PLANTBGC: LABEL-SCARCE CROSS-DOMAIN REPRESENTATION LEARNING FOR PLANT BIOSYNTHETIC GENE CLUSTER DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Plant biosynthetic gene cluster (BGC) discovery poses a label-scarce cross-domain learning problem: abundant microbial supervision is available, whereas plant BGC annotations are limited. We present PlantBGC, which combines a cross-domain representation learning framework with a standardized benchmark and evaluation protocol for plant BGC discovery. PlantBGC learns BGC-relevant Pfam context from microbial, adapts the representation to unlabeled plant genomes through masked language modeling, and further calibrates predictions through weak supervision derived from GO/KEGG annotations. On 34 curated plant BGCs, self-supervised plant adaptation increases strict 100% recovery from 29.4% to 67.6%; at matched genomic coverage, Stage 2 retains 61.8% recovery at 0.53%, versus 29.4% at 0.52% for Stage 1. Weak calibration further reduces residual primary-metabolism-like proxy burden beyond hard rule-only filtering without reducing curated-BGC recovery. Under matched per-genome candidate budgets, PlantBGC retains a 4.0-fold higher strict recovery than *plantiSMASH* while producing more compact candidate intervals. PlantBGC provides both an effective transfer framework and a controlled benchmark for label-scarce plant BGC discovery.

1 INTRODUCTION

Natural products from bacteria, fungi, plants, and marine organisms remain important sources of medicines, agricultural compounds, and biotechnology products (Dias et al., 2012; Newman & Cragg, 2016). Many bioactive molecules are synthesized by enzymes encoded by physically co-localized gene sets, known as biosynthetic gene clusters (BGCs), which underlie diverse specialized metabolites ranging from microbial antibiotics such as penicillin and erythromycin (Reeves et al., 1999) to therapeutic compounds produced by plants, including vinblastine and vincristine from *Catharanthus roseus* and paclitaxel-related taxanes from *Taxus* spp. (Hwang et al., 2025; Xiong et al., 2021). As genomic resources rapidly expand, scalable computational strategies are needed to discover and prioritize BGCs for downstream biological validation (Bauman et al., 2021).

BGC discovery has relied on mutagenesis screening, chemistry-guided isolation, and heterologous expression, which are definitive but difficult to scale (Adrio & Demain, 2006; Skinnider et al., 2015; Ning et al., 2022). Genome-mining systems such as antiSMASH and ClusterFinder enable large-scale nomination using hidden Markov models and heuristic rules, but handcrafted features and local matching limit sensitivity to atypical clusters and cross-species generalization (Blin et al., 2023; Cimermancic et al., 2014; Bauman et al., 2021).Recent machine learning methods, including RFBGCpred (Madival et al., 2025), DeepBGC and e-DeepBGC (Hannigan et al., 2019; Liu et al., 2022), GECCO (Carroll et al., 2021), BiGCARP (Rios-Martinez et al., 2023), and Transformer-based platforms (Kawano et al., 2025), aim to learn BGC representations; however, most remain microbe-centric, complicating transfer to eukaryotic systems.

Plant BGC discovery is particularly challenging because plant genomes are larger and heterogeneous, with distinct Pfam-domain co-occurrence patterns. Only a few dozen plant BGCs have high-confidence, reference-mappable annotations (Zdouc et al., 2025). In contrast, microbial BGC annotations are much more abundant, but their genomic context differs substantially from that of plants, creating a natural source–target domain shift. Plant-oriented tools such as *plantiSMASH* (Kautsar

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2017; Del Pup et al., 2026), PlantClusterFinder (Schläpfer et al., 2017), and PhytoClust (Töpfer et al., 2017) incorporate biological priors but remain largely signature-based and do not learn transferable representations from labeled source genomes. The combination of severe target-label scarcity, source–target distribution shift, and inconsistent locus-level evaluation motivates both transferable representation learning and a standardized evaluation setting for plant BGC discovery.

We study plant BGC discovery as a label-scarce cross-domain sequence-to-locus learning problem. To this end, we propose PlantBGC, a three-stage cross-domain representation learning framework that first learns contextual BGC-likeness from curated microbial BGC annotations using an encoder-only Transformer, then performs label-free plant adaptation through masked language modeling on unlabeled plant Pfam sequences, and finally uses GO/KEGG-derived functional annotations(Ashburner et al., 2000; Kanehisa & Goto, 2000) as confidence-weighted weak supervision to calibrate primary-metabolism-like candidates. The evaluation uses coordinate-resolved curated plant BGCs, graded locus coverage, genomic screening-budget controls, leave-species-out testing, and matched-budget comparison with existing tools. Together, PlantBGC provides a benchmark for studying label-scarce cross-domain transfer in plant BGC discovery, while narrowing the genomic search space and potentially facilitating the prioritization of unrecognized BGCs for experimental characterization.

Our contributions are threefold: (i) we formulate plant BGC discovery as a label-scarce cross-domain representation learning problem and establish a standardized, coordinate-aware evaluation protocol with graded recovery, screening-budget controls, and leave-species-out testing; (ii) we develop a three-stage reference framework spanning microbial source learning, self-supervised plant-domain adaptation, and weak functional calibration, and use controlled experiments to study how each component contributes to target-domain transfer; and (iii) we evaluate practical prioritization through matched-budget comparison with *plantiSMASH*, boundary compactness, and independent co-expression and cross-species collinearity evidence.

2 METHOD

PlantBGC combines a cross-domain representation learning framework with a standardized benchmark and evaluation protocol for label-scarce plant BGC discovery. The learning framework transfers BGC-relevant representations from a label-rich microbial source domain to a distribution-shifted plant target domain with limited BGC annotations. Figure 1 summarizes the three-stage PlantBGC learning framework. The benchmark provides coordinate-resolved plant BGCs for evaluating this transfer under consistent locus-level criteria.

Specifically, microbial genomes and plant coding sequences(CDSs) are first converted into ordered Pfam-token sequences and mapped to Pfam2vec embeddings (Fig. 1A); dataset construction and preprocessing are detailed in Section 2.1 and Appendix A. Stage 1 learns a source-domain BGC representation and scoring function from microbial annotations using an encoder-only Transformer (Fig. 1B, red). Stage 2 then performs self-supervised target-domain adaptation on unlabeled plant Pfam sequences through masked language modeling, without using plant BGC labels (Fig. 1B, green). At inference, token scores are pooled to CDS-level scores and aggregated along genomic order into compact candidate loci. Stage 3 uses GO/KEGG-derived proxy evidence as confidence-weighted weak supervision to conservatively calibrate primary-metabolism-like candidates and attach functional cues for prioritization, without treating these proxies as ground-truth validation (Fig. 1C).

2.1 MICROBIAL AND PLANT DATASET PROCESSING

We construct source and target datasets from MIBiG (Medema et al., 2015; Zdouc et al., 2025) and NCBI GenBank/RefSeq as ordered Pfam-domain sequences. Stage 1 uses 1,420 validated microbial BGC loci mapped from MIBiG to GenBank. For negative samples, we follow with GeneSwap strategy of DeepBGC (Hannigan et al., 2019), construct negative segments by replacing each gene in a positive BGC segment with a background gene sampled from microbial reference genomes, matching genes with similar numbers of Pfam domains. Regions with high nucleotide similarity to known MIBiG BGCs are first excluded from the background genomes using BLAST (Altschul et al., 1990), reducing the likelihood that known biosynthetic loci are introduced into the negative set. Microbial genes are predicted with Prodigal (Hyatt et al., 2010), while plant use CDS annotations directly to avoid ORF-calling uncertainty in intron-rich genomes. Proteins are assigned Pfam tokens by hmmscan (Eddy, 2011; El-Gebali et al., 2019) and mapped to 100-dimensional Pfam2vec embeddings (Mikolov et al., 2013). For plants, we curate 34 known BGC loci across 17 species from MIBiG and use RefSeq CDS annotations as target-domain input. Genome annotations map CDSs to genomic coordinates, and

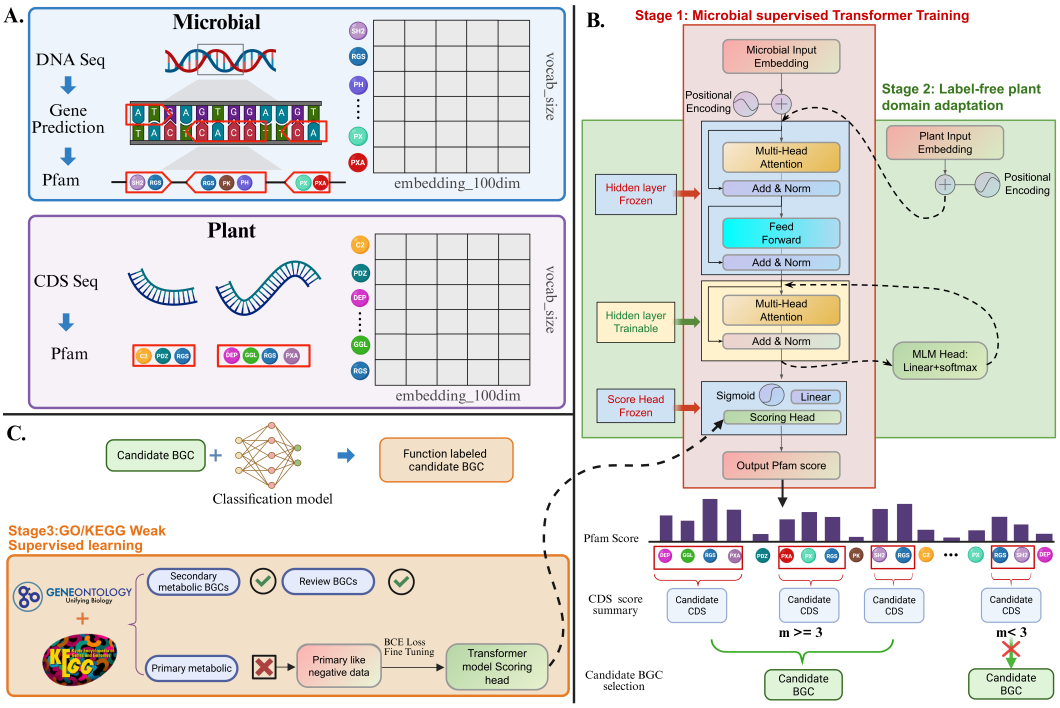


Figure 1: **Overview of PlantBGC learning framework.** (**A**) Genome sequences are encoded as ordered Pfam-domain tokens. (**B**) A source-domain Transformer learns BGC-likeness from microbial and adapted to unlabeled plant sequences via masked language modeling. (**C**) GO/KEGG-derived weak evidence calibrates plant predictions and supports compact candidate-locus prioritization.

plant loci are evaluated as base-pair intervals on matched reference assemblies. Additional dataset construction and preprocessing details are provided in Appendix A.

2.2 Transformer Architecture and Training Stages

Figure 1B illustrates the PlantBGC detector, an encoder-only Transformer operating on Pfam-token sequences and producing BGC-likeness scores at the token level. Transformer encoder blocks: Let $\mathbf{H}^{(0)} \in \mathbb{R}^{T\times d}$ denote the embedded token sequence of length T. The encoder stacks L standard Transformer blocks; for block $\ell \in \{1,\dots,L\}$, we apply multi-head self-attention (MHA) and a position-wise feed-forward network (FFN) with residual connections and layer normalization (LN):

$$\tilde{\mathbf{H}}^{(\ell)} = \mathrm{LN}\Big(\mathbf{H}^{(\ell-1)} + \mathrm{MHA}\Big(\mathbf{H}^{(\ell-1)}\Big)\Big), \qquad \mathbf{H}^{(\ell)} = \mathrm{LN}\Big(\tilde{\mathbf{H}}^{(\ell)} + \mathrm{FFN}\Big(\tilde{\mathbf{H}}^{(\ell)}\Big)\Big). \tag{1}$$

The final hidden state of token t is denoted by $\mathbf{h}_t \in \mathbb{R}^d$. Token-level scoring head: We predict a Pfam-token-level BGC-likeness score via a linear layer followed by a sigmoid:

$$s_t = \sigma(\mathbf{w}^\top \mathbf{h}_t + b), \tag{2}$$

where $s_t \in (0,1)$ is interpreted as the confidence that token t belongs to a BGC-like region. Token scores are later pooled to gene-level scores and aggregated into candidate BGC loci.

2.2.1 Stage 1: Microbial Supervised Training

In Stage 1, we train the encoder and scoring head on microbial BGC annotations to learn a strong, generalizable source-domain BGC scoring function. Because BGC annotations are typically available at locus-level interval resolution, we project locus boundaries to token labels: $y_t = 1$ if the gene hosting token t falls within any annotated BGC interval, and $y_t = 0$ otherwise. To mitigate class imbalance, we optimize weighted binary cross-entropy:

$$\mathcal{L}_{\mathrm{sup}} = -\sum_t \Big(\alpha\, y_t \log s_t + (1-y_t)\log(1-s_t)\Big), \tag{3}$$

where α upweights positive tokens based on training-set imbalance. This stage provides the transferable source representation used by Stage 2 adaptation and Stage 3 calibration.

2.2.2 STAGE 2: LABEL-FREE PLANT DOMAIN ADAPTATION VIA MLM

In Stage 2, plants lack sufficient high-confidence BGC labels for fully supervised training, so we adapt the microbial source representation to plant domain context using masked language modeling (MLM) on unlabeled plant Pfam-token sequences. A subset $\mathcal{M}$ of token positions is corrupted, and an auxiliary MLM head predicts the original Pfam identities:

$$\mathcal{L}_{\mathrm{MLM}} = -\sum_{t\in\mathcal{M}} \log p_\theta(d_t \mid \tilde{D}), \tag{4}$$

where $\tilde{D}$ is the corrupted (masked) token sequence and $p_\theta(\cdot)$ is produced by the MLM head. MLM encourages the encoder to model plant-specific Pfam co-occurrence and ordering patterns, aligning the microbial representation to the plant domain.

Let L denote the number of Transformer blocks in the encoder. We freeze the lower L_f blocks and update only the upper $L_u = L - L_f$ blocks during adaptation. In addition, we freeze the token-level scoring head and update the plant embedding table $\mathbf{E}^{(p)}$ together with the MLM head. This design preserves the BGC decision function learned under microbial supervision while allowing higher-level representations to align with plant-specific Pfam co-occurrence statistics captured by MLM. After adaptation, the MLM head is discarded, and Stage 3 addresses residual calibration mismatch through proxy-guided refinement.

2.3 CANDIDATE LOCUS AGGREGATION AND FUNCTIONAL LABELING

PlantBGC pools token-level BGC-likeness scores within each CDS and aggregates consecutive high-scoring CDS units into compact, gap-free candidate loci. Thus, compactness reflects contiguous high-scoring Pfam-domain context rather than post-hoc trimming. Coarse functional cues are assigned using multi-label random-forest classifiers trained on MIBiG annotations (Breiman, 2001); pooling, thresholding, and product-labeling details are provided in Appendix B.

2.4 GO/KEGG WEAK SUPERVISION

Stage 3 uses GO/KEGG-derived weak supervision to calibrate plant predictions by suppressing primary-metabolism-like loci while preserving secondary/mixed candidates for prioritization. Because comprehensive plant BGC ground truth is unavailable, GO/KEGG annotations are converted into conservative proxy labels; detailed annotation pipelines and proxy-label rules are provided in Appendix B.

GO and KEGG provide complementary primary-like evidence. For each candidate locus ℓ, the proxy-label pipeline produces primary-like confidence scores q_ℓ^{GO} and q_ℓ^{KEGG}. We fuse them with an agreement-boosted rule:

$$q_\ell = \min\!\left\{1,\ \max\!\left(q_\ell^{\mathrm{GO}}, q_\ell^{\mathrm{KEGG}}\right) + \gamma \cdot \min\!\left(q_\ell^{\mathrm{GO}}, q_\ell^{\mathrm{KEGG}}\right)\right\}, \tag{5}$$

where $\gamma \in [0,1]$ and the default is $\gamma = 0.5$. The outer clipping ensures fused confidence remains a valid soft-label confidence. Token-level soft targets within locus ℓ are assigned by $\tilde{y}_t = 1 - q_\ell$.

Starting from the Stage 2 checkpoint, we optimize a confidence-weighted soft-label BCE objective and combine it with the Stage 1 microbial supervised objective:

$$\mathcal{L}_{\mathrm{soft}} = -\sum_{\ell} w_\ell \sum_{t\in\mathcal{T}_\ell} \Big(\tilde{y}_t \log s_t + (1-\tilde{y}_t)\log(1-s_t)\Big), \mathcal{L}_{\mathrm{stage3}} = \mathcal{L}_{\mathrm{sup}} + \lambda\, \mathcal{L}_{\mathrm{soft}}. \tag{6}$$

During Stage 3, we freeze the lower encoder blocks and update the scoring head together with the upper encoder blocks, so primary-like loci receive downward gradients while uncertain cases are corrected conservatively.

3 EXPERIMENTS AND RESULTS

PlantBGC is a cross-domain representation learning framework and a standardized setting for plant BGC discovery. We ask whether representations learned from microbial supervision can be adapted to a distribution-shifted, label-scarce plant domain using unlabeled target data, while distinguishing

Table 1: **Stage 1 microbial supervised performance.**

Model	Leave-class-out					10-fold CV				
	AUC	Acc.	Prec.	Rec.	F1	AUC	Acc.	Prec.	Rec.	F1
Random Forest	0.852±0.03	0.862	0.720	0.565	0.632	0.923±0.02	0.924	0.786	0.583	0.669
DeepBGC-BiLSTM	0.945±0.02	0.910	0.838	0.721	0.775	0.980±0.02	0.981	0.933	0.900	**0.916**
BiGCARP	0.950±0.02	0.912	0.841	0.750	0.781	0.982±0.02	0.981	**0.934**	0.902	**0.916**
PlantBGC-Transformer	**0.979±0.01**	**0.973**	**0.906**	**0.852**	**0.878**	**0.988±0.02**	**0.981**	0.933	**0.904**	0.915

genuine adaptation from increased genomic nomination, species-specific exposure, or direct rule-based filtering. Specifically, **Q1** evaluates source-domain generalization, **Q2** tests self-supervised plant adaptation under controlled screening budgets, and **Q3** examines weak GO/KEGG calibration beyond hard rules. We further assess unseen-species transfer, matched-budget comparison with *plantiSMASH* (Kautsar et al., 2017), candidate compactness, and independent biological evidence.

Standardized plant evaluation protocol. The plant benchmark contains 34 coordinate-resolved curated BGCs across 17 species, which are reserved for evaluation rather than plant-side model training. For each known BGC, we select the predicted locus with maximum base-pair overlap and define recovery coverage as: $\mathrm{cov}(\mathrm{pred}, \mathrm{truth}) = \frac{|\mathrm{pred} \cap \mathrm{truth}|}{|\mathrm{truth}|}$. Recovery is reported under Any, $\geq 50\%$, $\geq 80\%$, and 100% coverage criteria, so that evaluation distinguishes partial localization from complete locus recovery. Unless otherwise stated, candidate loci are reconstructed with score threshold $\tau = 0.5$, minimum length $m = 3$, and gap tolerance 0. For budget-controlled analyses, genomic screening burden is measured as the fraction of reference-genome base pairs covered by the union of predicted candidate intervals.

3.1 Stage 1: Microbial supervised Transformer learning

Stage 1 addresses **Q1**: whether microbial supervision can provide a sufficiently generalizable source representation for later transfer to plants. No plant data are used at this stage, so the purpose is not to evaluate plant performance directly, but to establish that the source model captures BGC-relevant context beyond the classes on which it is trained.

We train Stage 1 with weighted BCE using Adam ($lr = 10^{-3}$), a maximum sequence length of 512, batch size 16, and early stopping; full settings are provided in Appendix H. We compare against three baselines: (i) **DeepBGC** (Hannigan et al., 2019): the BiLSTM uses 128 hidden units and a time-distributed sigmoid head, optimized with weighted BCE and Adam ($lr = 10^{-4}$) on fixed-length sequence chunks. (ii) **BiGCARP** (Rios-Martinez et al., 2023): we use its Pfam-sequence representation and BGC classification pipeline and evaluate it under the same microbial data splits as PlantBGC. (iii) **Random Forest**: we train a Random Forest classifier on hand-crafted Pfam features computed per training segment and window, using class balancing and hyperparameter tuning on the training folds. All methods are evaluated using the same data partitions. Our primary transfer-oriented test is leave-class-out (LCO) over major non-hybrid biosynthetic classes (PKS, NRPS, RiPP, and terpene), with model selection using only the remaining classes. We additionally report 10-fold cross-validation on paired MIBiG positives and GeneSwap negatives for in-domain source learning. All models use three random seeds with fixed data splits; ROC-AUC is reported as mean ± standard deviation, while accuracy, precision, recall, and F1 are three-seed means.

Table 1 shows the clearest difference under LCO. PlantBGC achieves an ROC-AUC of 0.979 ± 0.01 and F1 of 0.878, compared with 0.950 ± 0.02/0.781 for BiGCARP, 0.945 ± 0.02/0.775 for DeepBGC-BiLSTM, and 0.852 ± 0.03/0.632 for Random Forest. This indicates stronger generalization to biosynthetic classes excluded from training, the relevant prerequisite for subsequent cross-domain adaptation. Under 10-fold CV, PlantBGC also achieves the highest ROC-AUC (0.988 ± 0.02), while F1 remains comparable across the strongest deep baselines (0.915 for PlantBGC versus 0.916 for BiGCARP and DeepBGC). Thus, the advantage is more pronounced under the transfer-oriented LCO setting than under in-domain CV. Full ROC curves are provided in Appendix C and Fig. 6.

3.2 Stage 2: Self-Supervised Plant-Domain Adaptation

Stage 2 addresses the central transfer question (**Q2**): whether unlabeled plant sequences can adapt the microbial source representation and improve plant-BGC recovery without using plant BGC

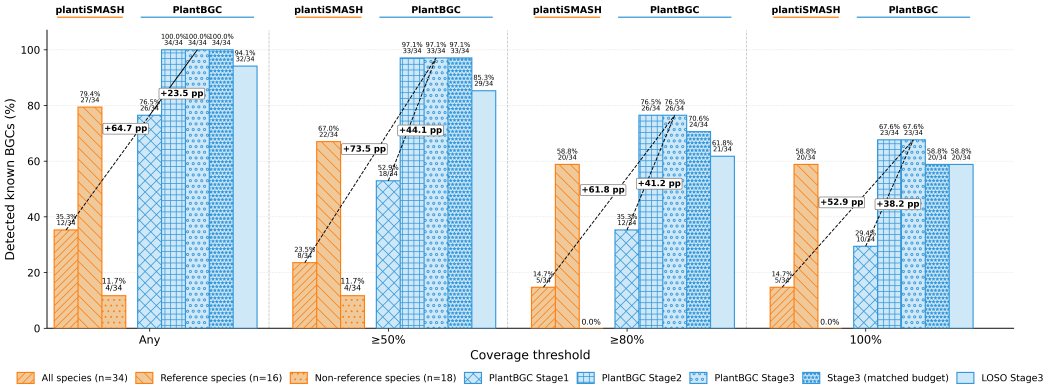


Figure 2: **Curated-BGC recovery across PlantBGC stages and evaluation controls.**

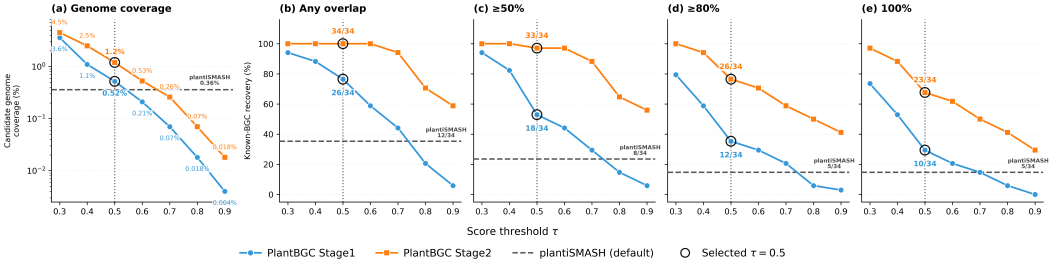


Figure 3: **Threshold τ sweep recovery.** (a) Candidate genome coverage. (b–e) Known-BGC recovery. The default $\tau = 0.5$ is circled, and dashed lines indicate *plantiSMASH* operating point.

labels. Plants differ substantially from microbial genomes in gene organization and Pfam-domain co-occurrence statistics, while only a few dozen coordinate-resolved plant BGCs are available for evaluation. We therefore compare the Stage 1 source model applied directly to plants with the Stage 2 model adapted through MLM on unlabeled plant Pfam sequences. The two models use the same BGC scoring head and candidate-locus reconstruction procedure. Thus, the comparison is designed to isolate the effect of self-supervised target-domain representation adaptation rather than changes in the downstream prediction rule.

As shown in Fig. 2, at the default $\tau = 0.5$, Stage 2 substantially improves curated-BGC recovery across all criteria: 26/34 to 34/34 under Any overlap, 18/34 to 33/34 at $\geq 50\%$, 12/34 to 26/34 at $\geq 80\%$, and 10/34 (29.4%) to 23/34 (67.6%) under strict 100% coverage. The larger gains at stricter thresholds indicate improved completeness rather than merely increased partial overlap. Additional Stage 2 diagnostics are provided in Appendix D. Across species, the adaptation gains are broadly distributed rather than driven by a small subset of genomes (Table 4; Figs. 7 and 8). Together, these results support **Q2**: self-supervised plant-domain adaptation substantially improves transfer of the microbial BGC representation without plant BGC supervision.

Recovery–budget analysis. A direct comparison at a fixed score threshold is not sufficient, however, because MLM adaptation may change the overall score distribution and therefore the amount of genomic sequence nominated. Higher recovery could then arise simply because Stage 2 screens more of the genome. To distinguish representation improvement from this screening-budget effect, we explicitly measure recovery as a function of nominated genomic space. We therefore sweep τ from 0.3 to 0.9 for Stage 1 and Stage 2, reconstruct loci using the same aggregation rule, and jointly measure genome coverage and recovery under the Any, $\geq 50\%$, $\geq 80\%$, and 100% criteria.

Figure 3 shows that Stage 2 shifts the recovery–budget trade-off favorably across the threshold sweep. Although Stage 2 often nominates more genomic sequence at the same τ, its advantage persists at comparable screening burdens. Stage 1 at $\tau = 0.5$ covers 0.52% of the genome and recovers 10/34 (29.4%) BGCs under strict 100% coverage, whereas Stage 2 at $\tau = 0.6$ covers a nearly identical 0.53% and recovers 21/34 (61.8%). Likewise, at approximately 0.07% genome coverage, strict recovery increases from 5/34 (14.7%) for Stage 1 to 14/34 (41.2%) for Stage 2. These matched-

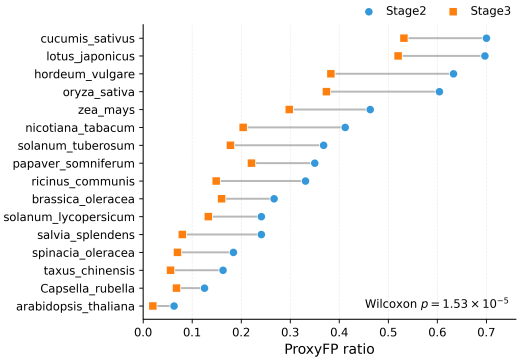


Figure 4: **Proxy Primary Ratio: Stage2 vs. Stage3**

Table 2: **Primary-like ratio and residual burden under GO/KEGG proxies.**

Method	Prim. ratio (r)	Prim. burden (b)
GO		
Stage 2	926/4243 (21.82%)	21.45%
Rule-only	0/3317	21.45%
Stage 3	223/2868 (7.75%)	6.12%
Stage 3 reduction	**64.5%**	**71.5%**
KEGG		
Stage 2	1370/3730 (36.73%)	12.63%
Rule-only	0/2360	12.63%
Stage 3	310/2037 (15.2%)	2.12%
Stage 3 reduction	**58.6%**	**83.2%**

coverage operating points show that the Stage 2 gain cannot be explained solely by nominating more genomic sequence. Instead, MLM adaptation improves the ability of the source-supervised representation to prioritize plant BGC regions under comparable genomic screening budgets.

3.3 Stage 3: GO/KEGG Weakly Supervised Calibration

Stage 3 addresses **Q3**: after target-domain representation adaptation, can weak functional evidence further calibrate plant predictions without sacrificing known-BGC recovery? Plant genomes contain gene-dense regions associated with primary metabolism that can still receive high BGC-likeness scores, while comprehensive plant BGC ground truth is unavailable for direct false-positive estimation. We therefore use GO (Ashburner et al., 2000) and KEGG (Kanehisa & Goto, 2000) as conservative proxies for primary-metabolism-like evidence rather than as biological ground truth. Stage-wise objectives and ablations are summarized in Appendix E and Table 5.

To quantify calibration, we use two complementary locus-level metrics. The *primary-like ratio* r measures the fraction of proxy-annotated candidate loci explicitly classified as primary-like. This metric alone is insufficient for evaluating learned calibration because a hard filtering rule can trivially make $r = 0$ by deleting every explicitly flagged locus. We therefore additionally define the *residual primary-like burden* b, which measures mixed or ambiguous primary-metabolism evidence among the non-primary candidates that remain:

$$r = \frac{N_{\text{primary}}}{N_{\text{proxy}}}, \qquad b = \frac{N_{\text{residual primary}}}{N_{\text{retained non-primary}}}, \qquad \text{Reduction}(x) = 100\frac{x_{\text{Stage2}} - x_{\text{Stage3}}}{x_{\text{Stage2}}},\ x \in \{r, b\}. \tag{7}$$

To determine whether Stage 3 simply reproduces the explicit GO/KEGG rules, we construct a *rule-only* control that directly removes all explicitly primary-like loci without additional training or model updates. By construction, this control drives r to zero, but it does not change the primary-like evidence carried by candidates that survive the hard rule. Comparison with Stage 3 therefore tests whether weak supervision produces learned calibration beyond literal rule execution.

Table 2 shows that Stage 3 reduces both explicit and residual primary-like signals, with the per-species trend broadly distributed across genomes (Fig. 4). Under GO, r decreases from 21.82% to 7.75% (64.5%) and b from 21.45% to 6.12% (71.5%); under KEGG, the corresponding reductions are 36.73% to 15.2% (58.6%) and 12.63% to 2.12% (83.2%). In contrast, rule-only filtering drives r to zero by construction but leaves b unchanged, showing that Stage 3 reduces residual primary-like evidence beyond literal rule execution.

Importantly, calibration does not reduce recovery of the curated BGCs: Stage 2 and Stage 3 remain identical across all four recovery criteria, including 23/34 (67.6%) under strict 100% coverage (Fig. 2). Because r and b are derived from GO/KEGG annotations, they should be interpreted as functional proxy measures rather than experimentally validated false positives. Fewer KEGG-supported loci are available than GO-supported loci, primarily because KO assignments are incomplete for a subset of RefSeq CDSs. Additional Stage 3 analyses are provided in Appendix F, including proxy-label changes and per-species filtering (Figs. 9–10; Table 6).

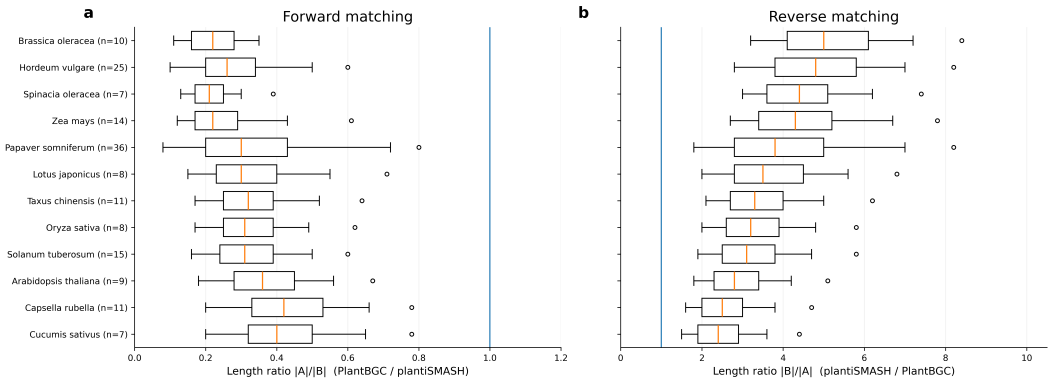


Figure 5: **Bidirectional compactness comparison with *plantiSMASH*.** Forward and reverse best matches compare boundary compactness within shared genomic neighborhoods.

Cross-species transfer. The standard Stage 2 and Stage 3 models are adapted using unlabeled plant data from multiple species. We therefore further test whether their performance depends on having observed the target species during plant-side training. In leave-one-species-out (LOSO) evaluation, each target species is excluded from both Stage 2 MLM adaptation and Stage 3 weak calibration before evaluation, and predictions are made only on the held-out species. Recovery decreases from 34/34 to 32/34 under Any overlap and from 23/34 (67.6%) to 20/34 (58.8%) under strict 100% coverage. Although some degradation is expected when the target species is completely unseen, most of the final-model recovery is retained. This result supports cross-species transfer of the plant-adapted representation rather than purely species-specific adaptation (Fig. 2).

3.4 Comparison with plantiSMASH

We compare PlantBGC with *plantiSMASH 2.0* (Del Pup et al., 2026) along two complementary axes: curated-BGC recovery under matched per-genome candidate budgets and boundary compactness within shared genomic neighborhoods.

Recovery of curated plant BGCs under matched budgets. Because the methods nominate different numbers of loci, for each genome we retain only the top-k PlantBGC candidates, where k is the number reported by *plantiSMASH*.

Under this controlled budget, PlantBGC recovers 34/34 (100%), 33/34 (97.1%), 24/34 (70.6%), and 20/34 (58.8%) loci at the Any, $\geq$50%, $\geq$80%, and 100% coverage thresholds, respectively. By comparison, *plantiSMASH* recovers 12/34 (35.3%), 8/34 (23.5%), 5/34 (14.7%), and 5/34 (14.7%). Notably, restricting PlantBGC to the same number of candidates reduces strict recovery only from 23/34 (67.6%) to 20/34 (58.8%), while remaining 4.0-fold higher than *plantiSMASH*. This indicates that the recovery advantage is not driven solely by a larger candidate set: high-confidence PlantBGC loci are themselves enriched for curated BGCs under an equal per-genome screening budget. Per-species results are provided in Appendix G.

Boundary compactness on shared genomic neighborhoods. The matched-budget analysis controls for the number of nominated loci, but does not address how much genomic sequence must be inspected within a nominated neighborhood. We perform a complementary matched-neighborhood analysis to test whether PlantBGC produces more compact intervals when both methods identify the same genomic region. We compare boundary compactness bidirectionally, with A_i denoting a PlantBGC candidate and B_j a *plantiSMASH* cluster. Best matches are defined by bp-IoU:

$$\mathrm{IoU}_{\mathrm{bp}}(A,B) = \frac{|A \cap B|}{|A \cup B|}, \qquad j^*(i) = \arg\max_j \mathrm{IoU}_{\mathrm{bp}}(A_i, B_j), \qquad i^*(j) = \arg\max_i \mathrm{IoU}_{\mathrm{bp}}(A_i, B_j).$$

We compute forward and reverse length ratios: $r_{\mathrm{fwd}} = \frac{|A_i|}{|B_{j^*(i)}|}, r_{\mathrm{rev}} = \frac{|B_j|}{|A_{i^*(j)}|}$.

Figure 5 shows that PlantBGC predictions are consistently more compact than *plantiSMASH* within shared genomic neighborhoods. The pooled forward median length ratio is 0.278, with 93.8% of matched PlantBGC intervals shorter than their *plantiSMASH* counterparts. Reverse matching yields

Table 3: **Collinearity and co-expression analysis.**

(a) Cross-species collinearity

Quantity	Observed	Reference
Candidate loci	2,033	9 genomes
Significant match	61.5%	5% nominal
Three-species clique	61.9%	2.6% null
Conserved groups (3-species)	507 (222)	1,236 loci
Median significant wJ	0.58	0.25 null

Representative conserved groups

Members	Mean wJ	Min wJ
Cs101/Sl114/Ps111	0.46	0.25
Os84/Zm82/Hv35	0.53	0.46
At105/Cr129/Bo282	0.47	0.21

(b) co-expression PlantBGC and PlantiSMASH

Quantity	PlantBGC	*plantiSMASH*
Testable loci	68/79	63/64
Individually significant	16/68 (25.0%)	14/63 (22.2%)
Local-null significant	15/68 (23.4%)	13/63 (20.6%)
Median z vs. size-matched null	0.440	0.642
Stratified van Elteren test	$p = 0.93$	

PlantBGC representative loci

Locus	$\bar{\mathbf{r}}$	$\mathbf{p}_{\mathrm{local}}$
BGC99	0.939	2.0×10^{-5}
BGC54	0.442	4.9×10^{-3}
BGC23	0.440	2.2×10^{-2}

the same overall pattern (median ratio = 3.92; 96.1% shorter PlantBGC intervals), indicating that the result is not an artifact of choosing only one matching direction.

3.5 Biological Evidence from Co-expression and Cross-species Collinearity

We evaluate whether PlantBGC-prioritized loci exhibit evolutionary and functional evidence independent of the model inputs (Table 3). PlantBGC inference uses only ordered Pfam-domain tokens; sequence homology, gene expression, and cross-species collinearity are not used for prediction.

Cross-species collinearity. We analyzed all 2,033 candidate loci across RefSeq genomes. Cross-species similarity was quantified by IDF-weighted Jaccard similarity (wJ) between Pfam-domain sets, and per-locus significance was assessed against a domain-shuffled null. Overall, 61.5% loci had a best cross-species match ($q < 0.05$; median $wJ = 0.58$ versus 0.25 under the null). Mutually similar loci formed 507 conserved groups, including 222 three-species groups; 61.9% of loci formed a three-species clique compared with only 2.6% under the null. Representative groups include a cucumber–tomato–opium poppy group containing the characterized cucurbitacin C cluster (MIBiG BGC0001315), as well as conserved groups in Poaceae and Brassicaceae (Table 3a).

Within-locus co-expression. We next assessed transcriptional coherence of *A. thaliana* candidates using ATTED-II v13 (Ath-r.c7-0) (Obayashi et al., 2022), comprising 19,851 RNA-seq runs. For each testable locus, mean pairwise Pearson correlation ($\bar{r}$) was evaluated against genome-wide random gene sets and a local null generated by 50,000 permutations within its ±25-gene neighborhood. Among the 79 PlantBGC loci, 68 were testable ($\geq$ 2 genes), of which 16/68 (23.5%) were individually significant and 15/68 (22.1%) remained significant against the local null (Table 3b).

As a control, we applied the identical pipeline to *plantiSMASH* loci. Of 64 loci, 63 were testable, with 14/63 (22.2%) individually significant and 13/63 (20.6%) significant against the local null. Because locus-size distributions differ substantially between the two methods, we use a size-stratified comparison rather than raw $\bar{r}$ values; the van Elteren test shows no detectable difference in per-locus co-expression support ($p = 0.93$). Thus, PlantBGC maintains transcriptional support comparable to *plantiSMASH* while producing more compact candidate intervals. Together with the independent collinearity analysis, these results provide complementary biological support for PlantBGC-prioritized loci and help narrow the candidate space for downstream validation.

4 Conclusion and Limitation

We presented PlantBGC, which combines a cross-domain representation learning framework with a standardized evaluation protocol for label-scarce plant BGC discovery. Self-supervised plant-domain adaptation increases strict recovery from 29.4% to 67.6%, with gains retained under controlled screening budgets and unseen-species evaluation. Weak GO/KEGG calibration further reduces primary-metabolism-like proxy burden without sacrificing curated-BGC recovery. Under matched candidate budgets, PlantBGC retains a 4.0-fold higher strict recovery than *plantiSMASH* while producing more compact candidate intervals. The evaluation remains limited by the small number of coordinate-resolved plant BGCs, annotation-version mismatches, and imperfect GO/KEGG proxy labels. PlantBGC may also miss atypical loci with weak Pfam-domain signatures, and predicted functions and boundaries require downstream biological validation.

AI USE STATEMENT

In this work, we used generative AI tools for (1) preliminary literature search and screening and (2) polish writing to improve the readability and (3) linguistic clarity of the manuscript. The authors independently reviewed the relevant literature and verified information against the original sources. AI-assisted language edits were reviewed and revised by the authors to ensure that they accurately reflected the intended scientific content.

Generative AI tools were not used to generate or modify experimental data, develop the proposed methodology, design or conduct experiments, or interpret experimental results. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

ETHICS STATEMENT

This study is based entirely on publicly available genomic, biosynthetic gene cluster, and bioinformatics resources. Plant genome and protein sequence data were obtained from the National Center for Biotechnology Information (NCBI), and experimentally characterized biosynthetic gene clusters were obtained from the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository. Co-expression data were obtained from the publicly available ATTED-II database, and cross-species collinearity analysis used the corresponding NCBI RefSeq genome annotations.

No human participants, personally identifiable information, clinical data, or newly collected biological samples were involved in this study. All analyses were computational and were conducted in accordance with the access and usage conditions of the respective public databases.

REPRODUCIBILITY STATEMENT

We provide the information and resources necessary to reproduce the experiments reported in this work. For the double-blind review stage, the Supplementary Materials include the complete source code, scripts, and Readme required to reproduce the experiments presented in the paper.

We have additionally developed a web-based interface for PlantBGC to make the method accessible to researchers, including users without extensive computational or bioinformatics expertise. To preserve author anonymity during double-blind review, the public website and the corresponding GitHub repository are not linked in the current manuscript. Upon acceptance, we will include links to both the PlantBGC website and the public GitHub repository in the main paper. The GitHub repository will provide the publicly accessible version of the code supplied with the current Supplementary Materials.

ACKNOWLEDGMENTS

We thank all collaborators and contributors for helpful discussions.

REFERENCES

Jose L. Adrio and Arnold L. Demain. Genetic improvement of processes yielding microbial products. *FEMS Microbiology Reviews*, 30(2):187–214, 2006. doi: 10.1111/j.1574-6976.2005.00009.x. URL https://doi.org/10.1111/j.1574-6976.2005.00009.x.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2. URL https://doi.org/10.1016/S0022-2836(05)80360-2.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556. URL https://doi.org/10.1038/75556.

Katherine D. Bauman, Keelie S. Butler, Bradley S. Moore, and Jonathan R. Chekan. Genome mining methods to discover bioactive natural products. *Natural Product Reports*, 38(11):2100–

2129, 2021. doi: 10.1039/d1np00025h. URL https://pubs.rsc.org/en/content/articlelanding/2021/np/d1np00025h.

Kai Blin, Simon Shaw, Hannah E. Augustijn, Zachary L. Reitz, Friederike Biermann, Mohammad Alanjary, Artem Fetter, Barbara R. Terlouw, William W. Metcalf, Eric J. N. Helfrich, Gilles P. van Wezel, Marnix H. Medema, and Tilmann Weber. 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. URL https://doi.org/10.1093/nar/gkad344.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Laura M. Carroll, Martin Larralde, Jonas S. Fleck, et al. Accurate de novo identification of biosynthetic gene clusters with gecco. *bioRxiv*, 2021. doi: 10.1101/2021.05.03.442509. URL https://doi.org/10.1101/2021.05.03.442509.

Peter Cimermancic, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, Paul A. Godfrey, Michael Koehrsen, Jon Clardy, et al. 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. URL https://www.cell.com/cell/fulltext/S0092-8674(14)00768-3.

Elena Del Pup, Charlotte Owen, Ziqiang Luo, Hannah E Augustijn, Arjan Draisma, Guy Polturak, Satria A Kautsar, Anne Osbourn, Justin JJ van der Hooft, and Marnix H Medema. plantismash 2.0: improvements to detection, annotation, and prioritization of plant biosynthetic gene clusters. *Journal of Molecular Biology*, pp. 169798, 2026.

Daniel A. Dias, Sylvia Urban, and Ute Roessner. A historical overview of natural products in drug discovery. *Metabolites*, 2(2):303–336, 2012. doi: 10.3390/metabo2020303. URL https://www.mdpi.com/2218-1989/2/2/303.

Sean R. Eddy. Accelerated profile hmm searches. *PLoS Computational Biology*, 7(10):e1002195, 2011. doi: 10.1371/journal.pcbi.1002195. URL https://doi.org/10.1371/journal.pcbi.1002195.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R. Eddy, et al. The pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, 2019. doi: 10.1093/nar/gky995. URL https://doi.org/10.1093/nar/gky995.

Gavin D. Hannigan, Daniel Prihoda, Adam Palicka, Jan Soukup, Ondrej Klempir, Leena Rampula, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, 47(18):e110, 2019. doi: 10.1093/nar/gkz654. URL https://doi.org/10.1093/nar/gkz654.

Jaewook Hwang, Jonathan Kirshner, Daniel A. R. Ramey Deschênes, Matthew B. Richardson, Steven J. Fleck, others, and Yang Qu. Ancient gene clusters govern the initiation of monoterpenoid indole alkaloid biosynthesis and c3 stereochemistry inversion. *Nature Communications*, 16:10495, 2025. doi: 10.1038/s41467-025-65543-z.

Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, 2010. doi: 10.1186/1471-2105-11-119. URL https://doi.org/10.1186/1471-2105-11-119.

Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27. URL https://doi.org/10.1093/nar/28.1.27.

Satria A. Kautsar, Hernando G. Suarez Duran, Kai Blin, Anne Osbourn, and Marnix H. Medema. plantismash: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, 45(W1):W55–W63, 2017. doi: 10.1093/nar/gkx305. URL https://doi.org/10.1093/nar/gkx305.

Tomoki Kawano, Taro Shiraishi, Tomohisa Kuzuyama, and Masayuki Umemura. A novel transformer-based platform for the prediction and design of biosynthetic gene clusters for (un)natural products. *bioRxiv*, 2025. doi: 10.1101/2025.06.02.657346. URL https://doi.org/10.1101/2025.06.02.657346.

Mingyang Liu, Yun Li, and Hongzhe Li. Deep learning to predict the biosynthetic gene clusters in bacterial genomes. *Journal of Molecular Biology*, 434(14):167463, 2022. doi: 10.1016/j.jmb.2022.167463. URL https://doi.org/10.1016/j.jmb.2022.167463.

Sharanbasappa D. Madival, Dwijesh Chandra Mishra, Krishna Kumar Chaturvedi, Neeraj Budhlakoti, Mohammad Samir Farooqi, Sudhir Srivastava, Anu Sharma, Shivadarshan S. Jirli, Alka Arora, Girish K. Jha, and Shesh N. Rai. Rfbgcpred: A random forest based tool for prediction of biosynthetic gene clusters. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1869(12):130859, 2025. doi: 10.1016/j.bbagen.2025.130859. URL https://doi.org/10.1016/j.bbagen.2025.130859.

Marnix H. Medema, Ralf Kottmann, Raphael Yilmaz, Markus Cummings, John B. Biggins, Kai Blin, et al. Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology*, 11:625–631, 2015. doi: 10.1038/nchembio.1890. URL https://www.nature.com/articles/nchembio.1890.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv*, 2013. URL https://arxiv.org/abs/1301.3781.

David J. Newman and Gordon M. Cragg. Natural products as sources of new drugs from 1981 to 2014. *Journal of Natural Products*, 79(3):629–661, 2016. doi: 10.1021/acs.jnatprod.5b01055. URL https://pubs.acs.org/doi/10.1021/acs.jnatprod.5b01055.

Yaodong Ning, Yao Xu, Binghua Jiao, and Xiaoling Lu. Application of gene knockout and heterologous expression strategy in fungal secondary metabolites biosynthesis. *Marine Drugs*, 20(11):705, 2022. doi: 10.3390/md20110705. URL https://doi.org/10.3390/md20110705.

Takeshi Obayashi, Hiroaki Hibara, Yusuke Kagaya, Yasunobu Aoki, and Kengo Kinoshita. Atted-ii v11: a plant gene coexpression database using a sample balancing technique by subagging of principal components. *Plant and Cell Physiology*, 63:869–881, 2022.

Andrew R. Reeves, R. Samuel English, J. S. Lampel, David A. Post, and Thomas J. Vanden Boom. Transcriptional organization of the erythromycin biosynthetic gene cluster of Saccharopolyspora erythraea. *Journal of Bacteriology*, 181(22):7098–7106, 1999. doi: 10.1128/JB.181.22.7098-7106.1999.

Carolina Rios-Martinez, Nitin Bhattacharya, A. P. Amini, Lucas Crawford, and Kevin K. Yang. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. *PLOS Computational Biology*, 19(5):e1011162, 2023. doi: 10.1371/journal.pcbi.1011162. URL https://doi.org/10.1371/journal.pcbi.1011162.

Pascal Schläpfer, Peifen Zhang, Chuan Wang, Taehyong Kim, Michael Banf, Lee Chae, Kate Dreher, Arvind K. Chavali, Ricardo Nilo-Poyanco, Thomas Bernard, Daniel Kahn, and Seung Y. Rhee. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, 173(4):2041–2059, 2017. doi: 10.1104/pp.16.01942. URL https://doi.org/10.1104/pp.16.01942.

Michael A. Skinnider, Chris A. Dejong, Philip N. Rees, Chad W. Johnston, Haoxin Li, Andrew L. H. Webster, Morgan A. Wyatt, and Nathan A. Magarvey. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Research*, 43(20):9645–9662, 2015. doi: 10.1093/nar/gkv1012. URL https://doi.org/10.1093/nar/gkv1012.

Nadine Töpfer, Lisa-Maria Fuchs, and Asaph Aharoni. The phytoclust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Research*, 45(12):7049–7063, 2017. doi: 10.1093/nar/gkx404. URL https://doi.org/10.1093/nar/gkx404.

Xingyao Xiong, Yan Li, Yanjun Zeng, et al. The Taxus genome provides insights into paclitaxel biosynthesis. *Nature Plants*, 7(8):1026–1036, 2021. doi: 10.1038/s41477-021-00963-5.

Mitja M. Zdouc, Kai Blin, Nico L. L. Louwen, Jorge Navarro, et al. Mibig 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Research*, 53(D1):D678–D690, 2025. doi: 10.1093/nar/gkae1115. URL https://doi.org/10.1093/nar/gkae1115.

A Data Construction and Preprocessing

Pfam-token representation. Along each chromosome or scaffold, let $C = \{c_1, \ldots, c_N\}$ denote the ordered list of CDS units. Each CDS c_i is represented by a Pfam-token sequence $D_i = (d_{i1}, \ldots, d_{iM_i})$. All Pfam tokens are mapped to vectors trained with the skip-gram objective (Mikolov et al., 2013). We use Pfam2vec embeddings as a compact continuous representation for discrete Pfam symbols. This design exposes domain similarity and co-occurrence patterns, avoids extremely high-dimensional one-hot vocabularies, and provides a fixed input dimension for the Transformer encoder. We set the embedding dimension to 100 as a lightweight trade-off between expressiveness and efficiency.

Microbial source-domain data. We build the microbial source-domain dataset from MIBiG (Zdouc et al., 2025) and NCBI GenBank/RefSeq. MIBiG reports 2437 active curated BGC entries in addition to pending and retired records. Because only a subset can be directly mapped to complete, coordinate-resolved genomic records in NCBI, we retain validated mapped loci, yielding 1420 microbial BGCs for Stage 1 supervised training.

Microbial negatives. We generate negative samples using GeneSwap (Hannigan et al., 2019), which preserves local gene-length and neighborhood structure while disrupting BGC-like signals. Replacement genes are sampled from genomes screened to avoid high nucleotide similarity to known MIBiG loci using BLAST-based filtering (Altschul et al., 1990). For each positive segment, we create three independently swapped variants, producing a negative set three times larger than the positives.

Plant target-domain data. From MIBiG, we curate 17 plant species as the target domain and collect 34 known plant BGC loci for locus-level evaluation. We use CDS sequences from NCBI RefSeq as the plant gene dataset. All predicted candidate BGC loci are represented and evaluated in the reference genome coordinate system using genome annotations.

Coordinate normalization. All predicted loci and curated plant BGCs are represented as base-pair intervals on the same reference assembly for each genome, so that coverage and IoU are measured consistently. External tool outputs, including *plantiSMASH*, are normalized to the same assembly version before overlap computation, ensuring that cross-method comparisons reflect boundary behavior rather than annotation-version artifacts.

B Candidate Locus Aggregation and Weak Proxy Construction

Token-to-CDS pooling. Let $C = \{c_1, \ldots, c_N\}$ denote the ordered list of CDS units on a chromosome or scaffold. Each CDS c_i contains M_i Pfam tokens with token scores $\{s_{ij}\}_{j=1}^{M_i}$. We summarize token scores within each CDS to obtain a unit-level score:

$$S_i = \sum_{j=1}^{M_i} w_{ij}\, s_{ij}, \qquad \sum_{j=1}^{M_i} w_{ij} = 1, \tag{8}$$

where w_{ij} are normalized pooling weights. We use uniform pooling by default, i.e., $w_{ij} = 1/M_i$.

Candidate locus construction. We select high-scoring CDS units by $S_i > \tau$ with default threshold $\tau = 0.5$. Candidate loci are defined as maximal strictly consecutive runs of selected CDS units with gap tolerance set to 0. To suppress spurious short runs, we require a minimum run length of $m = 3$ consecutive CDS units. Each locus interval is defined from the start coordinate of its first CDS to the end coordinate of its last CDS in the genome annotation.

Multi-label biosynthetic product classifier. Beyond detection, we provide coarse-grained functional labels for candidate BGC loci by predicting biosynthetic product classes and bioactivities. We use random-forest models (Breiman, 2001), which are lightweight and robust for sparse Pfam-derived features and limited labeled examples in MIBiG v1.3. We train two separate models using MIBiG annotations: one for biosynthetic product classes and one for product activities. Each BGC is represented by a binary multi-label vector indicating the presence of zero or more classes and activities.

After training, these classifiers are applied to plant candidate BGCs to obtain predicted functional profiles.

GO functional annotation pipeline. We map Pfam domains to GO terms using the official `pfam2go` file and take the union over domains within each CDS. Let $\mathcal{T}_{\text{sec}}^{\text{GO}}$ and $\mathcal{T}_{\text{prim}}^{\text{GO}}$ denote GO term sets for secondary and primary metabolism. For each candidate locus ℓ with genes $\mathcal{G}_\ell$, we compute:

$$\begin{aligned} \text{sec_score}_\ell^{\text{GO}} &= \sum_{g \in \mathcal{G}_\ell} \mathbb{I}\big[\text{GO}(g) \cap \mathcal{T}_{\text{sec}}^{\text{GO}} \neq \emptyset\big], \\ \text{pri_score}_\ell^{\text{GO}} &= \sum_{g \in \mathcal{G}_\ell} \mathbb{I}\big[\text{GO}(g) \cap \mathcal{T}_{\text{prim}}^{\text{GO}} \neq \emptyset\big]. \end{aligned} \tag{9}$$

GO proxy labels are assigned as:

$$\begin{aligned} \texttt{secondary_likely}: &\ \text{sec_score}_\ell^{\text{GO}} \geq 2 \wedge \text{pri_score}_\ell^{\text{GO}} = 0, \\ \texttt{secondary_tilt}: &\ \text{sec_score}_\ell^{\text{GO}} = 1 \wedge \text{pri_score}_\ell^{\text{GO}} = 0, \\ \texttt{primary_likely}: &\ \text{pri_score}_\ell^{\text{GO}} \geq 2 \wedge \text{sec_score}_\ell^{\text{GO}} = 0, \\ \texttt{primary_tilt}: &\ \text{pri_score}_\ell^{\text{GO}} = 1 \wedge \text{sec_score}_\ell^{\text{GO}} = 0, \\ \texttt{mixed_possible}: &\ \text{sec_score}_\ell^{\text{GO}} > 0 \wedge \text{pri_score}_\ell^{\text{GO}} > 0, \\ \texttt{review}: &\ \text{otherwise}. \end{aligned} \tag{10}$$

When ℓ is labeled `primary_likely`, we additionally flag it as `unlikely_bgc=True`. For the confidence scores used in Eq. 5, `primary_likely` and `primary_tilt` are mapped to 0.8 and 0.5, respectively.

KEGG functional annotation pipeline. We assign KO identifiers to genes using BlastKOALA[1] or an equivalent workflow such as KofamScan, and treat genes without KO as `review`. Let $\mathcal{K}_{\text{sec}}^{\text{KO}}$ and $\mathcal{K}_{\text{prim}}^{\text{KO}}$ denote KO sets for specialized and primary metabolism. For each candidate locus ℓ, we compute:

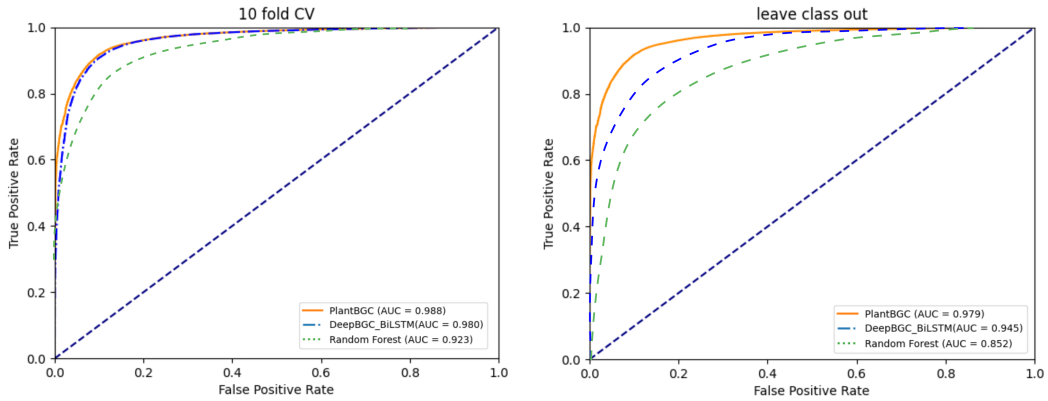
$$\begin{aligned} \text{kegg_sec}_\ell &= \sum_{g \in \mathcal{G}_\ell} \mathbb{I}\big[\text{KO}(g) \in \mathcal{K}_{\text{sec}}^{\text{KO}}\big], \\ \text{kegg_pri}_\ell &= \sum_{g \in \mathcal{G}_\ell} \mathbb{I}\big[\text{KO}(g) \in \mathcal{K}_{\text{prim}}^{\text{KO}}\big]. \end{aligned} \tag{11}$$

KEGG proxy labels are assigned as `secondary` if $\text{kegg_sec}_\ell > 0$ and $\text{kegg_pri}_\ell = 0$, `primary` if $\text{kegg_pri}_\ell > 0$ and $\text{kegg_sec}_\ell = 0$, `mixed_possible` if both are > 0, and `review` otherwise. For Eq. 5, KEGG primary-like labels are mapped to the corresponding primary-like confidence scores, while non-primary and uncertain labels are treated conservatively.

C Additional Stage 1 Microbial Benchmark Results

We provide full ROC curves for Stage 1 microbial supervised learning to complement the summary metrics in Table 1. The 10-fold cross-validation protocol evaluates in-domain microbial BGC detection, while the leave-class-out protocol evaluates whether the learned source-domain scoring function generalizes to unseen biosynthetic classes.

[1] BlastKOALA: `https://www.kegg.jp/blastkoala/`



(a) 10 fold CV ROC Curve

(b) leave class out ROC Curve

Figure 6: **Stage 1 ROC curves under two microbial evaluation protocols.** The curves compare PlantBGC, DeepBGC-BiLSTM, and Random Forest for token-level BGC detection on microbial data. (**a**) In 10-fold cross-validation, models are trained on nine folds of MIBiG-derived positive segments and GeneSwap-generated negatives and evaluated on the held-out fold; PlantBGC achieves AUC 0.988, compared with 0.980 for DeepBGC-BiLSTM and 0.923 for Random Forest. (**b**) In leave-class-out evaluation, all positive segments from one biosynthetic class are excluded during training and used for testing; PlantBGC achieves AUC 0.979, compared with 0.945 for DeepBGC-BiLSTM and 0.852 for Random Forest. The larger margin in the leave-class-out setting suggests that Stage 1 learns source-domain Pfam-context representations that generalize beyond within-class interpolation.

The leave-class-out protocol is particularly relevant for cross-domain transfer because it stresses whether the microbial source model captures transferable BGC context rather than only memorizing class-specific patterns. This supports the use of Stage 1 as the source-domain representation backbone for subsequent plant adaptation.

D Additional Stage 2 Plant Adaptation Analyses

This appendix provides additional diagnostics for Stage 2 label-free plant adaptation. While the main text reports aggregate recovery improvements over 34 curated plant BGCs, we further examine (i) changes in candidate confidence after adaptation and (ii) whether recovery gains are broadly distributed across plant species. These analyses complement the pooled results by testing whether Stage 2 improvements reflect systematic plant-domain adaptation rather than isolated high-scoring candidates or gains concentrated in a small subset of genomes.

D.1 Per-species Known-BGC Recovery

Figure 7 shows that the Stage 2 gains are broadly distributed across species. Stage 1 recovery declines sharply for many genomes as the coverage requirement becomes more stringent, whereas Stage 2 frequently maintains or improves recovery at the 50%, 80%, and 100% thresholds. Several species already recovered well by Stage 1 remain stable after adaptation, while others show substantial gains. Together, the per-species results support the pooled analysis in the main text and indicate that the benefit of label-free plant adaptation is not driven by only a few genomes.

D.2 Per-species candidate confidence statistics

Table 4 reports candidate-locus confidence statistics before and after Stage 2 adaptation. For each species, *All Candidate Mean* is the mean confidence score over all aggregated candidate BGC loci. *Stage 1 Mean* and *Stage 1 Max* summarize the mean and maximum locus scores among loci retained from the Stage 1 source-domain model, while *Stage 2 Mean* and *Stage 2 Max* report the corresponding statistics after label-free plant adaptation. Scores are model-assigned BGC-likeness

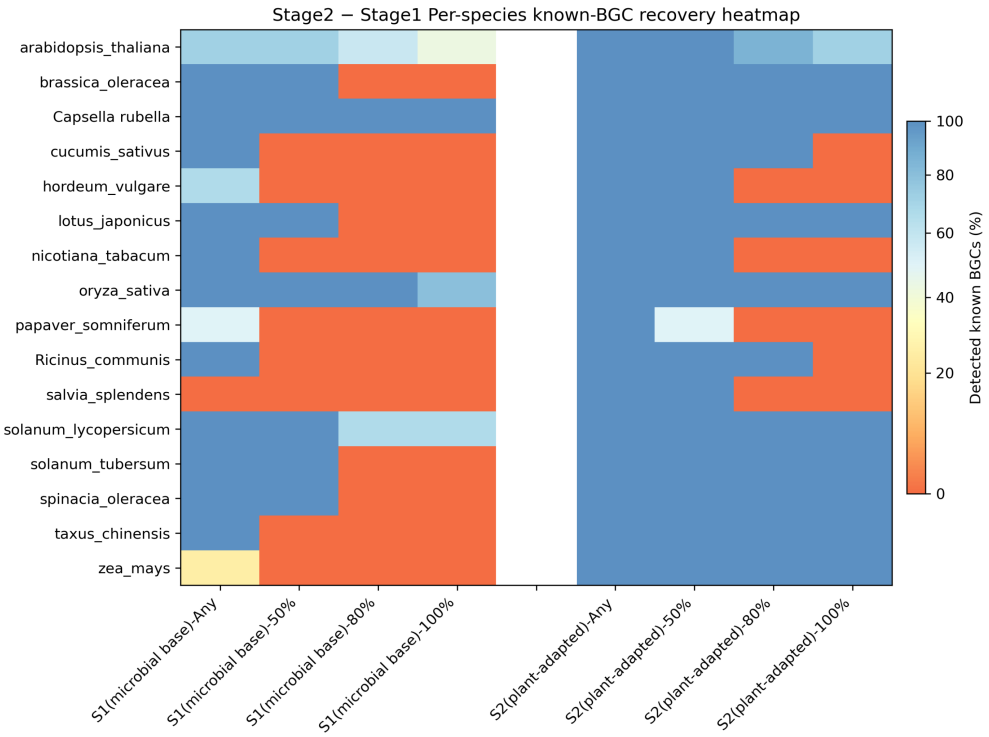


Figure 7: **Stage 2 adaptation improves known-BGC recovery across plant species.** Per-species recovery of curated plant BGCs before (Stage 1 microbial base model) and after (Stage 2 plant-adapted model) label-free adaptation, evaluated at any-overlap and 50%, 80%, and 100% coverage thresholds. Stage 2 maintains higher recovery across increasingly stringent coverage criteria for most evaluated species, indicating that the aggregate improvements reported in the main text are broadly distributed across genomes rather than driven by a small subset of species.

values on candidate loci produced by the same locus-aggregation procedure: strictly consecutive CDSs with score > 0.5, minimum length 3, and gap tolerance 0.

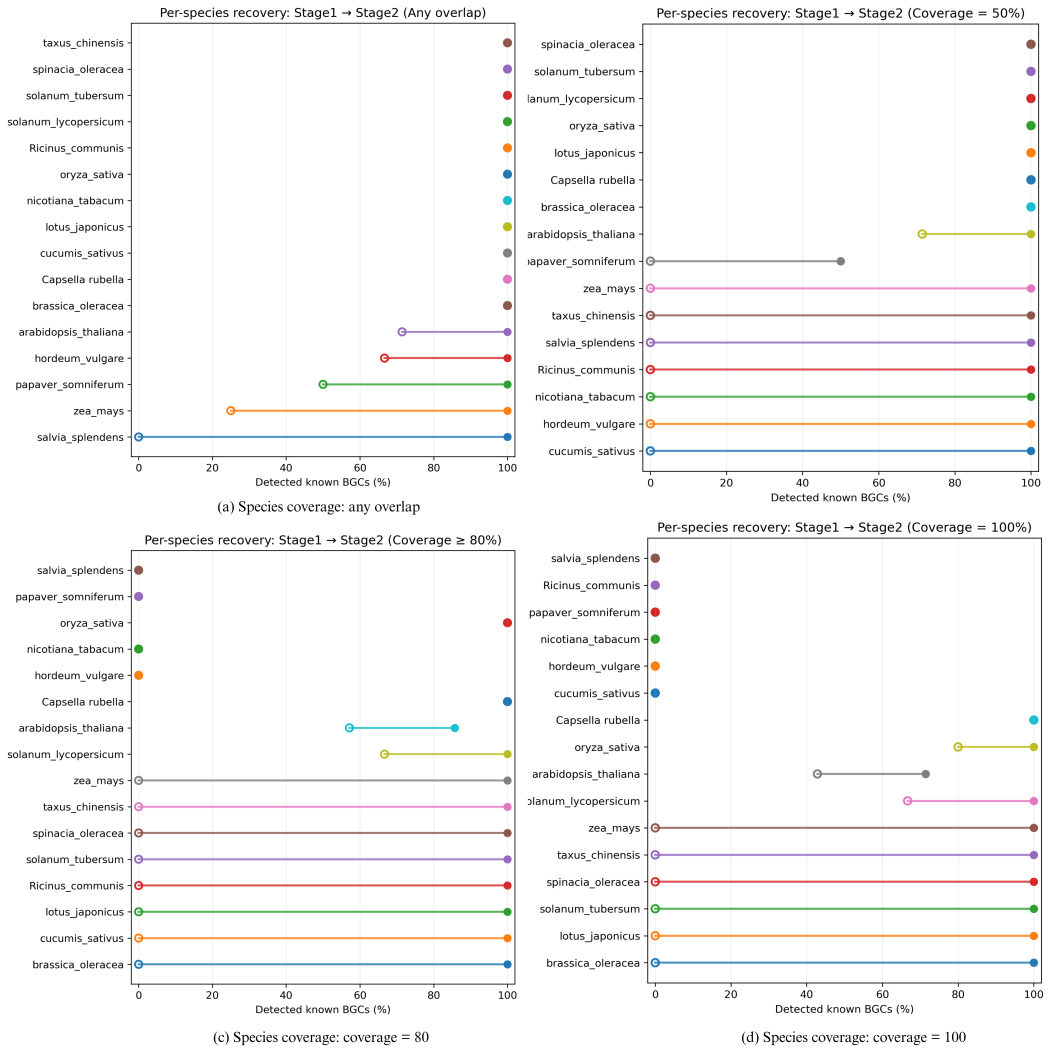
Table 4: **Stage 2 candidate confidence score statistics.** Stage 2 generally increases both mean and maximum candidate-locus confidence across plant species, suggesting stronger plant-domain scoring after label-free adaptation.

Species	All Candidate Mean	Stage 1 Mean	Stage 1 Max	Stage 2 Mean	Stage 2 Max
Arabidopsis thaliana	0.658	0.756	0.921	0.789	0.944
Brassica oleracea	0.695	0.763	0.931	0.795	0.952
Capsella rubella	0.582	0.632	0.852	0.692	0.855
Cucumis sativus	0.795	0.823	0.896	0.866	0.964
Hordeum vulgare	0.652	0.752	0.852	0.796	0.894
Lotus japonicus	0.648	0.763	0.985	0.795	0.984
Nicotiana tabacum	0.726	0.800	0.936	0.822	0.982
Oryza sativa	0.625	0.815	0.989	0.841	0.990
Papaver somniferum	0.682	0.760	0.853	0.775	0.862
Ricinus communis	0.702	0.730	0.859	0.802	0.900
Salvia splendens	0.723	0.763	0.862	0.795	0.902
Solanum lycopersicum	0.621	0.689	0.933	0.756	0.956
Solanum tuberosum	0.596	0.690	0.863	0.750	0.892
Spinacia oleracea	0.620	0.694	0.875	0.785	0.952
Taxus chinensis	0.625	0.702	0.896	0.752	0.952
Zea mays	0.596	0.659	0.932	0.795	0.962
Mean	0.653	0.733	0.904	0.788	0.932

Table 4 shows that Stage 2 increases the average retained-locus confidence from 0.733 to 0.788 across species. The average maximum locus score also increases from 0.904 to 0.932, indicating that adaptation modestly strengthens the highest-confidence predictions as well. Several species exhibit especially large gains in mean confidence, including *Zea mays* (0.659→0.795), *Spinacia oleracea* (0.694→0.785), *Solanum lycopersicum* (0.689→0.756), and *Hordeum vulgare* (0.752→0.796). These trends suggest that Stage 2 does not merely alter candidate counts or rely on a few extreme maxima; instead, it improves the confidence of retained candidate loci across a broad range of plant genomes. This supports the interpretation that masked language modeling adapts the source-domain representation toward plant-specific Pfam co-occurrence structure.

D.3 Per-species recovery under overlap and coverage thresholds

Figure 8 reports per-species known-BGC recovery before and after Stage 2 adaptation. For each species with curated known BGCs, recovery is computed under four criteria. The permissive criterion considers a known BGC recovered if at least one predicted locus overlaps its genomic span. The stricter criteria require the predicted locus to cover at least 50%, at least 80%, or 100% of the curated known-BGC interval. These stricter thresholds are important because they evaluate boundary completeness rather than simple neighborhood overlap, which matters for downstream biological review and experimental validation.



(a) Species coverage: any overlap

(c) Species coverage = 80

(d) Species coverage = 100

Figure 8: **Per-species known-BGC recovery under overlap and coverage thresholds.** For each plant species with curated known BGCs, recovery is computed for Stage 1 and Stage 2 candidate loci under four criteria: any overlap, at least 50% coverage, at least 80% coverage, and full 100% coverage of the known BGC interval. Hollow markers denote Stage 1 recovery and filled markers denote Stage 2 recovery; each horizontal segment shows the Stage 1→Stage 2 change for the same species and threshold. The stricter thresholds evaluate whether Stage 2 improves boundary completeness rather than merely producing loose overlaps near known BGCs.

Figure 8 shows that Stage 2 improves or preserves known-BGC recovery across multiple species and across increasingly strict recovery criteria. Under the permissive any-overlap criterion, several species move to full recovery after Stage 2, indicating that label-free adaptation helps the model retain biologically validated loci. The more important pattern appears under stricter coverage thresholds: Stage 2 often maintains or improves recovery when the predicted locus must cover at least 50%, 80%, or 100% of the known BGC span. This indicates that the improvement is not limited to detecting nearby regions, but also reflects better boundary-level recovery.

The per-species view is also useful because it shows that the aggregate Stage 2 improvement is not dominated by a single genome. Several species exhibit large Stage 1→Stage 2 gains, while many others preserve already high recovery after adaptation. This broad distribution supports the main conclusion of Stage 2: label-free MLM adaptation mitigates microbial-to-plant domain shift and improves plant-domain candidate recovery without using plant BGC labels.

E Ablation Analysis

Table 5: **Incremental contribution of each stage.**

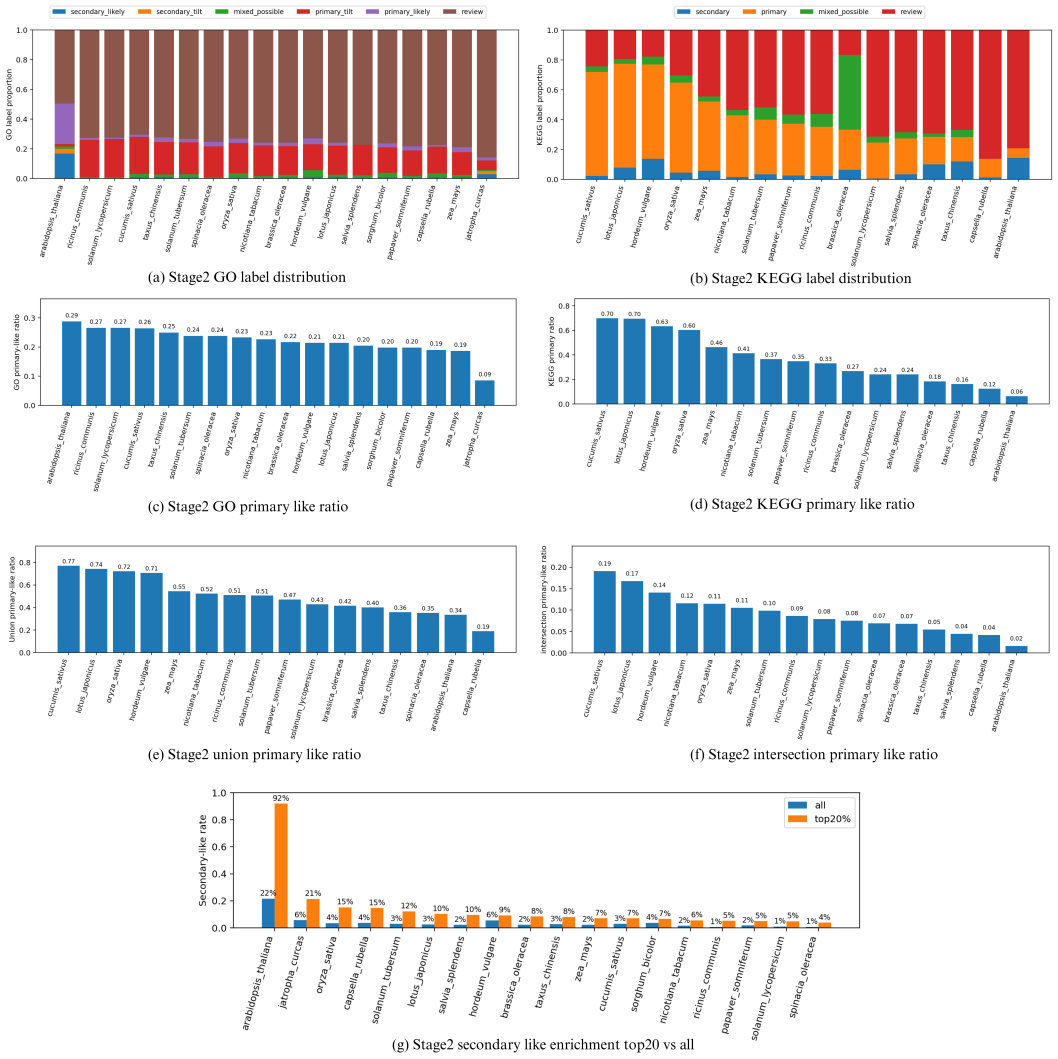
Stage	Primary objective	Key result
Stage 1	Learn a strong, generalizable source-domain BGC scoring function	Leave-class-out ROC-AUC improves from 0.945 to 0.979 and F1 from 0.775 to 0.878 over DeepBGC, supporting transfer-relevant generalization.
Stage 2	Mitigate microbial-to-plant domain shift without plant labels	Known-BGC recovery improves from 76.5% to 100% under any overlap and from 29.4% to 67.6% under 100% coverage.
Stage 3	Reduce ProxyFP while preserving ranking utility	Median proxy-primary ratio decreases from 0.35 to 0.16 across species (Wilcoxon $p = 1.53 \times 10^{-5}$).

F Additional Comparison Stage 3 Weak-Supervision and ProxyFP Analyses

This appendix provides additional diagnostics for Stage 3 GO/KEGG-guided weak supervision. The main text reports aggregate ProxyFP reduction, while the analyses here examine how GO- and KEGG-derived proxy labels are distributed across species before and after Stage 3 refinement, whether primary-like candidates are reduced consistently, and whether secondary/mixed evidence remains concentrated among top-ranked candidates. Together, these diagnostics clarify whether Stage 3 behaves as targeted calibration rather than indiscriminate candidate removal.

F.1 Proxy-label composition before weak calibration

Figure 9 summarizes GO/KEGG proxy-label diagnostics for Stage 2 candidate loci before weak-supervision refinement. Panels (a) and (b) report per-species proxy-label distributions under GO and KEGG mappings, respectively. The GO labels include `secondary_likely`, `secondary_tilt`, `mixed_possible`, `primary_tilt`, `primary_likely`, and `review`, while the KEGG labels use the manuscript-level grouping of `secondary`, `primary`, `mixed_possible`, and `review`. Panels (c) and (d) quantify the fraction of Stage 2 candidates flagged as primary-like by GO and KEGG, showing that primary-metabolism-like predictions remain non-negligible after label-free plant adaptation. Panels (e) and (f) compare union and intersection primary-like estimates: the union rule captures candidates flagged by either GO or KEGG and therefore gives a more conservative estimate of primary-like burden, while the intersection rule captures candidates supported by both resources and provides a stricter agreement-based estimate. Panel (g) compares secondary-like fractions among all candidates and the top-ranked 20% candidates, testing whether the confidence ranking already enriches specialized-metabolism evidence before Stage 3.



(a) Stage2 GO label distribution

(b) Stage2 KEGG label distribution

(c) Stage2 GO primary like ratio

(d) Stage2 KEGG primary like ratio

(e) Stage2 union primary like ratio

(f) Stage2 intersection primary like ratio

(g) Stage2 secondary like enrichment top20 vs all

Figure 9: **Stage 2 GO/KEGG proxy-label diagnostics before weak calibration.** The figure reports per-species proxy-label distributions, primary-like ratios, GO–KEGG union/intersection summaries, and secondary-like enrichment among top-ranked candidates.

The Stage 2 diagnostics show that label-free adaptation improves plant-domain recovery but does not fully remove primary-metabolism-like candidates. This motivates Stage 3: rather than increasing known-BGC recovery directly, Stage 3 targets residual calibration error by using GO/KEGG-derived primary-like evidence as weak negative feedback. The top-ranked enrichment panel is important because it verifies whether the ranking remains biologically useful: if high-confidence candidates retain secondary/mixed enrichment, then weak calibration can reduce screening burden without destroying prioritization utility.

F.2 Proxy-label composition after weak calibration

Figure 10 applies the same diagnostics to Stage 3 candidate loci after weak-supervision refinement. The repeated panel structure allows a direct before–after comparison with Fig. 9: panels (a)–(b) show GO/KEGG proxy-label distributions, panels (c)–(d) report primary-like ratios, panels (e)–(f) compare union/intersection primary-like estimates, and panel (g) evaluates secondary-like enrichment among the top-ranked 20% candidates. This design makes the refinement behavior auditable: Stage 3 should reduce primary-like burden while preserving or improving enrichment of secondary/mixed evidence near the top of the ranked list.



(a) Stage3 GO label distribution

(b) Stage3 KEGG label distribution

(c) Stage3 GO primary like ratio

(d) Stage3 KEGG primary like ratio

(e) Stage3 union primary like ratio

(f) Stage3 intersection primary like ratio

(g) Stage3 secondary like enrichment top20 vs all

Figure 10: **Stage 3 GO/KEGG proxy-label diagnostics after weak calibration.** The same diagnostics as Fig. 9 are computed after Stage 3 refinement to evaluate primary-like reduction and preservation of secondary/mixed enrichment.

Comparing Figs. 9 and 10 shows that Stage 3 behaves as a targeted refinement step. The primary-like candidate burden decreases after weak calibration, while the top-ranked candidates continue to show secondary/mixed enrichment. This distinction matters because reducing candidates alone would not be useful if the procedure also removed biologically plausible specialized-metabolism loci. Instead, the before–after diagnostics support the intended behavior of Stage 3.

F.3 Per-species Stage 3 filtering summary

Table 6 reports the per-species candidate counts before and after Stage 3 filtering. For each species, **Orig. total** is the number of aggregated candidate loci before Stage 3 adjustment, corresponding to the Stage 2 candidate set. **Kept total** is the number of loci retained after applying Stage 3 scaling factors, and **Removed** is the difference between the two. The scaling factors $\textbf{Scale}_{\text{prim}}$, $\textbf{Scale}_{\text{mixed}}$, and $\textbf{Scale}_{\text{review}}$ denote the multiplicative factors used to downsample primary-like, mixed-possible, and review loci.

Table 6 provides an auditable, per-genome view of the remaining candidate set after Stage 3. This is useful for interpreting practical screening workload: the method does not only report aggregate

Table 6: **Species-level candidate reduction after Stage 3.**

Species	Orig.	Kept	Removed	% Red.	$\text{Scale}_{\text{prim}}$	$\text{Scale}_{\text{mixed}}$	$\text{Scale}_{\text{review}}$
Arabidopsis thaliana	125	79	46	36.8	0.21	0.21	0.75
Brassica oleracea	295	184	111	37.6	0.21	0.22	0.75
Capsella rubella	168	112	56	33.3	0.36	0.20	0.75
Cucumis sativus	136	89	47	34.6	0.35	0.19	0.77
Hordeum vulgare	270	161	109	40.4	0.27	0.20	0.71
Lotus japonicus	191	128	63	33.0	0.36	0.20	0.76
Nicotiana tabacum	450	270	180	40.0	0.25	0.18	0.71
Oryza sativa	227	141	86	37.9	0.28	0.24	0.74
Papaver somniferum	399	256	143	35.8	0.37	0.24	0.72
Ricinus communis	139	91	48	34.5	0.27	0.21	0.80
Salvia splendens	361	222	139	38.5	0.20	0.23	0.73
Sorghum bicolor	373	244	129	34.6	0.30	0.21	0.76
Solanum lycopersicum	203	128	75	36.9	0.35	0.23	0.74
Solanum tuberosum	243	157	86	35.4	0.27	0.22	0.78
Taxus chinensis	184	114	70	38.0	0.24	0.22	0.75
Zea mays	209	148	61	29.2	0.38	0.24	0.79

Table 7: **Per-species known-BGC recovery of PlantBGC and *plantiSMASH*.** Recovery is reported under any overlap and increasingly stringent coverage thresholds. Species represented in the original *plantiSMASH* study are marked with †.

	Any overlap		≥50%		≥80%		100%	
Species	**pSMASH**	**PlantBGC**	**pSMASH**	**PlantBGC**	**pSMASH**	**PlantBGC**	**pSMASH**	**PlantBGC**
Arabidopsis thaliana†	80%	100%	71%	100%	62%	∼80%	∼62%	∼70%
Brassica oleracea	10%	100%	5%	100%	1%	100%	1%	100%
Capsella rubella	12%	100%	6%	100%	1%	100%	1%	100%
Cucumis sativus	7%	100%	8%	100%	1%	100%	1%	0%
Hordeum vulgare	8%	100%	8%	100%	0%	0%	0%	0%
Lotus japonicus	11%	100%	5%	100%	2%	100%	2%	100%
Nicotiana tabacum	5%	100%	6%	100%	0%	0%	0%	0%
Oryza sativa†	82%	100%	72%	100%	65%	100%	65%	100%
Papaver somniferum	10%	100%	6%	60%	0%	0%	0%	0%
Ricinus communis	6%	100%	5%	100%	2%	100%	2%	0%
Salvia splendens	6%	100%	2%	100%	0%	0%	0%	0%
Solanum lycopersicum†	85%	100%	63%	100%	54%	100%	54%	100%
Solanum tuberosum	10%	100%	6%	100%	1%	100%	1%	100%
Spinacia oleracea	11%	100%	5%	100%	0%	100%	0%	100%
Taxus chinensis	13%	100%	2%	100%	0%	100%	0%	100%
Zea mays†	79%	100%	63%	100%	62%	100%	60%	100%

ProxyFP reduction, but also shows how many candidate loci remain for each plant genome after weak-evidence filtering. The retained counts vary across species, reflecting differences in genome size, annotation content, and proxy-label composition, but the procedure consistently reduces the candidate set relative to Stage 2.

G Per-Species Comparison with *plantiSMASH*

This appendix provides the per-species results underlying the aggregate PlantBGC–*plantiSMASH* comparison in Section 3.4. We report known-BGC recovery under increasingly stringent coverage criteria and distinguish the four species represented in the original *plantiSMASH* study from the remaining benchmark species.

Table 7 reveals a pronounced species-dependent pattern. *plantiSMASH* achieves substantially higher recovery on the four species represented in its original study, but recovery drops sharply for most previously unrepresented species and approaches zero under the stricter 80–100% coverage criteria. PlantBGC, in contrast, retains known-BGC recovery across a broader range of species, including many not represented in the original *plantiSMASH* study. These results complement the pooled

benchmark in Fig. 2 and are consistent with improved cross-species generalization of the learned Pfam-context representation.

H Reproducibility, Limitations, and Responsible Use

Model architecture and training configuration. For reproducibility, Table 8 summarizes the model architecture and optimization settings used throughout the three training stages of PlantBGC. Unless otherwise specified, the same model configuration was used across all evaluated plant species. Data preprocessing, benchmark construction, and proxy-label generation are described in Appendices A–B.

Table 8: **Model architecture and training hyperparameters of PlantBGC.** Stage 1 denotes microbial supervised training, Stage 2 denotes plant-domain adaptation, and Stage 3 denotes weakly supervised refinement.

Component	Hyperparameter	Value
Model architecture		
	Pfam embedding dimension	100
	Transformer layers (L)	4
	Hidden dimension (d)	128
	Attention heads	4
	Feed-forward dimension	512
	Dropout	0.1
	Maximum sequence length	512
	Total parameters	$\sim$1M
Stage 1: microbial supervised training		
	Training objective	Token-level BCE
	Optimizer	Adam
	Learning rate	1×10^{-3}
	Batch size	16
	Early stopping	Yes 0.01
	Maximum epochs / patience	50 / 20
Stage 2: plant-domain adaptation		
	Training objective	Masked language modeling
	Masking probability	0.15
	Frozen lower blocks (L_f)	2
	Trainable upper blocks (L_u)	2
	Optimizer	Adam
	Learning rate	1×10^{-4}
	Batch size	16
	Training epochs	20
Stage 3: weakly supervised refinement		
	Soft-negative weight (γ)	0.5
	Loss weight (λ)	0.5
	Optimizer	Adam
	Learning rate	1×10^{-4}
	Batch size	16
	Training epochs	20

Limitations. PlantBGC is designed for computational prioritization rather than experimental validation. The current evaluation is limited by the small number of coordinate-resolved curated plant BGCs, possible annotation-version mismatches across genomic resources, and the fact that GO/KEGG-derived proxy labels are imperfect surrogates for true false positives. The method may miss atypical BGCs with weak Pfam-domain evidence or produce candidates whose biological function requires downstream experimental confirmation.

Compute. All experiments were run on NVIDIA RTX 4080 GPU, with 16GB memory. Stage 1 microbial supervised training required approximately 7 hours per run, Stage 2 plant-domain adaptation

required approximately 2 hours, and Stage 3 weak-supervision calibration required approximately 2.5 hours. Hyperparameters, data splits, and preprocessing details are provided in the main paper and Appendices A–B.

Data, code, and licenses. The study uses public genomic and functional annotation resources, including MIBiG, NCBI GenBank/RefSeq, Pfam, GO, KEGG, Prodigal, HMMER/hmmscan, GeneSwap/DeepBGC, and plantiSMASH, which are cited in the paper. We will release anonymized reproduction scripts and processed metadata where permitted by the original resource terms; raw data should be obtained from the original sources. Any released code and metadata will include documentation, dependency specifications, and license information.

Broader impact and responsible use. PlantBGC may support natural-product discovery by prioritizing plant genomic regions that may encode specialized metabolites with biomedical, agricultural, or biotechnological relevance. The main risk is overinterpretation: computationally predicted BGC candidates should not be treated as experimentally verified pathways. The method does not provide wet-lab protocols, synthesis instructions, or pathogenic engineering guidance, and candidate predictions should be interpreted as prioritization evidence requiring expert review and experimental validation.

Code availability. We provide an anonymized supplementary code package for review, including scripts for data preprocessing, model training, candidate aggregation, GO/KEGG proxy construction, and evaluation. The implementation will be de-anonymized and released as a public repository upon publication.