BGC-Master: Detecting Novel Biosynthetic Gene Clusters with DNA Foundation Models

Fenyi Liu [1] **Yuhong Sun** [1] **Fan Zhang** [1] **Siheng Chen** [1]

Abstract

Biosynthetic gene clusters (BGCs) encode the enzymatic machinery behind microbial natural products, yet current mining tools remain biased toward known biosynthetic motifs. This reliance limits the retrieval of clusters that diverge from canonical motifs in expanding metagenomic collections. We present **BGC-Master**, a method that uses frozen Evo2 7B embeddings and a compact one-dimensional U-Net decoder to localize biosynthetic gene clusters directly from genomic sequence. BGC-Master achieves an F1 of 0.521 on a held-out 9-genome benchmark, outperforming antiSMASH 7 (0.256), DeepBGC (0.117), and BGC-Prophet (0.393) under a shared overlap-based evaluator. Applied to the OER004256 marine metagenome, BGC-Master prioritizes 171 reference-unannotated candidates with biosynthetic evidence and low MIBiG similarity, including regions not recovered by antiSMASH expanded detection. By exposing detection and evidence-based prioritization as agent-ready tools, BGC-Master provides a modular substrate for future agentic natural-product discovery workflows.

1. Introduction

Microbial natural products constitute a major reservoir of bioactive chemical diversity, with broad relevance to medicine, agriculture, and industrial biotechnology (Newman & Cragg, 2020; Walsh & Tang, 2017). In bacteria and fungi, the enzymes that assemble these molecules are commonly encoded by *biosynthetic gene clusters* (BGCs): contiguous genomic loci whose detection is the first computational step in most natural-product discovery pipelines (Medema & Fischbach, 2015). With advances in genome and metagenome sequencing, natural-product discovery has increasingly relied on computational mining of BGC regions from isolate genomes and metagenome-assembled genomes (MAGs) (Parks et al., 2017). However, the sequence search space has grown far faster than experimentally validated BGC annotations and curated motif libraries, leaving many atypical or weakly conserved clusters difficult to recognize.

The current community standard, antiSMASH (Blin et al., 2023), locates BGCs by matching curated profile hidden Markov models (pHMMs) to established biosynthetic enzyme families. While highly precise, this rule-based paradigm is inherently circumscribed by its reliance on pre-defined motifs, often failing to detect clusters that lack canonical signatures. Such "dark" biosynthetic space frequently contains the most pharmacologically promising targets. Although deep-learning detectors such as DeepBGC (Hannigan et al., 2019), ClusterFinder (Cimermancic et al., 2014), GECCO (Carroll et al., 2021), and the recent transformer-based BGC-Prophet (Lai et al., 2025) aim to broaden this search, all of them operate over protein-domain (Mistry et al., 2021) or ESM2-style protein embedding features (Lin et al., 2023) rather than raw DNA. This design forces an upstream gene-calling (Hyatt et al., 2010) and Pfam-annotation step (Mistry et al., 2021) whose errors propagate into detection on draft metagenomic contigs, and discards nucleotide-level signals known to be informative for biosynthetic cluster identification, such as compositional bias (GC content and codon usage) (Sharp & Li, 1987), operon architecture, and proximity to mobile genetic elements (Frost et al., 2005).

We introduce **BGC-Master**, the first BGC detector that operates directly on raw DNA via a frozen genomic foundation model and bypasses the protein-level intermediate altogether. Among the recent wave of genomic foundation models (Nguyen et al., 2023; 2024; Dalla-Torre et al., 2024; Zhou et al., 2024; Brixi et al., 2025), we use Evo2 7B as the backbone because of its base-level tokenization and demonstrated transfer to functional genomic tasks (Shearer et al., 2025). BGC-Master pairs frozen Evo2 7B per-token embeddings with a lightweight 1D U-Net (Ronneberger et al., 2015) segmentation head trained under weak-negative binary cross-entropy (Bekker & Davis, 2020), followed by a deterministic HMM/MIBiG triage gate that produces wet-lab-ready candidate hypotheses.

[1]Shanghai Jiao Tong University. Correspondence to: Siheng Chen <sihengc@sjtu.edu.cn>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

On a held-out 9-genome benchmark, BGC-Master attains F1 0.521 (precision 0.620, recall 0.449), improving over antiSMASH 7 (0.256), DeepBGC (0.117), and BGC-Prophet (0.393) under a shared overlap-based evaluator. On the OER004256 marine metagenome, it nominates 171 reference-unannotated candidate regions, none of which are recovered even by antiSMASH in expanded detection mode (Blin et al., 2025). The full pipeline is exposed as a tool-callable REST service, providing a drop-in primitive for future closed-loop biology agents (Boiko et al., 2023; Zhu et al., 2026) in which both the detector and the agent's prioritization policy could be progressively refined as wet-lab outcomes accumulate.

2. Methods

2.1. Task Formulation and Overview

We define BGC discovery as the problem of mapping a raw genomic or metagenomic DNA sequence to candidate loci for downstream wet-lab triage. Given an input contig $S = (s_1, s_2, \ldots, s_L)$, where $s_i \in \{A, T, C, G, N\}$, the pipeline outputs a set of candidate regions $\mathcal{R} = \{(a_m, b_m, p_m, \tau_m)\}_{m=1}^{M}$. Here, $[a_m, b_m)$ denotes the genomic coordinates of the m-th candidate region, $p_m \in [0, 1]$ is a region confidence score, and τ_m denotes optional downstream annotations such as product type or triage evidence.

BGC-Master instantiates this mapping with a dense sequence segmentation detector followed by region decoding and triage. Each fixed-length DNA window is encoded by a frozen Evo2 backbone and compressed into 8-bp pooled latent tokens. A compact one-dimensional segmentation head predicts one BGC probability for each latent token. During genome-wide inference, local token-level predictions are projected back to genomic coordinates and averaged across overlapping windows to form a continuous BGC probability track. Candidate regions are then decoded from high-confidence contiguous segments of this track and annotated for downstream prioritization.

2.2. Evo2 Representations for DNA Windows

We use Evo2 7B as a frozen representation backbone for DNA sequence windows. Each input sequence is cleaned to contain only A, T, C, G, and N, with nonstandard characters mapped to N. In our extraction pipeline, cleaned DNA windows are tokenized with Evo2's byte-level tokenizer, giving one token per input base. This one-to-one alignment allows model representations and downstream predictions to be mapped back to genomic coordinates.

For each window, we extract token-level hidden states from the `blocks.20.mlp.l3` layer of Evo2 7B (the third linear projection of the MLP sublayer in transformer block 20) and use this layer consistently across all experiments. An auxiliary layer-probe study is provided in Appendix F. Evo2 parameters are kept frozen in all experiments. Freezing the backbone reduces training cost and forces the downstream segmentation head to rely strictly on the generalized representations learned during pretraining.

The raw Evo2 hidden states are high-dimensional, with 4,096 features per base token, making feature caching and dense genome-wide training expensive. We therefore compress each hidden state to 128 dimensions using a fixed Gaussian random projection before pooling. The projection matrix is sampled once with entries $R_{ij} \sim \mathcal{N}(0, 1/128)$ and then held fixed for all experiments. This non-parametric step substantially reduces memory and I/O cost and is motivated by the use of random projections as lightweight geometry-preserving compression (Johnson & Lindenstrauss, 1984; Achlioptas, 2003).

After projection, every eight consecutive base-level states are averaged to form one latent token. Thus, each 16,384 bp window is represented as a 2048×128 latent sequence. This representation provides a practical resolution for region-level BGC detection: it is fine enough to recover candidate boundaries at genomic scale, while compact enough to train and apply the segmentation head over large genomes and metagenomic contigs. Formally, for Evo2 hidden states $\mathbf{h}_i \in \mathbb{R}^{4096}$, a fixed projection matrix $\mathbf{R} \in \mathbb{R}^{4096\times128}$, and pooling factor $K = 8$, the downstream token representation is

$$\mathbf{z}_t = \frac{1}{8}\sum_{r=1}^{8} \mathbf{h}_{8(t-1)+r}\mathbf{R} \in \mathbb{R}^{128}. \tag{1}$$

Unless otherwise specified, we use "token" below to refer to this 8-bp pooled latent representation, rather than Evo2's original base-level tokenizer output.

2.3. One-Dimensional Segmentation Head

To transform Evo2 latent features into BGC probabilities, we use a lightweight 1D U-Net architecture (Ronneberger et al., 2015). Our goal is not to introduce a new segmentation architecture, but to use a simple inductive bias that matches the structure of BGC detection: mapping a $T \times 128$ latent sequence to one logit per 8-bp latent token while aggregating broad genomic context and preserving local boundary information. Let g_θ denote the segmentation head; it maps the token sequence to logits ℓ_t, which are converted to BGC probabilities $p_t = \sigma(\ell_t)$.

The network uses a four-level encoder-decoder structure. BGC detection benefits from context beyond short local motifs, so the encoder progressively downsamples the latent sequence by a total factor of 256 (4^4). For a standard 2,048-token window, this produces an 8-token bottleneck, allowing the deepest layers to summarize information over the input window at a coarse resolution. The decoder then

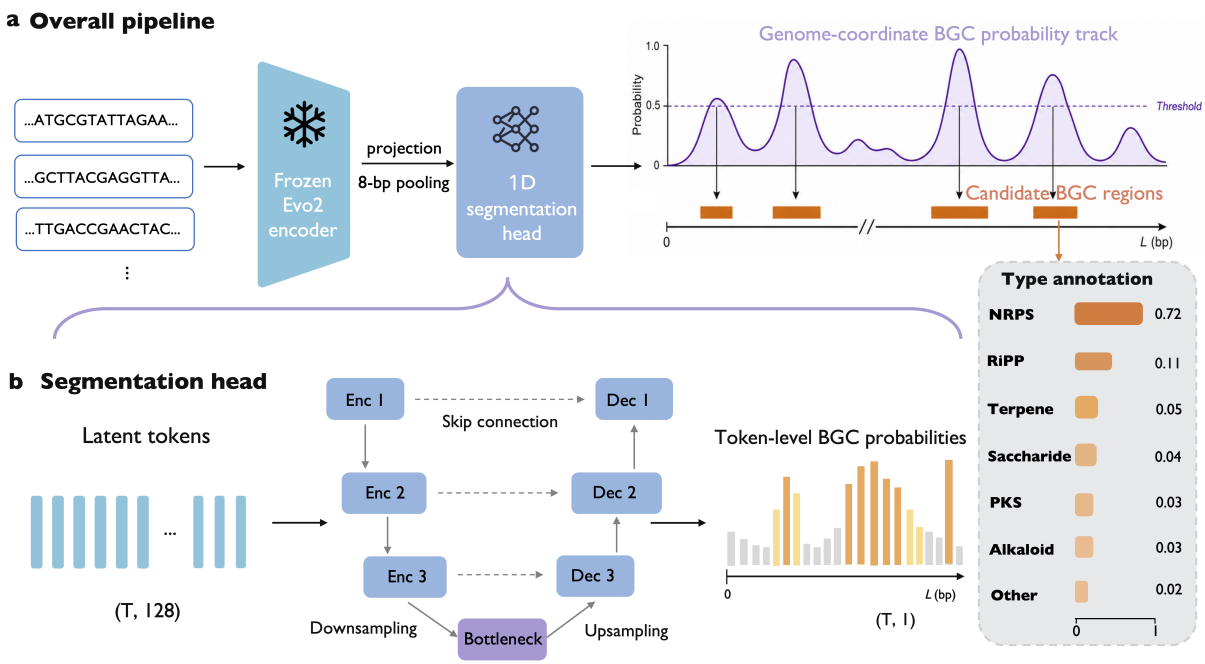

Figure 1. Overview of the proposed BGC discovery pipeline. **a**, Long genome or metagenomic contig sequences are encoded by a frozen Evo2 backbone. Evo2 hidden states are compressed through projection and 8-bp pooling, then passed to a lightweight 1D segmentation head to produce a genome-coordinate BGC probability track. Candidate BGC regions are decoded from high-confidence intervals and optionally assigned product-type annotations for downstream interpretation. **b**, The segmentation head maps Evo2-derived latent tokens to token-level BGC probabilities using a compact 1D U-Net decoder. The encoder aggregates broader genomic context through downsampling, while skip connections preserve local boundary cues for region-level BGC calling.

upsamples the representation back to token resolution, with skip connections reintroducing higher-resolution local signals for boundary localization.

The segmentation head contains approximately 320K trainable parameters, keeping the task-specific module compact relative to the frozen Evo2 backbone. A final 1×1 convolution produces one logit per latent token, and a sigmoid activation converts the logits into a token-level BGC probability track. Full architectural details, including block configurations and normalization layers, are provided in the Appendix.

2.4. Training Supervision and Objective

We train the segmentation head using genome windows paired with base-pair-level BGC region annotations derived from antiSMASH (Blin et al., 2023). The training corpus consists of bacterial genome assemblies identified by GCA/GCF accessions; the window metadata contains 1,064,568 16,384 bp windows from 1,996 assemblies, with windows assigned to train, validation, and test partitions at the genome level. For each window, annotated BGC intervals are converted into a dense binary mask over genomic coordinates. Positions overlapping annotated BGC regions are labeled as positive, positions outside annotated regions are treated as background, and padding positions are ignored. Splitting by genome, rather than by window, prevents highly similar windows from the same assembly from appearing in different partitions.

Because the decoder predicts at latent-token resolution, we downsample the base-pair-level mask to match the 8-bp latent tokens. A latent token is labeled positive if any base in its 8-bp span overlaps an annotated BGC. Equivalently, if $b_i \in \{0, 1\}$ denotes the base-pair-level BGC mask, then $y_t = \mathbb{I}[\sum_{r=1}^{8} b_{8(t-1)+r} > 0]$. Tokens overlapping padding or precomputed ambiguous boundary/proximity windows around antiSMASH regions are masked from the loss. We additionally define a binary loss mask m_t, where $m_t = 0$ for tokens overlapping padding or ambiguous tokens and $m_t = 1$ otherwise.

BGC annotations are incomplete: annotated regions provide reliable positives, whereas unannotated regions may still contain unknown or atypical clusters. We therefore train with a masked weak-negative weighted binary cross-entropy loss, an instance of the broader positive-unlabeled learning paradigm (Bekker & Davis, 2020). Positive tokens receive weight 1.0, unannotated background tokens receive weight 0.5, and masked tokens are excluded. This preserves negative supervision for calibration while reducing the penalty on

potentially BGC-like unannotated regions. For a minibatch of windows, let $L_{n,t} = \mathrm{BCEWithLogits}(\ell_{n,t}, y_{n,t})$ and let $w_{n,t}$ be 1.0 for positive tokens and 0.5 for background tokens. The objective is the masked weighted average

$$\mathcal{L}(\theta) = \frac{\sum_{n,t} m_{n,t} w_{n,t} L_{n,t}}{\sum_{n,t} m_{n,t} w_{n,t}}, \tag{2}$$

where masked tokens do not contribute to either the numerator or the normalization term.

Only the segmentation head is optimized during training. Evo2 features are extracted offline and cached, with no gradients propagated through the foundation model. We train with AdamW (Loshchilov & Hutter, 2019) using learning rate 1×10^{-3}, global batch size 64, seed 0, and early stopping with patience 6 over a maximum of 20 epochs. Validation uses token-level precision, recall, and F1 as diagnostic metrics, while final evaluation is reported at the region level.

2.5. Genome-Wide Inference and Region Calling

During inference, we scan long genomes and metagenomic contigs using overlapping sequence windows. Although the segmentation head is trained on 16,384 bp windows, genome-wide inference uses 8,192 bp windows with a 2,048 bp stride (75% overlap). This shorter inference window reduces feature-extraction cost and improves coverage near contig ends, while the fully convolutional segmentation head can be applied to variable-length token sequences without architectural changes. Each window is encoded and scored independently, producing token-level BGC probabilities within the local window. These probabilities are then mapped back to their original genomic coordinates and averaged across overlapping windows, yielding a continuous BGC probability track for each contig. This genome-coordinate track allows candidate regions to be decoded from the aggregated signal rather than from independent window-level decisions. For contig c, let $\mathcal{W}_c(i)$ be the set of scored windows covering genomic coordinate i, and let $p_w(i)$ be the probability assigned to that coordinate by window w after expanding token probabilities back to base-pair coordinates. The aggregated contig-level track is

$$P_c(i) = \frac{1}{|\mathcal{W}_c(i)|} \sum_{w \in \mathcal{W}_c(i)} p_w(i). \tag{3}$$

We decode candidate regions directly from this probability track. High-confidence contiguous segments are retained, nearby segments are merged when they likely correspond to the same cluster, and short isolated predictions are removed. Unless otherwise specified, decoding uses a probability threshold of 0.5, merges gaps shorter than 500 bp, and removes candidate intervals shorter than 2 kb. Each retained interval is reported as a candidate BGC with genomic coordinates and a region confidence score, computed from the average probability within the interval. Alternative operating thresholds and length filters are treated as readout choices and reported with each experiment.

After region detection, we optionally attach coarse product-type annotations for interpretation and candidate prioritization. We use lightweight one-versus-rest classifiers trained on Evo2-derived window embeddings and antiSMASH product labels. For each detected region, type probabilities from overlapping windows are aggregated to assign a primary type label and auxiliary type scores. This annotation layer is used downstream for triage and does not affect region detection.

3. Experiments and Results

3.1. Experimental Setup

9-Genome BGC Benchmark. A held-out 9-genome benchmark provides 305 curated ground-truth BGC regions across nine fully-sequenced bacterial isolate genomes (Cimermancic et al., 2014). We use region-level coverage criteria for evaluation. A recall true positive requires the best same-contig prediction to cover at least 50% of a ground-truth region ($\mathrm{overlap}/\mathrm{gt_len} \geq 0.5$), while a precision true positive requires a same-contig ground-truth region to cover at least 50% of the prediction ($\mathrm{overlap}/\mathrm{pred_len} \geq 0.5$). We report precision, recall, and F1 under this region-level protocol.

OER004256 Marine Metagenome. An internal test set of 1,976 antiSMASH 7 reference regions on contigs ≥ 10 kb (median length 13,573 bp). Discovery quality is reported as antiSMASH-overlap recall (AS-recall): an antiSMASH reference region is counted as recovered when the best same-contig prediction covers at least 50% of its length. Predictions that do not overlap any antiSMASH region on the same contig form the *reference-unannotated candidate* set used for downstream triage.

Triage Gate. For each reference-unannotated candidate we attach Pfam (Mistry et al., 2021) and biosynthetic HMM evidence (via HMMER (Eddy, 2011)), BGC-type annotations, and MIBiG 4.0 (Augustijn et al., 2024) similarity. A candidate is marked first-pass PASS when it satisfies all of: ≥ 3 biosynthetic HMM hits, ≥ 2 biosynthetic ORFs, region length ≥ 5 kb, and maximum MIBiG protein identity below 70%.

3.2. 9-Genome BGC Benchmark Result

We evaluate BGC-Master, a weak-negative latent-token U-Net detector over frozen Evo2 features, on the 9-genome BGC benchmark. Table 1 reports the operating point used

Table 1. 9-genome BGC benchmark readout. Precision uses overlap/pred_len ≥ 0.5 and recall uses overlap/gt_len ≥ 0.5. Total ground-truth count: 305 BGC regions across 9 genomes.

Method	Precision	Recall	F1
antiSMASH 7	0.166	0.557	0.256
DeepBGC	0.064	**0.721**	0.117
BGC-Prophet	0.453	0.348	0.393
BGC-Master	**0.620**	0.449	**0.521**

for this readout (threshold 0.70, minimum region length 8 kb), together with antiSMASH 7 (Blin et al., 2023), DeepBGC (Hannigan et al., 2019), and BGC-Prophet (Lai et al., 2025). Under this protocol, BGC-Master obtains precision 0.620, recall 0.449, and F1 0.521.

At this operating point, BGC-Master emits 221 regions on the 9-genome benchmark. Recall is highest for longer regions (40/61 for 20–50 kb and 16/22 for $\geq$ 50 kb) and lower for sub-5 kb ground-truth regions (6/48), matching the known small-region blind spot.

The compared tools show different precision–recall profiles. DeepBGC recovers the largest fraction of reference regions (recall 0.721) but emits 1,132 calls, yielding low precision (0.064). BGC-Prophet is more conservative, with precision 0.453 and recall 0.348. antiSMASH 7 recovers 170 of 305 ground-truth regions, but its broader region boundaries lead to precision 0.166 under this coverage-based protocol.

3.3. Metagenomic Discovery on OER004256

We evaluate BGC-Master on the OER004256 marine metagenome at three operating points. Table 2 summarizes the results. The discovery operating point (threshold 0.50, minimum length 2 kb) recovers 1,127 of 1,976 antiSMASH regions (AS-recall 0.570) while emitting 2,199 regions. Of these predictions, 579 do not overlap any antiSMASH region on the same contig and therefore enter the AS-unannotated discovery bucket. More conservative operating points shrink the handoff bucket: at threshold 0.70 with an 8 kb minimum length the model emits 617 regions with AS-recall 0.251 and 33 AS-unannotated candidates.

The discovery readout is strongest for shorter antiSMASH regions: at threshold 0.50 and minimum length 2 kb, AS-recall is 759/1150 (0.660) for 10–20 kb antiSMASH regions, 296/621 (0.477) for 20–50 kb regions, and 72/205 (0.351) for regions $\geq$ 50 kb.

Across the 2,199 detected regions at the discovery operating point, the post-hoc type annotation layer assigns labels with the following composition: terpene 1,000, RiPP 474, other 373, polyketide 224, NRP 125, and saccharide 3. This annotation layer is used only for downstream interpretation and triage; it does not affect U-Net region detection.

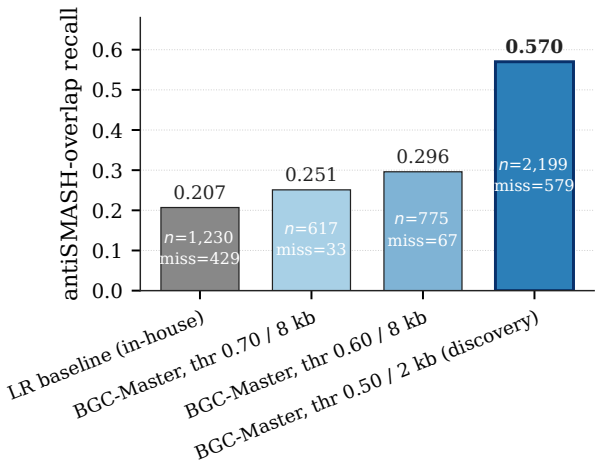

Figure 2. antiSMASH-overlap recall on OER004256 across operating points. Bar labels show the number of region calls (n) and the AS-unannotated count for each operating point.

3.4. Wet-Lab Triage Funnel

We apply the deterministic triage gate of Section 3.1 to the 579 AS-unannotated candidates produced at the discovery operating point. The funnel is summarized in Figure 3. Marginally, 357 of the 579 candidates carry strong biosynthetic HMM evidence ($n_{\text{HMM}} \geq 3$); a further 78 carry moderate evidence (1–2 hits); and 144 are filtered out at this stage as having weak or no informative HMM support. Independently, 304 candidates fall below the 70% MIBiG identity ceiling (170 likely-novel, 134 fully novel), suggesting that a substantial fraction of the AS-unannotated bucket is distant from known MIBiG entries even before applying biosynthetic evidence filters.

The full conjunction of all four criteria (HMM hits, biosynthetic ORF count, region length, and MIBiG identity) yields **171 first-pass PASS candidates**, about 30% of the AS-unannotated bucket, which we forward as the wet-lab handoff list. Two properties make the gate practical: it is fully deterministic, so the handoff is reproducible from the regions table without re-running the model, and it is auditable per row, so a wet-lab collaborator can inspect why a specific candidate was kept or dropped.

Table 2. OER004256 metagenomic discovery with BGC-Master. AS-recall counts antiSMASH regions for which the best same-contig prediction covers at least 50% of the reference length. The "AS-unannotated" column counts predictions with no antiSMASH overlap on the same contig and defines the discovery bucket.

Operating point	Regions	AS-recall	AS-unannotated	Median len (bp)
thr 0.50, min 2 kb (discovery)	**2,199**	**0.570**	**579**	**8,192**
thr 0.60, min 8 kb	775	0.296	67	10,122
thr 0.70, min 8 kb	617	0.251	33	10,240

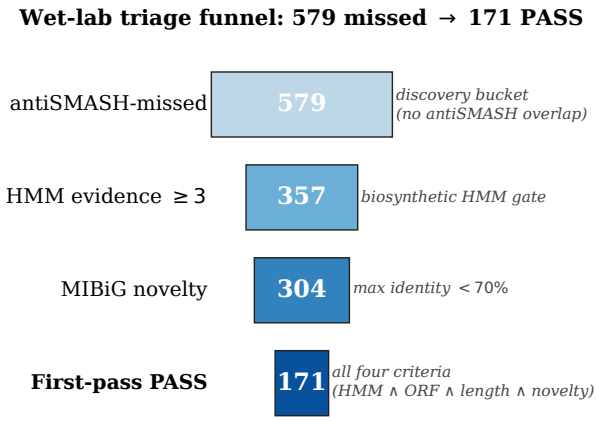

Figure 3. Wet-lab triage funnel applied to the 579 AS-unannotated regions from BGC-Master on OER004256. Marginal counts at each filter are shown; the first-pass PASS bucket is the conjunction of all four conditions (HMM evidence, biosynthetic ORF count, region length, and MIBiG novelty).

4. Conclusion

BGC-Master demonstrates that a frozen genomic foundation model paired with a parameter-efficient segmentation head trained on top outperforms established rule-based and deep-learning BGC detectors on isolate genomes under a shared overlap-based evaluator. Crucially, it surfaces a large bucket of reference-unannotated biosynthetic loci that elude both default and expanded antiSMASH classification, expanding the candidate set available for wet-lab follow-up. The modular nature of this method allows the detection and prioritization logic to function as a standardized building block, ready for future integration into autonomous discovery pipelines and downstream cluster-similarity tooling such as BiG-SCAPE (Navarro-Muñoz et al., 2020). This design ensures that as experimental validation data accumulates, the system can be iteratively enhanced without requiring structural modifications. We view this methodology as a scalable template for embedding foundation models within evolving scientific workflows, bridging the gap between computational genome mining and wet-lab natural-product discovery.

Impact Statement

BGC-Master accelerates discovery of novel biosynthetic gene clusters that may yield new antibiotics or other bioactive compounds. Foundation-model representations may concentrate predictions on taxa over-represented in pretraining, biasing wet-lab follow-up; we mitigate this with explicit MIBiG-distance gating that rewards divergence from known chemistry. The system processes publicly available genome data only and does not provide any synthesis guidance for the predicted compounds. We will release the service as open infrastructure to the GenBio community to enable collaborative benchmarking and downstream agentic composition.

References

Achlioptas, D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. doi: 10.1016/S0022-0000(03)00025-4.

Augustijn, H. E., Pereira, L., Loureiro, C., Augustijn, W., Maw, A., Reitz, Z. L., Selem-Mojica, N., et al. MIBiG 4.0: Advancing biosynthetic gene cluster curation through community engagement. *Nucleic Acids Research*, 2024. doi: 10.1093/nar/gkae1115.

Bekker, J. and Davis, J. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109:719–760, 2020. doi: 10.1007/s10994-020-05877-5.

Blin, K., Shaw, S., Augustijn, H. E., Reitz, Z. L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B. R., Metcalf, W. W., Helfrich, E. J. N., van Wezel, G. P., Medema, M. H., and Weber, T. antiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research*, 51 (W1):W46–W50, 2023. doi: 10.1093/nar/gkad344.

Blin, K., Shaw, S., Vader, L., Demir, F., Reitz, Z. L., Augustijn, H. E., Hansen, S. H., Klau, L. J., Gren, T., Booth, T., Donadio, S., Iorio, M., Sosio, M., Medema, M. H., and Weber, T. antiSMASH 8.0: Extended gene cluster detection capabilities and updated visualisation. *Nucleic Acids Research*, 2025. doi: 10.1093/nar/gkaf334.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., et al. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918.

Buchfink, B., Reuter, K., and Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, 2021. doi: 10.1038/s41592-021-01101-x.

Carroll, L. M., Larralde, M., Fleck, J. S., Ponnudurai, R., Milanese, A., Cappio Barazzone, E., and Zeller, G. Accurate de novo identification of biosynthetic gene clusters with GECCO. *bioRxiv*, 2021. doi: 10.1101/2021.05.03.442509.

Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Linington, R. G., and Fischbach, M. A. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, 158(2):412–421, 2014. doi: 10.1016/j.cell.2014.06.034.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., and Pierrot, T. Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, 2024. doi: 10.1038/s41592-024-02523-z.

Eddy, S. R. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10):e1002195, 2011. doi: 10.1371/journal.pcbi.1002195.

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. Mobile genetic elements: The agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005. doi: 10.1038/nrmicro1235.

Hannigan, G. D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D. J., Woelk, C. H., and Bitton, D. A. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, 47(18):e110, 2019. doi: 10.1093/nar/gkz654.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, 2010. doi: 10.1186/1471-2105-11-119.

Johnson, W. B. and Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Lai, Q., Yao, S., Zha, Y., Zhang, H., Zhang, H., Ye, Y., Zhang, Y., Bai, H., and Ning, K. Deciphering the biosynthetic potential of microbial genomes using a BGC language processing neural network model. *Nucleic Acids Research*, 53(7):gkaf305, 2025. doi: 10.1093/nar/gkaf305.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Medema, M. H. and Fischbach, M. A. Computational approaches to natural product discovery. *Nature Chemical Biology*, 11(9):639–648, 2015. doi: 10.1038/nchembio.1884.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021. doi: 10.1093/nar/gkaa913.

Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L. C., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L. T. D., Goering, A. W., Thomson, R. J., Metcalf, W. W., Kelleher, N. L., Barona-Gomez, F., and Medema, M. H. A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology*, 16:60–68, 2020. doi: 10.1038/s41589-019-0400-9.

Newman, D. J. and Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, 83(3):770–803, 2020. doi: 10.1021/acs.jnatprod.9b01285.

Nguyen, E., Poli, M., Faizi, M., Thomas, A. W., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S. A., and Ré,

C. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723): eado9336, 2024. doi: 10.1126/science.ado9336.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., and Tyson, G. W. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2:1533–1542, 2017. doi: 10.1038/s41564-017-0012-7.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Sharp, P. M. and Li, W.-H. The codon adaptation index — a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15 (3):1281–1295, 1987. doi: 10.1093/nar/15.3.1281.

Shearer, C. A., Teufel, F., Orenbuch, R., Steinmetz, C. J., Ritter, D., Xie, E., Gazizov, A., Spinner, A., Frazer, J., Dias, M., Notin, P., and Marks, D. S. A genomic language model for zero-shot prediction of promoter indel effects. In *Proceedings of the 2nd ICML Workshop on Generative AI and Biology (GenBio)*, 2025.

Walsh, C. T. and Tang, Y. *Natural Product Biosynthesis: Chemical Logic and Enzymatic Machinery*. Royal Society of Chemistry, 2017.

Wu, Y. and He, K. Group normalization. In *European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhu, X., Cai, Y., Liu, Z., Cheng, W., Li, F., Jin, W., Liu, W., Bing, Z., Zheng, B., Chai, J., Tang, S., Ye, R., Du, Y., Pang, X., Du, Y., Miao, T., Zhang, Y., Liao, R., Ding, Z., Zhang, L., Wang, Y., E, W., and Chen, S. Evomaster: A foundational evolving agent framework for agentic science at scale. *arXiv preprint arXiv:2604.17406*, 2026.

A. Implementation and Compute

The BGC-Master detector uses a compact one-dimensional U-Net segmentation head on top of frozen Evo2 representations. The head has approximately 320K trainable parameters and follows a four-level encoder-decoder design. Each encoder level downsamples the latent-token sequence by a factor of 4, giving a total downsampling factor of 256. Residual blocks use GroupNorm (Wu & He, 2018) and SiLU activations. A final 1×1 convolution produces one logit for each 8-bp latent token.

Evo2 7B is kept frozen throughout training and inference. For each 16,384-bp DNA window, Evo2 hidden states are projected to 128 dimensions and pooled every 8 bp, yielding a 2048×128 latent sequence. These latent features are extracted offline and cached as float16 tensors before training the segmentation head. This caching step avoids repeated Evo2 forward passes during detector training.

The segmentation head is trained with AdamW using learning rate 1×10^{-3}, weight decay 10^{-4}, gradient clipping at $\ell_2 = 1.0$, global batch size 64, and seed 0. Training uses 4-way DDP on A800 GPUs. We apply early stopping with patience 6 over a maximum of 20 epochs; typical runs converge after 8 to 12 epochs.

B. Triage Protocol Details

The triage layer is applied only after neural region detection. Its purpose is to convert AS-unannotated model predictions into an auditable handoff list for downstream biological inspection. The gate does not affect the segmentation model, the genome-wide probability track, or the decoded region boundaries.

For each AS-unannotated candidate region, we first predict open reading frames (ORFs) on the corresponding contig sequence using Prodigal (Hyatt et al., 2010) and retain ORFs that overlap the candidate interval. Protein sequences from these ORFs are searched against a curated biosynthetic HMM library and against Pfam (Mistry et al., 2021) for contextual annotation. Biosynthetic HMM evidence is computed with HMMER 3.3.2 (Eddy, 2011) using `hmmscan`. A biosynthetic HMM hit is counted when it passes the reporting cutoff used in our pipeline ($E < 10^{-5}$). The curated biosynthetic HMM library contains profiles for major secondary-metabolism signatures, including NRPS condensation and adenylation domains, PKS ketosynthase and acyltransferase domains, terpene cyclases, lanthipeptide-associated domains, RiPP recognition elements, tailoring enzymes, and transport-related biosynthetic features. Pfam matches are retained for interpretation but are not by themselves sufficient for PASS status unless they also correspond to a biosynthetic HMM profile in the curated library.

For a candidate region r, we compute two HMM-derived quantities: $n_{\mathrm{HMM}}(r)$, the number of biosynthetic HMM hits in the region, and $n_{\mathrm{ORF}}(r)$, the number of distinct ORFs carrying at least one biosynthetic HMM hit. These two quantities capture different evidence. The first measures the total amount of biosynthetic-domain support, while the second ensures that the signal is distributed across at least two coding sequences rather than arising from a single isolated domain hit.

Novelty relative to known BGCs is estimated by comparing candidate ORF proteins against the MIBiG 4.0 protein release (Augustijn et al., 2024). We run DIAMOND BLASTp (Buchfink et al., 2021) between candidate-region ORFs and MIBiG proteins, and define the MIBiG identity ceiling of a region as the maximum protein percent identity over all retained ORF-level matches:

$$I_{\mathrm{MIBiG}}(r) = \max_{o \in r,\ q \in \mathrm{MIBiG}} \mathrm{pident}(o, q).$$

Regions with no retained MIBiG protein match are treated as having no detectable MIBiG analogue and pass the identity-ceiling criterion.

The first-pass PASS gate is a deterministic conjunction of four criteria:

$$\mathrm{PASS}(r) = \mathbb{I}\left[n_{\mathrm{HMM}}(r) \geq 3 \ \wedge \ n_{\mathrm{ORF}}(r) \geq 2 \ \wedge \ \mathrm{len}(r) \geq 5{,}000 \ \wedge \ I_{\mathrm{MIBiG}}(r) < 70\right].$$

Here, $\mathrm{len}(r)$ is computed as the half-open genomic span $\mathrm{end}(r) - \mathrm{start}(r)$ in base pairs. The 5 kb length criterion removes very short isolated predictions, the HMM and ORF criteria require interpretable biosynthetic evidence, and the MIBiG identity ceiling filters out candidates that are close protein-level variants of known MIBiG entries.

Product-type annotations are attached after region detection and are used for interpretation and prioritization, not for the PASS decision. The type annotation layer is a lightweight one-versus-rest classifier trained on Evo2-derived region embeddings and antiSMASH product labels. For each candidate, we report the top predicted product type. When the type

classifier and HMM evidence disagree, we treat the HMM evidence as the operative signal for triage, because the PASS gate is defined entirely by deterministic sequence and profile-search evidence.

The triage gate was fixed before counting PASS candidates. Therefore, the final PASS count is reproducible from the released region table and does not require re-running the neural detector. Each row in the handoff table records the values of n_{HMM}, n_{ORF}, region length, and I_{MIBiG}, allowing downstream collaborators to audit why a candidate was retained or filtered.

C. Top Candidate Loci for Wet Lab Handoff

To make the AS-unannotated PASS set biologically inspectable, Table 3 lists the ten highest priority candidate loci from OER004256. These candidates all pass the deterministic triage gate described in Appendix B. They are ranked by the priority score used in the released handoff table, which combines model confidence, biosynthetic evidence, region length, and MIBiG identity ceiling.

Several entries show low similarity to known MIBiG proteins. The first four candidates have maximum MIBiG protein identity below 30%, and two candidates have no retained MIBiG protein match. This pattern suggests that these loci are distant from known MIBiG entries at the protein sequence level. It should not be interpreted as proof of biochemical novelty, because metagenomic assembly errors, fragmented loci, and unmodeled background sequence can also produce low similarity.

The table also illustrates why the type annotation layer is treated as supplementary. The post hoc Evo2 type label provides a coarse product class, while the HMM and Pfam evidence provide the more direct sequence based rationale for triage. When these two signals disagree, the HMM evidence should be treated as the operative evidence for handoff decisions. For example, some candidates receive a Terpene label from the type classifier while their representative domains are more consistent with PKS like or tailoring enzyme evidence. Candidate 6 is a clear RiPP associated example, carrying `Peptidase_C39` together with `Lant_dehydr_N`/`Lant_dehydr_C`, a characteristic lanthipeptide associated profile.

Table 3. Top 10 candidate BGC loci for wet lab handoff on OER004256, ordered by priority score. "Type" is the post hoc Evo2 type label. "Representative domains" lists selected biosynthetic or contextual domains from HMMER and Pfam after deduplication. "Closest MIBiG" shows the highest scoring MIBiG 4.0 hit and the maximum protein identity (%id) to any predicted ORF in parentheses. "None" denotes no retained MIBiG protein match. Coordinates are zero based and half open. Sample IDs all begin with `OER0042`. Full profiles for all 171 PASS candidates accompany the release.

#	Locus (sample/contig:kb)	Len	Type	Representative domains	Closest MIBiG (%id)
1	5671 / k141_1357626 : 1.1–10.2	9.2	Terpene	ADSI-DH, ADH_N, MannoseP	tilivalline NRPS (25)
2	5668 / k141_10677679 : 28.7–34.8	6.1	Terpene	ADSI-DH, ADH_N, T2TS, PKS_ER	aurantinin PKS (28)
3	5670 / k141_9311326 : 4.1–10.4	6.3	RiPP	NTP_transf_5, Wzy_C, Lasso_RRE	None
4	5668 / k141_9159969 : 0.0–8.2	8.2	RiPP	NTP_transf_5, Poly_export	None
5	5671 / k141_5790616 : 10.2–20.5	10.2	RiPP	TIGR03104, Asn_synthase, GATase_7	gladiofungin PKS (37)
6	5673 / k141_11681567 : 8.2–14.3	6.1	RiPP	Peptidase_C39, Lant_dehydr_N/C*	pinensins RiPP (45)
7	5672 / k141_1644009 : 8.2–16.4	8.2	Terpene	α-am_amid, AMP-binding, novH	scytocyclamide (33)
8	5670 / k141_5835553 : 15.4–22.5	7.1	RiPP	DegT_DnrJ, TunD, RmlD_sub_bind	leucomycin PKS (39)
9	5670 / k141_2228349 : 1.4–8.2	6.8	Other	novJ, EntA, fabH, adh_short	saccharothrixin (39)
10	5670 / k141_2407267 : 40.9–48.6	7.7	RiPP	TIGR04103, PF04055 (rSAM), subtilosin	WGK RiPP (40)

* Lanthipeptide associated domain combination.

D. Per-Genome Region-Level Metrics

To examine whether the aggregate 9-genome result is driven by a small number of genomes, we report per-genome precision, recall, and F1 under the same region-level overlap protocol and operating point used in Table 1. This breakdown is intended as a diagnostic view of performance heterogeneity across genomes. Each genome contains only 26–47 annotated BGC regions, so the per-genome values should be interpreted with caution.

BGC-Master performs well on most genomes but shows a clear failure case on `GCA_000154945.1`, which has the shortest median ground-truth BGC length in the benchmark. antiSMASH 7 also drops sharply on the same genome, suggesting that this case is challenging for both learned and rule-based detection. This pattern is consistent with the small-region blind spot

discussed in Section 3.2.

Table 4. Per-genome region-level Precision / Recall / F1 on the held-out 9-genome benchmark, under the same overlap-based evaluator as Table 1. "n_{GT}" is the number of ground-truth BGC regions on that genome. "med len" is the median ground-truth cluster length in kb. Best F1 per row is bolded.

#	Genome	n_{GT}	med len (kb)	antiSMASH 7 P / R / F1	DeepBGC P / R / F1	BGC-Master P / R / F1
1	GCA_000158915	47	12.8	.28 / .79 / .41	.08 / .85 / .14	**.90 / .66 / .76**
2	GCA_000158875	32	9.5	.06 / .56 / .10	.05 / .72 / .09	**.50 / .56 / .53**
3	GCA_000154945	34	7.1	.04 / .06 / .05	**.03 / .29 / .06**	.06 / .00 / .00
4	GCA_000158895	35	9.1	.07 / .43 / .12	.03 / .71 / .06	**.40 / .43 / .41**
5	GCA_000156435	35	10.6	.10 / .43 / .16	.03 / .60 / .06	**.27 / .23 / .25**
6	GCA_000568915	31	12.3	.17 / .52 / .25	.10 / .74 / .18	**.62 / .39 / .48**
7	GCA_000158815	27	18.4	.33 / .78 / .47	.13 / .96 / .23	**.83 / .52 / .64**
8	GCA_000568255	38	13.7	.19 / .71 / .29	.06 / .89 / .11	**.80 / .58 / .67**
9	GCA_000156475	26	9.8	.21 / .73 / .32	.11 / .69 / .18	**.80 / .65 / .72**

E. Auxiliary Probe for Evo2 Layer Selection

We performed an auxiliary layer-probe study to guide the choice of Evo2 features used in the main detector. This study was conducted before training the final segmentation head and is intended as a sanity check for layer selection, rather than as an ablation of the full U-Net detector.

We compared single-layer Evo2 representations from layers 13, 20, and 27, as well as their concatenation. For each setting, window-level mean-pooled features were standardized on the training split and used to train a logistic regression classifier with class-balanced loss. All probes were trained on the same 200K stratified subset and evaluated on the held-out 9-genome benchmark using the same region-level overlap criteria as the main evaluation.

Table 5. Auxiliary layer-probe results on the 9-genome benchmark. All settings use the same 200K training subset and identical region-level readout. F1 is computed from the reported precision and recall under the same overlap-based protocol used in the main text.

Feature	Precision	Recall	F1
L13	0.329	0.574	0.418
L20	0.342	0.567	0.427
L27	0.325	0.521	0.400
L13+L20+L27	0.172	0.600	0.267

The three single-layer probes achieve similar region-level performance. Layer 20 gives the highest F1 among the single-layer probes, although the differences are small. In contrast, concatenating layers 13, 20, and 27 increases feature dimensionality and recall, but substantially reduces precision, leading to a lower F1.

Based on this probe, we use layer 20 as a representative intermediate Evo2 layer for the main segmentation model. We do not claim that layer 20 is globally optimal for BGC segmentation; the probe only indicates that intermediate layers contain comparable BGC-relevant signal under a simple window-level readout, while multi-layer concatenation does not provide a favorable precision-recall trade-off.

F. Loss-Function Ablation

We evaluate four training objectives under a matched experimental protocol: the same frozen Evo2 backbone, the same 1D U-Net segmentation head, the same training data, global batch size 64, seed 0, AdamW with learning rate 1×10^{-3}, and early stopping over a maximum of 20 epochs. Only the loss function differs across runs.

- **Pure BCE** ($\alpha = 1.0$): treats every unannotated token as a full-weight negative.
- **Weak-negative BCE** ($\alpha = 0.5$): assigns unannotated background tokens half the weight of annotated positive tokens.

- **Dice + BCE**: combines a standard segmentation Dice loss with token-level BCE.
- **Tversky + BCE**: combines token-level BCE with Tversky loss, using $\alpha_T = 0.7$ and $\beta_T = 0.3$ to place greater penalty on false negatives.

All runs are evaluated on the held-out 9-genome benchmark using the same region-level overlap protocol as Table 1. Results are reported at two readout settings: a low-threshold discovery setting (threshold 0.50, minimum length 2 kb) and the conservative 9-genome setting used in the main benchmark comparison (threshold 0.70, minimum length 8 kb).

Table 6. Loss-function ablation on the held-out 9-genome benchmark. F1 is computed from precision and recall under the same overlap-based evaluator used in the main text. "n_{pred}" is the total number of predicted regions.

Loss	**thr 0.50, min 2 kb**			**thr 0.70, min 8 kb**		
	F1	**P / R**	n_{pred}	**F1**	**P / R**	n_{pred}
Pure BCE ($\alpha = 1.0$)	0.343	.244 / .574	542	0.498	.505 / .492	277
Weak-negative BCE ($\alpha = 0.5$)	0.419	.347 / .528	427	**0.521**	**.620** / .449	221
Dice + BCE	**0.440**	**.375** / .534	408	0.503	.564 / .453	243
Tversky + BCE	0.389	.297 / **.564**	479	0.511	.552 / **.475**	261

The ablation shows two trends. First, treating all unannotated tokens as full-weight negatives gives the weakest overall behavior: Pure BCE produces the largest number of predictions at the low-threshold setting and has the lowest F1 among the four objectives at both readout settings. This supports the use of a weak-negative objective, since unannotated genomic sequence may contain unknown or atypical BGCs.

Second, Dice + BCE gives the highest F1 at the low-threshold discovery setting, while weak-negative BCE gives the highest F1 and precision at the conservative 9-genome setting. We therefore use weak-negative BCE in the main experiments because it provides the best high-confidence region-level performance under the benchmark operating point, while still maintaining competitive recall at the discovery setting used for metagenomic candidate generation.