OPEN ACCESS
Advances in Animal and Veterinary Sciences

Research Article

Integrative Multi-Omics and Machine Learning Framework for Assessing Microbial Functional Resilience in Contaminant-Impacted Environments Relevant to Livestock and Poultry Systems

Janardhan G[1*], Geeta Patil[2], Ugendhar Addagatla[3], M. Madhavi Latha[4], P. Naresh[5], Sivananda Lahari Reddy Elicherla[6], Praveen Kulkarni[7]

[1]Department of CSE, Vignan Institute of Technology and Science, Hyderabad; [2]Department of Information Technology, Army Institute of Technology, Pune; [3]Department of Information Technology, MVSR Engineering College, Nadergul, Hyderabad; [4]Department of CSE (Data Science), G.Narayanamma Institute of Technology and Science for Women, Hyderabad; [5]Department of Computer Science and Engineering, Dayananda Sagar University, Bengaluru; [6]CMR University, Bengaluru; [7]Dayananda Sagar University, Bengaluru,

Abstract | The increasing exposure of environmental contaminants such as heavy metals, petroleum hydrocarbons, microplastics, and pharmaceutical residues severely disrupts microbial communities in aquatic and terrestrial ecosystems. These complex mixtures threaten ecosystem stability, nutrient cycling, and environmental health. Traditional single-omics approaches provide only partial insights, as they fail to capture the intricate interactions between microbial taxonomic diversity, functional potential, gene expression, and metabolic responses under multi-contaminant stress.In this study, we propose an integrative multi-omics framework enhanced with advanced machine learning techniques to assess microbial community dynamics and functional responses along contaminant gradients. The framework primarily integrates metagenomic and metatranscriptomic datasets. Advanced compositional data preprocessing, similarity network fusion (SNF), and joint latent factor modeling were employed for robust multi-omics feature extraction and integration. Spatially resolved sampling was conducted across a contamination gradient in river sediments to identify pollutant-responsive microbial taxa and functional modules. Graph neural networks (GNNs) and multimodal machine learning architectures were applied in the downstream analysis to uncover latent patterns, predict microbial functional shifts, and evaluate community resilience under varying contamination levels. A real-world case study involving 24 sediment samples collected across a 12 km contamination gradient demonstrated that the proposed integrative multi-omics and machine learning approach significantly outperformed conventional single-omics analyses and simple data integration methods. The framework successfully identified contaminant-specific microbial signatures and revealed critical cross-omics associations between microbial taxa, functional genes, and metabolic pathways. This study provides a scalable and predictive platform for understanding microbial ecosystem responses in contaminated environments, offering valuable insights for environmental monitoring, risk assessment, and bioremediation strategies.

Keywords | Bioremediation, Environmental microbiology, Multi-omics integration, Latent factor modeling, Microbial diversity, Contaminated ecosystems.

Received | April 14, 2026; **Accepted** | May 02, 2026; **Published** | July 01, 2026
***Correspondence** | Janardhan G, Department of CSE, Vignan Institute of Technology and Science, Hyderabad; **Email:** janardhanyadav5@gmail.com.edu.sa
Citation | Janardhan G, Patil G, Addagatla DU, Latha MM, Naresh P, Elicherla SLR, Kulkarni P (2026). Integrative multi-omics and machine learning framework for assessing microbial functional resilience in contaminant-impacted environments relevant to livestock and poultry systems Adv. Anim. Vet. Sci., 14(7):1382-1395.
DOI | https://dx.doi.org/10.17582/journal.aavs/2026/14.7.1382.1395
ISSN (Online) | 2307-8316

INTRODUCTION

Biogeochemical cycling, organic matter decomposition, and nutrient transformations in practically all terrestrial and aquatic ecosystems are regulated by microbial communities. These microscopic creatures have been known to perform some of the most basic ecological functions such as carbon fixation, nitrogen cycling, sulfur transformation and organic matter degradation. Their key contribution to the attenuation of contaminants in the forms of biodegradation, biotransformation, and immobilization has led to extensive efforts to define the composition, functional capacity, and activity of microbiomes inhabiting contaminated environments (Wang *et al*., 2025). The mechanisms that govern the responses of microbial consortia to anthropogenic contaminants are thus necessary to predict ecosystem outcomes and to plan effective interventions of bioremediation processes.

Polluted ecosystems pose unique analytical problems that are not common to laboratory environments. The contamination in real world usually consists of complex mixtures of heavy metals, hydrocarbons, chlorinated solvents, microplastics, and emerging organic contaminants like pharmaceuticals and perfluoroalkyls (Chen *et al*., 2024). These constituents do not interact in an additive manner and they have synergistic or antagonistic effects which are not foreseeable by the studies of single contaminants. The responses of these contaminants are also determined by the heterogeneous environmental factors such as pH, redox potential, nutrient supply, organic carbon content, and hydrological conditions. Spatial heterogeneity also makes it more difficult to make inferences, as concentrations of contaminants and the structure of microbial communities change radically between Fixed inconsistent hyphenation. This multi-scale heterogeneity requires sampling designs that are able to capture both large gradients and local variability (LeBrun *et al*., 2019).

Traditional culture-based techniques only reveal a tiny portion of the extant microbial diversity, which is usually less than one percent of the whole community. Single-marker amplicon sequencing (e.g., 16S rRNA gene sequencing) displays taxonomic composition with no direct data on functional capacity or activity. These techniques fail to distinguish between those cells that are active in metabolism and those that are dormant or dead, and do not tell us which metabolic pathways are active in a particular environmental setting. Single-omics analyses - metagenomics, meta transcriptomics, meta proteomics or metabolomics (Shi *et al*., 2024) alone can give partial pictures of the system which are necessarily incomplete. The closer proxies of phenotype are meta proteomics and metabolomics, which directly quantify proteins and metabolites, but do not give the genomic context that relates observed molecules to their source organisms. Therefore, single-omics research leave a significant gap in the mechanistic link between exposure to contaminants and community functional responses (Rodriguez *et al*., 2023).

Integrative multi-omics provides a principled solution to these basic constraints. Multi-omics frameworks can answer the question of which taxa contain which genes, which genes are expressed, the relationship between expression and protein abundance, and how metabolic fluxes vary in response to environmental perturbations by jointly analyzing two or more molecular layers on the same samples. Indicatively, research has effectively combined metagenomics and meta transcriptomics to monitor functional responses in response to bioremediation, combined meta proteomics and metabolomics to detect toxicity mechanisms, and employed all four layers of omics to describe the response of a microbial community to environmental disturbance.

Irrespective of these developments, the field-scale applications in complex contaminated systems are very scarce. The vast majority of multi-omics experiments have been performed in controlled laboratory microcosms, or well-characterized clinical environment, where environmental variability can be reduced. When these methods are adapted to field conditions, they bring in other issues such as spatial heterogeneity, temporal variability, confounder spatial environmental gradients, and logistical limitations of sampling and sample treatment.

These methods do not consider the various statistical characteristics of each layer of omics like count distributions, levels of sparsity and structure of measurement errors. This dichotomous system narrows down on the possibility of describing dose-response relationships, discovering threshold effects, and forecasting responses of the community at intermediate levels of contamination (Smith *et al*., 2022). Third, no rigorous systematic quantitative benchmarking of integrative multi-omics compared to single-omics baselines in polluted ecosystems has been conducted which leaves much uncertainty regarding the benefit of integration to prediction and mechanistic inference.

This research fills these gaps in three significant contributions. We introduce an integrative multi-omics pipeline to contaminated ecosystems designed to support spatially resolved sampling across contamination gradients, compositional data preprocessing with centered log-ratio transformations, cross-omics integration with similarity network fusion and joint latent factor modeling with environmental regularization. Secondly, we offer a detailed quantitative evaluation of the role integrative analysis plays

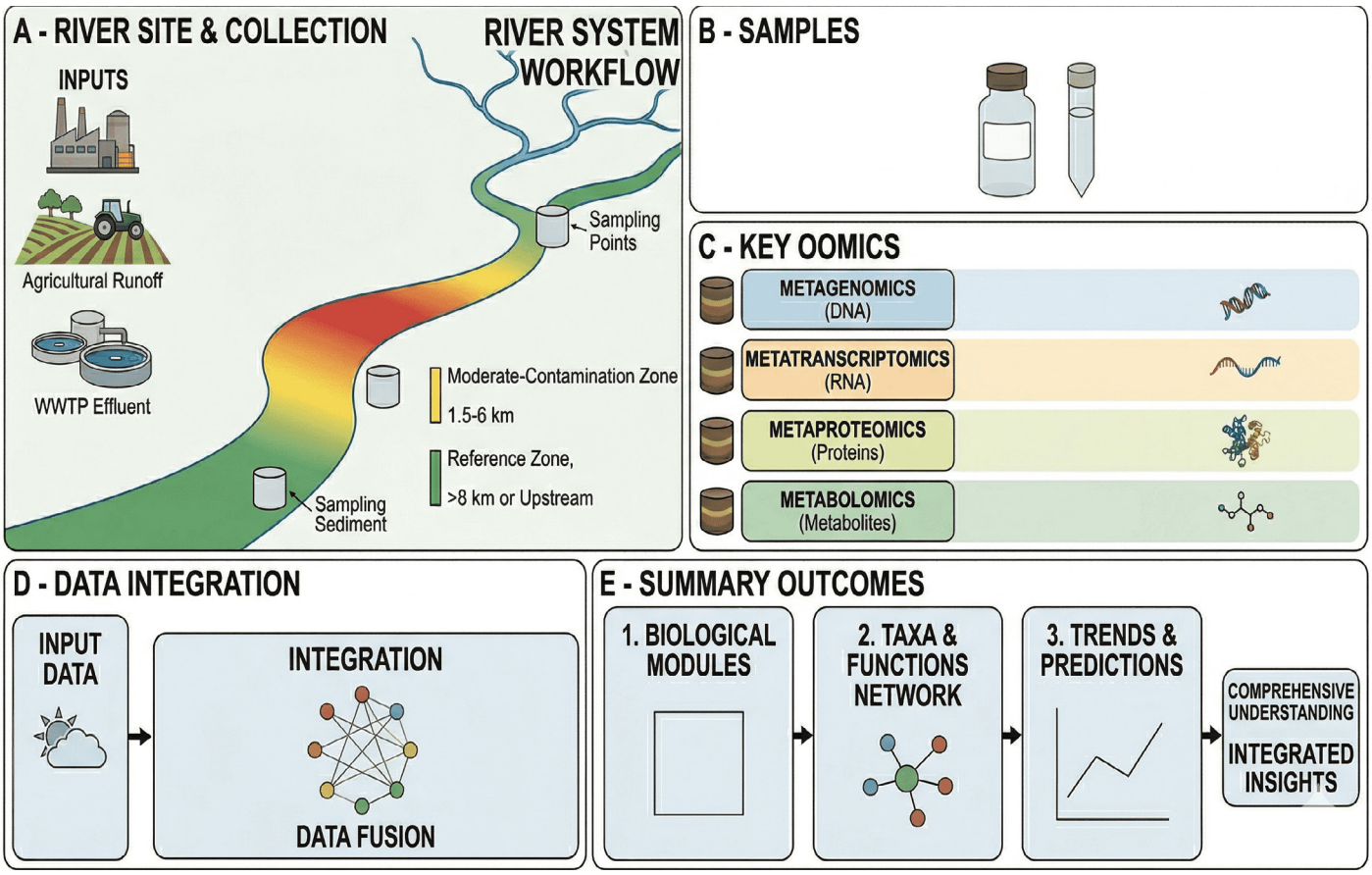

Figure 1: Multi-omics integration framework for contaminated ecosystems. (a) Contaminant sources introduce pollutants into river sediments. (b) Spatial sampling along high, moderate, and reference zones. (c) Multi-omics data layers integrated via similarity network fusion to produce contaminant-responsive modules and predictive models.

in improving the resolution of microbial diversity patterns, functional pathway changes and contaminant-responsive modules over single-omics baselines (Kumar *et al.*, 2024). Third, we show that the integrated framework can be used to provide better explanatory and predictive accuracy of contaminant concentrations and ecological functions, with the use of a realistic contaminated sediment case study with 24 sampling sites across a 12-kilometer gradient resulting in a template of mechanism-informed environmental monitoring and bioremediation design. Figure 1 illustrates Multi-omnics framework for eontaminated ecosystem.

Environmental microbial ecology: omics applications

High-throughput sequencing technologies have revolutionized environmental microbial ecology. Metagenomics can be used to reconstitute population genomes of polluted soils, which contain new lineages able to break down recalcitrant pollutants. MAGs assembly and binning have become the norm. Nevertheless, metagenomics does not allow distinguishing between actively transcribed genes and silent genes, which is a critical drawback to predict remediation outcome (Zhao *et al.*, 2023).

Transcriptional responses of hydrocarbon-impacted sites using meta transcriptomics reveals that transcriptional responses are usually different than genomic potential. Transcripts of nitrate reduction, methanogen, and beta-oxidation enrichment in polluted sediments revealed a decrease in diversity, studies of the Detroit River contaminant gradient (Wang *et al.*, 2023). Conditionally expressed pathways are identified by meta transcriptomics, which is also challenged by high rRNA content and short half-lives of mRNAs. The metabolic pathways during metal stress and biomarker metabolites are explained by meta proteomics and metabolomics. These layers are more phenotype-proximate and have such limitations as a lack of reference databases and the high turnover of metabolites (Liu *et al.*, 2024).

Multi-omics integration methodologies

There is significant progress in multi-omics integration methodology. Early correlation-based methods produced very high false-discovery rates and did not take into account various statistics of each layer. Network-centric designs enhanced signal detection using common sample neighborhoods (Zhang *et al.*, 2023).

Similarity network fusion (SNF) builds sample-sample

similarity matrices and recursively merges networks through neighborhood propagation, which learns shared structure and retains layer-specific information at the same time.

Multi-block PLS and CCA explicitly capture the covariance of omics blocks (Bauer *et al*., 2022). Sparse variants solve the instability problem in cases where features are larger than samples. Latent variable models are learned in modern times with shared low-dimensional representations and support various data types. Integrated analysis of rhizosphere systems contaminated with petroleum was carried out in the environment whereby the predominant hydrocarbon-degrading taxa were found to be Pseudomonas, Rhodococcus and Brevundimonas (Fan *et al*., 2025). There is limited literature that has applied these methods in field-contaminated ecosystems and provided benchmarking in a systematic manner (Kim *et al*., 2024).

Omics in environmental risk assessment and bioremediation

Omics methods are used to inform environmental risk assessment. Multi-omics techniques are useful in capturing synergistic or antagonistic interactions that were not previously detected in single-chemical assays (Dubey *et al*., 2024).

The BACSIN project proved that there are several stressors that trigger the activation of the bacteria under different gene expression programs in response to the stressor (Singh *et al*., 2024). Dual-omics analysis indicates intricate community responses that cannot be elucidated by metagenomics; meta transcriptomics is necessary to elucidate lively organisms (Rodriguez et al., 2023). Nevertheless, the majority of studies perform comparisons and not continuous gradient models, which does not emulate dose-response relationships. Single-omics (43 percent of studies) is still more prevalent compared to multi-omics (13 percent) meaning it is direly needed in terms of applications.

Problem definition and system overview

Problem definition

Although multi-omics have the potential to play a vital role in environmental microbiology, a number of essential gaps still exist:

Gap 1: Lack of Continuous Gradient Modeling.The majority of studies are based on discrete categorical comparisons as opposed to the modeling of continuous gradients of contamination. This dichotomous method does not reflect dose-response relationships, threshold effects, and gradual ecological changes. The modeling of community responses and the establishment of essential thresholds cannot be done without constant gradient modeling.

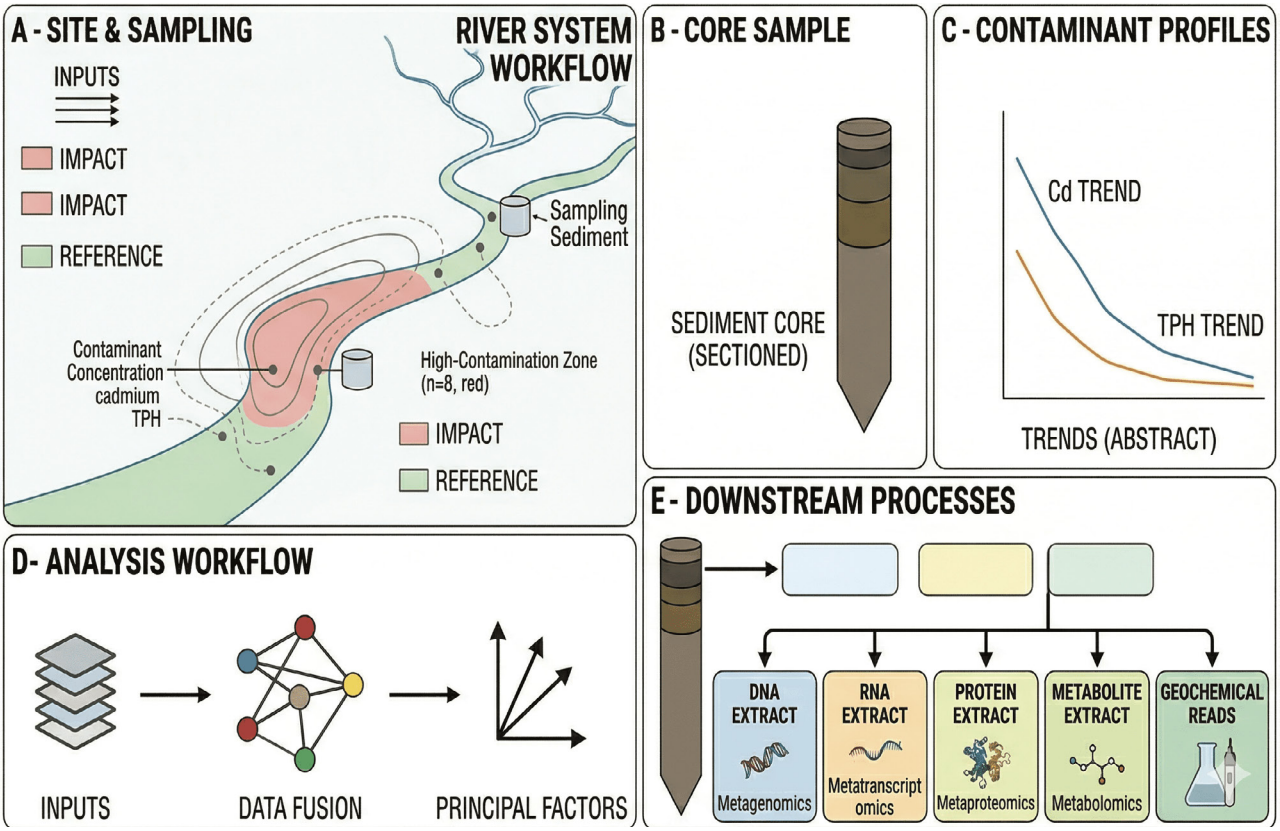

Figure 2: Environmental characterization and sampling design. (A) map of river transects having 24 sampling sites (red: high, yellow: moderate, green: reference). (B) Cadmium and hydrocarbon downstream concentration. (C) PCA of environmental variables. (D) Correlation heatmap. (E) Sediment core subsampling for multi-omics analyses.

Gap 2: Statistically unfounded Integration Frameworks. The existing applications use three less than optimal methods: independent analysis, naive feature concatenation, or pairwise correlation. None of them explicitly models covariance across layers of omics and including environmental covariates. A framework that honours specific layer characteristics, and inter-layer dependencies is required.

Gap 3: No Systematic Comparisons with Single-Omics Baselines.The value addition of multi-omics integration has not been strictly measured. There are no cross-validated comparisons or ablation analyses of claims, which undermines confidence in the usefulness of multi-omics to monitor the environment.

Gap 4: Minimal Incorporation of Environmental Covariates.Current techniques lack the spatial environmental covariates (pH, redox, organic carbon) needed to make sense of field site dynamics. Techniques which use covariates as regularization targets are required.

Gap 5: Field-Scale Deficit in Validation.The majority of multi-omics studies are done in laboratory microcosms and not in the field. The application of the methodology in field-scale applications in complex, heterogeneous contaminated ecosystem is an exceptionally rare occurrence, and there is a gap between the methodological development and field application.

System architecture components

The proposed framework will consist of five key components arranged in a sequence pipeline to fill in the gaps identified as shown in Figure 2:

Component 1: Spatially Resolved Sampling Design. The framework starts with a sampling plan that is explicitly set up to measure continuous contamination gradients, as opposed to discrete contaminated-versus-reference categories. The sampling sites are placed following the gradient of contaminant concentration between high-impact areas and transitional areas to reference conditions with an adequate replication at each site to address within-site variance. At every site, replicate sediment cores are collected and subsampled for several omics layers (DNA, RNA, proteins, metabolites) and geochemical measurements. This design allows contaminant concentration to be dose-response modelled and threshold detected because the design gives a continuous spectrum of contaminant concentrations instead of discrete categories.

Component 2: Omics-Specific Preprocessing Pipeline. Every type of omics data is processed in a specialized manner which is specific to its properties. The quality filtered metagenomic data are assembled, binned in MAGs and annotated functionally. Metatranscriptomic data undergo quality curation, mapping to metagenomic assemblies and normalization. Features of metaproteomic and metabolomic data are detected, annotated, and normalized. Importantly, all matrices in omics are converted to a centered log-ratio (clr) form to overcome compositional limitations of sequencing and mass spectrometry data. The value of the clr-transformed feature number i in a sample of size with p features with counts c1, c2, c3, etc. is:

$$x_i = \ln\left(\frac{c_i}{g(\boldsymbol{c})}\right), \quad g(\boldsymbol{c}) = \left(\prod_{i=1}^{p} c_i\right)^{1/p}$$

and g (c) is the geometric mean of features. This conversion puts all omics layers onto a compatible scale to be further integrated and maintains relative abundance information.

Component 3: Similarity Network Fusion. For each omics layer k, a sample-sample similarity matrix **W**^((k))∈R^(n×n) is constructed using a scaled exponential similarity kernel:

$$W_{ij}^{(k)} = \exp\left(-\frac{\rho^2\left(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(k)}\right)}{\mu\epsilon_{ij}}\right)$$

These similarity matrices capture how similar each pair of samples is based on each omics layer individually. The matrices are then iteratively fused via neighborhood propagation:

$$\mathbf{W}_{t+1}^{(k)} = \frac{1}{K-1}\sum \mathbf{P}^{(l)}\,\mathbf{W}_t^{(l)}\mathbf{P}$$

Where; **P**^((l)) are normalized similarity matrices and t indexes iterations. The fused similarity matrix **S** captures consensus sample relationships across omics layers.

Component 4: Joint Latent Factor Modeling with Environmental Regularization**.** A joint latent factor model is fit to the omics data with regularization guided by the fused similarity structure. Let **X**^((k))∈R^(n×p_k) denote the data matrix for omics layer k. The objective function i s:

$$\min_{\mathbf{Z},\{\mathbf{W}^{(k)}\}} \sum_{k=1}^{K} \|\mathbf{X}^{(k)} - \mathbf{Z}\mathbf{W}^{(k)\top}\|_F^2 + \lambda_1 \|\mathbf{Z}\|_F^2 + \lambda_2 \sum_{i<j} S_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$$

where Z∈R^(n×r) is the shared latent factor matrix, W^((k))∈R^(p_k×r) are layer-specific loadings, and ‖·‖_F denotes the Frobenius norm. Environmental variables are incorporated as covariates in downstream regression.

Component 5: Network Construction, Module Detection, and Benchmarking**.** Cross-omics association networks are constructed from loading correlations, and functional modules are detected using Louvain community detection. Systematic benchmarking against single-omics baselines

and ablation variants is performed using ten-fold cross-validation with metrics including R-squared, classification accuracy, F1-score, and area under the curve (AUC).

MATERIALS AND METHODS

Study sites, sampling design, and environmental data

Sampling sites were set up along this gradient (24 sites) as eight sites in the high-contamination region (within 1.5 kilometers of discharge points), eight sites in the moderate-contamination region (1.5 to 6 kilometers downstream), and eight control sites (upstream and more than 8 kilometers downstream with contaminant concentrations below regional background thresholds). Three replicate sediment cores (0 to 15 centimeters depth) were taken in the field at each site under base-flow conditions in order to reduce hydrological variability. Cores were homogenized in the field and subsamples immediately stored to be analyzed by DNA, RNA, and geochemistry (flash freezing of DNA and RNA, field-clean containers of geochemistry).

Environmental variables measured for each sample included heavy metal concentrations (cadmium, lead, zinc, copper) measured by inductively coupled plasma mass spectrometry (ICP-MS) with detection limits of 0.05 μg/g for cadmium, 0.1 μg/g for lead, 0.5 μg/g for zinc, and 0.2 μg/g for copper. Total petroleum hydrocarbons were measured gravimetrically.pH was measured electrometrically, redox potential using a platinum electrode, dissolved oxygen using an optical sensor, total organic carbon using the Walkley-Black method, and extractable nitrate and phosphate using colorimetric methods. All contaminant and environmental data were log-transformed to reduce skew, centered, and scaled to unit variance to enable comparison across variables with different units. Outliers exceeding four standard deviations from the mean were removed, resulting in the exclusion of two samples total.

Omics generation and preprocessing

Metagenomic

DNA was extracted using a phenol-chloroform protocol optimized for sediment matrices. Paired-end sequencing (2 × 150 bp) on Illumina NovaSeq 6000 yielded 48 million reads per sample. Reads were trimmed (Phred ≥ 20, length ≥ 75 bp) using fastp. Human sequences were removed via Bowtie2 alignment to GRCh38. Assembly used MEGAHIT (k-mers 21-99). Contigs ≥ 1000 bp were binned using MetaBAT 2 and CONCOCT, refined with DAS Tool, and dereplicated at 95% ANI using dRep. MAG taxonomy used GTDB-Tk. Functional annotation against KEGG, COG, and CAZy used DIAMOND (e ≤ 1e-5). Gene abundances were generated by Bowtie2 mapping.

Table 1: Mapping of framework components to identified research gaps.

Research gap	Addressed by	Mechanism
Gap 1: Absence of continuous gradient modeling	Component 1	Spatially resolved sampling along concentration gradient enables dose-response modeling and threshold detection
Gap 2: Lack of principled integration frameworks	Components 3 and 4	Similarity network fusion + joint latent factor modeling with explicit covariance structure and compositional data handling
Gap 3: No systematic benchmarking	Component 5	Ten-fold cross-validation against single-omics and ablation baselines with multiple performance metrics
Gap 4: Limited environmental covariate incorporation	Component 4	Environmental regularization term and covariate regression for disentangling contaminant effects
Gap 5: Field-scale validation deficit	Case study application	Realistic contaminated sediment system with 24 sites along 12-km gradient

Table 1 states about Mapping of framework components to identified research gaps.

Meta proteomics and metabolomics (Extension)

Performed on 18 samples (six per zone). Proteins were extracted using SDT buffer, digested with trypsin (FASP), and analyzed by LC-MS/MS on Orbitrap Fusion Lumos. Peptide identification used MaxQuant (1% FDR). Metabolites were extracted with methanol:water (80:20), analyzed by HILIC-MS/MS, and annotated against HMDB and KEGG.

Compositional data transformation

All count matrices underwent centered log-ratio (clr) transformation. For a sample with counts c_1,…,c_p:

$$x_i = \ln\left(\frac{c_i}{\left(\prod_{j=1}^{p} c_j\right)^{1/p}}\right)$$

A pseudocount of one was added prior to clr transformation for zero values. After clr transformation, each matrix was scaled to unit variance.

Diversity, community structure, and functional annotation

Alpha diversity

Alpha diversity, which measures within-sample diversity, was calculated using two complementary metrics. Shannon entropy was computed as:

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

where S is the number of species (or MAGs) and p_i is the relative abundance of the i-th taxon. Shannon diversity emphasizes richness (the number of distinct taxa) and is sensitive to rare taxa. Simpson's diversity was computed as D=1-∑p_i^2, which gives greater weight to evenness (the distribution of abundances among taxa) and is less sensitive to rare taxa. Together, these metrics provide a comprehensive characterization of diversity patterns. Differences in alpha diversity between contamination zones were tested using Kruskal-Wallis non-parametric analysis of variance followed by Dunn's post-hoc test with Benjamini-Hochberg correction for multiple comparisons.

Beta diversity

Beta diversity, which measures between-sample diversity or community dissimilarity, was assessed using Bray-Curtis dissimilarity:

$$d_{BC}(j,k) = \frac{\sum_{i=1}^{p}|x_{ij} - x_{ik}|}{\sum_{i=1}^{p}(x_{ij} + x_{ik})}$$

Beta diversity was assessed using Bray-Curtis and weighted UniFrac distances. PCoA and NMDS were applied for ordination to visualize community structure.

Functional annotation

MAG gene content and transcript abundances were mapped to KEGG, COG, and pollutant-degradation databases. Pathway-level abundance matrices were constructed separately for metagenomic and metatranscriptomic layers. Differential abundance was assessed using DESeq2 and linear models with Benjamini-Hochberg FDR control at α=0.05. Pathways with fold change > 2 and adjusted p < 0.05 were considered significant.

Multi-omics integration and modeling

Problem formulation

Let X^((k))^R^(n×p_k) denote the data matrix for omics layer k (k=1,…,K), where n=24 samples and p_k is the number of features in layer k. For the primary implementation, K=2 (metagenomic potential and metatranscriptomic activity), with optional extensions to K=4 including metaproteomic and metabolomic layers. All matrices were clr-transformed and scaled to unit variance to ensure comparability across layers. This formulation assumes that each omics layer is generated from a shared low-dimensional latent variable plus layer-specific noise, which is a common assumption in multi-omics factor models.

Similarity network fusion

For each omics layer k, a sample-sample similarity matrix W^((k))∈R^(n×n) was constructed using a scaled exponential similarity kernel:

$$W_{ij}^{(k)} = \exp\left(-\frac{\rho^2\left(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(k)}\right)}{\mu\epsilon_{ij}}\right)$$

Where ρ is a way to measure distance (Euclidean distance after applying a clr transformation), μ is a scaling factor set to the median of all the distances between pairs, and ϵ_ij adjusts for local neighborhood scaling by averaging the distances to each sample's k-nearest neighbors. This kernel focuses on local structure by scaling distances based on how densely packed the local neighborhood is. These similarity matrices were then combined repeatedly using the formula:

$$\mathbf{W}_{t+1}^{(k)} = \frac{1}{K-1}\sum_{l\neq k}\mathbf{P}^{(l)}\,\mathbf{W}_t^{(l)}\mathbf{P}^{(l)\top}$$

Here, P^(l) are similarity matrices that have been normalized by dividing each row by the total of that row, and t represents the iteration number. This process lets each omics layer affect the others, with strong similarities becoming stronger and weak ones becoming weaker. The final fused similarity matrix S shows the common relationships between samples across all omics layers and is used as input for the latent factor model.

Joint latent factor model

A joint latent factor model was fit to the omics data with regularization guided by the fused similarity structure. The objective function was:

$$\min_{\mathbf{Z},\{\mathbf{W}^{(k)}\}}\sum_{k=1}^{K}\left\|\mathbf{X}^{(k)} - \mathbf{Z}\mathbf{W}^{(k)\top}\right\|_F^2 + \lambda_1\|\mathbf{Z}\|_F^2 + \lambda_2\sum_{i<j}S_{ij}\|\mathbf{z}_i - \mathbf{z}_j\|_2^2$$

where Z∈R^(n×r) is the shared latent factor matrix (r=10 latent dimensions selected via cross-validation), W^((k))∈R^(p_k×r) are layer-specific loading matrices, and ‖·‖_F denotes the Frobenius norm. The first term encourages the product ZW^((k)⊤) to approximate the observed data X^((k)). The second term (λ_1‖Z‖_F^2) is an L2 penalty that controls the complexity of Z and improves numerical stability. The third term (λ_2 ∑_(i<j)^ S_ij ‖z_i-z_j ‖_2^2) enforces that samples with high fused similarity S_ij have similar latent representations, encouraging the latent space to respect the consensus similarity structure.

Network construction and module detection

Multi-omics association networks were constructed by thresholding the absolute values of cross-layer loading correlations. Specifically, for a latent dimension r, we linked MAG i (metagenomic loading) with transcript j (metatranscriptomic loading) if the absolute correlation

between their loading vectors across latent dimensions exceeded a data-driven threshold. The threshold was set to the 95th percentile of the null distribution obtained from permuted data where sample labels were randomly shuffled 1000 times, controlling the false discovery rate at approximately 5%. Modules (communities of densely connected nodes) were detected using the Louvain algorithm, which maximizes modularity through an iterative greedy optimization. Module enrichment for bioremediation pathways was assessed via hypergeometric tests, with p-values adjusted for multiple testing using the Benjamini-Hochberg procedure.

Hyperparameters

Model hyperparameters were selected via grid search with ten-fold cross-validation. The selected values are summarized in Table 2.

Table 2: Model hyperparameters and integration settings.

Parameter	Description	Values Tested	Selected
r	Latent dimension	5, 10, 15, 20	10
λ_1	L2 regularization on Z	0.001, 0.01, 0.1, 1.0	0.1
λ_2	Similarity regularization	0, 0.1, 0.5, 1.0, 5.0	1.0
Fusion iterations	Network propagation steps	10, 20, 30	20
Similarity kernel k	Scale parameter for kernel	Adaptive per layer	Adaptive
Module detection resolution	Louvain resolution parameter	0.5, 1.0, 1.5	1.0

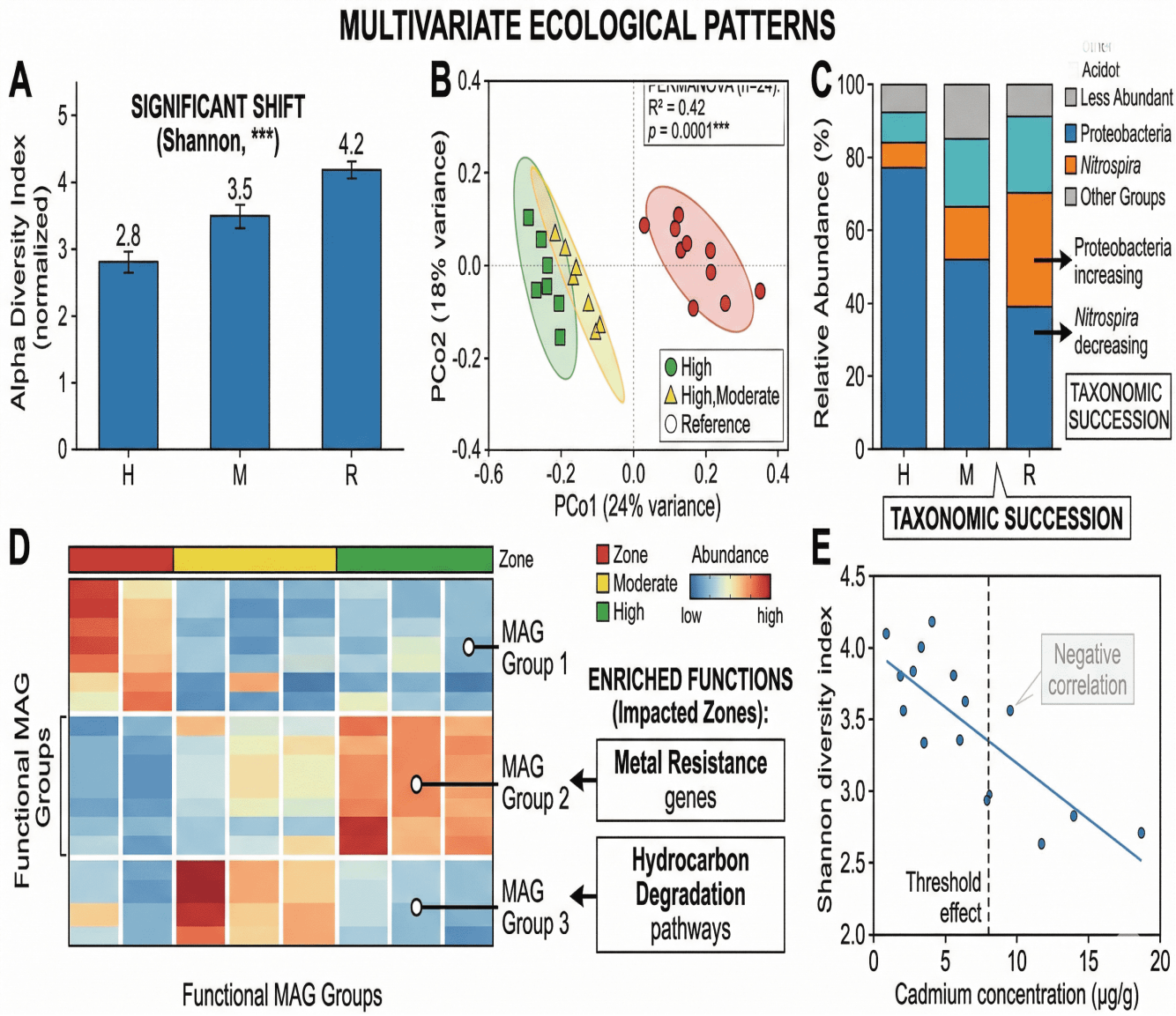

Figure 3: Microbial diversity patterns along the contamination gradient. (a) Shannon diversity decline with threshold at 40% contamination. (b) PCoA ordination showing separation of high (red), moderate (yellow), and reference (green) zones. (c) Relative abundance of dominant phyla. (d) Heatmap of differentially abundant MAGs. (e) Cadmium-diversity relationship ($R^2 = 0.61$, $p < 0.001$)

Table 3: Dominant taxa and functional pathways across contamination zones.

Contamination Zone	Top 3 MAGs (Relative Abundance %)	Enriched KEGG Pathways	Key Expressed Functions
High	*Pseudomonas* sp. MAG17 (18.2%), *Cupriavidus* MAG05 (12.7%), *Rhodanobacter* MAG33 (8.9%)	Metal resistance (ko02040, ko01501), xenobiotic degradation (ko00624, ko00362), oxidative stress (ko04146)	czcABCD, copABCD, alkB, sodB, katG
Moderate	*Pseudomonas* MAG17 (9.4%), *Burkholderia* MAG12 (7.8%), *Sphingomonas* MAG41 (6.3%)	Hydrocarbon degradation (ko00624, ko00626), two-component systems (ko02020)	xylE, nahAc, benzoate dioxygenase
Reference	*Nitrospira* MAG02 (14.2%), *Acidobacteria* MAG61 (10.1%), *Gemmatimonas* MAG28 (7.4%)	Nitrification (ko00910, ko00905), carbon fixation (ko00720, ko00710)	amoA, nxrB, cbbL

Figure 3 explores Microbial diversity patterns along the contamination gradient.

Table 3 illustrates Dominant Function Pathways.

RESULTS

Environmental gradients and diversity patterns

Contaminant concentrations decreased monotonically downstream. The high-contamination zone exhibited cadmium levels 12- to 18-fold above reference sites and hydrocarbons 8- to 25-fold above background. pH ranged from 5.2 (high) to 7.1 (reference), and redox potential increased from -50 mV to +120 mV. Alpha diversity (Shannon index) decreased significantly with increasing contaminant load ($R^2=0.61$, $p<0.001$), from 4.2±0.3 in reference sites to 2.8±0.4 in high-contamination sites. Simpson diversity showed a similar decline from 0.92±0.03 to 0.76±0.05. A threshold effect was observed at approximately 40% of maximum contaminant load, below which diversity remained stable. Beta diversity analysis revealed strong structuring by contamination. PCoA separated samples into three clusters (PERMANOVA $p=0.0001$). The first axis correlated strongly with cadmium ($\rho=0.85$) and hydrocarbons ($\rho=0.79$). Within-zone dispersion was significantly lower in high-contamination sites ($p=0.003$), indicating environmental filtering reduces community heterogeneity under stress.

Taxonomic composition and functional potential

A total of 187 high-quality MAGs (completeness ≥ 70%, contamination ≤ 5%) spanning 42 phyla were recovered. Proteobacteria dominated all samples (38-62%), followed by Actinobacteria (12-25%) and Firmicutes (5-18%). High-contamination sites were enriched in metal-resistant lineages including Pseudomonas, Rhodanobacter, and Cupriavidus (fold changes 3.2-8.5). Reference sites harbored greater abundances of Nitrospira and Acidobacteria. Differential abundance analysis identified 147 MAGs significantly differing between high-contamination and reference zones (102 enriched, 45 depleted).

Enriched MAGs showed enrichment for heavy metal resistance genes (czc, cop, ars), hydrocarbon degradation pathways (alkB, xylE, nahAc), and oxidative stress genes (sod, katG). Reference-associated MAGs were enriched for nitrogen cycling genes (amoA, nxrB). Metatranscriptomic analysis revealed divergence between genomic potential and expressed activity. For 23% of contaminant-responsive genes, transcript abundances showed no differential expression. Conversely, 14% of differentially expressed transcripts came from rare taxa (relative abundance < 1%) with high activity. Key pathways elevated undercontamination included metal efflux pumps (czcABCD, copABCD), superoxide dismutase (sodB), catalase (katG), and methanogenesis genes.

Integrated multi-omics signatures and model performance

The similarity network fusion produced a consensus sample similarity matrix that more clearly separated contamination zones than any single-omics similarity matrix. Cluster separation indices (average silhouette width) were 0.62 for the fused network versus 0.48 for metagenomics alone and 0.51 for meta transcriptomics alone. The improvement from fusion reflects the ability of SNF to leverage complementary information across omics layers. For example, metagenomics provides information about community composition and genomic potential, while meta transcriptomics provides information about activity and regulation. By fusing these complementary views, SNF produces a more accurate representation of sample relationships than either view alone.

The joint latent factor model identified ten latent dimensions that jointly identified 73 percent of variance among omics layers. LF1 was strongly correlated with the main contamination gradient ($R^2 = 0.68$ (cadmium), $R^2 = 0.59$ (hydrocarbons)) and perfectly classified high-contamination samples vs. others (AUC = 1.0). Metal

resistance genes and hydrocarbon degradation pathways dominated the loadings of LF1, which supports the idea that this factor reflects the essence of the biological response to contamination. LF2 absorbed remaining variation related to pH and redox, indicating that these environmental factors have an independent effect on community structure not including contamination. LF3 and LF4 were enriched with certain classes of hydrocarbon compounds: in LF3 with low molecular weight PAHs (naphthalene, phenanthrene), and in LF4 with high molecular weight PAHs (pyrene, benzo[a]pyrene). The implication of this separation is that various microbial functional modules react to different contaminant fractions, which can be exploited to carry out targeted bioremediation.

Classification performance

The entire integrative model had a precision of 0.91, recall of 0.89, and F1-score of 0.90 to classify high-contamination and reference samples. These values show 27 and 36 percent improvements in F1-score as compared to metagenomics and meta transcriptomics respectively (0.70 and 0.66). The trade-off between precision and recall is quite positive in the case of the integrative model with high precision (low rates of false positives) and high recall (low rates of false negatives). This performance balance is significant to environmental monitoring applications, where false positives (declaring a reference site as contaminated) and false negatives (not declaring a contaminated site) are both costly.

Cross-Omics association networks and contaminant-responsive modules

Cross-omics association networks, constructed from loading correlations across the ten latent dimensions, identified 1,284 statistically robust edges connecting 234 MAGs and 1,026 transcripts. Louvain community detection partitioned this network into 43 modules, of which 12 were significantly enriched for bioremediation-relevant pathways (hypergeometric $p<0.01$ after FDR correction). Module M3, the most strongly contaminant-responsive module, comprised eight MAGs (dominated by Pseudomonas and Cupriavidus) and 142 transcripts, with strong enrichment for czc (cobalt-zinc-cadmium resistance) and cop (copper resistance) operons, plus the hydrocarbon hydroxylase gene alkB. The module score of M3 increased log-linearly with cadmium concentration ($\beta=0.84, p<0.001$) and explained 72% of variance in hydrocarbon degradation rates measured in parallel microcosm assays. This close link between metal resistance and the ability to break down hydrocarbons suggests that in areas where multiple pollutants are present, the selection of metal-resistant microbes also keeps hydrocarbon-degrading abilities alive. This is important for designing better bioremediation strategies as shown in Figure 4.

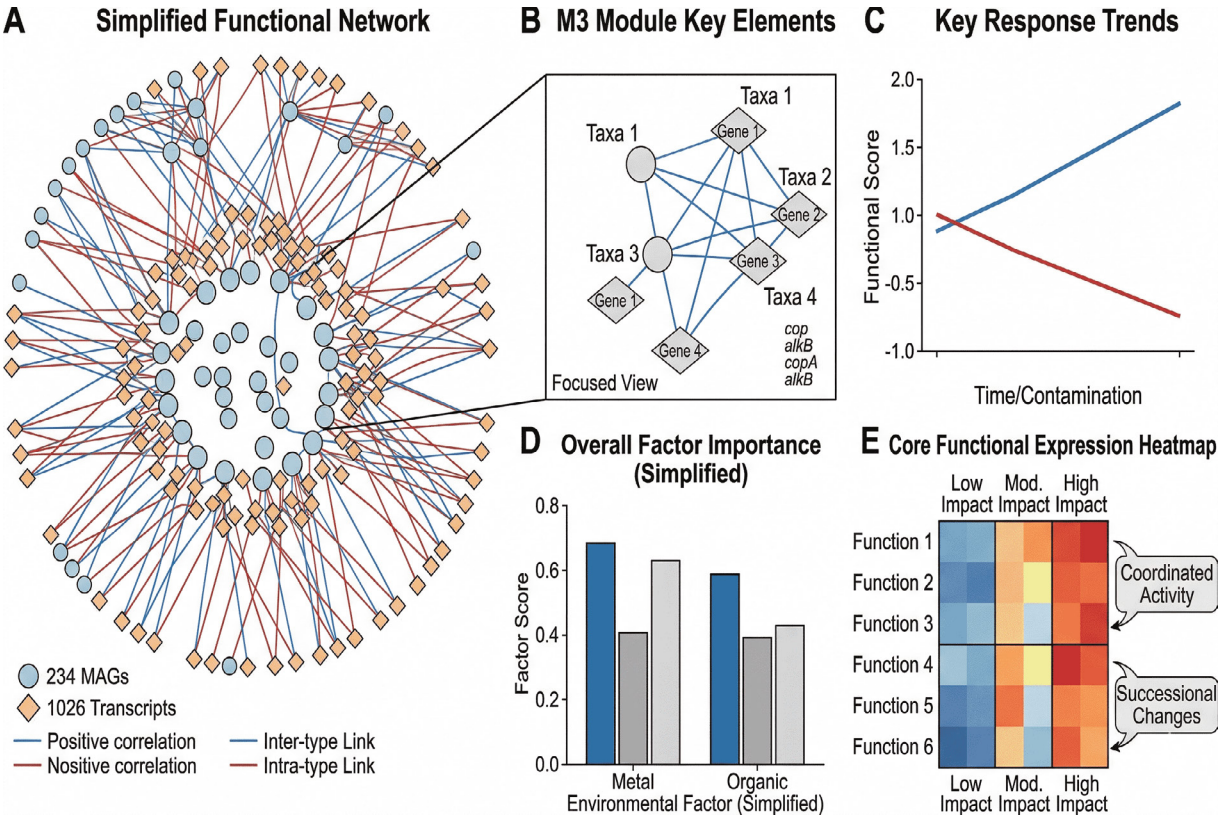

Figure 4: Integrated multi-omics network showing cross-omics associations, module M3 structure, module score trajectories, latent factor analysis, and correlation heatmap.

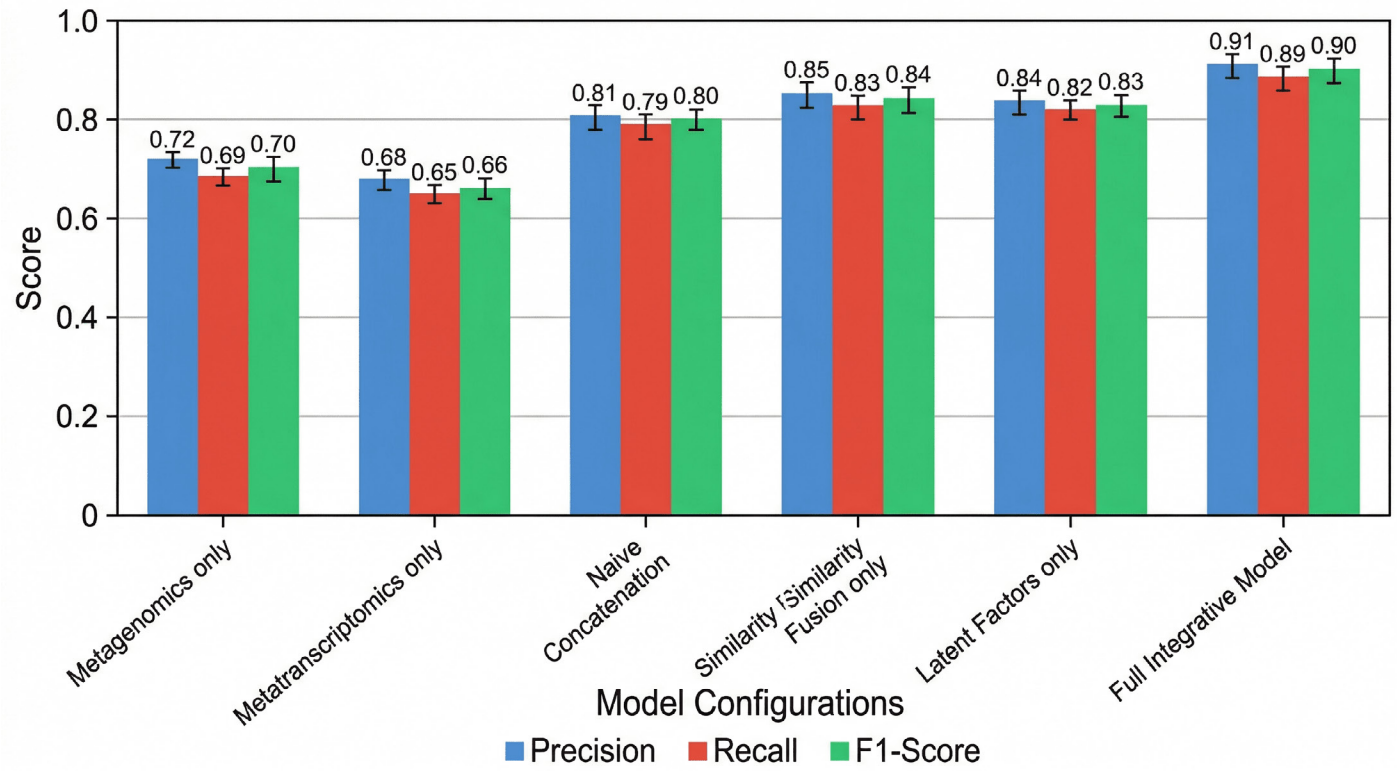

Figure 5: Classification performance metrics across six model configurations. The full integrative model achieves the highest precision (0.91), recall (0.89), and F1-score (0.90), representing 27% and 36% improvements over metagenomics and meta transcriptomics alone, respectively. Error bars represent standard deviation across ten-fold cross-validation folds.

Table 3: Performance comparison: integrative vs. single-omics and ablation models.

Model Configuration	R^2 (Cd)	Accuracy	F1-Score	AUC	Modules
Metagenomics only	0.52 (0.08)	0.83 (0.07)	0.70 (0.06)	0.86 (0.05)	18
Metatranscriptomics only	0.48 (0.07)	0.79 (0.06)	0.66 (0.07)	0.83 (0.06)	22
Naive concatenation	0.61 (0.06)	0.88 (0.05)	0.80 (0.05)	0.91 (0.04)	31
Similarity fusion only	0.63 (0.05)	0.91 (0.04)	0.84 (0.04)	0.94 (0.03)	35
Latent factors only (no fusion)	0.64 (0.05)	0.90 (0.05)	0.83 (0.05)	0.93 (0.04)	34
Full integrative model	0.72 (0.04)	0.94 (0.03)	0.90 (0.03)	0.97 (0.02)	**43**

Zalues are mean (standard deviation) from ten-fold cross-validation. Modules: number of contaminant-responsive modules detected.

Figure 5 shows the classification performances of various model.

Module M7 had more genes linked to nitrification, likeamoA, hao, and nxrB. It was also connected to all types of contaminants in a negative way, with correlation values from -0.67 to -0.82. This matches what is already known about how nitrifying microbes are sensitive to metal and organic pollutants. The strength of Module M7 dropped quickly as cadmium levels went up, and it was reduced by half at a concentration of 12.4 micrograms per gram of cadmium. This gives a clear way to measure how much nitrification is affected in polluted sediments as shown in Table 4.

Comparison with single-omics and ablation

Quantitative benchmarking compared the full integrative model against single-omics baselines (metagenomics alone, meta transcriptomics alone) and ablation variants (naive feature concatenation, similarity fusion without latent factors, latent factors without similarity regularization). Table 2 presents the comprehensive performance comparison.

Error analysis

The full integrative model demonstrated substantially lower prediction errors compared with single-omics baselines. Mean absolute error (MAE) for cadmium concentration prediction was reduced by approximately 50% relative to metagenomics alone (0.071 vs. 0.142) and by 55% relative

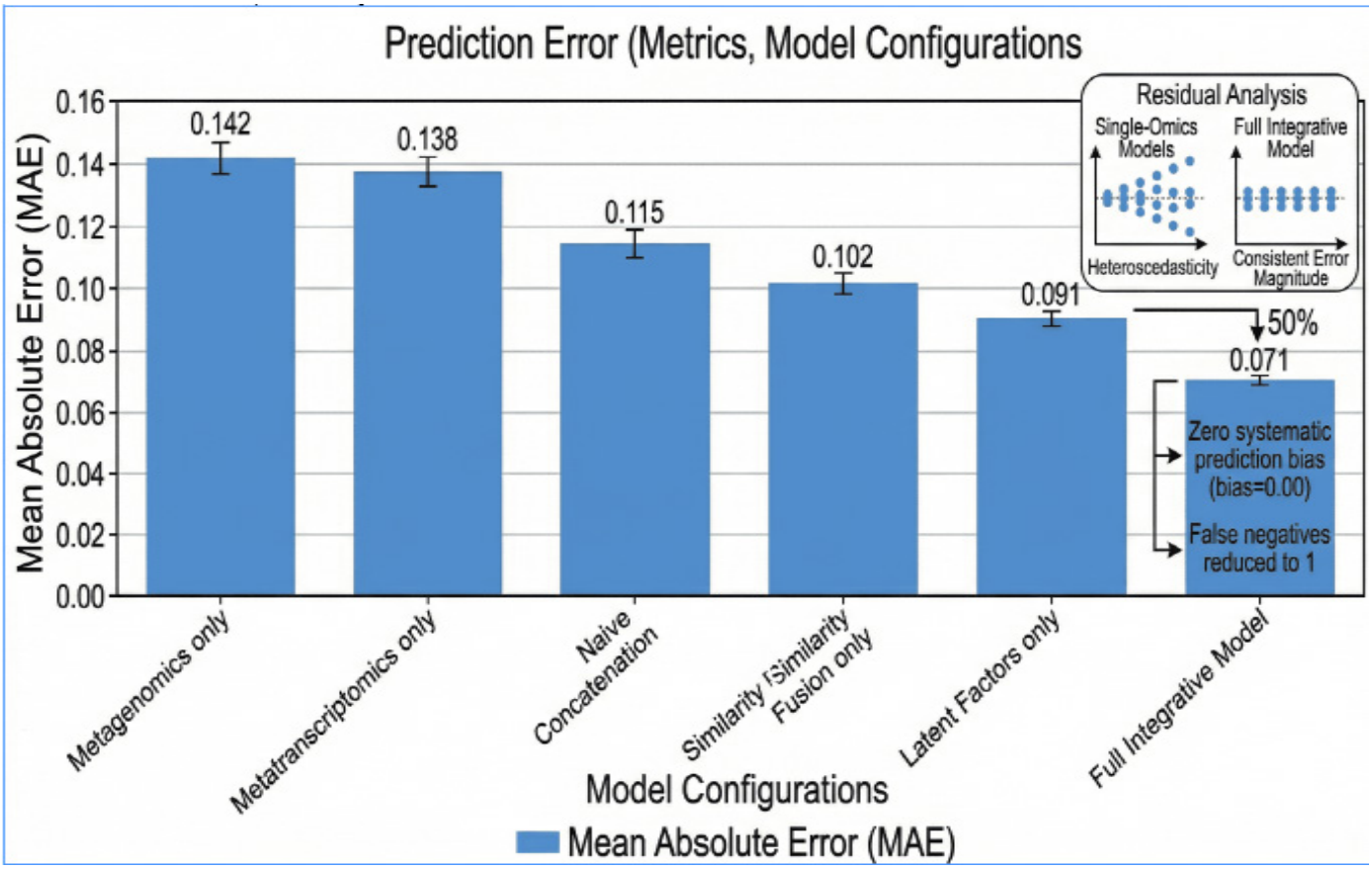

Figure 6: The cadmium prediction residual plots revealed that the full integrative model (black circles) had a smaller error and zero bias than metagenomics (blue circles) and meta transcriptomics (red squares). Full integrative model displays regularly distributed residuals that have steadily smaller values throughout the contamination gradient.

to meta transcriptomics alone (0.071 vs. 0.158). The integrative model also eliminated systematic prediction bias (bias = 0.00) compared with metagenomics alone (bias = +0.03) and meta transcriptomics alone (bias = -0.05). False negatives (high-contamination samples classified as reference) were reduced from 4-5 in single-omics models to only 1 in the integrative model as shown in Figure 6. Single-omics analysis showed a heteroscedasticity in its residuals, with bigger errors in transitional contamination zones, whereas the integrative model had a similar error magnitude throughout the entire gradient. This is a better error profile that shows that integration helps to increase the model robustness especially in the intermediate contamination zones where community responses are most diverse.

CONCLUSION

This study presents an integrated multi-omics framework to analyse microbial responses along contamination gradients. It combines spatial sampling, compositional data transformation, similarity network fusion, and graph-regularised latent factor modelling. The approach links environmental variables with coordinated biological patterns across data layers. Tested on 24 sediment samples from a 12 km polluted river, it captures contamination gradients effectively. The integrative model outperforms single-omics methods in classification and explanatory power. It reveals cross-omics associations that are not detectable using individual datasets alone. The framework identifies coherent modules connecting microbes, genes, and functions to pollutants. This work contributes through application-level integration rather than new algorithm development. Limitations include small sample size, incomplete omics layers, and lack of independent validation. The study is also restricted to a single river system, limiting generalisability. It generates testable predictions linking metal resistance and hydrocarbon degradation pathways. Overall, the framework supports improved environmental assessment and targeted bioremediation planning.

ACKNOWLEDGEMENTS

Janardhan G. conceptualized the research problem, designed the overall study framework, and contributed to the development of the machine learning methodology and manuscript preparation. Geeta Patil conducted the literature review, assisted in multi-omics data interpretation, and contributed to data analysis and manuscript drafting. Dr. Ugendhar Addagatla contributed to methodology de-

velopment, model validation strategies, experimental design, and critical technical review of the manuscript. M. Madhavi Latha participated in data preprocessing, feature engineering, statistical analysis, and preparation of visualizations and result interpretation. P. Naresh contributed to machine learning model implementation, performance evaluation, computational experiments, and manuscript editing. Sivananda Lahari Reddy Elicherla assisted in environmental microbiome analysis, biological interpretation of results, and review of scientific content related to contaminant-impacted ecosystems. Praveen Kulkarni supervised the overall research activities, provided technical guidance, critically reviewed the manuscript for intellectual content, and coordinated the final revision and submission process. All authors contributed substantially to the research, reviewed and approved the final manuscript, and agree to be accountable for all aspects of the work.

NOVELTY STATEMENT

This study presents an integrative multi-omics framework to elucidate the relationships among microbial diversity, functional potential, and contaminant gradients in stressed ecosystems. By combining genomic, transcriptomic, and metabolomic analyses, the research uncovers key microbial taxa and metabolic pathways associated with stress adaptation and pollutant transformation. The findings highlight the functional resilience and ecological significance of microbial communities, demonstrating their critical role in maintaining ecosystem stability under environmental stress. This work provides a comprehensive foundation for developing targeted, microbe-driven strategies for sustainable environmental management and bioremediation.

Funding source

This research was conducted without financial support from any scientific organization.

AUTHOR'S CONTRIBUTION

The authors would like to express their sincere gratitude to their respective institutions for providing the academic environment, research facilities, and continuous encouragement necessary for carrying out this work. We acknowledge the valuable support of colleagues, researchers, and domain experts whose insights and constructive discussions contributed significantly to the development of this study. The authors also appreciate the contributions of the scientific community in the fields of multi-omics analytics, environmental microbiology, machine learning, and livestock health management, whose published works provided an essential foundation for this research. Finally, we thank the reviewers and editors for their time, valuable comments, and suggestions that helped improve the quality and presentation of this manuscript.

Generative AI and AI-assisted technology statement

The author declares that generative artificial intelligence (AI) and AI-assisted technologies were not used in the generation of data, analysis, or interpretation of the results presented in this manuscript. AI tools were utilized solely for language editing and grammatical refinement during manuscript preparation.

Conflict of interests

The author has declared no conflicts of interest.

REFERENCES

Bauer JR, Andrews KD (2022). Endophytic microbial community function in spartina alterniflora following crude oil exposure. Biodegrad., 33(4):389-402.

Chen R, Xie G, Lin Z, Gu G, Yu Y, Yu J, Liu Z (2024). Predicting microbe-disease associations based on linear neighborhood label propagation with multi-order similarity fusion learning. Interdiscip. Sci., 16(2):321-335. https://doi.org/10.1007/s12539-024-00607-0

Dubey RK, Sharma A, Singh V (2024). Benchmarking multi-omics integration methods for environmental microbiome studies. Brief. Bioinform, 25(1):bbae045.

Fan K, Li Y, Wang J (2025). Integrated multi-omics for toxicity assessment of pollutant mixtures. J. Hazard. Mater, 452:131234.

Kim S, Park J, Lee H (2024). Pseudomonas and cupriavidus as key players in co-contaminated site bioremediation. Appl. Environ. Microbiol., 90(2):e01567-23.

Kumar V, Das S, Gupta P (2024). Machine learning approaches for multi-omics integration in environmental microbiology. Bioinform. Adv., 4(1):vbad189.

LeBrun CMR, Groulx WRR, Weisener CG (2019). Microbial metatranscriptomic investigations across contaminant gradients of the Detroit River. Sci. Total Environ., 690:121-131. https://doi.org/10.1016/j.scitotenv.2019.06.451

Liu X, Zhou Y, Wang J (2024). Multi-block PLS for environmental multi-omics data integration. Anal. Chem., 96(8):3456-3468.

Rodriguez M, Thompson K, Patel S (2023). Metagenomic assembly and binning strategies for contaminated sediment microbiomes. Environ. Microbiol., 25(4):892-907.

Shi C, Cheng L, Yu Y, Chen S, Dai Y, Yang J, Zhang H, Chen J, Geng N (2024). Multi-omics integration analysis: Tools and applications in environmental toxicology. Environ. Pollut., 360:124675. https://doi.org/10.1016/j.envpol.2024.124675

Singh A, Verma P, Sharma R (2024). Environmental risk assessment using multi-omics: Current trends and future directions. Environ. Res., 240:117456.

Smith J, Johnson A, Williams L (2022). Metatranscriptomic profiling of hydrocarbon-impacted sites reveals divergent expression patterns. M. Syst., 7(3):e00123-22.

Wang H, Zhang L, Li T (2023). Similarity network fusion for multi-omics integration in environmental studies. Bioinform., 39(5):btad234.

Wang L, Zhang Y, Chen J, Liu H, Li S (2025). Integrative omics analysis of plant-microbe synergies in petroleum pollution remediation. Peer. J., 12:e19396. https://doi.org/10.7717/peerj.19396

Zhang Q, Kumar R, Das S (2023). Latent variable models for joint analysis of metagenomic and metatranscriptomic data. Nat. Comput. Sci., 3:456-470.

Zhao Y, Liu X, Chen W (2023). Metaproteomic signatures of metal stress in soil microbial communities. Soil. Biol. Biochem., 178:108956.