PLOS Computational Biology

RESEARCH ARTICLE

A novel transformer-based platform for the prediction and design of biosynthetic gene clusters for (un)natural products

Tomoki Kawano[1], Taro Shiraishi[1,2], Tomohisa Kuzuyama[1,2], Maiko Umemura[3]*

1 Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan,
2 Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Tokyo, Japan,
3 Department of Applied Biology, Kyoto Institute of Technology, Kyoto, Japan

* umemura-m@kit.ac.jp

Citation: Kawano T, Shiraishi T, Kuzuyama T, Umemura M (2026) A novel transformer-based platform for the prediction and design of biosynthetic gene clusters for (un)natural products. PLoS Comput Biol 22(2): e1013181. https://doi.org/10.1371/journal.pcbi.1013181

Editor: Boyang Ji, BioInnovation Institute, DENMARK

Received: May 29, 2025

Accepted: February 9, 2026

Published: February 23, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1013181

Abstract

Biosynthetic gene clusters (BGCs), comprising sets of functionally related genes responsible for synthesizing complex natural products, are a rich source of bioactive compounds with pharmaceutical potential. Here, we present a transformer-based framework that models functional domains as linguistic units to capture and predict their positional relationships within genomes. Using a RoBERTa architecture, we trained models on four progressively broader datasets: bacterial BGCs, Actinomycetes genomes, bacterial genomes, and bacterial plus fungal genomes. Evaluation using 2,492 experimentally-validated BGCs from the MIBiG database showed that more than 50% of true domains were ranked first and over 75% within the top 10 candidates. Our models also achieved classification accuracies exceeding 70% for major compound classes including polyketides (PKs) and terpenes. To explore model-guided BGC design, we compared predictions from the BGC-trained and genome-trained models using the BGC for the bacterial diterpenoid cyclooctatin as a case study. The genome-trained model uniquely predicted several domains absent from both the original BGC and the prediction by the BGC-trained model. Heterologous expression of one of those predicted domains in *Streptomyces albus*, together with the biosynthetic genes for cyclooctatin, yielded an unknown cyclooctatin derivative. This framework not only provides a novel BGC prediction method using machine learning but also facilitates rational design of artificial BGCs. Future integration of transcriptomic, protein structural, and phylogenetic data will enhance the models' predictive and generative capabilities, supporting accelerated discovery and engineering of natural products.

Data availability statement: All processed genome datasets, model construction codes, statistical analysis scripts, trained models, and other materials are available on Zenodo (https://doi.org/10.5281/zenodo.17577731). The source codes and small accompanying data files are also available on GitHub (https://github.com/umemura-m/bgc-transformer/).

Funding: This work was supported by Grant-in-Aid for Transformative Research Areas (22H05119 to TK, 23H04566 and 25H01599 to MU) and Grant-in-Aid for Challenging Research (Pioneering) (23K18120 to MU) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

BGCs encode diverse natural products, including antibiotics and anticancer agents. Identifying and designing BGCs in microbial genomes is crucial for discovering new bioactive compounds. In this study, we developed a transformer-based deep learning model that treats protein domains as language-like tokens and learns how they are arranged in genomes. By training on both known BGCs and whole genomes, the model successfully predicts biologically plausible combinations of domains, including those absent in known BGCs. We experimentally validated one such prediction by expressing a newly identified gene alongside known cyclooctatin biosynthetic genes, confirming the production of an unknown cyclooctatin derivative. Our results demonstrate how language models can uncover hidden biosynthetic potential and offer a promising new AI tool for natural product discovery and synthetic biology.

Introduction

Natural products have provided many important drugs such as penicillin, cyclosporine, tacrolimus, and paclitaxel, making them invaluable pharmaceutical resources. These compounds, primarily produced by microorganisms and plants, possess complex chemical structures that are often challenging to synthesize using conventional methods. The genes responsible for producing these compounds typically co-localize in biosynthetic gene clusters (BGCs), representing genomic regions where functionally related genes cooperate to produce specific natural products.

The organization of BGCs reflects their evolutionary history. In prokaryotes, multiple genes in a BGC often form operons regulated by one or a few promoters [1]. In contrast, eukaryotes lack operon structures, but BGC gene expression is frequently coordinated by transcription factors encoded within the cluster [2,3]. Recent experimental work by Kanai et al. [4] showed that DNA rearrangement via transposases facilitates operon formation, providing insights into BGC evolution. These evolutionary events leave genomic footprints as differences in the gene organization within cluster structures, as seen in cyanobactin BGCs across *Spirulina* species (Fig 1). Cyanobactins are cyclic peptides produced by cyanobacteria, some of which exhibit antitumor activity [5]. In the figure, the top BGC contains a transposase gene and four endonuclease genes, whereas the bottom three lack these elements. The second and third BGCs retain endonucleases but lack transposases. Additionally, gene orientation shifts from random to uniform directionality. These patterns imply that the evolutionary history of BGC formation is embedded in their genomic context, analogous to a recording device.

Recent advances in genome sequencing and computational methods have significantly improved BGC identification [6]. Current rule-based tools like antiSMASH [7] use hidden Markov models to detect BGCs through functional genes specific to natural compound biosynthesis, but they largely depend on known domain patterns

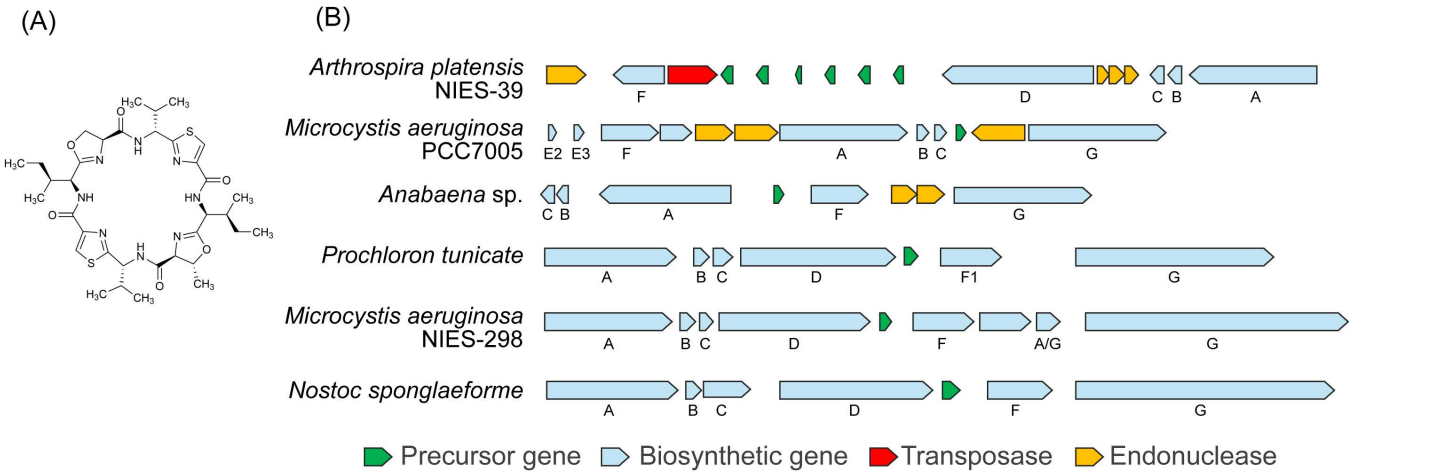

Fig 1. Gene organization in cyanobactin biosynthetic gene clusters (BGCs) from cyanobacterial strains. (A) Structure of patellamide A, a representative cyanobactin produced by *Prochloron didemni*. **(B)** Comparison of cyanobactin BGCs from six different cyanobacterial strains, showing variation in gene content and arrangement.

https://doi.org/10.1371/journal.pcbi.1013181.g001

and are limited in their ability to discover novel BGC architectures. Machine-learning-based tools, including ClusterFinder [8], DeepBGC [9], TOUCAN [10], and BGC-Prophet [11], have been developed to address these limitations, learning BGC features directly from domain organization patterns and enabling broader generalization beyond rule-based tools, although they still rely on pre-defined BGC and non-BGC labels for supervised learning.

Advances in artificial intelligence, particularly in natural language processing, provide new opportunities for analyzing genomic sequences. Transformer-based models that are superior at learning long-range dependencies in text have proven useful in biology, especially in protein structure prediction and functional annotation [12–14]. Large-scale models such as Evo have demonstrated the ability to predict and generate genomic sequences at megabase scales [15]. These examples suggest that transformer models may be well-suited for capturing the positional and contextual relationships among functional domains in genomes.

In this study, we introduce a transformer-based framework for predicting and designing BGCs by learning the sequential organization of functional domains as a proxy for evolutionary constraints. Functional domains are treated as language-like tokens, and their arrangements are modeled using natural language processing techniques. A distinguishing feature of our approach is that it learns from genomic sequences in an unsupervised manner, without predefined labels of BGC or non-BGC regions, allowing the model to capture contextual domain relationships across entire genomes. The present study focuses on enabling the discovery of novel domain combinations, while demonstrating a representative case of a variant of an existing pathway. Our models, currently trained solely on static representations of functional domain sequences in genomes, could be extended with multimodal information related to evolutionary direction, such as transcriptomic data and compound productivity, thereby enabling the design of functionally optimized BGCs for specific biosynthetic objectives.

Results

Pretraining on Genomes Using Functional Domain-Based Tokenization

We adopted functional domains within genes as token units for modeling genomic sequences, based on their established importance in BGC characterization as utilized in bioinformatics tools such as antiSMASH. We employed the RoBERTa architecture [16] for this purpose, configured with 8 hidden layers and 16 attention heads, and a maximum token length of

512 to cover the context of entire BGC domain sequences (Fig 2). We chose the BERT-based architecture due to its bidirectional attention mechanism, which captures context in both directions from each token [17,18]. As gene order in a BGC is often not essential for compound biosynthesis, GPT-style models with unidirectional attention are less suitable for BGC modeling. As shown in Fig A in S1 Data, both the RoBERTa and GPT-1 models demonstrated effective convergence on Dataset II (described in the following paragraph). However, the GPT-1 model exhibited earlier saturation of the validation

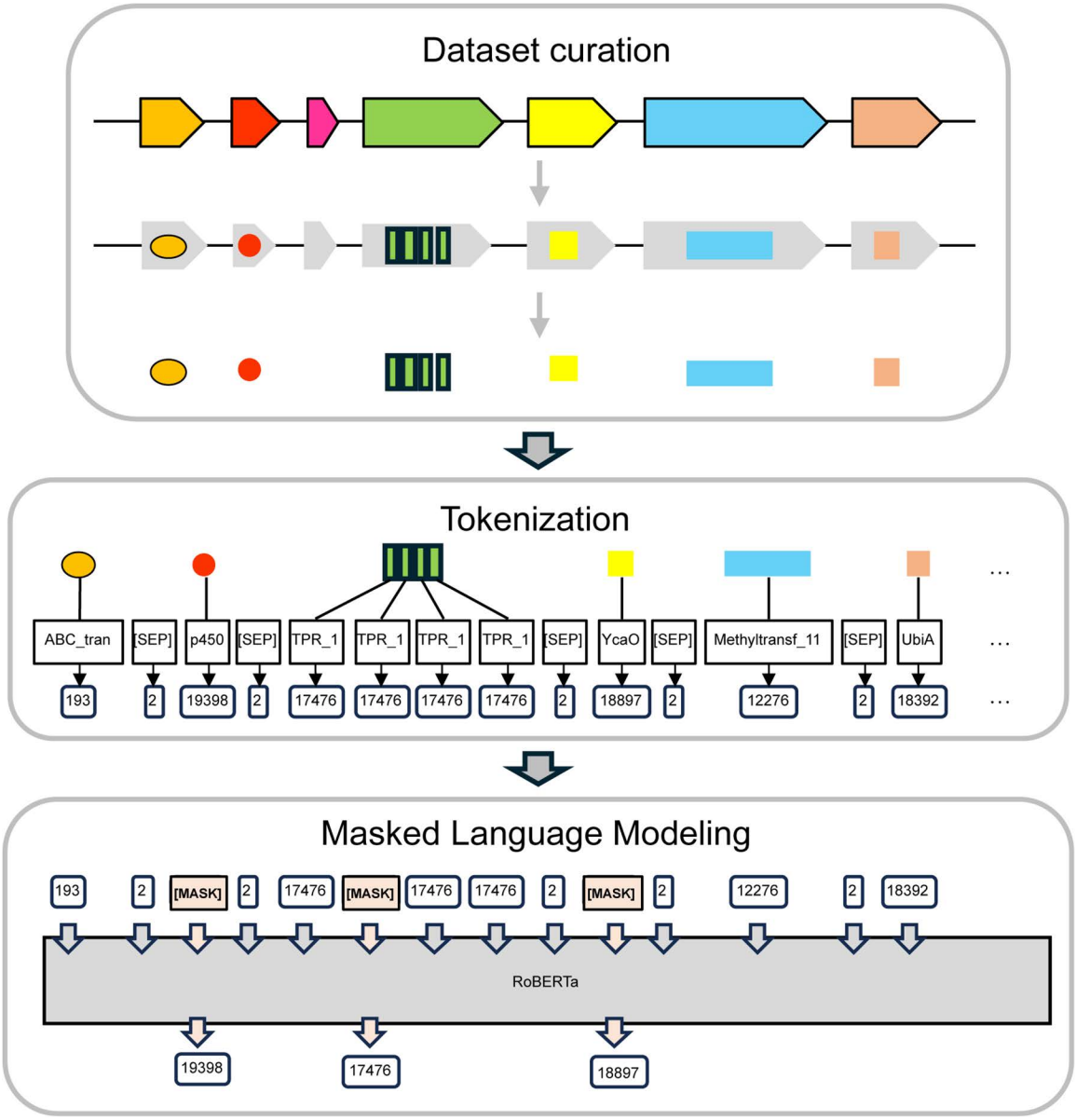

Fig 2. Workflow of masked language modeling of genomic data using RoBERTa. The workflow illustrates the process of tokenizing genomic data and training the RoBERTa language model. In the data curation phase, genomic sequences are retrieved from NCBI and analyzed using HMMer to identify Pfam domains. The tokenization phase converts all identified Pfam domains into discrete tokens. During the masked language modeling phase, these tokenized sequences serve as input for RoBERTa model training, where the model learns to predict masked tokens based on their surrounding context information.

https://doi.org/10.1371/journal.pcbi.1013181.g002

loss and failed to produce meaningful predictions for domain tokens in BGC sequences. These results indicate that the bidirectional attention mechanism in RoBERTa offers an advantage in learning the genomic arrangement of functional domains.

To investigate the impact of training data on model performance, we prepared four datasets of increasing taxonomic coverage: (I) bacterial BGCs from the antiSMASH database; (II) complete genomes of Actinomycetes; (III) complete bacterial genomes; and (IV) complete bacterial plus fungal genomes (Table 1). Actinomycetes and fungi represent two major taxa responsible for the production of diverse natural compounds. The bacterial dataset also included other lineages, such as cyanobacteria, which are likewise known to produce a wide range of natural products. This stepwise expansion (Fig 3) allowed us to explore how different genomic contexts influence model generalization while maintaining the domain-based tokenization approach consistent.

Training on all four datasets resulted in a consistent decrease in both training and validation losses, with the validation loss exceeding the training loss in all cases (Fig 4). This behavior indicates that the model successfully learned the genomic arrangement of functional domains across datasets. The BGC-only model (Dataset I) exhibited the most rapid convergence, likely reflecting its narrower functional diversity and more homogeneous domain contexts. In contrast, the genome-based models (Datasets II, III, and IV) converged more slowly, suggesting that the broader diversity of domain combinations and the presence of non-BGC regions in genome-wide data increase the complexity of the learning process and influence convergence behavior during training.

Model Performance in BGC Classification

To evaluate the classification performance of the constructed models across different BGC types, we conducted a compound class prediction task using class labels from the antiSMASH database (Fig 5). Our model achieved more than 70% classification accuracy for 37 of 75 compound classes, with highest performance observed in well-characterized classes

Table 1. Datasets used for training transformer-based models.

#	Organism	#Data	#Domain tokens[1]
I	Bacterial BGC[2]	239,021 BGCs	8,597,242
II	Actinomycetes genome	2,664 strains	13,750,809
III	Bacterial genome	9,748 strains	48,941,605
IV	Bacterial plus fungal genome	11,884 strains	70,453,078

[1]Separated into 80% training and 20% test datasets.

[2]Predicted by antiSMASH.

https://doi.org/10.1371/journal.pcbi.1013181.t001

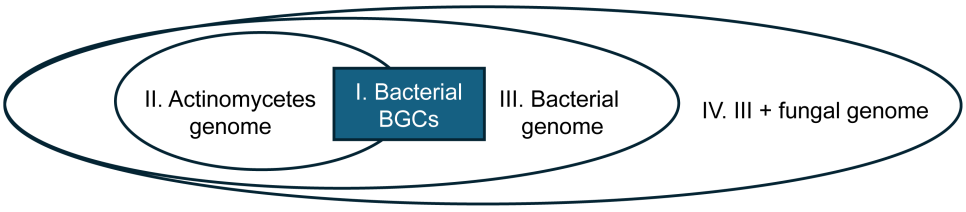

Fig 3. The relationships among four datasets used in this study. The outermost dataset (IV) comprises bacterial and fungal genomes (70.5M tokens from 11,884 strains), encompassing the bacterial genome dataset (III; 48.9M tokens, 9,748 strains) and the Actinomycetes dataset (II; 13.8M tokens, 2,664 strains). The BGC dataset from antiSMASH database (I; 8.6M tokens from 239,021 BGCs) overlaps with Datasets II, III, and IV, reflecting its specialized nature in secondary metabolite biosynthesis.

https://doi.org/10.1371/journal.pcbi.1013181.g003

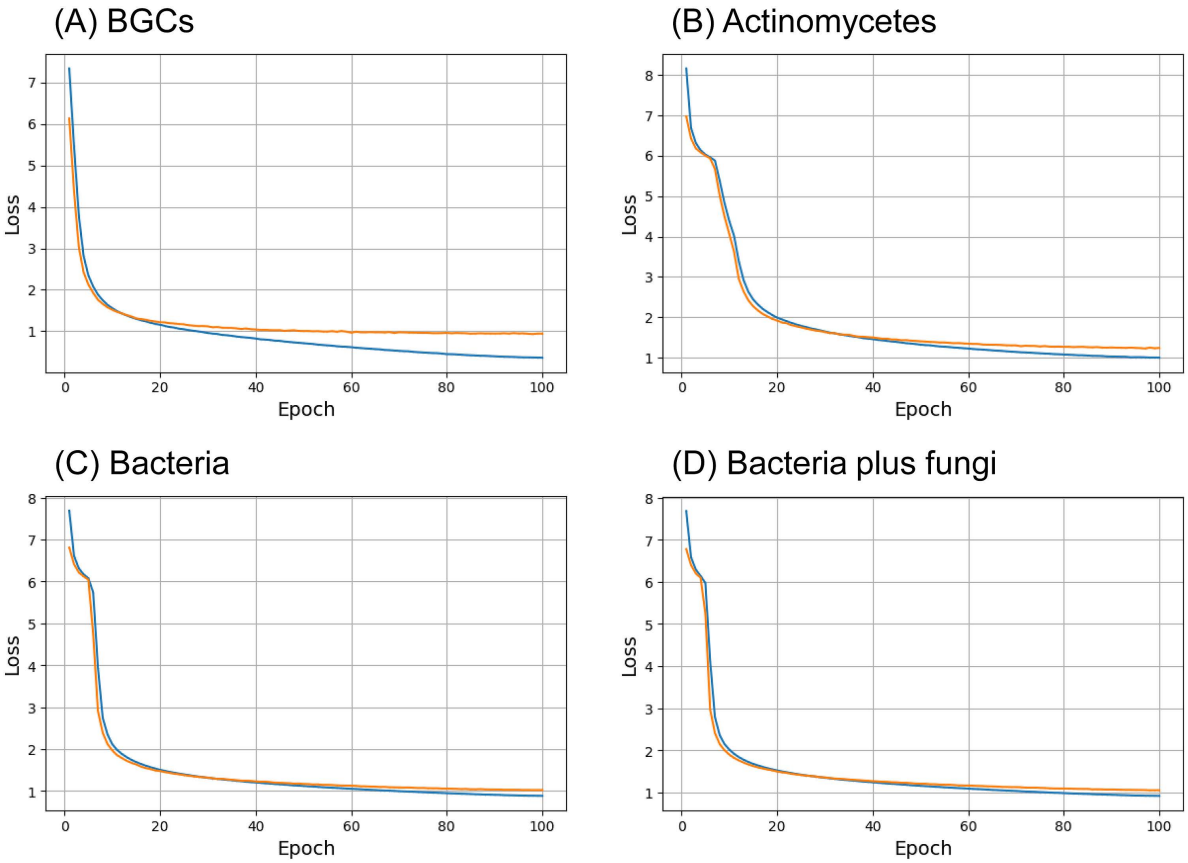

Fig 4. Training and validation loss curves during model training. Training (blue) and validation (orange) loss trajectories are plotted over training epochs for models trained on (A) Dataset I: biosynthetic gene clusters (BGCs), (B) Dataset II: Actinomycetes genomes, (C) Dataset III: bacterial genomes, and (D) Dataset IV: bacterial plus fungal genomes. All models exhibited decreases in loss during training, and the validation loss became larger than the training loss, indicating that the models effectively captured the combinatorial structure of domain arrangements.

https://doi.org/10.1371/journal.pcbi.1013181.g004

such as the Type II polyketide synthase class. The prediction accuracy per class was not correlated with the number of BGCs per class, indicating that the model distinguishes BGC classes based on the combination of functional domains rather than sample size.

For comparison, we implemented a simple Bag-of-Domains (BoD) baseline classifier using decision trees with depths of 1 and 3 (Table 2). The transformer-based model substantially outperformed the BoD classifiers in BGC class prediction; an overall accuracy and micro-F1 score of 0.765, with a macro-F1 of 0.577 and a weighted-F1 of 0.784, indicating balanced and robust performance across both major and minor classes. In contrast, the BoD classifiers showed markedly lower performance, with accuracies of 0.350 and 0.255 for tree depths of 1 and 3, respectively, and macro-F1 scores below 0.05. These results demonstrate that the transformer model effectively captures functional relationships among domains, whereas simple frequency-based representations in the BoD approach fail to generalize across diverse BGC classes. A per-class classification summary for our model and the BoD classifiers is provided in S1 Table.

In the antiSMASH database used for class annotation, hybrid BGCs are represented as separate class-specific entries (e.g., a PKS-NRPS hybrid BGC is recorded as two entries, one for PKS and one for NRPS). Among 2,762 hybrid BGCs with a total of 5,775 entries in the training set, 90% (2,490/2,762) were correctly predicted when accuracy was evaluated at the BGC level, that is, when a prediction was considered correct if the predicted class of any entry within the same BGC

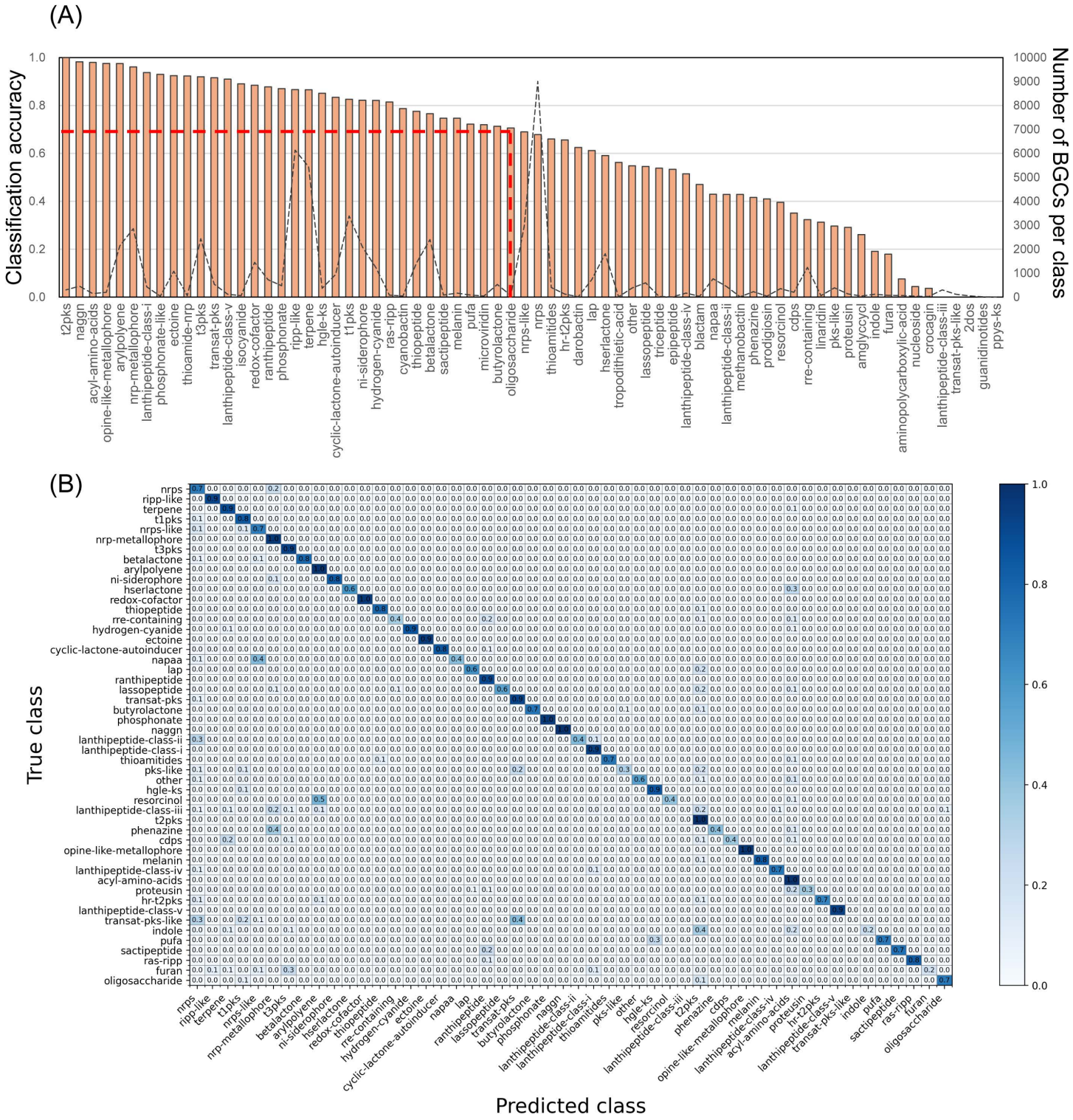

Fig 5. Classification performance across compound classes in the antiSMASH database. (A) Prediction accuracy per class. Bars represent the prediction accuracy of the model for each biosynthetic gene cluster (BGC) class. The dashed black line indicates the number of BGCs per class in the test set. The highest accuracies were observed for classes such as T2PKS and terpene (>0.8). The red dashed line denotes an accuracy reference of 0.7 for visual context; 37 of 75 classes containing >10 BGCs exceed this value. **(B)** Confusion matrix of class predictions. Values are normalized by the number of true samples per class. Off-diagonal cells with colors represent incorrect true-predicted pairs.

https://doi.org/10.1371/journal.pcbi.1013181.g005

Table 2. Prediction performance of the transformer-based model and Bag-of-Domains (BoD) classifiers with tree depths of 1 and 3.

Metric	Transformer model	BoD (depth = 1)	BoD (depth = 3)
Accuracy	0.765	0.350	0.255
Macro-F1	0.577	0.044	0.012
Micro-F1	0.765	0.350	0.255
Weighted-F1	0.784	0.272	0.136

https://doi.org/10.1371/journal.pcbi.1013181.t002

matched its true class. This percentage is substantially higher than the per-entry accuracy of 60% (3,477/5,775), reflecting the multiplicative effect of multiple predictions per hybrid region (S2 Table). The confusion matrix revealed several frequent class-level misclassifications, such as *resorcinol*/*arylpolyene*, *NAPAA*/*NRPS-like*, *transAT-PKS-like*/*transAT-PKS*, and *indole*/*T2PKS* (Fig 5B). However, these misclassified pairs were not specific to hybrid BGCs, as only 44 out of 5,775 hybrid entries showed cases where the incorrectly predicted class appeared in other hybrid entries.

The observed disagreements with antiSMASH class labels likely reflects differences in labeling logic between the two approaches: antiSMASH assigns hierarchical subclasses based on predefined domain rules, whereas our transformer learns domain-context relationships without such constraints, occasionally merging or reinterpreting related classes in a biologically plausible manner.

To visualize the model's capability to distinguish between BGC types, we projected learned feature representations of 2,492 experimentally validated BGCs from the MIBiG database [19] using t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction [20] (Fig 6). In Fig 6A, major compound classes such as NRP, saccharide, and ribosomally synthesized and post-translationally modified peptide form clearly separated clusters, reflecting the model's ability to capture class-specific features. In Fig 6B, BGCs associated with polyketide synthases (PKSs) further resolved into Type I (blue) and Type II (red) subclasses. These results support the model's capability to capture functional and structural characteristics relevant to compound class classification. Hybrid BGCs are broadly distributed across the map; however, many single-class BGCs also appear dispersed, suggesting that hybrid clusters are not the sole factor influencing the dispersion of single-class BGCs (Fig 6C).

Regarding BGC classification performance and embedding structure quality across MIBiG compound classes, our transformer-based Model IV outperformed DeepBGC's pfam2vec embeddings, achieving higher classification accuracy (0.81 vs 0.78) and macro-F1 (0.74 vs 0.69), along with improved local cluster purity across layers (Table 3). These results indicate that contextual learning along genomic sequences enhances both the separability and internal coherence of functional domain embeddings. Although the silhouette and Calinski–Harabasz indices decreased, suggesting weaker boundary delineation, this likely reflects the model's broader representation of domain combinations shared among functionally related BGC classes.

Model Performance in Functional Domain Prediction

To evaluate how well the developed models understand domain context, we conducted a masked prediction task using 2,492 BGCs from the MIBiG database. One domain at a time was masked, and models were evaluated on their ability to rank the true domain among the prediction.

Across all models, the median prediction rank of true domains was #1, more than 50% of true domains were top-ranked, and over 75% appeared within the top ten predictions (Table 4). As training data expanded from Actinomycetes (Model II) to all bacteria (Model III) and then to bacteria plus fungi (Model IV), the average and standard deviation of the prediction ranks improved (Fig 7). These results indicate that the models effectively capture BGC context and positional relationships among functional domains, and that greater diversity and size of training data enhanced model accuracy. When the four models were evaluated using only the 1,042 *Actinomycetota* BGCs from MIBiG, both the mean and

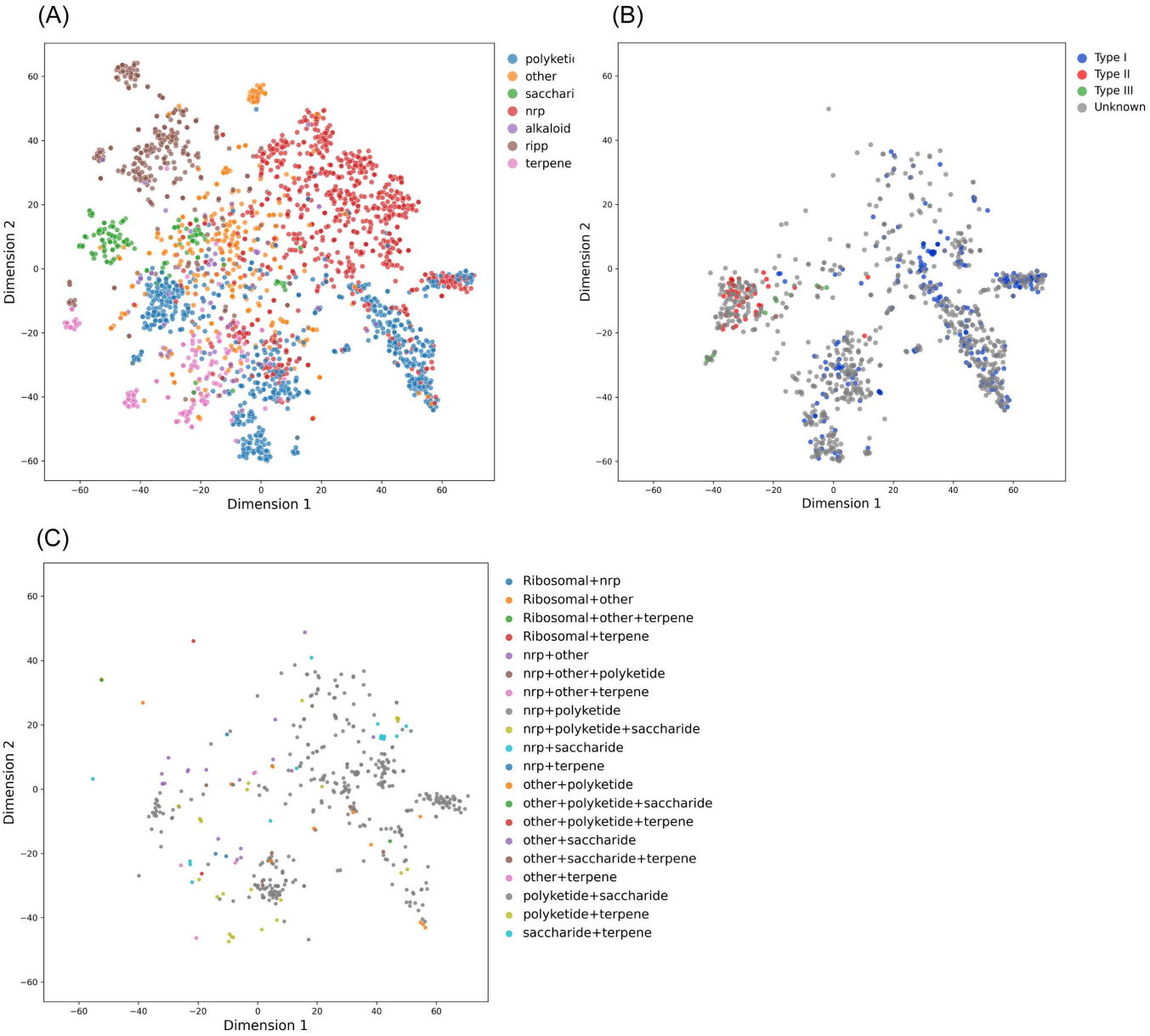

Fig 6. t-SNE visualization of biosynthetic gene cluster (BGC) feature representations. Feature embeddings of BGCs from the MIBiG database were visualized using t-distributed stochastic neighbor embedding (t-SNE). (A) Transformer-based embeddings (2,492 BGCs) colored by seven compound classes. Nonribosomal peptides (*nrp*, red), ribosomally synthesized and post-translationally modified peptides (*ripp*, brown), and saccharides (*saccharide*, green) formed distinct clusters. (B) Polyketide BGCs visualized by subclass. A clear separation between Type I and Type II polyketide synthases (PKSs) highlights subclass-level resolution. (C) Hybrid BGCs colored by their class combination. The most frequent hybrid type, PKS-NRPS ("nrp+polyketide", grey) is widely distributed across the t-SNE map.

https://doi.org/10.1371/journal.pcbi.1013181.g006

standard deviation of prediction ranks improved substantially in Models I, II, and III compared with evaluations using all 2,492 BGCs, whereas the improvement in Model IV was limited and its performance fell below that of Models I and III (Table A in S1 Data).

Table 3. Quantitative evaluation of learned domain embeddings compared with DeepBGC.

	DeepBGC	Model IV[1]		
Embedding	**pfam2vec**	**Early layer**	**Middle layer**	**Final layer**
BGC classification performance				
Lp-macro-F1	0.6939	0.7307 (0.0057)	0.7381 (0.0078)	0.7318 (0.0050)
Lp-accuracy	0.7816	0.8005 (0.0055)	0.8066 (0.0036)	0.8051 (0.0014)
Cluster purity @5	0.7495	0.7732 (0.0038)	0.7679 (0.0041)	0.7755 (0.0038)
Cluster purity @10	0.7295	0.7514 (0.0040)	0.7456 (0.0037)	0.7547 (0.0041)
Cluster purity @20	0.7110	0.7273 (0.0035)	0.7190 (0.0026)	0.7318 (0.0041)
Recall @1	0.0021	0.0021 (0.0000)	0.0021 (0.0000)	0.0021 (0.0000)
Recall @5	0.0094	0.0098 (0.0001)	0.0097 (0.0001)	0.0098 (0.0001)
Recall @10	0.0179	0.0187 (0.0001)	0.0184 (0.0001)	0.0187 (0.0002)
Embedding structure metrics				
Silhouette coefficient	0.0879	0.0388 (0.0023)	0.0375 (0.0025)	0.0351 (0.0013)
Calinski–Harabasz index	163.7	75.7 (1.1)	71.9 (1.8)	60.8 (1.7)
Davies–Bouldin index	3.8	4.5 (0.0)	4.6 (0.1)	5.1 (0.1)

[1]Standard deviations over five independent runs are in parentheses.

https://doi.org/10.1371/journal.pcbi.1013181.t003

Table 4. True domain rank statistics across four training datasets using 2,492 biosynthetic gene clusters (BGCs) in the MIBiG database.

#	I	II	III	IV
Data	**Bacterial BGCs**	**Actinomycetes genomes**	**Bacterial genomes**	**Bacterial plus fungal genomes**
Mean	129	343	143	121
Median	1	1	1	1
SD	1105	1804	1093	1031
Minimum	1	1	1	1
Maximum	19503	19711	19706	19692
#1	71.2%	50.7%	57.2%	58.3%
Within #10	88.3%	75.8%	81.9%	83.0%
Above #1000	1.8%	5.3%	2.3%	2.3%
# Token	8.6 M	14 M	49 M	70 M

https://doi.org/10.1371/journal.pcbi.1013181.t004

The MIBiG database is one of the most comprehensive collections of experimentally validated BGCs [19]. Excluding these entries from the training data would therefore remove valuable biological information that defines known BGC architectures. Instead of omitting them to avoid potential data leakage, we implemented a similarity-tagging procedure in which each MIBiG domain was labeled according to its similarity to the training data; L1 (exact match), L2 (near match), L3 (family-level similarity), or None (no detectable similarity). As shown in Fig 8, most L1

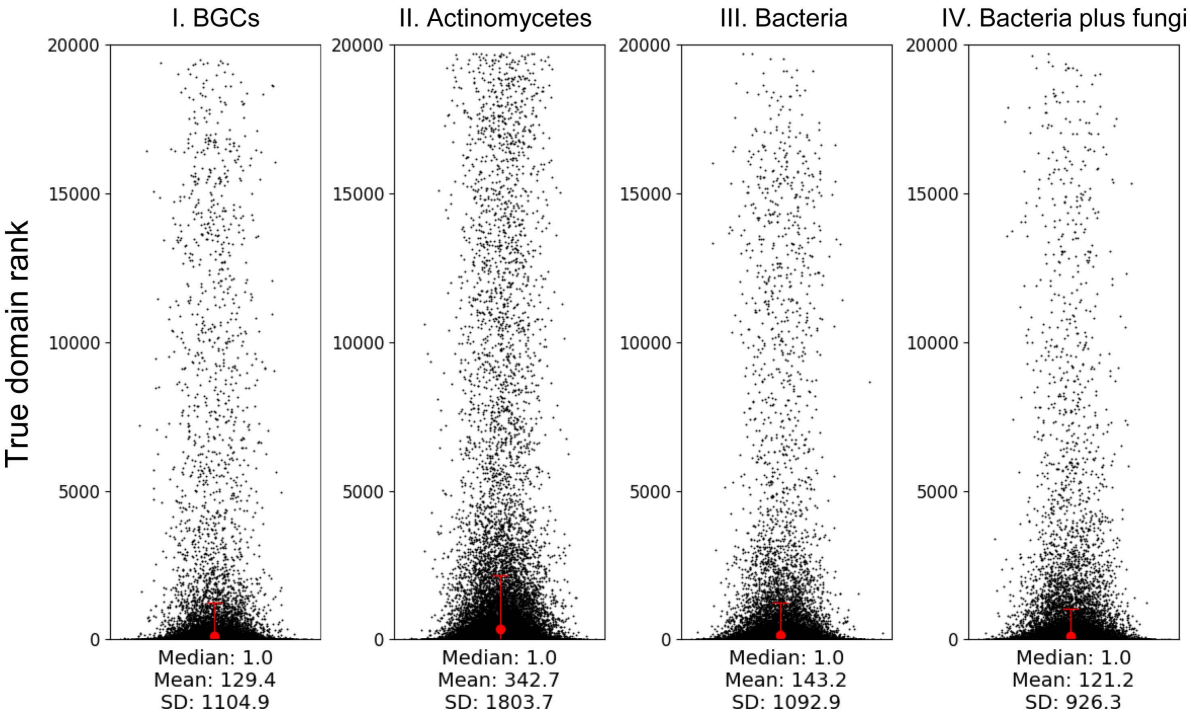

Fig 7. Distribution of prediction rankings for 2,492 biosynthetic gene clusters (BGCs) from the MIBiG database across four training datasets. Shown are the distributions of true functional domain ranks predicted by models trained on I. antiSMASH-predicted bacterial BGCs, **II.** Actinomycetes genomes, III. bacterial genomes, and IV. bacterial plus fungal genomes. Red dots represent the average prediction rank, and red bars indicate the standard deviation (SD) for each dataset.

https://doi.org/10.1371/journal.pcbi.1013181.g007

domains were predicted as rank #1, while a substantial number of None domains were also ranked #1, indicating that the model generalizes its predictive accuracy beyond the known BGCs. The statistics of true domains per similarity tag in each model, for both the 2,492 MIBiG BGCs and the 1,092 *Actinomycetota* BGCs, is summarized in S3 Table.

For the 3,157 unique functional domains among the 14,455 total unique domains in the training dataset of Model IV, the mean true domain rank showed a positive correlation, while the average probability of being ranked as top-1 exhibited a weak decreasing trend, with increasing median rank (Fig 9A and 9B). In contrast, token counts per domain in the training and evaluation datasets did not show apparent such trends (Fig 9C and 9D), indicating that the model learned genomic domain organization independently of token frequency. Domains tend to be predicted as functionally related types; for example, one of the most frequent domains in datasets, ketoacyl-synt, catalyzes the condensation reaction that extends polyketide chains, and is often predicted within the top ten as KR, Ketoacyl-synt_C, Acyl_transf_1, or PKS_DE, which are known to participate in polyketide synthesis [21]. This indicates that the model successfully captures functional relationships among biosynthetically related domain types.

Context dependency of domain prediction

To determine how input context affects domain prediction by the constructed models, we performed a systematic perturbation analysis in which input domains were randomly altered. We observed the prediction probability of the true domain at a masked position domain by Model IV as increasing the number of altered domains (Fig 10).

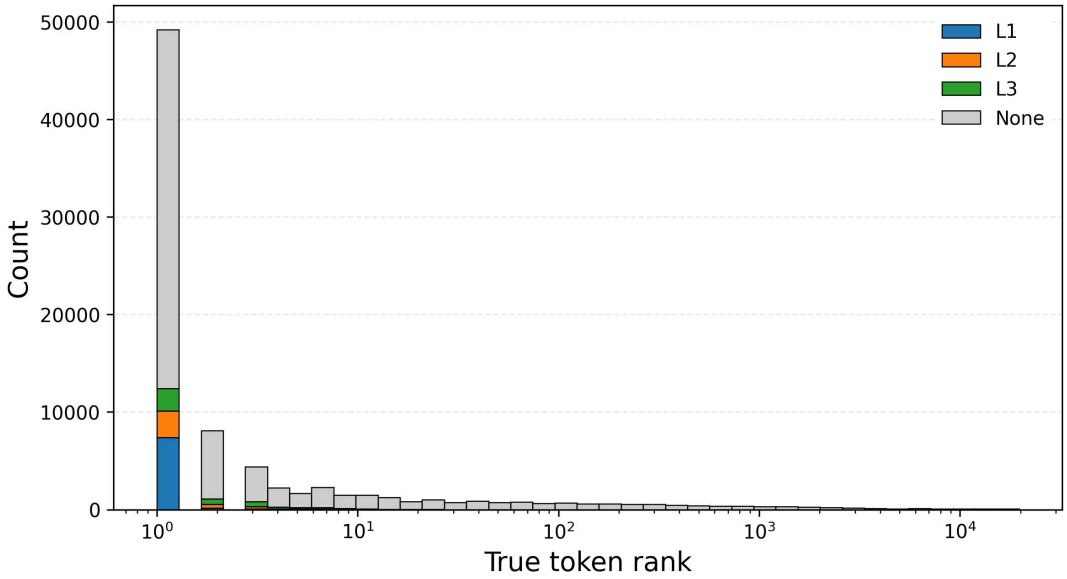

Fig 8. Histogram of prediction ranks with similarity tags for MIBiG domains. Shown is the distribution of true functional-domain ranks predicted by Model IV, trained on bacterial plus fungal genomes. Each bar represents the count of domains grouped by similarity tag (L1, L2, L3, or None). Most L1 domains were correctly predicted as rank #1, while a substantial number of None domains (no detectable similarity) were also predicted as rank #1, indicating that the model generalized its accuracy beyond training examples.

https://doi.org/10.1371/journal.pcbi.1013181.g008

The average prediction probability decreased monotonically as replacement rate increased, across BGCs containing 6, 11, 16, and 21 functional domains from the MIBiG database (Fig 10B). These results indicate that unmasked domains serve as critical contextual cues, effectively guiding the model in predicting masked BGC domains.

Design of Artificial BGCs from Model Prediction

We next tested the model's generative capability to propose plausible domain combinations. Compared to the model trained solely on known BGCs, the model trained on comprehensive genome datasets is expected to predict broader and more diverse sets of biologically plausible domain combinations, including those not observed in known BGCs. To illustrate this, we compared the candidate lists generated by Model I (trained only on BGCs) and Model IV (trained on bacterial plus fungal genomes) using the cyclooctatin biosynthetic pathway as a case study. Cyclooctatin is a C_{20} diterpene produced by *Streptomyces melanosporofaciens*, whose biosynthesis is catalyzed by four enzymes, CotB1 through CotB4 [22] (Fig 11). CotB1 synthesizes geranylgeranyl diphosphate (GGDP), while CotB2, the main terpene synthase, catalyzes the cyclization and hydroxylation of GGDP to produce cyclooctat-9-en-7-ol (Fig 11B). Subsequently, CotB3 and CotB4, both P450 hydroxylases, introduce hydroxyl groups at the C5 and C18 positions, respectively, to yield cyclooctatin (Fig 11B).

When introducing a masked domain after the *cotB4* gene in the cyclooctatin BGC, the genome-trained model (Model IV) predicted functional domains such as polyprenyl_synt, FAD_binding_3 and Methyltransf_2, which were not ranked highly by the BGC-trained model (Model I) (Table 5). These differences likely reflect that Model IV learned a broader diversity of functional domain combinations from whole genomes, including those that may represent transitional or partially integrated domains under the evolution of BGC membership. Thus, the newly predicted domains represent plausible candidate functions for modifying cyclooctatin or its intermediates.

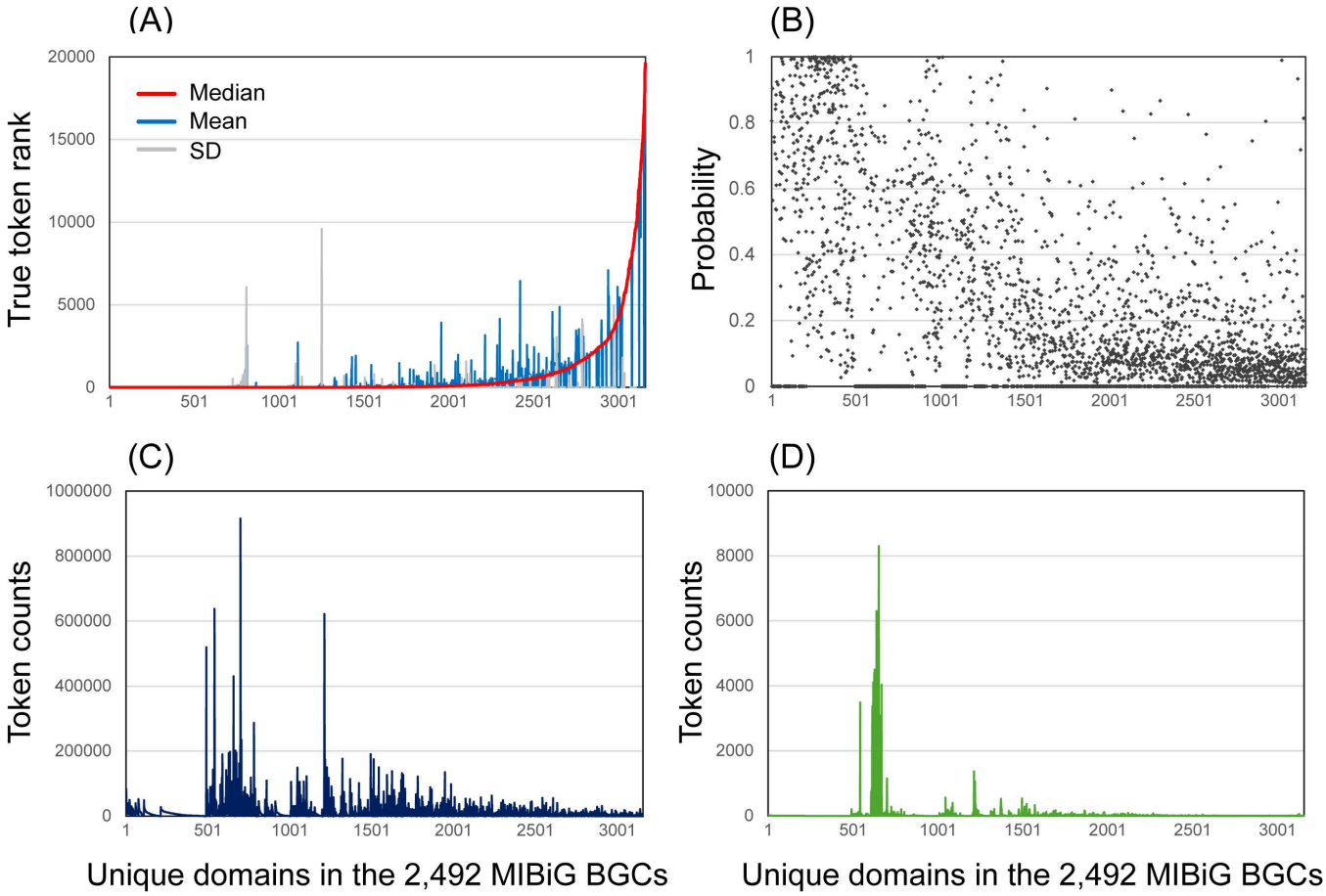

Fig 9. Prediction summary per functional domain in Model IV. (A) Mean, median, and standard deviation of true token ranks. **(B)** Average probability of being ranked as top-1. **(C)** Token counts in the training dataset. **(D)** Token counts in the 2,492 MIBiG BGCs used for evaluation. Tokens are ordered by the median of true domain ranks. The mean true domain rank shows a consistent increasing trend, while the average top-1 probability exhibits a weak degreasing trend with increasing median rank. In contrast, token counts in the training and evaluation datasets show no apparent correlation with prediction rank statistics.

https://doi.org/10.1371/journal.pcbi.1013181.g009

Experimental validation of model prediction

To experimentally validate these model-generated domain suggestions, we need to obtain appropriate amino acid sequences of proteins containing suggested domains, because currently our model solely depends on domain sequences. For this purpose, we used the EFI-GNT web resource [23,24] to search BGCs containing both a CotB1 homolog and a gene encoding domains uniquely predicted by Model IV. We identified such candidate clusters containing the FAD_binding_3, Bac_luciferase, or Glyoxalase domains; however, in the latter two cases, the predicted domain-containing genes were separated by at least two intervening genes from the CotB1 homolog. To prioritize contiguous gene arrangement, we selected a BGC from *Streptomyces* sp. ISL86, which possesses a CotB1 homolog (E-value: 5e-119) and a gene encoding the predicted domain, separated by one intervening gene from the CotB1 homolog.

Based on the amino acid sequence translated from the gene encoding the FAD_binding_3 domain, we constructed a plasmid harboring this gene and co-expressed it in *S. albus* G153 with a plasmid harboring CotB1 and CotB2, CotB1 through CotB3, or CotB1 through CotB4 (Fig 12A). Subsequent LC-MS analysis revealed trace amounts of unknown metabolites specifically produced by the *S. albus* G153 transformant co-expressing FAD_binding_3 with CotB1 and

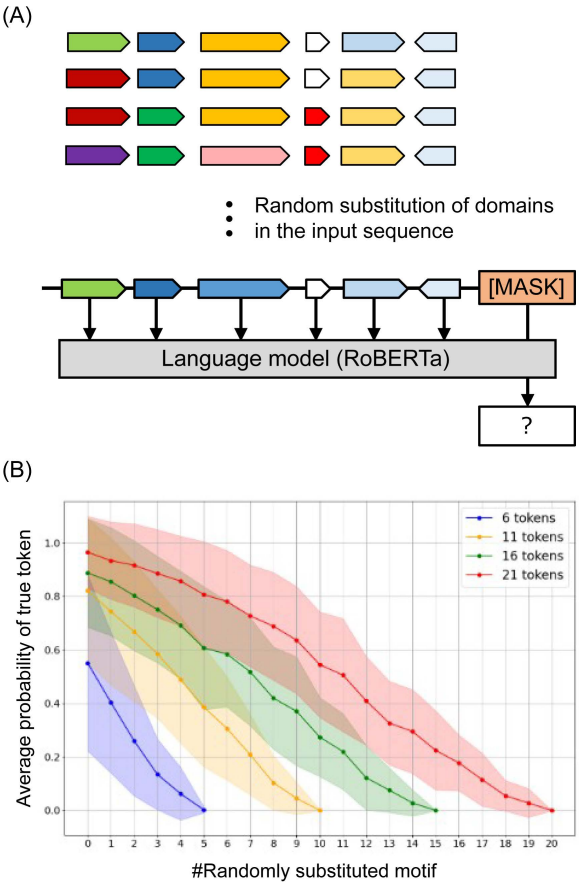

Fig 10. Effect of token substitutions on domain prediction accuracy. (A) Schematic of the perturbation experiment to investigate how prediction accuracy for a masked token changes when randomly substituting other tokens in the input sequence (excluding [SEP] tokens). Colored arrows represent functional domains in proteins. **(B)** Prediction results for domain clusters consisting of 6, 11, 16, and 21 tokens (sample sizes of 81, 66, 71, and 65 clusters, respectively). The y-axis shows the model's prediction probability for the correct token at the [MASK] position, while the x-axis indicates the number of randomly substituted tokens. Shaded areas represent standard deviations across samples.

https://doi.org/10.1371/journal.pcbi.1013181.g010

CotB2, CotB1 through CotB3, or CotB1 through CotB4 (Figs 12B and B in S1 Data). High-resolution MS analysis deduced that one of the unknown metabolites shares the same molecular formula as cyclooctat-9-ene-5,7-diol (Fig 12C). Nevertheless, their retention times under the chromatographic conditions were clearly different, suggesting that they represent distinct structural isomers of cyclooctat-9-ene-5,7-diol.

This finding validates the model's ability to propose novel, biologically plausible domain combinations that can be functionally expressed in a heterologous host.

Discussion

This study demonstrates that our transformer-based model can effectively learn and predict BGC patterns by treating functional domains as linguistic units. This approach provides both conceptual insights into BGC organization and practical tools for natural product discovery. Notably, divergent prediction patterns between the BGC-trained and genome-trained models suggest that broader training contexts may reveal alternative or previously unexplored biosynthetic trajectories, as exemplified in the analysis of cyclooctatin BGC.

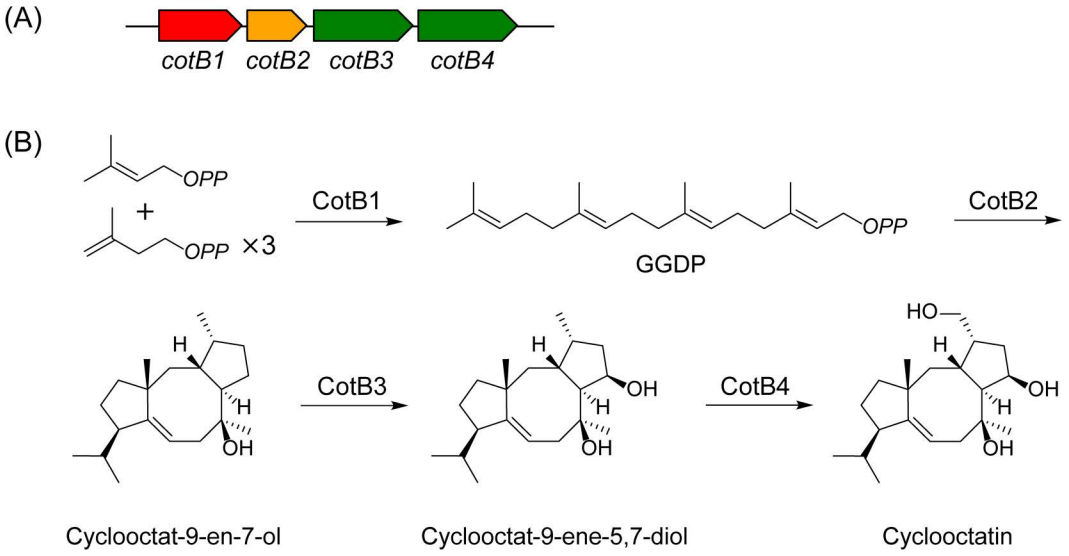

Fig 11. Cyclooctatin biosynthetic gene cluster and biosynthetic mechanism. (A) Gene cluster responsible for cyclooctatin biosynthesis, comprising four genes, *cotB1* to *cotB4*. **(B)** Mechanism of cyclooctatin biosynthesis [22].

https://doi.org/10.1371/journal.pcbi.1013181.g011

Table 5. Comparison of top-7 predicted domains for an additional domain to the cyclooctatin biosynthetic gene cluster by Model I and Model IV.

	Model I			Model IV		
Rank	**Pfam domain**	**Probability**	**Rank in Model IV**	**Pfam domain**	**Probability**	**Rank in Model I**
1	p450	0.1928	2	Terpene_syn_C_2	0.2592	14
2	Flavodoxin_1	0.1563	>100	p450	0.1133	1
3	AMP-binding	0.1224	42	polyprenyl_synt	0.0144	>100
4	adh_short	0.0470	7	MFS_1	0.0137	>100
5	cNMP_binding	0.0426	>100	FAD_binding_3	0.0132	35
6	AMP-binding_C	0.0387	>100	Methyltransf_2	0.0100	54
7	Prenyltrans	0.0232	98	adh_short_C2	0.0089	26

https://doi.org/10.1371/journal.pcbi.1013181.t005

We experimentally validated a model-predicted domain combination through heterologous co-expression of the FAD_binding_3 gene with cyclooctatin biosynthetic genes. The FAD_binding_3 domain is found in various enzymes, typically monooxygenases and hydroxylases, where it binds and utilizes flavin adenine dinucleotide as a redox cofactor. Although detected in trace amounts, a putative isomer of cyclooctat-9-ene-5,7-diol was specifically observed in transformants co-expressing CotB1-B2, CotB1-B3, or CotB1-B4 with the FAD_binding_3 protein (Figs 12B and B in S1 Data). This observation aligns with the model's prediction that such domain combination could lead to the production of unknown metabolites. While structural elucidation of the unknown isomer remains necessary, this form of metabolomic validation offers critical feedback for iterative model refinement.

Compared to existing tools such as antiSMASH, which rely on known BGC patterns, our model learns from both BGC and non-BGC genomic regions and can propose novel domain combinations. The use of a transformer architecture enables the capture of long-distance dependencies among domains, which is crucial for understanding cluster organization. These strengths are reflected in the models' high prediction accuracy and its ability to propose biologically plausible, previously uncharacterized domain assemblies.

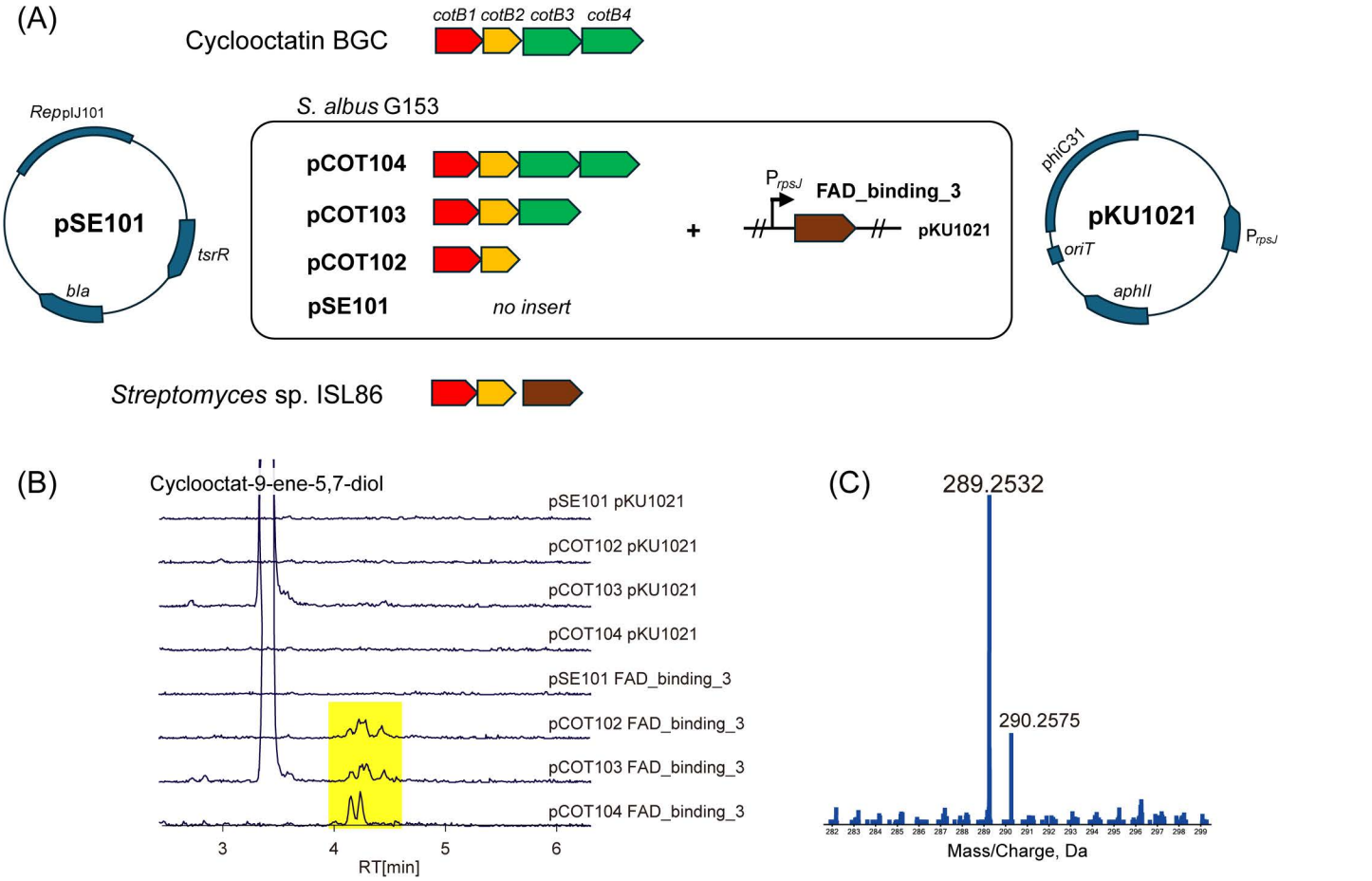

Fig 12. Functional analysis of the domain predicted by the model. (A) A plasmid encoding the gene with the predicted FAD_binding_3 domain from *Streptomyces* sp. ISL86 was constructed and co-expressed with plasmids harboring *cotB1* to *cotB4* in *S. albus* G153. **(B)** Extracted ion chromatograms for *m*/*z* 289.252 in the liquid chromatography–mass spectrometry analysis. The culture extracts were analyzed from the *S. albus* transformants harboring pCOT102/103/104 with and without the FAD_binding gene. Cyclooctat-9-ene-5,7-diol (*m*/*z* 289.2526 $[M-H_2O+H]^+$) was abundantly detected, eluting at 3.3 min in the extracts from *S. albus* pCOT103 pKU1021 and *S albus* pCOT103 FAD_binding_3. In addition, several previously undetected signals, highlighted in yellow, were observed in trace amounts at around 4.2 minutes in the extracts of *S. albus* pCOT102 FAD_binding_3, *S. albus* pCOT103 FAD_binding_3, and *S. albus* pCOT104 FAD_binding_3. Notably, these signals shared the same *m*/*z* 289.2526 as cyclooctat-9-ene-5,7-diol. **(C)** High resolution MS spectrum of one of the unknown metabolites in the extract from *S. albus* pCOT102 FAD_binding_3.

https://doi.org/10.1371/journal.pcbi.1013181.g012

Nonetheless, several limitations remain. First, although the current model effectively predicts domain-level organization, it does not generate amino acid sequences optimized for protein-protein interactions or catalytic compatibility. This is because the model is based exclusively on genomic co-occurrence patterns, without incorporating nucleotide or amino acid sequence information. We plan to address this issue by integrating protein sequence design tools such as ProteinMPNN [25] or ProGen [26], as well as nucleic acid generation tools such as Evo [15]. Second, the model cannot autonomously generate candidate artificial BGCs; instead, candidate domains must be selected by comparing predictions across models trained on different datasets. This limitation could be overcome by incorporating additional computational modules, including rule-based heuristics for candidate selection and neural network-based components for domain sequence generation and translation.

In the per-class accuracies (Fig 5A), certain classes such as *transAT-PKS-like* and *PpyS-KS* showed zero accuracy. Upon inspection, these discrepancies primarily resulted from overly fine subclass annotations rather than a failure of

model learning. For example, *transAT-PKS-like* clusters were often predicted as *transAT-PKS* or *PKS*, which are chemically and biosynthetically related categories. This observation suggests that the current BGC annotations granularity may exceed the resolution supported by the available training data. A possible solution would be to adopt hierarchical classification schemes, allowing evaluation at both major-class and subclass levels to more accurately reflect model performance. In addition, in some cases such as *PpyS-KS*, the correct class appeared as the second- to fourth- ranked prediction, which was still counted as incorrect. A rank-weighted accuracy metric could provide a more comprehensive assessment, better capturing near-correct predictions.

The model's ability to predict biologically plausible domain combinations can prioritize BGCs for experimental validation. Differences in predictions between BGC- and genome-trained models may serve as indicators of evolutionarily atypical clusters with potential for novel bioactivity. Future developments should include integrating transcriptomic data to capture co-expression patterns, using protein structure predictions to assess domain compatibility, and incorporating phylogenetic and metabolic network data to improve biological plausibility. These enhancements can be achieved through a pre-training and fine-tuning framework that balances computational efficiency with multimodal data integration. While this study emphasizes generative design of new domain architectures, the framework may also be adapted for de novo genome mining to identify previously uncharacterized BGC types, which will be the subject of future work.

In conclusion, this study demonstrates the potential of applying natural language modeling techniques to understand and design BGCs. While the current framework focuses on domain-level prediction, future integration of sequence design, gene expression patterns, and structural constraints will enable generation of more evolutionarily informed and functionally robust BGCs. This transformer-based strategy provides a scalable platform for both theoretical studies of BGC evolution and practical applications in the discovery and engineering of natural products.

Although our validation focused on the cyclooctatin BGC, we anticipate that the broader utility of the platform will be further demonstrated as the biochemical community applies it to diverse BGCs. Such cumulative applications will not only reinforce confidence in its universality but also enhance the long-term value of this study.

Methods

Data collection and preprocessing

Genomic Data. Microbial genome data was downloaded from the National Center for Biotechnology Information. For bacterial genomes, we selected 12,186 complete genomes annotated in RefSeq. For fungal genomes, 2,670 assemblies were collected using GenBank submitter annotations. For each genome, both gene location files (GFF format) and protein amino acid sequences (protein-FASTA format) were retrieved.

BGC Data. We obtained 239,021 predicted bacterial BGC data (GenBank format) from the antiSMASH database. Protein sequences of genes located within BGC regions were extracted from each file. Each BGC was treated as a linearized sequence of functional domains for model input.

Pfam Domain Identification. Pfam domains were annotated using HMMer version 3.3.1 [27] with the Pfam-A.hmm database (Pfam v35.0) [28]. The E-value threshold was set to 1e-10. Domain annotations were integrated with GFF files to determine domain coordinates on genomes. In cases where domain annotations overlapped in a region, the domain with the better alignment score was retained. A special token [SEP] was inserted between genes to preserve gene boundary information in domain sequences.

Construction of Tokenization system

Pfam domain names were extracted from the Pfam-A.hmm database. A tokenizer containing 19,523 unique domain tokens was constructed using the PreTrainedTokenizerFast module from the HuggingFace Transformers library (https://huggingface.co/docs/transformers/index). Each domain was treated as a discrete token, allowing BGCs and genomic sequences to be represented as tokenized sequences for language model input.

Model architecture and training

RoBERTa Model. We used the RoBERTa architecture implemented in the Hugging Face Transformers library. A limited hyperparameter search was conducted on a subset of the bacterial dataset (3×10^6 tokens) to identify stable configurations. The search space included variations in the number of hidden layers (6–10), hidden size (768–1024), number of attention heads (12–16), learning rate (3×10^{-4} to 5×10^{-5}), weight decay (0.01–0.02), warmup ratio (0.04–0.06), and learning-rate scheduler type (cosine or linear). The final configuration was selected based on the lowest validation loss and stable convergence across three independent runs, as follows:

- Number of hidden layers: 8
- Number of attention heads: 16
- Hidden layer dimension: 1,024
- Maximum input sequence length: 512 tokens
- Learning rate: 3×10^{-4} (5×10^{-5} for Dataset I)
- Weight decay: 0.02
- Warmup ratio: 0.06
- Learning-rate scheduler type: linear
- Total number of trainable parameters: 105,781,515

Training was conducted using the Masked Language Modeling (MLM) tasks. For each input sequence, 15% of tokens were randomly selected for masking, and among them, 80% were replaced with [MASK] tokens, 10% with random tokens, and the remaining 10% were kept unchanged. The model learned to predict the original tokens based on surrounding context.

Training Datasets. Four models were trained on datasets with increasing taxonomic scopes:

- Model I: bacterial BGCs
- Model II: Actinomycetes genomes
- Model III: bacterial genomes
- Model IV: bacterial plus fungal genomes

For each dataset, 80% of samples were used for training and 20% reserved for testing in a leakage-safe manner. Specifically, non-overlapping domain chunks were generated from Pfam domain sequences, and their pairwise similarities were computed using 6-gram MinHash signatures (128 permutations, 32 bands) following the locality-sensitive hashing (LSH) approach [29,30]. Pairs of chunks with Jaccard similarity≥ 0.8 were connected to form a similarity graph, with each connected component treated as a similarity group [31]. The data were then partitioned using grouped K-fold splitting [32], ensuring that all chunks within a group were assigned to the same partition and thereby preventing chunk-level leakage between the training and test sets.

To assess model reproducibility, five independent rounds of data splitting and training were performed for Dataset IV. All runs exhibited smooth decreases in both training and evaluation losses, along with consistent reproduction of true domain-rank statistics (Fig C and Table B in S1 Data).

Evaluation methods

BGC Classification Task. To assess model performance in compound class prediction, a token indicating the BGC class (*e.g.*, T1PKS, T2PKS, NRPS, terpene) as defined by antiSMASH was appended to the beginning of each domain

sequence from Dataset I including 191,216 bacterial BGCs retrieved from the antiSMASH database. During training, this label token was masked as part of the MLM task to infer the correct class from domain context. Model performance was evaluated on an independent set of 47,805 bacterial BGCs from the same dataset that were not used for training.

For comparison, we implemented a Bag-of-Domains (BoD) baseline classifier using scikit-learn modules [33], analogous to a Bag-of-Words approach [34], treating domain tokens as words and predicting BGC classes using decision trees with depth of 1 and 3. Model performance was evaluated using four standard multi-class classification metrics: accuracy, macro-F1, micro-F1, and weighted-F1. Accuracy measures the proportion of correctly predicted samples among all predictions. Macro-F1 computes the unweighted mean of per-class F1 scores, giving equal importance to all classes. Micro-F1 aggregates true and false predictions across classes to evaluate overall correctness, while weighted-F1 averages per-class F1 scores weighted by class frequency. Together, these metrics provide complementary views of overall and class-specific prediction performance.

Embedding Evaluation Metrics. To quantitatively assess the quality of domain embeddings in comparison with the DeepBGC embedding model, pfam2vec, we evaluated both supervised classification and semi-supervised structural metrics using scikit-learn modules [33]. Embeddings were extracted from the early (2nd), middle (5th), and final (8th) transformer encoder layers, corresponding to progressively deeper contextual representations, from local domain composition to global BGC architectural organization.

For supervised evaluation, a linear probe classifier (logistic regression) was trained on frozen embeddings to assess class separability using 2,492 MIBiG BGCs annotated with seven compound classes (alkaloid, NRP, polyketide, RiPP, saccharide, terpene, and other). Performance was measured by Lp-accuracy (classification accuracy) and Lp-macro-F1 (mean of per-class F1 scores, equally weighting all classes). Local neighborhood consistency was evaluated using cluster purity @k (k = 5, 10, 20), representing the proportion of top-k nearest neighbors sharing the same class, and recall @k (k = 1, 5, 10), measuring the probability that a true neighbor appears among the top-k retrieved embeddings.

For semi-supervised structure evaluation, we used three common clustering metrics: the silhouette coefficient [35], the Calinski–Harabasz (CH) index [36], and the Davies–Bouldin (DB) index [37], which respectively quantify cluster cohesion versus separation, between- versus within-cluster dispersion, and average inter-cluster similarity. Clusters were defined according to the seven compound classes in the MIBiG database.

Together, these indices provide a comprehensive assessment of both the discriminative power and structural organization of the learned embedding spaces.

Visualization of Feature Representations. To visualize learned feature representations, we performed t-SNE dimensionality reduction using embeddings from Model IV:

1. The domain sequence of each of the 2,492 MIBiG BGCs was input into Model IV.
2. For each domain, a 1,024-dimensional output feature vector from the model was extracted.
3. The domain vectors for each BGC were averaged to obtain a single feature representation per BGC.
4. Feature vectors were compressed to two dimensions using the t-SNE algorithm (scikit-learn implementation, perplexity = 30, random_state = 42).
5. BGCs were visualized in 2D space, colored according to their compound class annotations.

MIBiG domain tagging for the MLM domain prediction task

To assess potential redundancy between the model training data and the MIBiG domains used for evaluation, we implemented a similarity-based tagging procedure using 5-gram MinHash/LSH similarity between domain-token sequences [29,30]. Each MIBiG domain sequence was converted into a token window (11 tokens; half-width $W = 5$), for which 5-gram

shingles were generated and compared against the MinHash index constructed from the training dataset. The maximum Jaccard similarity [31] between each MIBiG domain and any training-domain window was then used to assign a similarity tag:

- L1 (exact): identical 5-gram signature found in training,
- L2 (near): high-similarity signature (shared Jaccard ≥ 0.8),
- L3 (family): moderate similarity (0.7 <shared Jaccard < 0.8),
- None: no detectable overlap with any training chunk.

These tags were used only for analysis and visualization to evaluate potential data leakage and contextual redundancy in the MIBiG evaluation dataset.

Experimental Validation

General. Biochemicals and enzymes for genetic manipulation were purchased from NEB (Tokyo, Japan). Oligonucleotide primers used for genetic manipulation were purchased from Fasmac Co., Ltd. (Kanagawa, Japan). HR-ESI-MS spectra were collected using an SCIEX X500R QTOF system equipped with a UPLC Nexera system (Shimadzu, Kyoto, Japan). **Strain Construction.** The gene encoding the FAD_binding_3 domain was PCR-amplified using the primers FAD3_fw (GCCAAGCTTTCAGGAGCCGGC) and FAD3_rv(TCCTCTAGAATGCAGCAGCGCCC) from the synthetic gene (synthesised by Twist Bioscience, codon-optimized for expression in *Streptomyces coelicolor* A3(2), Table C in S1 Data). The PCR-amplified gene was digested with XbaI and HindIII and cloned downstream of the *rpsJ* promoter (P*rpsJ*) on the pKU1021 vector to yield pKU1021_FAD_binding_3. The resultant vector was then used for the transformation of *Streptomyces albus* G153 with polyethylene glycol-mediated protoplast transformation. Protoplasts were prepared using the standard protocol. The resultant strain was further transformed using pSE101, pCOT102, pCOT103 or pCOT104 with the same method, which were previously constructed [22].

Metabolic analysis. Each heterologous expressing strain was inoculated into 10 mL TSB and incubated with shaking (300 rpm) at 30°C for 2 d. A total of 2 mL of the preculture was inoculated into 100 mL of TSB medium and incubated continuously with shaking (180 rpm) at 27°C for 3 d. After fermentation, two times volume of acetone was added to the culture broth. After the extraction, the acetone was evaporated, and the remaining aqueous fraction was extracted with ethyl acetate. The ethyl acetate fraction was recovered and evaporated in vacuo. The remaining residue was dissolved in methanol and analyzed using HR-ESI-LC-MS equipped with a CAPCELL PAK C18 IF column (2.0 φ x 50 mm; Shiseido, Tokyo, Japan). LC conditions were as follows: mobile phase A, water + 0.1% formate; mobile phase B, acetonitrile + 0.1% formate; 10%–90% B over 5 min, 90% B for 2.5 min and then 10% A for 2.5 min, at a flow rate of 0.4 mL/min.

Supporting information

S1 Table. Summary of class prediction performance by the transformer-based model and Bag-of-Domains classifiers (tree depths = 1 and 3).
(XLSX)

S2 Table. Class prediction results for hybrid biosynthetic gene clusters (BGCs).
(XLSX)

S3 Table. Summary statistics of true domains categorized by similarity tag.
(XLSX)

S1 Data. Table A. Domain prediction performance across four training datasets using Actinomycetota 1,042 biosynthetic gene clusters (BGCs) in the MIBiG database. Table B. True domain rank statistics across five independent runs using Dataset IV for 2,492 biosynthetic gene clusters (BGCs) in the MIBiG database. Table C. Nucleotide and amino acid sequences of the synthetic gene encoding the FAD_binding_3 domain used in this study. Fig A. Training and validation loss curves of RoBERTa and GPT-1 models on Dataset II. Fig B. Extracted ion chromatograms for cyclooctatin and its intermediates from *Streptomyces albus* transformants harboring pCOT102/103/104, with or without the FAD_binding_3 gene. Fig C. Cross validation of loss behavior during model training.
(PDF)

Acknowledgments

We thank Dr Totai Mitsuyama at National Institute of Advanced Industrial Science and Technology and Dr Yuki Kanai at University of Tokyo for valuable discussion.

Author contributions

Conceptualization: Maiko Umemura.

Data curation: Tomoki Kawano.

Funding acquisition: Tomohisa Kuzuyama, Maiko Umemura.

Investigation: Tomoki Kawano, Taro Shiraishi.

Methodology: Tomoki Kawano, Taro Shiraishi, Maiko Umemura.

Project administration: Tomohisa Kuzuyama.

Software: Tomoki Kawano.

Supervision: Tomohisa Kuzuyama, Maiko Umemura.

Validation: Tomoki Kawano, Taro Shiraishi.

Visualization: Tomoki Kawano.

Writing – original draft: Tomoki Kawano, Taro Shiraishi.

Writing – review & editing: Taro Shiraishi, Tomohisa Kuzuyama, Maiko Umemura.

References

1. Bibb MJ. Regulation of secondary metabolism in streptomycetes. Curr Opin Microbiol. 2005;8(2):208–15. https://doi.org/10.1016/j.mib.2005.02.016 PMID: 15802254
2. Osbourn AE, Field B. Operons. Cell Mol Life Sci. 2009;66(23):3755–75. https://doi.org/10.1007/s00018-009-0114-3 PMID: 19662496
3. Nützmann H-W, Scazzocchio C, Osbourn A. Metabolic Gene Clusters in Eukaryotes. Annu Rev Genet. 2018;52:159–83. https://doi.org/10.1146/annurev-genet-120417-031237 PMID: 30183405
4. Kanai Y, Tsuru S, Furusawa C. Experimental demonstration of operon formation catalyzed by insertion sequence. Nucleic Acids Res. 2022;50(3):1673–86. https://doi.org/10.1093/nar/gkac004 PMID: 35066585
5. Schmidt EW, Nelson JT, Rasko DA, Sudek S, Eisen JA, Haygood MG, et al. Patellamide A and C biosynthesis by a microcin-like pathway in Prochloron didemni, the cyanobacterial symbiont of Lissoclinum patella. Proc Natl Acad Sci U S A. 2005;102(20):7315–20. https://doi.org/10.1073/pnas.0501424102 PMID: 15883371
6. Kenshole E, Herisse M, Michael M, Pidot SJ. Natural product discovery through microbial genome mining. Curr Opin Chem Biol. 2021;60:47–54. https://doi.org/10.1016/j.cbpa.2020.07.010 PMID: 32853968
7. Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res. 2023;51(W1):W46–W50. https://doi.org/10.1093/nar/gkad344
8. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158(2):412–21. https://doi.org/10.1016/j.cell.2014.06.034 PMID: 25036635

9. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 2019;47(18):e110. https://doi.org/10.1093/nar/gkz654 PMID: 31400112
10. Almeida H, Palys S, Tsang A, Diallo AB. TOUCAN: a framework for fungal biosynthetic gene cluster discovery. NAR Genom Bioinform. 2020;2(4):lqaa098. https://doi.org/10.1093/nargab/lqaa098 PMID: 33575642
11. Lai Q, Yao S, Zha Y, Zhang H, Zhang H, Ye Y, et al. Deciphering the biosynthetic potential of microbial genomes using a BGC language processing neural network model. Nucleic Acids Res. 2025;53(7):gkaf305. https://doi.org/10.1093/nar/gkaf305 PMID: 40226917
12. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844
13. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15):e2016239118. https://doi.org/10.1073/pnas.2016239118 PMID: 33876751
14. Zhang S, Fan R, Liu Y, Chen S, Liu Q, Zeng W. Applications of transformer-based language models in bioinformatics: a survey. Bioinform Adv. 2023;3(1):vbad001. https://doi.org/10.1093/bioadv/vbad001 PMID: 36845200
15. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, et al. Sequence modeling and design from molecular to genome scale with Evo. Science. 2024;386(6723):eado9336. https://doi.org/10.1126/science.ado9336 PMID: 39541441
16. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. 2019. https://doi.org/arXiv:1907.11692
17. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. 2018. https://doi.org/arXiv:1810.04805
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. arXiv. 2017. https://doi.org/arXiv:1706.03762
19. Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. 2023;51(D1):D603–10. https://doi.org/10.1093/nar/gkac1049 PMID: 36399496
20. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
21. Khosla C, Herschlag D, Cane DE, Walsh CT. Assembly line polyketide synthases: mechanistic insights and unsolved problems. Biochemistry. 2014;53(18):2875–83. https://doi.org/10.1021/bi500290t PMID: 24779441
22. Kim S-Y, Zhao P, Igarashi M, Sawa R, Tomita T, Nishiyama M, et al. Cloning and heterologous expression of the cyclooctatin biosynthetic gene cluster afford a diterpene cyclase and two p450 hydroxylases. Chem Biol. 2009;16(7):736–43. https://doi.org/10.1016/j.chembiol.2009.06.007 PMID: 19635410
23. Zallot R, Oberg N, Gerlt JA. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry. 2019;58(41):4169–82. https://doi.org/10.1021/acs.biochem.9b00735 PMID: 31553576
24. Oberg N, Zallot R, Gerlt JA. EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. J Mol Biol. 2023;435(14):168018. https://doi.org/10.1016/j.jmb.2023.168018 PMID: 37356897
25. Wang J, Lisanza S, Juergens D, Tischer D, Watson JL, Castro KM, et al. Scaffolding protein functional sites using deep learning. Science. 2022;377(6604):387–94. https://doi.org/10.1126/science.abn2100 PMID: 35862514
26. Madani A, McCann B, Naik N, Keskar NS, Anand N, Eguchi RR, et al. ProGen: language modeling for protein generation. Nat Mach Intell. 2023;5:316–29. https://doi.org/10.1038/s42256-023-00650-9
27. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(Web Server issue):W29-37. https://doi.org/10.1093/nar/gkr367 PMID: 21593126
28. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res. 2010;38(Database issue):D211-22. https://doi.org/10.1093/nar/gkp985 PMID: 19920124
29. Broder AZ. On the resemblance and containment of documents. In: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171). 21–9. https://doi.org/10.1109/sequen.1997.666900
30. Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun ACM. 2008;51(1):117–22. https://doi.org/10.1145/1327452.1327494
31. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press. 2008. https://doi.org/10.1017/cbo9780511809071
32. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the International Joint Conference on Artificial Intelligence, 1995. 1137–43.
33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30. https://doi.org/10.5555/1953048.2078195
34. Harris ZS. Distributional Structure. WORD. 1954;10(2–3):146–62. https://doi.org/10.1080/00437956.1954.11659520

35. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

36. Calinski T, Harabasz J. A dendrite method for cluster analysis. Comm in Stats - Theory & Methods. 1974;3(1):1–27. https://doi.org/10.1080/03610927408827101

37. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell. 1979;PAMI-1(2):224–7. https://doi.org/10.1109/tpami.1979.4766909