

Tier 1: Exploratory | Natural Products & Microbial Genomics | Signal Rating: 2/5 Beakers | Nov 1, 2026 | 13 min read

Biosynthetic Gene Cluster Detection: Benchmarking a Domain-Architecture Rule Engine Against GECCO on Five Genomes with Curated MIBiG Loci

Comparing 34-family Pfam rule-based cluster calling with a conditional-random-field detector across 27.5 Mb of bacterial and fungal sequence and 33 experimentally curated clusters

We benchmarked a lightweight domain-architecture rule engine for biosynthetic gene cluster (BGC) detection against GECCO, a machine-learning detector, on five genomes (27.5 Mb) containing 33 experimentally curated MIBiG loci. The rule engine overlapped 15 of 33 loci with 29 predicted clusters, versus 27 of 33 with 108 for GECCO, trading sensitivity for fewer unconfirmed calls.

Update (1 November 2026): a follow-up calibration, Lab Discovery #60, scored this detector on five genomes withheld from tuning. On those genomes the detector described below overlapped 9 of 27 curated loci and called 66 clusters, of which 13.6% overlapped a curated locus, against 12.9% for GECCO (24 of 27 loci, 178 clusters). The larger overlap share reported in this article (14 of 29 clusters versus 24 of 108) therefore did not hold on new genomes and should be read as specific to this panel. The caveat that MIBiG is incomplete, so overlap is only a lower bound on precision, applies to both tools.

The follow-up also found that the sensitivity gap narrows after calibration (19 of 27 loci for the revised detector) and that the revised detector called 1 cluster in four BGC-free genomes against 18 for GECCO. This article is otherwise unchanged.

1. Why Cluster Detection Needs Domain Architecture, Not Just Domain Presence

Bacteria and fungi make antibiotics, siderophores and other specialised metabolites with assembly-line enzymes whose genes sit next to one another in biosynthetic gene clusters (BGCs). Two enzyme families dominate: polyketide synthases (PKS), built from ketosynthase (KS), acyltransferase (AT) and acyl-carrier-protein modules, and nonribosomal peptide synthetases (NRPS), built from adenylation (A), peptidyl-carrier (PCP) and condensation (C) modules (Fischbach & Walsh, 2006). A detector that simply counts these domains makes a predictable mistake: every bacterium also carries fatty-acid synthase (FAS) genes with KS, AT and carrier domains, and acyl-CoA ligases that contain an adenylation domain, yet neither produces a natural product.

Two design philosophies address this. GECCO scores every gene by its full Pfam domain profile and feeds the sequence of scores to a conditional random field (CRF), which learns cluster boundaries from known BGCs (Carroll et al., 2021). Our rule engine instead uses only 34 hand-chosen biosynthetic Pfam families and explicit architecture rules: a modular PKS gene must fuse a KS to an AT, reductase or dehydratase; an NRPS gene must carry a C domain or fuse A to a carrier protein.

Neighbouring genes then contribute a windowed support score $S_i = \sum_j w_j$ over the nine genes centred on gene i , and a cluster is seeded only when a core gene reaches $S_i \geq 1.8$ (Figure 1).

Generated by: Native Biosynthetic Gene Cluster (MIBiG / ARTS) & Amphipathic AMP Biophysics Miner * Native 4-Stage 6-Frame ORF Graph, Dicodon Hexamer Markov & 28-Bin Shine-Dalgarno RBS EM Engine * Native 4-Stage Plan-7 Profile HMM, Forward-Backward Log-Odds, Optimal Accuracy & Null2 Bias Engine

2. Benchmark Design: Five Genomes and Experimentally Curated Ground Truth

We selected five complete genomes that carry MIBiG 4.0 loci with genomic coordinates (Zdouc et al., 2025): *Streptomyces coelicolor* A3(2), *Bacillus velezensis* FZB42, *Pseudomonas aeruginosa* PAO1, *Escherichia coli* K-12 MG1655 (a near-negative control whose only curated locus is enterobactin), and chromosome II of the fungus *Aspergillus nidulans*. The fungal genome is deliberately outside the regime of both tools: GECCO was trained on bacterial clusters, and our gene caller uses a prokaryotic model that cannot reconstruct spliced multi-exon genes. After collapsing records with identical coordinates, the five genomes contribute 33 curated loci.

Two cautions apply to every number below. First, 32 of the 33 MIBiG records for these loci carry the curation flag questionable and several have unknown completeness, so coordinates are approximate. Second, MIBiG is far from exhaustive: these genomes contain additional real clusters that are simply not curated. The fraction of predicted clusters that overlap a MIBiG locus is therefore a lower bound on precision, not precision itself. GECCO was run at its documented default posterior cut-off (0.8) and at a permissive cut-off (0.3) to show how its call count scales.

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

Across all five genomes the rule engine overlapped 15 of 33 curated loci and GECCO at its default cut-off overlapped 27 of 33 (13 of 28 versus 22 of 28 in the four bacterial genomes). GECCO reached 30 of 33 at the permissive cut-off, at the cost of 155 clusters. The two tools differ most in how many clusters they report: 29 for the rule engine against 108 for GECCO (0.8), of which 14 and 24 respectively overlap a curated locus (Table 1). Because MIBiG is incomplete, the unconfirmed clusters in either list may include real but uncurated BGCs; the data cannot say which.

Genome	Organism	Kingdom	Genome Size (Mb)	Genes Called	MIBiG Loci	Native Hits (rules v2)	Native Hits (rules v1)	GECCO Hits (cut-off 0.8)	GECCO Hits (cut-off 0.3)	Native Clusters	GECCO Clusters (0.8)	GECCO Clusters (0.3)	Native Clusters Overlapping MIBiG	GECCO (0.8) Clusters Overlapping MIBiG	Base-Pair Jaccard (Native vs GECCO 0.8)
AL645882.2	<i>Streptomyces coelicolor</i> A3(2)	Prokaryota	8.67	7738	15	7	7	14	15	12	46	73	6	13	0.288
CP000560.1	<i>Bacillus velezensis</i> FZB42	Prokaryota	3.92	3741	6	4	4	5	5	9	10	12	4	4	0.825

| AE004091.2 || Pseudomonas aeruginosa PAO1 || Prokaryota || 6.26 || 5681 || 6
|| 1 || 1 || 2 || 4 || 3 || 9 || 18 || 1 || 2 || 0.370 |

| U00096.3 || Escherichia coli K-12 MG1655 || Prokaryota || 4.64 || 4319 || 1 |
| 1 || 0 || 1 || 1 || 1 || 2 || 7 || 1 || 1 || 0.424 |

| BN001302.1 || Aspergillus nidulans FGSC A4 chr II || Eukaryota || 4.01 || 5094
|| 5 || 2 || 2 || 5 || 5 || 4 || 41 || 45 || 2 || 4 || 0.068 |

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

3. Sensitivity Depends on Cluster Class

Breaking the 33 loci down by chemistry shows where the rule engine's gaps come from (Table 2). Its definitions cover PKS, NRPS, terpene and RiPP cores, so loci annotated as saccharide or other (a catch-all category for chemistry that does not fit the named classes) cannot be seeded by design. Within NRPS and hybrid NRPS/PKS clusters, the engine overlapped 11 of 13 loci against 13 of 13 for GECCO. PKS-only loci are the weakest class for the rule engine (2 of 7), partly because the fungal PKS loci sit on a genome where gene calling itself fails.

| MIBiG Cluster Class || Loci || Native Hits || GECCO Hits (0.8) || GECCO Hits (0.3) || Class Has Core Definition In Native Engine |

| Nonribosomal peptide synthetase (NRPS) || 11 || 9 || 11 || 11 || Yes |

| Polyketide synthase (PKS) || 7 || 2 || 6 || 7 || Yes |

| Hybrid NRPS/PKS || 2 || 2 || 2 || 2 || Yes |

| Terpene || 3 || 1 || 2 || 3 || Yes |

| Ribosomal peptide (RiPP) || 2 || 0 || 2 || 2 || Yes |

| Saccharide || 1 || 0 || 0 || 1 || No |

| Other / uncategorised || 7 || 1 || 4 || 4 || No |

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

4. A Worked Failure: Enterobactin and the Stand-Alone Adenylation Pattern

The E. coli K-12 enterobactin locus (MIBiG BGC0002476) is the clearest single illustration of how a rule gap, rather than a weak search, produced a miss.

Enterobactin is assembled from a stand-alone adenylation enzyme (EntE, 536 residues), a stand-alone carrier protein (EntB, 285 residues) and a multi-domain NRPS (EntF) that carries a condensation domain (Raymond et al., 2003). The native engine found the EntE and EntB domain signatures with high confidence, but under the first rule set a lone adenylation gene was not a core gene, so no cluster was called (Figure 3).

We then added an NRPS-like rule (a lone adenylation gene with a carrier-protein gene within three genes). This was introduced after examining this miss, so the enterobactin recovery is a diagnosis-and-repair result, not an independent test. The

rule added 3 clusters across the five genomes (26 to 29); only the enterobactin cluster overlaps a curated locus, while the other 2 (in *Bacillus velezensis* and *Aspergillus nidulans*) do not overlap any MIBiG record and so cannot be confirmed or rejected here. The total of curated loci overlapped therefore moved from 14 to 15 of 33. The same panel also shows a sensitivity gap: GECCO assigns a reductase (Pfam KR) domain to the third gene of the block at $E = 4 \times 10^{-12}$, which the native profile misses (Section 5).

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

5. Domain-Level Concordance: High Precision, Moderate Recall

To separate profile quality from cluster-calling logic, we compared the per-gene domain calls of the native engine with the calls GECCO makes for the same 34 Pfam families on the same genes (genes matched by stop-codon coordinate). Of 1013 native calls, 983 were also found by GECCO (0.970 of native calls confirmed). Of 1490 GECCO calls, 983 were recovered by the native engine (recall 0.660). The native engine is therefore rarely wrong but misses a third of GECCO's calls (Table 3).

Two effects probably contribute, and the data here cannot apportion them. First, GECCO keeps a domain hit when its p-value is below 10^{-9} (its default), a single statistical cut-off, whereas the native engine applies each family's curated Pfam gathering threshold; for broad families such as the short-chain-reductase-like KR domain (PF08659) the two criteria can differ widely, and that family shows the largest gap. Second, the native profiles are built from the small seed alignments (as few as 5-8 sequences for Pfam families PF00195, PF02797 and PF00698) with a simplified prior, whereas the released Pfam models use hmmbuild's mixture priors and entropy-based sequence weighting (Eddy, 2011; Mistry et al., 2021). The type III PKS families, which have the smallest seeds, are recovered least well (Table 3), and the one type III PKS locus in this benchmark (flaviolin, MIBiG BGC0002127) is missed.

Pfam Family	Name	Seed Sequences	Model Length (aa)	Curated GA Cut-off (bits)	Native Calls	GECCO Calls	Shared Calls	Native Recall vs GECCO
-------------	------	----------------	-------------------	---------------------------	--------------	-------------	--------------	------------------------

PF00109	ketoacyl-synt	75	253	20.8	57	61	55	0.90
---------	---------------	----	-----	------	----	----	----	------

PF02801	Ketoacyl-synt_C	75	119	20.8	52	53	52	0.98
---------	-----------------	----	-----	------	----	----	----	------

PF00698	Acyl_transf_1	8	320	32.7	18	28	18	0.64
---------	---------------	---	-----	------	----	----	----	------

PF08659	KR	42	180	22.6	42	181	41	0.23
---------	----	----	-----	------	----	-----	----	------

PF14765	PS-DH	59	168	23.7	16	25	16	0.64
---------	-------	----	-----	------	----	----	----	------

PF00550	PP-binding	150	67	20.8	89	89	87	0.98
---------	------------	-----	----	------	----	----	----	------

PF00501	AMP-binding	136	385	29.9	106	115	106	0.92
---------	-------------	-----	-----	------	-----	-----	-----	------

PF00668	Condensation	21	457	30.6	29	41	29	0.71
---------	--------------	----	-----	------	----	----	----	------

PF00975		Thioesterase		16		231		26.4		20		31		20		0.65
PF13193		AMP-binding_C		222		76		24.5		72		77		70		0.91
PF00195		Chal_sti_synt_N		5		225		27.0		1		5		1		0.20
PF02797		Chal_sti_synt_C		6		151		27.0		2		4		2		0.50
PF05147		LANC_like		45		352		23.2		3		3		3		1.00
PF04738		Lant_dehyd		45		651		28.0		2		2		2		1.00
PF02624		YcaO		181		325		25.7		3		3		3		1.00
PF04183		IucA_IucC		160		240		25.0		4		4		4		1.00
PF00067		p450		50		463		22.8		31		45		31		0.69
PF01494		FAD_binding_3		19		357		22.0		34		71		30		0.42
PF08241		Methyltransf_11		141		96		21.2		109		118		94		
0.80																
PF13489		Methyltransf_23		40		166		27.0		31		100		31		0.31
PF00155		Aminotran_1_2		48		365		19.7		59		114		59		0.52

Generated by: Native 4-Stage Plan-7 Profile HMM, Forward-Backward Log-Odds, Optimal Accuracy & Null2 Bias Engine

Generated by: CRV Science Labs Native Bioinformatics Suite

6. Audit of Every Locus the Rule Engine Did Not Recover

Table 4 lists all 20 curated loci covered by less than half of their length by the native clusters. Of these, 4 are in the fungal genome (out of regime for the gene caller), 8 are bacterial loci in classes that have no core definition (saccharide or other), and 8 are bacterial loci in classes the engine is meant to detect. That last group is the actionable one: it is where profile sensitivity and rule design should be improved. GECCO's coverage of the same loci is shown for context; it too misses several of the saccharide and other-class loci.

Genome	MIBiG Locus	Class	Locus Length (bp)	Native Coverage (fraction)	GECCO 0.8 Coverage (fraction)	Interpretation
--------	-------------	-------	-------------------	----------------------------	-------------------------------	----------------

AL645882.2	BGC0000325	NRPS	22626	0.00	1.00	Core NRPS architecture not recovered at the curated cut-offs, or cluster support below the seeding threshold.
------------	------------	------	-------	------	------	---

AL645882.2	BGC0002127	PKS	3014	0.00	1.00	Type III-like or small cluster whose seed-built domain profile scored below the curated Pfam gathering cut-off.
------------	------------	-----	------	------	------	---

AL645882.2	BGC0000910	other	1420	0.00	1.00	Cluster class outside the 34 signature families (no core definition for this chemistry).
------------	------------	-------	------	------	------	--

AL645882.2	BGC0000940	other	7328	0.00	1.00	Cluster class outside the 34 signature families (no core definition for this chemistry).
------------	------------	-------	------	------	------	--

| AL645882.2 || BGC0000660 || terpene || 2468 || 0.00 || 1.00 || Terpene locus not recognised by the two terpene architectures in the engine (terpene synthase, or paired cyclase domains). |

| AL645882.2 || BGC0001181 || terpene || 2181 || 0.00 || 0.00 || Terpene locus not recognised by the two terpene architectures in the engine (terpene synthase, or paired cyclase domains). |

| AL645882.2 || BGC0000849 || other || 1710 || 0.05 || 1.00 || Cluster class outside the 34 signature families (no core definition for this chemistry). |

| AL645882.2 || BGC0000551 || ribosomal || 6728 || 0.00 || 1.00 || Core NRPS architecture not recovered at the curated cut-offs, or cluster support below the seeding threshold. |

| AL645882.2 || BGC0002128 || other || 8672 || 0.00 || 0.87 || Cluster class outside the 34 signature families (no core definition for this chemistry). |

| CP000560.1 || BGC0000616 || ribosomal || 4173 || 0.00 || 1.00 || Core NRPS architecture not recovered at the curated cut-offs, or cluster support below the seeding threshold. |

| CP000560.1 || BGC0001184 || other || 7298 || 0.00 || 0.00 || Cluster class outside the 34 signature families (no core definition for this chemistry). |

| AE004091.2 || BGC0000922 || PKS || 13136 || 0.00 || 0.00 || Type III-like or small cluster whose seed-built domain profile scored below the curated Pfam gathering cut-off. |

| AE004091.2 || BGC0002037 || NRPS || 16597 || 0.00 || 1.00 || Core NRPS architecture not recovered at the curated cut-offs, or cluster support below the seeding threshold. |

| AE004091.2 || BGC0002645 || other || 6095 || 0.00 || 0.00 || Cluster class outside the 34 signature families (no core definition for this chemistry). |

| AE004091.2 || BGC0000725 || saccharide || 16652 || 0.00 || 0.00 || Cluster class outside the 34 signature families (no core definition for this chemistry). |

| AE004091.2 || BGC0002489 || other || 5082 || 0.00 || 0.00 || Cluster class outside the 34 signature families (no core definition for this chemistry). |

| BN001302.1 || BGC0001515 || NRPS || 64035 || 0.40 || 1.00 || Fungal gene structure: spliced multi-exon genes are fragmented by a bacterial-style gene caller, so domain evidence is split or missing. |

| BN001302.1 || BGC0002758 || PKS || 30659 || 0.00 || 1.00 || Fungal gene structure: spliced multi-exon genes are fragmented by a bacterial-style gene caller, so domain evidence is split or missing. |

| BN001302.1 || BGC0001121 || PKS || 9908 || 0.00 || 1.00 || Fungal gene structure: spliced multi-exon genes are fragmented by a bacterial-style gene caller, so domain evidence is split or missing. |

| BN001302.1 || BGC0000107 || PKS || 6819 || 0.00 || 0.88 || Fungal gene structure: spliced multi-exon genes are fragmented by a bacterial-style gene caller,

so domain evidence is split or missing. |

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

7. Runtime

The native engine was not faster. Its total wall-clock time across the five genomes was 163.7 s, of which 100.9 s was gene calling and 62.7 s was the 34-family domain scan; GECCO at its default cut-off took 127.4 s in total, including a scan against the full Pfam library with a compiled search library. The ratio is a property of these two implementations on one Windows machine rather than of the algorithms, but it does mean the rule engine's benefit here is selectivity, not speed.

Discussion & Key Takeaways

1. Selectivity versus sensitivity. On this panel the rule engine overlapped fewer curated loci than GECCO (15 versus 27 of 33) but reported far fewer clusters (29 versus 108), and a larger share of its clusters overlap a curated locus (14 of 29 versus 24 of 108). Because MIBiG is incomplete, this is a lower bound on precision, and the unconfirmed clusters from either tool may include genuine uncurated BGCs. The result points to a trade-off worth testing further; it does not show one approach to be better.
2. Failures are mostly explicable and partly fixable. Of the 20 missed loci, 12 lie outside the engine's declared scope (fungal gene structure, or chemistry without a core definition). The enterobactin miss was a rule gap that a single added rule repaired, although that repair was made after seeing the failure. The remaining in-scope misses are consistent with profiles built from small seed alignments and a stricter cut-off than GECCO uses.
3. Signal Rating Assessment (2 Technical Validation). The benchmark spans five genomes (27.5 Mb) and 33 curated loci, with one near-negative control and one deliberately out-of-regime fungal genome. It is a reproducible technical benchmark rather than a discovery: $n = 5$ genomes is too few to estimate how performance varies across bacterial clades, MIBiG labels are mostly flagged questionable, the rule revision was informed by the same data it was scored on, and GECCO was used without retraining. We therefore rate it as a technical validation and not as a comparative synthesis.
4. Candidate directions for downstream genome mining. Combining a high-selectivity architecture filter with a high-sensitivity learned detector is a testable strategy for ranking clusters before expensive chemistry or sequencing work. Any such prioritisation is a computational hypothesis for follow-up in the laboratory, not evidence that a cluster produces 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] 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>

* [6] Raymond KN, Dertz EA, Kim SS (2003). Enterobactin: an archetype for microbial iron transport. *Proceedings of the National Academy of Sciences* 100(7):3584-3588. <https://doi.org/10.1073/pnas.0630018100>

* [7] 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](https://doi.org/10.1016/0022-2836(94)90032-9)

* [8] White BP (2026). Calibrating a Biosynthetic Gene Cluster Detector: Released Profiles, Added Marker Families and a Reserve-Genome Test Against GECCO. *CRV Science Labs, Lab Discovery #60* (follow-up to this article). <https://crvscience.com/labs/discoveries/bgc-reserve-genomes>