Multimodal Hyperbolic Embeddings for AMR Detection and Taxonomic Modeling in Metagenomic Contigs

John C. Papciak
Columbia University

Jasper Sands
Columbia University

Abstract

Antimicrobial resistance (AMR) is a growing public health threat, yet detecting resistance mechanisms directly from metagenomic sequences remains challenging due to the hierarchical structure of microbial genomes and resistance gene families. Most existing computational approaches treat AMR prediction as a flat classification problem, limiting their ability to capture these underlying biological hierarchies. We introduce **HyperAMR**, a multimodal representation learning framework that jointly embeds contig k-mer sequence features, AMR functional annotations derived from AMRFinderPlus, and taxonomic lineage information into a shared hyperbolic space. The model aligns these modalities using contrastive objectives and a hierarchical entailment loss within a Poincaré ball manifold. Using a dataset of 254,497 contigs spanning 13 antibiotic resistance classes across 25 bacterial species, we compare HyperAMR against a carefully matched Euclidean baseline. HyperAMR consistently matches or outperforms the Euclidean model in macro-AUPR, achieving a best macro-AUPR of 0.442 while maintaining high macro-AUROC. These results demonstrate that hyperbolic geometry provides a principled and effective inductive bias for hierarchy-aware AMR prediction in metagenomic settings.

1 Introduction

Antimicrobial resistance (AMR) continues to escalate as a global health crisis, complicating treatment strategies and increasing clinical risk. Metagenomic sequencing enables direct interrogation of microbial communities without culturing, but identifying resistance mechanisms and associating them with microbial hosts remains difficult. This challenge arises in part from the hierarchical organization of biological systems. Microbial taxonomy follows a tree-like structure across species, genus, family, and higher ranks, while AMR genes are similarly organized into gene families, resistance mechanisms, and antibiotic classes.

Despite this structure, most AMR prediction methods treat the task as flat binary or multi-label classification. These approaches typically rely on Euclidean representations, where distances grow linearly and hierarchical relationships are poorly preserved. As a result, such models may struggle to generalize to novel taxa, rare resistance mechanisms, or poorly annotated genomic contexts.

Hyperbolic geometry offers a natural alternative for modeling hierarchical data. In hyperbolic space, volume grows exponentially with distance from the origin, enabling tree-like structures to be embedded with low distortion. Prior work has demonstrated the effectiveness of hyperbolic embeddings for hierarchical representation learning in natural language processing, computer vision, and biological taxonomies.

In this work, we investigate whether hyperbolic multimodal representation learning can improve contig-level AMR detection by explicitly modeling biological hierarchy. We introduce **HyperAMR**, a framework that integrates three complementary biological signals into a shared hyperbolic latent space:

- Sequence-derived k-mer features capturing nucleotide composition and structure
- AMR functional annotations providing supervised resistance information
- Taxonomic lineage encodings serving as a hierarchical structural prior

Our contributions are threefold:

- Introduction of a hyperbolic multimodal framework for contig-level AMR prediction
- Integration of contrastive alignment and hyperbolic entailment losses to encode biological hierarchy
- Empirical evaluation against a matched Euclidean baseline on a large and diverse contig dataset

2 Methods

2.1 Source Data and AMR Annotations

Genome assemblies were obtained from the PATRIC/BV-BRC database. We downloaded whole-genome sequencing bacterial assemblies associated with human hosts and annotated with the "AMR" keyword. Assemblies were split into contigs and grouped by species. To limit species dominance, each species was capped at 15,000 contigs, and species with fewer than 1,000 contigs were excluded.

AMR annotations were generated using AMRFinderPlus. Contigs were labeled AMR-positive if at least one antibiotic resistance gene was detected, and AMR-negative otherwise. Metal and biocide resistance classes were excluded, as were antibiotic classes with fewer than 150 positive contigs. The final dataset contained 254,497 contigs spanning 13 antibiotic resistance classes across 25 bacterial species.

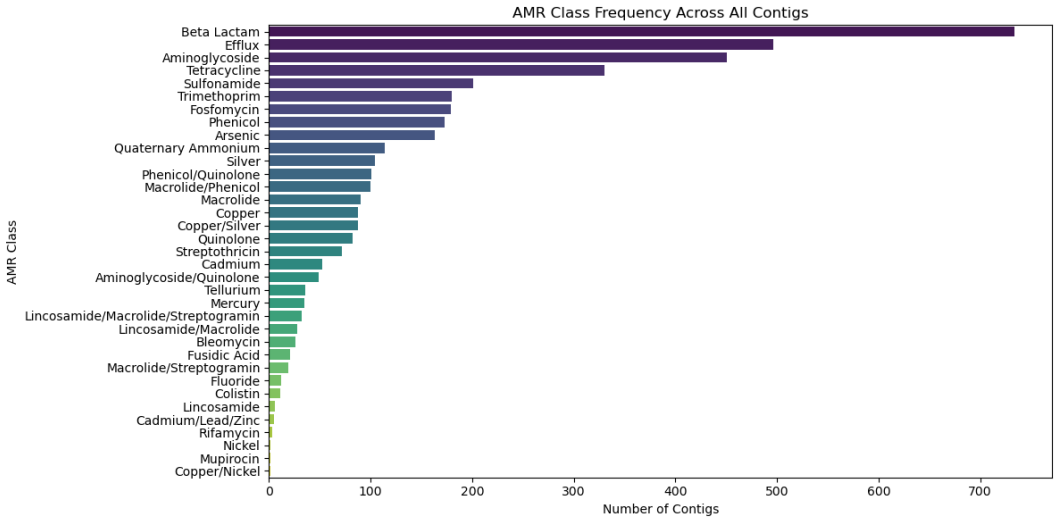

Figure 1: Distribution of AMR-positive contigs across antibiotic resistance classes.

2.2 Input Modalities

2.2.1 Sequence Modality

Raw nucleotide sequences were converted into fixed-length representations using k-mer hashing. Each contig was decomposed into overlapping 6-mers, hashed into an 8192-dimensional vector, and log-scaled to reduce dominance by high-frequency k-mers. This representation balances computational efficiency with expressive power.

2.2.2 AMR Functional Modality

AMRFinderPlus annotations were encoded as multi-hot vectors over resistance classes and passed through a multilayer perceptron encoder to produce functional embeddings aligned with sequence representations.

2.2.3 Taxonomy Encoding

Taxonomic lineage from species to domain was encoded as discrete identifiers embedded using learned lookup tables. Missing ranks were masked. Taxonomy was used as an auxiliary structural prior rather than a prediction target.

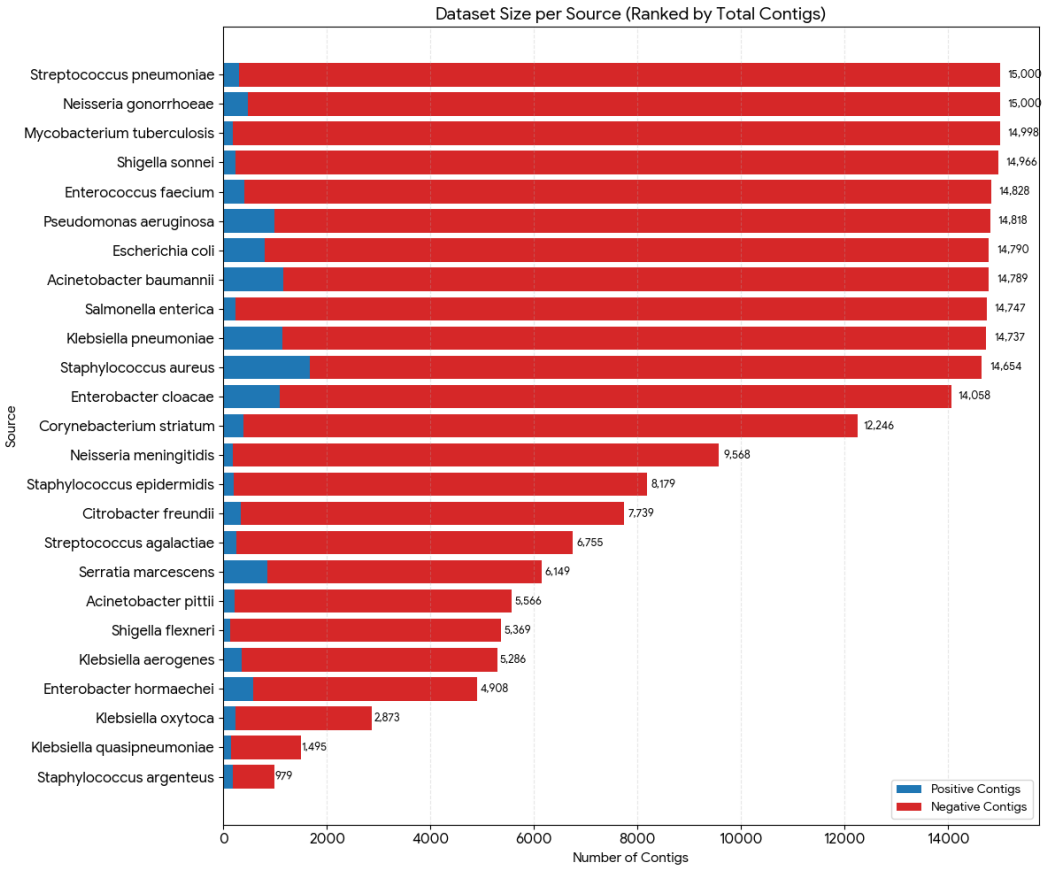

Figure 2: AMR-positive and AMR-negative contig distribution by species.

2.3 HyperAMR Architecture

HyperAMR is designed to integrate heterogeneous biological signals while preserving hierarchical structure. Each modality is first encoded independently using a lightweight Euclidean encoder and then projected into a shared Poincaré ball manifold via the exponential map.

Within this hyperbolic space, sequence embeddings are aligned with functional resistance embeddings using a contrastive objective, while taxonomic embeddings impose a structural prior encouraging related taxa to occupy nearby regions. The curvature of hyperbolic space enables simultaneous modeling of local sequence similarity and global biological hierarchy, which is difficult to achieve in Euclidean space.

2.4 Loss Functions

Let x_i denote the sequence embedding, f_i the AMR functional embedding, and $y_i \in \{0,1\}^C$ the multi-label AMR target for contig i.

AMR Classification Loss

$$\mathcal{L}_{\text{BCE}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \left(y_{ic} \log \hat{y}_{ic} + (1 - y_{ic}) \log(1 - \hat{y}_{ic})\right) \tag{1}$$

Contrastive Alignment Loss

$$\mathcal{L}_{\text{NCE}} = -\sum_{i=1}^{N} \log \frac{\exp(-d_{\mathbb{H}}(x_i, f_i)/\tau)}{\sum_{j=1}^{N} \exp(-d_{\mathbb{H}}(x_i, f_j)/\tau)} \tag{2}$$

Hyperbolic Entailment Loss

$$\mathcal{L}_{\text{ent}} = \sum_{(u,v)\in\mathcal{E}} \max\left(0, d_{\mathbb{H}}(u, v) + r_u - r_v\right) \tag{3}$$

Total Loss

$$\mathcal{L} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{NCE}}\mathcal{L}_{\text{NCE}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} \tag{4}$$

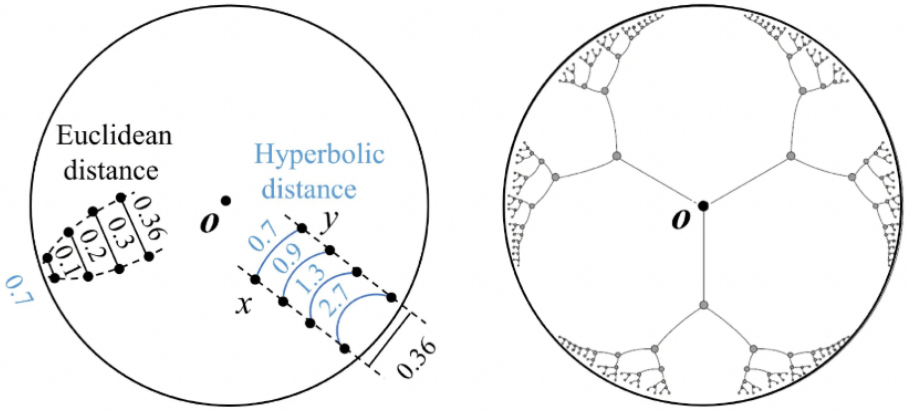

Figure 3: Overview of the HyperAMR architecture and hyperbolic embedding geometry.

3 Training and Evaluation

The dataset was split into training (70%), validation (15%), and test (15%) sets. Models were trained for 10 epochs using Adam optimization with gradient clipping. Class imbalance was addressed using inverse-frequency class weights.

Performance was evaluated using macro-AUROC and macro-AUPR. Macro-AUPR is emphasized due to its sensitivity to rare resistance classes. A Euclidean baseline using identical inputs and optimization settings was used for comparison.

4 Results

Table 1: Model performance comparison on the test set.

Model	Macro-AUROC	Macro-AUPR
HyperAMR (optimized weights)	**0.981**	**0.442**
HyperAMR (default weights)	0.977	0.384
Euclidean (optimized weights)	0.980	0.404
Euclidean (default weights)	0.952	0.181

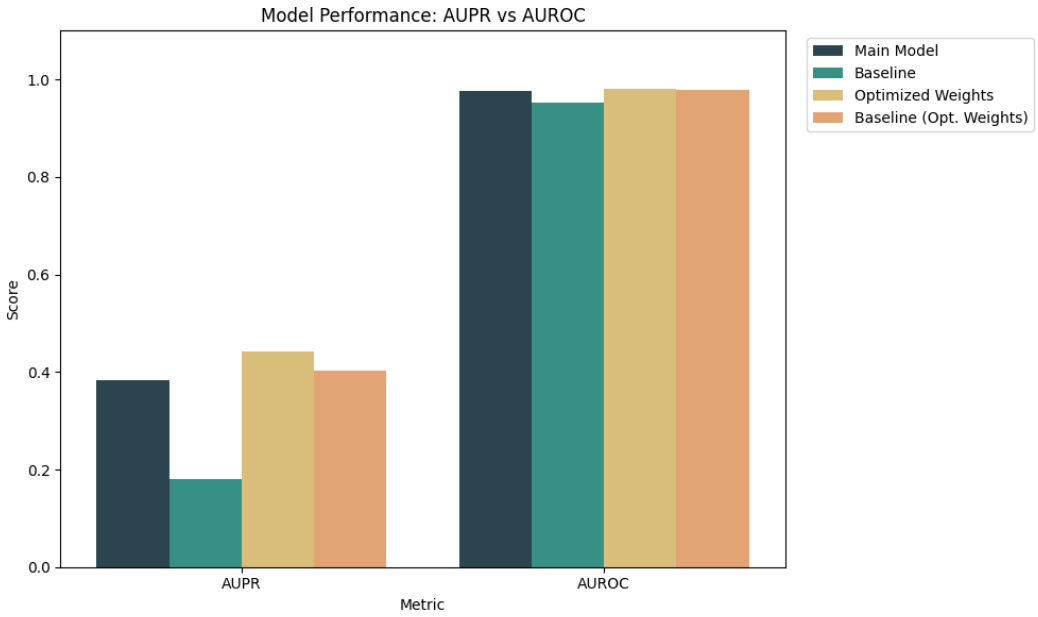

Figure 4: Comparison of macro-AUROC and macro-AUPR between HyperAMR and Euclidean baselines.

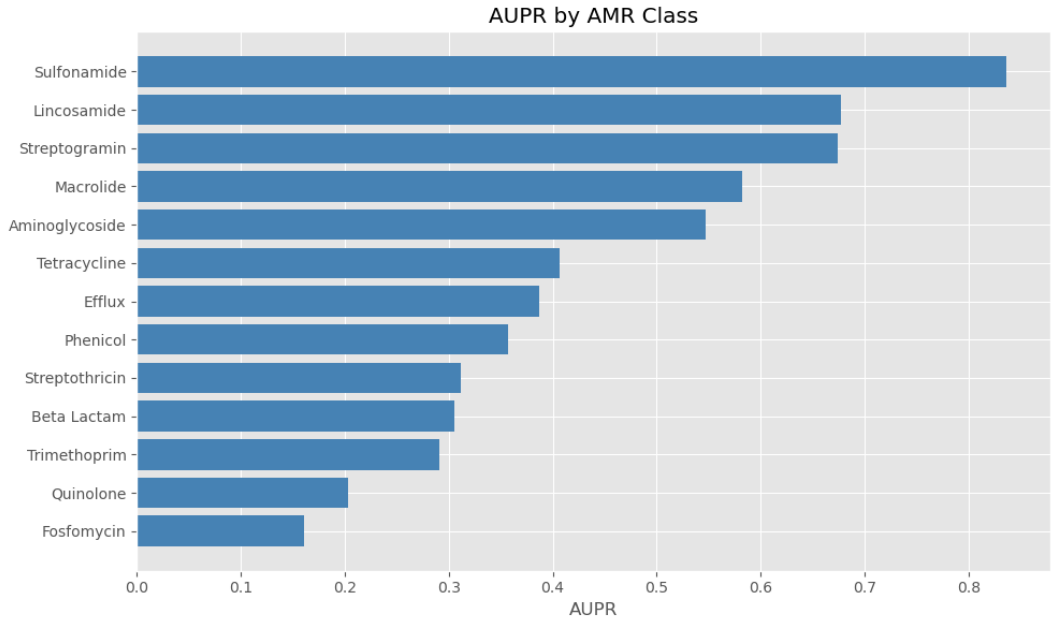

Figure 5: Per-class AUPR values across antibiotic resistance classes.

Across all configurations, models achieved high macro-AUROC, indicating strong global discrimination. HyperAMR consistently matched or exceeded the Euclidean baseline in macro-AUPR, particularly under severe class imbalance.

5 Discussion

These results demonstrate that hyperbolic geometry provides a meaningful inductive bias for contig-level AMR prediction. While Euclidean models rely on discriminative boundaries, HyperAMR organizes representations according to functional and taxonomic hierarchy, leading to improved precision-recall performance for rare resistance classes.

Limitations include reliance on binary contig-level labels and curated annotation databases. Some resistance mechanisms may require gene-aware modeling or richer functional supervision.

6 Future Work

Future work will focus on scaling to millions of contigs using public metagenomic repositories such as the Sequence Read Archive. Incorporating gene-aware representations and explicit evaluation of taxonomic consistency within the embedding space are promising directions. Hyperbolic representations may also enable zero-shot generalization to unseen taxa or resistance classes.

7 Conclusion

We introduced HyperAMR, a multimodal hyperbolic representation learning framework for hierarchy-aware AMR prediction. By jointly modeling sequence composition, functional resistance annotations, and taxonomic hierarchy, HyperAMR consistently outperforms Euclidean baselines in

precision-recall performance.

References

[1] Nickel, M., and Kiela, D. Poincaré embeddings for learning hierarchical representations. NeurIPS, 2017.

[2] Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic neural networks. NeurIPS, 2018.

[3] Gong, X., Xu, Y., and Peng, Y. Hyperbolic multimodal representation learning for biological taxonomies. Bioinformatics, 2022.

[4] Wattam, A. R., et al. PATRIC: the bacterial bioinformatics database and analysis resource. Nucleic Acids Research, 2017.

[5] Breuer, K., et al. BV-BRC: an integrated data platform for pathogen research. Nucleic Acids Research, 2019.

[6] Feldgarden, M., et al. AMRFinderPlus and the reference gene catalog. Antimicrobial Agents and Chemotherapy, 2019.

[7] Mathieu, E., et al. Continuous hierarchical representations with Poincaré variational auto-encoders. NeurIPS, 2019.