A survey of 3,974 archaeal biosynthetic gene clusters and a substrate predictor for the detectable fraction – Extended Data and Supplementary

Extended Data

Extended Data Table 1. Complete literature references for 12 archaeal bioactive compound classes, including producing organisms, documented bioactivities, biosynthetic pathway organization, and detection status by antiSMASH 7.0, PRISM 4, BiG-SLiCE, and DeepSEMS.

Extended Data Table 2. Non-BGC pathway detection results across three validation genomes (*H. volcanii* DS2, *M. acetivorans* C2A, *S. acidocaldarius* DSM 639), showing detected enzymes per pathway, completeness scores, and genomic coordinates.

Extended Data Table 3. Hypothetical protein fractions for the eight archaeal genomes analysed (*H. salinarum* NRC-1, *H. mediterranei* ATCC 33500, *H. volcanii* DS2, *M. acetivorans* C2A, *Natrinema* sp. J7-2, *P. furiosus* COM1, *S. acidocaldarius* DSM 639, *T. kodakarensis* KOD1), including total protein count, hypothetical count, percentage and phylum. [ACCESSIONS NEEDED for five of the eight; see Methods, Reference genomes.]

Extended Data Table 4. CLEAN and sequence-based DeepFRI analysis of *M. acetivorans* hypothetical proteins: full list of high-confidence enzyme predictions, EC numbers, GO terms and confidence scores, with the per-tool and union membership of each protein.

Extended Data Table 5. ProTrek similarity scores for all fifteen functional descriptions queried against MA_4560 (queried believing it to be TfuA; corrected in Results — MA_4560 is not TfuA), ranked.

Extended Data Table 6. Seven computational methods applied to MA_4560, retained to document the superseded benchmark described in Results. Predictions are as measured; the grading columns reflect the original assignment of MA_4560 to TfuA and its correction.

Method	Tool	Prediction	Correct for TfuA?	Correct for what MA_4560 actually is?
Sequence homology	BLAST/Pfam	PF08350 (FilR1) — transcriptional regulator	No (wrong protein tested)	Plausibly yes — matches MA_4560's own UniProt Pfam call
Structural search	Foldseek [9]	963 hits, all transcriptional regulators	No (wrong protein tested)	Plausibly yes
Enzyme prediction	CLEAN [8]	Score 0.000 — no enzyme function detected	No (wrong protein tested)	Plausibly correct abstention (MA_4560 is not an enzyme)
Protein language model	ESM-2 [13]	Rank 1903/4540 — not discriminative	No (wrong protein tested)	Uninformative either way
Tri-modal contrastive	ProTrek [24]	"regulatory" rank #1 (-0.035); "thioamidation" rank #3 (-0.054)	No (wrong protein tested)	Plausibly yes on rank #1
Structure-function (sequence-only)	DeepFRI [27]	DNA binding (GO:0003677) at confidence 0.669	No (wrong protein tested)	Plausible — MA_4560 has a winged-helix DNA-binding domain
Custom biosynthetic classifier	This work	Score 0.000 — rank 3718/4540	No (wrong protein tested)	Plausibly correct (MA_4560 is not biosynthetic)

Extended Data Figure 1. (a) Phylogenetic tree of 1,681 archaeal genomes colored by number of detected BGCs. (b) BGC density (BGCs per Mb) across archaeal phyla. (c) Distribution of BGC sizes (bp) compared to bacterial BGCs from MIBiG.

Extended Data Figure 2. (a) BGCast training curves (loss and accuracy). (b) Per-substrate accuracy breakdown (top 20 substrates). (c) Confusion matrix for most common substrates. (d) Ablation study: 34 pocket positions against the 10 Stachelhaus positions, re-run for this paper over five seeds under the configuration reported here. 34 positions give 74.74% ± 0.32 top-1 and the 10 Stachelhaus positions 74.41% ± 0.46; the paired difference is +0.34 percentage points with a standard deviation of 0.50 and an inconsistent sign across seeds. The extra 24 positions are not distinguishable from the 10 at this sample size. This supersedes a previously quoted 74.9% / 73.8% pair, which belonged to a different model configuration and was cited for a +1.1-point advantage this rerun does not reproduce (`bin/audit_substrate_ablation.py`). (e) Sequence identity distribution for archaeal versus bacterial thiopeptide alignments.

Extended Data Figure 3. CLEAN and sequence-based DeepFRI analysis of *M. acetivorans* hypothetical proteins. (a) Distribution of CLEAN confidence scores. (b) EC class distribution of high-confidence CLEAN predictions. (c) Recovery rate as a function of CLEAN confidence threshold. (d) Distribution of DeepFRI confidence scores for EC and GO predictions. (e) Overlap between CLEAN and DeepFRI high-confidence predictions over the 2,455 proteins both tools scored: CLEAN only 82, DeepFRI only 263, both 11, neither 2,099 (Jaccard 0.03).

Extended Data Figure 4. TfuA-associated loci analysis, from the 500-loci database query described in Methods and Results, each locus a distinct BGC within the QC'd population (498 of them also genus-classified). (a) Genomic context of 12 highest-scoring loci showing precursor peptide candidates and absence of MCR. (b) Distribution of secondary metabolism scores across the 500 loci. (c) Phylogenetic distribution of TfuA loci with and without YcaO partners (partner search confined to each locus's called cluster boundary, not the flanking contig).

Extended Data Figure 5. Comparison of the seven computational methods on the MA_4560 benchmark, originally run believing MA_4560 to be TfuA (corrected in Results: it is an unrelated FilR1-family regulator; the real TfuA is MA_0164). (a) BLAST alignment scores and top hit annotations. (b) Foldseek structural search results colored by functional annotation. (c) CLEAN confidence distribution for *M. acetivorans* proteins highlighting MA_4560's position. (d) ESM-2 cosine similarity ranking. (e) ProTrek similarity scores for functional descriptions. (f) DeepFRI predicted GO terms and confidence scores.