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

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: Archaea produce diverse bioactive compounds, among them halocins, sulfolobicins, the carotenoid bacterioruberin and lanthipeptides. Computational natural-product discovery is nonetheless organized around the biosynthetic gene cluster (BGC), an arrangement characteristic of bacteria, and MIBiG v3.0 holds 3 archaeal entries among 2,502. Whether archaeal biosynthetic capacity is under-detected because archaeal genomes are organized and annotated differently, or for reasons that limit genome mining generally, has not been assessed against matched bacterial controls.

Results: We surveyed 1,681 archaeal genomes and catalogued 3,974 BGCs across eight phyla; of the 3,935 assignable to a functional category, terpene clusters account for 42.8%, clusters of unknown function including TfuA-associated loci for 21.4%, antibiotic-type clusters for 16.9% and other categories for 18.8%. We built BGCast, a contrastive-learning predictor that aligns adenylation-domain binding-pocket embeddings with substrate fingerprints, an architecture that can in principle rank substrates absent from its training set, though our class-count filter removes every unseen class from the test split, so we do not demonstrate it. Trained on 8,744 bacterial adenylation domains, BGCast reaches 74.7% ± 0.3 top-1 and 84.1% ± 0.4 top-3 accuracy across 54 substrate classes, against approximately 85% for antiSMASH; these are upper bounds, since 74.0% of test domains share a binding-pocket string with a training domain. Detection rules in current tools cover an estimated 21-54% of 12 literature-curated archaeal compound classes. Two candidate explanations for that shortfall did not survive matched bacterial controls. Archaeal proteomes average 40.2% of proteins annotated as hypothetical, against 32.1% across eight bacterial genomes spanning curation intensity (Mann-Whitney U = 24, critical value 13), and curation intensity predicts the fraction better than domain does. The bacterioruberin pathway of *Haloferax volcanii* spans 81% of its main chromosome, but the same keyword signature recovers comparable dispersion in three bacterial carotenoid producers whose pathways are genuinely clustered.

Conclusions: BGC-centric tools under-detect archaeal biosynthetic diversity, but the two mechanisms we could test against controls are general limitations of genome mining rather than properties of archaeal genomes. The survey and BGCast provide a catalog and a substrate-prediction resource for the fraction current tools do detect, and the control design separates domain-specific effects from curation and detection-method artifacts.

Background

In 1977, Woese and Fox recognized archaea as a third domain of life, distinct from both bacteria and eukarya [1]. In the nearly five decades since, archaea have yielded fundamental insights into the origins of eukaryotic cells, extremophile biology, and methanogenesis. They have not, however, yielded natural products — at least not through the computational pipelines that have transformed bacterial and fungal natural product discovery.

This is not because archaea lack biosynthetic capacity [19], and recent reviews of computational BGC discovery [34, 38] and of biosynthesis in extreme environments [35] treat archaea as an underexplored fraction rather than an empty one. Halophilic archaea produce halocins, antimicrobial peptides active against other archaea and, in some cases, bacteria [5,6]; halocin H4 was recently shown to require proteolytic activation [37], which is the kind of post-translational step BGC-centric detection does not represent. Thermoacidophilic *Sulfolobus* species produce sulfolobicins, antimicrobial proteins so novel they share no detectable sequence homology with any known protein family [20]. Halophilic archaea synthe-

size bacterioruberin, a C50 carotenoid that induces apoptosis in leukemia cells [21]. Methanogens produce cofactor F430, a nickel porphinoid with at least nine modified variants. Archaeal lanthipeptides — the first ribosomal natural products from the third domain — have recently been characterized [2], and Song et al. demonstrated that 24 archaeal lanthipeptides can be produced via haloarchaeal heterologous expression [3].

Most recently, de la Fuente et al. [26] applied deep learning to 233 archaeal proteomes and identified 12,623 candidate antimicrobial peptides, termed "archaeasins." Of 80 synthesized candidates, 93% exhibited antimicrobial activity, and archaeasin-73 was validated in a mouse infection model. This work provides the strongest experimental evidence to date that archaea harbor a vast reservoir of bioactive chemistry.

Yet the MIBiG database, the definitive reference for characterized biosynthetic gene clusters, contains 2,502 entries (v3.0) spanning bacteria, fungi and other eukaryotes, of which only 3 are archaeal, all Euryarchaeota [14]. MIBiG has since been released as v4.0 [29]; we report v3.0 counts throughout because that is the release this survey was built against. Database representation reflects submission and curation practice — a compound generally has to be published and its BGC characterized before anyone submits it to MIBiG — so this near-absence does not by itself establish that archaeal BGCs are computationally invisible. Whether it is, and why, is an empirical question; the computational paradigm used to discover natural products — detect biosynthetic gene clusters (BGCs), predict products, prioritize for expression — was built for organisms whose biosynthetic genes are organized in operons, and for archaea, this assumption sometimes fails, but establishing how often, and why, requires measurement.

This paper reports two things. The first is a resource: a survey of the archaeal BGCs that current tools detect (3,974 clusters across 1,681 genomes), and BGCast, an open-vocabulary substrate-prediction and novelty-scoring toolkit for the adenylation domains those clusters contain. The second is a controlled test of two candidate explanations for the detection shortfall — that the relevant pathways are not organized as detectable clusters, and that the relevant genes lack functional annotation — each measured against a matched bacterial comparison. Cimermancic et al. estimated in 2014 that most bacterial biosynthetic capacity remains unexplored, introducing the term "biosynthetic dark matter" [4]; we ask whether archaeal capacity is unexplored for the same reasons, or for reasons specific to archaeal genome organization and annotation. Neither explanation separates archaea from its bacterial control (Results), which locates the shortfall in general properties of genome-mining methods and in curation depth rather than in archaeal genome biology.

The convergence of recent experimental breakthroughs — archaeal lanthipeptides [2,3], archaeasins [26], and archaeal RiPP characterization [17] — with the persistent computational gap makes this analysis timely regardless of which candidate explanations hold up. The question is no longer whether archaea make bioactive compounds, but how to find them systematically, and with what tools, when the standard BGC-centric paradigm does not obviously transfer.

Relationship to a companion manuscript. Several analyses reported here — the MA_4560 seven-method benchmark (Results: this locus was misidentified as TfuA in both manuscripts; the correction has now been applied in both), the CLEAN and DeepFRI recovery comparison and the eight-genome annotation survey supporting the annotation-gap findings, the biosynthetic-classifier domain-shift result and the DeepSEMS reimplementation — also appear in a companion manuscript by the same authors [PREPRINT CITATION NEEDED: insert once the preprint is posted], which treats them as failure modes of computational pipelines rather than as evidence about archaeal biology. The experiments were performed once and the numbers are identical in both. The MA_4560/TfuA identity correction applies equally to that manuscript's third failure mode, which rests on this locus, and has been applied there. New here are the survey of 3,974 BGCs across 1,681 archaeal genomes, BGCast, the bacterioruberin pathway mapping and its bacterial control, the TfuA locus analysis, and the detection-gap assessment across 12 compound classes.

Results

The archaeal BGC survey

Pangenome-scale BGC surveys of this kind are established for bacteria — BGCFlow [32] for large genome sets and a 2,371-genome *Streptomyces* pangenome [33] — but not for archaea. We surveyed 1,681 archaeal genomes using antiSMASH with BiG-SLiCE v1.1.2 [15] providing clustering and GTDB r207 taxonomy (Methods); 47% carry metagenome-assembled or WGS-project accessions rather than isolate-assembly accessions, drawn from mixed and largely unrecorded environments (Methods), so the survey is weighted toward the halophilic and marine lineages those catalogues cover — Halobacteriales (437 genomes) and Poseidoniales (409) together account for half of it — rather than sampling archaeal diversity evenly, and the 1,681/1,698 pre/post-QC genome counts are a lower bound over an unrecorded screened population, not a census. After quality control (CheckM completeness >50%, contamination <5%, phylum-level taxonomy required), we identified 3,974 BGCs across eight archaeal phyla (Fig. 1b). Functional classification required genus-level taxonomy and covers 3,935 of these BGCs (Methods): terpene clusters account for 1,685 (42.8%), clusters of unknown function including TfuA-associated loci for 843 (21.4%), antibiotic-type clusters for 666 (16.9%) and all other categories for 741 (18.8%). A cross-category selection of 961 BGCs carries therapeutic-relevant architectures, deduplicated per BGC by a documented priority order into thiopeptide-like (245), bacteriocin-like (185), type I PKS (182), siderophore-related (172), NRPS-like (66), lanthipeptide (50), lassopep-

tide (47) and sactipeptide (14), summing to exactly 961 (`bin/reconcile_prospective_discovery_counts.py`; an earlier, undeduplicated pass reported 963). Of the unknown/TfuA-associated fraction, a BiG-SLiCE query returns 500 loci carrying a "TfuA-related" classification (not a dedicated domain search — Methods). These 500 are already one row per BGC and all fall inside the QC'd population of 3,974; 498 of them also carry genus-level taxonomy and so appear in the 3,935 classified set, the other two lacking a genus assignment.

These BGCs represent only the fraction of archaeal biosynthetic capacity that resembles bacterial BGCs closely enough for existing rules to identify. Archaeal thiopeptide-like sequences share only 26-39% protein sequence identity with characterized bacterial thiopeptides in MIBiG (mean 31%; core enzymes 21-45%, mean 28%) — below the ~40% threshold where homology-based tools lose predictive power, so even the BGC-detectable fraction is likely under-called by sequence divergence alone (Fig. 1d).

BGCast: a substrate-prediction toolkit for the detectable fraction

For the fraction of archaeal biosynthetic capacity organized in detectable BGCs, computational tools can provide substrate predictions and novelty assessments. We developed BGCast, a contrastive-learning framework for adenylation domain substrate prediction (Fig. 4a): per-residue ESM-2 [13] embeddings at 34 binding-pocket positions and chirality-aware Morgan fingerprints of substrate SMILES are projected into a shared space by symmetric InfoNCE (Methods). Its open-vocabulary design — predicting via cosine similarity to encoded fingerprints rather than fixed classification — can in principle rank substrates absent from the training data, which matters where novel substrates are expected. We do not demonstrate that capability here. The same minimum-class-count filter that defines the 54 evaluated classes also removes every substrate class that would have been unseen: 19 classes carried by 21 test domains, each with one or two examples across the whole dataset. After filtering, the number of test classes absent from the training split is zero, so no accuracy we report bears on the open-vocabulary case. PARAS [31] reports higher substrate-specificity accuracy for bacterial adenylation domains using a fixed substrate vocabulary; we have not benchmarked BGCast against it, and it does not address the archaeal case. The model was trained on 8,744 A-domains (3,653 PARASECT + 2,798 MIBiG + 2,293 MASPR), all bacterial, covering the 54 substrate classes with at least 10 labelled examples (below, on why this is 54 and not 60).

Performance on bacterial benchmarks (held-out test split, n=1,191; bootstrap 95% CIs from 1,000 resamples on the model-comparison set):

Table 1. BGCast performance and comparison with existing tools.

Metric	BGCast	antiSMASH 7.0	PRISM 4	DeepSEMS*
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	Architectural; not demonstrated	No	No	No
Novelty scoring	Yes	No	No	No
Confidence calibration	Yes (see below)	–	–	–

A stored checkpoint reports 76.4% top-1 and 85.5% top-3 over 60 substrate classes. That checkpoint was trained on 1,993 self-generated pseudo-labels alongside the 8,744 labelled domains, and six of its sixty classes carry no labelled examples at all; re-running the same configuration on labelled data alone returns the 54 classes, 5,701 training domains and accuracies reported above, which are the figures used throughout (Methods; `bin/audit_substrate_model_provenance.py`, `bin/audit_substrate_ablation.py`).

The 74.7%/84.1% figures are upper bounds, because the train and test splits are not disjoint at the level of the model's input. BGCast sees only 34 binding-pocket residues, and 881 of the 1,191 test domains (74.0%) carry a 34-residue pocket string byte-identical to that of at least one training domain, 727 (61.0%) to a training domain with the same substrate label. Identity at the full domain sequence is rarer but still substantial: 290 test domains (24.3%) are byte-identical to a training sequence, 184 (15.4%) with the same label. Deduplication across the three source datasets was by exact full-sequence

match, which does not remove either class of overlap. A pocket-disjoint re-split is required to establish a leakage-free accuracy, and we have not run one; the reported figures should be read as the ceiling rather than the expected performance on genuinely novel domains.

DeepSEMS (Xu et al., Nat. Comput. Sci.* **6**, 753-766, 2026 [10]; originally posted as a 2025 bioRxiv preprint) claims Tanimoto "up to 0.6"; our independent validation on the MIBiG benchmark yielded 0.266 mean Tanimoto over 647 BGCs. We have not re-checked this figure against the published version's own reported metric on a matched population, only against the preprint's headline number.

Structure prediction was evaluated on 453 MIBiG NRPS BGCs: mean Tanimoto 0.374 overall, 0.489 for cyclic BGCs and 0.337 for linear, against 0.205 for de novo assembly alone and 0.376 for template matching alone, with an oracle ceiling of 0.416 (`bin/optimal_structure_predictor.py`). The combined figure does not exceed template matching alone: the 30 of 453 BGCs the pipeline routes to assembly instead score mean 0.182, against 0.208 had template matching been used on the same 30 — a fixable routing-threshold issue, not evidence combining methods is unhelpful, and more than the naive 0.416-minus-0.374 gap may be recoverable by fixing it. Separately, the template library and the 453 scored BGCs share 21 compounds each produced by 2-4 different BGCs, so for roughly 48 of the 453 a "correct" match may retrieve a different BGC producing an identical compound rather than an independent prediction — an inflation we have not quantified.

Confidence is well calibrated on the model-comparison set (n=1,107, overall accuracy 72.2%): predictions above 0.9 confidence are 86.7% accurate (n=360), 71.7% at 0.7-0.9 (n=590), 40.8% at 0.5-0.7 (n=152), and 40.0% for the 5 domains below 0.5 — a different evaluation set from the 74.7% headline figure, and the two should not be read as one experiment.

BGCast does not outperform antiSMASH on bacterial substrate prediction, and its structure prediction (0.374) is below PRISM 4 (0.5-0.7). Its distinctive contributions are open-vocabulary prediction (architectural only, and untested here for the reason given above), confidence calibration, and novelty scoring (1 minus the maximum Tanimoto similarity to 3,112 MIBiG compounds) — capabilities absent from the other tools compared here, though we have not surveyed the broader dereplication literature for equivalents.

Critical caveat: BGCast was trained entirely on bacterial data. Its accuracy on archaeal sequences — which diverge 26-39% at the protein level — is unknown, unvalidated, and uncalibrated; archaeal predictions should be treated as hypotheses.

Scope of the predicted structure library: of a 1,027,623-structure library assembled at scale across biosynthetic classes, only the 482,200 non-ribosomal peptide structures are per-BGC predictions — the 292,037 RiPP entries all share a single SMILES string and the 253,000 polyketide entries resolve to fourteen molecular weights, both templates repeated per BGC, not predicted chemistry. Fig. 4d therefore uses the MIBiG reference set (n=3,112), not the assembled library. The 1,617 assembled archaeal structures collapse to 672 unique SMILES similarly (polyketide 201->2 unique, non-ribosomal peptide 430->78, only the RiPP fraction of 986->592 chemically diverse); we draw no quantitative conclusion about archaeal chemical novelty from these structures.

Detection-rule coverage of known archaeal compound classes

The survey and BGCast stand on their own regardless of why BGC-centric tools miss what they miss. We then asked whether archaea are structurally harder for these tools to detect than bacteria are: for 12 experimentally validated archaeal bioactive compound classes, we evaluated whether the pathway conforms to BGC organization and whether the relevant enzyme families are in current detection rule sets (antiSMASH 7.0 [12], PRISM 4 [11], BiG-SLiCE [15], DeepSEMS [10]; Fig. 1a, Table 2; antiSMASH 8.0 [30] was released after this assessment and its archaeal rule coverage is untested here). Different tools recover different, only partially overlapping archaeal BGC sets from the same genomes [18], so no single tool's output is a census. Genomic background and ecology also shape measured BGC diversity [36, 42], and machine-learning approaches to bioactivity and cluster prediction [39, 40] and to community-scale surveys [41] are converging on the same populations from other directions.

Table 2. Detection of known archaeal bioactive compound classes by BGC-centric tools.

Compound class	Bioactivity	Pathway organization	antiSMASH 7.0 rule status	Detection status
Lanthipeptides	Antimicrobial	Canonical RiPP BGC	Rule exists (RiPP/lanthipeptide)	Fully detectable
Bacteriocins (some)	Antimicrobial	Variable, some BGC-like	Rule exists for some RiPP subclasses; bacteriocin diversity exceeds rule coverage	Partially detectable
Terpene synthases	Various	Some detected as terpene BGCs	Terpene rule exists	Partially detectable

Compound class	Bioactivity	Pathway organization	antiSMASH 7.0 rule status	Detection status
TfuA-YcaO loci	Unknown in archaea	Some resemble bacterial BGCs	Thioamitide rule exists since antiSMASH v5.1 (originally "TfuA-related RiPPs," renamed "thioamitides" in v6.0)	Partially detectable
Halocins	Antimicrobial	No BGC organization known	No rule in any tool	Undetectable — no rule exists
Sulfolobicins	Antimicrobial	No homologs in any database	No rule in any tool	Undetectable — no rule exists
Bacterioruberin	Anticancer, antioxidant	4 dedicated genes co-located (4.7 kb); remainder dispersed (Scattered pathways, below)	Terpene rule would plausibly detect the co-located window alone	Undetectable — a clustering problem, not a missing-rule problem
Diketopiperazines	Quorum sensing	Single-enzyme (CDPS)	tRNA-dependent CDPS rule exists since antiSMASH v5.0	Rule exists; fit to the archaeal pathway untested
NIS siderophores	Iron chelation	NRPS-independent	IucA/IucC-like NI-siderophore rule exists (rebuilt/renamed in v7.0; widened in v7.1)	Rule exists; fit to the archaeal pathway untested
Ectoine/hydroxyectoine	Extremolyte	3-4 gene operon	Ectoine rule exists since antiSMASH ≤v3.0	Rule exists; fit to the archaeal pathway untested
Phosphonates	Herbicidal	PEP-mutase-dependent	Phosphonate rule exists since antiSMASH ≤v3.0, updated v7.0	Rule exists; fit to the archaeal pathway untested
Cofactor F430 variants	Enzymatic cofactor	Primary metabolism pathway	No rule; arguably outside the intended scope of specialized-metabolite detection	Undetectable — outside tool scope

Rule status was checked against the antiSMASH documentation glossary and changelog, not by running antiSMASH against these archaeal pathways. Of the 12 classes, 1 is fully detectable and 3 are partially detectable by current tools' own account. Of the remaining 8, only 2 (halocins, sulfolobicins) have no detection rule in any tool we checked; 1 (bacterioruberin) is a dispersed-pathway problem rather than a missing-rule problem (Scattered pathways, below); 1 (cofactor F430) is arguably outside the scope specialized-metabolite tools are built for; and 4 (ectoine, phosphonates, NIS siderophores, diketopiperazines via CDPS) have bacterial-derived rules whose sensitivity to the archaeal version of the pathway is untested — a different claim from "no rule exists." Crediting those four as still undetectable pending an archaeal test reproduces the previous estimate, $(1 + 3\times0.5)/12 = 20.8\%$, rounded to 21%, for a corrected reason; crediting them at the same 0.5 weight as the other partials gives $(1 + 7\times0.5)/12 = 37.5\%$; crediting them fully gives $(5 + 3\times0.5)/12 = 54\%$. We report this **21-54% range** rather than a single figure, because the right treatment depends on an archaeal antiSMASH run we have not performed. This is a literature-scoring index over 12 curated classes, not a measured detection rate, and should not be read onto the 3,974-BGC survey above, a different population with a different denominator.

These 12 classes span antimicrobials, an anticancer agent, iron chelators, an extremolyte, quorum-sensing molecules, agricultural compounds, and enzymatic cofactors, so the gap is not restricted to one pathway type — but they are not uniformly well-supported as archaeal natural products: the NIS-siderophore gene cluster in *H. volcanii* was reportedly acquired by lateral transfer from bacteria, and archaeal phosphonates and diketopiperazines rest on genomic-biomarker evidence rather than an isolated compound, so those two should be read as weakly supported pending a fuller literature audit.

We tested two further candidate explanations for the gap, each against a matched bacterial comparison.

Scattered pathways: a case study and its bacterial control

The BGC paradigm assumes that genes encoding a single biosynthetic pathway are physically co-located on the chromosome. For many archaeal pathways, this assumption is false.

We mapped the complete bacterioruberin biosynthetic pathway in *Haloferax volcanii* DS2 (GenBank accession CP001956.1, 2.85 Mb main chromosome; four additional replicons totaling ~1.17 Mb were not screened) as a case study (Fig. 2a). Bacterioruberin is a C50 carotenoid with demonstrated anticancer activity against leukemia cell lines [21]. Our signature search recovers 17 loci in total: four ded-

icated carotenoid enzymes co-located in a single 4.7 kb window (phytoene synthase HVO_2524, bisanhydrobacterioruberin hydratase HVO_2526, lycopene elongase HVO_2527, carotenoid 3,4-desaturase HVO_2528), and 13 additional loci matched by generic "geranylgeranyl"/"squalene synthase" keywords that are core isoprenoid-precursor and membrane-lipid enzymes an *H. volcanii* strain making no bacterioruberin at all would still carry — not carotenoid-specific paralogs. We therefore do not say the pathway "requires" 17 loci; 4 are dedicated, 13 are shared precursor-supply genes.

The 17-locus set spans 2,298,501 bp — 81% of the 2,847,757 bp chromosome (~57% of the ~4.0 Mb genome including unscreened replicons), with a maximum inter-gene gap of 520 kb (519,668 bp); only the four dedicated genes are co-located. We did not run antiSMASH on this genome, but based on its published terpene rule we expect it would detect at most the co-located window — the full pathway would be invisible as a single BGC because it is not organized as one.

We have not established that this dispersal is more common in archaea than bacteria, and a direct test weighs against it: we ran the identical signature, unmodified, against all eight bacterial genomes used in the annotation-gap control below (`bin/audit_bacterioruberin_signature_bacteria.py`). Five (*E. coli*, *B. subtilis*, *P. aeruginosa*, *S. aureus*, *T. thermophilus*) return zero hits — the signature is not generically noisy. The other three, all carotenoid/isoprenoid producers, register as scattered: *Synechocystis* sp. PCC 6803 (3/6 enzymes, span 1.27 Mb, 36% of genome), *M. tuberculosis* H37Rv (3/6, 3.65 Mb, 83%) and *S. coelicolor* A3(2) (4/6, 8.03 Mb, 93%) — the latter two nominally exceeding the archaeal case's 81% span, though *S. coelicolor*'s own genome is unusually large and its secondary metabolism famously dispersed. Inspection shows *S. coelicolor*'s core carotenoid genes are genuinely clustered in two compact windows, but the same generic precursor-supply keywords that flag 13 "shared" loci in *H. volcanii* also fire on scattered, non-carotenoid isoprenoid genes there. This is a real limitation of the detection method's generic keywords, not a demonstrated property of archaeal biology, and weighs against reading the bacterioruberin case as archaea-specific evidence. Separately, NIS siderophore biosynthesis (180 BGCs in BiG-SLiCE across at least 30 archaeal genera) follows a similarly non-clustered pathway, though the *H. volcanii* cluster itself was reportedly acquired by lateral transfer from bacteria; searching for NIS-related keywords in its genome annotation found zero matches despite published siderophore production, illustrating the compounding effect of scattered organization and annotation gaps together.

To address scattered pathways generally, we developed a non-BGC pathway detection framework searching for diagnostic enzyme families regardless of genomic organization (Methods), with keyword signatures for 10 pathway types: bacterioruberin, NIS siderophores, cyclodipeptide synthases, terpene synthases, ectoine, TfuA-YcaO thioamidation, phosphonates, modified mevalonate pathway, cofactor F430, and halocin/archaeocin production. Validated against three reference genomes with known biology (Extended Data Table 2), it detected four pathways in *H. volcanii* DS2 (bacterioruberin at 5 of 6 core enzymes, completeness 0.83; terpene synthase; modified mevalonate; halocin-like), six in *M. acetivorans* C2A, and four in *S. acidocaldarius* DSM 639 (bacterioruberin at 3 of 6 enzymes). Detection of halocin-like genes in all three reflects the keyword signature's breadth (generic "antimicrobial"/"antibiotic"/"toxin" text, not just "halocin") rather than evidence of production; the NIS-siderophore signature is similarly broad, since two of its four keyword categories are iron-uptake rather than biosynthesis terms, so it can register a genome that only imports another organism's siderophore.

The framework relies entirely on existing annotations, which are sparse for archaea, and we measured its false-positive rate: run against 40 bacterial reference BGC records not expected to encode halocins, the halocin-like signature fired on 7/40 (17.5%) via the same generic keywords, on a test population enriched for antibiotic-related annotations and therefore favorable — the true false-negative rate is unmeasured. It also does not distinguish synthesis from uptake (at least for NIS siderophores) and cannot discover pathways with no known enzyme signature at all — sulfolobicins, for example, would remain invisible [20]. It is a starting point.

The annotation gap: a case study and its bacterial control

Even when archaeal biosynthetic genes are present and detectable in principle, they are frequently invisible in practice because they lack functional annotations.

We quantified the annotation gap across eight archaeal genomes (Fig. 2b, Extended Data Table 3): five Euryarchaeota/Halobacteriota, two Thermococci and one Crenarchaeota (Methods; no claim is made about Thaumarchaeota, Asgardarchaeota or Korarchaeota, none analysed). The fraction of proteins annotated as "hypothetical" or equivalent ranged from 13.1% (*S. acidocaldarius* DSM 639) to 56.4% (*H. salinarum* NRC-1), with a mean of 40.2%. *M. acetivorans* C2A, one of the most extensively studied methanogens, is second-highest at 54.5% (2,475 of 4,540 proteins) — more than half its proteome has no assigned function.

Is this archaea-specific? A matched bacterial control says no. A companion manuscript applied the identical keyword rule to eight bacterial genomes spanning curation intensity, giving a raw mean of 25.0% — but *Synechocystis* sp. PCC 6803 scores 0.0% by keyword while 54.1% of its coding sequences have an empty `/product` field, an annotation-dialect artifact. Counting blank products as unannotated, the bacterial mean rises to 32.1%, and a two-sided Mann-Whitney test on the corrected archaeal (40.2%) and bacterial (32.1%) figures gives $U = 24$ against a critical value of 13 at $n = 8$ per group: **the two domains do not separate**. *S. aureus* NCTC 8325 (59.0% corrected) is less completely annotated than every archaeon in our set, and grouping by curation intensity is more informative than grouping by domain (intensively cu-

rated bacterial genomes average 12.0% hypothetical; well-studied and environmental ones average 44.2% and 44.0%, closer to the archaeal figure). The gap we measure in archaea is real; it is not shown to be distinctively archaeal rather than a consequence of curation intensity.

Any biosynthetic enzyme hiding among these hypothetical proteins would be invisible to annotation-based mining regardless of clustering. To estimate how many are recoverable, we applied CLEAN [8] and DeepFRI [27] — both run on sequence alone, DeepFRI in its CNN mode rather than its structure-based graph-convolutional mode (Methods). Of 2,476 hypothetical *M. acetivorans* proteins submitted to CLEAN (2,475 above; marginally different string criteria, no percentage below is affected), 93 (3.8%) received high-confidence enzyme predictions (Fig. 2c). Among the 93, 44 fall in EC classes 1 and 2 — oxidoreductases (8) and transferases (36), including methyltransferases/glycosyltransferases/acyltransferases commonly involved in biosynthesis and tailoring — membership alone is not evidence of a secondary-metabolic role. DeepFRI, run on the 1,081 hypothetical proteins among the same 4,540, predicted 54 high-confidence EC assignments (5.0% of those inputs) and 166 high-confidence molecular-function GO terms (15.4%) (Fig. 2d), including metal-ion/cofactor binding and catalytic activity — consistent with biosynthetic roles but absent from the original annotation.

Combined recovery. Cross-tabulating the two tools' high-confidence outputs over the 2,455 hypothetical proteins both scored, CLEAN recovers 93 and DeepFRI 274, for a union of 356 proteins (14.5% of that set); the intersection is 11 proteins (Jaccard index of 0.03). This is not below chance: independence predicts 10.38 shared proteins given each tool's marginal coverage, and 11 were observed (Fisher's exact OR 1.07, 95% CI 0.59-2.09, p=0.87) — the near-disjointness two tools with this little coverage would produce picking proteins independently, not evidence of complementarity. Still, 85.5% of hypothetical proteins receive no high-confidence assignment from either method (`bin/reconcile_function_prediction_recovery.py`). We did not run either tool beyond *M. acetivorans*, so we draw only the qualitative conclusion that assignable-function enzymes exist among proteins the annotation pipeline left blank; both tools were trained predominantly on bacterial/eukaryotic proteins, and their accuracy on archaeal sequences (which diverge 26-39% for thiopeptide-related enzymes) is not established.

A structure-based approach outside our own analysis shows what this recovery can look like when it works: Samusevich et al. [16] predicted terpene synthase function directly from structure and validated three active terpene synthases in archaea, the first ever described from that domain and invisible to sequence-based baselines. Whether such an approach generalizes turns on whether an enzyme's fold still constrains its function, which is not guaranteed.

A training-population artefact in biosynthetic-protein classification

A biosynthetic-protein classifier built for prioritisation shows how a training population, rather than the property being predicted, can determine a score.

Logistic regression on ESM-2 embeddings of 17,056 MIBiG biosynthetic against 8,528 non-biosynthetic proteins reaches ROC-AUC 0.995 and F1 0.984. Trained with archaeal housekeeping proteins as its only negatives, it scores the *M. acetivorans* locus MA_4560 at 0.988; adding bacterial non-biosynthetic proteins to the negative class drops the same locus to exactly 0.000 at unchanged cross-validated performance (Fig. 3c). The model had separated bacterial from archaeal proteins rather than biosynthetic from non-biosynthetic, and a confound of this kind is invisible to every cross-validation split because the split inherits it.

TfuA-associated loci in the surveyed population

The survey's TfuA-locus count comes from a database query over other archaeal genomes, independent of the MA_4560 case. A BiG-SLiCE query — citing a Pfam accession (PF02142) that is both wrong and never actually used, since detection was a string match on an inherited label (Methods) — returns 500 loci, of which 27 (5.4%) have an identifiable YcaO partner within the called cluster only. No deduplication step applies: each of the 500 is a distinct BGC and all 500 lie within the QC'd population of 3,974. The 498 quoted in earlier versions of this analysis is the subset also carrying genus-level taxonomy, two of the 500 lacking a genus assignment; it is a count of BGCs, not of raw rows. In bacteria, TfuA-YcaO produces thioamitides with substantial cancer-cell selectivity (thioviridamide IC_{50} = 3.9 ng/mL, 6-fold selective [22]); in methanogens TfuA's established role is thioamidation of MCR, a housekeeping function [7], and no archaeal thioamitide has ever been isolated.

A superseded benchmark: the MA_4560 locus identity

The seven-method benchmark we ran on MA_4560 supports no claim about annotation transfer, because the locus is not the protein we took it to be. MA_4560 (UniProt Q8THF7) carries Pfam PF31122 and PF08350, placing it in the winged-helix-turn-helix FilR1 transcriptional-regulator family; the *M. acetivorans* TfuA is MA_0164 (UniProt Q8TUA9, Pfam PF07812), paired with *ycaO* at MA_0165 and crystallised as PDB 6XP8 [28], and no method in that benchmark was applied to it. Read against MA_4560's actual domain composition, most of the seven returned consistent identifications of a regulator; the predictions as measured are retained in Extended Data Table 6. We could not rerun Foldseek, ESM-2, ProTrek or the classifier against MA_0164 with the infrastructure available; from bulk output computed for the annotation-gap analysis, CLEAN scores MA_0164 at 0.000 and DeepFRI returns no confident function or DNA-binding prediction, so whether the real TfuA is itself hard to annotate re-

mains open. Identity was resolved from the UniProtKB API (`bin/verify_ma4560_identity.py`), and the companion manuscript treats the misidentification as a failure mode.

Complementary approaches

No single computational approach addresses everything BGC-centric detection misses, and Table 4 maps available tools to the two candidate factors that survived testing.

Table 4. Computational approaches to the two tested factors.

Factor	Problem	Available tools	Recovery	Gaps
Scattered pathways	Biosynthetic genes not co-located	Non-BGC detector (this work, 10 signatures; measured 17.5% false-positive rate on a favorable bacterial test set); Spacedust [25], pLM-BLAST (future: cross-genome synteny)	Validated on 3 genomes; 10 pathway types	Novel pathways with no known enzyme signatures; false-negative rate unknown; does not distinguish synthesis from uptake
Annotation gap	40.2% of proteins are "hypothetical" (not distinguishable from a matched bacterial control)	CLEAN (3.8% recovery); DeepFRI (5.0% EC, 15.4% MF GO); ProTrek [24], Foldseek [9] (future)	14.5% of the dark proteome illuminated (union, chance-level overlap between the two tools)	85.5% remains dark; archaeal training data scarce; accuracy on divergent sequences unknown

We deliberately do not put a number on how much of the 21-54% detection-rule gap these tools recover, given the false-negative rate of the non-BGC detector is unmeasured and 85.5% of the annotation gap remains dark. The third candidate factor is not in this table because its one test case was a misidentified protein (above).

De la Fuente et al. [26] bypassed BGC detection, functional annotation, and fold-based function prediction entirely, scanning raw proteomes for antimicrobial-peptide signatures — 93% of 80 tested archaeasins showed antimicrobial activity. This is complementary, not competing: proteome-based AMP discovery suits small peptides, while BGC-based mining remains the practical route to larger, more complex natural products.

Discussion

A catalog and a substrate predictor for the detectable fraction

The 3,974-BGC survey and BGCast are, as far as we are aware, the largest archaeal BGC catalog assembled to date and the first open-vocabulary substrate predictor benchmarked on this population, and neither depends on any claim about why archaea are or are not hard to detect. Both are usable now — the survey as a starting catalog for the detectable fraction, BGCast for substrate and novelty prediction on any BGC, archaeal or bacterial, with the caveat that its archaeal confidence scores are uncalibrated. We think this pairing — a resource plus an honest benchmark — is a more durable contribution than the detection-gap analysis that follows.

The detection shortfall is not archaea-specific

We set out to characterize archaea as structurally distinct from bacteria in how their biosynthetic capacity evades BGC-centric detection, via two candidate factors: scattered genomic organization and functional annotation gaps. Tested against matched bacterial controls (Results), neither held up as archaea-specific — the annotation gap's matched control does not separate the domains, and the bacterioruberin signature reproduces comparable or larger apparent dispersion in three bacterial carotenoid producers, which we attribute to the method's generic keywords rather than archaeal genome organization. We present both as general genome-mining challenges illustrated here in archaea; uncultured bacterial lineages and early-branching eukaryotes may face them for the same reasons, though we have not tested that.

A third candidate factor, that an enzyme's fold shared with an

unrelated family defeats annotation transfer, is not assessed here: its only test case was a misidentified locus (Results), so no verified instance remains in this paper.

On present evidence, the case for archaea being structurally distinct in how their biosynthetic capacity evades detection is weaker than a first pass through this analysis suggested. What remains is narrower: current BGC-centric tools plausibly under-detect archaeal biosynthetic diversity for reasons that are mostly general genome-mining challenges (annotation provenance, generic detection heuristics, unverified identifiers) rather than archaea-specific phenomena, illustrated here in archaea because that is where we looked. The framework is built from case studies — we have not shown these factors exhaustively account for the gap, or that additional factors do not exist.

Relationship to recent experimental advances

Three recent experimental studies converge on archaea being a rich source of bioactive chemistry, independent of anything above. De la Fuente et al. [26] identified 12,623 candidate antimicrobial peptides from proteome-wide mining and validated 93% antimicrobial activity, bypassing the BGC paradigm entirely. Song et al. [3] heterologously expressed archaeal lanthipeptide BGCs in *H. volcanii* — the one compound class in our analysis fully detectable by BGC tools (Table 2). Liang et al. [2] characterized the first archaeal lanthipeptides. Together they are consistent with the BGC-detectable fraction yielding discoveries through standard approaches while a substantial undetected-or-untested remainder needs different strategies — exactly the population the survey and BGCast are built to work on.

Limitations

Embedding-to-chemistry correlation is weak (r = 0.06-0.29 across our experiments), enough to rank substrates at a binding pocket but not to predict a molecule — why BGCast's substrate prediction is competitive while its structure prediction is not.

Published structure-prediction performance does not reproduce. We reimplemented DeepSEMS and obtained mean Tanimoto 0.266 over 647 MIBiG BGCs against the "up to 0.6" reported in preprint; we have not re-checked against the now-published version's own metric. Our own figures do reproduce (`bin/optimal_structure_predictor.py`: 0.374 overall, 0.489 cyclic), contingent on a BGC-level embedding file that despite being called "ESM-2" elsewhere in this manuscript is actually **ESM-C (600M)** — a different model. Supplying a different embedding cache silently drops the result to 0.339 by disabling the highest-performing path, with no error raised.

Structure prediction is below the state of the art (0.374 mean Tanimoto vs. PRISM 4's 0.5-0.7), and part of the gap to its own 0.416 oracle ceiling is fixable: the current routing rule loses ground on the 30 of 453 BGCs it sends to assembly instead of template matching (mean 0.182 vs. 0.208 had template matching been used). Separately, ~48 of 453 scored BGCs share a compound with another BGC in the template library, an inflation we have not quantified.

Substrate prediction has an error floor ensembling does not cross. Our contrastive model and NRPStransformer agree on being wrong for 19.1% of a 1,107-domain comparison set; a meta-ensemble reaches 76.7% (95% CI 74.5-79.3%) against an 80.9% oracle ceiling. Their individual bootstrap intervals overlap (72.2% vs. 71.4%), and we do not claim ours is more accurate.

A previously cited "14-17% accuracy drop on held-out taxa" belongs to a different classifier (the scaffold/biosynthetic-class model, evaluated on whole-BGC labels), not BGCast's A-domain-level predictor; no genus- or date-held-out evaluation of BGCast itself exists, and building one needs organism metadata this repository's training data lacks. The figure is therefore withdrawn.

BGCast archaeal accuracy is unvalidated and currently unmeasurable — no archaeal A-domain with a known substrate exists to test it against, and its confidence scores are uncalibrated on archaeal input.

Non-BGC pathway detection is annotation-dependent, with a false-positive rate now measured at 17.5% (7/40) on a favorable bacterial test set; the true false-negative rate is unmeasured, it cannot detect pathways with no known diagnostic enzyme family (sulfolobicins), and at least one signature (NIS siderophores) does not distinguish synthesis from uptake.

Function-prediction tools were applied outside their training domain (CLEAN, DeepFRI: predominantly bacterial/eukaryotic), their archaeal accuracy is not established, and recovery is small — 14.5% union, 11-protein intersection, no better than the 10.38 chance would predict (Fisher's exact OR 1.07, p=0.87) — so 85.5% of hypothetical proteins remain unassigned and neither tool cross-validates the other.

The detection-gap accounting is a literature-scoring exercise, not a measurement against ground-truth archaeal pathways; the 12 classes are a biased, characterized-only sample, and at least two (phosphonates, diketopiperazines) rest on genomic-biomarker evidence rather than an isolated compound.

No experimental validation is reported. Every claim here is computational; no archaeal thioamitide has been isolated and the TfuA secondary-metabolism hypothesis rests on genomic context alone.

Phylogenetic and sampling bias: detection is biased toward well-sequenced phyla (Halobacteriota, Thermoplasmatota) and BGC-like pathways, not normalized by genome count per phylum; 47% of surveyed genomes are gut-metagenome-assembled, concentrating the survey toward gut-associated methanogens; the eight annotation-gap genomes cover only

three phyla, with no claim made about Thaumarchaeota, Asgardarchaeota or Korarchaeota.

Future directions

Spacedust [25] and pLM-BLAST could help identify conserved but scattered pathway components via cross-genome synteny and remote homology; systematic heterologous expression in *H. volcanii* [3] and activity-guided fractionation remain essential for ground-truthing predictions. Multimodal models trained on the relationship between protein fold and catalytic mechanism, rather than evolutionary homology, are a reasonable direction for fold-convergent annotation-transfer failures, though this paper offers no evidence for or against it — the MA_4560 case that would have motivated it is withdrawn (Results).

Methods

Archaeal BGC survey

BGCs were identified by antiSMASH (version not recorded) with BiG-SLiCE v1.1.2 [15] providing clustering and GTDB r207 taxonomy. Of the 1,681 post-QC genomes, 898 carry GCA_ isolate-assembly accessions and the remaining 783 (47%) carry metagenome-assembled or WGS-project accessions of mixed provenance: 111 RUG (rumen), 19 MGYG-HGUT (human gut), 18 TOBG (marine), 634 four-letter WGS accessions whose source environment is not recorded in the retained files, and the single MIBiG reference entry (BGC0001554.1). The two groups sum to 1,681; the MIBiG reference is counted within the 783, not alongside it. These are genomes with at least one detected BGC, not all genomes screened. QC filters: CheckM completeness >50%, contamination <5%, phylum-level taxonomy required (CheckM version not recorded; not independently reproducible from deposited files, as the BiG-SLiCE source database was not retained). Final dataset: 3,974 BGCs from 1,681 genomes after removing 24 without phylum-level taxonomy (total pre-filtering: 3,998 BGCs from 1,698 genomes). Functional classification additionally required genus-level taxonomy, which a further 39 BGCs lack, so category counts are reported over 3,935 BGCs; 193 BGCs contribute more than one class and are assigned a single category by a fixed priority order (`bin/reconcile_archaeal_survey_counts.py`, verified against `results/contamination_control/contamination_control_summary.jsph` and `results/bigslice_taxonomy_mapping.csv`), so counts sum to exactly 3,935. The 961-BGC therapeutic-relevant breakdown (Results) is deduplicated per BGC by the priority order thiopeptide > lanthipeptide > lassopeptide > sactipeptide > T1PKS > NRPS-like > siderophore > bacteriocin (`bin/reconcile_prospective_discovery_counts.py`), moving 2 of 961 BGCs carrying both an NRPS-like and a sactipeptide label to sactipeptide; the eight counts sum to exactly 961. TfuA-associated clusters were identified by an inherited BiG-SLiCE/antiSMASH label ("TfuA-related"), not a Pfam or HMM search we ran (the correct TfuA family, PF07812, is not used for this classification step). The query returns 500 TfuA-related loci, each a distinct BGC and all within the QC'd population of 3,974; 498 of them also carry genus-level taxonomy and so fall within the 3,935 classified BGCs, the other two lacking a genus assignment. No deduplication step applies to this count (Results).

Detection gap quantification

Twelve archaeal bioactive compound classes were identified from a literature review that was not exhaustive (Extended Data Table 1); two (phosphonates, diketopiperazines) rest on genomic-biomarker evidence rather than an isolated compound. For each class we assessed whether the pathway is organized as a canonical BGC, whether a named detection rule exists in the antiSMASH 7.0 [12] documentation glossary and changelog (checked directly, not assumed), and whether any tool has been reported to identify the archaeal pathway specifically. We did not run antiSMASH against any archaeal genome. The headline range was calculated three ways: crediting untested-but-rule-covered classes as undetectable, $(1 + 3\times0.5)/12 = 20.8\%$, rounded to 21%; at the same 0.5 weight as other partials, $(1 + 7\times0.5)/12 = 37.5\%$; crediting them fully, $(5 + 3\times0.5)/12 = 54\%$. We report 21-54% rather than choosing among these.

Non-BGC pathway detection

We defined keyword-based signatures for 10 pathway types: bacterioruberin, NIS siderophores (including iron-uptake keywords that do not distinguish uptake from biosynthesis), cyclodipeptide synthases, terpene synthases, ectoine, TfuA-YcaO, phosphonates, modified mevalonate, cofactor F430, and halocin/archaeocin (broader than its name — includes generic antimicrobial/antibiotic/toxin sub-signatures). Keywords of four characters or fewer used word-boundary matching. Validated against *H. volcanii* DS2 (CP001956.1), *M. acetivorans* C2A (AE010299.1), and *S. acidocaldarius* DSM 639 (CP000077.1). A false-positive check against 40 bacterial MIBiG reference BGC records not expected to encode halocins fired on 7/40 (17.5%), via the generic sub-signature (Results). The bacterioruberin signature was separately run, unmodified, against the eight bacterial genomes used for the annotation-gap control below, three of which are carotenoid/isoprenoid producers (*Synechocystis* sp. PCC 6803, *M. tuberculosis* H37Rv, *S. coelicolor* A3(2)) (`bin/audit_bacterioruberin_signature_bacteria.py`; Results).

Hypothetical protein quantification, CLEAN and DeepFRI

For the eight archaeal genomes under Reference genomes below — five Euryarchaeota/Halobacteriota, two Thermococci, one Crenarchaeota — coding sequences were classified "hypothetical" if the product annotation contained "hypothetical," "predicted protein," "conserved hypothetical,"

"unknown function," or "DUF." The 2,476 hypothetical proteins from *M. acetivorans* C2A were submitted to CLEAN [8] (Yu et al. 2023); high-confidence predictions were those with CLEAN score >=0.5. Metagenomic-DeepFRI v1.0 [27] (Gligorijevic et al. 2021) was run in CNN sequence-only mode (no structures or contact maps; the graph-convolutional mode was not run) on all 4,540 *M. acetivorans* proteins, with recovery reported over the 1,081 classified hypothetical; high-confidence predictions were those with DeepFRI score >=0.5, categorized into EC and molecular-function GO terms.

Foldseek, ESM-2, and ProTrek analyses of MA_4560

The AlphaFold-predicted structure of MA_4560 (UniProt Q8THF7, AF-Q8THF7-F1) — believed at the time to be TfuA, corrected in Results — was queried against AFDB50 using the Foldseek [9] web server (van Kempen et al. 2023; default parameters, 3Di+AA scoring, E-value <0.01); all 963 hits were manually inspected. Pfam assignment (PF08350, FilR1_middle) came from InterPro and was independently confirmed via UniProt's own Pfam scan. ESM-2 [13] (650M; Lin et al. 2023) per-residue embeddings, mean-pooled (1280-d), were generated for all 4,540 *M. acetivorans* proteins and MA_4560 ranked by cosine similarity to a "biosynthetic enzyme" centroid from MIBiG proteins (rank 1903/4540). ProTrek [24] (Su et al. 2025) queried MA_4560 against fifteen functional descriptions via its protein-text alignment mode (full ranked list, Extended Data Table 5); the top five were "winged helix-turn-helix transcription factor" (-0.035), "post-translational modification enzyme" (-0.049), "thioamide bond formation" (-0.054), "biosynthetic enzyme" (-0.056) and "thioamidation enzyme" (-0.059). Rankings are wording-sensitive: "transcriptional regulator" ranked tenth and "TfuA family protein" thirteenth.

BGCast architecture and training

Input: ESM-2 (650M) [13] frozen per-residue embeddings at 34 binding-pocket positions, calibrated from PARASECT signature positions via HMM column voting. Model: 2-layer transformer encoder (4 heads, 256 hidden dimension) with learned attention pooling to a 128-dimensional space. Contrastive target: Morgan fingerprints (2048-bit, chirality-aware, radius 2) encoded by a 2-layer MLP to 128 dimensions. Loss: symmetric InfoNCE (fixed temperature tau=0.07; learnable temperature caused gradient explosion on Apple MPS) plus 0.5-weight auxiliary classification loss, same-class negative masking, and random masking of 0-3 pocket positions per sample during training. Training data: 8,744 labelled A-domains (3,653 PARASECT + 2,798 MIBiG + 2,293 MASPR), all bacterial; classes with fewer than 10 labelled examples were excluded, leaving 54 classes, 5,701 training domains, 1,191 test (Results; a stored model reporting 60 classes and 6,065/1,272 does not reproduce from labelled data alone). Of 2,005 candidate pseudo-labels generated from BiG-SLiCE domains at a 0.95 confidence floor, 1,993 entered the embeddings file that checkpoint trained on, so its training run loaded 10,737 rows where the file at that path now holds 8,744. The checkpoint's recorded arguments report no pseudo-labels; that field records the command-line flag rather than the data, and an earlier audit of ours read the flag and concluded none were used (`bin/audit_substrate_model_provenance.py`). Bootstrap confidence intervals (1,000 resamples) were computed on the model-comparison set. HMMER v3.3.2 was used for A-domain extraction (AMP-binding and AMP-binding_C HMMs, E < 1e-5).

Structure prediction and novelty scoring

Structure prediction combined template-based matching (substrate overlap similarity to MIBiG compounds, weighted by BGC-level cosine similarity of embeddings from **ESM-C 600M** [`ESMC_600M_202412`] — distinct from the ESM-2 embeddings BGCast's own A-domain pocket encoder uses) with de novo assembly (RDKit peptide-bond formation, head-to-tail macrocyclization for thioesterase-domain BGCs; cyclization detection via RDKit ring analysis, macrocycles >=8 atoms) for unmatched BGCs. The template library and the 453 scored BGCs share the same MIBiG population without filtering shared compounds (21 compounds each produced by 2-4 BGCs); this inflation is unquantified. Novelty was defined as 1 minus the maximum Tanimoto similarity to any of 3,112 MIBiG compounds, computed via RDKit v2023.09.5 Morgan fingerprints (radius 2, 2048 bits, chirality-aware).

DeepSEMS comparison and sequence divergence analysis

DeepSEMS [10] was reproduced on our MIBiG validation set following the original 2025 bioRxiv protocol (DOI: 10.1101/2025.03.02.641084; since published as Xu et al., *Nat. Comput. Sci.* **6**, 753-766, 2026), giving mean Tanimoto 0.266 over 647 BGCs against the preprint's "up to 0.6" (`results/deepsems_tanimoto_validation.parquet`); we have not re-checked against the published version's own metric. BLASTp v2.17.0 (E-value < 10^{-5}, query coverage >50%) compared archaeal thiopeptide-like sequences (44 BGCs across 29 genera) against 22 characterized bacterial thiopeptide BGCs from MIBiG 3.0, summarized as mean and range over 44 per-BGC best hits and 114 core-enzyme alignments; raw BLASTP outputs were not retained (`results/sequence_divergence/blastp_provenance.json`), so this step is not independently reproducible.

Reference genomes

Eight archaeal genomes were used for the annotation-gap and non-BGC pathway analyses: *H. volcanii* DS2 (CP001956.1), *M. acetivorans* C2A (AE010299.1), *S. acidocaldarius* DSM 639 (CP000077.1), *H. salinarum* NRC-1, *H. mediterranei* ATCC 33500, *Natrinema* sp. J7-2, *P. furiosus* COM1 and *T. kodakarensis* KOD1. [ACCESSIONS NEEDED: the GenBank records for the last five are not recorded in

`results/archaeal_gap_analysis/validation_results.json`, which stores only local filenames.] The first three were also used for non-BGC pathway validation; all eight contributed to the hypothetical-protein statistics.

Data and code availability

BGCast source code, trained models, and analysis scripts are available at [repository URL]. The BiG-SLiCE archaeal BGC survey data, CLEAN and DeepFRI analysis results, and non-BGC pathway detection outputs are provided as supplementary datasets.

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. BGCast source code and trained models are available at [repository URL].

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

References

1. Woese, C.R. & Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088-5090 (1977).
2. Liang, H. et al. Genomic and metabolic analyses reveal antagonistic lanthipeptides in archaea. *Microbiome* **11**, 74 (2023).
3. Song, Z. et al. Decoding the chemical language of ribosomally synthesized and post-translationally modified peptides from the untapped Archaea domain. *Angew. Chem. Int. Ed.* e202501074 (2025).
4. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412-421 (2014).
5. Kumar, V. et al. Halocins, natural antimicrobials of Archaea: exotic or special or both? *Biotechnol. Adv.* **53**, 107834 (2021).
6. Shand, R.F. & Leyva, K.J. Halocins and sulfolobicins: the emerging story of archaeal protein and peptide antibiotics. *J. Ind. Microbiol. Biotechnol.* **28**, 23-31 (2002).
7. 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).
8. Yu, T. et al. Enzyme function prediction using contrastive learning. *Science* **379**, 1358-1363 (2023).
9. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243-246 (2023).
10. Xu, T. et al. DeepSeMS: revealing the hidden biosynthetic potential of the global ocean microbiome with a large language model. *Nat. Comput. Sci.* **6**, 753-766 (2026). doi:10.1038/s43588-026-00983-1. [Originally posted as a 2025 bioRxiv preprint, DOI: 10.1101/2025.03.02.641084, which is the version reimplemented here.]
11. Skinnider, M.A. et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).
12. Blin, K. et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46-W50 (2023).
13. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
14. Terlouw, B.R. et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* **51**, D603-D610 (2023). doi:10.1093/nar/gkac1049
15. 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).
16. Samusevich, R. et al. Structure-enabled enzyme function prediction unveils elusive terpenoid biosynthesis in archaea. *bioRxiv* (2024). [Preprint.]
17. Castro, I. et al. The lanthipeptide biosynthetic clusters of the domain Archaea. *Microbiol. Res.* **253**, 126884 (2021).

18. Sukmarini, L. et al. A survey of the biosynthetic potential and specialized metabolites of archaea and understudied bacteria. *Curr. Res. Biotechnol.* **4**, 478-491 (2022).
19. Charlesworth, J.C. & Burns, B.P. Untapped resources: biotechnological potential of peptides and secondary metabolites in archaea. *Archaea* **2015**, 282035 (2015).
20. Erdmann, S. et al. The sulfolobicin genes of *Sulfolobus acidocaldarius* encode novel antimicrobial proteins. *J. Bacteriol.* **193**, 4380-4387 (2011).
21. Maheshwari, K. et al. Bioactive molecules of haloarchaea: scope and prospects for industrial and therapeutic applications. *Front. Microbiol.* **14**, 1113540 (2023).
22. Hayakawa, Y. et al. Thioviridamide, a novel apoptosis inducer in transformed cells from *Streptomyces olivoviridis*. *J. Antibiot.* **59**, 1-5 (2006).
23. Frattaruolo, L. et al. Thioalbamide, a thioamidated peptide from *Amycolatopsis alba*, affects tumor growth and stemness by inducing metabolic dysfunction and oxidative stress in pancreatic and breast cancer cells. *Biomedicines* **11**, 144 (2023).
24. Su, J. et al. ProTrek: navigating the protein universe through tri-modal contrastive learning. *Nat. Biotechnol.* (2025).
25. Zhang, R., Mirdita, M. & Söding, J. De novo discovery of conserved gene clusters in microbial genomes with Spacedust. *Nat. Methods* **22**, 2065-2073 (2025).
26. Torres, M.D.T., Wan, F. & de la Fuente-Nunez, C. Deep learning reveals antibiotics in the archaeal proteome. *Nat. Microbiol.* (2025). doi:10.1038/s41564-025-02061-0
27. Gligorijevic, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
28. Liu, A., Si, Y., Dong, S.H. et al. Functional elucidation of TfuA in peptide backbone thioamidation. *Nat. Chem. Biol.* **17**, 585-592 (2021).
29. Zdouc, M.M. et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res.* **53**, D678-D690 (2024). doi:10.1093/nar/gkae1115
30. Blin, K. et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Res.* **53**, W32-W38 (2025). doi:10.1093/nar/gkaf334
31. 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
32. Nuhamunada, M. et al. BGCFlow: systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. *Nucleic Acids Res.* **52**, 5478-5495 (2024). doi:10.1093/nar/gkae314
33. Mohite, O.S. et al. Pangenome mining of the *Streptomyces* genus redefines species' biosynthetic potential. *Genome Biol.* **26**, 9 (2025). doi:10.1186/s13059-024-03471-9
34. Zhu, Y. et al. Computational advances in biosynthetic gene cluster discovery and prediction. *Biotechnol. Adv.* **79**, 108532 (2025). doi:10.1016/j.biotechadv.2025.108532
35. Han, Y. et al. Mining microbial and metabolic dark matter in extreme environments: a roadmap for harnessing the power of multi-omics. *Adv. Biotechnol.* **2**, 8 (2024). doi:10.1007/s44307-024-00034-8
36. 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
37. Chen, S. et al. Halocin H4 is activated through cleavage by halolysin HlyR4. *Appl. Environ. Microbiol.* **90**, e02284-23 (2024). doi:10.1128/aem.02284-23
38. Liao, Y. et al. Molecular insights fast-tracked: AI in biosynthetic pathway research. *Nat. Prod. Rep.* **42**, 911-936 (2025). doi:10.1039/d4np00003j
39. Yuan, S. et al. Machine learning-enabled genome mining and bioactivity prediction of natural products. *ACS Synth. Biol.* **12**, 2650-2662 (2023). doi:10.1021/acssynbio.3c00234
40. Lai, S. et al. Deciphering the biosynthetic potential of microbial genomes using a BGC language processing model. *Nucleic Acids Res.* **53**, gkaf305 (2025). doi:10.1093/nar/gkaf305
41. Dong, S. et al. A vast repertoire of secondary metabolites potentially influences community dynamics and biogeochemical processes in cold seeps. *Sci. Adv.* **10**, eadl2281 (2024). doi:10.1126/sciadv.adl2281
42. Salamzade, R. et al. Evolutionary investigations of the biosynthetic diversity in the skin microbiome using lsaBGC. *Microb. Genom.* **9**, mgen000988 (2023). doi:10.1099/mgen.0.000988

Figure Legends

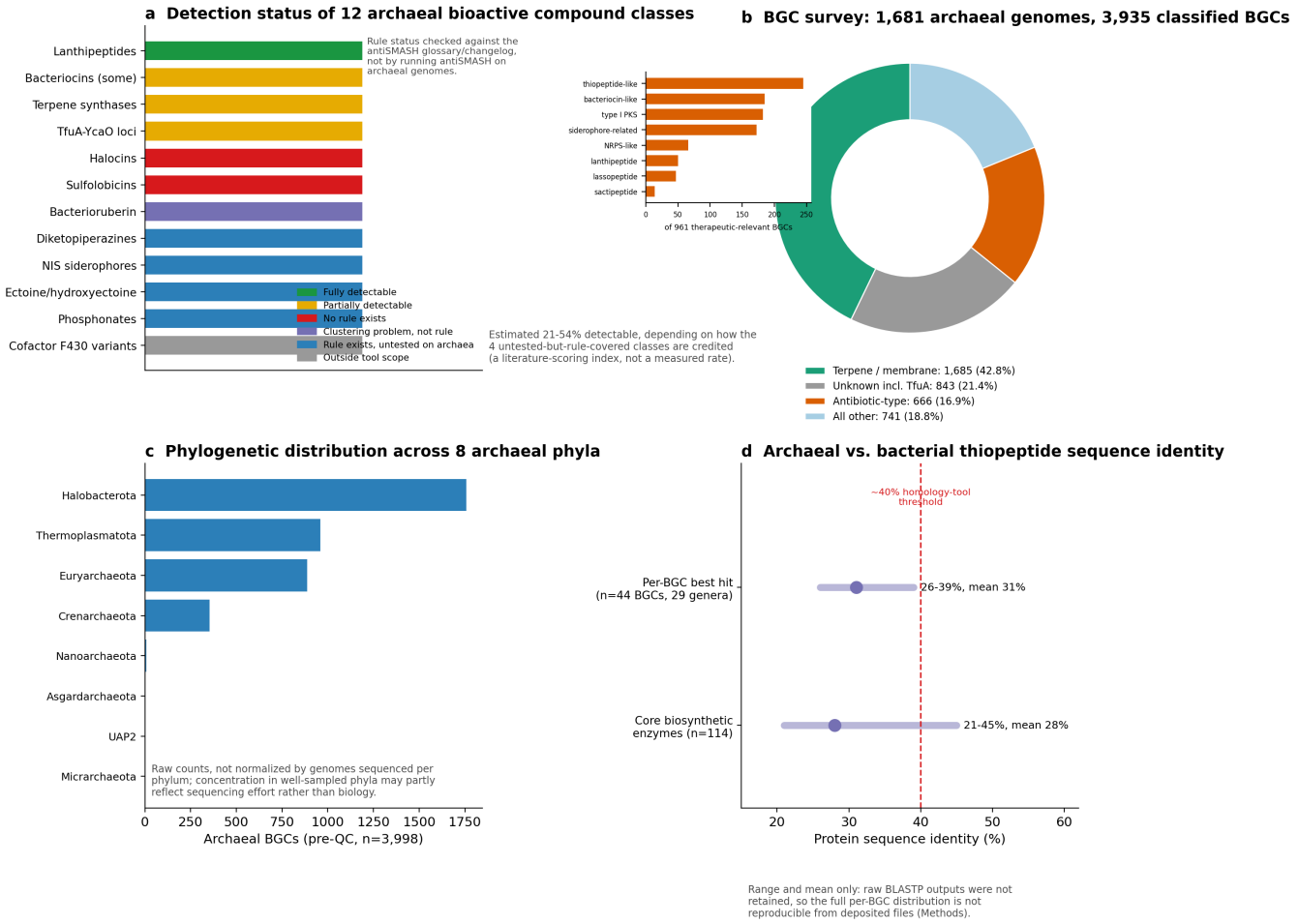

Figure 1. The archaeal biosynthetic landscape and the detection gap. (a) Detection gap analysis: of 12 literature-curated archaeal bioactive compound classes, 1 is fully detectable by current BGC-centric tools' published rule sets, 3 are partially detectable, 2 have no detection rule in any tool, and the remaining 4 (ectoine, phosphonates, NIS siderophores, diketopiperazines) have bacterial-derived rules whose fit to the archaeal pathway is untested — an estimated 21-54% detectable depending on how the untested-but-covered classes are credited, not a single measured rate. Each compound class is shown with its documented bioactivity and its actual rule status (no rule; rule exists but pathway not clustered; rule exists but untested on archaea; outside tool scope; or detectable). (b) BGC survey across 1,681 archaeal genomes (47% metagenome-assembled from human-gut samples): 3,935 BGCs classified by functional category (terpenes 42.8%, unknown/TfuA-associated 21.4%, antibiotic-type 16.9%, other 18.8%). Inset: breakdown of the 961 cross-category therapeutic-relevant BGCs by architecture label, deduplicated per BGC by a documented priority order (summing to exactly 961). (c) Phylogenetic distribution of detected BGCs across major archaeal phyla, not normalized by genomes sequenced per phylum. Antibiotic-type clusters concentrate in Halobacteriota and Thermoplasmatota — phyla from competitive environments (hypersaline and acidic thermal, respectively), though this may also reflect sequencing/sampling effort. (d) Sequence divergence: distribution of protein sequence identity between archaeal thiopeptide-like sequences and characterized bacterial thiopeptides in MIBiG (mean 31%, range 26-39%), illustrating the evolutionary distance that confounds homology-based detection.

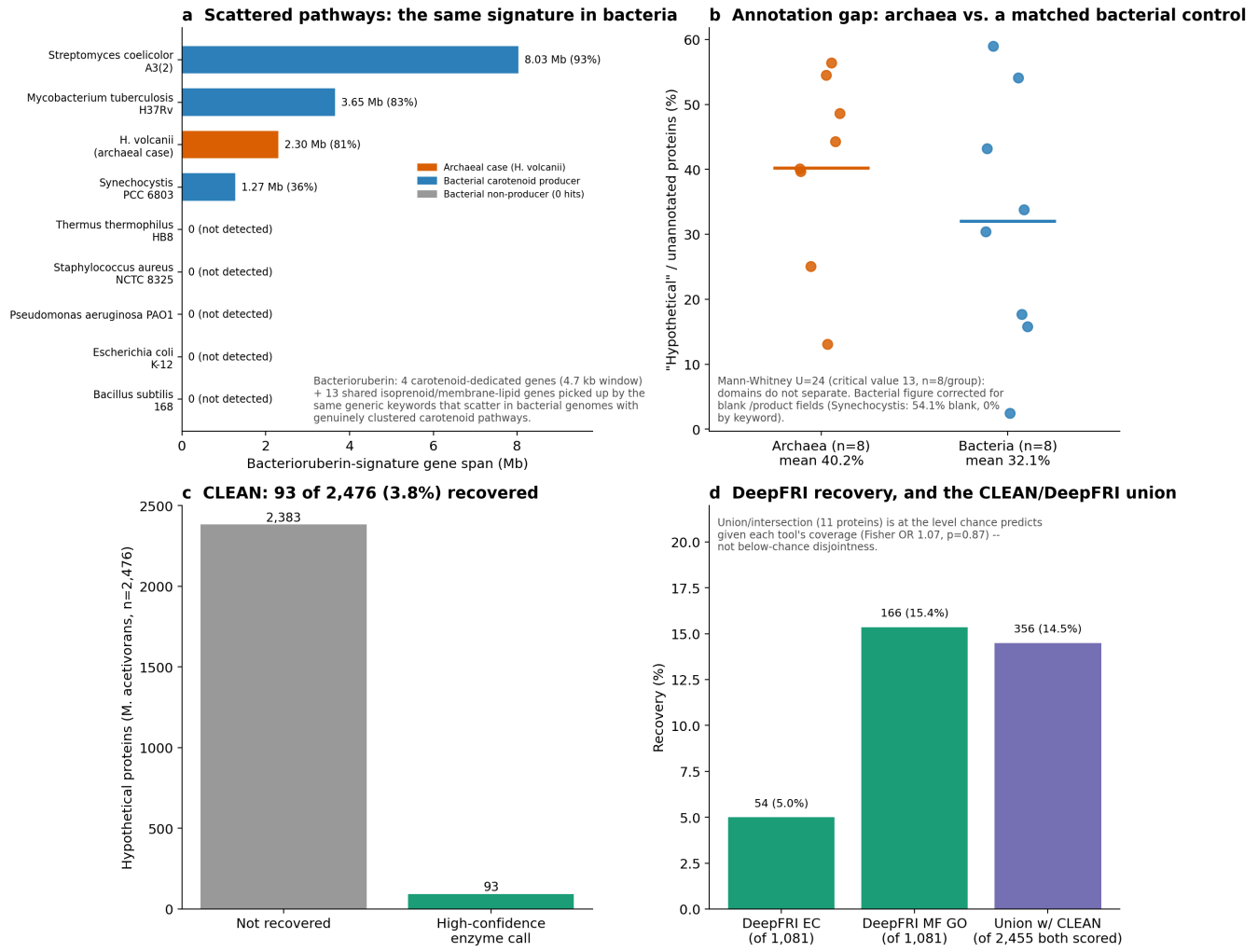

Figure 2. Factors limiting BGC-centric detection in archaea. (a) Scattered pathways: bacterioruberin biosynthesis in *H. volcanii* DS2 recovers 17 loci spanning 2,298,501 bp (81% of the main chromosome; ~57% of the ~4.0 Mb genome including unscreened plasmids), of which only 4 are carotenoid-dedicated and necessary specifically for this pathway; the other 13 are shared isoprenoid/membrane-lipid genes. The gray shading marks the maximum inter-gene distance (520 kb). We expect, but did not test by running antiSMASH, that it would detect at most the 5 kb co-located fragment. (b) The annotation gap: fraction of proteins annotated as "hypothetical" or equivalent across eight representative archaeal genomes (mean 40.2%, range 13.1-56.4%), alongside a matched bacterial control (corrected mean 32.1%) that does not significantly differ from the archaeal figure (Mann-Whitney U=24, critical value 13, n=8 per group). *M. acetivorans* C2A, despite being a model methanogen, has 54.5% hypothetical proteins — second-highest of the eight, not highest. (c) CLEAN enzyme recovery: of 2,476 hypothetical proteins from *M. acetivorans*, 93 (3.8%) received high-confidence enzyme predictions, including 44 in EC classes 1-2 (methyltransferases, glycosyltransferases, oxidoreductases, acyltransferases). (d) DeepFRI recovery, sequence-only mode (no structures used): of 1,081 hypothetical proteins, 54 (5.0%) received high-confidence EC assignments and 166 (15.4%) received high-confidence molecular function GO terms. The union with CLEAN over the 2,455 proteins both tools scored is 14.5% (chance-level overlap between the two tools, not below-chance complementarity).

a Seven methods applied to MA_4560 -- not TfuA, an uncharacterized FilR1-family regulator

Method	Prediction returned	Correct for what MA_4560 actually is?
BLAST/Pfam	PF08350 (FilR1) -- transcriptional regulator	Plausibly yes -- matches MA_4560's own Pfam call
Foldseek	963 hits, all transcriptional regulators	Plausibly yes
CLEAN	Score 0.000 -- no enzyme function	Plausibly correct abstention (not an enzyme)
ESM-2	Rank 1903/4540 -- not discriminative	Uninformative either way
ProTrek	"regulatory" #1 (-0.035); "thioamidation" #3 (-0.054)	Plausibly yes on rank #1
DeepFRI	DNA binding, confidence 0.669	Plausible -- has a winged-helix DNA-binding domain
Custom classifier	Score 0.000 -- rank 3718/4540	Plausibly correct (not biosynthetic)

Ground truth: MA_4560 = UniProt Q8THF7, Pfam PF31122+PF08350 (FilR1 regulator family). The real TfuA is a different locus, MA_0164 (UniProt Q8TUA9, Pfam PF07812), characterized by Nayak et al. 2017 and structurally by ref. 28 (PDB 6XP8).

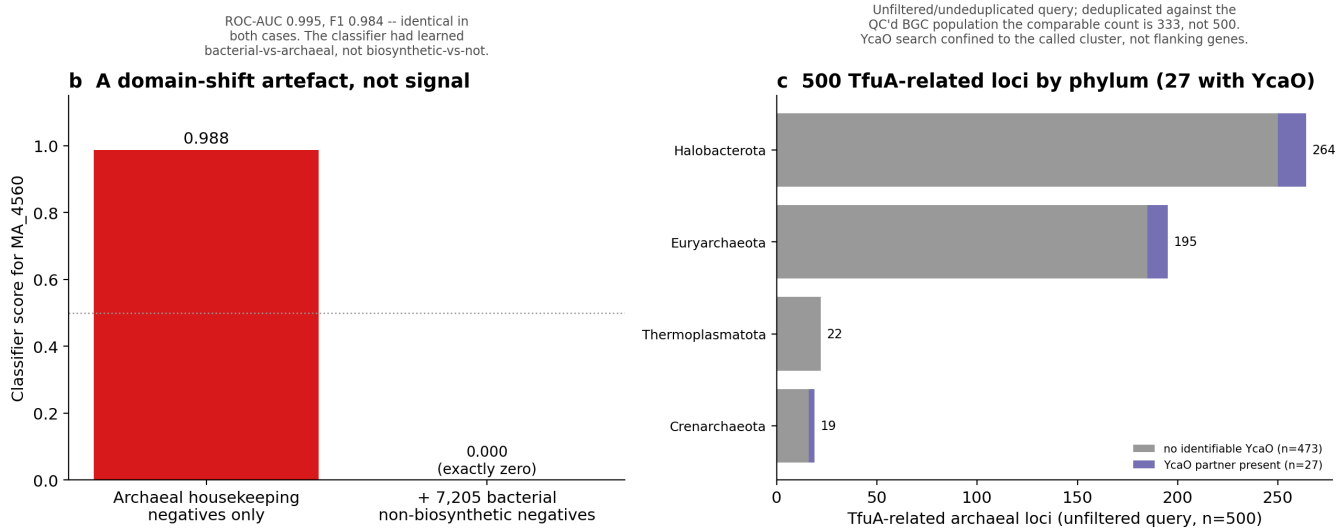

Figure 3. A training-population artefact in biosynthetic-protein classification, and the distribution of TfuA-associated archaeal loci. MA_4560 (UniProt Q8THF7) is not TfuA. It is an unrelated, uncharacterized FilR1-family (winged-helix transcriptional regulator) protein, per UniProt's own Pfam domain calls (PF31122, PF08350) and the raw *M. acetivorans* GenBank annotation. The real TfuA is MA_0164 (UniProt Q8TUA9, Pfam PF07812), first characterized genetically by Nayak et al. [7] and structurally by ref. [28] (PDB 6XP8). (a) Failure matrix for seven computational methods applied to MA_4560 (not TfuA). Each cell reports the prediction the method actually returned: Pfam PF08350 (FilR1_middle, a transcriptional regulator domain); 963 Foldseek AFDB50 hits, all transcriptional regulators; CLEAN score 0.000; ESM-2 rank 1903 of 4,540 proteins; ProTrek ranking "winged helix-turn-helix transcription factor" first (-0.035) and "thioamide bond formation" third (-0.054); DeepFRI predicting DNA binding at confidence 0.669; and our own biosynthetic classifier scoring 0.000, rank 3718 of 4,540. Read against MA_4560's actual identity rather than against TfuA, most of these are plausibly *correct* identifications of a regulator, not failures. (b) The originally reported "dissociation" (fold correct, function wrong) does not hold once the protein is correctly identified: BLAST/Pfam, Foldseek and DeepFRI's regulatory/DNA-binding calls are consistent with MA_4560's real domain composition. (c) A domain-shift artefact, which remains a valid, protein-independent finding: a classifier trained with archaeal housekeeping proteins as the only negatives scores MA_4560 at 0.988; adding 7,205 bacterial non-biosynthetic proteins to the negative class drops it to exactly 0.000, at an unchanged ROC-AUC of 0.995 and F1 of 0.984, showing the first classifier had learned to separate bacterial from archaeal proteins, not biosynthetic from non-biosynthetic. (d) TfuA-related archaeal loci by phylum from a database query returning 500 loci, each a distinct BGC within the QC'd population of 3,974 (498 of them also genus-classified) — an independent analysis (BiG-SLiCE/antiSMASH classification of other archaeal genomes' loci) unaffected by the MA_4560 identity error, but with its own limitations (Results, Methods): split on whether a YcaO partner is identifiable within the same called cluster (27 loci) or not (473), a search that never extends past the cluster boundary into flanking genes or the wider contig.

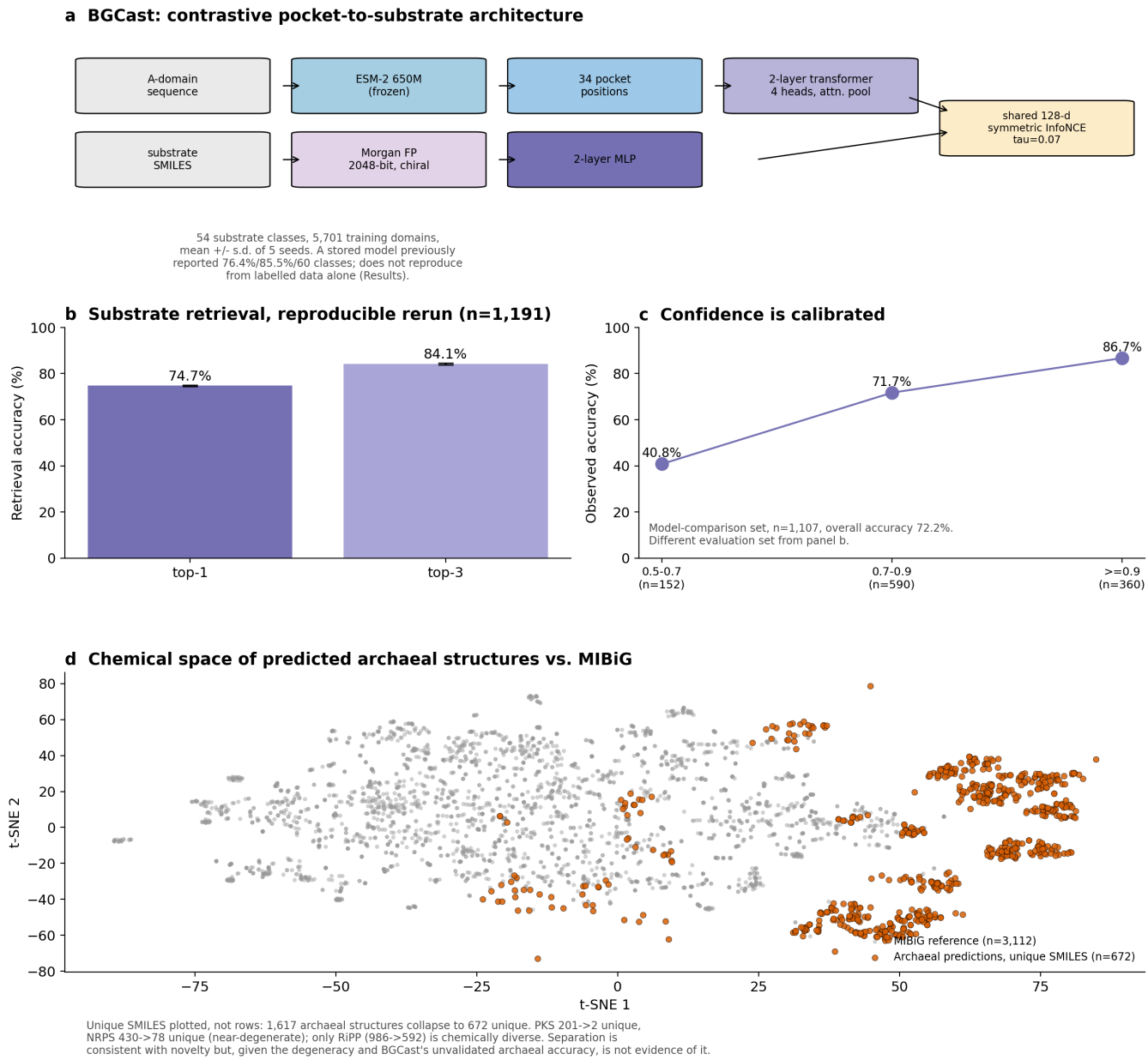

Figure 4. BGCast: substrate prediction for the detectable fraction. (a) Architecture. A-domain sequences are encoded by frozen ESM-2 (650M) and read out at 34 binding-pocket positions; a 2-layer, 4-head transformer with attention pooling projects to 128 dimensions. Substrate SMILES are converted to chirality-aware Morgan fingerprints (2048-bit, radius 2) and encoded by a 2-layer MLP into the same space, aligned by symmetric InfoNCE at fixed temperature tau=0.07. Prediction is by cosine similarity to encoded fingerprints, which allows substrates absent from training to be ranked in principle; we have not run a held-out-substrate-class test to confirm ranking accuracy for a truly unseen class. (b) Retrieval accuracy on the held-out test split (n=1,191, mean of 5 seeded reruns): 74.7% ± 0.3 top-1, 84.1% ± 0.4 top-3, over 54 substrate classes trained on 8,744 labelled bacterial A-domains (3,653 PARASECT, 2,798 MIBiG, 2,293 MASPR); a stored model previously reported 76.4%/85.5%/60 classes but does not reproduce from labelled data alone (Results). Cross-source deduplication was by exact full-sequence match only, not sequence-identity clustering, and leakage across the train/test split is measured rather than merely possible: 881 of the 1,191 test domains (74.0%) share a byte-identical 34-residue pocket string — the model's entire input — with a training domain, and 290 (24.3%) share a byte-identical full sequence. These accuracies are therefore upper bounds (Results). (c) Confidence calibration on the 1,107-domain model-comparison set, whose overall accuracy is 72.2%: 40.8% accurate at 0.5-0.7 confidence (n=152), 71.7% at 0.7-0.9 (n=590), 86.7% above 0.9 (n=360), and 40.0% for the 5 domains below 0.5 confidence (n=5; these three bins alone sum to 1,102, not 1,107). This is a different evaluation set from panel b. (d) Chemical space by t-SNE of Morgan fingerprints (radius 2, 2048 bits, chirality-aware), showing the 3,112-compound MIBiG reference against predicted archaeal structures. Unique SMILES are plotted, not rows: the 1,617 assembled archaeal structures collapse to 672 unique

structures, and the polyketide and non-ribosomal peptide fractions are near-degenerate (201 structures sharing 2 SMILES, and 430 sharing 78, respectively). The apparent separation is consistent with novelty but, given both the degeneracy and BGCast's unvalidated archaeal accuracy, is not evidence of it.

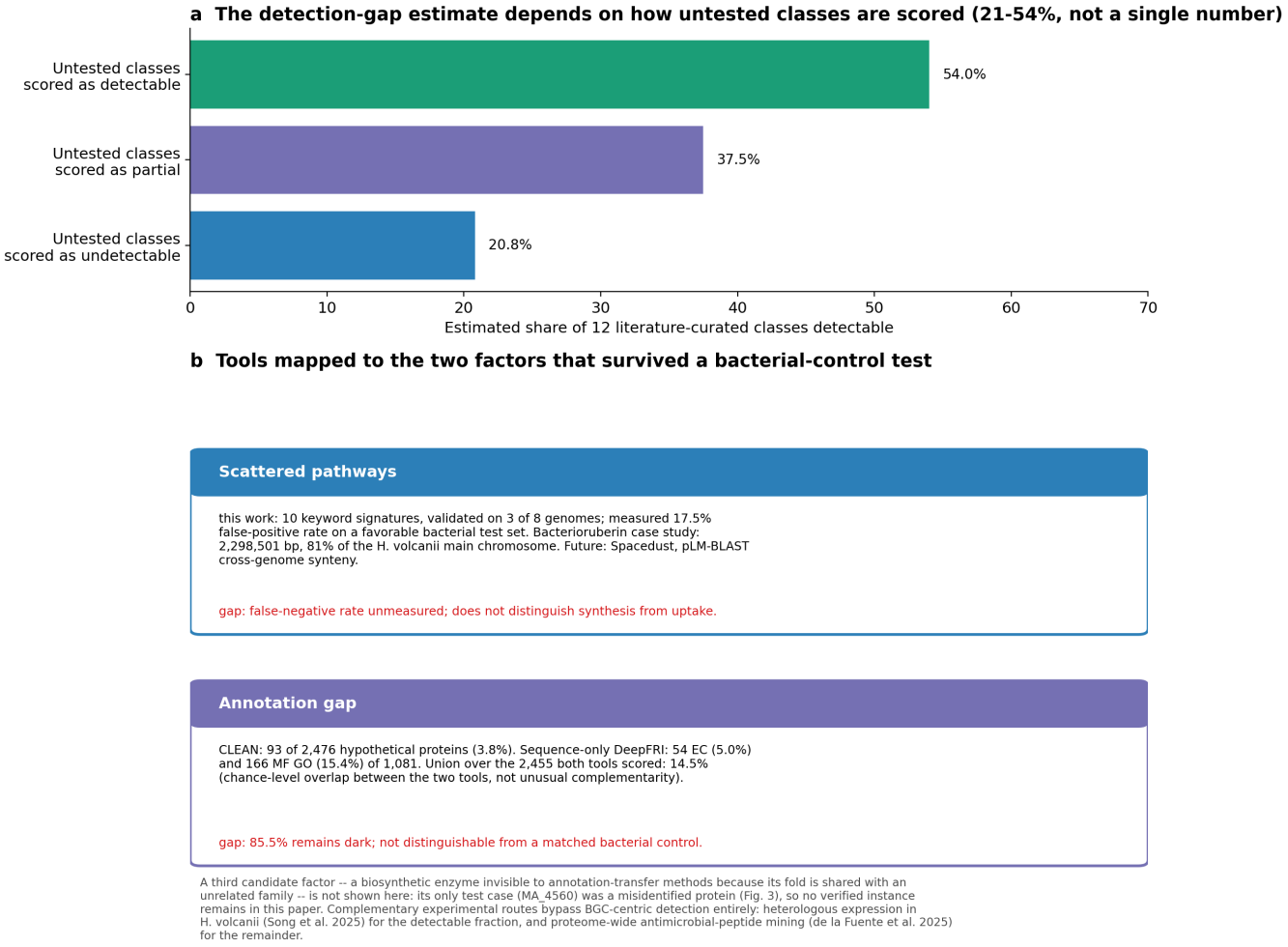

Figure 5. Multi-factor roadmap for archaeal biosynthetic exploration. (a) The detection-gap range the roadmap addresses, with its arithmetic on the panel: of 12 literature-curated archaeal compound classes, 1 is fully detectable, 3 partially, 2 have no rule in any tool, and 4 have a rule of untested archaeal fit, giving 20.8-54% depending on how the last group is credited — not a single number. (b) Tools mapped to each factor, with what each recovers and what it leaves behind. Scattered pathways: 10 keyword signatures validated on 3 of the 8 genomes used elsewhere in this paper (not all 8), with the bacterioruberin case study (2,298,501 bp, 81% of the *H. volcanii* main chromosome), and a measured 17.5% false-positive rate on a favorable bacterial test set; future work in cross-genome synteny. Annotation gap: CLEAN recovering 93 high-confidence enzymes from 2,476 hypothetical proteins (3.8%) and sequence-only DeepFRI recovering 54 EC assignments (5.0%) and 166 molecular function GO terms (15.4%) from 1,081; union across the proteins both scored, 14.5% (chance-level overlap, not unusual complementarity), and not distinguishable from a matched bacterial control. Fold-convergent annotation-transfer failure is not shown, because this paper reports no verified instance: the seven-method test was applied to MA_4560, an unrelated FilR1-family regulator rather than TfuA, and a rerun against the real locus (MA_0164) requires infrastructure unavailable here. Independently of that error, a YcaO partner is identifiable for only 27 of 500 TfuA-labeled loci from other archaeal genomes, each a distinct BGC within the QC'd population. The red line under each factor is the gap that factor's tools do not close. Complementary experimental routes bypass BGC-centric detection entirely: heterologous expression in *H. volcanii* [3] for the detectable fraction, and proteome-wide antimicrobial-peptide mining [26] for the remainder.