Metagenomic Binning using Graph Neural Networks

Bachina Pranav Aaryan Ajay Sharma Sriteja Reddy Pashya
Team 10: Graphs are All you Need

Abstract

Metagenomic Binning (MB) is essential for analyzing environmental microbial communities by grouping mixed DNA fragments into bins that represent individual or related species. This project advances the field by applying Graph Representation Learning (GRL) techniques to the assembly graph of these fragments, a method traditionally overshadowed by the focus on genomic features. Building on the foundational architecture of RepBin (Xue et al., 2022), we systematically replace components of the RepBin framework with alternative methods to evaluate their impact on graph-based learning and the binning process. We use the Sim-5G dataset to elucidate how various architectural modifications affect MB performance.

1 Introduction

Metagenomic Binning (MB) is a crucial process for analyzing environmental microbial communities by categorizing mixed DNA fragments into bins representing individual or related species. The challenges in MB include managing the immense diversity and unknown nature of microbial species, the short length and high similarity of DNA sequences, and the lack of reference genomes (Pasolli et al., 2019), which complicate the accurate assignment and categorization of DNA fragments. RepBin introduces constraint-based learning and binning methods and graph representation learning that preserves both homophily relations and heterophily constraints, moving beyond traditional approaches that primarily rely on contig composition and coverage.

Contributions: (1) We reproduced the results achieved by RepBin. (2) We showed the importance of constraint-based binning and experimented with various loss functions as alternatives to the exponential contrastive loss. (3) Additionally, we extensively tested alternatives to the Graph Diffusion Convolution (GDC) operator and achieved results comparable to the original model. (4) Furthermore, we propose some promising novel ideas, which we hope to explore in future research.

2 RepBin: Constraint-based Graph Representation Learning for MB

RepBin aims to learn node representations that capture the global structure of the entire graph by maximizing the mutual information between node-level and graph-level features (cf. 1). On top of this, it also tries to learn node representations that capture the heterophily information i.e., contigs having the same Single Copy Marker Gene(SCG) belong to different bins (cf. 2). $\mathcal{M}$ is a set of pairs of contigs of all heterophily constraints. The entire framework of RepBin is shown in Figure **??**. Equation 3 is the final objective function used to update the parameters of the RepBin model.

$$\mathcal{L}_g = -\frac{1}{2n}[\sum_{i=1}^{n} \log \mathcal{D}(\mathbf{h_i}, \mathbf{s}) + \sum_{j=1}^{n} \log(1 - \mathcal{D}(\tilde{\mathbf{h_j}}, \mathbf{s}))] \tag{1}$$

$$\mathcal{L}_c = \frac{1}{|\mathcal{M}|} \sum_{m(i,j)\in\mathcal{M}} \exp^{-||\mathbf{h}_i - \mathbf{h}_j||_2} \tag{2}$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

$$\mathcal{L} = \mathcal{L}_g + \lambda \cdot \mathcal{L}_c \tag{3}$$

3 Experiments

Experiments were conducted on the Sim-5G dataset. F-1, ARI, and NMI scores were used as evaluation metrics for binning quality. The results reported in the main paper were successfully reproduced. Attempts were made to improve performance by replacing the GNN engine of RepBin with more powerful variants. The removal of the GDC operator or the constraint-based binning component from the framework resulted in significant performance degradation, indicating their critical importance to the overall framework's effectiveness. All values are reported in Table 5

3.0.1 Diffusion operator

Given a graph $\mathcal{G}$, the node-homophily score for $\mathcal{G}$ is defined in Lim et al. (2021) as follows:

$$\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \frac{|\{u \in \mathcal{N}(v) : y_v = y_u\}|}{|\mathcal{N}(v)|}, \tag{4}$$

where $\mathcal{V}$ is the set of nodes, $\mathcal{N}(v)$ is the set of neighbours of node v, and y_v is the class of node v. Assembly graphs are known to have high node-homophily score (Xue et al., 2022).

Message-Passing Neural Networks (MPNNs) (e.g., GCN, GraphSAGE) mainly aggregate information from its one-hop neighbourhood. Owing to this, MPNNs become limited in capturing higher order structural information, particularly in graphs with high node-homophily score like assembly graphs. In contrast, diffusion has been found to perform better in graphs with high node-homophily score (Gasteiger et al., 2019). To this end, RepBin uses the GDC operator to capture multi-hop structural information.

To this end, we did ablation on three different graph diffusion operators: (1) **Personalized PageRank (PPR) (Jeh and Widom, 2003)**: The GDC operator was derived by modifying the graph diffusion introduced in PPR. Therefore checking performance of PPR graph diffusion in MB becomes imperative. (2) **Heat Diffusion Kernel (HDK) (Lafferty et al., 2005)**: The heat kernel leverages the local structure of target node under heat diffusion to determine its neighboring nodes flexibly, without the constraint of order suffered by previous methods like GCN, GraphSAGE etc (Xu et al., 2020). Ergo, HDK may prove to be more effective than GDC. (3) **Adaptive Diffusion (AD) (Zhao et al., 2021)**: The neighborhood size in GDC is manually tuned for each graph by conducting grid search over the validation set, making its generalization practically limited. AD can potentially address this problem.

3.0.2 Loss function

We did ablation on three different loss functions by replacing the exponential loss function in Eq. 2 with the following three loss functions: (1) Cosine Embedding loss (Pytorch, 2023b) (2) Contrastive loss (Pytorch, 2023a) (3) Margin-based Pairwise Ranking loss (Pytorch, 2023c).

4 Novel ideas that worked/didn't work

The work by Tan et al. (2024) proved contrastive learning to be equivalent to spectral clustering on similarity graphs when done using the standard InfoNCE loss. Using this theoretical insights, they introduce Kernel-InfoNCE loss. Several spectral clustering and contrastive learning methods (Wang et al., 2019; Zhang et al., 2019; Wang et al., 2014; Velickovic et al., 2019; Xue et al., 2022) have been applied to MB. Nevertheless, none of the works have adapted the Kernel-InfoNCE loss— an approach unifying the two standard approaches to MB. This leaves an opportunity to adapt it to the task of MB. We sought to exploit this opportunity but were limited due to the lack of availability of standardised and pre-processed datasets. We hope to further explore this direction of research in the future.

5 Figures/Plots

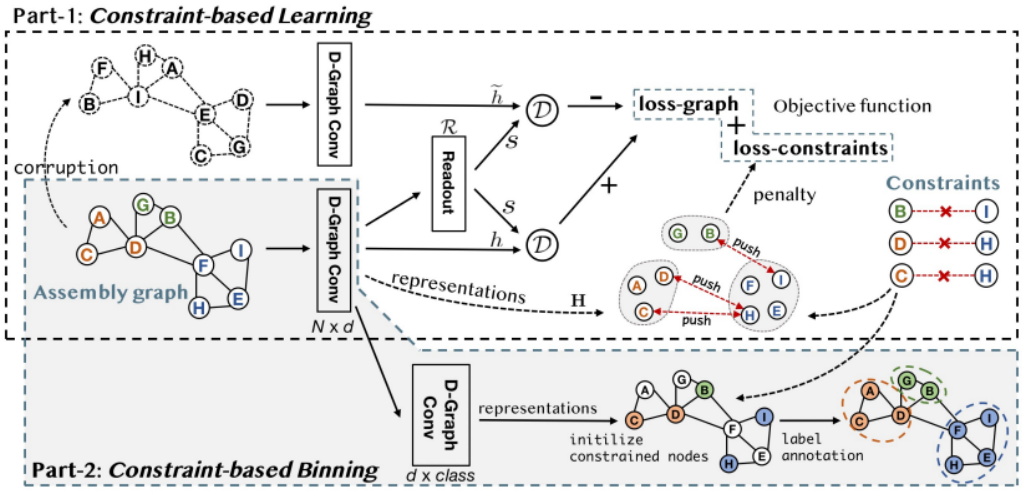
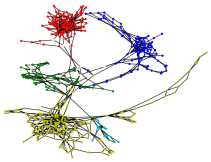
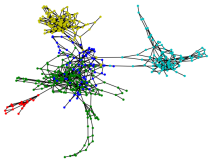


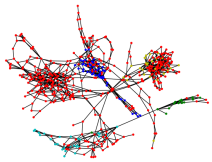
Figure 1: RepBin Framework



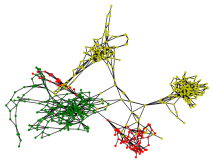
(a) Ground Truth



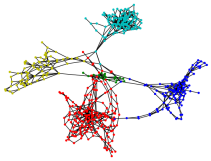
(b) RepBin



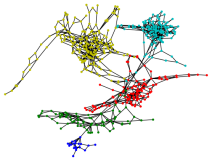
(c) RepBin without GDC



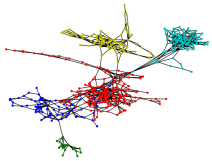
(d) RepBin without $\mathcal{L}_c$



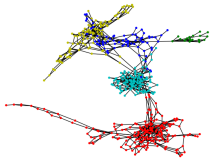
(e) RepBin with degree-based adaptive diffusion



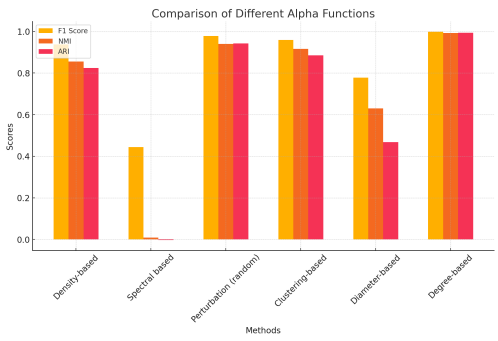
(f) RepBin with HDK



(g) RepBin with PPR-diffusion



(h) RepBin with margin-based contrastive loss

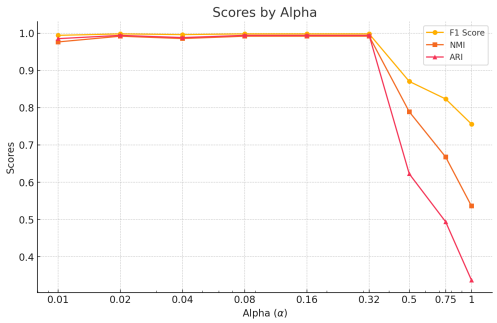


(i) Comparison of different alpha functions

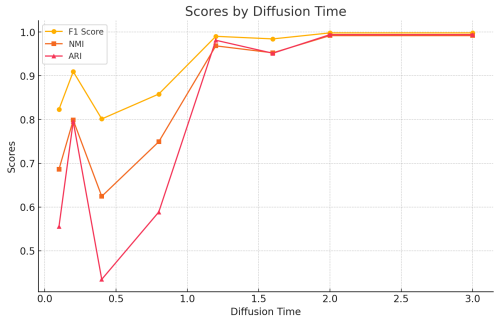
(j) Comparison of different contrastive losses

Figure 2: Performance of RepBin across various ablations.

(a) Variation of performance with the threshold of contrastive loss



(b) Performance distribution with variation in the teleport probability α using PPR.



(c) Performance distribution with variation in diffusion time in HDK.

Figure 3: Performance variation of RepBin with contrastive loss function, and different diffusion operators like PPR and HDK.

Table 1: Impact of various modifications to RepBin on binning performance

Manipulation	F-1	ARI	NMI
Reported values in the paper	99.80	99.40	99.20
Reproduced result	99.80	99.40	99.18
Using GraphSAGE	72.69	38.48	62.08
Using Graph Transformer	69.40	19.95	39.42
Without $\mathcal{L}_c$	96.84	95.36	92.93
Without GDC	70.33	24.97	41.27
Without Constraint-based binning	74.02	45.53	56.26
Using GIN READOUT	99.61	99.11	98.46

Table 2: Impact of various adaptive diffusion functions on binning performance

Function Type	F-1	ARI	NMI
Density-based	93.71	82.42	85.50
Spectral based	44.41	-0.34	0.89
Perturbation (random)	97.84	94.26	93.99
Clustering-based	95.87	88.56	91.57
Diameter-based	77.80	46.75	63.04
Degree-based	99.80	99.40	99.18

Table 3: Variation of binning performance at different diffusion times (Heat Kernels)

Diffusion Time	F-1	ARI	NMI
0.1	0.8232	0.5554	0.6863
0.2	0.9096	0.7957	0.7987
0.4	0.8014	0.4348	0.6245
0.8	0.8583	0.5887	0.7493
1.2	0.9902	0.9812	0.9684
1.6	0.9843	0.9515	0.9525
2.0	0.9980	0.9940	0.9918
3.0	0.9980	0.9940	0.9918

Table 4: Variation of binning performance at different alpha values (PPR diffusion)

Alpha value	F-1	ARI	NMI
0.01	0.9941	0.9852	0.9764
0.02	0.9980	0.9940	0.9918
0.04	0.9961	0.9881	0.9855
0.08	0.9980	0.9940	0.9918
0.16	0.9980	0.9940	0.9918
0.32	0.9980	0.9940	0.9918
0.50	0.8705	0.6225	0.7891
0.75	0.8234	0.4938	0.6677
1.00	0.7561	0.3371	0.5361

Table 5: Variation of binning performance with different loss functions for heterophily learning

Loss Type	F-1	ARI	NMI
Exponential loss (original)	0.9980	0.9940	0.9918
Cosine Embedding loss	0.8729	0.7320	0.7628
Margin Based Pair-wise loss	0.9941	0.9852	0.9764
Contrastive loss	0.9790	0.9672	0.9585

References

[1] Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. *Advances in neural information processing systems*, 32, 2019.

[2] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279, 2003.

[3] John Lafferty, Guy Lebanon, and Tommi Jaakkola. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6(1), 2005.

[4] Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Advances in Neural Information Processing Systems*, 34:20887–20902, 2021.

[5] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662, 2019.

[6] Pytorch. Contrastive loss, 2023a. URL `https://kevinmusgrave.github.io/pytorch-metric-learning/losses/#contrastiveloss`.

[7] Pytorch. Cosine embedding loss, 2023b. URL https://pytorch.org/docs/stable/generated/torch.nn.CosineEmbeddingLoss.html.

[8] Pytorch. Margin loss, 2023c. URL https://pytorch.org/docs/stable/generated/torch.nn.MarginRankingLoss.html#torch.nn.MarginRankingLoss.

[9] Zhiquan Tan, Yifan Zhang, Jingqin Yang, and Yang Yuan. Contrastive learning is spectral clustering on similarity graph. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hLZQTFGToA.

[10] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *ICLR (Poster)*, 2(3):4, 2019.

[11] Xiang Wang, Buyue Qian, and Ian Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28:1–30, 2014.

[12] Ziye Wang, Zhengyang Wang, Yang Young Lu, Fengzhu Sun, and Shanfeng Zhu. Solidbin: improving metagenome binning with semi-supervised normalized cut. *Bioinformatics*, 35(21):4229–4238, 2019.

[13] Bingbing Xu, Huawei Shen, Qi Cao, Keting Cen, and Xueqi Cheng. Graph convolutional networks using heat kernel for semi-supervised learning. *arXiv preprint arXiv:2007.16002*, 2020.

[14] Hansheng Xue, Vijini Mallawaarachchi, Yujia Zhang, Vaibhav Rajan, and Yu Lin. Repbin: Constraint-based graph representation learning for metagenomic binning. In *AAAI*, 2022.

[15] Xiaotong Zhang, Han Liu, Qimai Li, and Xiao-Ming Wu. Attributed graph clustering via adaptive graph convolution. *arXiv preprint arXiv:1906.01210*, 2019.

[16] Jialin Zhao, Yuxiao Dong, Ming Ding, Evgeny Kharlamov, and Jie Tang. Adaptive diffusion in graph neural networks. *Advances in neural information processing systems*, 34:23321–23333, 2021.