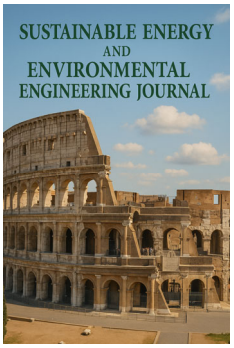
UNIVERSITY OF WEST ATTICA

eISSN:3136-8174

Sustainable Energy and Environmental Engineering Journal

Vol. 4, No. 1, January, 2026



Topology Preserving Hyena Operator Prediction Model for Functional Annotation of Microbial Communities

Nika Potočnik[1,*], Tim Korošec[1], Matic Kavčič[1], Ema Hribar[1]

[1] *Department of Environmental Engineering, Faculty of Natural Sciences and Engineering, University of Ljubljana, Ljubljana, 1000, Slovenia*

Received 24 October 2025 Revised 30 January 2026; Accepted 03 February 2026; Published 09 February 2026

Abstract

Functional annotation of the microbial community is required to interpret metagenomic data, identify changed metabolic pathways, and provide support for ecological or clinical hypotheses. Prediction is still difficult because taxonomic abundance, gene-family evidence, pathway topology, community interactions and sample environment may all affect the label. This paper proposes a topology-preserving Hyena Operator model for microbial community functional annotation. Taxon abundance, marker genes, predicted enzyme families, pathway modules, co-occurrence edges, habitat metadata, sequencing-depth indicators and uncertainty tags are aligned into abundance-function windows. The topology-preserving graph layer keeps the neighbourhood relationships and pathway dependencies of microbes, thus distinguishing between isolated gene evidence and community signals supported by pathways. The Hyena Operator then encodes long-range abundance and functional relationships with a lower computational cost than full attention, which is suitable for high-dimensional metagenomic feature spaces. Experiments are carried out on a simulated microbial community dataset from soil, gut, marine environments, fermentation, engineered sources, etc. The dataset contains various taxonomic profiles, sparse gene-family evidence, pathway-module labels, co-occurrence topology, habitat metadata, and incomplete annotation masks. Source-aware testing and unseen-community evaluation are used to determine if the model can generalise beyond random sample splits. The multi-task head predicts pathway class, enzyme profile, missing-function risk, confidence, dominant evidence and review priority. The framework proposed here aims to support interpretable microbial function screening by indicating whether a predicted function is supported by abundance, gene evidence, topology, source context or weak uncertain signals. This Design helps the researchers prioritise pathway annotations that are reliable, provisional or need further database review.

Keywords: Microbial Functional Annotation; Hyena Operator; Topology Preserving Learning; Metagenomics; Community Network

1. Introduction

Recent microbial community function prediction research shows that many gene products from metagenomes remain difficult to annotate with conventional similarity search [1]. Deep-learning based metagenomic tools show that functional interpretation increasingly depends on representations that can handle fragmented contigs, sparse labels and heterogeneous samples [2]. Shotgun metagenomic pipelines [3] demonstrate that reproducible annotation requires coordinated taxonomic, gene-family and pathway-level processing. Hyena-based biological sequence modelling [4] also indicates that long sequence operators can reduce computational burden while retaining long-range dependency information. These studies suggest that microbial functional annotation should move beyond isolated gene matching toward a topology-aware and long-context prediction framework. This shift is important because microbial

Corresponding author: nika.poto@ntf.uni-lj.si

https://doi.org/10.68406/seeej.2026.vol4iss1nz3:28-40

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

communities rarely express functions through one clean observation. A function may be supported by several low-abundance organisms, by partially assembled gene fragments or by a pathway module whose members are distributed across taxa. A model that treats these signals as unrelated features can assign plausible but unstable labels, especially when the functional database contains uneven coverage across habitats.

The task in this paper is not a general taxonomic profile. Taxonomic profiling estimates which organisms may be present, while functional annotation estimates which community-level activities can be supported by genes, enzymes and pathway modules. Deep learning methods in metagenomics can support functional interpretation beyond exact alignment [5], and metagenomic resistance prediction often needs to identify weak or novel signals rather than only recover known labels [6]. The revised text therefore clarifies that abundance, gene evidence, pathway context and community topology must be interpreted together, with uncertainty shown when evidence is incomplete.

Another problem is that the same function may be distributed among several taxa, while the same taxon may perform different roles in different environments. A comprehensive gut microbiome catalogue expands genome-resolved functional interpretation [7], rumen metagenomic analysis shows environment-related functional-gene adjustment [8], and fermentation metagenomics demonstrates simultaneous microbial succession and functional-potential change [9]. The revised manuscript therefore argues that samples should not be treated as independent feature counts. Community structure must remain visible so distributed functions and context-dependent taxa can be reviewed.

This paper proposes a topology-preserving Hyena Operator prediction model for microbial community functional annotation. Chapter 2 reviews microbial functional annotation, metagenomic binning, co-occurrence networks, long-sequence modelling and graph topology learning. Chapter 3 introduces abundance-function window construction, topology-mask design, Hyena long-range encoding and multi-task functional scoring. Chapter 4 presents the simulated dataset, source and pathway statistics, baseline comparison, ablation analysis, cross-community transfer and error interpretation. Chapter 5 concludes the paper and discusses application scenarios in metagenomic annotation triage.

2. Related Work

2.1. Functional Annotation of Microbial Communities

Functional annotation of microbes has moved from single-gene lookups to community-level interpretation. Co-occurrence network analysis across a tomato domestication gradient shows that microbial relationships vary with host and environment [10]. Contrastive multi-view metagenomic binning improves genome-resolved recovery in complex samples [11], and benchmarking studies show that upstream binning quality strongly affects downstream annotation [12]. The revised related-work section therefore separates taxonomic recovery, functional-gene evidence and topology-aware interpretation, while motivating a multi-source window that stores these signals together.

Function prediction also needs to handle missing and ambiguous labels. Machine-learning models for microbial genome annotation can identify virulence and functional patterns, but results remain sensitive to feature selection and label quality [13]. Bayesian community detection supports latent community structure in microbiome data [14], and soil microbial network studies show that associations can mediate ecosystem-level functions [15]. The revised text therefore presents functional annotation as a multi-task problem that predicts pathway category, enzyme profile, missing-function risk and confidence at the same time.

At the same time, there is a lack of microbial functional evidence. Some genes are observed directly, some are inferred from related taxa, and some are missing because of sequencing depth or assembly uncertainty. Keystone microbe detection from co-occurrence networks [16] indicates that a small number of taxa can regulate the network's behaviour. Vegetation disturbance studies [17] have also found that the strength of network interaction increases under different nutrient conditions. Based on the above observations, topology-preserving evidence can be used instead of independent feature vectors. In actual annotation, this means that a rare taxon should not be ignored automatically if it is connected to a functionally important pathway neighbourhood. It also means that a very abundant organism should not be predicted if its pathway evidence is inconsistent with the surrounding network. The two situations can both be displayed in the model by the above representation.

Thus, the relevant studies suggest three modelling requirements. The input should retain taxonomic abundance and functional gene evidence, the model should represent long-range relationships among taxa and pathways, and the output should report confidence when evidence is missing. Drought and restoration studies show that environment can shift microbial co-occurrence networks [18], and long-read metagenomic binning can strengthen genome recovery [19]. The revised text connects these observations with source-aware windows and topology-preserving evidence rather than relying on one fixed database match.

2.2. Long Sequence Models and Graph Topology Learning

Long-sequence modelling is suitable for high-dimensional taxon, contig, gene-family and pathway features in metagenomic samples. Deep learning and alignment comparisons for environmental resistomes show that learned annotations can supplement traditional methods when resistance genes are diverse or poorly characterized [20]. The revised discussion explains that graph topology and long-range functional sequence evidence should be combined, because topology preserves biological relationships while Hyena-style encoding efficiently mixes distant feature blocks.

Topology learning is also required because microbial functions rarely occur in isolation. A sample may contain taxa that do not have a complete pathway individually but jointly support a community-level function. A topology-preserving model retains co-occurrence neighbourhoods and pathway adjacency, while the Hyena Operator encodes long-range ordered features. The revised text clarifies that the topology mask prevents biological relationships among distant positions from being lost after long-range sequence mixing, and it helps link low-abundance taxa with pathway modules appearing elsewhere in the same sample.

The Hyena Operator is employed to model long feature windows without the full quadratic cost of attention. Microbial samples may contain thousands of taxa, contigs, gene-family indicators and pathway features. The topology-preserving component prevents these entries from being treated as interchangeable tokens, while Hyena encoding supplies efficient long-range mixing. The revised wording therefore presents the architecture as an efficiency-and-review design for large metagenomic windows rather than as a generic long-sequence encoder.

3. Methodology

3.1. Abundance Function Window Construction

Each microbial community sample is represented as an abundance-function window. The window contains taxon abundance, gene-family counts, enzyme commission indicators, pathway module coverage, sample metadata, sequencing-depth tags and missing-label masks. Features are organized by taxonomic group and pathway dependency so neighbouring positions have biological meaning. presents the topology-preserving Hyena encoding workflow, where abundance-function windows, pathway dependencies, co-occurrence edges and graph masks are aligned before long-sequence modelling.

$$x_i = [a_i, g_i, e_i, p_i, m_i] \tag{1}$$

The abundance term is normalized within source type because soil, gut, marine, fermentation and engineered communities have different sequencing depth and diversity ranges. This step reduces source-level scale differences while preserving relative community structure. The normalized abundance is used together with gene-family evidence rather than replacing it. The revised explanation clarifies that source normalization is intended to control measurement scale, not remove biological source context.

$$a_i^* = \frac{\log(1 + a_i)}{\sum_j \log(1 + a_j)} \tag{2}$$

Functional pathway coverage is estimated from observed enzyme indicators and module dependency. A pathway is treated as weakly supported when only isolated enzymes appear, and strongly supported when several required steps are present in a consistent community neighbourhood. The revised text emphasizes that pathway evidence is not a simple count; it depends on whether enzymes, taxa and module positions support one another within the topology-aware window.

$$p_k = \frac{\sum_r M_{kr} e_r}{\max(1, \sum_r M_{kr})} \tag{3}$$

The three contents of the resulting window are taxonomic abundance, functional gene evidence and pathway structure. Determine if the predicted function is supported by a single dominant taxon, a distributed community pattern, or a pathway module shared by multiple organisms in the organisation. The same window also stores a missing-evidence mask to inform the model that a zero is indeed missing and thus unobserved. Without this division, the sparse functional annotations may be considered missing, and thus rare pathways and newly discovered gene families in the model will be disadvantaged.

Quality supervision during construction precedes training. Samples with very low sequencing depth, inconsistent pathway identifiers or missing source labels are excluded from the simulated training set. Only extreme feature values are clipped so strong ecological or functional signals are not removed. Habitat type, sequencing mode and annotation

confidence are kept as explicit fields because they affect how much trust should be placed in a predicted function. The revised wording makes the input more reproducible for review-oriented annotation.

3.2. Topology Preserving Hyena Operator Model

The model first builds a topology mask from microbial co-occurrence and pathway adjacency. Edges connect taxa that repeatedly co-occur, gene families that participate in the same pathway and taxa linked to gene families through genome or contig evidence. Figure 1 presents the topology-preserving Hyena encoding flowchart, where abundance-function windows and graph masks are aligned before long-sequence modelling.

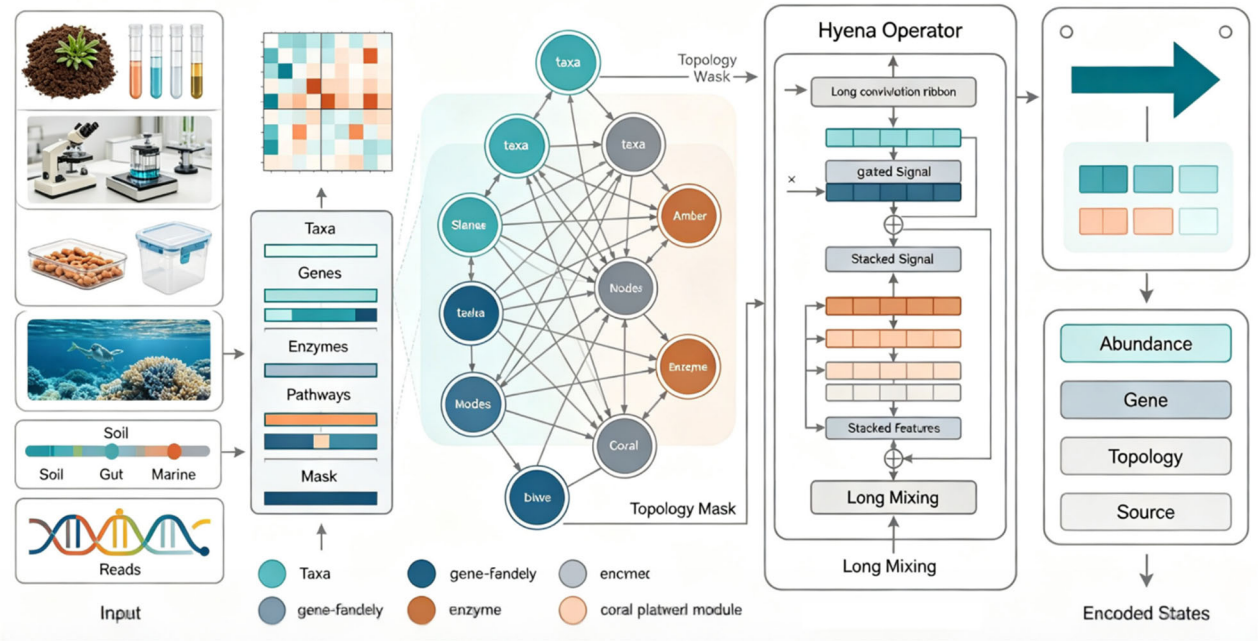


Figure 1. Topology preserving Hyena Operator encoding flowchart.

Topology preservation is introduced as a graph-weighted modulation term. The model does not force every connected node to share the same label; instead, it increases the influence of connected evidence when the connection is supported by abundance consistency and pathway support. The revised text presents the topology mask as soft evidence, which is important because microbial co-occurrence networks may contain indirect associations, shared habitat effects and sampling artifacts.

$$T_{ij} = \text{sigmoid}\left(w_c C_{ij} + w_p P_{ij} + w_g G_{ij}\right) \tag{4}$$

The Hyena Operator then encodes the ordered window with implicit long convolution and gating. This step captures long-range relations among rare taxa, distributed enzymes and pathway modules while avoiding the computational cost of full attention over the entire feature set. The revised description separates the two roles: topology preserves biological structure, and Hyena gating mixes distant feature blocks efficiently.

$$h_t = \text{Hyena}(x_t, T, \theta) \tag{5}$$

The encoded states are pooled by function group and source group. This makes the output less sensitive to feature order noise and allows the review card to state whether a prediction is driven by abundance, gene evidence, topology or source context. The architecture therefore keeps efficiency and interpretability in the same prediction path. During training, the topology mask is not treated as a fixed truth table. Edge weights are softened by abundance consistency and pathway support, so unreliable co-occurrence relations have limited influence. This is necessary because microbial networks inferred from abundance data can contain indirect associations, shared-environment effects and sampling artifacts.

The Hyena module uses gated long convolution to mix distant positions in the window. Compared with full attention, this design is more practical for large metagenomic feature spaces because it can process long ordered vectors without building all pairwise token relations. The topology mask supplies biological structure while the Hyena operator supplies efficient long-range mixing. The model can therefore learn that a weak enzyme signal near one part of the window becomes meaningful when a connected pathway module and a related taxon group appear elsewhere in the same sample.

3.3. Multi task Functional Scoring Output

The output layer predicts pathway class, enzyme profile, missing-function risk and confidence. Figure 2 shows the multi task scoring flowchart, where encoded states are separated into annotation probability, continuous enzyme score, missing-function risk and review action. The confidence output is penalized when the model predicts a function from weak topology alone or from sparse gene evidence without pathway support, so the review output separates label prediction from evidence reliability.

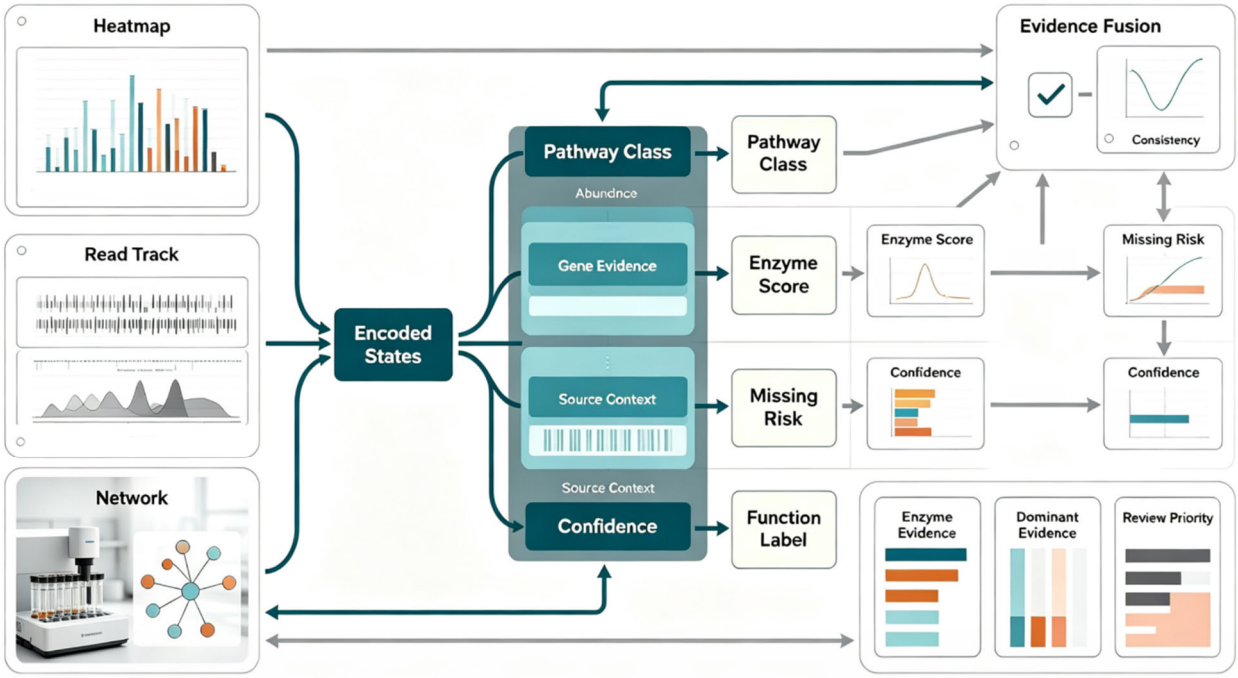


Figure 2. Multi task functional scoring output flowchart.

Training uses a joint objective that combines pathway classification, enzyme regression, missing-label reconstruction and calibration. The reconstruction term is important because many simulated records contain incomplete annotation labels, so the model must recover plausible functions without becoming overconfident. The revised text clarifies that the joint objective is needed because functional annotation requires both correct labels and reliable uncertainty for manual database review.

$$L = L_{\text{path}} + \alpha L_{\text{ec}} + \beta L_{\text{miss}} + \gamma L_{\text{cal}} \tag{6}$$

The final annotation card reports functional class, enzyme evidence, confidence, dominant evidence and review priority. A high-confidence pathway prediction requires consistent support from gene evidence and topology, while a low-confidence prediction is marked for database review or additional assembly verification. The card is designed for triage. It does not replace curated databases or expert interpretation, but it highlights where a functional label is supported and where it may be fragile.

4. Results and Discussion

4.1. Dataset Composition and Functional Class Statistics

The simulated dataset has 18,400 microbial community windows from soil, gut, marine, fermentation and engineered community sources. Each record contains taxon abundance, gene-family indicators, enzyme commission labels, pathway module coverage, co-occurrence edges, habitat tags and label-confidence masks. Metagenomic Hi-C protocols [21] support the joint application of taxonomic annotation and interaction network visualisation, and microbiological profiling of cultural materials [22] can be used to study source-specific microbial community structure. Therefore, the dataset employs source-aware splits and unseen-community testing instead of using only random records.

Figure 3 shows the simulated source and label composition. Figure 3(a) shows community-source distribution, where soil and gut samples are large groups while marine and fermentation records create cross-source variation. Figure 3(b) shows pathway-label distribution, where carbohydrate metabolism, nitrogen cycling, stress response, biosynthesis and antimicrobial resistance appear with different frequencies. The revised text explains that this imbalance is realistic for metagenomic repositories and makes rare functional categories difficult to annotate without calibrated topology-aware evidence.

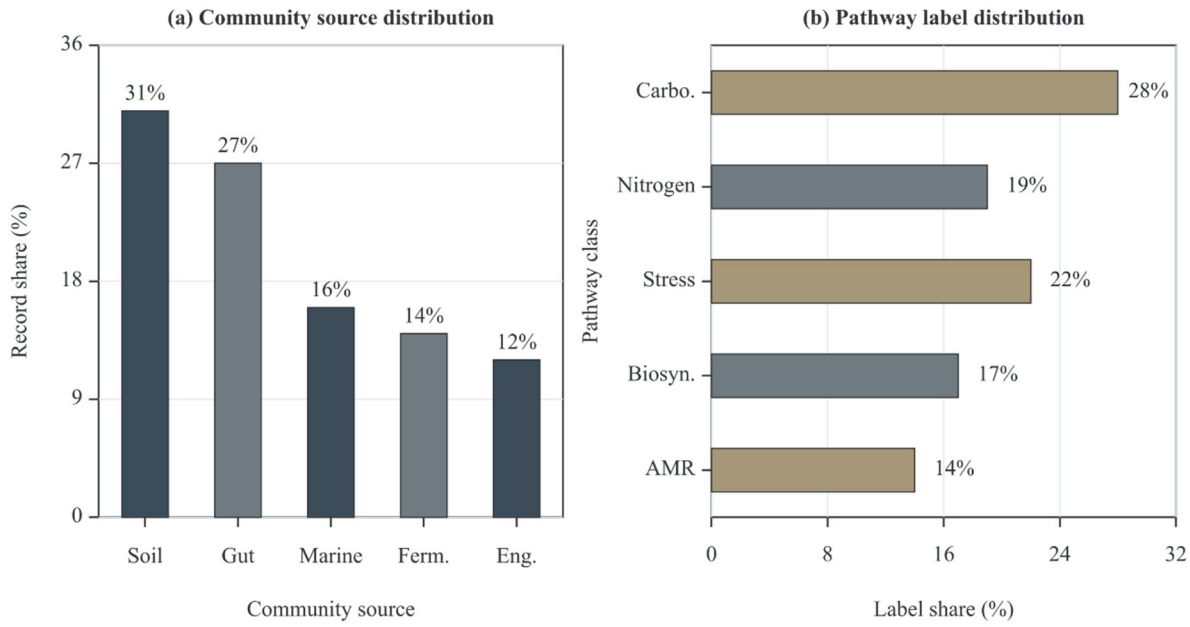


Figure 3. Dataset composition and functional class statistics. (a) community source distribution. (b) pathway label distribution.

The simulated labels also contain missing-function masks. Some records have direct enzyme evidence; others have incomplete gene evidence and are mainly supported by taxon-topology. Intercropping network analysis [23] shows that community relations change in an agricultural context, and thus the simulated windows have topology differences that cannot be explained by abundance alone. Therefore, the experiment will determine whether topology preservation is advantageous in the case of sparse annotations. The model does not reward all paths; therefore, it must be sensitive enough and well-calibrated to flag uncertain predictions for manual examination. The proportions of direct and inferred labels in the source groups vary. Soil samples have more network-supported functions, gut samples have more well-annotated metabolic labels, and fermentation samples show stronger temporal succession signals. Therefore, the simulated data are more representative of the mixed conditions that occur in metagenomic analysis.

Class statistics can be used to find class imbalance. Carbohydrate metabolism and stress response are common pathways that appear frequently; specialised biosynthesis and resistance-related labels are less frequent. If the model only optimises the general accuracy of classification, it may disregard less frequent but biologically significant labels. Therefore, the evaluation reports pathway F1, enzyme-profile error and calibration instead of using a single aggregate accuracy number. Therefore, the dataset is suitable for comparing the prediction results and review accuracy. Imbalance Setting can also be used to test the recovery of missing functions, as rare labels are more likely to have incomplete supporting evidence. In light of the above, a model needs to decide whether to keep the weak evidence as provisional or to cautiously annotate based on topology and neighbouring enzyme signals.

4.2. Multi task Prediction and Ablation Performance

Table 1 shows the multi-task prediction and ablation results. The table has been reduced to five model rows and four columns for more convenient comparison; Pathway F1 shows class prediction, EC MAE shows enzyme-profile error, and calibration measures confidence and reliability. Daqu storage community assembly analysis [24] and oral microbiome driver-taxon analysis [25] both show that microbial network complexity changes in response to the environment or disease, so a good model should be able to handle topology variation rather than fitting a single source distribution. Therefore, the table presents a concise comparison rather than a large-scale metric grid. The goal is to determine if the new Design can improve the three required outputs of the review: categorical pathway assignment, continuous enzyme-profile estimation, and the reliability of confidence scores.

Table 1. Multi task prediction and ablation performance on the simulated microbial dataset.

Model	Pathway F1	EC MAE	Calibration
Taxon baseline	0.712	0.184	0.091
Graph CNN	0.756	0.161	0.074
Hyena only	0.781	0.149	0.063
Topo-Hyena	0.824	0.126	0.047

Ablated topology	0.793	0.143	0.066

Figure 4 summarizes the main multi task prediction results. Figure 4(a) compares pathway classification performance, Figure 4(b) shows enzyme-profile regression, Figure 4(c) reports missing-function recovery, and Figure 4(d) shows confidence calibration. The topology-preserving Hyena model improves most in sparse functional categories because long-range evidence can connect low-frequency taxa with pathway modules. The revised interpretation treats these panels as a combined assessment of annotation quality rather than a single performance score.

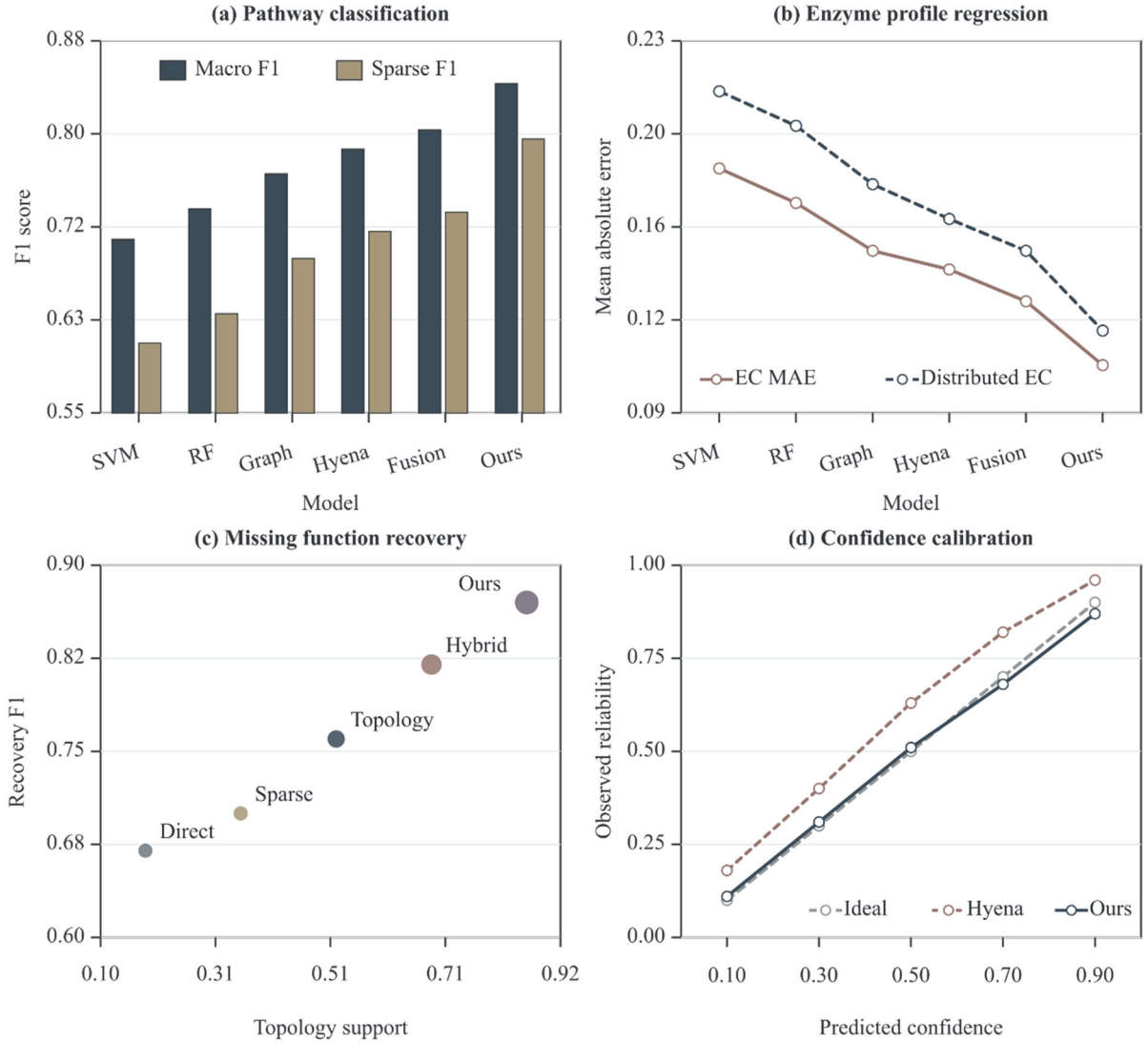


Figure 4. Multi task functional annotation performance. (a) pathway classification. (b) enzyme profile regression. (c) missing function recovery. (d) confidence calibration.

Ablation analysis removes topology masking, Hyena gating, source normalisation and calibration loss in turn. Extreme drought microbiome research [26] shows that community disruption alters both abundance and interaction patterns; thus, topology-related ablation is required. Biliary microbiota profiling [27] has also shown that functional differences are source-specific, so source normalisation should be tested rather than assumed. Removing the topology mask results in a larger drop in missing-function recovery, and removing Hyena gating mainly affects long-range pathway prediction. Without calibration, the mean pathway F1 is not significantly different, but high-confidence false annotations increase. The two are not the same; therefore, the annotation model may seem accurate but is unreliable in practice. A wrong high-confidence pathway assignment may lead to an incorrect biological interpretation, and a low-confidence assignment can be flagged for further assembly, database search or expert review.

Figure 5 shows the ablation and sensitivity results. Figure 5(a) compares module removal, and the full model retains the best balance of pathway prediction, enzyme regression and calibration. Figure 5(b) is topology-weight sensitivity; that is, adding reliable co-occurrence and pathway edges improves performance, but too dense a mask will reduce it. Figure 5(c) is the Hyena length sensitivity, and longer windows do not add more information after redundant low-abundance features. Figure 5(d) shows the effect of source normalisation; that is, normalised windows improve transfer across soil, gut and fermentation sources without losing source-specific biological signals. Ablation experiments show that topology and long-range encoding solve different problems. Topology enhances biological

plausibility and addresses missing functions; the Hyena operator improves the connection capability of distant feature blocks. Therefore, the two are more effective together than separately.

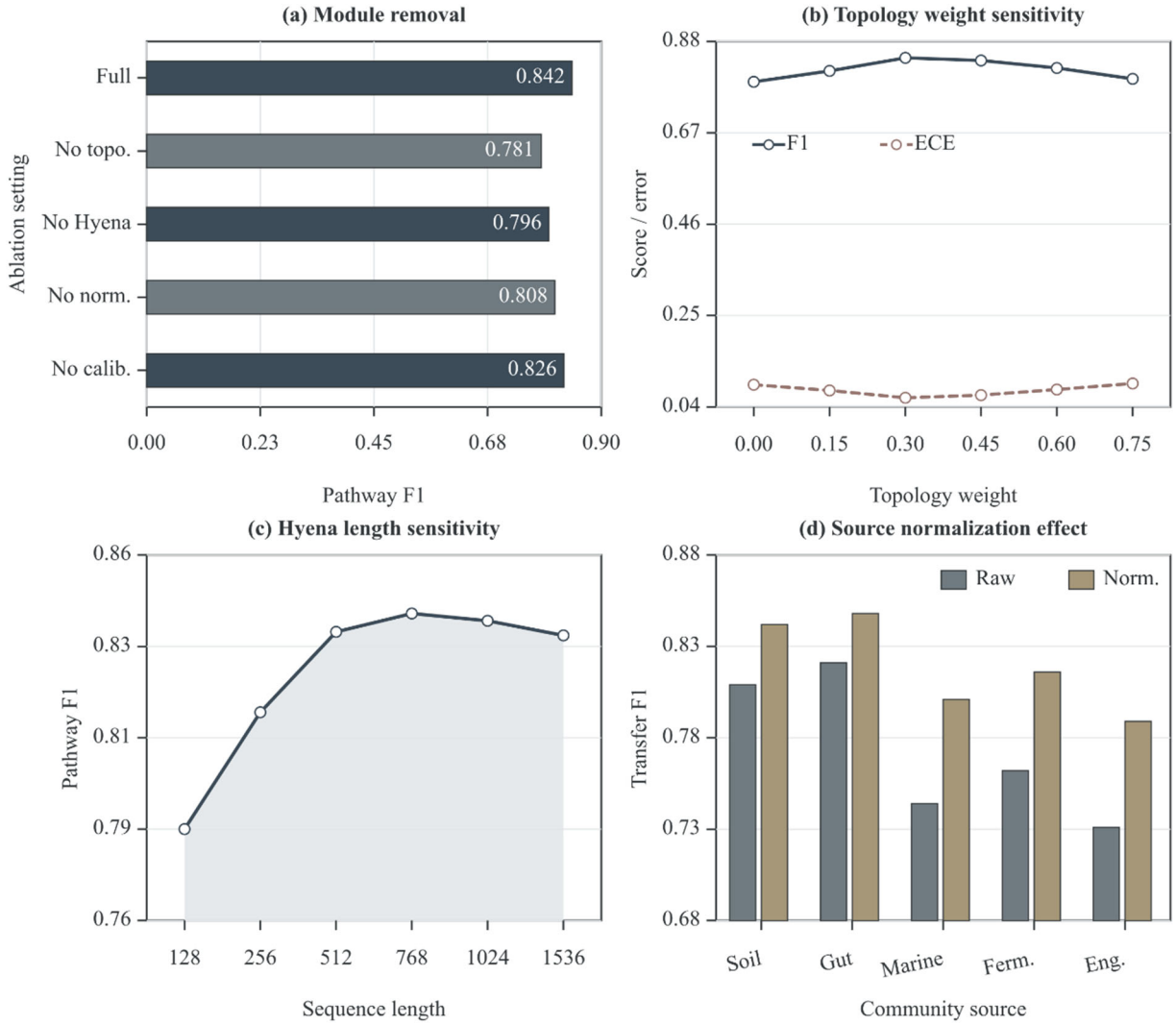


Figure 5. Ablation and sensitivity analysis. (a) module removal. (b) topology weight sensitivity. (c) Hyena length sensitivity. (d) source normalization effect.

The error profile changes after ablation. Without topology, the model may assign pathways to abundant taxa even when enzyme support is weak. Without the Hyena gate, local feature blocks dominate and distributed functions are harder to recover. Without calibration, low-evidence annotations receive overly high confidence. The revised text therefore presents the model as a prediction-and-review system: pathway labels and confidence must be considered together before ecological or clinical interpretation.

4.3. Cross Community Transfer and Error Interpretation

Cross-community transfer tests whether the trained model can annotate unseen community sources and perturbed networks. Metagenomics of food waste fermentation [28] has shown that microbial succession can change functional potential, and colorectal adenoma metagenomic profiling [29] has indicated that disease-associated communities may have different functions. Therefore, transfer tests are required to avoid overestimating the performance because of the presence of similar communities in both training and testing. In the simulated unseen-source test, the proposed model maintains higher pathway F1 and lower calibration error than graph-only and Hyena-only baselines. The improvements are not uniform in all modules. Common metabolic pathways are easily transferred because they are frequently supported by multiple sources, but specialised resistance or biosynthesis genes are difficult to spread when the unseen community contains unfamiliar taxon combinations.

Figure 6 shows cross-community transfer. Figure 6(a) reports unseen-source performance, where topology constraints improve stability for marine and engineered community records. Figure 6(b) reports topology-perturbation error, showing that random edge addition is more harmful than random edge removal. The revised interpretation explains why false biological connections are risky: they can support incorrect pathway labels, while a few missing weak co-occurrence edges may be handled by gene evidence and calibration.

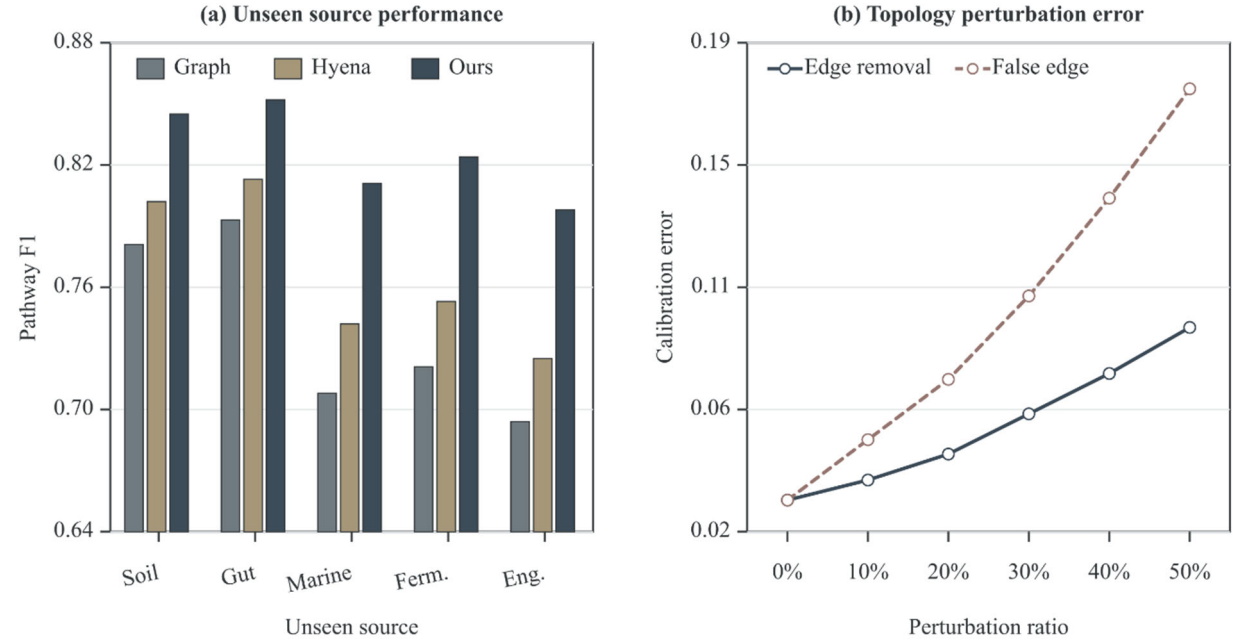


Figure 6. Cross community transfer and topology perturbation. (a) unseen source performance. (b) topology perturbation error.

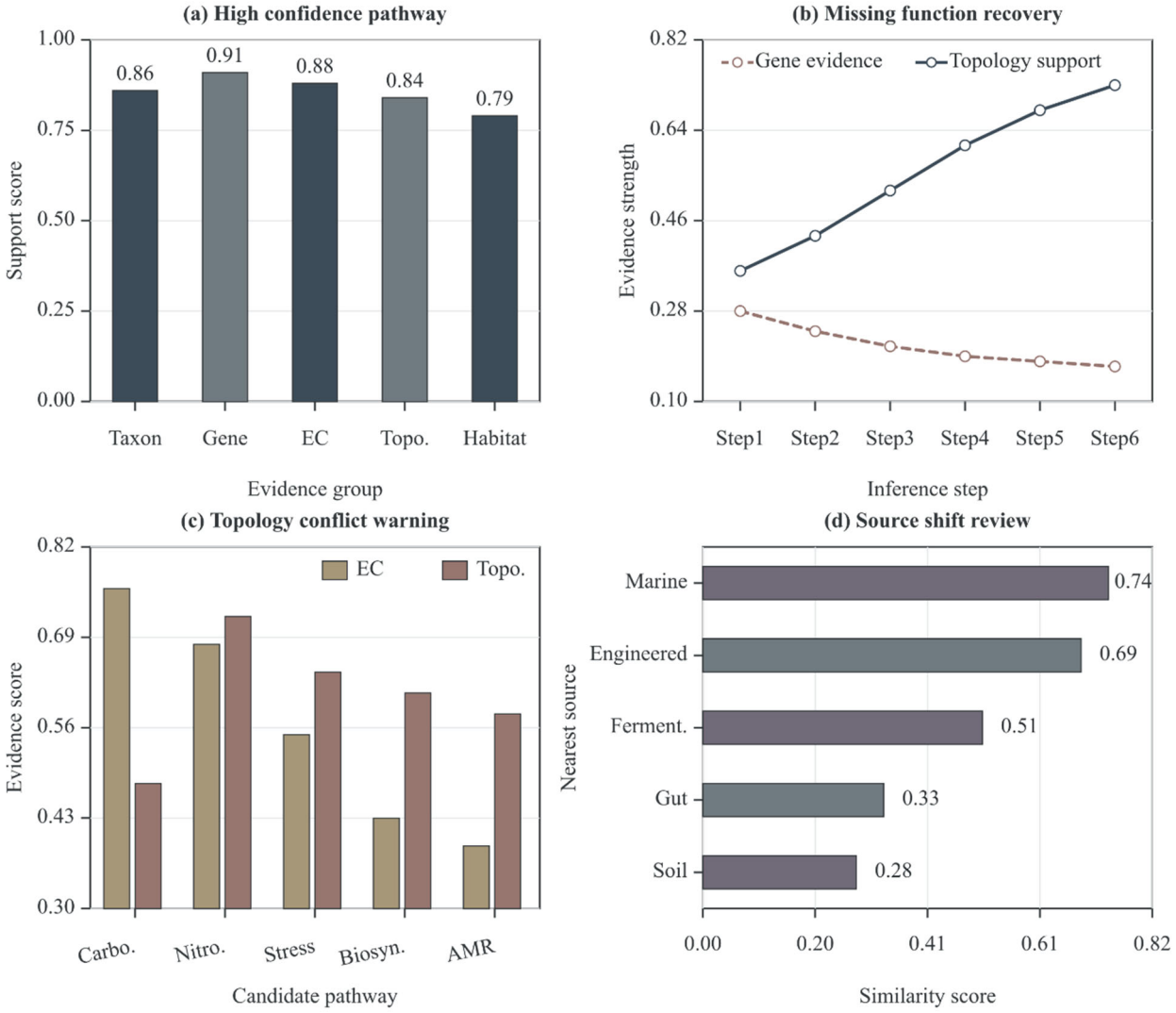


Figure 7. Annotation review card and dominant evidence output. (a) high confidence pathway. (b) missing function recovery. (c) topology conflict warning. (d) source shift review.

Error interpretation focuses on cases where the model assigns an incorrect path or provides a low-confidence annotation for a correct annotation. Deep learning reviews for microbiome analysis [30] support the general idea that neural network models can identify weak nonlinear signals, but microbial annotation still fails when gene-family evidence is sparse, topology edges are noisy or pathway modules overlap. Therefore, the review output is designed to indicate whether the error is likely due to an abundance shift, missing gene evidence, topology conflict or source-context mismatch. The two are not the same; therefore, the subsequent measures will differ. Abundance shift may require source-specific normalisation; missing gene evidence may require deeper assembly; topology conflict may require network filtering; and source-context mismatch may require adding comparable training communities.

Figure 7 is the annotation review card and dominant-evidence output. Figure 7(a) shows high-confidence metabolic pathway prediction with agreement among abundance, enzyme evidence and topology. Figure 7(b) shows cautious missing-function recovery when direct gene evidence is unavailable. Figure 7(c) shows a topology-conflict warning, and Figure 7(d) shows a low-confidence source-shift case. The revised text explains that ordinary annotation acceptance, database review, network filtering and extra verification should be handled differently.

Therefore, topology preservation is more suitable for functional evidence distributed among several taxa rather than being concentrated in a single abundant organism. Hyena encoding is most useful when long-range feature interactions need to be preserved without making the model too costly for large metagenomic windows. Therefore, the combined model is suitable for annotation triage; it can prioritise high-confidence pathway assignments, recover plausible missing functions, and identify cases that require provisional annotation. The first residual risk is that the topology inferred from abundance may be affected by shared habitat conditions. Therefore, the model does not require topology as a strict condition, and the review card only flags predictions that are heavily based on network evidence without sufficient gene-family support.

5. Conclusion

This paper proposed a topology-preserving Hyena Operator prediction model for functional annotation of microbial communities. The model builds abundance-function windows, preserves microbial and pathway topology, encodes long-range functional dependencies, and outputs pathway annotation, enzyme profile, missing-function risk, confidence and review priority. The Design is for metagenomic review environments where functional evidence is scarce, dispersed and source-dependent. By keeping the abundance, gene evidence and pathway structure in the same window, the model does not consider microbial functions as independent labels and supports a more realistic interpretation of community-level activity.

Based on the results of the simulated experiments, topology preservation improves missing-function recovery and reduces false high-confidence annotations; the Hyena Operator models long-range associations among taxa, genes and pathway modules. Ablation and sensitivity results show that topology masks, source normalisation and calibration are all required for stable annotation. Cross-community transfer analysis shows that the model is also more stable when it can distinguish between useful community structure and noisy or misleading edges. The above results are very suitable for large-scale studies that need to compare annotations in different habitats, sequencing modes and incomplete functional databases.

Future work will link the simulated framework with large-scale real metagenomic databases, curated pathway databases and experimentally validated community functions. The model can also be extended to include assembly-quality scores, strain-resolved signals and active learning, and uncertain annotations will be selected for expert review. These extensions would make topology-preserving long-sequence models more applicable to microbial ecology, clinical microbiome analysis, fermentation monitoring and engineered microbiome design. Another way is to link the annotation score to a particular time; community functions may be relatively stable at that time but change rapidly after a disturbance, treatment or succession. In this case, the Hyena Operator can be applied at various positions of features and also to ordered sampling windows.

6. Declarations

6.1. Author Contributions

Nika Potočnik contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Tim Korošec contributes to validation, analysis, investigation, data collection, draft preparation. Matic Kavčič contributes to conceptualization, validation, analysis. Ema Hribar contributes to investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

6.2. Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3. Institutional Review Board Statement

Not applicable.

6.4. Informed Consent Statement

Not applicable.

6.5. Data Availability

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6.7. Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized ChatGPT-4o for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Zhang, Y., Bhosle, A., Bae, S., Eckenrode, K., Huang, X., Tang, J., ... & Huttenhower, C. (2026). Predicting functions of uncharacterized gene products from microbial communities. Nature Biotechnology, 44(7), 1126-1141. https://doi.org/10.1038/s41587-025-02813-7

[2] Levy Karin, E., & Steinegger, M. (2025). Cutting-edge deep-learning based tools for metagenomic research. National Science Review, 12(6), nwaf056. https://doi.org/10.1093/nsr/nwaf056

[3] Bai, D., Chen, T., Xun, J., Ma, C., Luo, H., Yang, H., ... & Liu, Y.-X. (2025). EasyMetagenome: A user-friendly and flexible pipeline for shotgun metagenomic analysis in microbiome research. iMeta, 4(1), e70001. https://doi.org/10.1002/imt2.70001

[4] Zhang, Y., Bian, B., & Okumura, M. (2025). Hyena architecture enables fast and efficient protein language modeling. iMetaOmics, 2(1), e45. https://doi.org/10.1002/imo2.45

[5] Roy, G., Prifti, E., Belda, E., & Zucker, J.-D. (2024). Deep learning methods in metagenomics: A review. Microbial Genomics, 10(4), 001231. https://doi.org/10.1099/mgen.0.001231

[6] Rannon, E., Shaashua, S., & Burstein, D. (2025). DRAMMA: A multifaceted machine learning approach for novel antimicrobial resistance gene detection in metagenomic data. Microbiome, 13(1), 67. https://doi.org/10.1186/s40168-025-02055-4

[7] Castillo-Fernandez, J., Gilroy, R., Jones, R. B., Honaker, R. W., Whittle, M. J., Watson, P., ... & Amos, G. C. A. (2026). Waltham catalogue for the canine gut microbiome: A complete taxonomic and functional catalogue of the canine gut microbiome through novel metagenomic based genome discovery. Microbiome, 14(1), 25. https://doi.org/10.1186/s40168-025-02265-w

[8] Mousa, A. A., Zhang, H., Duan, H., Zhang, J., & Mao, S. (2025). Metagenomic analysis reveals rumen microbiome enrichment and functional genes adjustment in carbohydrate metabolism induced by different sorting behavior in mid-lactation dairy cows. Animal Microbiome, 7(1), 82. https://doi.org/10.1186/s42523-025-00439-3

[9] Yin, Z., Cao, K., Duan, N., & Zhang, Z. (2025). Microbial community dynamics and functional potential during the natural fermentation of rose: A metagenomic and volatile compound analysis. Enzyme and Microbial Technology, 190, 110703. https://doi.org/10.1016/j.enzmictec.2025.110703

[10] Dixon, M. M., Afkairin, A., Manter, D. K., & Vivanco, J. (2024). Rhizosphere microbiome co-occurrence network analysis across a tomato domestication gradient. Microorganisms, 12(9), 1756. https://doi.org/10.3390/microorganisms12091756

[11] Wang, Z., You, R., Han, H., Liu, W., Sun, F., & Zhu, S. (2024). Effective binning of metagenomic contigs using contrastive multi-view representation learning. Nature Communications, 15(1), 585. https://doi.org/10.1038/s41467-023-44290-z

[12] Han, H., Wang, Z., & Zhu, S. (2025). Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes. Nature Communications, 16(1), 2865. https://doi.org/10.1038/s41467-025-57957-6

[13] Qian, J., Jin, P., Yang, Y., Ma, N., Yang, Z., & Zhang, X. (2024). Protein function annotation and virulence factor identification of Klebsiella pneumoniae genome by multiple machine learning models. Microbial Pathogenesis, 193, 106727. https://doi.org/10.1016/j.micpath.2024.106727

[14] Lutz, K. C., Neugent, M. L., Bedi, T., De Nisco, N. J., & Li, Q. (2025). A generalized Bayesian stochastic block model for microbiome community detection. Statistics in Medicine, 44(3-4), e10291. https://doi.org/10.1002/sim.10291

[15] Zhang, Y., Wang, Y., Zhou, G., Revillini, D., Liu, H., Wu, S., ... & Yuan, X. (2025). Soil microbial networks mediate long-term effects of nitrogen fertilization on ecosystem multiservices. Journal of Ecology, 113(9), 2539-2558. https://doi.org/10.1111/1365-2745.70112

[16] Brauner, M., Cohen, J., Briggs, B. R., & Hennon, G. M. M. (2025). Identifying potential keystone microbes from co-occurrence networks in the Gulf of Alaska. Frontiers in Marine Science, 12, 1646100. https://doi.org/10.3389/fmars.2025.1646100

[17] Gao, B., Xiao, D., Yang, K., Sun, M., Luo, S., Zhang, W., & Wang, K. (2025). Increased vegetation disturbance intensity reduces soil nutrients while enhancing microbial network interactions. Frontiers in Microbiology, 16, 1634424. https://doi.org/10.3389/fmicb.2025.1634424

[18] Yang, X., Loik, M. E., Wu, X., Luong, J. C., Wei, X., & Li, L.-J. (2025). Drought and vegetation restoration lead to shifts in soil microbial diversity and co-occurrence networks in California coastal prairie. Plant and Soil, 512(1-2), 175-189. https://doi.org/10.1007/s11104-024-06625-7

[19] Xue, W., Liu, Z., Zhang, Y., Raza, W., Li, Y., Jiang, L., ... & Wei, Z. (2025). LorBin: Efficient binning of long-read metagenomes by multiscale adaptive clustering and evaluation. Nature Communications, 16(1), 9353. https://doi.org/10.1038/s41467-025-64916-8

[20] Serrana, J. M., Nascimento, F. J. A., & Posselt, M. (2026). Metagenomic mining of environmental resistomes: Comparative performance of deep learning and alignment-based annotation methods. iMetaOmics, 3(2), e70107. https://doi.org/10.1002/imo2.70107

[21] Du, Y., Wang, Y., & Sun, F. (2026). Metagenomic Hi-C protocols for viral genome binning, taxonomic annotation, and interaction network visualization. Current Protocols, 6(3), e70341. https://doi.org/10.1002/cpz1.70341

[22] Keles, E., & Celik, O. (2026). Metagenomic and microbiological analyses of historical manuscripts for bacterial community profiling and bacteria-related biodeterioration assessment. Microbial Cell, 13, 117-130. https://doi.org/10.15698/mic2026.03.871

[23] Li, L. X. (2025). Effect of intercropping on soil microbial diversity and community network. Frontiers in Microbiology, 16, 1588559. https://doi.org/10.3389/fmicb.2025.1588559

[24] Lei, Z., Huang, Z., Zhang, Y., Lyu, C., Huang, J., Xiang, C., ... & Zhou, R. (2026). Stochastic process dominated community assembly during high-temperature Daqu storage, leading to reduced microbial co-occurrence network complexity. International Journal of Food Microbiology, 449, 111594. https://doi.org/10.1016/j.ijfoodmicro.2025.111594

[25] Xu, T., Wei, Z., Zeng, Q., Feng, Q., & Zhi, M. (2026). Unveiling potential driver taxa in subgingival plaque and their roles in mediating periodontitis progression. Journal of Oral Microbiology, 18(1), 2609452. https://doi.org/10.1080/20002297.2025.2609452

[26] Philipp, L., Blagodatskaya, E., Tarkka, M., & Reitz, T. (2025). Soil microbial communities are more disrupted by extreme drought than by gradual climate shifts under different land-use intensities. Frontiers in Microbiology, 16, 1649443. https://doi.org/10.3389/fmicb.2025.1649443

[27] Yu, Y., Wang, C., Pan, X., Ding, C., & Chen, J. (2026). Metagenomic profiling of biliary microbiota reveals distinct microbial and functional features in cholelithiasis and cholecystic polyps. Medicine, 105(24), e49251. https://doi.org/10.1097/md.0000000000049251

[28] Parab, C., Yadav, K. D., & Prajapati, V. (2026). Metagenomic insights into microbial community succession and its functional changes during natural fermentation of food waste. Applied Biochemistry and Biotechnology, 198(3), 2053-2081. https://doi.org/10.1007/s12010-025-05575-9

[29] Guo, Z., Liu, J., Xie, Y., Yan, F., Ren, D., Zhang, Q., & Li, X. (2026). Fecal metagenomic profiling in patients with colorectal adenomas to characterize gut microbial composition and functional potential. Frontiers in Microbiology, 17, 1842365. https://doi.org/10.3389/fmicb.2026.1842365

[30] Przymus, P., Rykaczewski, K., Martin-Segura, A., Truu, J., Carrillo De Santa Pau, E., Kolev, M., ... & Nechyporenko, A.

(2025). Deep learning in microbiome analysis: A comprehensive review of neural network models. Frontiers in Microbiology, 15, 1516667. https://doi.org/10.3389/fmicb.2024.1516667