Designing Biosynthetic Gene Clusters (BGCs) using Conditional Variational Autoencoder (cVAE)

Course: PHA6935 (AI for Drug Discovery)

Institution: University of Florida

Contact: palash.sethi@ufl.edu, josephtsenum@ufl.edu

Group 3 Members: Palash Sethi and Joseph Luper Tsenum

Outline: -

1.1 Introduction

Biosynthetic gene clusters (BGCs) are physically clustered groups of two or more genes that together encode a biosynthetic pathway for producing a secondary metabolite. Refactoring existing BGCs by modifying gene organization can significantly enhance the expression and yield of target compounds. Many natural BGCs remain silent or are poorly expressed under standard laboratory conditions. Developing new regulatory systems or reconstructing these BGCs can activate silent pathways, thereby unlocking their potential for the production of novel metabolites and advancing drug discovery efforts.

Traditionally, BGCs have been discovered by 'find and grind' workflow of extracting from natural sources, chemically isolating, purifying and then testing compounds. Further, various data mining tools such as antiSMASH which use profile hidden Markov models were adopted to identify BGCs. Recently, deep learning methods such as deepBGC and BigCARP have been developed from identification and classification of BGCs and their product type. Deep learning models outperform profile hidden Markov models (HMMs) in BGC classification by capturing the semantic relationships within clusters more effectively. The availability of extensive BGC datasets further provides a robust foundation for training these deep learning models.

In this research, we designed novel biosynthetic gene clusters (BGCs) by using a conditional variational autoencoder (cVAE), BGC-GenVAE, a generative AI model. This approach generates gene clusters by conditioning the model on a specified BGC class as input, enabling the generation of tailored and diverse clusters aligned with desired biosynthetic properties. To the best of our knowledge, this is the first method designed to create new BGCs. Our approach is developed on a sequence-to-sequence cVAE model, offering a cutting-edge framework for generating novel and diverse BGCs.

1.2 Literature Review

Biosynthetic gene clusters (BGCs) are found in bacteria, fungi, and some plant species. Common types of BGCs include clusters for nonribosomal peptide synthetase (NRPS),

polyketide synthase (PKS), terpenes, and ribosomally synthesized and post-translationally modified peptides (RiPPs). These secondary metabolites are a valuable source of small molecule drug candidates with applications in antimicrobial drugs, anticancer therapies, and immunomodulatory agents (Hannigan et al.,). BGCs typically contain multiple operons that are regulated in a coordinated manner. These enzyme systems function as "molecular assembly lines," orchestrating the sequential construction of specific types of molecules with remarkable precision.

To our knowledge, only BGC detection and product type classification methods exist. antiSMASH, a bioinformatics tool based on profile hidden Markov models (pHMMs), was developed to facilitate the precise characterization of biosynthetic gene clusters (Medema et al., 2011). It enables a comprehensive understanding of their biosynthetic potential, evolutionary context, and regulatory mechanisms within microbial genomes. By aligning identified gene cluster regions with their closest relatives in a robust database of known clusters, antiSMASH provides insights into gene cluster relationships. Additionally, it integrates a wide range of secondary metabolite-specific gene analysis tools into a unified, interactive interface for streamlined analysis. DeepBGC is a deep learning-based model using a bi-directional LSTM network designed for biosynthetic gene cluster (BGC) identification. It utilizes pFAM domains to represent proteins, converting these domains into vectors using the pFAMtoVec technique (Hannigan et al., 2019). It incorporates a binary start/end feature to capture domain context. For each domain, DeepBGC predicts the probability of its presence within a BGC, thereby providing a powerful tool for analyzing and identifying biosynthetic gene clusters. e-DeepBGC is an advanced model built on the foundation of DeepBGC, incorporating additional domain summary information extracted from PFAM (Liu et al., 2022). This includes details such as domain annotations (e.g., PF00001: 7 transmembrane receptor, rhodopsin family), clan affiliations (e.g., PF00001: CL0192), and related Pfam domains (e.g., PF00001: PF05296, PF10320, PF10323, PF10324, PF10328, PF13853g). These enhancements result in a reduced false positive rate and improved sensitivity for detecting biosynthetic gene clusters (BGCs). BiGCARP, developed by Martinez et al. in 2023, employs a convolutional self-supervised model for pre-training on biosynthetic gene cluster (BGC) data. The model appends a BGC class token at the beginning of the input sequence to provide class-specific context. Proteins are represented by their constituent

domains, which are vectorized using the ESM1b embedding technique, enabling precise characterization and analysis of BGCs.

Supervised deep learning techniques have demonstrated significant success in addressing various recognition tasks in machine learning and computer vision. While these models excel at approximating complex many-to-one functions given sufficient training data, they lack probabilistic inference capabilities, which limits their effectiveness in modeling complex structured output representations. Conditional Variational Autoencoders (cVAEs), a class of deep generative models, address this limitation by employing Gaussian latent variables to represent structured output variables, enabling more probabilistically informed predictions. (Sofi et al., 2023). To generate drug molecules with specific properties, a VAE is insufficient. cVAE models are widely used for de novo drug molecule generation due to their ability to generate molecular structures with specified properties. By conditioning both the encoder and decoder on specific properties, the model learns conditional distributions, enabling the generation of drug molecules with desired properties (e.g., solubility, partition coefficient, growth inhibition), for more targeted design.

1.3 Results

1.3.1 Synthetic BGCs generated by BGC-GenVAE resemble naturally occurring BGCs

We train BGC-GenVAE (Fig 1), a conditional variational autoencoder based on a transformer sequence-to-sequence model, on 2050 samples of T1PKS and T3PKS BGC classes (Fig 2). These classes were chosen due to their comparable dataset sizes and shared protein domains, despite differing in complexity—T1PKS are large, modular, multifunctional enzymes, while T3PKS are small, homodimeric, and lack modularity. BGC-GenVAE treats each PFAM domain name as a token and extends its vocabulary with tokens such as [CLS], [SEP], and [EOS]. Proteins in a BGC are represented by their constituent PFAM domains and separated by [SEP], with a [CLS] token and a BGC product type token prepended to the sequence. The latent embedding for [CLS] is passed through the reparameterization trick to ensure it follows a normal distribution. This latent embedding is then appended to the decoder input, which consists of the BGC product type and the input BGC sequence, ending with an [EOS] token. BGC-GenVAE is

trained for 50 epochs using a loss function that combines reconstruction loss and KL divergence. After training, synthetic BGCs are generated by sampling a normally distributed vector *z* and a BGC product class token, which are fed into the decoder of BGC-GenVAE.

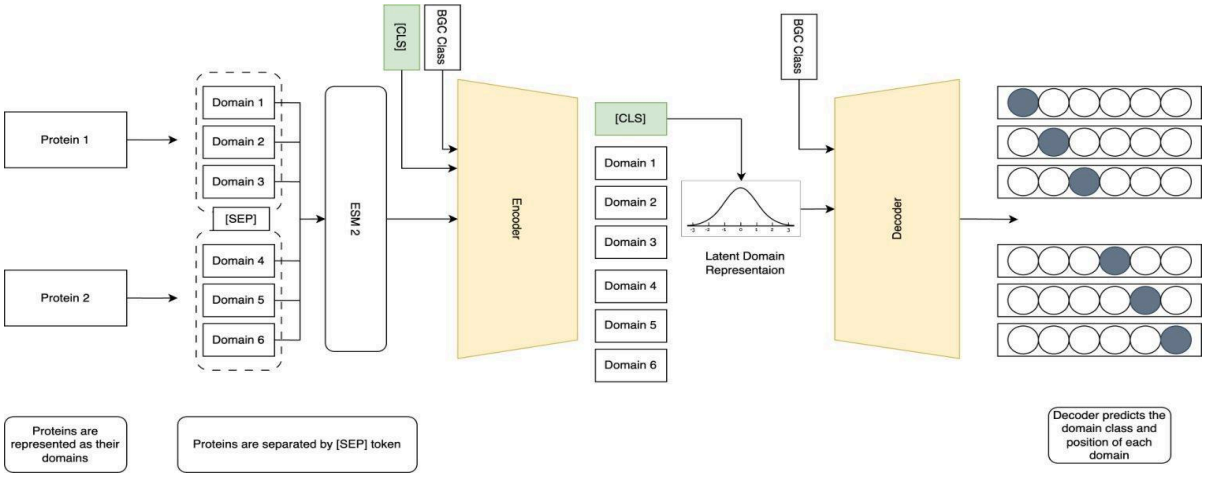

Figure 1 BGC-GenVAE model, where proteins, represented as their constituent domains and separated by a [SEP] token, are processed through an ESM2 model and an encoder to generate a latent domain representation conditioned on a BGC class. The decoder then predicts the domain class and position for each domain based on this representation, enabling domain reconstruction.

To validate the quality and biological relevance of synthetic BGCs generated by BGC-GenVAE, we propose a novel use of an oracle model based on ESM1b embeddings. Previous studies have primarily focused on the detection and classification of naturally occurring BGCs from the MIBIG dataset and BGCs identified through antiSMASH. Our analysis reveals that ground truth BGCs, represented by their PFAM domain embeddings derived from ESM1b, can be efficiently classified into T1PKS and T3PKS classes with high accuracy (Figure 3a). Using a logistic regression classifier trained on ESM1b-derived PFAM domain embeddings—averaged across the domain dimension—we demonstrate accurate classification of ground truth BGCs into T1PKS and T3PKS categories. Extending this approach, we applied the classifier to synthetic BGCs generated by BGC-GenVAE. Remarkably, synthetic T3PKS BGCs achieved an accuracy of 72.84%, while synthetic T1PKS BGCs achieved an accuracy of 83.97% (Fig 3c). These results suggest that BGCs generated by BGC-GenVAE closely resemble ground truth BGCs in terms of their ESM1b PFAM domain embeddings. This demonstrates the ability of BGC-GenVAE to

generate biologically relevant and classifiable synthetic BGCs, making it a useful generative tool in BGC research.

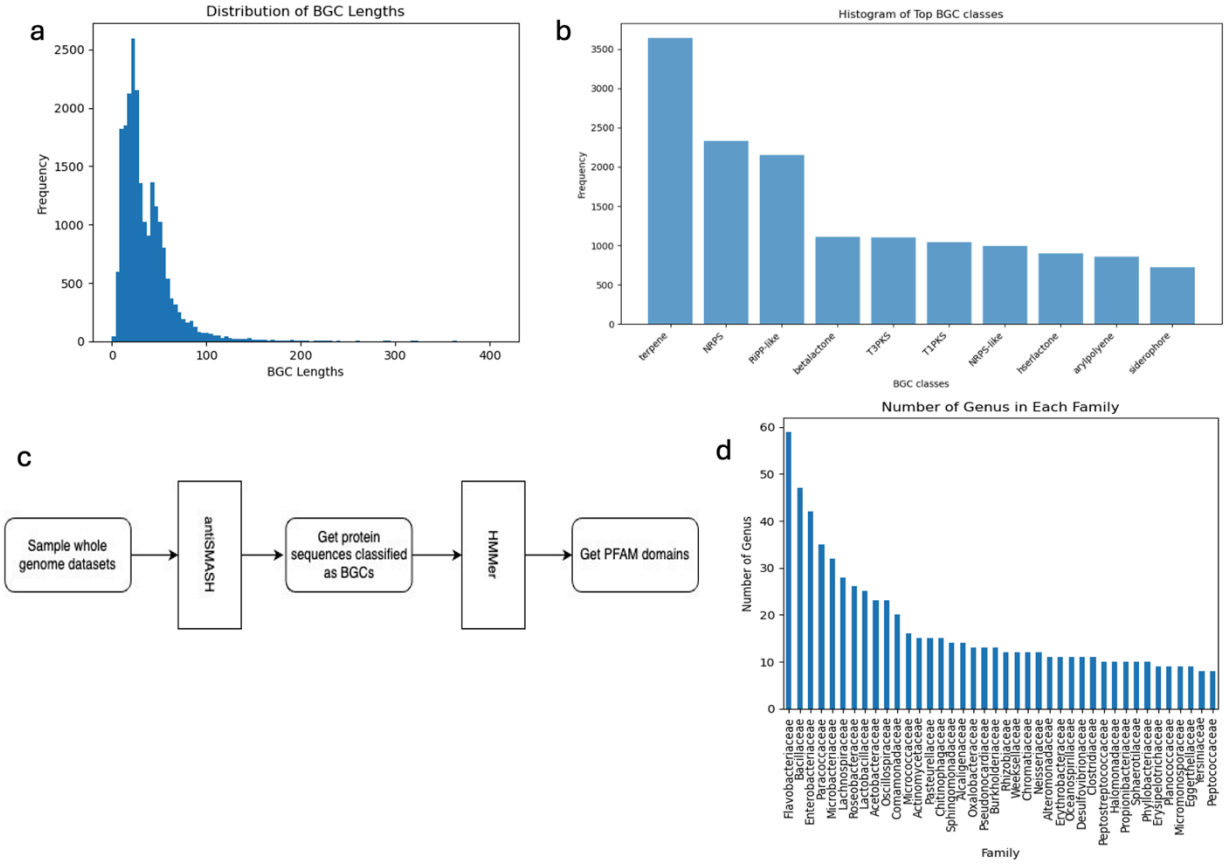

Figure 2 Data analyses related to BGCs **Fig. a:** The histogram displays the distribution of BGC lengths, showing that most BGCs have relatively short lengths **Fig. b:** Frequency of top BGC classes **Fig. c:** Data generation workflow, PFAM domains were extracted from whole genome datasets using antiSMASH for BGC detection, followed by HMMER for domain identification. **Fig. d:** Number of genus in different microbial families.

1.3.2 Functional Domain Analysis Reveals Conservation and Novelty in Synthetic BGCs

Functional domains within a BGC are essential for producing secondary metabolites, and these domains must remain conserved to maintain BGC functionality. To assess whether the synthetic BGCs generated by BGC-GenVAE preserve this critical conservation, we conducted a single-domain and pairwise-domain correlation analysis between naturally occurring BGCs and synthetic BGCs (Fig 4a and Fig 4b). First, we calculated the probability of individual domain occurrences in the training dataset and extended this to pairwise domain occurrence probabilities

for both the training and synthetic datasets. Our analysis revealed that 40% of pairwise domains present in the synthetic dataset are novel, indicating that BGC-GenVAE can generate previously unseen domain combinations. This ability to discover novel pairwise domains highlight the generative model's capacity for innovation while still producing biologically meaningful BGCs.

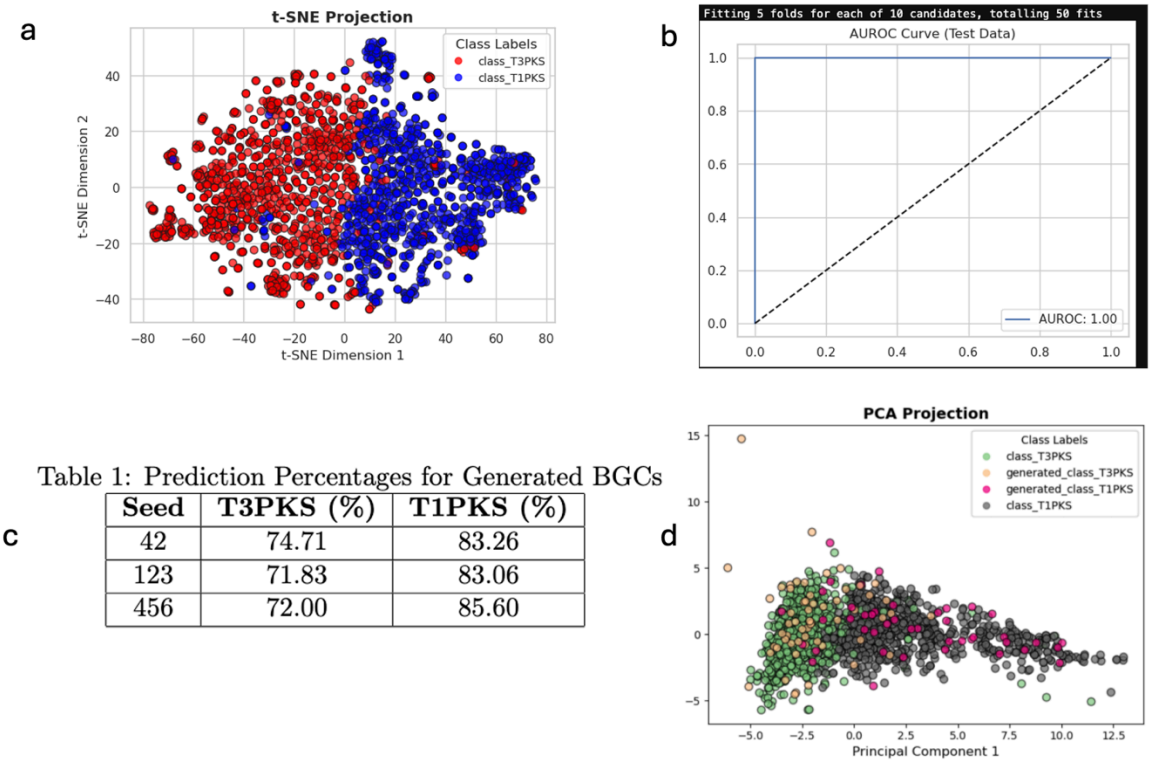

Table 1: Prediction Percentages for Generated BGCs

Seed	T3PKS (%)	T1PKS (%)
42	74.71	83.26
123	71.83	83.06
456	72.00	85.60

Figure 3 ESM1b embeddings analysis for training and synthetic data **Fig. a:** t-SNE visualization of ESM1b embeddings for train-set BGCs showing the separation between two classes, T3PKS and T1PKS. **Fig. b:** AUROC curve for a logistic regression model fitted on ESM1b embeddings for train-set BGCs. **Fig. c:** Prediction accuracies of the generated BGCs for T3PKS and T1PKS classes across different random seeds. The percentages demonstrate consistent performance, with higher prediction accuracy for T1PKS compared to T3PKS. **Fig. d:** PCA projection of both original and generated BGCs. The overlap and separation between the generated and actual data points for T3PKS and T1PKS classes indicates BGC-GenVAE's ability to generate realistic representations.

For the remaining individual and pairwise domains shared between the training and synthetic datasets, we observed a high Pearson correlation (0.97 and 0.90, respectively), demonstrating that the domain frequencies in the synthetic BGCs closely resemble those in the training data. These results collectively suggest that while BGC-GenVAE generates synthetic BGCs with conserved functional domains, it also introduces novel domain combinations, potentially expanding the diversity of secondary metabolite production. We further analyze the top 10 pairwise domains in the training and synthetic BGCs based on their frequency. Among these, we

identify a set of domains essential for maintaining BGC functionality in T1PKS and T3PKS clusters. Notably, we observe the formation of a domain interaction graph (Fig 4c), which reveals the intricate interrelationships between domains. This graph underscores the structural and functional coherence of both natural and synthetic BGCs, highlighting that BGC-GenVAE preserves critical domain associations necessary for producing valid secondary metabolites.

1.4 Discussion

To the best of our knowledge, this is the first method developed to design novel biosynthetic gene clusters (BGCs). Our approach introduces enhanced domain representations specifically tailored for BGCs, which have not been utilized in prior deep learning methods. This methodology is built upon a transformer and attention-based conditional variational autoencoder (cVAE) architecture, providing a cutting-edge framework for generating new and diverse BGCs. To validate the designed BGCs, predicted domains were classified based on ESM1b embeddings. Additionally, validation was conducted by performing correlation analysis between the domains of train-set and generated BGCs. Predicted domains were also queried against their gene ontology information to confirm functional similarity with the replaced domains. The approach remained flexible and open to incorporating additional validation methods.

1.5 Limitations

BGC-GenVAE was trained on a limited set of data, consisting only of two product classes. Further, the data used for training is derived from antiSMASH, a rule-based method for BGC detection, which may restrict the diversity and representativeness of the dataset. Furthermore, additional evaluation metrics need to be developed to ensure the accurate classification and validation of BGCs generated by the seq2seq model.

1.6 Future Work

In the short term, we aim to develop a new metric based on multiple sequence alignment (MSA) of PFAM domains to evaluate not only the presence of functional domains in synthetic BGCs but also their correct sequential order. Additionally, we plan to expand the training dataset by

incorporating more naturally occurring BGCs to improve the model's robustness and generalizability. An ablation study will also be conducted using random domain embeddings as a baseline to assess the impact of using ESM1b embeddings.

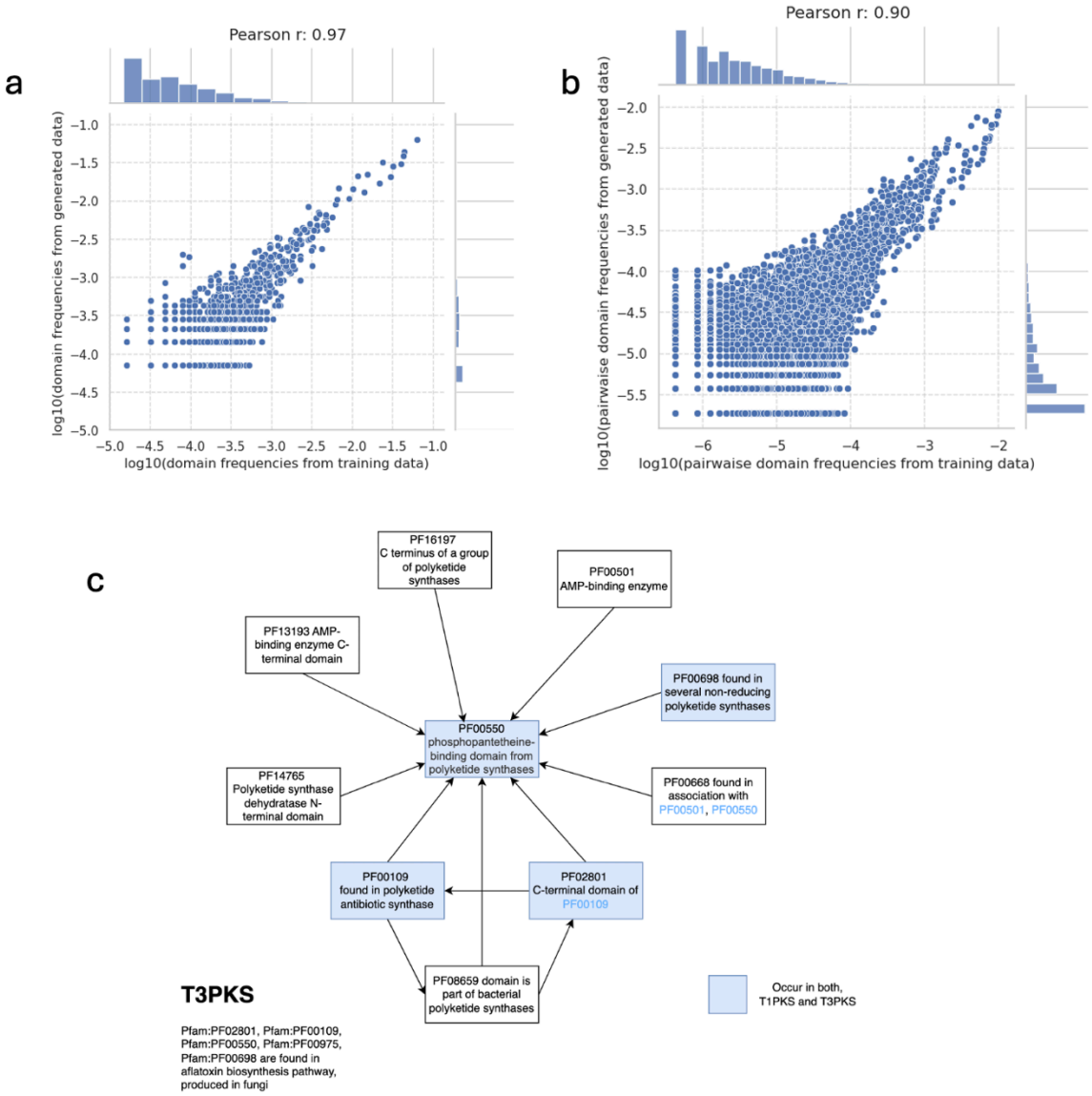

Figure 4 Functional domain conservation and novelty in synthetic BGCs. **Fig. a:** This scatter plot compares the log-transformed frequencies of individual domains between the generated and training datasets, showing a strong positive correlation with a Pearson coefficient of 0.97 for T1PKS. **Fig. b:** Similar to Fig (a), this scatter plot compares pairwise domain co-occurrence frequencies between the generated and training datasets, with a Pearson coefficient of 0.90. **Fig. c:** This diagram illustrates the functional relationships between conserved and novel

domains within T3PKS and T1PKS classes. Central domains like PF00550 (phosphopantetheine-binding domain) connect to other functionally significant domains, highlighting their role in core biosynthetic pathways.

In the long term, our goal is to develop a model capable of modifying plant BGCs to closely mimic yeast BGCs, thereby enabling the "yeastizing" of plant enzymes for use in microbial cell factories. This approach has the potential to facilitate the integration of plant-derived pathways into microbial hosts, enhancing biosynthesis capabilities.

1.7 Conclusion

This study successfully developed and demonstrated novel approaches to biosynthetic gene cluster (BGC) design using transformer-based and attention-based conditional variational autoencoder (cVAE) architectures. Through the introduction of enhanced domain representations specific to BGCs and addressing limitations in traditional rule-based methods like antiSMASH, these methods provide robust frameworks for generating new and diverse BGCs. However, challenges remain, including limited training data for the model, a restricted range of BGC classes, and the need for additional evaluation metrics to ensure accurate classification and validation of generated BGCs. These findings represent a significant advancement in generative AI for BGC design, with promising applications in drug discovery and natural product engineering.

1.8 Materials and Methods

Datasets

A total of 32,000 whole bacterial genomes spanning multiple taxa were downloaded from NCBI. To ensure taxonomic diversity, five genomes were randomly sampled from each genus. BGC genomic regions and their corresponding product types were identified using antiSMASH. The genomic regions were then annotated with PFAM domains using HMMer and pfam_scan, resulting in the identification of 20,000 BGCs across 55 product classes.

Due to missing amino acid sequences for some PFAM domains, we supplemented the dataset with amino acid sequences from the bigCARP paper, which provides a comprehensive set of

PFAM domains and their sequences. PFAM domains identified in BGCs that did not match the bigCARP dataset were removed to ensure consistency in annotations. This cleaned dataset was subsequently used for pre-processing and BGC-GenVAE training.

The data was preprocessed as follows:

Step	Description
1. Load PFAM Sequences	Input the sequences from the provided dataset.
2. Preprocess Sequences	Clean and filter the sequences to prepare for embedding.
3. Apply ESM Model	Use the ESM model to generate embeddings for sequences.
4. Extract Embeddings	Obtain the numerical embeddings for downstream tasks.
5. Save Embeddings	Store the embeddings in a structured format for modeling.

Figure 5: Data preprocessing steps for PFAM embeddings and esm1b embeddings.

The MiBIG dataset contained a total of 3,685 PFAM IDs, while the final_pfams.fasta file included 19,092 PFAM IDs, with 3,685 of these being common between the two datasets. The shape of the PFAM embeddings (pfam_dense) was (2024, 19092), and the shape of the ESM embeddings (esm_dense) was (19450, 1280). The MiBIG dataset contained 2,024 rows. To align the PFAM embeddings with the ESM embeddings, 17,426 rows were padded, resulting in a concatenated embedding shape of (19450, 20372).

Model training

BGC-GenVAE is a sequence-to-sequence based conditional variational autoencoder, which treats each PFAM domain name as a token and extends its vocabulary with tokens such as [CLS], [SEP], and [EOS]. Proteins in a BGC are represented by their constituent PFAM domains and separated by [SEP], with a [CLS] token and a BGC product type token prepended to the sequence. The latent embedding for [CLS] is passed through the reparameterization trick to ensure it follows a normal distribution. This latent embedding is then appended to the decoder input, which consists of the BGC product type and the input BGC sequence, ending with an [EOS] token. BGC-GenVAE is trained for 50 epochs using a loss function that combines reconstruction loss and KL divergence. After training, synthetic BGCs are generated by sampling

a normally distributed vector *z* and a BGC product class token, which are fed into the decoder of BGC-GenVAE. The model was developed in pytorch and trained on 1 A100 GPU.

Transformer-based and attention-based conditional variational autoencoder (cVAE) architectures were also employed in this work, methods and results about which were included in our presentation slides.

1.9 References

1. Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Breitling, R., & Takano, E. (2011). antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr466
2. Sofi, M. A., Singh, D., & Teli, T. A. (2023). Attention-based Conditional VAE for Lung Cancer Drug Generation. *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 924–928.
3. Zhai, S. (2021). Journal of Manufacturing Systems. https://doi.org/10.1016/j.jmsy.2021.02.0006
4. Masoodi, F., Quasim, M., Bukhari, S., Dixit, S., & Alam, S. (2023). Applications of Machine Learning and Deep Learning on Biological Data. *CRC Press*.
5. Dollar, O., Joshi, N., Beck, D. A. C., & Pfaendtner, J. (2021). Attention-based generative models for de novo molecular design. *Chemical Science*, 12(24), 8362–8372.
6. Cheng, J., Dong, L., & Lapata, M. (2016). Long Short-Term Memory-networks for machine reading. *arXiv [cs.CL]*.
7. Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *CoRR*.
8. Hannigan, G. D., et al. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, 47, e110.
9. Liu, M., Li, Y., & Li, H. (2022). Deep Learning to Predict the Biosynthetic Gene Clusters in Bacterial Genomes. *Journal of Molecular Biology*, 434, 167597.

10. Rios-Martinez, C., Bhattacharya, N., Amini, A. P., Crawford, L., & Yang, K. K. (2023). Deep self-supervised learning for biosynthetic gene cluster detection and product classification. *PLOS Computational Biology*, 19, e1011162.
11. Palaniappan, K., et al. (2020). IMG-ABC v.5.0: An update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Research*, 48, D422–D430.
12. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D., & Medema, M. H. (2021). BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *GigaScience*, 10, giaa154.
13. Kwon, M. J., Steiniger, C., Cairns, T. C., Wisecaver, J. H., Lind, A. L., Pohl, C., Regner, C., Rokas, A., & Meyer, V. (2021). Beyond the Biosynthetic Gene Cluster Paradigm: Genome-Wide Coexpression Networks Connect Clustered and Unclustered Transcription Factors to Secondary Metabolites.

2.0 Data Availability

The datasets used in this study are publicly available:

- AntiSMASH dataset: https://antismash-dbv2.secondarymetabolites.org/#!/start
- **Bi**osynthetic **G**ene **CARP** (BiGCARP) dataset: https://zenodo.org/records/6857704

2.1 Contributions and acknowledgements

JLT worked on manuscript writing and two additional models that were presented in the accompanying slides but were not included in the written report due to space constraints. These models are the bgc-cVAE ESM1b Model and the attention-based bgc-cVAE model. **PS** worked on manuscript, study design, data generation, BGC-GenVAE scripting, training and analysis. We

would like to thank all the professors of PHA6935 for their useful feedback on the study. We would also like to thank Dr. Juannan Zhou and Dr. Xiao Fan for their useful advice.