoto over Morgan fingerprints (radius 2, 2048 bits, chirality-aware; RDKit). Pearson/Spearman over all pairs and within cosine strata; 95% CIs from 1,000 bootstrap resamples over BGCs. Retrieval view: each BGC's single highest-cosine neighbour (excluding itself) against a random permutation of the same compounds. `bin/measure_embedding_chemistry_correlation.py`.

Substrate prediction: per-residue ESM-2 650M embeddings at 34 HMM-calibrated binding-pocket positions, encoded by a 2-layer transformer (4 heads, 256 hidden, attention pooling to 128-d), aligned by symmetric InfoNCE (τ=0.07) against chirality-aware Morgan fingerprints via a 2-layer MLP, 0.5-weighted auxiliary classification, same-class negative masking. Training data: 8,744 bacterial adenylation domains (3,653 PARASECT, 2,798 MIBiG, 2,293 MASPR); classes under 10 examples excluded, leaving 60 classes over 8,639 domains (split 6,065/1,299/1,272; no pseudo-labelled data used). HMMER 3.3.2 (AMP-binding/AMP-binding_C, E<1e-5) for domain extraction.

Structure prediction: template matching by substrate-overlap

similarity weighted by ESM-C cosine, then RDKit *de novo* assembly for unmatched BGCs, with head-to-tail macrocyclisation where a thioesterase domain and an RDKit-confirmed ring (≥8 atoms) are present. Evaluated on 453 MIBiG NRPS BGCs with module annotations and a reference compound. Requires an HDF5 store keyed `<bgc_id>|<gene_id>`; a differently-keyed store yields zero matches and silent degradation to 0.339 (Failure 6, Discussion). `bin/optimal_structure_predictor.py`.

DeepSEMS reimplementation

Reimplemented from the 2025 bioRxiv preprint (DOI 10.1101/2025.03.02.641084) with Pfam version mapping, evaluated on our MIBiG validation set (mean Tanimoto 0.266, 647 BGCs). Since peer-reviewed [9]; the reimplementation predates it and was not re-run against it.

Numerical verification

Every quantitative claim in this manuscript is re-derived from files under `results/` by `bin/verify_manuscript_numbers.py`, which reports each claim as matching, mismatching, or having no artefact, and exits non-zero on any mismatch or unfilled placeholder. The artefacts it reads are pinned by SHA-256 in `manuscript_package/ARTIFACT_CHECKSUMS.sha256`.

Data and code availability

All code is available at [REPOSITORY URL NEEDED]. Analysis outputs, the checking script and the artefact checksum manifest are included. Genome records are retrievable from NCBI under the accessions listed above.

References

1. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. & Wichmann, F.A. Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673 (2020). doi:10.1038/s42256-020-00257-z
2. Nevers, Y. et al. Quality assessment of gene repertoire annotations with OMArk. *Nature Biotechnology* **43**, 124–133 (2024). doi:10.1038/s41587-024-02147-w
3. Nayak, D.D. et al. Post-translational thioamidation of methyl-coenzyme M reductase, a key enzyme in methanogenic and methanotrophic archaea. *eLife* **6**, e29218 (2017). doi:10.7554/eLife.29218
4. Yu, T. et al. Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023). doi:10.1126/science.adf2465
5. Gligorijević, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nature Communications* **12**, 3168 (2021). doi:10.1038/s41467-021-23303-9
6. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**, 243–246 (2023). doi:10.1038/s41587-023-01773-0
7. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). doi:10.1126/science.ade2574
8. Su, J. et al. ProTrek: navigating the protein universe through tri-modal contrastive learning. *Nature Biotechnology* (2025). [VERIFY volume, pages and DOI against the published version before submission.]
9. Xu, T. et al. DeepSeMS: revealing the hidden biosynthetic potential of the global ocean microbiome with a large language model. *Nature Computational Science* **6**, 753–766 (2026). doi:10.1038/s43588-026-00983-1. Reimplemented here from the earlier preprint, doi:10.1101/2025.03.02.641084.
10. Blin, K. et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research* **51**, W46–W50 (2023). doi:10.1093/nar/gkad344
11. Skinnider, M.A. et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature Communications* **11**, 6058 (2020). doi:10.1038/s41467-020-19986-1
12. Terlouw, B.R. et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research* **51**, D603–D610 (2023). doi:10.1093/nar/gkac1049
13. Kautsar, S.A. et al. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *GigaScience* **10**, giaa154 (2021). doi:10.1093/gigascience/giaa154
14. Carroll, L.M., Larralde, M., Fleck, J.S., Ponnudurai, R., Milanese, A., Cappio, E. & Zeller, G. Accurate de novo identification of biosynthetic gene clusters with GECCO. *bioRxiv* (2021). doi:10.1101/2021.05.03.442509. [Preprint, not peer reviewed.]
15. Hannigan, G.D. et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research* **47**, e110 (2019). doi:10.1093/nar/gkz654
16. Blin, K. et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Research* **53**, W32-W38 (2025). doi:10.1093/nar/gkaf334
17. Zdouc, M.M. et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Research* **53**, D678-D690 (2024). doi:10.1093/nar/gkae1115
18. Yang, J. et al. Opportunities and challenges for machine learning-assisted enzyme engineering. *ACS Central Science* **10**, 226-241 (2024). doi:10.1021/acscentsci.3c01275
19. Kouba, P. et al. Machine learning-guided protein engineering. *ACS Catalysis* **13**, 13863-13895 (2023).

doi:10.1021/acscatal.3c02743
20. Terlouw, B.R. et al. PARAS: high-accuracy machine learning of substrate specificities in nonribosomal peptide synthetases. *JACS Au* **6**, 2315-2336 (2026). doi:10.1021/jacsau.5c01636
21. Jha, A. et al. Gaia: an AI-enabled genomic context-aware platform for protein sequence annotation. *Sci. Adv.* **11**, eadv5109 (2025). doi:10.1126/sciadv.adv5109
22. Zhu, Y. et al. Computational advances in biosynthetic gene cluster discovery and prediction. *Biotechnology Advances* **79**, 108532 (2025). doi:10.1016/j.biotechadv.2025.108532
23. Salamzade, R. et al. Context matters: assessing the impacts of genomic background and ecology on microbial biosynthetic gene cluster evolution. *mSystems* **10**, e01538-24 (2025). doi:10.1128/msystems.01538-24

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files. Source code is available at [REPOSITORY URL NEEDED].

Competing interests

G.R. is the founder of Omic Inc. All other authors are employees of Omic Inc. and declare no other competing interests.

Funding

This work was funded by Omic Inc. The funder had no role in the design of the study, the analysis, or the decision to publish.

Authors' contributions

All authors contributed to the design of the study, the analysis and the manuscript, and approved the submitted version.

Acknowledgements

Acknowledgements

Figure Legends

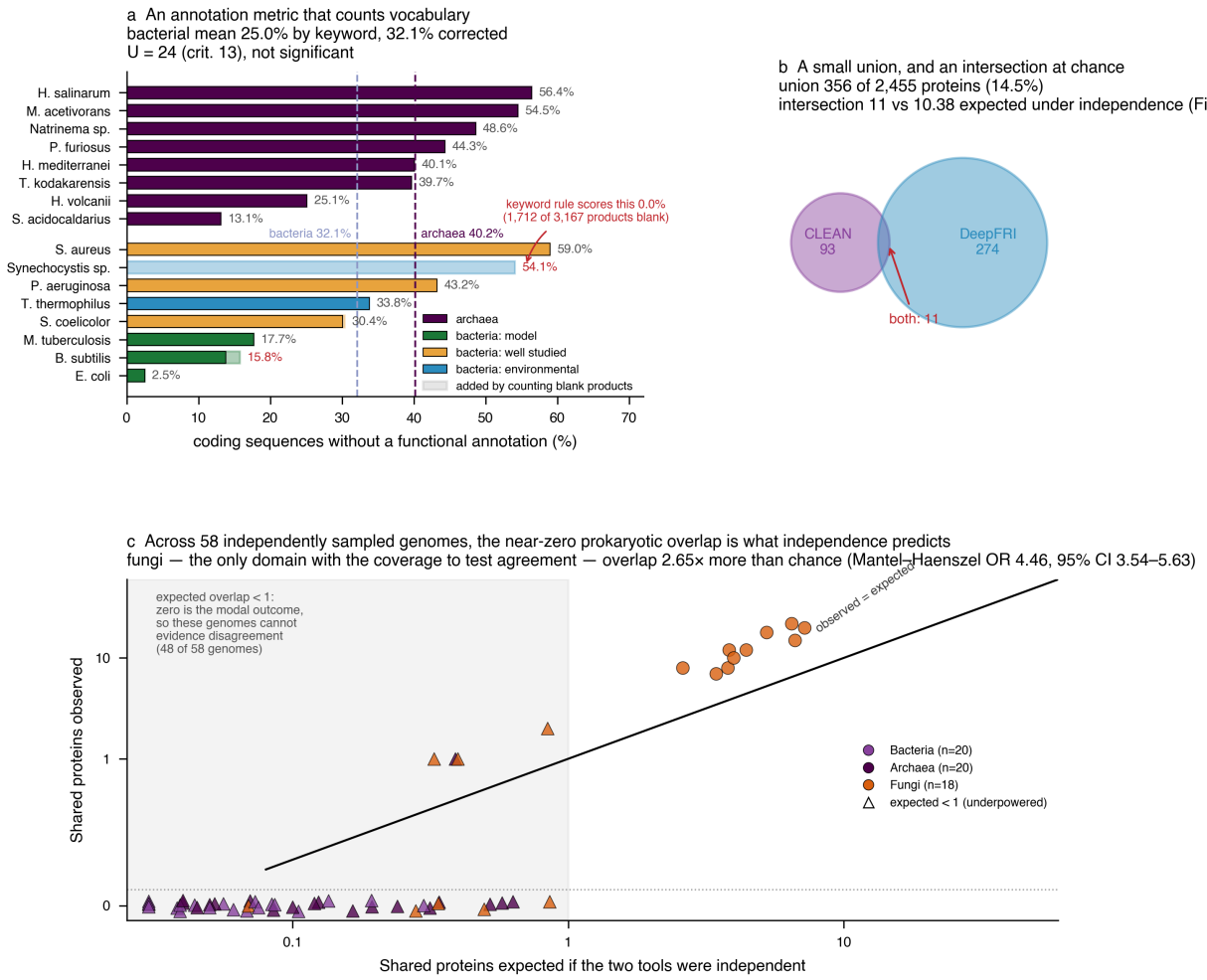


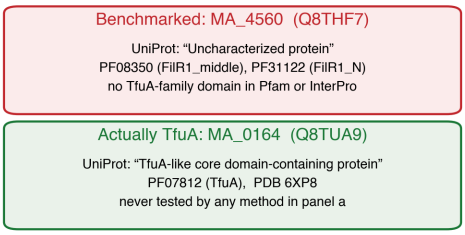
Figure 1. Two failures of aggregate reporting. (a) Unannotated coding sequences across eight archaeal and eight bacterial genomes, under the keyword rule (open bars) and counting blank `/product` fields (filled). *Synechocystis* sp. PCC 6803 scores 0.0% by keyword and 54.1% when blank products are counted; no archaeal genome has any blank products. Archaeal and bacterial means are 40.2% and 32.1% respectively under the corrected rule, with ranges overlapping and Mann-Whitney U = 24 against a critical value of 13. Bacterial genomes are shaded by curation tier, which orders them better than domain does. (b) CLEAN and DeepFRI high-confidence recovery over the 2,455 proteins both tools scored: union 356 (14.5%), intersection 11, Jaccard 0.03. (c) The same measurement repeated on 58 independently sampled genomes, plotted against the overlap each genome's coverage rates predict under independence rather than as a raw Jaccard. Bacteria and archaea sit at or below expectation on counts too small to test (0 observed against 1.50 expected, and 1 against 4.11); fungi sit well above it (136 against 51.3, Mantel-Haenszel odds ratio 4.46). Every one of the 44 genomes with zero observed overlap had an expected overlap below 1; the 48 genomes whose expected overlap is below 1 are drawn as triangles on a shaded background, because zero is the modal outcome there under independence and those genomes cannot evidence disagreement either way. Genomes with no observed overlap are drawn on the axis band marked 0. Panel (c) replaces an earlier version plotting per-domain mean Jaccard, a statistic bounded by the marginal coverage rates and therefore uninformative about agreement at these rates.

a Seven methods on MA_4560, regraded against what the protein actually is

Method	Prediction returned	Regraded
Sequence homology BLAST / Pfam	`PF08350 (FilR1_middle)` `transcriptional regulator`	correct
Structural search Foldseek	`963 AFDB50 hits,` `all transcriptional regulators`	correct
Enzyme prediction CLEAN	`score 0.000` `no enzyme function`	abstains
Protein language model ESM-2 (650M)	`rank 1903/4540` `not discriminative`	abstains
Tri-modal contrastive ProTrek	`#1 "winged helix-turn-helix …" (-0.035)` `#3 thioamide bond (-0.054)`	correct (at rank #1)
Structure-function DeepFRI	`DNA binding` `confidence 0.669`	correct
Purpose-trained classifier (this work)	`score 0.000` `rank 3718/4540`	abstains

Every prediction is as measured; the grading is not. Scored against a ground truth taken from an internal code comment, all seven read as failures. Scored against UniProt and InterPro, 4 of 7 return the winged helix-turn-helix assignment those databases support and 3 correctly decline to call it. Methanosarcina acetivorans C2A, UniProt Q8THF7.

b The benchmark tested the wrong protein



c The score moved three orders of magnitude; the protein stayed in the same half of its proteome

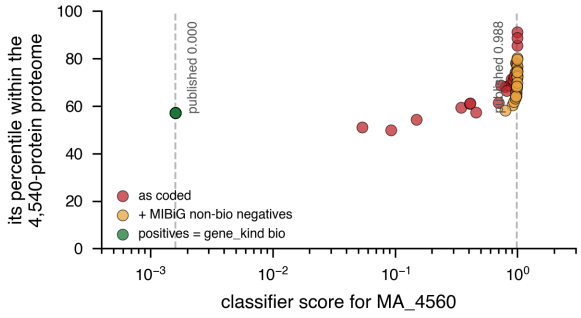


Figure 2. A benchmark whose key was wrong, and a classifier whose score was not about the protein. (a) Seven methods on MA_4560, each cell showing the prediction actually returned. Every prediction is as measured; the grading is the part that changed. Scored against the ground truth taken from an internal code comment, all seven read as failures. Scored against UniProt and InterPro, four return the winged helix-turn-helix assignment those databases support — BLAST/Pfam (PF08350, FilR1_middle), Foldseek (963 AFDB50 and 477 PDB100 hits, all regulators), ProTrek (its rank-1 string) and DeepFRI (DNA binding, confidence 0.669) — and three abstain, two at exactly 0.000 and one at a near-random rank — CLEAN, our own classifier at rank 3718 / 4540, and ESM-2 at rank 1903/4540 — which on the corrected key is the right call in each case. (b) Why the key was wrong. MA_4560 (UniProt Q8THF7) is annotated "Uncharacterized protein" with Pfam PF08350 (FilR1_middle) and PF31122 (FilR1_N) and carries no TfuA-family domain in either Pfam or InterPro. The TfuA of *M. acetivorans* is MA_0164 (UniProt Q8TUA9), Pfam PF07812 (TfuA core), the protein in PDB 6XP8 — and it was never tested by any method in panel (a). Resolved from the UniProtKB API by `bin/verify_ma4560_identity.py`. (c) The rebuilt Failure 4 classifier, 75 fits across three arms, plotting MA_4560's score against its percentile within the 4,540-protein *M. acetivorans* proteome. The score spans three orders of magnitude, from 0.0016 to 0.9996, while the percentile never leaves the 49.9–91.2 band. Dashed lines mark the two published values, neither of which the rebuild reproduces in the arm it was reported from. Panels (a) and (c) replace earlier versions that encoded the withdrawn grading and the two published scores; panel (b) replaces one drawn to show a fold/function dissociation that does not exist in this protein.

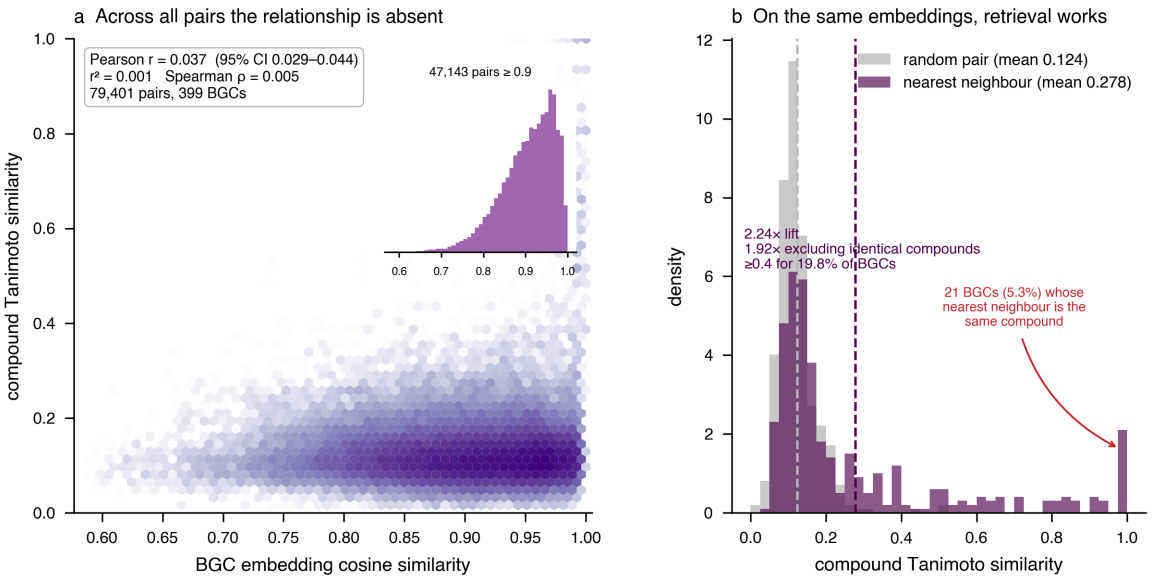


Figure 3. A relationship that no single coefficient describes. (a) Embedding cosine against compound Tanimoto for 79,401 MIBiG BGC pairs: Pearson r = 0.037, $r^2 = 0.001$. The inset shows the cosine distribution, with 47,143 pairs above 0.9 — mean-pooled embeddings occupy a narrow cone. (b) The same embeddings under nearest-neighbour retrieval: mean Tanimoto 0.278 against 0.124 for random pairs, a 2.24-fold lift, with 19.8% of BGCs above 0.4. The spike at 1.0 is 21 BGCs (5.3%) whose nearest neighbour is a different MIBiG entry for the same compound; excluding them the mean is 0.237 and the lift 1.92-fold.

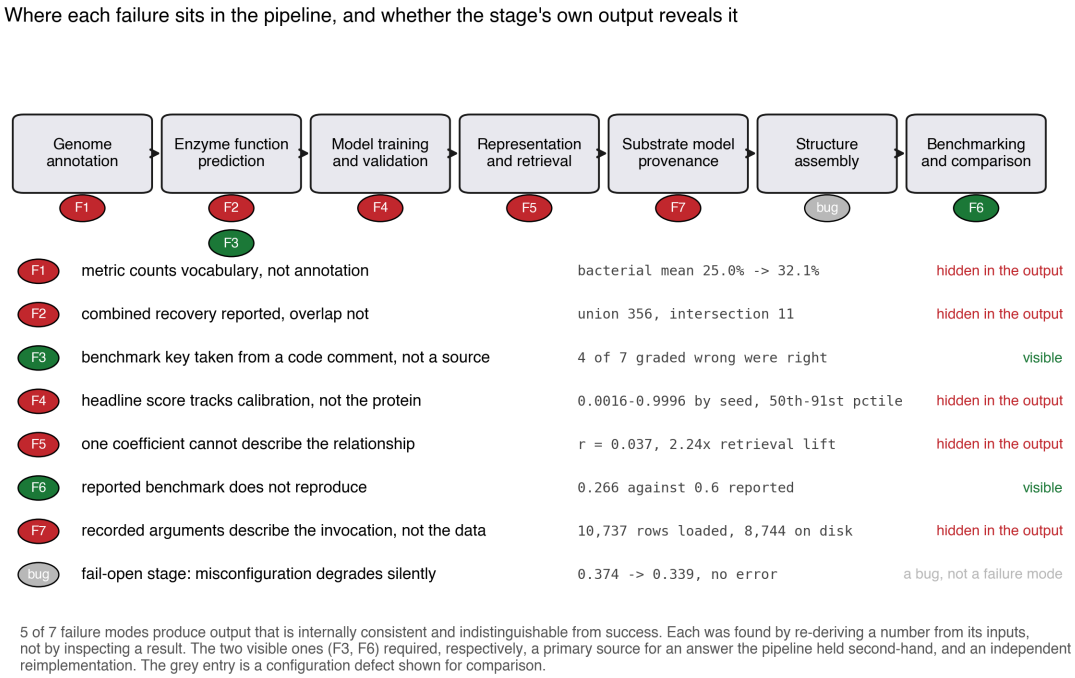


Figure 4. Where each failure sits in the pipeline, and which are visible in the artefact the stage produces. Five of seven are not; the fail-open embedding-store defect discussed in the text is shown in grey as an eighth entry that is a bug rather than a failure mode.