DATA-DRIVEN METHODS FOR COMPARATIVE METAGENOMICS

TOPOLOGY-AWARE DIMENSIONALITY REDUCTION AND GRAPH-BASED CLUSTERING

GROUP 16

TEAM MEMBERS

Malshan P.G.P.
E/20/244

Jananga T.G.C.
E/20/158

Prasadinie H.A.M.T.
E/20/300

Supervisors

Dr. Damayanthi Herath
Senior Lecturer
Department of Computer Engineering
University of Peradeniya

Dr.Rajith Vidanaarachchi
Research Fellow, NHMRC
Faculty of Architecture, Building and Planning

Dr.Vijini Mallawaarachchi
Research Fellow in Bioinformatics, FAME
College of Science and Engineering
Flinders University

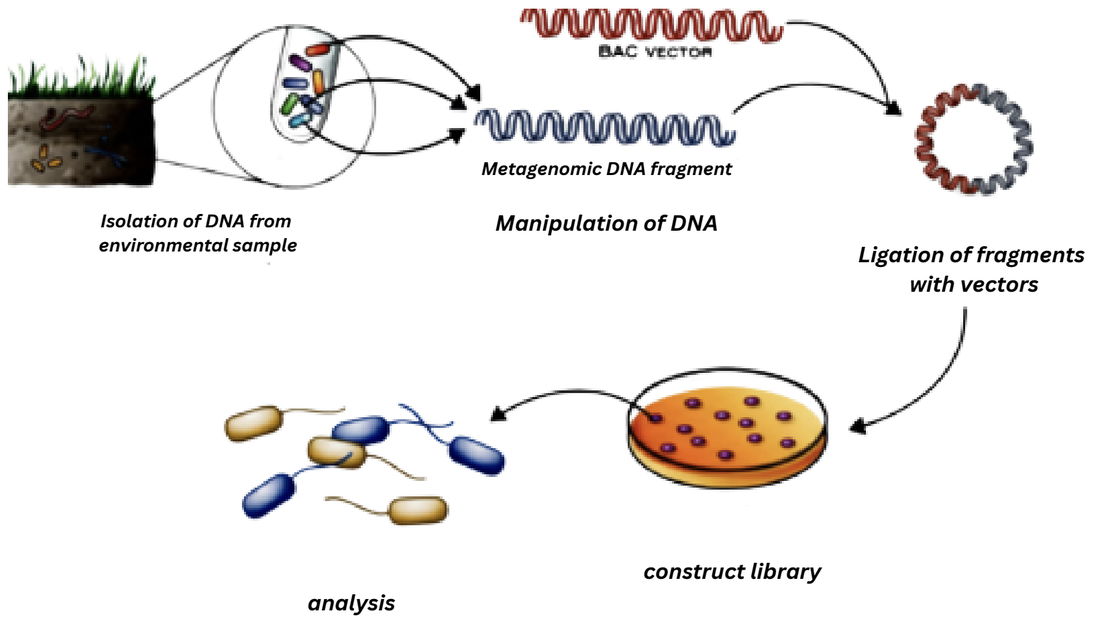
INTRODUCTION

Metagenomics

Metagenomics = Meta (collective) + Genomics (Study of Genomes)

Metagenomics is the culture-independent analysis of a mixture of microbial genomes (metagenome) using an approach based either on expression or on sequencing

(Riesenfeld et al., 2004; Schloss et al., 2003; Susannah et al., 2005; Patrick et al., 2005)



Steps involved in a metagenomics experiment

Source: Zeyaullah et al., Metagenomics

- **Who is there?**
- **What are they doing?**
- **How do they change?**

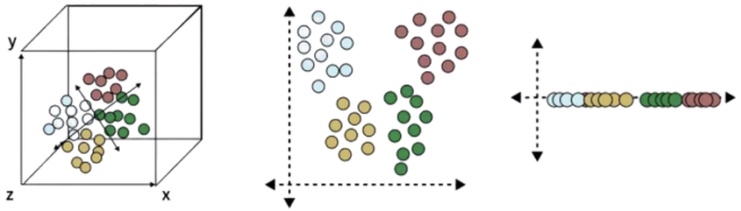
INTRODUCTION(cont.)

Data Driven methods

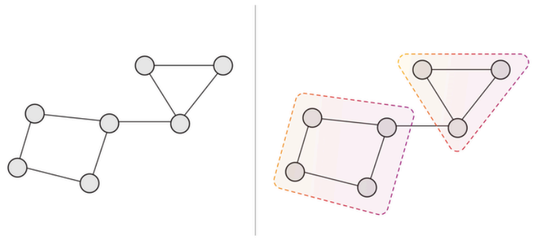
- Uncover patterns or make decisions directly from the data.
- Do not require predefined biological assumptions or predefined rules.

Which Data Driven Methods?

Dimensionality Reduction



Graph Based Clustering

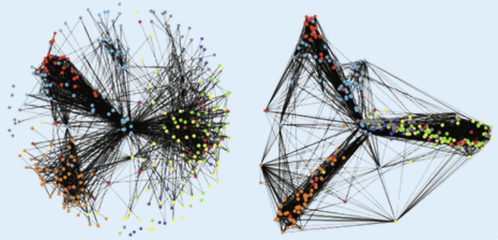


INTRODUCTION TO PROBLEM

The Metagenomic Data Challenge

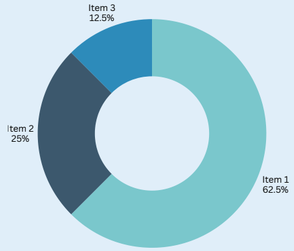
High Dimensional

Thousands of distinct taxa (genes, OTUs, or ASVs) creating a massive feature space



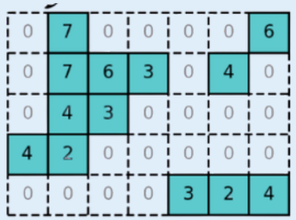
Compositional

Requires specialized statistical methods, unless creating false correlations.



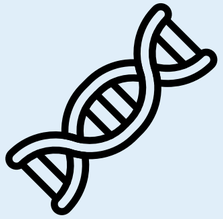
Sparse

Most taxa are not present in most samples



Bio-structural

Ecological and phylogenetic relationships among taxa must be preserved.



Armstrong et al. (2022)

RESEARCH PROBLEM

In Metagenomics data

Which dimensionality reduction method best preserves topological structure?

How graphs and GNNs improve clustering quality?

RESEARCH OBJECTIVE

- Analyze dimensionality reduction for high-dimensional metagenomic data
- Build a preprocessing pipeline for sparse biological datasets
- Benchmark DR methods using multiple evaluation metrics
- Evaluate graph-based and GNN approaches for clustering improvement

PHASE 1

Evaluating Topology Preservation in Dimensionality Reduction Methods for Metagenomic Data

- Developed compositional preprocessing pipeline (nzCLR + matrix completion)
- Benchmarked 8 dimensionality reduction techniques
- Multi-metric evaluation framework (topological + biological metrics)
- Validated across human, soil, and marine metagenomic datasets

METHODOLOGY

Implementation pipeline for Data Preprocessing

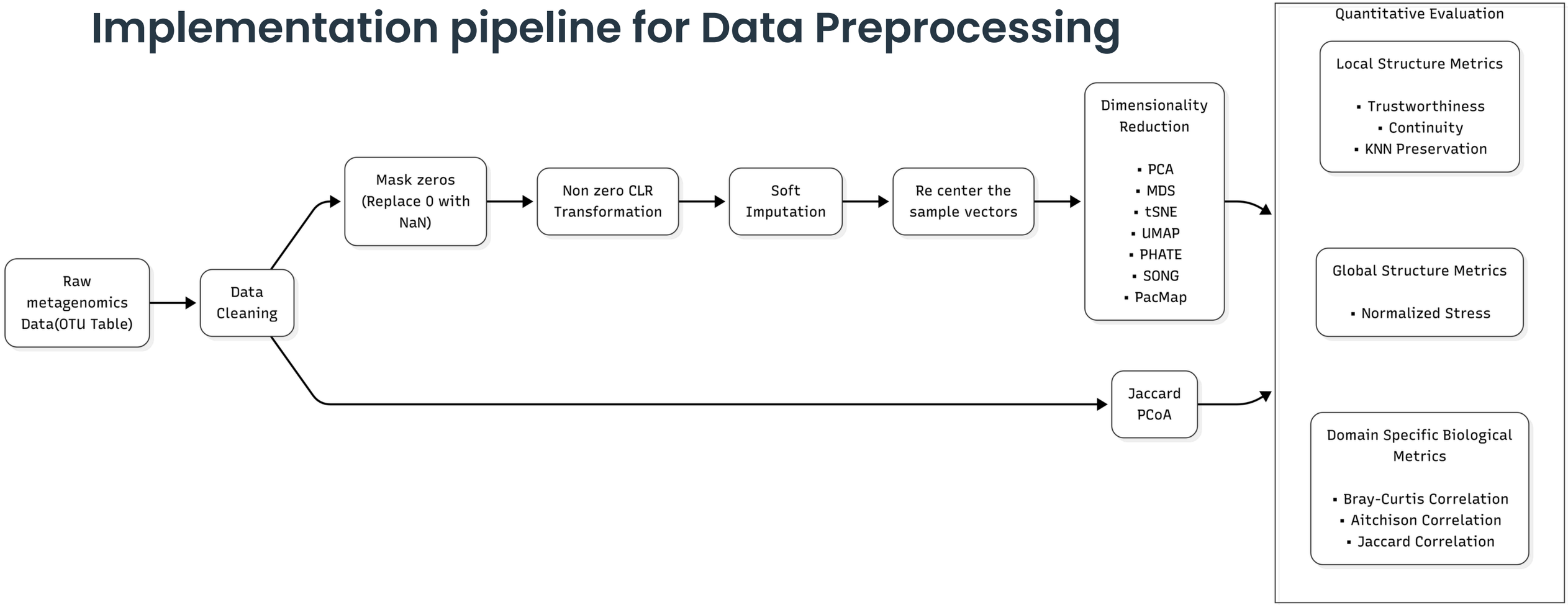
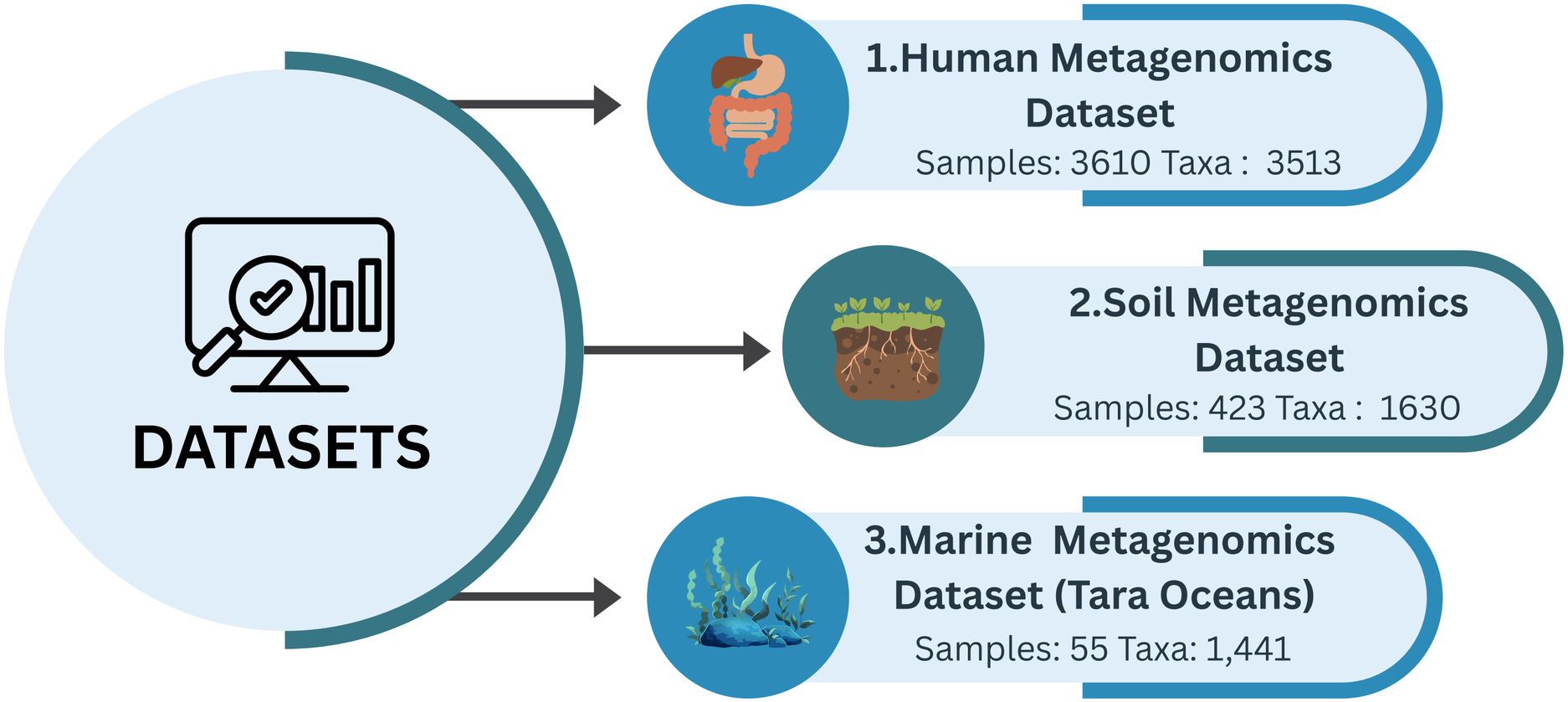


Fig. 1. End-to-end workflow for benchmarking dimensionality reduction methods on compositional metagenomic data, including preprocessing, imputation, and multi-metric evaluation

IMPLEMENTATION

- We experimented across 3 different metagenomics datasets



IMPLEMENTATION-VISUALIZATION

2D Visualization of Human Metagenomics Data

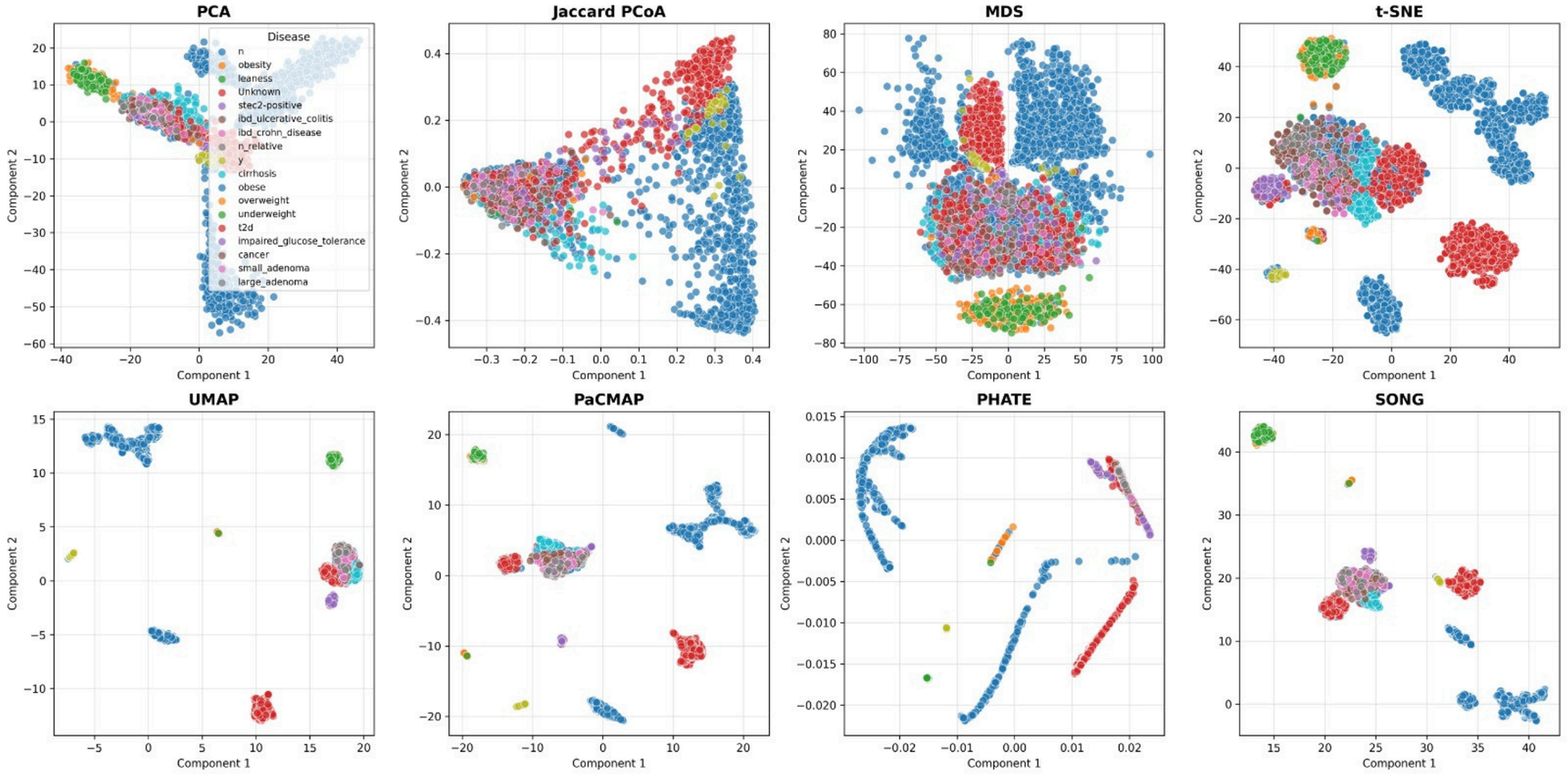


Fig. 2. Visualization of the human metagenomics dataset using eight dimensionality reduction (DR) methods.

IMPLEMENTATION-EVALUATION

Evaluation Metrics for Human Metagenomics Data

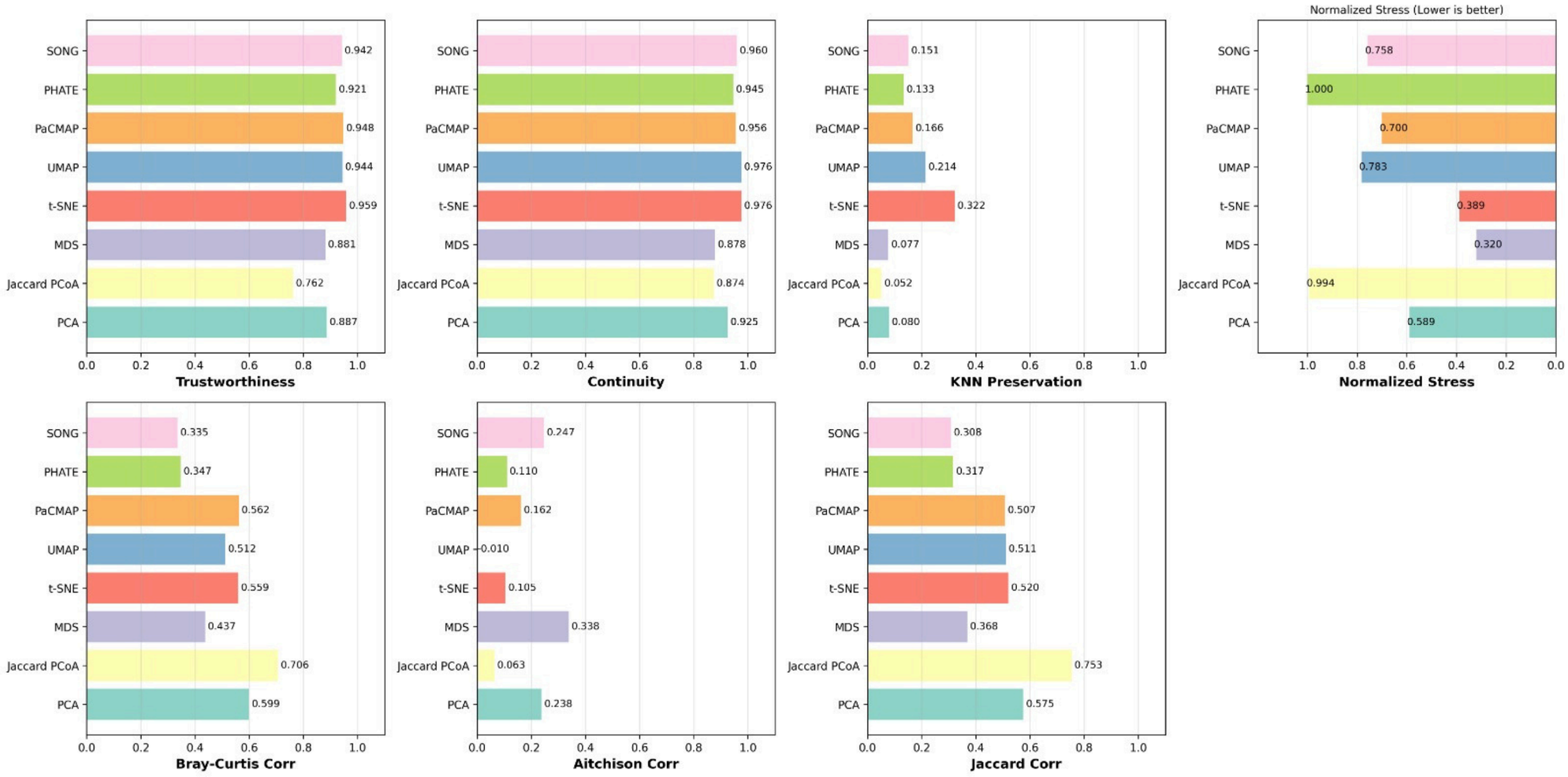


Fig. 3. Comparative visualization of the human metagenomics dataset using multimetric evaluation metrics.

IMPLEMENTATION-VISUALIZATION

2D Visualization of Soil Metagenomics Data

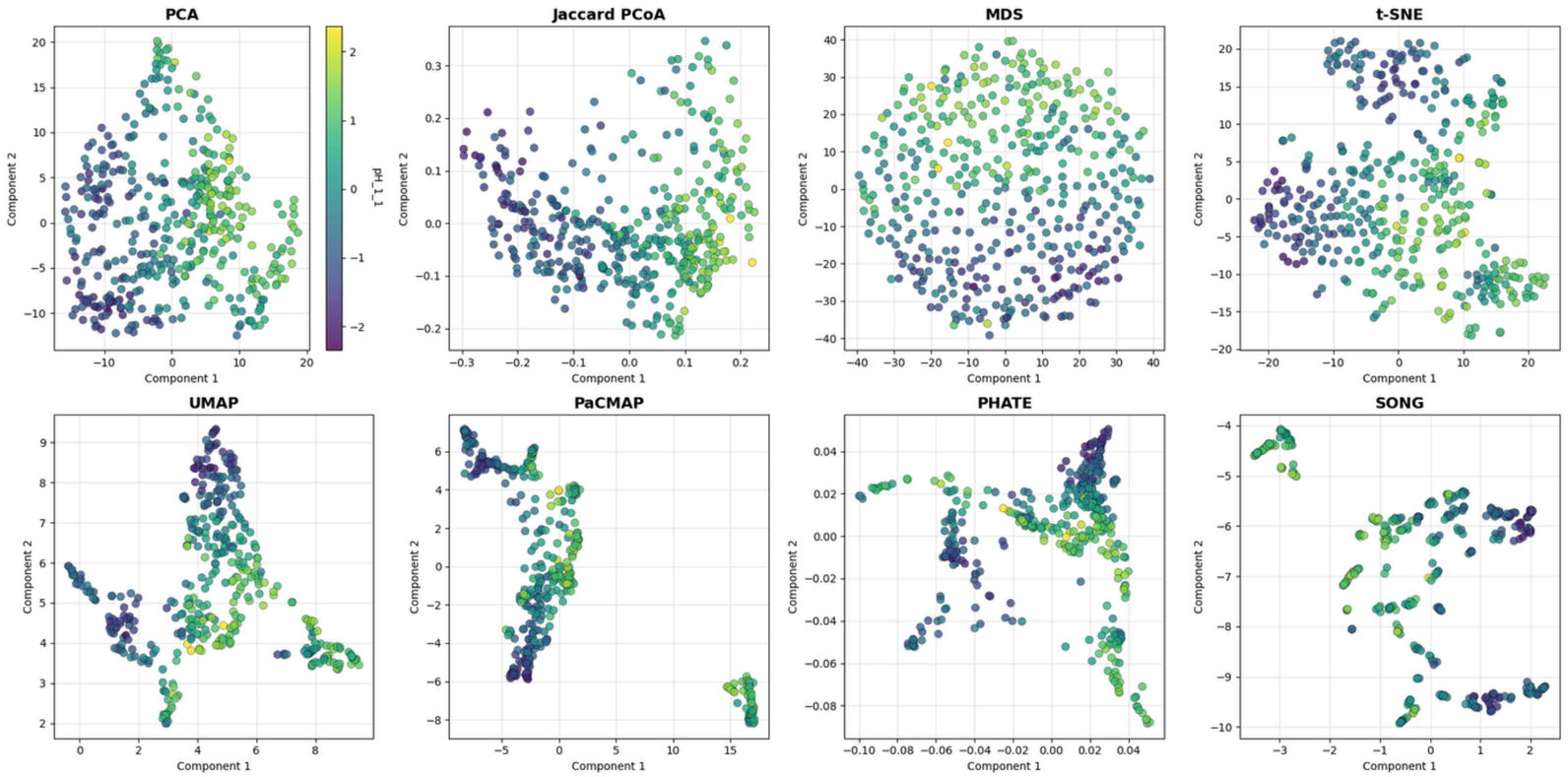


Fig. 4. Visualization of the soil metagenomics dataset using eight dimensionality reduction (DR) methods.

IMPLEMENTATION-EVALUATION

Evaluation Metrics for Soil Metagenomics Data

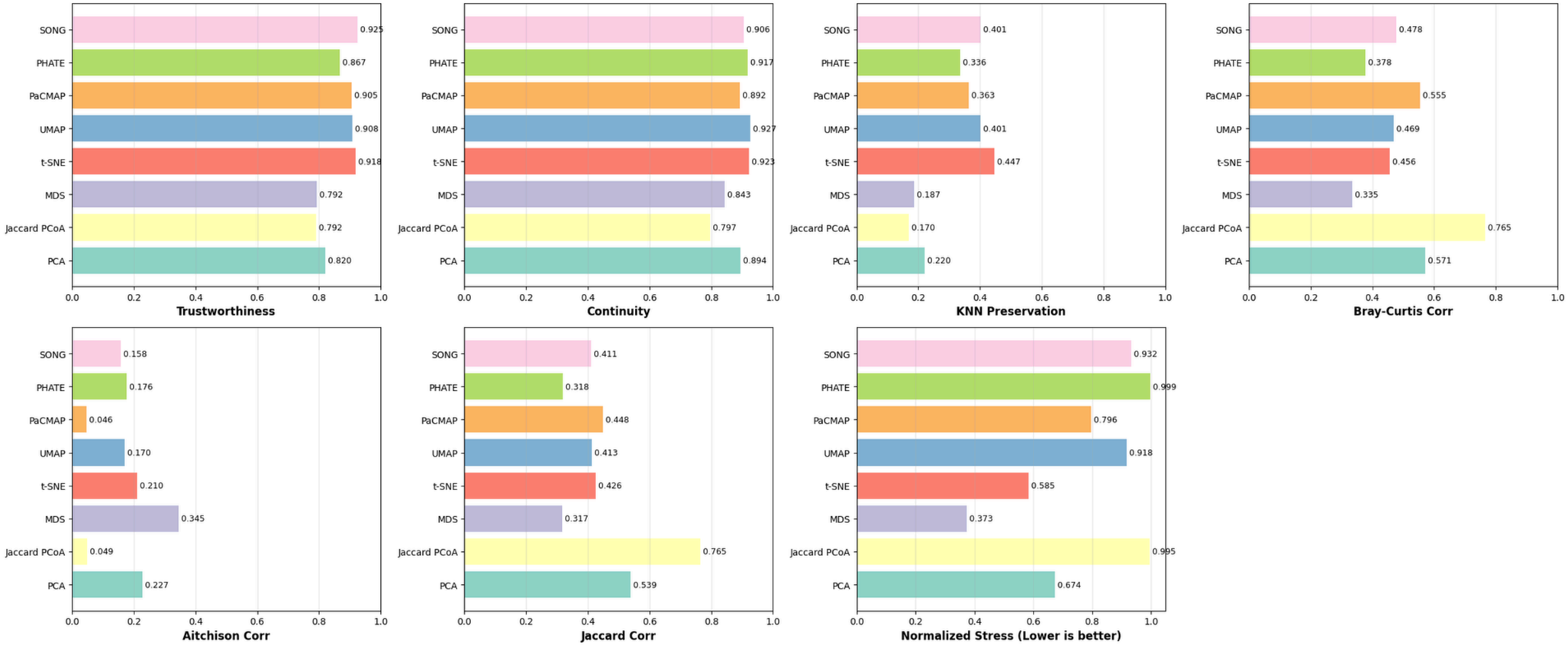


Fig 5. Comparative visualization of the soil metagenomics dataset using multimetric evaluation metrics.

IMPLEMENTATION-VISUALIZATION

2D Visualization of Marine Metagenomics Data

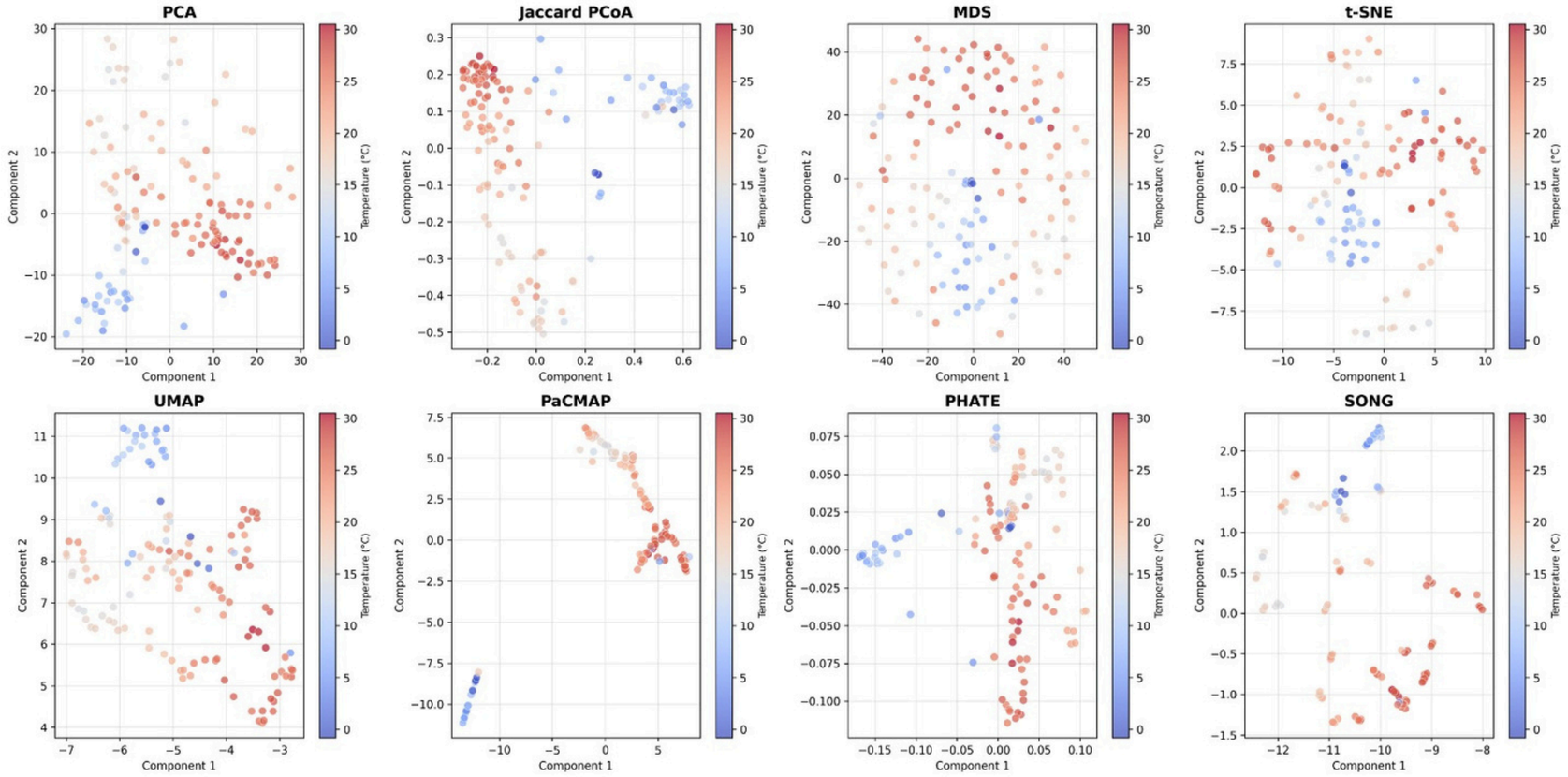


Fig. 6. Visualization of the marine metagenomics dataset using eight dimensionality reduction (DR) methods.

IMPLEMENTATION-EVALUATION

Evaluation Metrics for Marine Metagenomics Data

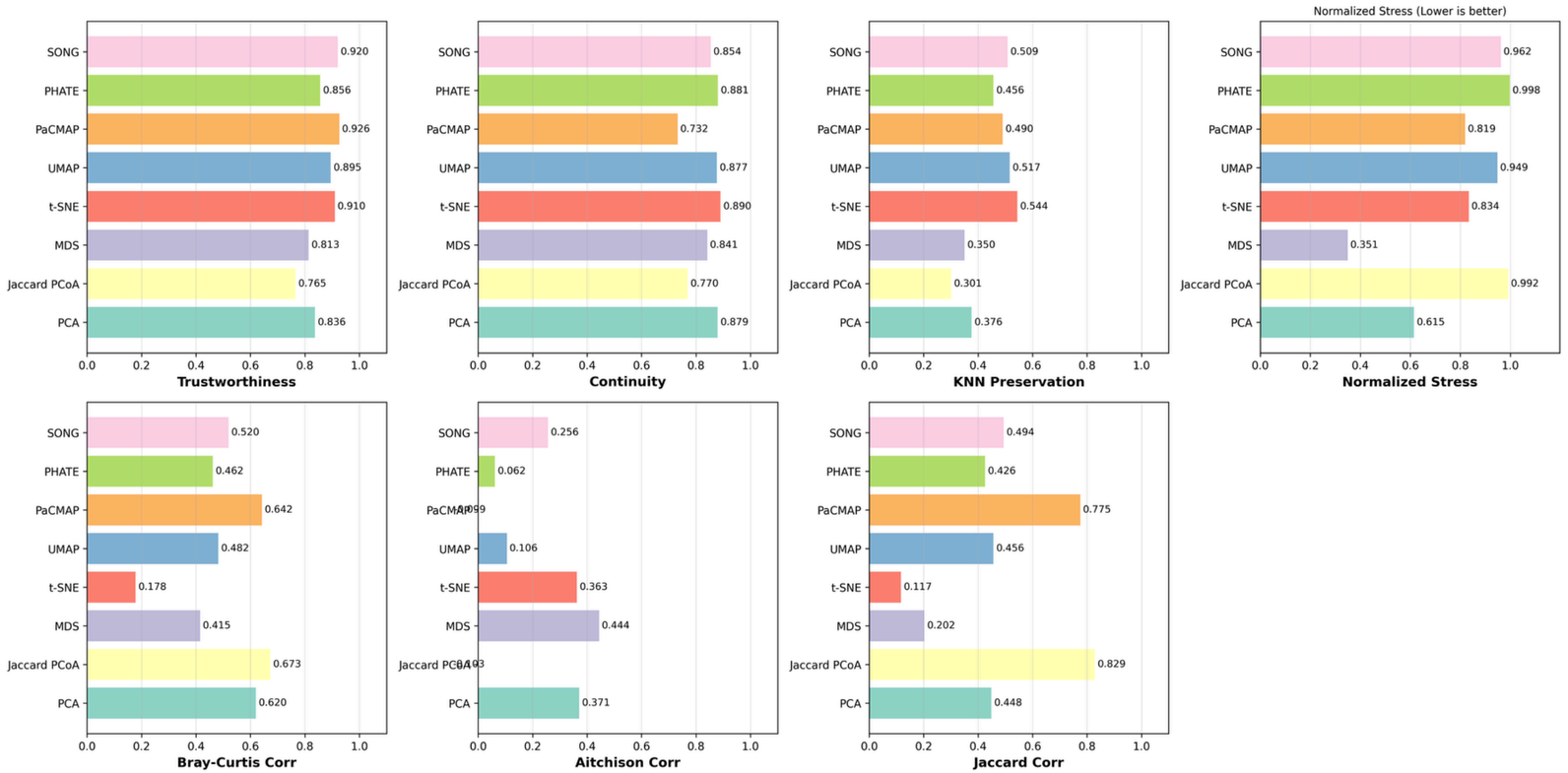


Fig. 7. Comparative visualization of the Marine metagenomics dataset using multimetric evaluation metrics.

KEY FINDINGS

- Quantitative evaluation across datasets confirms clear performance trade-offs among dimensionality reduction methods.
- Linear methods (PCA, Jaccard PCoA) preserve high-dimensional proximity but fail to resolve subtle non-linear structures.
- Manifold methods (t-SNE, UMAP) produce clear cluster separation but often at the cost of global metric fidelity.

KEY FINDINGS (CONTD...)

- PHATE and PaCMAP effectively capture continuous environmental gradients, while SONG preserves mixed discrete and continuous structures.

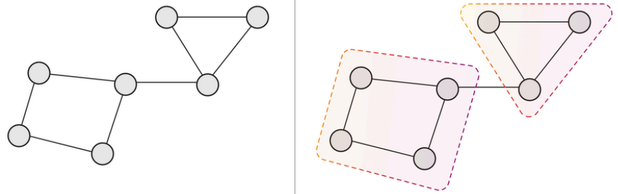
- Threshold-based evaluation shows that no single method satisfies all local, global, and biological distance criteria.

Overall, The best technique depends on whether the analysis prioritizes cluster separation, global structure preservation, or ecological distance relationships.

PHASE 2

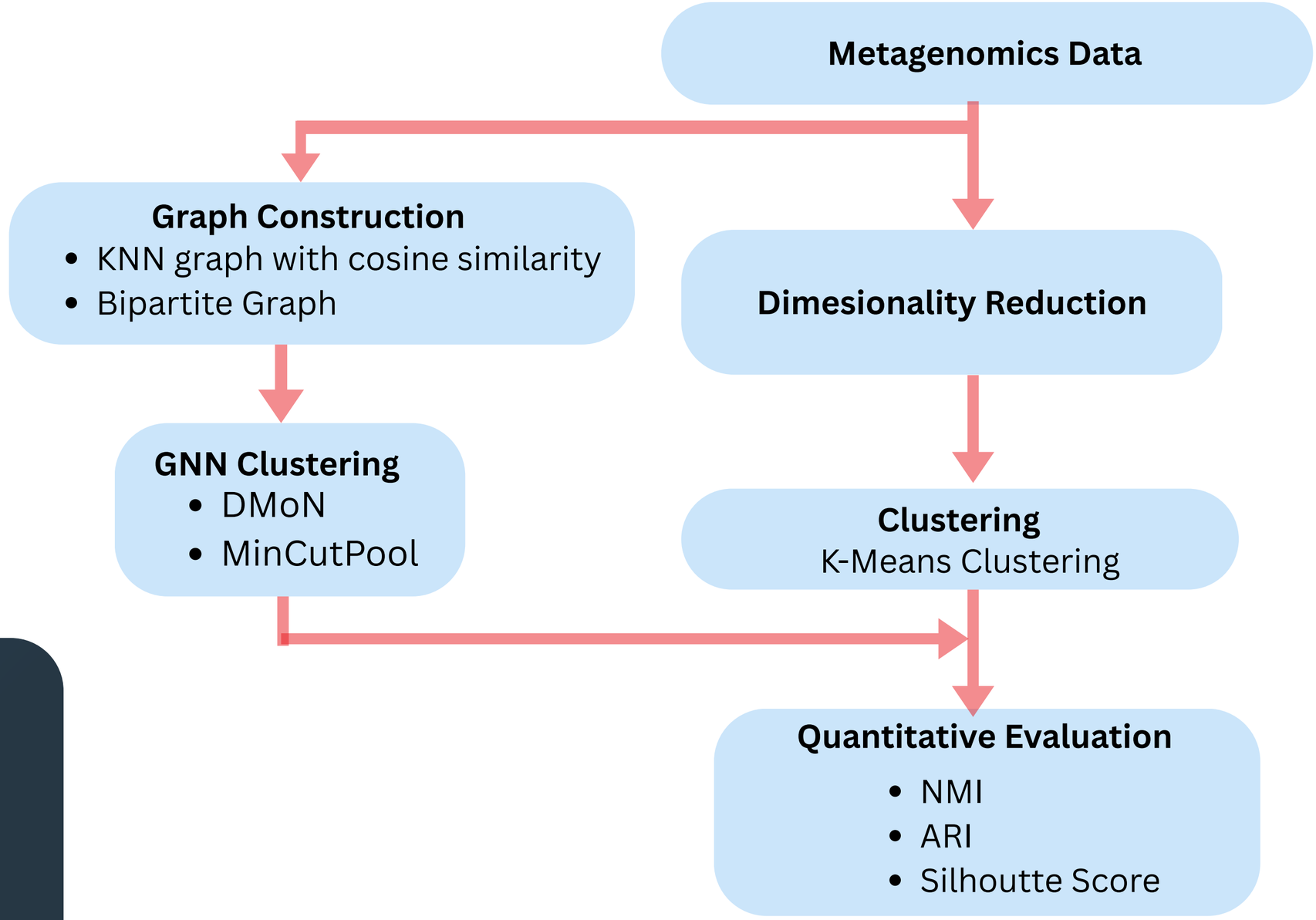
Graph Neural Network Based End-to-End Clustering for Metagenomic Data

- Construct Graph representations
- Implement end-to-end GNN clustering models
- Compare against traditional clustering pipeline
- Clustering evaluation framework



PROPOSED METHODOLOGY

Phase 2 – GNN Based Clustering



KNN WITH COSINE SIMILARITY GRAPH FOR HUMAN METAGENOMICS DATASET

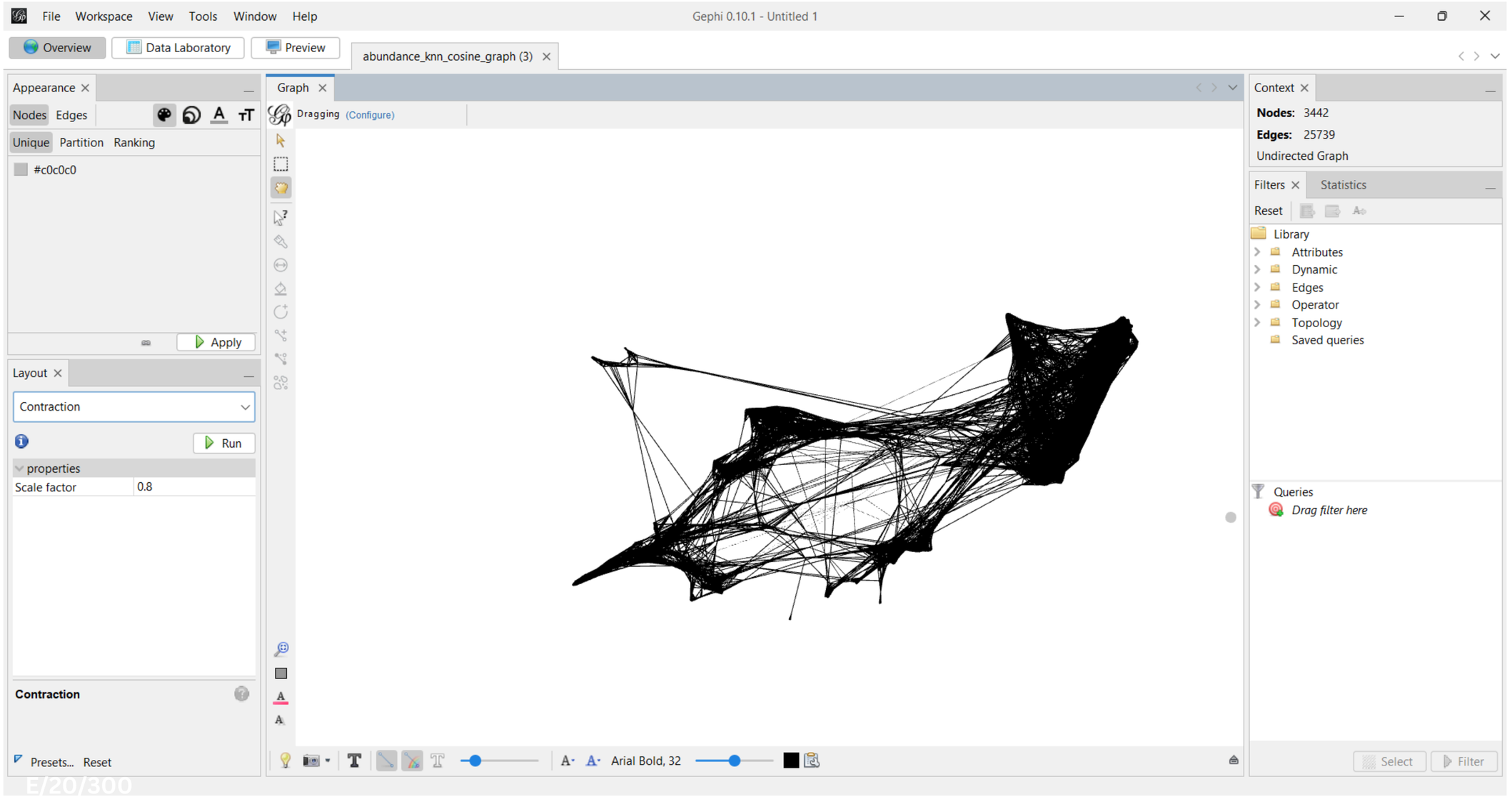


Fig. 8. KNN with Cosine Similarty graph for the human metagenomics dataset

KNN WITH COSINE SIMILARITY GRAPH FOR SOIL METAGENOMICS DATASET

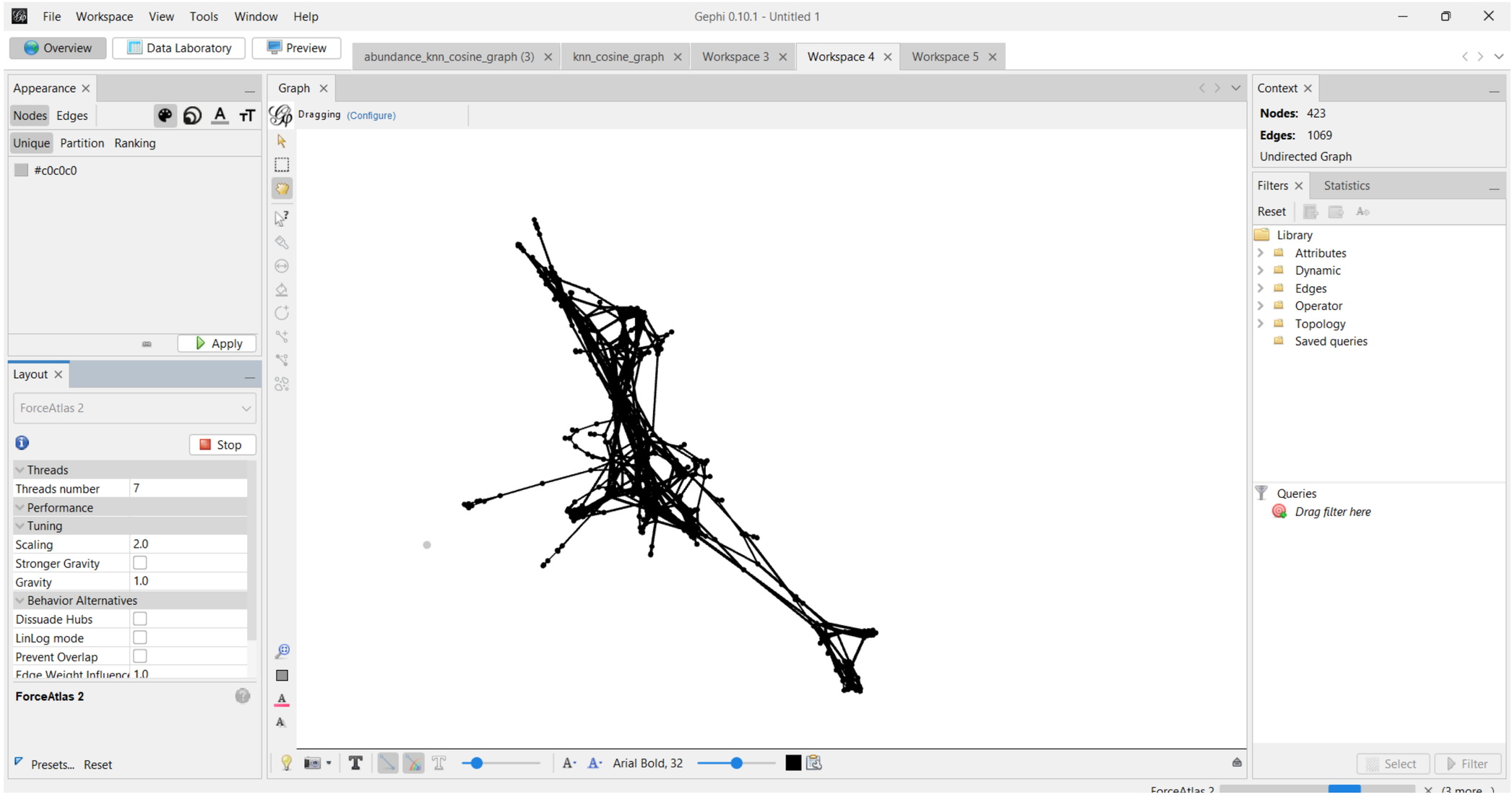
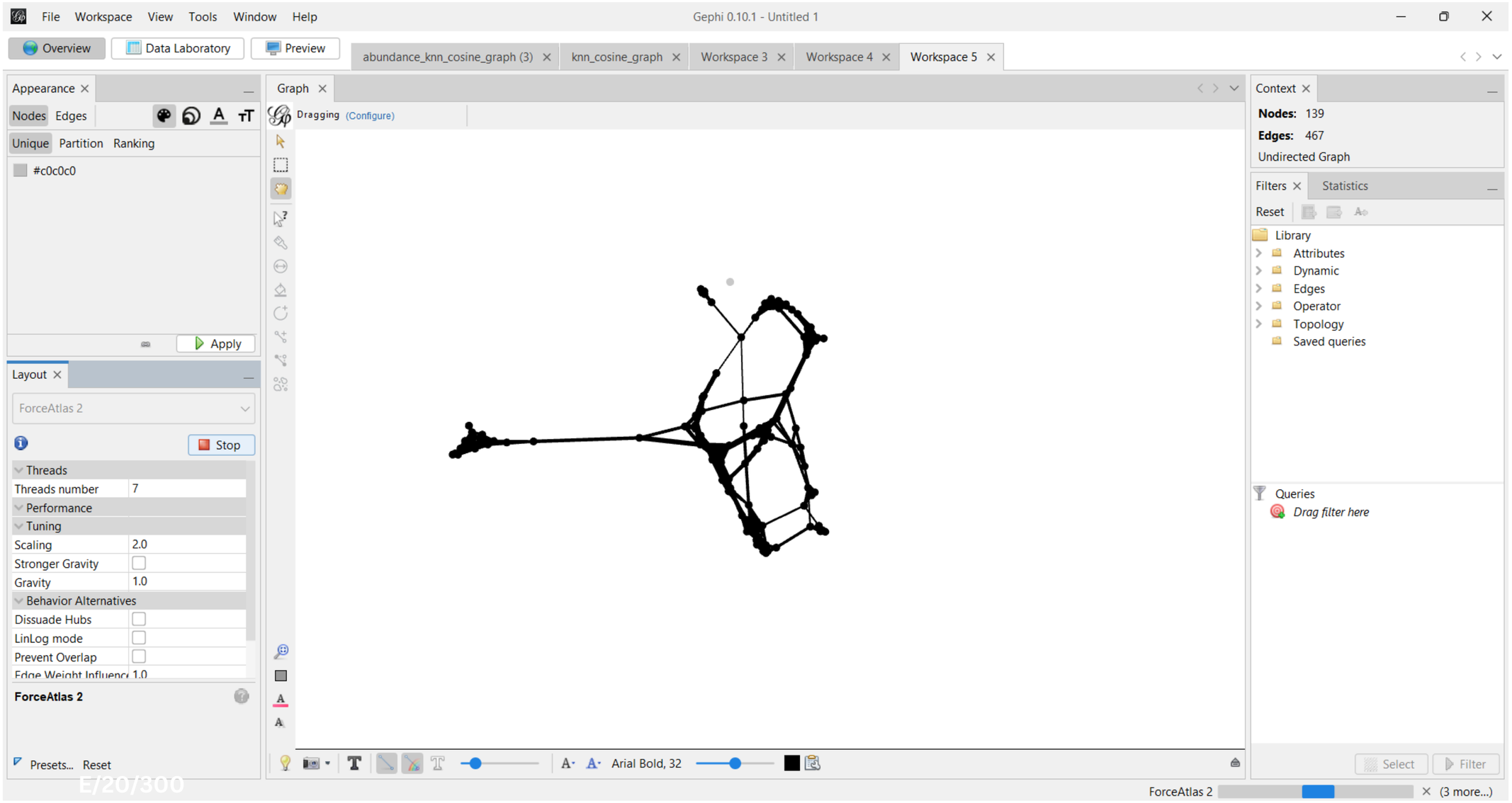


Fig. 8. KNN with Cosine Similarty graph for the soil metagenomics dataset

KNN WITH COSINE SIMILARITY GRAPH FOR MARINE METAGENOMICS DATASET

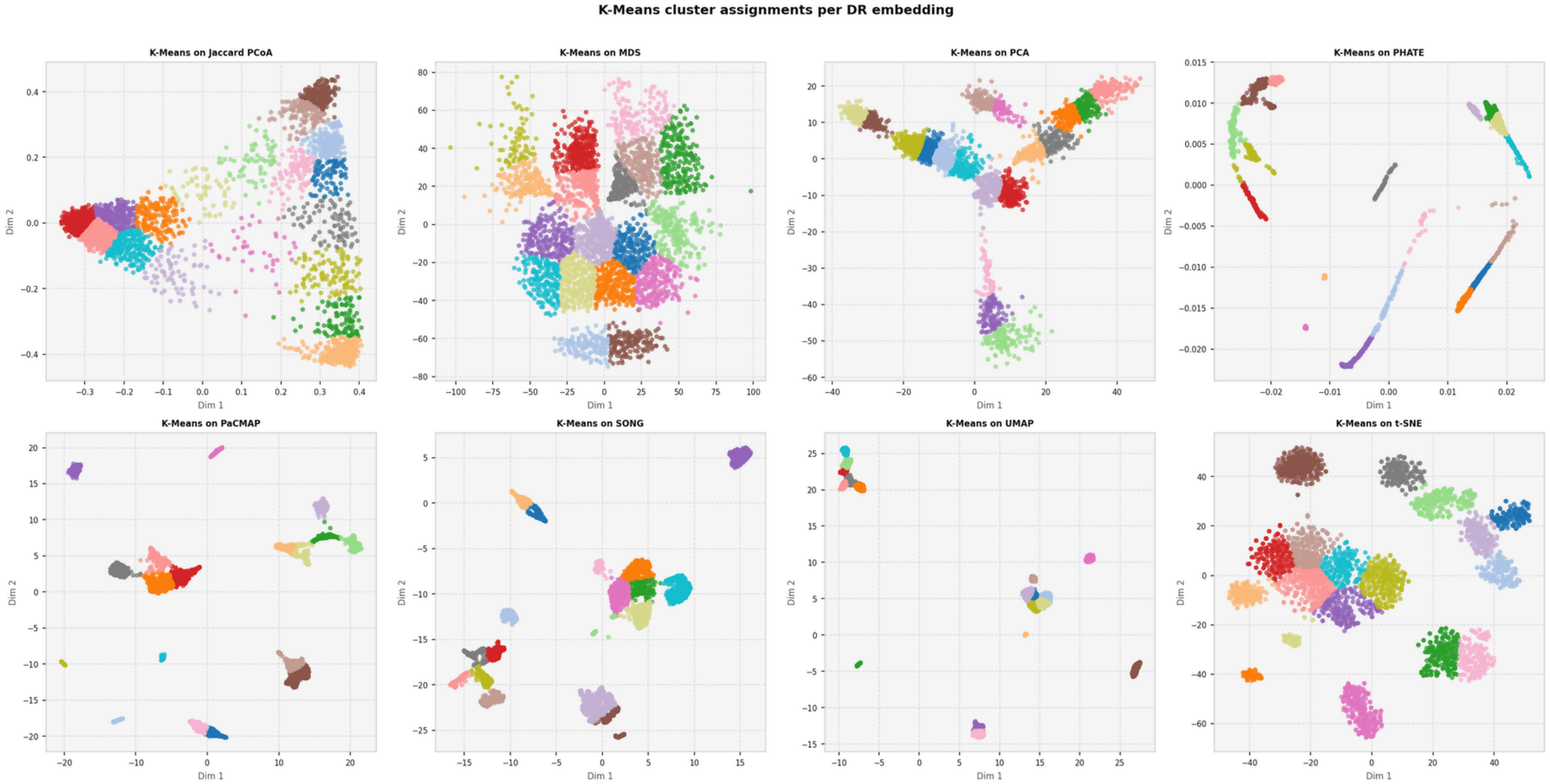


E/20/300

Fig. 8. KNN with Cosine Similarty graph for the marine metagenomics dataset

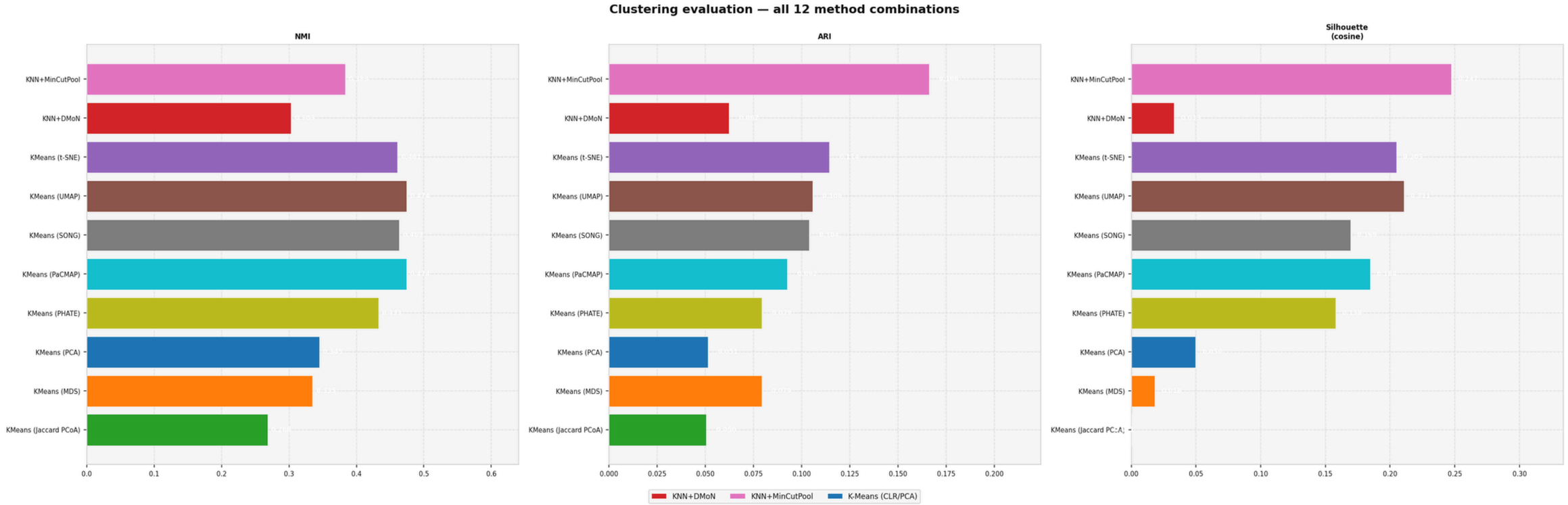
IMPLEMENTATION-VISUALIZATION

Cluster Visualization of Human Metagenomics Data



IMPLEMENTATION-EVALUATION

Evaluation Metrics for Human Metagenomics Data

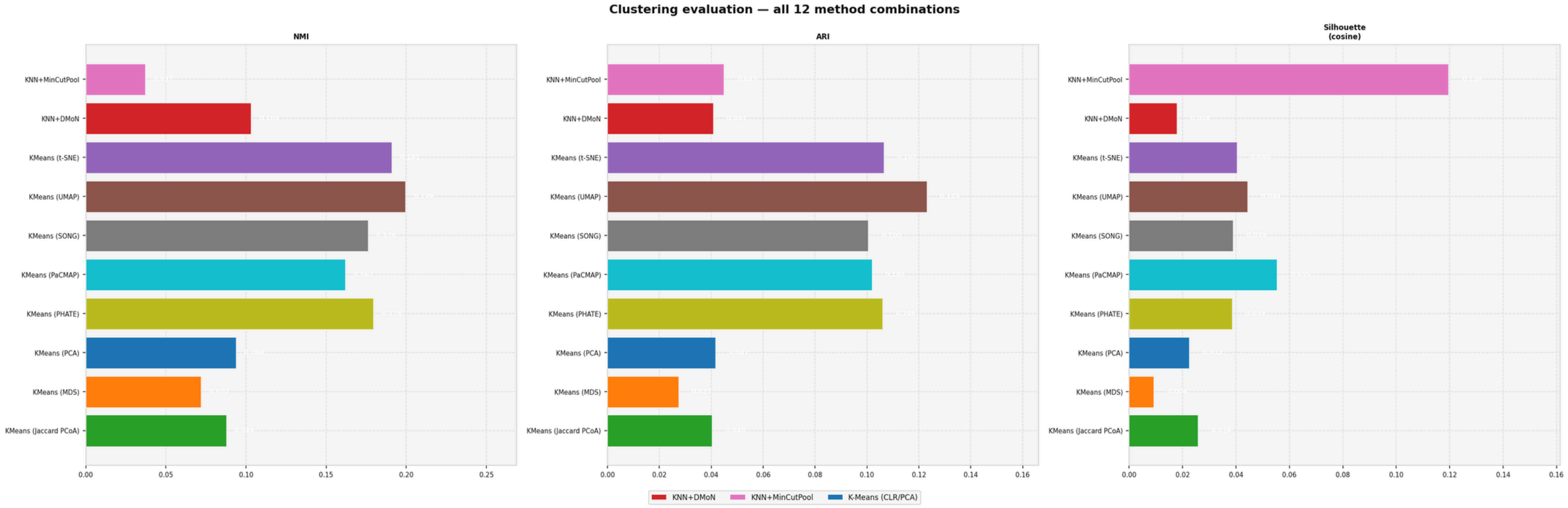


IMPLEMENTATION-VISUALIZATION

Cluster Visualization of Soil Metagenomics Data

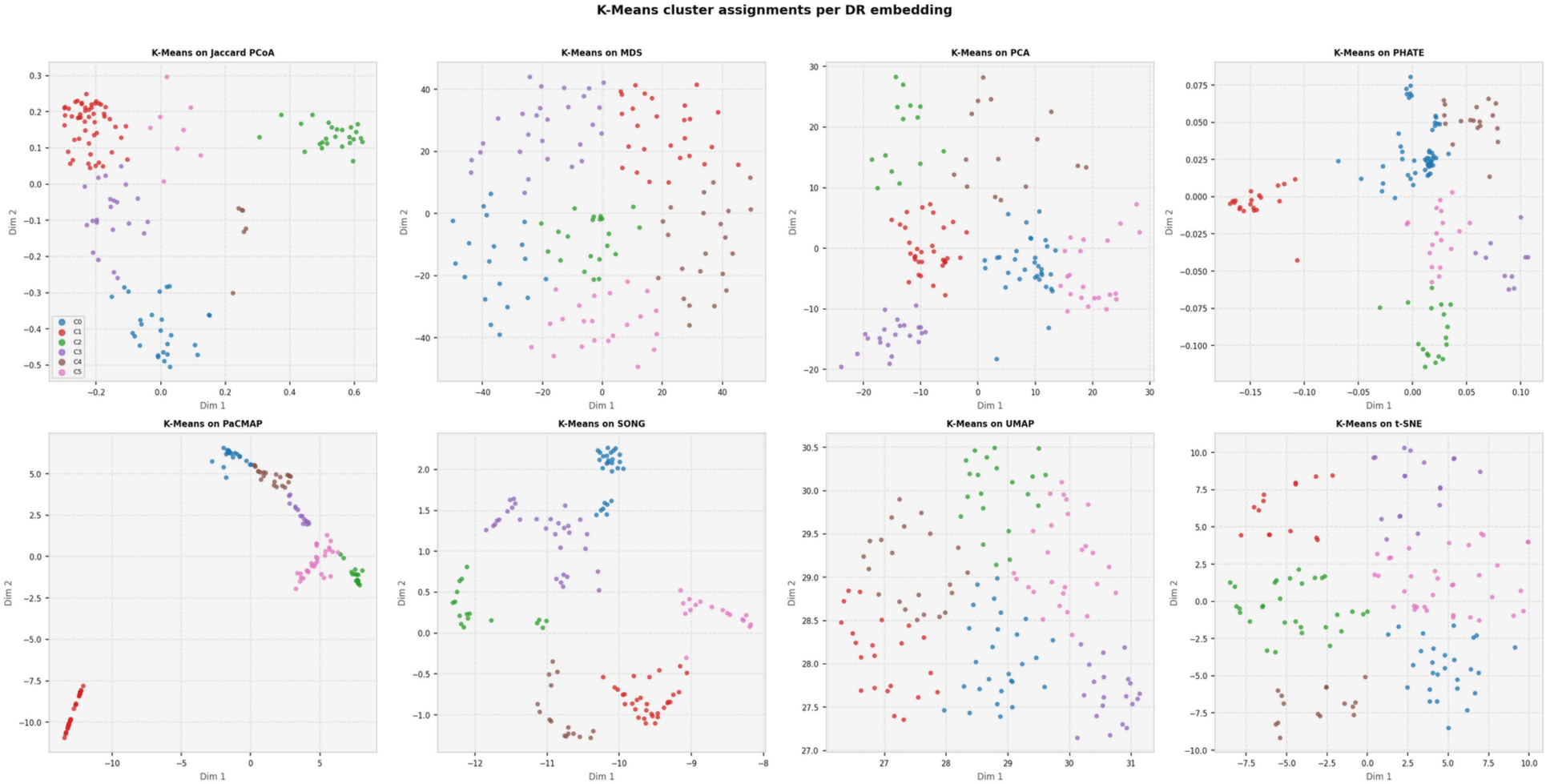
IMPLEMENTATION-EVALUATION

Evaluation Metrics for Soil Metagenomics Data



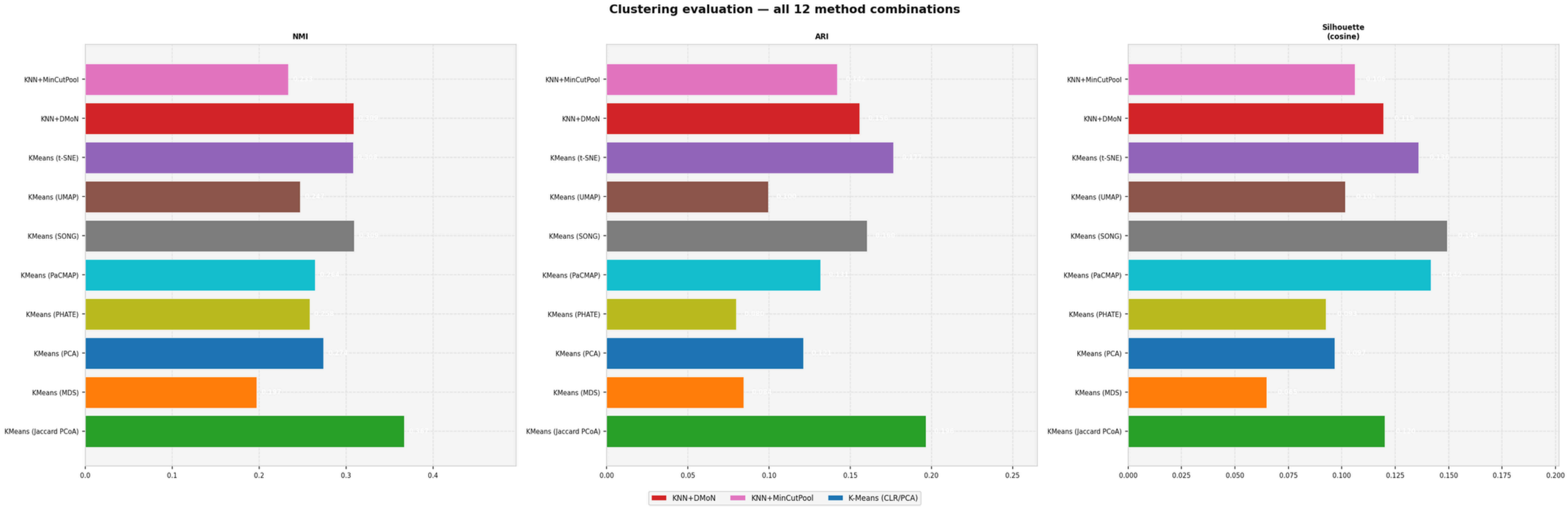
IMPLEMENTATION-VISUALIZATION

Cluster Visualization of Marine Metagenomics Data



IMPLEMENTATION-EVALUATION

Evaluation Metrics for Marine Metagenomics Data



KEY FINDINGS

Human Data

GNN (MinCutPool) clearly outperforms all methods (NMI, ARI, Silhouette).

Soil Data

Mixed results — GNN strong in Silhouette, while UMAP/PHATE competitive in NMI & ARI.

Marine Data

Traditional methods (Jaccard PCoA) perform better in NMI & ARI.

GNN Performance

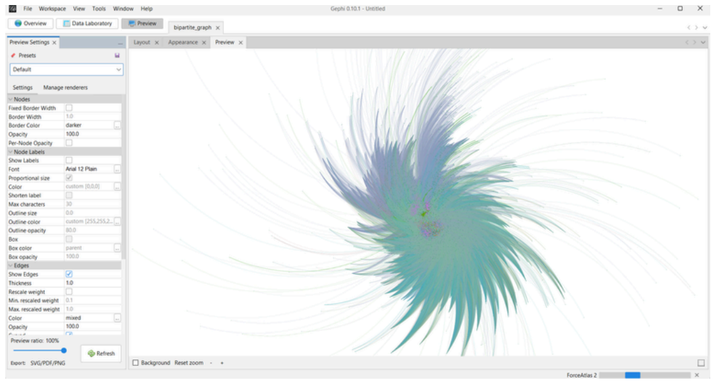
MinCutPool > DMoN; produces more compact clusters.

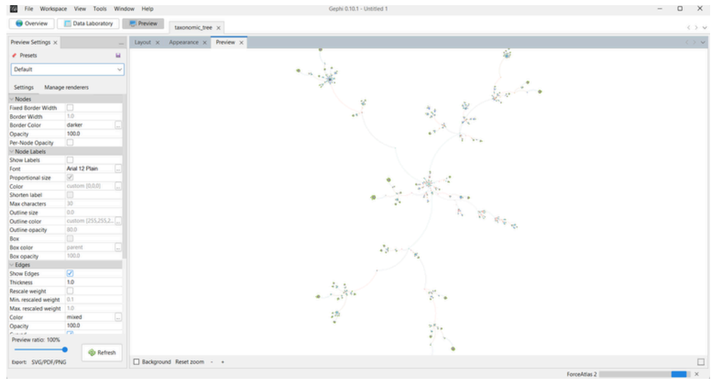
Overall Insight

GNNs are robust and effective, especially for metagenomics datasets, but not always the best.

LIMITATIONS

- Phylogenetic Graph could not be effectively constructed for the given sample data
- Bipartite Graph produced inaccurate or misleading relationships
- Results highly depend on target selection.





RESEARCH PUBLICATION – IEEE XPLORE

Evaluating Topology Preservation in Dimensionality Reduction Methods for Metagenomic Data: A Comparative Analysis

1st Chamuditha Jananga
Department of Computer Engineering
University of Peradeniya
Sri Lanka
tgcjananga@gmail.com

2nd Tharushika Prasadinie
Department of Computer Engineering
University of Peradeniya
Sri Lanka
prasadinietharushika@gmail.com

3rd Pasindu Malshan
Department of Computer Engineering
University of Peradeniya
Sri Lanka
pasindumalshan237@gmail.com

4th Vijini Mallawaarachchi
College of Science and Engineering
Flinders University
Australia
vijini.mallawaarachchi@flinders.edu.au

5th Rajith Vidanaarachchi
Faculty of Architecture
University of Melbourne
Australia
rajith.v@unimelb.edu.au

6th Damayanthi Herath
Department of Computer Engineering
University of Peradeniya
Sri Lanka
damayanthiherath@eng.pdn.ac.lk

***Abstract*—Metagenomics employs high-throughput sequencing and computational methods to analyze microbial communities across diverse environments. This research comparatively evaluates dimensionality reduction (DR) methods for metagenomics by addressing the inherent computational challenges of high sparsity, compositionality, and dimensionality. We present a comparative study evaluating 8 DR techniques: Principal Component Analysis (PCA), Multidimensional Scaling (MDS), t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE), Self-Organizing Nebulous Growths (SONG), Pairwise Controlled Manifold Approximation Projection (PaCMAP), and Jaccard Principal Coordinates Analysis (PCoA). The data were processed via a robust preprocessing pipeline integrating non-zero Center Log-Ratio (nzCLR) transformation with iterative matrix completion. We utilize a comprehensive evaluation framework that incorporates both local and global topological preservation metrics: trustworthiness, continuity, normalized stress, k-nearest neighbor preservation. For non-Gaussian and sparse metagenomic data, we incorporated correlation metrics: Bray-Curtis dissimilarity, Aitchison distance, and Jaccard index. Our findings indicate that no single DR method is universally optimal; effectiveness depends on the specific scientific question and the type of biological distance being preserved for metagenomics data.**

***Index Terms*—Dimensionality Reduction, Metagenomics, Comparative Metagenomics, Topological Preservation**

I. Introduction

Microbiome research employs metagenomics—the culture-independent sequencing and analysis of genetic material directly from environmental or host-associated microbial communities to generate high-dimensional feature tables through amplicon sequencing or shotgun metagenomics that characterize microbial community composition and function [10]. These datasets present fundamental analytical challenges: (1) high dimensionality with thousands of features relative to sample size [11], (2) extreme sparsity with 90–99% zero entries due to zero-inflation—the phenomenon where zeros arise from multiple distinct processes, and (3) compositionality, where relative abundances sum to a constant, creating mathematical dependencies between features [11], [19].

Dimensionality reduction (DR) is essential to address these challenges. By transforming high-dimensional data into lower-dimensional representations that preserve biological variation, DR techniques reduce noise, prevent overfitting, and enable meaningful visualization and interpretation of microbial community structures [1], [12], [18]. This facilitates robust downstream analyses including clustering, classification, and ecological interpretation [12], [18].

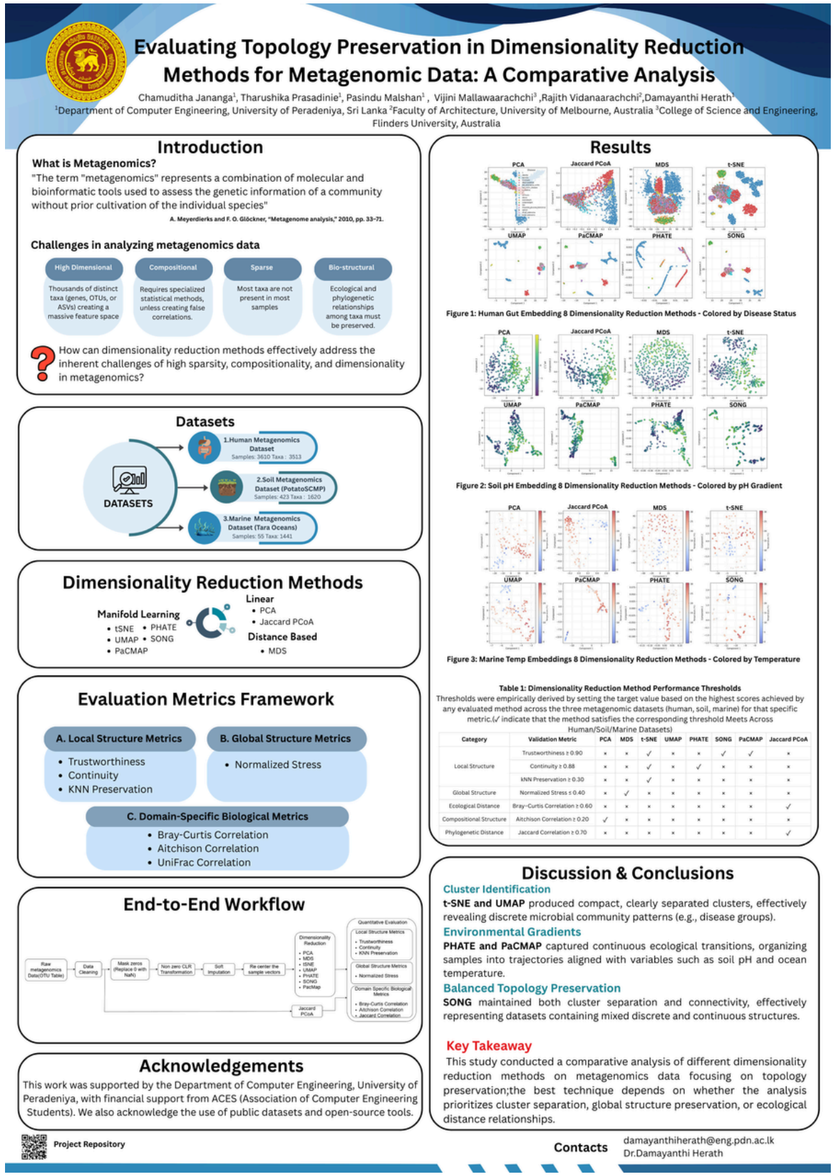
II. Related Work

Dimensionality reduction (DR) techniques have become essential for analyzing and visualizing high-dimensional metagenomic data. Classical linear methods PCA [1] and MDS [8] provide computationally efficient embeddings but often struggle to capture nonlinear relationships in complex biological data. Manifold learning methods have emerged to address these limitations: t-SNE [4] reshaped visualization through probability distribution matching and local neighborhood preservation, while UMAP [2], [3] incorporated topological data analysis principles for improved computational efficiency and global structure preservation. Recent advances include PaCMAP [6], which balances local and global structure through dynamic graph component selection, SONG [9] for incremental data visualization with noise tolerance, and PHATE, which applies diffusion processes to preserve continuous trajectories and branching structures common in developmental biology.

ICIPRoB 2026
International Conference on Image Processing and Robotics

Paper Accepted at the **ICIPRoB 2026**
To be Published in IEEE Xplore

RESEARCH POSTER



Evaluating Topology Preservation in Dimensionality Reduction Methods for Metagenomic Data: A Comparative Analysis

Chamuditha Jananga[1], Tharushika Prasadinie[1], Pasindu Malshan[1] , Vijini Mallawaarachchi[3] ,Rajith Vidanaarachchi[2],Damayanthi Herath[1]
[1]Department of Computer Engineering, University of Peradeniya, Sri Lanka [2]Faculty of Architecture, University of Melbourne, Australia [3]College of Science and Engineering, Flinders University, Australia

Introduction

What is Metagenomics?

"The term "metagenomics" represents a combination of molecular and bioinformatic tools used to assess the genetic information of a community without prior cultivation of the individual species"

A. Meyerdierks and F. O. Glöckner, "Metagenome analysis," 2010, pp. 33–71.

Challenges in analyzing metagenomics data

- **High Dimensional**: Thousands of distinct taxa (genes, OTUs, or ASVs) creating a massive feature space
- **Compositional**: Requires specialized statistical methods, unless creating false correlations.
- **Sparse**: Most taxa are not present in most samples
- **Bio-structural**: Ecological and phylogenetic relationships among taxa must be preserved.

How can dimensionality reduction methods effectively address the inherent challenges of high sparsity, compositionality, and dimensionality in metagenomics?

Datasets

DATASETS

1. Human Metagenomics Dataset — Samples: 3610 Taxa : 3513
2. Soil Metagenomics Dataset (PotatoSCMP) — Samples: 423 Taxa : 1620
3. Marine Metagenomics Dataset (Tara Oceans) — Samples: 59 Taxa: 1441

Dimensionality Reduction Methods

Manifold Learning
- tSNE
- PHATE
- UMAP
- SONG
- PaCMAP

Linear
- PCA
- Jaccard PCoA

Distance Based
- MDS

Evaluation Metrics Framework

A. Local Structure Metrics
- Trustworthiness
- Continuity
- KNN Preservation

B. Global Structure Metrics
- Normalized Stress

C. Domain-Specific Biological Metrics
- Bray-Curtis Correlation
- Aitchison Correlation
- UniFrac Correlation

End-to-End Workflow

Raw metagenomics Data(OTU Table) → Data Cleaning → Mask zeros (Replace 0 with NaN) → Non zero CLR Transformation → Soil Imputation → Re-center the sample vectors → Dimensionality Reduction (PCA, MDS, tSNE, UMAP, PHATE, SONG, PaCMap; Jaccard PCoA) → Quantitative Evaluation (Local Structure Metrics: Trustworthiness, Continuity, KNN Preservation; Global Structure Metrics: Normalized Stress; Domain Specific Biological Metrics: Bray-Curtis Correlation, Aitchison Correlation, Jaccard Correlation)

Acknowledgements

This work was supported by the Department of Computer Engineering, University of Peradeniya, with financial support from ACES (Association of Computer Engineering Students). We also acknowledge the use of public datasets and open-source tools.

Project Repository

Results

Figure 1: Human Gut Embedding 8 Dimensionality Reduction Methods - Colored by Disease Status

Figure 2: Soil pH Embedding 8 Dimensionality Reduction Methods - Colored by pH Gradient

Figure 3: Marine Temp Embeddings 8 Dimensionality Reduction Methods - Colored by Temperature

Table 1: Dimensionality Reduction Method Performance Thresholds

Thresholds were empirically derived by setting the target value based on the highest scores achieved by any evaluated method across the three metagenomic datasets (human, soil, marine) for that specific metric.(✓ indicate that the method satisfies the corresponding threshold Meets Across Human/Soil/Marine Datasets)

Category	Validation Metric	PCA	MDS	t-SNE	UMAP	PHATE	SONG	PaCMAP	Jaccard PCoA
Local Structure	Trustworthiness ± 0.90	×	×	✓	×	×	✓	✓	×
	Continuity ± 0.88	×	×	✓	×	✓	×	×	×
	KNN Preservation ± 0.30	×	×	✓	×	×	×	×	×
Global Structure	Normalized Stress ± 0.40	×	✓	×	×	×	×	×	×
Ecological Distance	Bray–Curtis Correlation ± 0.60	×	×	×	×	×	×	×	✓
Compositional Structure	Aitchison Correlation ± 0.20	✓	×	×	×	×	×	×	×
Phylogenetic Distance	Jaccard Correlation ± 0.70	×	×	×	×	×	×	×	✓

Discussion & Conclusions

Cluster Identification

t-SNE and UMAP produced compact, clearly separated clusters, effectively revealing discrete microbial community patterns (e.g., disease groups).

Environmental Gradients

PHATE and PaCMAP captured continuous ecological transitions, organizing samples into trajectories aligned with variables such as soil pH and ocean temperature.

Balanced Topology Preservation

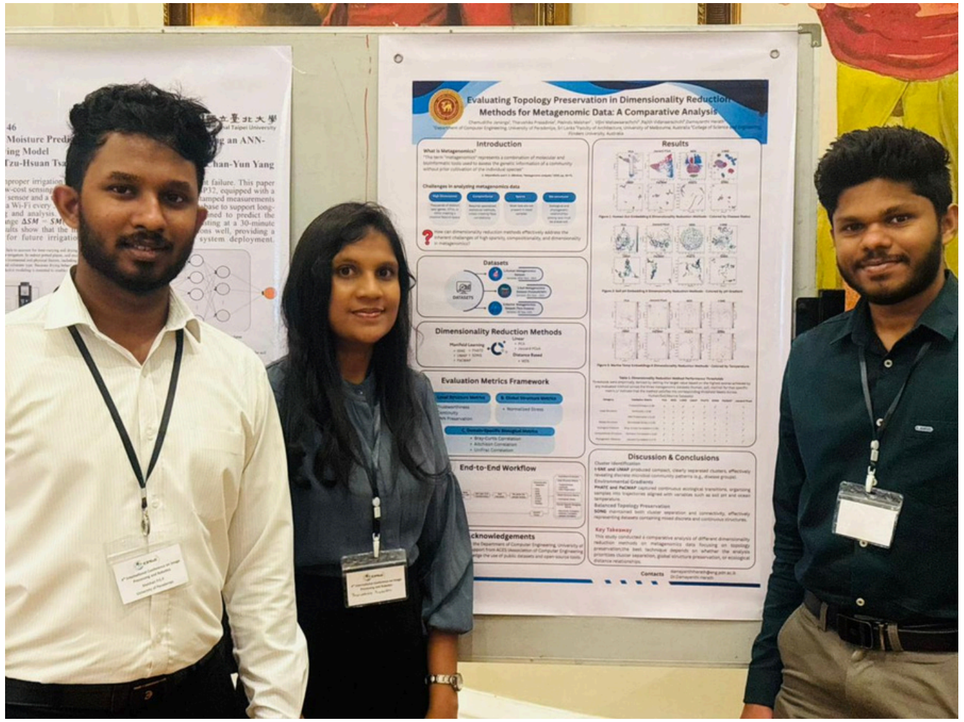
SONG maintained both cluster separation and connectivity, effectively representing datasets containing mixed discrete and continuous structures.

Key Takeaway

This study conducted a comparative analysis of different dimensionality reduction methods on metagenomics data focusing on topology preservation;the best technique depends on whether the analysis prioritizes cluster separation, global structure preservation, or ecological distance relationships.

Contacts damayanthiherath@eng.pdn.ac.lk
Dr.Damayanthi Herath

ICIPRoB 2026
International Conference on Image Processing and Robotics



EXTENSIVE WORK

Adapted our Final Year Research into creating a ML model for

- **Non-Invasive AI-Based Early Liver Cirrhosis Detection**

using preprocessing pipeline and a dataset we used

🏆 BioFusion Hackathon 2026

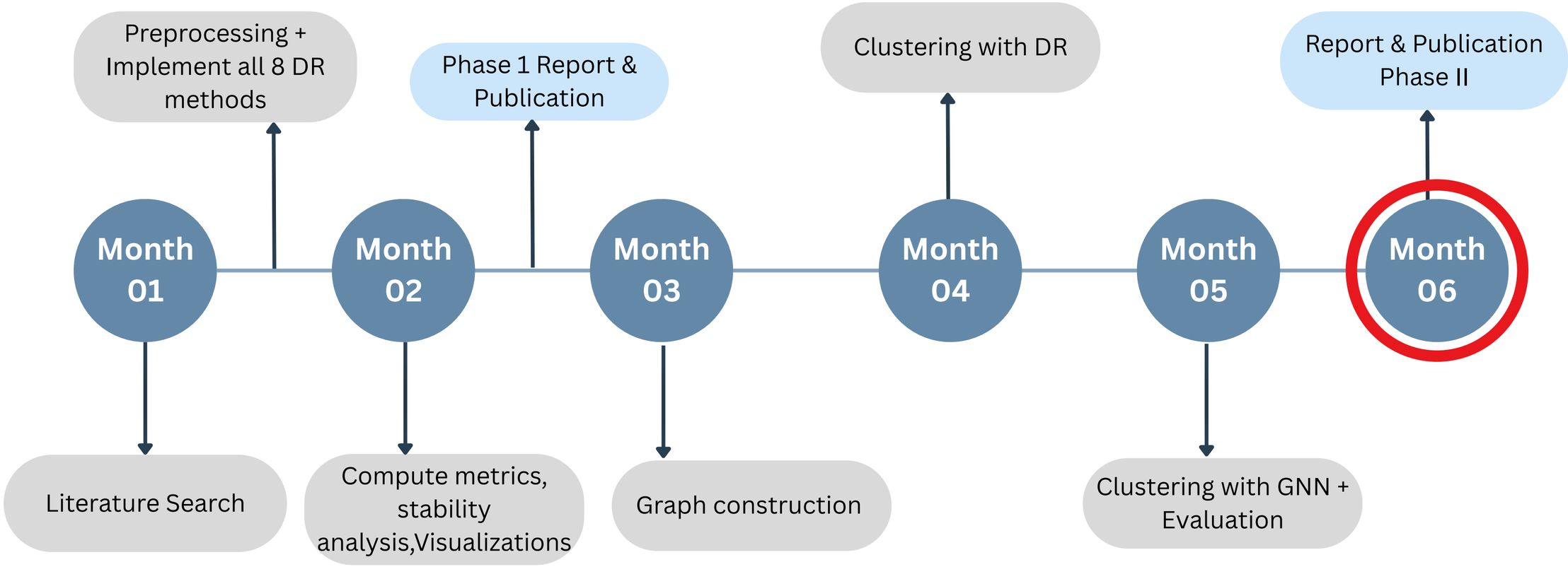
A Medical AI/ML Competition – BioFusion Science Convention

Second Runner-Up (Top 3 / 65 Teams)

FUTURE OF THE RESEARCH AREA

Develop a Novel Clustering Framework for Metagenomics to address context-specific challenges and improve clustering accuracy.

TIMELINE



Semester CS2
18 Aug - 13 Nov 2025

Semester 07
2025 Nov - 2026 Mar

Q & A


THANK YOU