

Seven Silent Failure Modes in Genome-Mining Pipelines – Extended Data and Supplementary

Extended Data

Extended Data Table 1. Unannotated coding sequences under the keyword rule and under the corrected rule.

Domain	Genome	CDS	Keyword	Blank product	Unannotated
Archaea	<i>Halobacterium salinarum</i> NRC-1	2,058	56.4%	0.0%	56.4%
Archaea	<i>Methanosarcina acetivorans</i> C2A	4,540	54.5%	0.0%	54.5%
Archaea	<i>Natrinema</i> sp. J7-2	4,239	48.6%	0.0%	48.6%
Archaea	<i>Pyrococcus furiosus</i> COM1	2,064	44.3%	0.0%	44.3%
Archaea	<i>Haloferax mediterranei</i> ATCC 33500	2,942	40.1%	0.0%	40.1%
Archaea	<i>Thermococcus kodakarensis</i> KOD1	2,306	39.7%	0.0%	39.7%
Archaea	<i>Haloferax volcanii</i> DS2	2,990	25.1%	0.0%	25.1%
Archaea	<i>Sulfolobus acidocaldarius</i> DSM 639	2,223	13.1%	0.0%	13.1%
Bacteria	<i>Staphylococcus aureus</i> NCTC 8325	2,892	59.0%	0.0%	59.0%
Bacteria	<i>Synechocystis</i> sp. PCC 6803	3,167	0.0%	54.1%	54.1%
Bacteria	<i>Pseudomonas aeruginosa</i> PAO1	5,571	43.2%	0.0%	43.2%
Bacteria	<i>Thermus thermophilus</i> HB8	1,973	33.8%	0.0%	33.8%
Bacteria	<i>Streptomyces coelicolor</i> A3(2)	7,825	30.1%	0.3%	30.4%
Bacteria	<i>Mycobacterium tuberculosis</i> H37Rv	4,031	17.7%	0.0%	17.7%
Bacteria	<i>Bacillus subtilis</i> 168	4,333	13.8%	2.1%	15.8%
Bacteria	<i>Escherichia coli</i> K-12 MG1655	4,318	2.5%	0.0%	2.5%
	Archaeal mean		40.2%	0.0%	40.2%
	Bacterial mean		25.0%	7.1%	32.1%

Extended Data Table 2. Disagreement between the keyword rule and blank /product fields, by annotation provenance.

Population	Genomes	With blank products	Max disagreement	Synechocystis-like
RefSeq, PGAP-annotated	1,999	0 (0.00%)	0.0 pp	0
GenBank, PGAP-annotated	1,164	1 (0.09%)	—	0
GenBank, submitter-annotated	393	34 (8.65%)	22.1 pp	0
GenBank, other institution	14	0 (0.00%)	—	0
All	3,570	35 (0.98%)	22.1 pp	0

Extended Data Table 3. CLEAN/DeepFRI overlap across 58 independently sampled genomes, against the overlap their coverage rates predict. Expected intersections are the sum over genomes of $n(\text{CLEAN}) \times n(\text{DeepFRI}) / n(\text{scored})$. Odds ratios are Mantel-Haenszel, stratified by genome; bacteria admit no odds ratio because no genome has a shared protein.

Domain	Genomes	Observed shared proteins	Expected if independent	Obs/exp	MH odds ratio (95% CI)	Genomes with zero intersection	Of those, expected < 1
Bacteria	20	0	1.50	0.00	— ($p = 0.22$)	20 (100%)	20
Archaea	20	1	4.11	0.24	0.23 (0.03–1.67)	19 (95%)	19
Fungi	18	136	51.27	2.65	4.46 (3.54–5.63)	5 (28%)	5
All	58	137	56.89	2.41	—	44 (76%)	44

Extended Data Table 4. The Failure 4 classifier rebuilt, three arms, 25 seeds each. Positives and negatives are disjoint in every arm. “Percentile” is MA_4560’s position in the score distribution over all 4,540 *M. acetivorans* proteins; “proteome median” is the median score those proteins receive. Artifacts: results/failure4_recreation/{recreation_results,seed_sensitivity,sc

Arm	Positives	Negatives	ROC-AUC	MA_4560 score	Its percentile	Proteome median	Its rank
As coded: all MIBiG CDS vs archaeal	36,434	1,495	0.9985	0.054– 0.9996	49.9–91.2	0.024–0.265	398–2277
Balanced: add MIBiG non- biosynthetic negatives	30,780	7,149	0.9882	0.793– 0.9997	58.2–80.4	0.291–0.614	892–1899
Label- corrected: positives are gene_kind biosynthetic	16,701	7,149	0.9982	0.0016	57.2	0.0001	1942

Extended Data Table 5. GECCO and DeepBGC on populations distinguished by training coverage. “Genus/phylum covered” means the BGC’s genus (GECCO: phylum) occurs in the tool’s published training population; “novel” means it does not. All post-training BGCs are drawn from MIBiG entries added after each tool’s training cutoff.

Tool	Population	n	Detection recall	Type-match accuracy
GECCO	Trained-on BGCs	60	80.0%	84.9% (45/53)
GECCO	Novel BGC, genus covered	120	75.8%	76.2% (80/105)
GECCO	Novel BGC, genus novel	120	73.3%	72.3% (73/101)
GECCO	Novel BGC, phylum entirely absent	120	76.7%	67.6% (75/111)
DeepBGC	Trained-on BGCs	60	98.3%	67.9% (36/53)
DeepBGC	Novel bacterial BGC, genus covered	110	95.5%	70.4% (69/98)
DeepBGC	Novel bacterial BGC, genus novel	108	88.0%	49.5% (48/97)
DeepBGC	Novel fungal BGC, genus covered	110	98.2%	63.2% (67/106)
DeepBGC	Novel fungal BGC, genus novel	68	98.5%	67.7% (42/62)

Supplementary

Supplementary Table 1. ProTrek similarity scores for all fifteen functional descriptions queried against MA_4560, ranked.

Supplementary Table 2. Per-genome annotation counts for the sixteen control genomes: total CDS, keyword-flagged, blank product, unannotated, curation tier.

Supplementary Table 4. Per-genome results for all 3,570 sampled RefSeq and GenBank genomes: accession, domain, annotation provider, total CDS, keyword-flagged, blank products, and the 429 assemblies carrying no CDS features.

Supplementary Table 3. Per-protein CLEAN and DeepFRI high-confidence assignments over the 2,455 shared proteins, with union and intersection membership.