UF UNIVERSITY of FLORIDA

PHA6935 (AI for Drug Discovery)

Conditional Design of Novel Biologics with Generative AI

(Designing Biosynthetic Gene Clusters using cVAE)

Group 3 Members: Joseph L. Tsenum and Palash Sethi

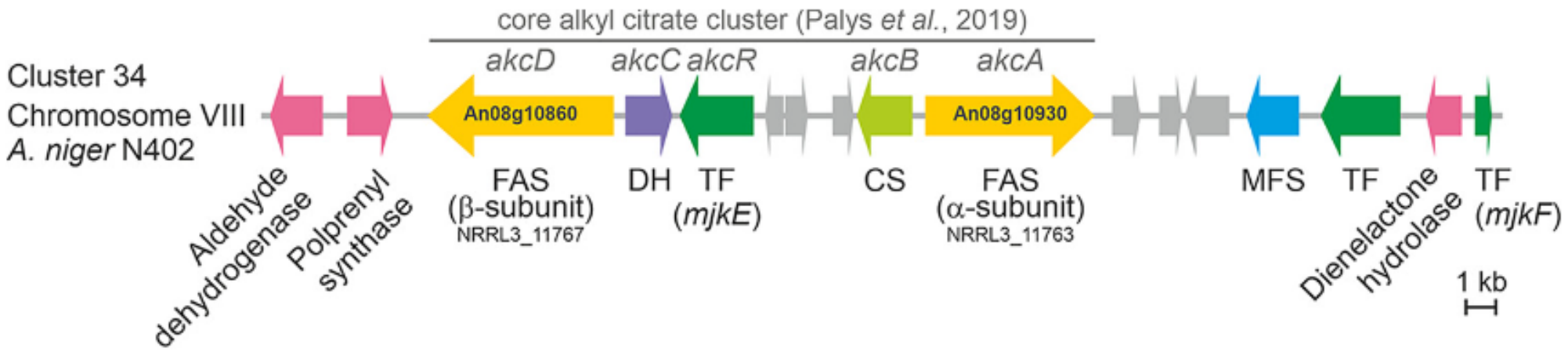
Outline

- BGCs
- Literature Review
- Learnings from Literature Review
- Proposed Method and Novelty
- Data
- Model
- Potential Issues
- References

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.

Biosynthetic Gene Clusters (BGCs)



An alkyl-citrate producing cluster from Kwon *et al.* **FAS (Fatty Acid Synthase), NRPS (Non-Ribosomal Peptide Synthetase), and PKS (Polyketide Synthase)** are enzyme systems involved in the biosynthesis of complex natural products. They act like "molecular assembly lines," building specific types of molecules.

Why do we need to design BGCs

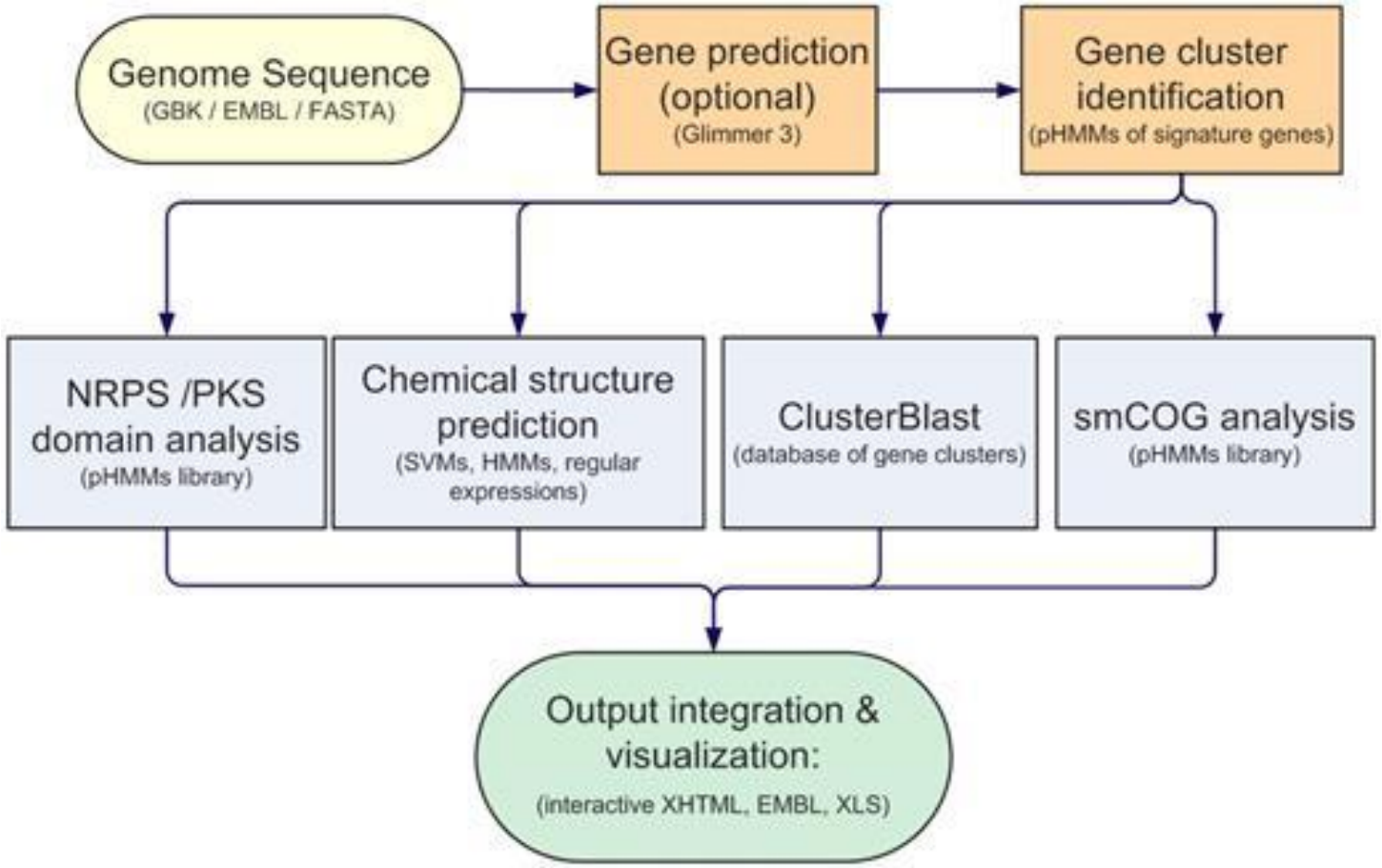
- Refactoring existing BGCs by modifying regulatory elements, promoters, or gene organization can enhance the expression and yield of desired compounds.
- Many natural BGCs are silent or poorly expressed under standard laboratory conditions. Designing new regulatory systems or reconstructing BGCs can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

Literature Review

- To our knowledge, only BGC detection and product type classification methods exist.
- We discuss the following papers for BGC classification -
  - antiSMASH (bioinformatic tool based on pHMMS)
  - DeepBGC (bi-LSTM based)
  - e-deepbgc (bi-LSTM + metadata based)
  - BigCARP (self-supervised learning based)

antiSMASH workflow

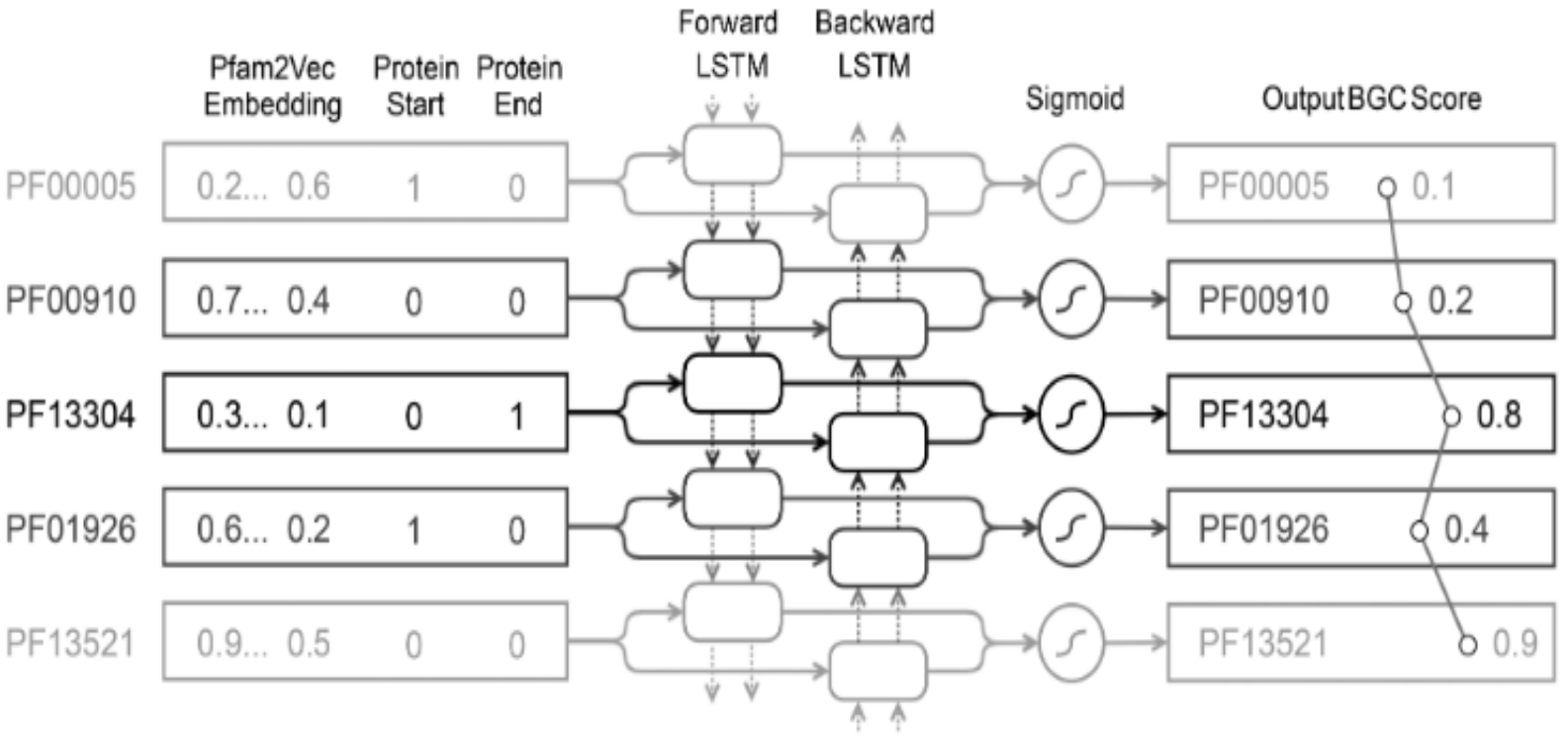
- Designed to enhance the precise characterization of BGCs, allowing for a detailed understanding of their biosynthetic potential, evolutionary context, and regulatory mechanisms within microbial genomes.
- It aligns the identified gene cluster regions with their closest relatives from a comprehensive database of known clusters and integrates all previously available secondary metabolite-specific gene analysis tools into one interactive view.



Medema et al. (2011)

DeepBGC

- A bi-LSTEM network
- Uses pFAM domains to represent a protein.
- pFAM domain is converted to vector using pFAMtovec
- Appends a start/end binary feature
- For each domain, predicts the probability of it being present in a BGC



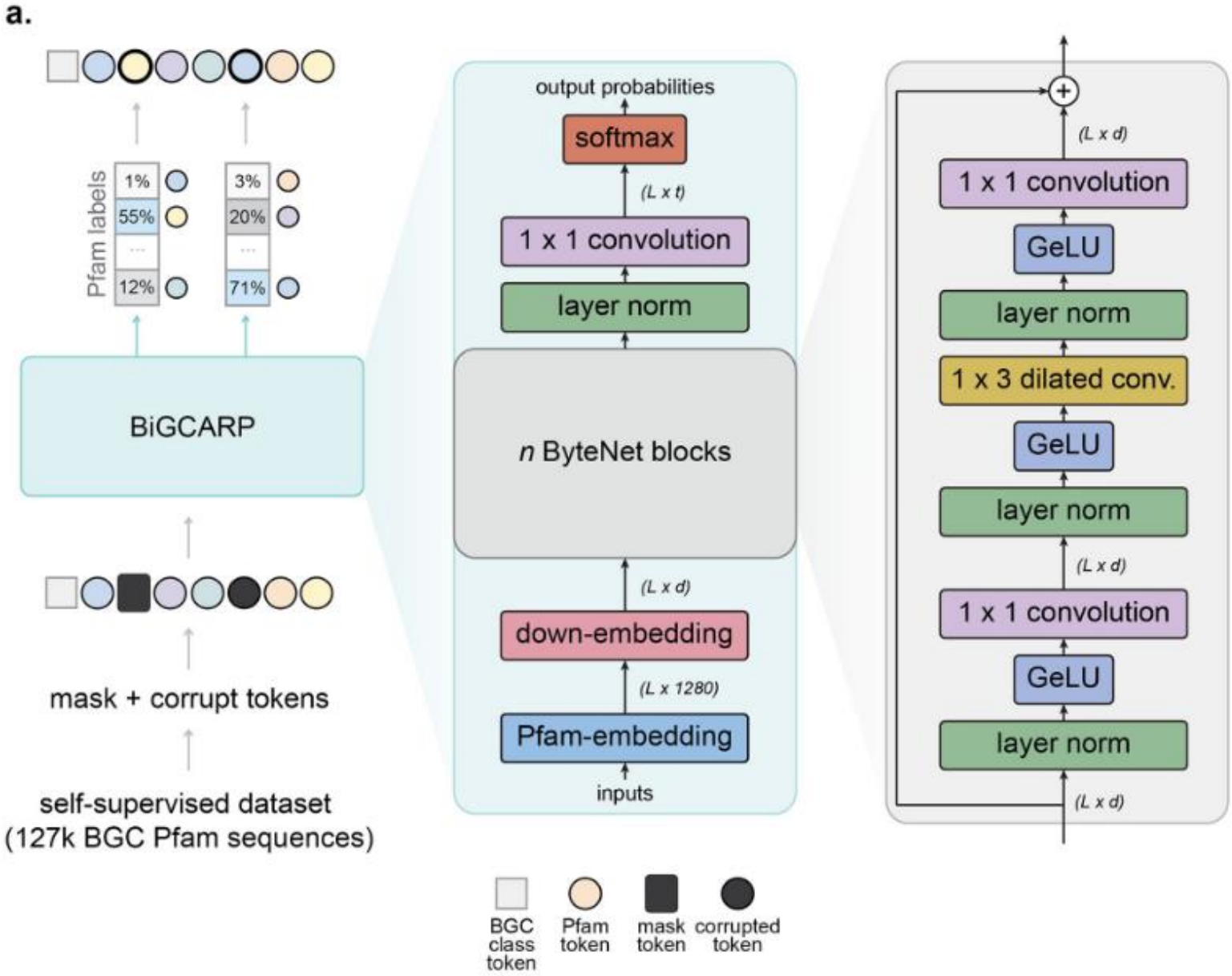
	DeepBGC	e-DeepBGC
		Prediction at the Pfam level
Precision	0.834(0.0172)	0.830(0.0148)
Recall	0.745(0.0082)	0.821(0.0052)
F1	0.787(0.0067)	0.825(0.0071)
AUC	0.867(0.0038)	0.904(0.0024)

e-DeepBGC

- A model similar to DeepBGC but also uses domain summary information extracted from PFAM (e.g., PF00001: 7 transmembrane receptor (rhodopsim family)), clan (e.g. PF00001: CL0192), and a set of similar Pfam domains (e.g. PF00001: PF05296, PF10320, PF10323, PF10324, PF10328, PF13853g).
- Reduced false positive rate and higher sensitivity for BGC detection.

BiGCARP

- Uses a convolutional self-supervised model for pre-training on BGC data
- A BGC class token is appended in the front of inputs.
- A protein is represented by its constituent domains which are further vectorised using ESM2



Learnings from Literature Survey

- Deep learning models learn the semantics of BGCs and perform better than profile HMMs for classification
- A lot of BGC data exist for training AI models
- Proteins in a BGC can be represented by the PFAM domains
- To convert PFAM domains to vectors, protein language models such as ESM2 can be used [tokenisation]
- Models need domain location along with input

Proposed Method and Novelty

- Better domains specific to BGCs not used in previous deep learning methods.
- To our knowledge, the first method to design new BGCs.
- Transformer based conditional variational autoencoders architecture

Data

- AntiSMASH (Medema *et al.*)
- IMG-ABC v.5.0 (Palaniappam *et al.*)

- antiSMASH and IMG-ABC, the 2 largest BGC databases, jointly comprise 565,096 BGCs predicted from 85,221 bacterial genomes with 55 metabolites classes such as NRPS/PKS/RiPPS etc.

Representation details for BGCs and prediction – Biosynthetic PFAM + sub-PFAM domains

- While a protein can be represented by its constituent domains, it is difficult to generate a protein sequence by the predicted domains since it is not a one-to-one map. Additionally, there are ~19,500 valid domains in PFAM, making our vocabulary size large.
- We use the domains covered in BiG-Slice(Kautser *et al.*), since they are exclusive to BGCs data. Additionally, the domains are further subdivided into sub-PFAM domains, making our search space to map predicted domains to protein sequence smaller.
- Vocab size – 2027 biosynthetic PFAM + 3889 sub-PFAM

Deep Conditional Generative Models (CGM) for Structured Output Prediction

- Supervised deep learning has been successfully applied in solving various recognition tasks across machine learning and computer vision.
- While supervised deep learning can effectively approximate complex many-to-one functions with sufficient training data, its **lack of probabilistic inference** poses challenges when modeling complex structured output representations.
- Conditional Variational Autoencoders (cVAE) are deep generative models that uses **Gaussian latent variables** to model structured output variables.

Sofi et al. (2023)

Conditional Variation Autoencoders (cVAE)

- 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.

Sofi et al. (2023)

Types of variables in cVAE

These include:

1. **Input variables (x)**: The observed data or features that serve as input to the encoder.
2. **Latent variables (z)**: Hidden variables sampled from a Gaussian distribution, capturing the underlying structure of the data.
3. **Conditioning variables (c)**: External variables or labels (e.g., class labels or properties) that influence both the encoder and decoder to generate conditioned outputs.
4. **Output variables (y or x hat)**: The reconstructed or generated data produced by the decoder, conditioned on both z and c.

Sofi et al. (2023)

cVAE Architecture

Suppose VAE consists of two parameterized neural nets:

a probabilistic encoder,

$q_\theta(z|x)$ and a probabilistic decoder,

$p_\phi(x|z)$, with

θ and

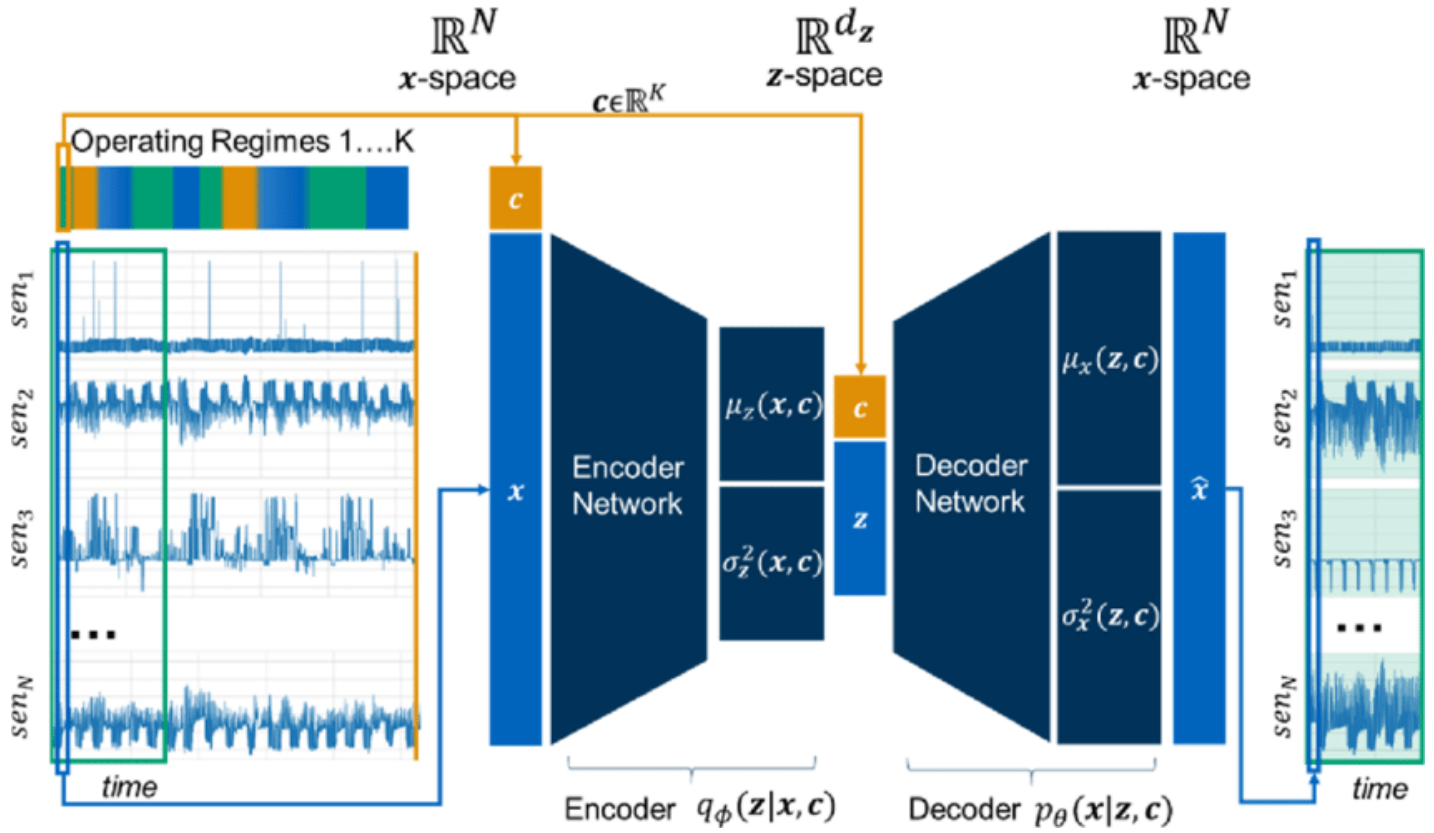
ϕ, as parameters for the encoder Gaussian distribution and likelihood

cVAE conditions these two distributions with a condition **vector (c)**, so the encoder and decoder learn conditional distributions as **$q_\theta(z|x,c)$,$p_\phi(x|z,c)$**, respectively.

Output is generated from the distribution **$p_\phi(y|x,z)$.**

Sofi et al. (2023)

cVAE



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

Maximizing the marginal likelihood in cVAE

To generate novel drug molecules with a desired anti-cancer properties (e.g., growth inhibition), probabilistic encoder and decoder networks are trained by **maximizing the marginal likelihood**, as direct maximization is computationally difficult.

Evidence lower bound (ELBO) as an objective function can be shown:

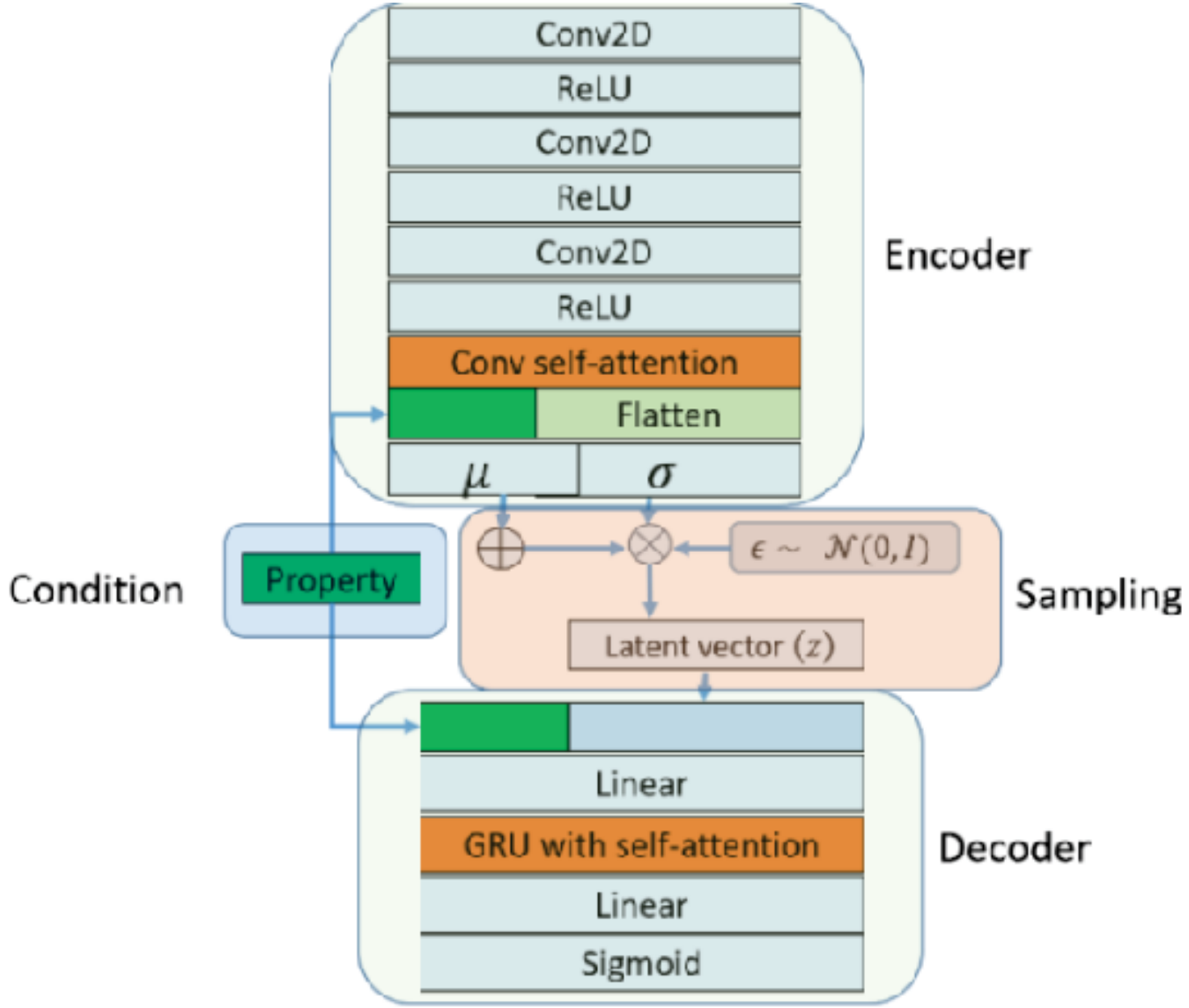
$$\mathbf{E_{n\sim q_\theta(n|l,m)}[\log p_\phi(l|n,m)] - D_{KL}[q_\theta(n|l,m) \| p_{prior}(n|m)]} \quad \mathbf{(1)}$$

- ELBO objective has two terms – Loss functions (Kingma and Welling (2013)):
  - Optimizing the first term ensures that **reconstruction loss is minimized**, i.e to ensure the decoded data is close to the input data.
  - The second term is the **Kullback-Leibler (KL) divergence** between the approximate posterior and the prior., acts as a **regularizer** (which regularizes the latent space to follow a known distribution, like Gaussian distribution).

Sofi et al. (2023)

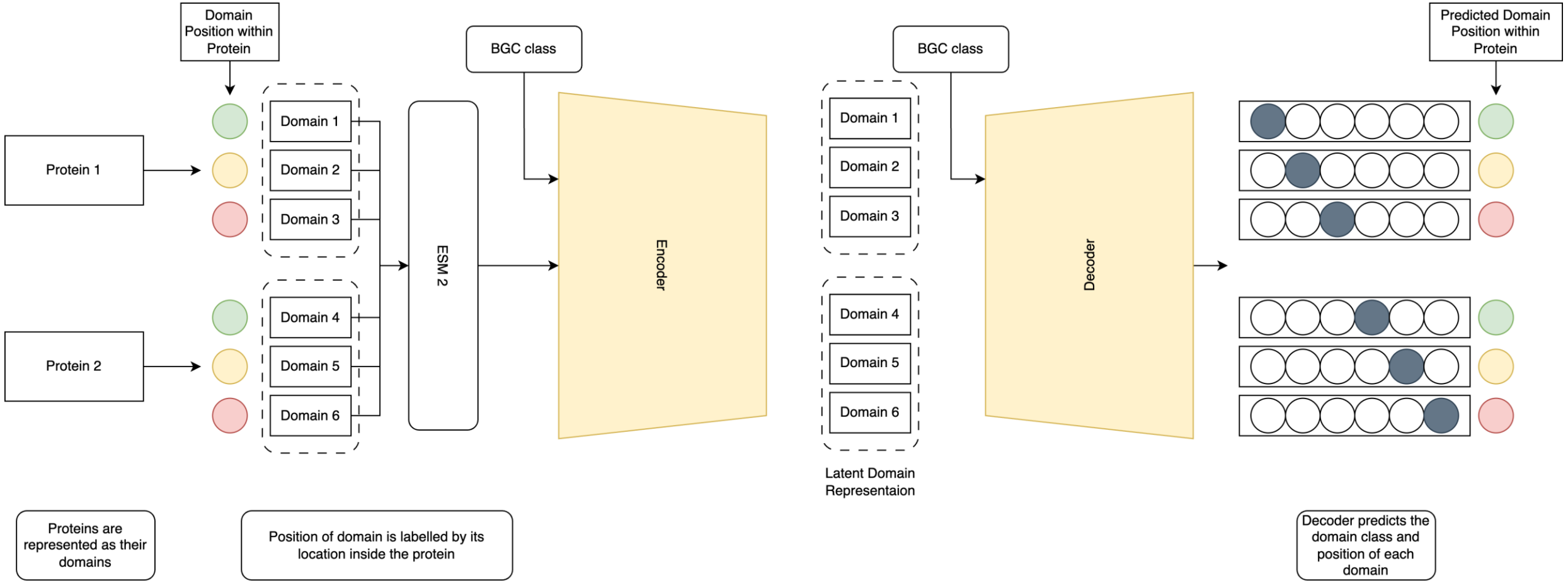
Attention-based Conditional VAE

Can we build 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)

BGCs design using cVAE



Methods for validation of generated BGCs

- To validate the designed BGCs, we will use PFAM to identify predicted domains and query antiSMASH to verify if similar BGCs exist in the database. Further we can verify our BGCs with deepBGC, BiGCARP , and clusterfinder.
- The predicted domains can be queried against their gene ontology information to verify if the domain has similar function with the replaced domain.
- Open to more ideas

Potential Issues

- We need to come up with a valid train-test split methodology.
- We need to control for sequence similarity of the designed BGC with the input BGC.
- What if the designed BGC cannot be expressed in bacteria/yeast?

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).

20. 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).

21. 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).

22. 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:e00898-21.https://doi.org/10.1128/Spectrum.00898-21

Thanks