Tier 1: Exploratory | Natural Products & Microbial Genomics | Signal Rating: 2/5 Beakers | Nov 1, 2026 | 16 min read

Calibrating a Biosynthetic Gene Cluster Detector: Released Profiles, Added Marker Families and a Reserve-Genome Test Against GECCO

A staged improvement scored once on 9 genomes withheld from tuning, with paired statistics, negative controls and documented failure modes

We calibrated a lightweight biosynthetic gene cluster detector and scored it once on 5 genomes never examined during tuning. Using released Pfam profiles, 14 added marker families and a development-set grid search, locus recall rose from 9 to 19 of 27 curated clusters, versus 24 for GECCO, while the detector called 1 cluster in four BGC-free genomes against 18 for GECCO.

1. Why a Second, Stricter Calibration Was Needed

Our earlier benchmark of a rule-based biosynthetic gene cluster (BGC) detector against GECCO, a conditional-random-field detector (Carroll et al., 2021), reported in Lab Discovery #59 , found that the rule-based tool called far fewer clusters but a larger share of them overlapped experimentally curated loci in the MIBiG database (Zdouc et al., 2025). A follow-up on ten new genomes weakened that conclusion: the apparent precision advantage shrank when the rules were applied to genomes they had not been tuned on, and an automated choice between detector settings came down to an arbitrary tie-break. Two lessons followed. Any change to the detector has to be judged on genomes that were not used to make it, and the improvements need to be separable so each can be credited or discarded on its own.

This article reports that calibration. We changed three things in sequence (the source of the protein-family profiles, the set of marker families and core-gene rules, and the cluster-calling parameters), froze the result on a development set, and only then scored it on 9 reserve genomes (Figure 1). Throughout, the quantity we can measure is overlap with curated loci, not true precision: MIBiG is far from exhaustive, so a cluster that overlaps no curated locus may still be a real, uncurated BGC.

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

2. Step One: Released Profiles Close Most of the Domain-Level Gap

The detector recognises biosynthetic enzymes by scanning every predicted protein against profile hidden Markov models (HMMs) of protein families from Pfam (Mistry et al., 2021), searched with the forward algorithm described by Eddy (2011). In the first version those profiles were rebuilt from the small Pfam seed alignments, some containing only 5-8 sequences, using a simplified prior. We exported the released Pfam profiles for the same families and loaded them into the identical search engine. On the development genomes the fraction of GECCO’s domain calls that the detector also made rose from 0.661 to 0.891, with precision against GECCO essentially unchanged (0.971 versus 0.962); on the second-round bacterial genomes it

rose from 0.728 to 0.895 (Table 1). Locus recall changed far less (13 to 14 of 28 on the development genomes), so most of the domain-level gap did not translate into missed clusters, which are called from several genes at once.

The profile source was the only variable in this step, but it also changed the Pfam release (the released profiles shipped with GECCO come from an earlier Pfam version than the seed alignments we downloaded), so some of the effect could reflect that difference. One family, the SPASM domain, is absent from the released set and keeps its seed-built profile. A related variable, the domain cut-off rule (Pfam curated gathering threshold, a fraction of it, or GECCO’s p-value filter), changed domain agreement slightly but locus recall by at most one locus in either direction.

| Profile Source | | Development Loci Overlapped | | Development Loci Total | | Development Domain Precision vs GECCO | | Development Domain Recall vs GECCO | | Round-2 Held-Out Loci Overlapped | | Round-2 Held-Out Loci Total | | Round-2 Held-Out Domain Precision | | Round-2 Held-Out Domain Recall |

| Seed-built profiles | | 13 | | 28 | | 0.971 | | 0.661 | | 15 | | 36 | | 0.985 | | 0.728 |

| Released Pfam profiles | | 14 | | 28 | | 0.962 | | 0.891 | | 16 | | 36 | | 0.968 | | 0.895 |

Generated by: CRV Science Labs Native Bioinformatics Suite

3. Step Two: Fourteen Added Marker Families and New Core-Gene Rules

Several curated loci were missed because they are built around families the detector did not model. To choose additions without guessing, we asked which Pfam domains that GECCO detects are over-represented in genes lying inside the 73 curated loci of the development genomes, relative to all other genes in the same genomes, and kept families that were both strongly enriched and biologically sensible for specialised metabolism (Table 2). The shortlist includes the terminal reductase domain of nonribosomal peptide synthetases, the MbtH-like partner protein of those enzymes, a bacterial terpene synthase family, the starter-unit transacylase of fungal polyketide synthases, pathway-regulator and siderophore-enzyme domains, and the melanin pathway cofactor. Because the enrichment was computed on the same genomes used for development, it is in-sample evidence; the reserve set tests whether it generalises.

Three new core-gene definitions use these families: an adenylation domain fused to a terminal reductase counts as an NRPS gene, a stand-alone adenylation gene beside an MbtH-like gene counts as an NRPS-like cluster seed, and the bacterial terpene synthase and the fungal starter-unit transacylase fused to a ketosynthase each define a core gene. The safeguards from the first version are unchanged: a lone ketosynthase or a lone adenylation gene still cannot seed a cluster. The MbtH family illustrates what these small marker proteins look like in the seed alignment (Figure 2).

| Pfam | | Family | | Curated Loci Containing It (of 73) | | Genes Inside Loci | | Genes Outside Loci | | Enrichment Inside vs Outside | | Role in Revision 3 |

| PF19086 | | Terpene_syn_C_2 | | 4 | | 4 | | 7 | | 26.8 | | Core gene: bacterial

terpene synthase |

| PF07993 | | NAD_binding_4 | | 13 | | 15 | | 65 | | 11.5 | | Core gene when fused to an adenylation domain (NRPS terminal reductase) |

| PF03621 | | MbtH | | 4 | | 4 | | 14 | | 13.9 | | Partner of a stand-alone adenylation gene (NRPS-like cluster) |

| PF03704 | | BTAD | | 9 | | 11 | | 32 | | 17.0 | | Accessory (pathway-specific regulator domain) |

| PF16073 | | SAT | | 4 | | 5 | | 6 | | 38.7 | | Core gene when fused to a ketosynthase (fungal PKS starter unit) |

| PF13434 | | Lys_Orn_oxgnase | | 4 | | 4 | | 29 | | 6.8 | | Accessory (siderophore pathway) |

| PF00425 | | Chorismate_bind | | 6 | | 6 | | 31 | | 9.6 | | Accessory (siderophore pathway) |

| PF00264 | | Tyrosinase | | 5 | | 5 | | 3 | | 71.9 | | Accessory (tyrosinase) |

| PF06236 | | MelC1 | | 4 | | 4 | | 0 | | 402.5 | | Core gene (melanin pathway cofactor) |

| PF08541 | | ACP_syn_III_C | | 7 | | 8 | | 29 | | 13.6 | | Accessory (ketosynthase III) |

| PF08545 | | ACP_syn_III | | 4 | | 5 | | 24 | | 10.3 | | Accessory (ketosynthase III) |

| PF16363 | | GDP_Man_Dehyd | | 9 | | 12 | | 107 | | 5.6 | | Accessory (deoxysugar biosynthesis) |

| PF00108 | | Thiolase_N | | 11 | | 20 | | 104 | | 9.6 | | Accessory (thiolase) |

| PF00891 | | Methyltransf_2 | | 4 | | 5 | | 32 | | 7.7 | | Accessory (O-methyltransferase) |

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

Generated by: CRV Science Labs Native Bioinformatics Suite

4. Step Three: Parameter Search on the Development Set Only

The cluster caller has five tunable settings: how many genes on each side contribute support to a gene (window), the support score a core gene needs to seed a cluster, the largest gap between genes during extension, and the padding added to each cluster end. We scored 288 combinations of rule revision and settings on the development set. The selection rule was fixed beforehand: allow at most 1 cluster in the BGC-free development genomes, then maximise a balanced score (the harmonic mean of locus recall and the fraction of predicted clusters overlapping a curated locus), then the mean overlap of cluster boundaries with curated boundaries, then prefer fewer clusters. The winner was revision 3 with a window of 6 genes, a seed score of 1.2, a gap of 5,000 bp and 1,000 bp of padding. The configuration was written to disk before any reserve genome was scanned.

Two caveats apply to this choice. The winning window and seed score sit at the edge of the grid (the widest window and the lowest seed score tried), so a better setting may lie outside it and the optimum is not established. Many neighbouring settings scored identically (Table 3), which suggests the objective is flat in this region and the choice among them carries little weight.

| Rank | | Rule Revision | | Window (genes) | | Seed Score | | Gap (bp) | | Padding (bp) | | Loci Overlapped | | Loci Total | | Clusters | | Share Overlapping MIBiG | | Locus Recall | | Balanced Score | | Mean Boundary IoU | | Negative-Control Clusters |

| 1 | | 3 | | 6 | | 1.2 | | 5000 | | 1000 | | 46 | | 64 | | 117 | | 0.376 | | 0.719 | | 0.494 | | 0.467 | | 1 |

| 2 | | 3 | | 6 | | 1.2 | | 5000 | | 2000 | | 46 | | 64 | | 117 | | 0.376 | | 0.719 | | 0.494 | | 0.451 | | 1 |

| 3 | | 3 | | 6 | | 1.2 | | 5000 | | 4000 | | 46 | | 64 | | 117 | | 0.376 | | 0.719 | | 0.494 | | 0.421 | | 1 |

| 4 | | 3 | | 6 | | 1.2 | | 3000 | | 1000 | | 46 | | 64 | | 119 | | 0.370 | | 0.719 | | 0.488 | | 0.467 | | 1 |

| 5 | | 3 | | 6 | | 1.2 | | 3000 | | 2000 | | 46 | | 64 | | 119 | | 0.370 | | 0.719 | | 0.488 | | 0.451 | | 1 |

| 6 | | 3 | | 6 | | 1.2 | | 3000 | | 4000 | | 46 | | 64 | | 119 | | 0.370 | | 0.719 | | 0.488 | | 0.421 | | 1 |

| 7 | | 3 | | 6 | | 1.2 | | 8000 | | 1000 | | 46 | | 64 | | 114 | | 0.368 | | 0.719 | | 0.487 | | 0.456 | | 1 |

| 8 | | 3 | | 6 | | 1.2 | | 8000 | | 2000 | | 46 | | 64 | | 114 | | 0.368 | | 0.719 | | 0.487 | | 0.440 | | 1 |

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

5. Reserve-Set Results

The reserve set contains 5 bacterial genomes with 27 curated loci (Xenorhabdus nematophila , Streptomyces scabiei , Streptomyces avermitilis , Saccharopolyspora erythraea and Salinispora arenicola) and four BGC-free controls (Buchnera aphidicola , Chlamydia trachomatis , Treponema pallidum and the fission yeast Schizosaccharomyces pombe). On these genomes the published detector overlapped 9 of 27 curated loci (19-52% 95% interval). Released profiles raised that to 12, the revision-3 rules to 15, and the frozen parameters to 19 of 27 (70.4%, interval 52-84%), against 24 of 27 for GECCO (72-96%) (Table 4, Figure 3).

The gain came with more clusters: the frozen configuration called 105 clusters against 66 for the published detector, and GECCO called 178. The fraction of predicted clusters that overlap a curated locus was 18.1% for the frozen configuration and 12.9% for GECCO, a smaller difference than the one seen in the first benchmark, and one that cannot be interpreted as a precision difference because most clusters in every list are uncurated. Mean overlap between predicted

and curated boundaries (computed over the loci each tool found) was 0.592 for the frozen configuration, 0.746 for the published detector and 0.550 for GECCO; the published detector’s higher value reflects that it only found the easiest loci, and the frozen configuration’s boundaries are looser than before.

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

| Configuration | | Loci Overlapped | | Loci Total | | Loci At Least Half Covered | | Clusters Called | | Clusters Overlapping MIBiG | | Share Overlapping MIBiG | | Locus Recall | | Balanced Score | | Mean Boundary IoU | | Negative-Control Clusters |

| Published detector (seed profiles, revision 2) | | 9 | | 27 | | 9 | | 66 | | 9 | | 0.136 | | 0.333 | | 0.194 | | 0.746 | | 1 |

| Released Pfam profiles, revision 2 | | 12 | | 27 | | 12 | | 73 | | 12 | | 0.164 | | 0.444 | | 0.240 | | 0.675 | | 1 |

| Revision 3 rules, default parameters | | 15 | | 27 | | 15 | | 84 | | 15 | | 0.179 | | 0.556 | | 0.270 | | 0.709 | | 1 |

| Revision 3, frozen parameters | | 19 | | 27 | | 19 | | 105 | | 19 | | 0.181 | | 0.704 | | 0.288 | | 0.592 | | 1 |

| GECCO v0.11.0 (default cut-off 0.8) | | 24 | | 27 | | 24 | | 178 | | 23 | | 0.129 | | 0.889 | | 0.226 | | 0.550 | | 18 |

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

For comparison, the same configurations on the development genomes, where the settings were chosen and the numbers are therefore optimistic, are in Table 5; the improvement there was of similar size (30 to 46 of 64 loci), and GECCO overlapped 52.

| Configuration | | Loci Overlapped | | Loci Total | | Loci At Least Half Covered | | Clusters Called | | Clusters Overlapping MIBiG | | Share Overlapping MIBiG | | Locus Recall | | Balanced Score | | Mean Boundary IoU | | Negative-Control Clusters |

| Published detector (revision 2) | | 30 | | 64 | | 27 | | 83 | | 27 | | 0.325 | | 0.469 | | 0.384 | | 0.536 | | 0 |

| Revision 3, default parameters | | 36 | | 64 | | 33 | | 92 | | 33 | | 0.359 | | 0.563 | | 0.438 | | 0.517 | | 0 |

| Revision 3, frozen parameters | | 46 | | 64 | | 44 | | 117 | | 44 | | 0.376 | | 0.719 | | 0.494 | | 0.467 | | 1 |

| GECCO v0.11.0 (default cut-off 0.8) | | 52 | | 64 | | 52 | | 187 | | 48 | | 0.257 | | 0.813 | | 0.390 | | 0.511 | | 14 |

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

Generated by: CRV Science Labs Native Bioinformatics Suite

| Genome | | Organism | | Role | | Size (Mb) | | Curated Loci | | Published Loci | | Published Clusters | | Released-Profile Loci | | Released-Profile Clusters | | Revision 3 Loci | | Revision 3 Clusters | | Frozen Loci | | Frozen Clusters | | GECCO Loci | | GECCO Clusters |

| FN667742.1 | | Xenorhabdus nematophila ATCC 19061 | | bgc-rich | | 4.43 | | 6 | | 3 | | 10 | | 3 | | 11 | | 3 | | 13 | | 4 | | 13 | | 5 | | 13 |

| FN554889.1 | | Streptomyces scabiei 87.22 | | bgc-rich | | 10.15 | | 6 | | 3 | | 15 | | 4 | | 15 | | 5 | | 18 | | 5 | | 22 | | 5 | | 47 |

| BA000030.4 | | Streptomyces avermitilis MA-4680 | | bgc-rich | | 9.03 | | 5 | | 1 | | 18 | | 1 | | 18 | | 1 | | 20 | | 2 | | 26 | | 4 | | 49 |

| AM420293.1 | | Saccharopolyspora erythraea NRRL 2338 | | bgc-rich | | 8.21 | | 5 | | 1 | | 10 | | 1 | | 13 | | 2 | | 15 | | 4 | | 22 | | 5 | | 42 |

| CP000850.1 | | Salinispora arenicola CNS-205 | | bgc-rich | | 5.79 | | 5 | | 1 | | 13 | | 3 | | 16 | | 4 | | 18 | | 4 | | 22 | | 5 | | 27 |

| NC_002528.1 | | Buchnera aphidicola APS | | negative-control | | 0.64 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 1 |

| NC_000117.1 | | Chlamydia trachomatis D/UW-3 | | negative-control | | 1.04 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 1 |

| NC_000919.1 | | Treponema pallidum Nichols | | negative-control | | 1.14 | | 0 | | 0 | | 1 | | 0 | | 1 | | 0 | | 1 | | 0 | | 1 | | 0 | | 1 | | 0 | | 1 |

| NC_003421.2 | | Schizosaccharomyces pombe chr III | | negative-control | | 2.45 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 15 |

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

6. Paired Comparison With GECCO and Which Loci Remain Missed

Because both tools were scored on the same 27 loci, they can be compared pairwise. 19 loci were overlapped by both, 0 by the frozen configuration alone, 5 by GECCO alone, and 3 by neither. The exact McNemar test on the discordant loci gives p = 0.063, so with this few loci a real difference in sensitivity can be neither established nor excluded at the conventional 0.05 level; the interval for GECCO (72-96%) overlaps that of the frozen configuration. The 5 loci found only by GECCO are of the classes 2 other, 1 terpene, 2 ribosomal, which are the classes the detector models least well: ribosomal peptides, terpenes and loci annotated ‘other’ (Table 7). Within nonribosomal peptide and polyketide classes, which the detector was designed around, the two tools agreed on every reserve locus (Table 8).

| MIBiG Class | | Loci | | Frozen Configuration Overlapped | | GECCO Overlapped |

| NRPS | | 8 | | 8 | | 8 |

| other | | 4 | | 0 | | 2 |

| ribosomal | | 4 | | 1 | | 3 |

| PKS | | 4 | | 4 | | 4 |

| terpene | | 4 | | 3 | | 4 |

| NRPS/PKS | | 1 | | 1 | | 1 |

| PKS/saccharide | | 1 | | 1 | | 1 |

| PKS/NRPS | | 1 | | 1 | | 1 |

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

| Genome | | MIBiG Locus | | Class | | Length (bp) | | Frozen Coverage (fraction) | | Frozen Boundary IoU | | GECCO Coverage (fraction) | | GECCO Boundary IoU |

| FN667742.1 | | BGC0001757 | | other | | 3222 | | 0.00 | | 0.00 | | 1.00 | | 0.57 |

| FN667742.1 | | BGC0001054 | | NRPS/PKS | | 41309 | | 1.00 | | 0.75 | | 1.00 | | 0.87 |

| FN667742.1 | | BGC0001132 | | NRPS | | 12417 | | 1.00 | | 0.15 | | 1.00 | | 0.14 |

| FN667742.1 | | BGC0000416 | | NRPS | | 15337 | | 1.00 | | 0.43 | | 1.00 | | 0.53 |

| FN667742.1 | | BGC0000465 | | NRPS | | 10043 | | 1.00 | | 0.46 | | 1.00 | | 0.57 |

| FN667742.1 | | BGC0002927 | | ribosomal | | 5676 | | 0.00 | | 0.00 | | 0.00 | | 0.00 |

| FN554889.1 | | BGC0002367 | | NRPS | | 32902 | | 1.00 | | 0.71 | | 1.00 | | 0.73 |

| FN554889.1 | | BGC0002568 | | other | | 11038 | | 0.00 | | 0.00 | | 0.00 | | 0.00 |

| FN554889.1 | | BGC0002089 | | NRPS | | 18465 | | 1.00 | | 0.75 | | 1.00 | | 0.27 |

| FN554889.1 | | BGC0001157 | | ribosomal | | 17167 | | 0.97 | | 0.86 | | 1.00 | | 0.98 |

| FN554889.1 | | BGC0002808 | | PKS | | 31376 | | 1.00 | | 0.83 | | 0.94 | | 0.83 |

| FN554889.1 | | BGC0000423 | | NRPS | | 30020 | | 1.00 | | 0.85 | | 1.00 | | 0.55 |

| BA000030.4 | | BGC0000683 | | terpene | | 1008 | | 0.00 | | 0.00 | | 1.00 | | 0.06 |

| BA000030.4 | | BGC0000059 | | PKS | | 80370 | | 1.00 | | 0.81 | | 1.00 | | 0.87 |

| BA000030.4 | | BGC0000678 | | terpene | | 13437 | | 0.90 | | 0.80 | | 1.00 | | 0.75 |

| BA000030.4 | | BGC0001587 | | other | | 8255 | | 0.00 | | 0.00 | | 0.00 | | 0.00 |

| BA000030.4 | | BGC0001552 | | ribosomal | | 4562 | | 0.00 | | 0.00 | | 1.00 | | 0.36 |

| AM420293.1 | | BGC0000055 | | PKS/saccharide | | 54612 | | 1.00 | | 0.96 | | 1.00 | | 0.98 |

| AM420293.1 | | BGC0000902 | | PKS | | 3903 | | 1.00 | | 0.25 | | 1.00 | | 0.25 |

| AM420293.1 | | BGC0000349 | | NRPS | | 24145 | | 1.00 | | 0.32 | | 1.00 | | 0.77 |

| AM420293.1 | | BGC0000659 | | terpene | | 1937 | | 1.00 | | 0.10 | | 1.00 | | 0.09 |

| AM420293.1 | | BGC0000513 | | ribosomal | | 6203 | | 0.00 | | 0.00 | | 1.00 | | 0.44 |

| CP000850.1 | | BGC0000137 | | PKS | | 91573 | | 0.76 | | 0.72 | | 1.00 | | 0.74 |

| CP000850.1 | | BGC0002969 | | terpene | | 3828 | | 1.00 | | 0.13 | | 1.00 | | 0.03 |

| CP000850.1 | | BGC0000827 | | other | | 24300 | | 0.00 | | 0.00 | | 1.00 | | 0.60 |

| CP000850.1 | | BGC0001007 | | PKS/NRPS | | 23687 | | 0.78 | | 0.38 | | 1.00 | | 0.38 |

| CP000850.1 | | BGC0000333 | | NRPS | | 47477 | | 0.98 | | 0.98 | | 1.00 | | 0.81 |

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner

7. Negative Controls and Known Failure Modes

In the 4 reserve controls the frozen configuration called 1 cluster (in T. pallidum) against 18 for GECCO, which included 15 in the fission yeast chromosome. In the 4 development controls the counts were 1 and 14. The detector’s one reserve call is a 334-residue adenylation gene with a 78-residue carrier-protein gene about 1.9 kb away, the pattern the NRPS-like rule is designed to catch but which can also arise in primary metabolism. No curated cluster is known in these genomes, but absence of a curated cluster is not proof that none exists.

The lower seed score also admits single-gene calls that are probably housekeeping. Applied to E. coli K-12 with the frozen configuration, the detector reports the enterobactin locus plus two RiPP-class clusters: one built on a conserved YcaO-family gene beside a radical-SAM gene, the other on a fused radical-SAM/SPASM gene (the detector does not search for precursor peptides, so neither call has that supporting evidence). We treat both as likely false positives. They were found after the configuration was frozen, so we have not changed the rules in response; doing so would require a new reserve set. The same loosening explains the larger cluster counts and the looser boundaries noted above.

Discussion & Key Takeaways

1. The improvement held up on genomes not used for tuning. On the reserve set, locus recall rose from 33.3% to 70.4% (9 to 19 of 27). Each change contributed: released profiles, the added families and rules, and the parameter search added 3, 3 and 4 loci respectively. The detector did not overlap any locus that GECCO missed, so on this panel it behaves like a more conservative subset of GECCO’s calls.

2. A gap to GECCO remains and its size is uncertain. GECCO overlapped 24 of 27 reserve loci against 19 for the frozen detector (p = 0.063, paired). The gap is concentrated in ribosomal, terpene and uncategorised loci. Against that, the detector called almost nothing in BGC-free genomes where GECCO called 32 clusters across the eight controls we used, which makes the two tools complementary rather

than interchangeable.

3. Signal Rating Assessment (2 Technical Validation). The calibration covers 24 genomes (119 Mb), but the out-of-sample test is 5 bacterial genomes with 27 curated loci, mostly actinomycetes, with no curated eukaryotic reserve genome. The marker families were shortlisted on the development loci, MIBiG labels are mostly flagged questionable , the frozen settings sit at a grid edge, and two likely false positives appeared after freezing. This is a reproducible technical validation of a tuning procedure, not evidence that the detector generalises across bacterial diversity, so we rate it as a technical validation.

4. What a next round should do. Tightening single-gene RiPP seeds, adding precursor-peptide evidence, and widening the parameter grid are testable fixes, but each needs a fresh reserve set because the current one has now informed our understanding of the failures. Overlap with curated loci should be reported alongside, not instead of, the detector’s calls in uncurated regions, which remain candidate hypotheses for laboratory follow-up rather than evidence of a compound.

References & Primary Sources

* [1] Carroll LM, Larralde M, Fleck JS, Ponnudurai R, Milanese A, Cappio E, Zeller G (2021). Accurate de novo identification of biosynthetic gene clusters with GECCO. bioRxiv (preprint). https://doi.org/10.1101/2021.05.03.442509

* [2] Zdouc MM, Blin K, Louwen NLL, et al. (2025). MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Research 53(D1):D678-D690. https://doi.org/10.1093/nar/gkae1115

* [3] Mistry J, Chuguransky S, Williams L, et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Research 49(D1):D412-D419. https://doi.org/10.1093/nar/gkaa913

* [4] Eddy SR (2011). Accelerated profile HMM searches. PLoS Computational Biology 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195

* [5] Blin K, Shaw S, Augustijn HE, et al. (2023). antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Research 51(W1):W46-W50. https://doi.org/10.1093/nar/gkad344

* [6] Fischbach MA, Walsh CT (2006). Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chemical Reviews 106(8):3468-3496. https://doi.org/10.1021/cr0503097

* [7] White BP (2026). Biosynthetic Gene Cluster Detection: Benchmarking a Domain-Architecture Rule Engine Against GECCO on Five Genomes with Curated MIBiG Loci. CRV Science Labs, Lab Discovery #59 (the earlier benchmark this article calibrates). https://crvscience.com/labs/discoveries/bgc-rules-vs-gecco-mibig

* [8] Henikoff S, Henikoff JG (1994). Position-based sequence weights. Journal of Molecular Biology 243(4):574-578. https://doi.org/10.1016/0022-2836(94)90032-9