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

PHA6935 (AI for Drug Discovery)

Group 3 Members: Palash Sethi and Joseph L. Tsenum

Outline

- Objectives + Recap
- Study Design
- Datasets
- Modeling
- Limitations
- Future Work
- References
- Q/A

Objectives

- Design novel BGCs using conditional variational autoencoder – a model based on deep learning that generates gene clusters when given a BGC class as input.

Biosynthetic Gene Clusters (BGCs)

What are they?

- Found in bacteria, fungi and some plant species, BGCs are physically **clustered groups of two or more genes** that together encode a biosynthetic pathway for producing a secondary metabolite. Common types include clusters for nonribosomal peptide synthetase (NRPS), polyketide synthase (PKS), terpenes, and ribosomally synthesized and post-translationally modified peptides (RiPPs).
- These secondary metabolites represent a rich reservoir of small molecule drug candidates utilized as antimicrobial drugs, anticancer therapies, and immunomodulatory agents. (Hannigan et al.)

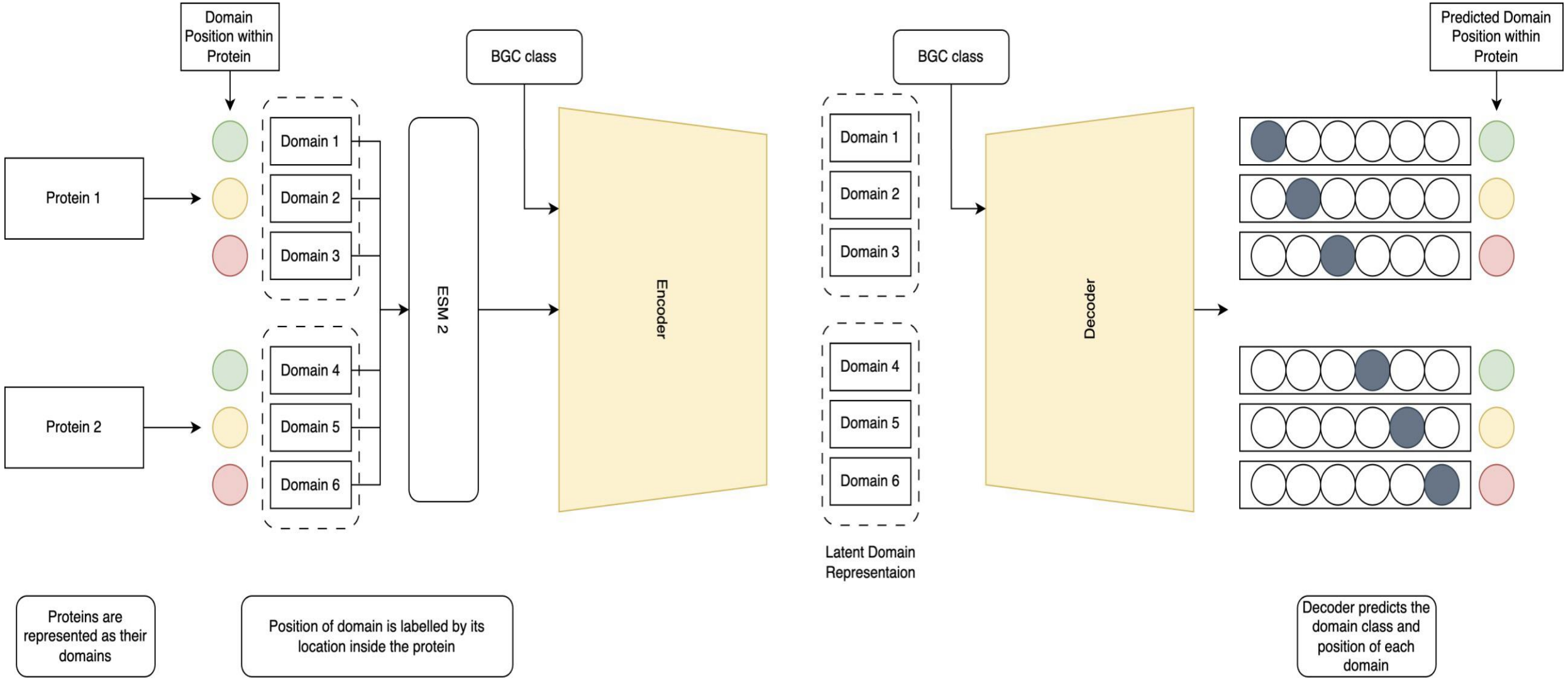
How do they work – Biologically (Operons, transcription)

- BGCs often contain several operons that are coordinately regulated.

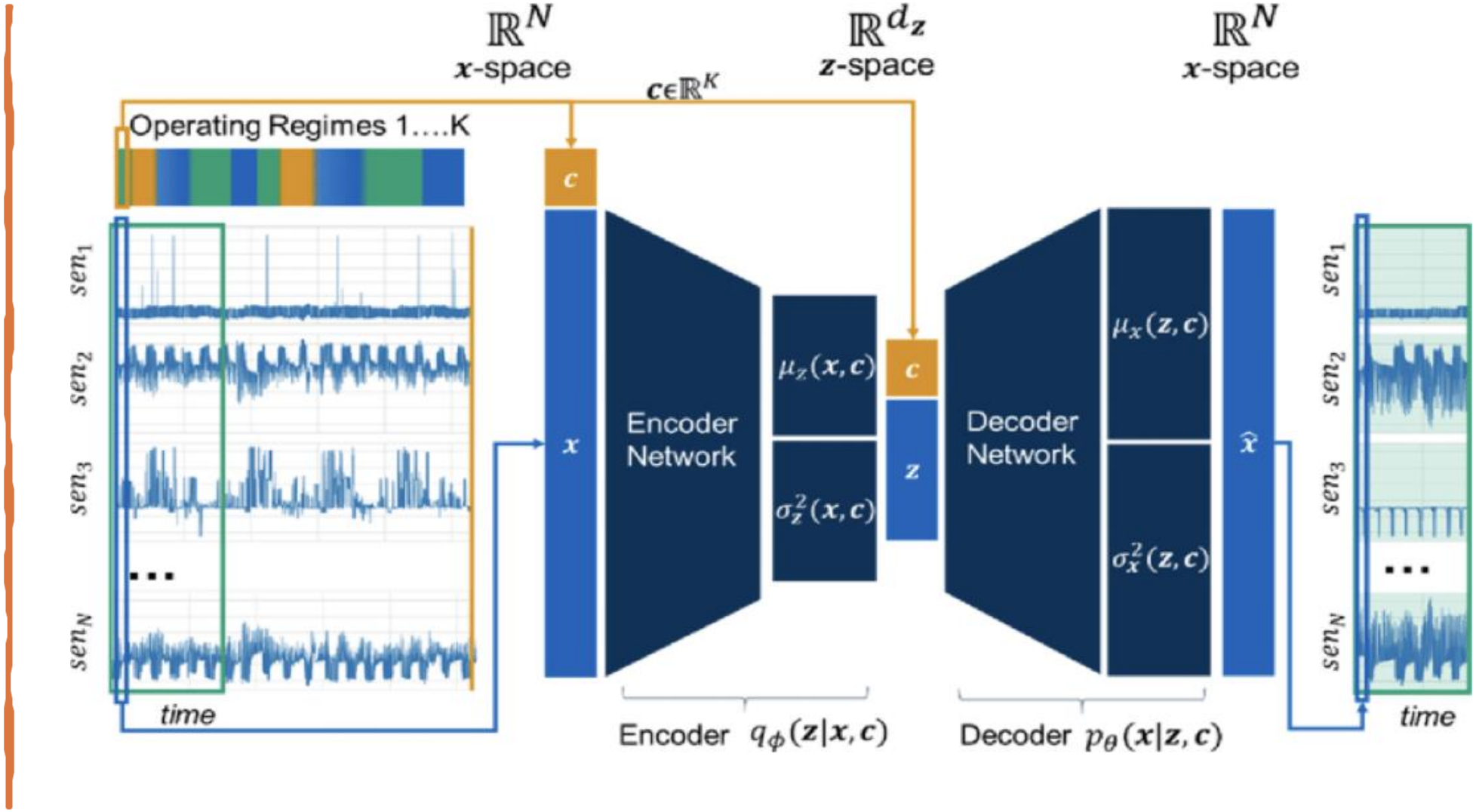
Why do we need to design BGCs

- Refactoring existing BGCs by modifying regulatory elements, promoters, or gene organization can enhance the expression and yield of desired compounds.
- Many natural BGCs are silent or poorly expressed under standard laboratory conditions. Designing new regulatory systems or reconstructing BGCs can activate these silent pathways, unlocking their potential for producing novel metabolites and drug discovery.

BGCs design using cVAE



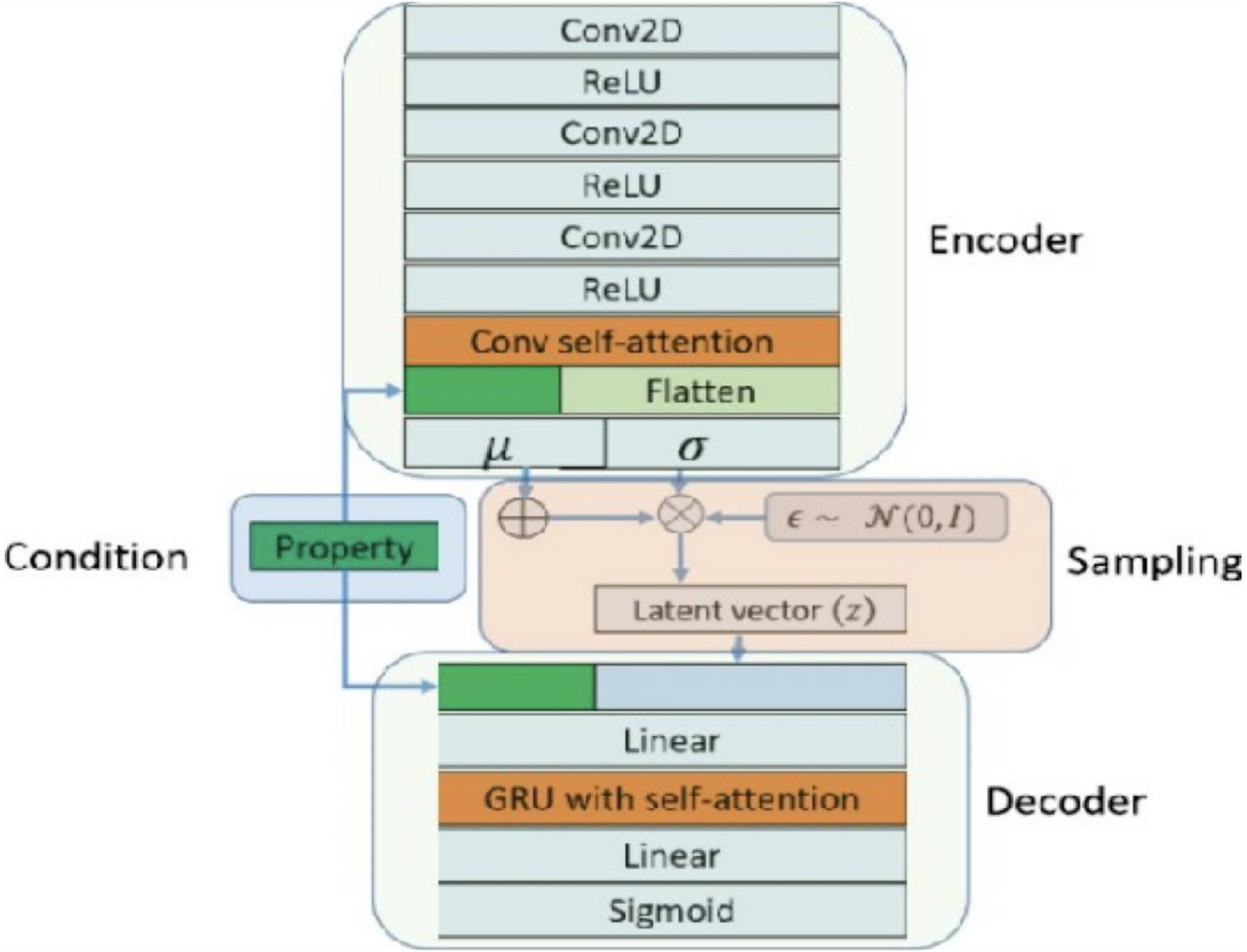
cVAE



Source: http://dx.doi.org/10.1016/j.jmsy.2021.02.006

Attention-based cVAE

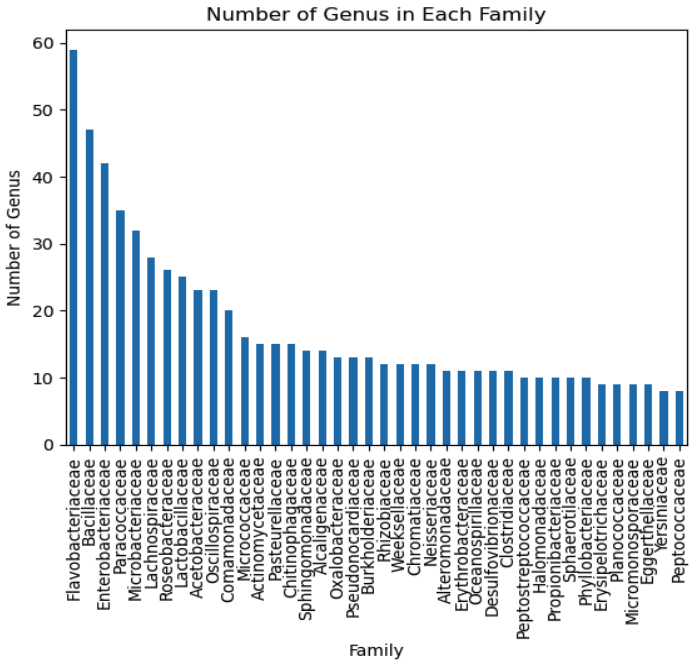
Sofi et al built a model using 2D convolutional layers in the encoder, followed by a convolutional-based self-attention layer (SAL), while employing an attention- based Recurrent Neural Network (RNN) for the decoder

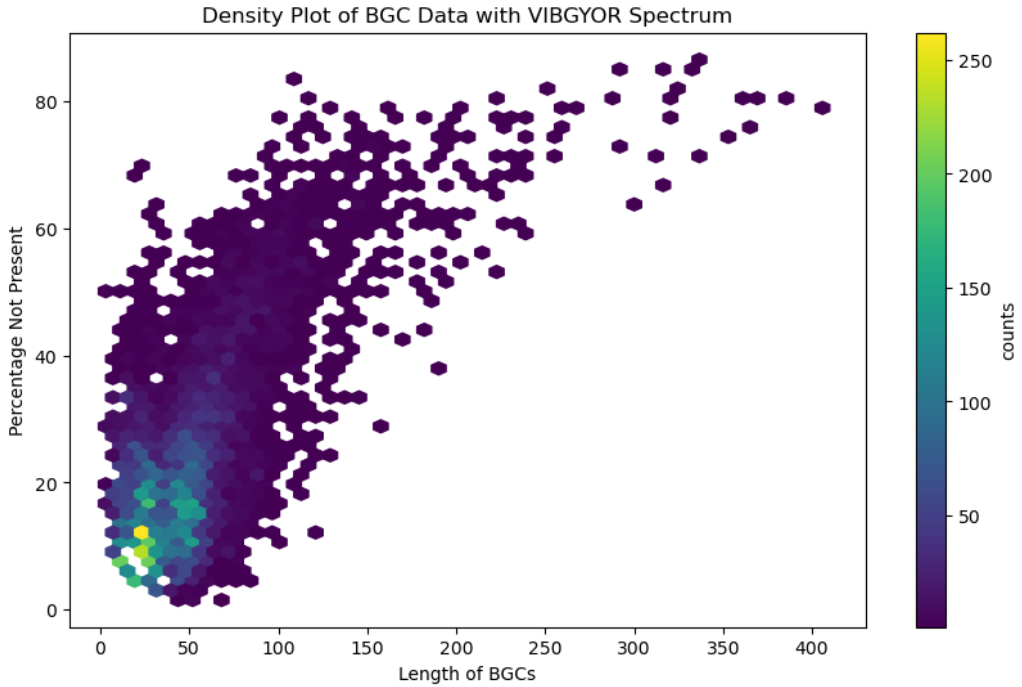


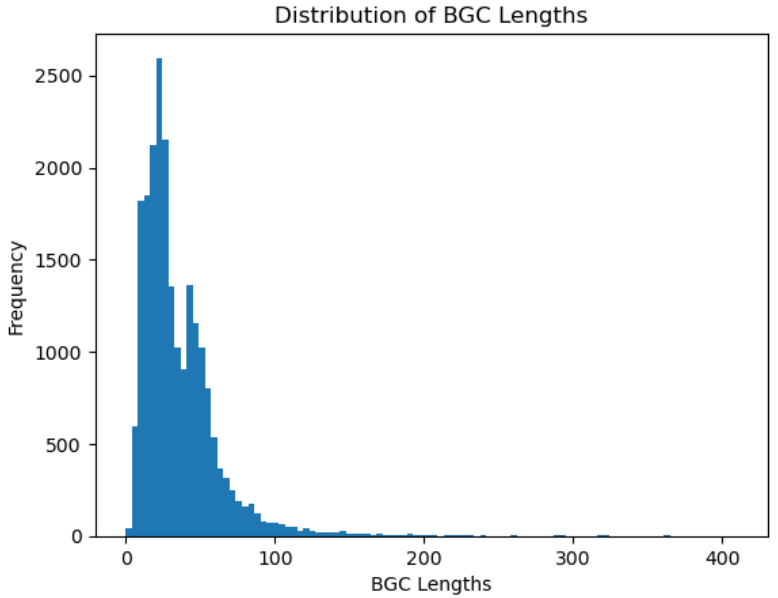
Sofi et al. (2023)

Datasets

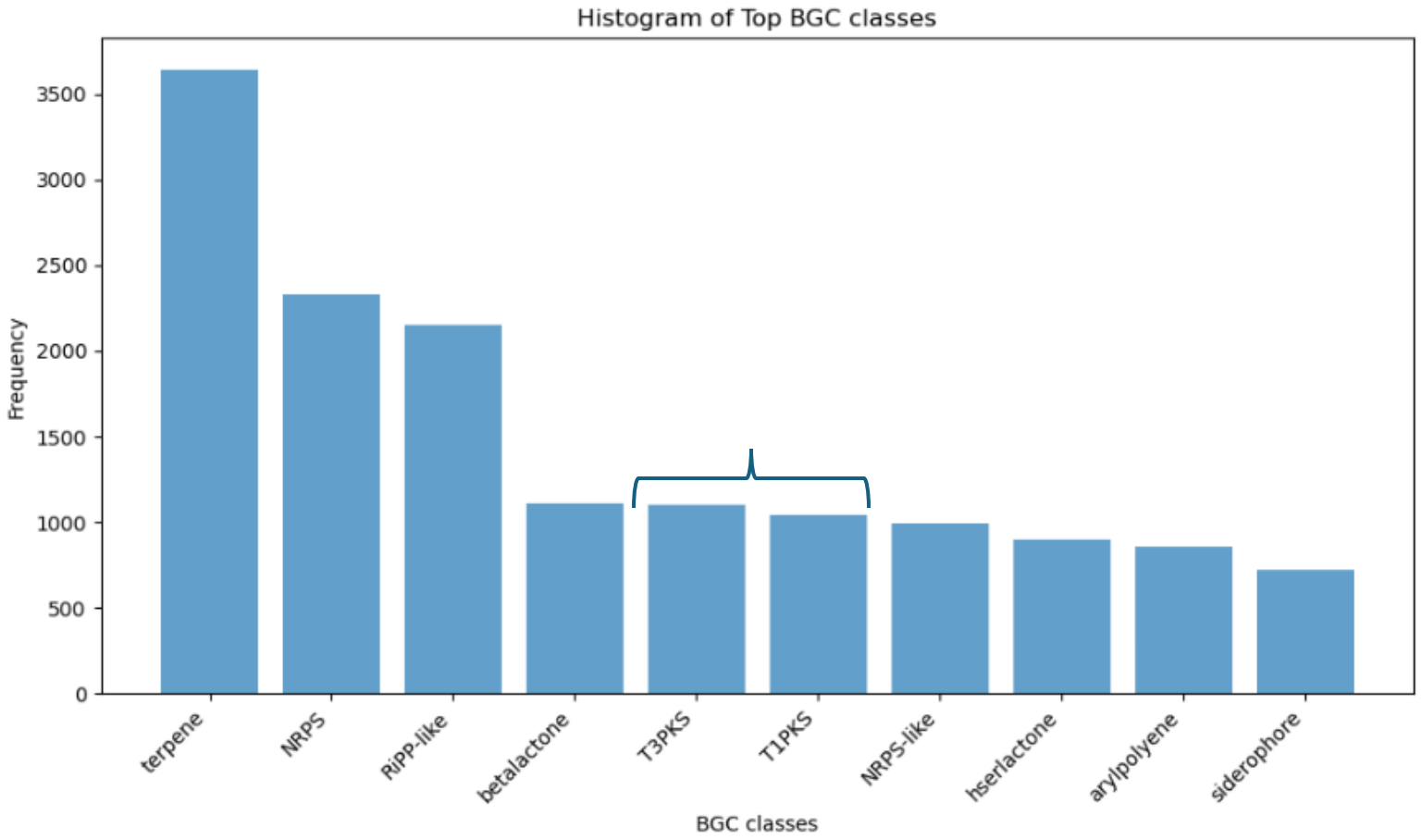

Sample whole genome datasets
antiSMASH
Get protein sequences classified as BGCs
HMMer
Get PFAM domains


Number of Genus in Each Family
Number of Genus
Family
Flavobacteriaceae
Bacillaceae
Enterobacteriaceae
Paracoccaceae
Microbacteriaceae
Lachnospiraceae
Roseobacteraceae
Lactobacillaceae
Acetobacteraceae
Oscillospiraceae
Comamonadaceae
Micrococcaceae
Actinomycetaceae
Pasteurellaceae
Chitinophagaceae
Sphingomonadaceae
Alcaligenaceae
Oxalobacteraceae
Pseudonocardiaceae
Burkholderiaceae
Rhizobiaceae
Weeksellaceae
Chromatiaceae
Neisseriaceae
Alteromonadaceae
Erythrobacteraceae
Oceanospirillaceae
Desulfovibrionaceae
Clostridiaceae
Peptostreptococcaceae
Halomonadaceae
Propionibacteriaceae
Sphaerotilaceae
Phyllobacteriaceae
Erysipelotrichaceae
Planococcaceae
Micromonosporaceae
Eggerthellaceae
Yersiniaceae
Peptococcaceae


Density Plot of BGC Data with VIBGYOR Spectrum
Percentage Not Present
Length of BGCs
counts


Distribution of BGC Lengths
Frequency
BGC Lengths

T1PKS-T3PKS ~2000 Samples



Type I and Type III polyketide synthases (PKSs) differ mainly in their structure and mechanism:

- **Type I PKS**: Large, modular, multifunctional enzymes. Each module is responsible for a single elongation step, and the modules are arranged in a linear sequence. Found in *bacteria*.
- **Type III PKS**: Small, homodimeric enzymes. They lack modularity and catalyze iterative chain extension using a single active site. Found in *plants, fungi, and bacteria*.

Modeling

- Bgc-cVAE ESM1b Model
- Attention-Based bgc-cVAE Model
- Seq2Seq bgc-cVAE

Bgc-cVAE ESM1b Model

Data Preprocessing

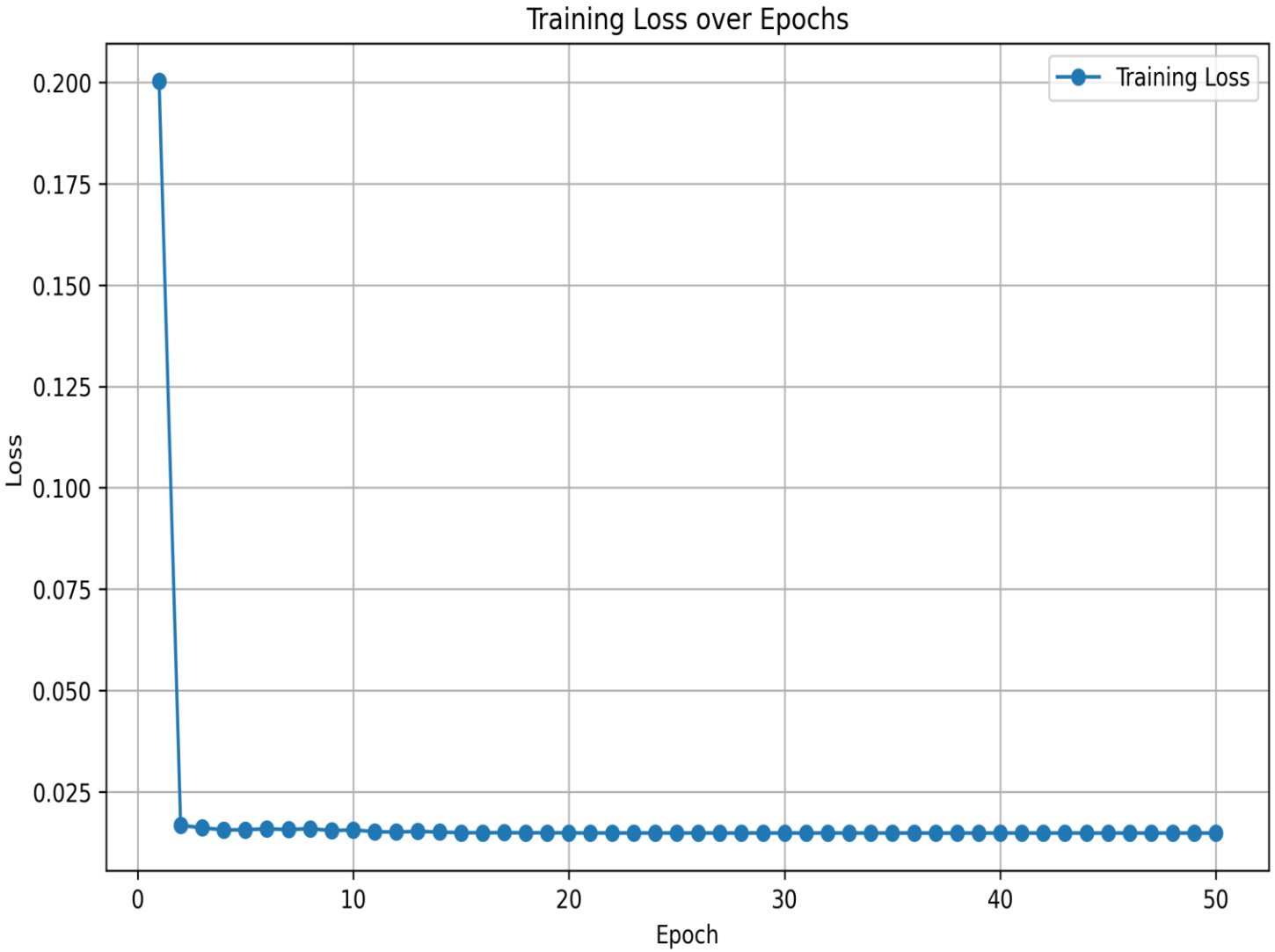
Step	Description
1. Load PFAM Sequences	Input the sequences from the provided dataset.
2. Preprocess Sequences	Clean and filter the sequences to prepare for embedding.
3. Apply ESM Model	Use the ESM model to generate embeddings for sequences.
4. Extract Embeddings	Obtain the numerical embeddings for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

- Total PFAM IDs in MiBIG: 3685
- Total PFAM IDs in final_pfams.fasta: 19092
- Common PFAM IDs: 3685

- Shape of PFAM embeddings (pfam_dense): (2024, 19092)
- Shape of ESM embeddings (esm_dense): (19450, 1280)
- Number of rows in MiBIG dataset: 2024
- Padding PFAM embeddings with 17426 rows to match ESM embeddings…
- Concatenated embedding shape: (19450, 20372)

Model Architecture and Parameters

	Component	Details
0	Model Type	Conditional Variational Autoencoder (cVAE)
1	Encoder Input Dimensions	(batch_size=64, input_dim=1280)
2	Latent Space Dimensions	latent_dim=128
3	Encoder Layers	1024 -> 256 -> latent_dim
4	Decoder Input Dimensions	(latent_dim=128 + condition_dim=10)
5	Decoder Output Dimensions	(batch_size=64, output_dim=1280)
6	Embedding Data Shape	(19450, 1280)
7	Condition Data Shape	(19450, 3)
8	Number of BGC Classes	3
9	Activation Functions	ReLU (encoder), Sigmoid (decoder)
10	Latent Sampling Equation	z = mu + exp(0.5 * logvar) * epsilon
11	Loss Function	MSE + KL Divergence
12	Optimizer	Adam Optimizer
13	Learning Rate	0.001
14	Batch Size	64
15	Epochs	50
16	Gradient Clipping	1.0
17	Learning Rate Scheduler	ReduceLROnPlateau (factor=0.1, patience=5)
18	Training Focus	MLP layers for both Encoder/Decoder, no CNN/RNN



Loss Function Components:

- **Reconstruction Loss (MSE)**: Measures how well the reconstructed output matches the input embeddings.

- **KL Divergence**: Encourages the latent space to follow a Gaussian distribution, facilitating smooth sampling.

- **Combined Loss**:

Total Loss = Reconstruction Loss + β × KL Divergence

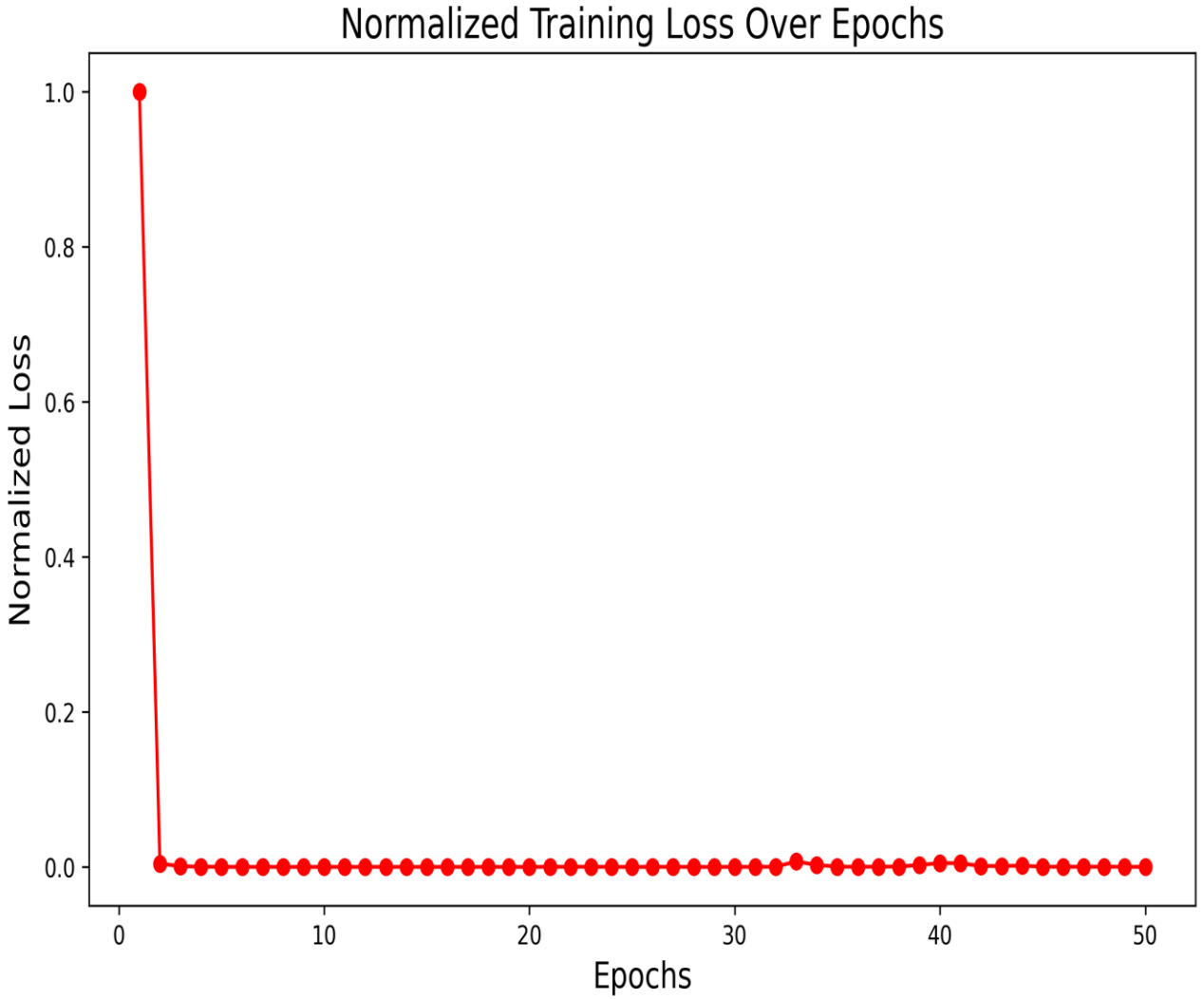
(where β = 1.0 in this case unless otherwise specified).

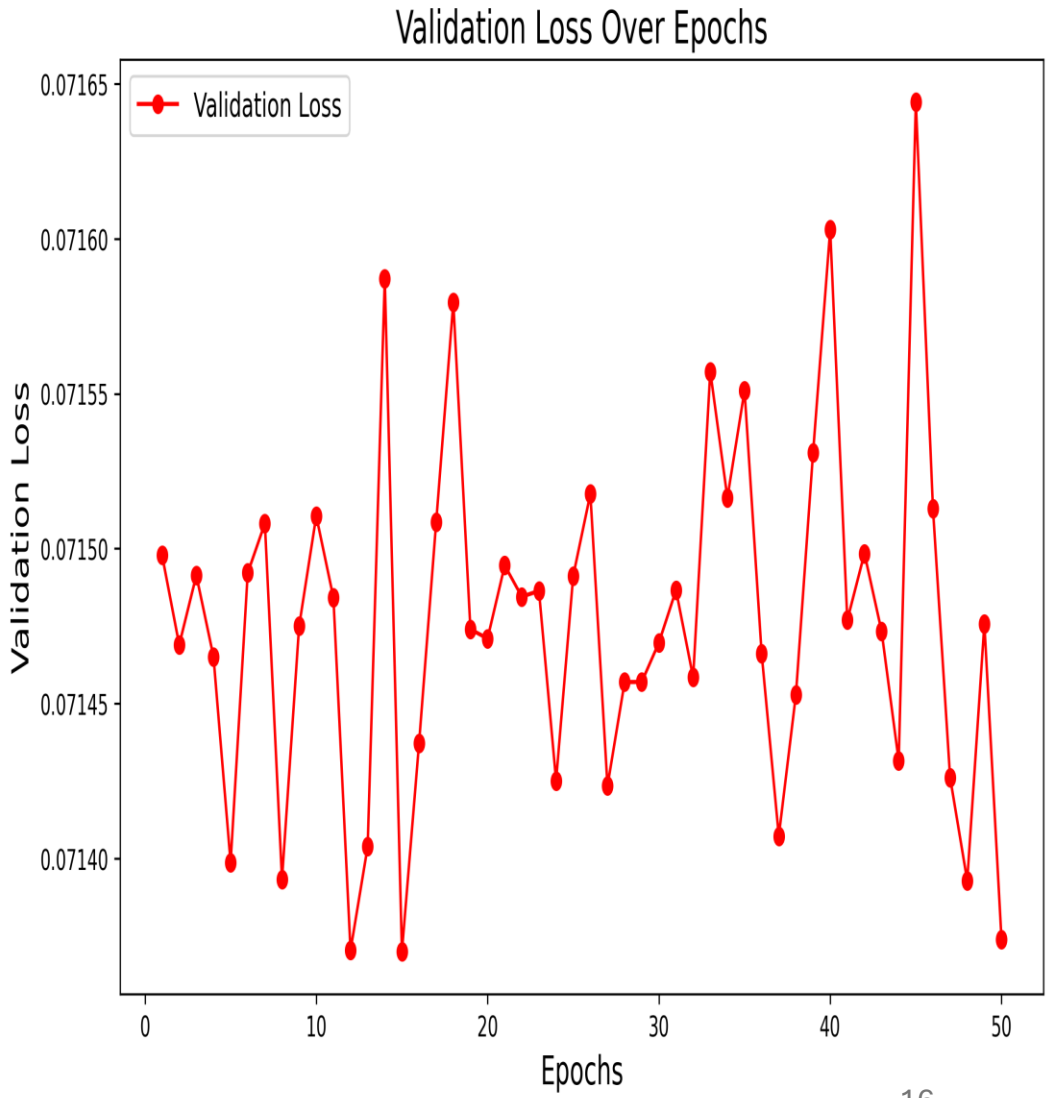
Attention-Based bgc-cVAE Model

Model Architecture and Results

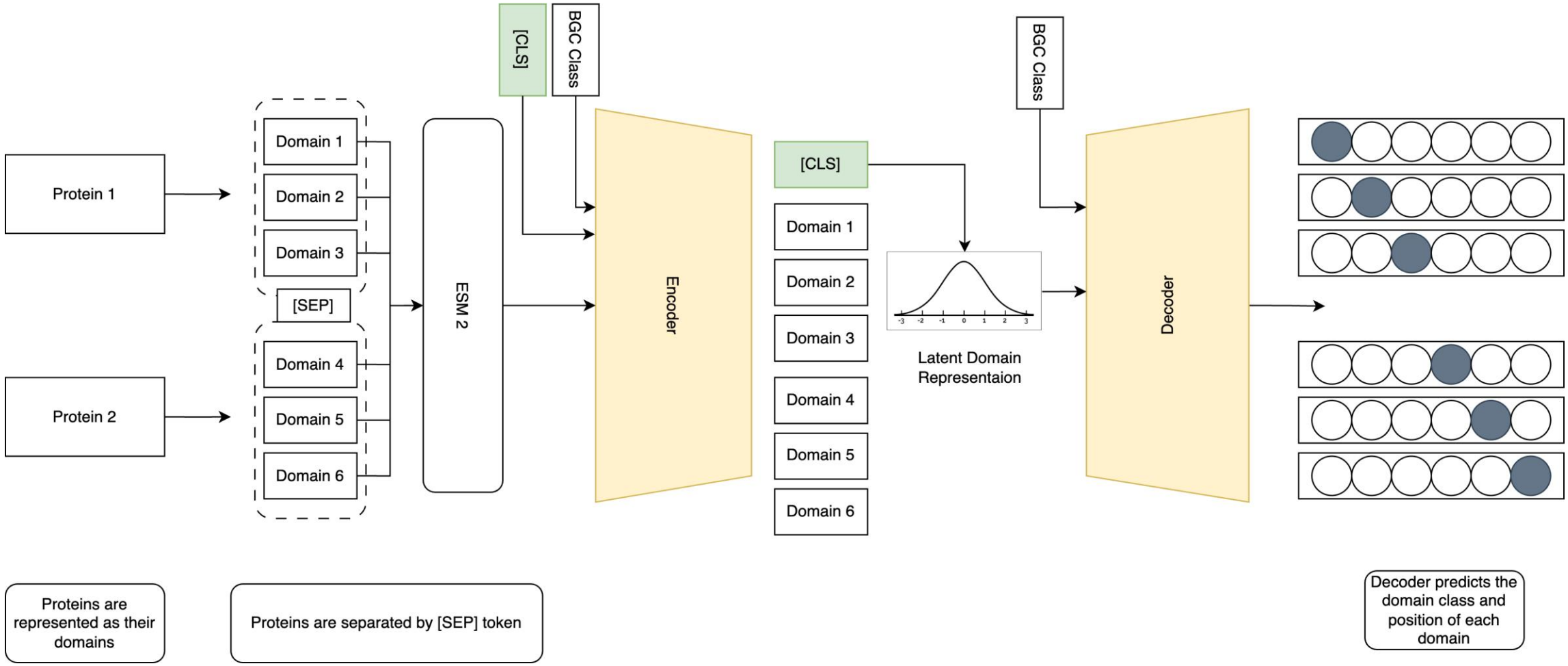
	Component	Details
0	Model Type	Conditional Variational Autoencoder (CVAE)
1	Encoder Input Dimensions	(batch_size=32, input_dim=1280)
2	Padded Embeddings Shape	torch.Size([32, 130, 1280])
3	Latent Space Dimensions	latent_dim=128
4	Encoder Layers	1D CNN: Conv1d -> MaxPool1d -> Fully Connected
5	Decoder Input Dimensions	(latent_dim=128 + condition_dim=10)
6	Decoder Output Dimensions	(batch_size=32, output_dim=1280)
7	Embedding Data Shape	(2145, 1280)
8	Condition Data Shape	(2145, 3)
9	Number of BGC Types	2145
10	Train Size	1716
11	Validation Size	214
12	Test Size	215
13	Shape of First Embedding	(148, 1280)
14	Activation Functions	ReLU (encoder), Sigmoid (decoder)
15	Latent Sampling Equation	z = mu + exp(0.5 * logvar) * epsilon
16	Loss Function	MSE + KL Divergence
17	Optimizer	Adam Optimizer
18	Learning Rate	0.001
19	Batch Size	32
20	Epochs	50
21	Gradient Clipping	1.0
22	Learning Rate Scheduler	ReduceLROnPlateau (factor=0.1, patience=5)
23	Training Focus	1D CNN for Encoder, LSTM + Attention for Decoder
24	BLEU Score	0.8187

Training vs Validation Loss

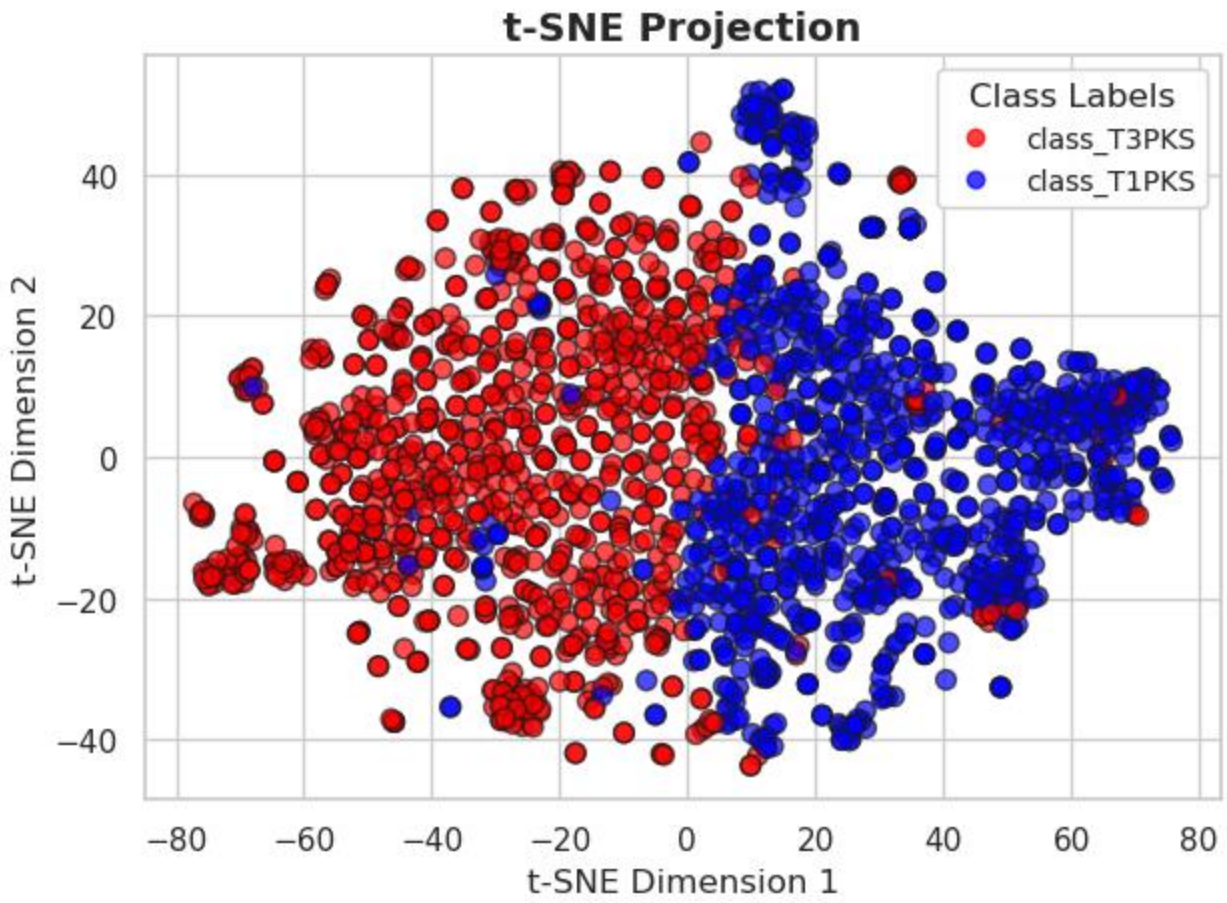


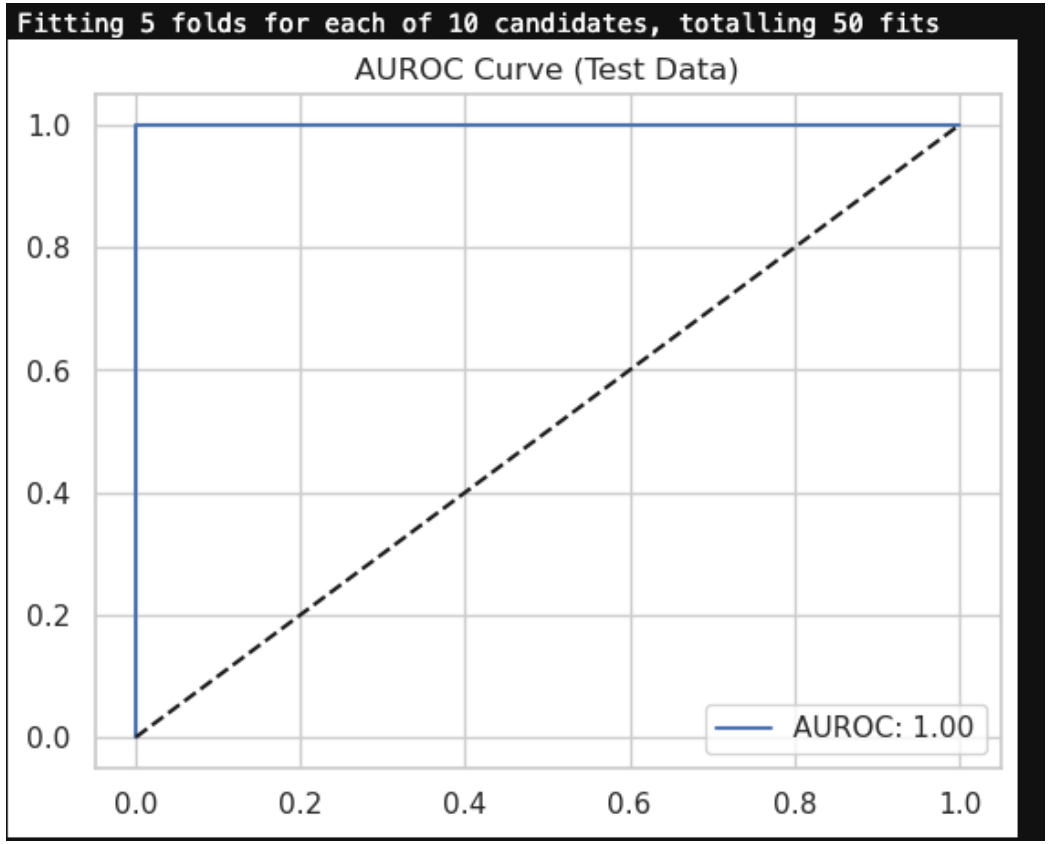


Seq2Seq bgc-cVAE



Seq2Seq bgc-cVAE – ESM1b Oracle Model

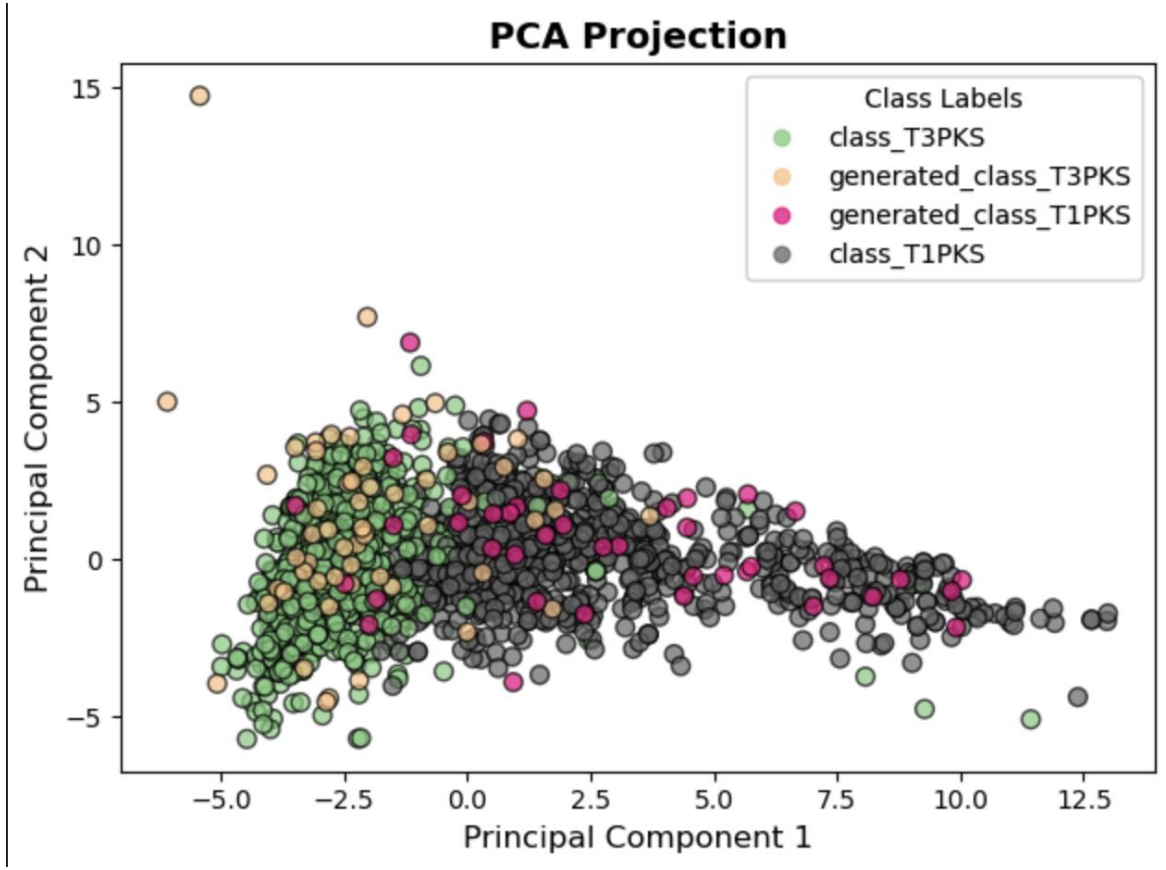




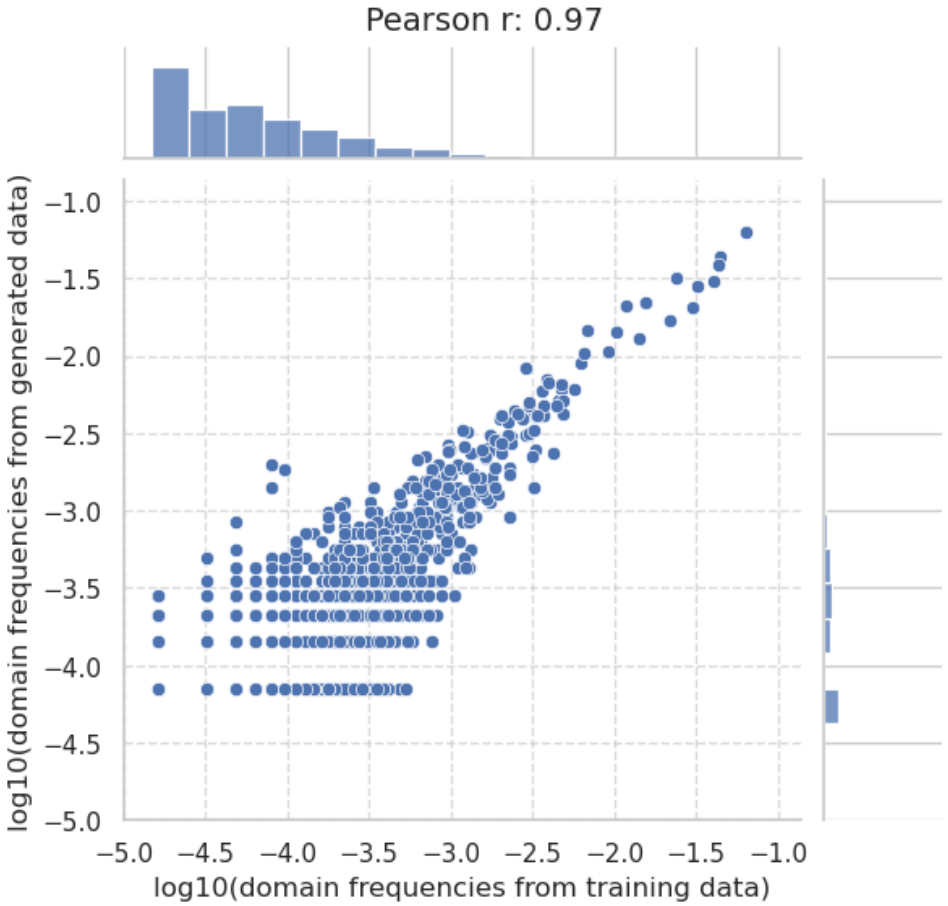
Seq2Seq bgc-cVAE - Results

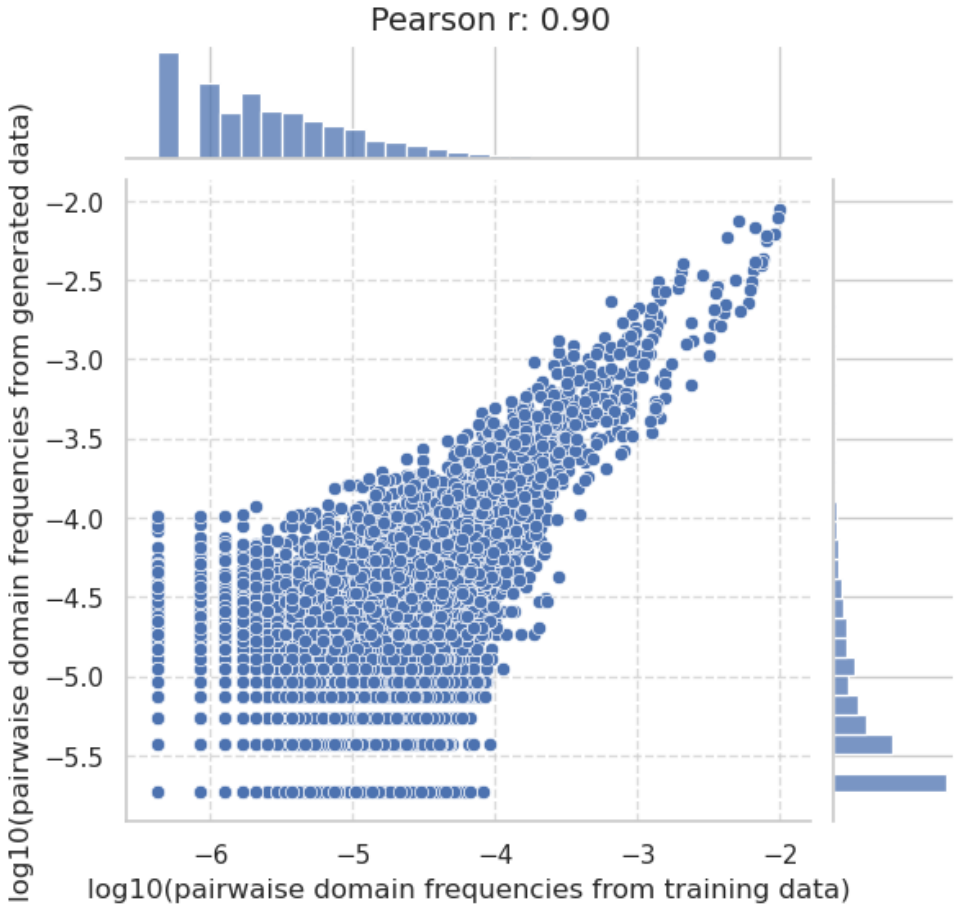
Table 1: Prediction Percentages for Generated BGCs

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



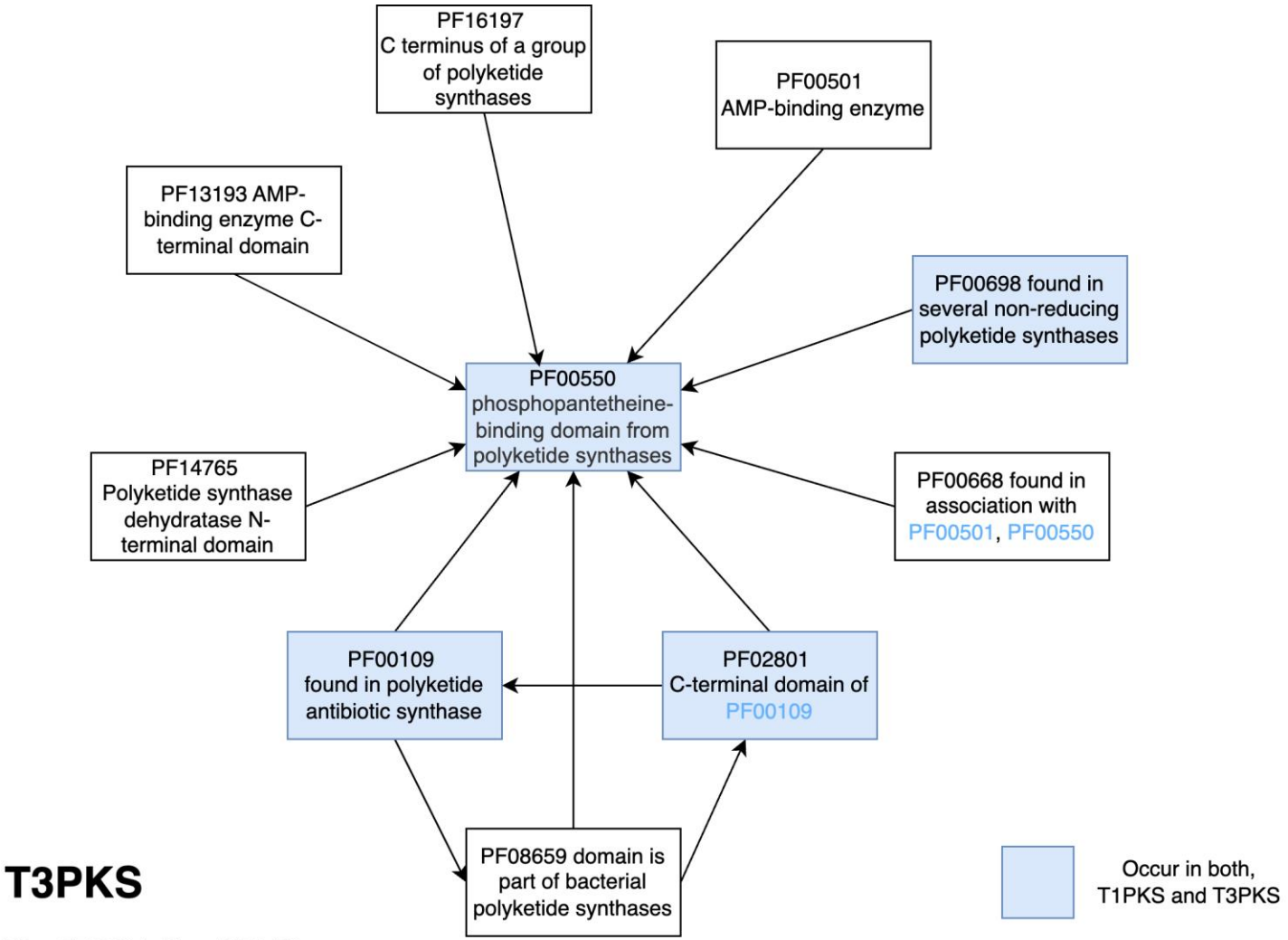
Seq2Seq bgc-cVAE - Results





Category	Percentage (%)
Paired-domains in train not in synthetic	85.36
Paired-domains in synthetic not in train	40.02

Seq2Seq bgc-cVAE - Results



Limitations

- Trained on limited data, with only 2 BGC class types
- Data is generated from antiSMASH, which is a rule-based method to detect BGCs
- More metrics need to be deviced for accurate classification of BGCs generated via seq2seq model

Future Work

- Short-term
  - While occurrence of functional domains in synthetic BGCs is important, it is equally important that the domains occur in correct order – need for a new metric (MSA based on pfam domains)
  - Train on more data with naturally occurring BGCs
  - Ablation study using random domain embeddings instead of ESM1b
- Long-term
  - Model that learns to modify plant BGCs to closely mimic yeast BGCs (yeastizing plant enzymes + microbial cell factories)

References

1. antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters. Marnix H. Medema, Kai Blin, Peter Cimermancic, Victor de Jager, Piotr Zakrzewski, Michael A. Fischbach, Tilmann Weber, Rainer Breitling & Eriko Takano Nucleic Acids Research (2011) doi: 10.1093/nar/gkr466.

2. M. A. Sofi, D. Singh and T. A. Teli, "Attention-based Conditional VAE for Lung Cancer Drug Generation," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 924-928.

3. Simon Zhai, Journal of Manufacturing Systems, https://doi.org/10.1016/j.jmsy.2021.02.0006.

4. F . Masoodi, M. Quasim, S. Bukhari, S. Dixit and S. Alam, Applications of Machine Learning and Deep Learning on Biological Data, CRC Press, 2023.

5. O. Dollar, N. Joshi, D. A. C. Beck and J. Pfaendtner, "Attention-based generative models for de novo molecular design", Chern. Sci., vol. 12, no. 24, pp. 8362-8372, 2021.

6. J. Cheng, L. Dong and M. Lapata, "Long Short-Term Memory-networks for machine reading", arXiv [cs.CL], 2016.

References

7. D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes' CoRR, 2013.

8. Hannigan, G. D. et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Research 47, e110 (2019)

9. Liu, M., Li, Y. & Li, H. Deep Learning to Predict the Biosynthetic Gene Clusters in Bacterial Genomes. Journal of Molecular Biology 434, 167597 (2022).

10. Rios-Martinez, C., Bhattacharya, N., Amini, A. P., Crawford, L. & Yang, K. K. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. PLOS Computational Biology 19, e1011162 (2023).

11. Palaniappan, K. et al. IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. Nucleic Acids Research 48, D422–D430 (2020).

12. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience 10, giaa154 (2021).

13. Kwon MJ, Steiniger C, Cairns TC, Wisecaver JH,Lind AL,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 Metabolic Pathways. Microbiol Spectr9:e00898-

14. https://doi.org/10.1128/Spectrum.00898-21

15. Van Gelder, K., Lindner, S.N., Hanson, A.D. & Zhou, J. (2024) Strangers in a foreign land: 'Yeastizing' plant enzymes. *Microbial Biotechnology*, 17, e14525. Available from: https://doi.org/10.1111/1751-7915.14525

Thank you!

Q/A