Exploring sequence landscape of biosynthetic gene clusters with protein language models

Tatiana Malygina [1], Olga V. Kalinina[1,2,3]

[1] Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Saarbruecken, Germany
[2] Center for Bioinformatics, Saarland University, Saarbruecken, Germany
[3] Faculty of Medicine, Saarland University, Homburg, Germany

NaturalPPLuM
(**Natural P**roduct
Protein **L**ang**u**age **M**odel)

Introduction

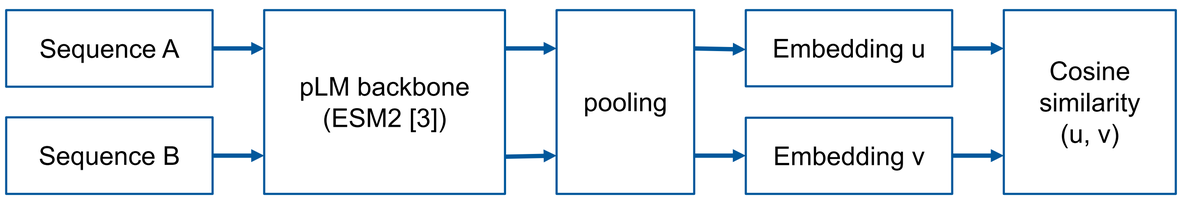
Many organisms, such as bacteria, fungi, and plants, produce intricate chemicals that are not needed for their growth and reproduction, and thus are called secondary metabolites or natural products (NPs). NPs are a rich source of drugs, with most antibiotics being derivatives of NPs. In a producer organism, NPs are synthesized by a set of enzymes encoded by genes that often lie near each other on the chromosome and are called a biosynthetic gene cluster (BGC).

In this work, we explore the capability of protein language models (PLMs) to produce meaningful representations of BGCs. We employ transfer learning to train models to predict the chemical class of the produced compound and explore the topological properties of the produced embeddings.

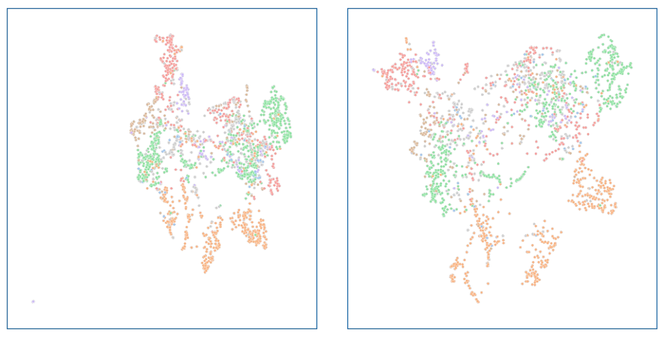
The code is available at project's GitHub repository:
https://github.com/kalininalab/NaturalPPLuM.

Training scheme

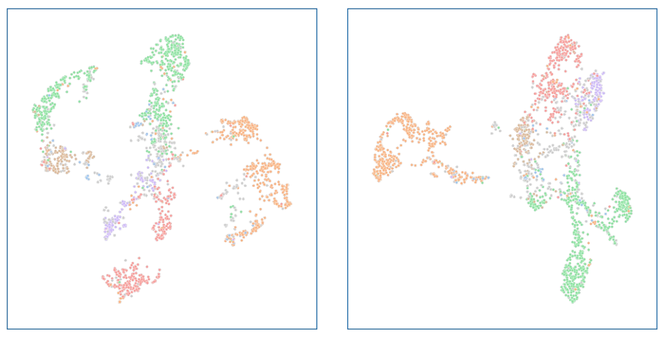
- Task: For a pair of BGCs, predict if they belong to the same BGC class or not
- Input: pairs of proteins or domains of a BGC
- Data split: stratification or LOCO (leave one class out)
- Contrastive learning with Siamese network [2] and cosine similarity loss



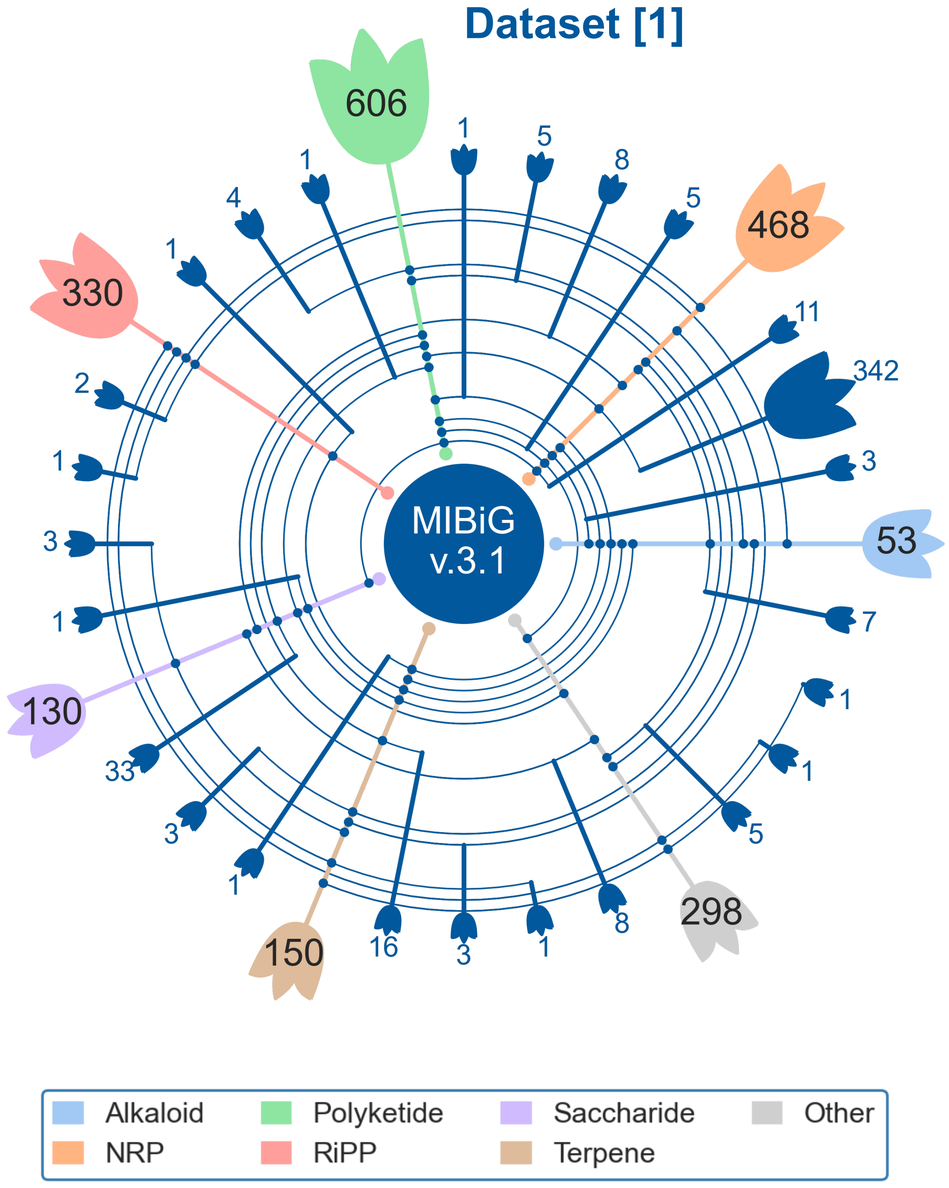
Projection of the embeddings spaces with UMAP for vanilla PLM. Single-domain input on the left and pair-domain input on the right. Each dot represents a BGC, colored by biosynthetic class.



Projection of the embeddings spaces with UMAP for the fine-tuned model with a stratified split. Single-domain input on the left and pair-domain input on the right. Each dot represents a BGC, colored by biosynthetic class.



Dataset [1]



Average Euclidean distance from embeddings of BGCs from a specific biosynthetic class X to all other classes from models fine-tuned with LOCO splits, where that particular class X was held out for the test set. Only distances for models with a pair-domain input are shown.

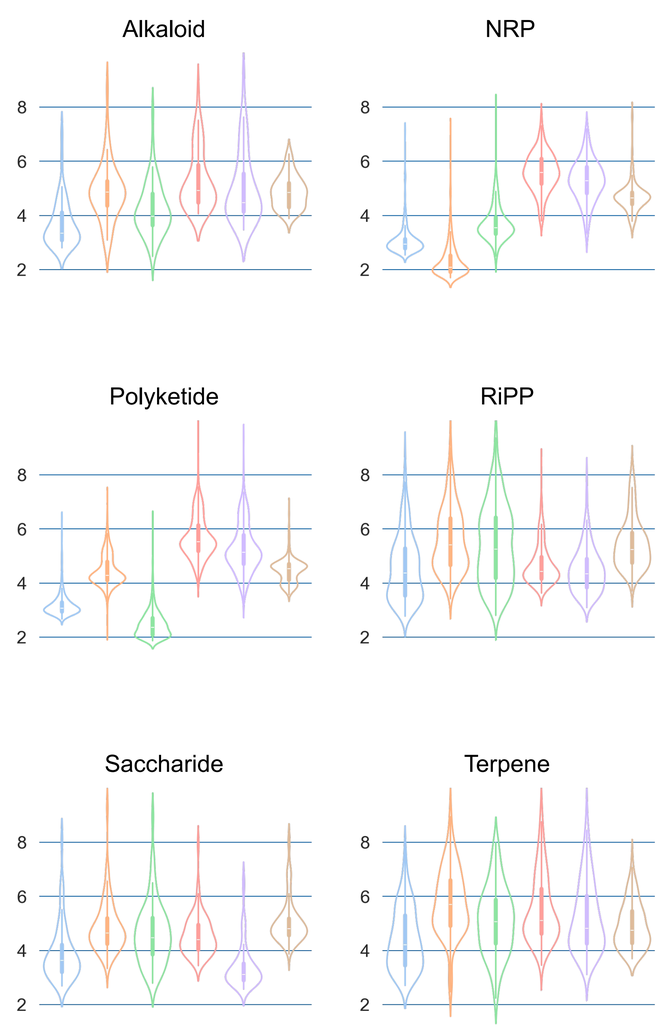
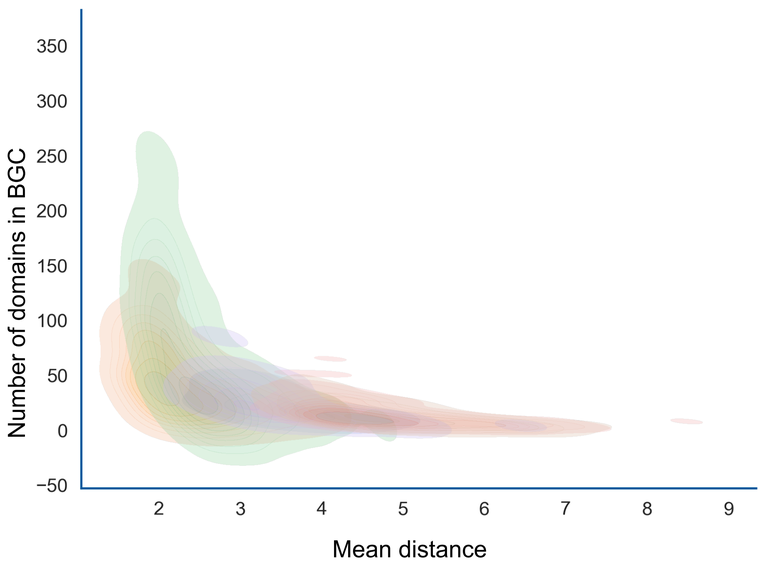


Table below shows the mean silhouette distance:

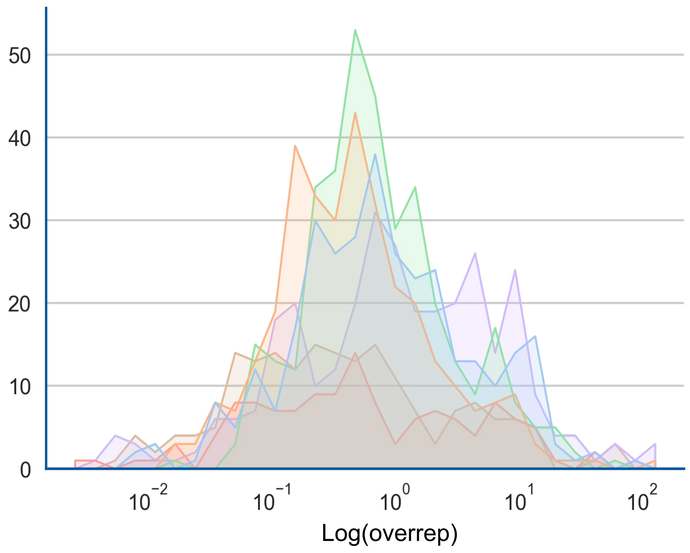
	NaturalPPLuM (our model), silhouette distance		BiG-SCAPE [4] distances
	Pairs of domains	Single domains	
Without fine-tuning	0.098	-0.1364	NA
Stratified split	0.1723	0.0048	NA
LOCO split			
Alkaloid	-0.0117	0.1249	0.0022
NRP	0.2272	0.1385	0.0840
Polyketide	0.219	0.1726	-0.0133
RiPP	-0.1649	-0.1573	0.0698
Saccharide	0.0634	0.0772	0.0288
Terpene	-0.1899	-0.1134	0.0352
Other	-0.0995	-0.0943	0.0220

Relationship between the number of domains per BGC (vertical axis) and the average distance to other BGCs (horizontal axis) from the same class for embeddings generated by LOCO models for each biosynthetic class.



Overrepresentation of domains in biosynthetic gene clusters (smoothed histogram). We calculate the overrepresentation of a Pfam domain d in a class C as

$$Over_C^d = \frac{N_C^d / N_C}{\sum_C N_C^d / N_C}$$



Conclusions and outlook

- Protein language models can be used for meaningful classification of biosynthetic gene clusters based on their embeddings in the latent space
- Further fine-tuning can improve separation of biosynthetic classes
- NRP and Polyketide classes are best separated from other, despite shared homologous domains
- Saccharide BGCs contain many domains that are overrepresented in this class, and can be also well separated from the others in the latent space
- **New unknown BGC classes can be identified?**

References

[1] Terlouw *et al.*, *Nucleic Acids Res*, 51(D1): D603-D610, 2022. https://doi.org/10.1093/nar/gkac1049
[2] Reimers *et al.*, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 11:, 2019. http://arxiv.org/abs/1908.10084
[3] Lin *et al.*, *Science*, 379: 1123-1130, 2023. https://www.science.org/doi/10.1126/science.ade2574
[4]: Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W. et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16, 60–68 (2020). https://doi.org/10.1038/s41589-019-0400-9

Project updates at:

kalininalab/NaturalPPLuM