MeSH-HyRerank: Hyperbolic Ontology Adaptation for Biomedical Foundation Model Retrieval

Shihong Huang
Faculty of Sciences
Nanjing Agricultural University
Nanjing, China
23123122@stu.njau.edu.cn

Chen Zhang*
Nanjing University
Nanjing, China
231098084@smail.nju.edu.cn

Abstract

Biomedical foundation model systems often access scientific knowledge through retrieval, yet their evidence interfaces are usually scored in lexical or Euclidean spaces while biomedical concepts are organized by curated hierarchies. We propose MeSH-HyRerank, a plug-in hyperbolic ontology adapter for frozen biomedical retrievers. It embeds Medical Subject Headings (MeSH) in a Poincare ball and reranks a retriever's top-K candidates with a geometry-induced query-evidence affinity, leaving both the retriever and downstream reader unchanged. On PubMedQA evidence retrieval, MeSH-HyRerank is the strongest deployable ontology adapter in our controlled comparisons, improving a dense LSA retriever from 0.785 to 0.826 MRR@10 across five splits while randomized MeSH assignments stay at the base-retriever level. For MedCPT, a strong biomedical pre-trained retriever, the adapter targets correctable mistakes and raises hard-subset MRR@10 from 0.372 to 0.434. A frozen Qwen2.5-7B reader probe further shows that answer gains concentrate when ontology geometry recovers missing evidence, with rescue accuracy rising from 0.257 to 0.571 ($p = .0034$). These results support curated non-Euclidean ontology geometry as a lightweight adaptation layer for biomedical foundation model retrieval interfaces.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; • **Computing methodologies** → *Natural language processing*.

Keywords

foundation model retrieval, biomedical RAG, MeSH, hyperbolic embeddings, evidence reranking

1 Introduction

Biomedical foundation models increasingly answer questions by reading retrieved scientific evidence rather than relying only on parametric memory. In medical RAG systems, a retriever selects PubMed abstracts or clinical literature and a downstream language model reads that evidence to produce an answer [32]. This retrieval layer is therefore the interface through which a foundation model sees biomedical knowledge. If the interface ranks evidence from the wrong biomedical neighborhood, the reader can be grounded in the wrong facts even when the generator is unchanged.

The geometry of this interface is often mismatched to the geometry of biomedical knowledge. Standard retrievers score lexical overlap or Euclidean embedding similarity, whereas MeSH, ICD, SNOMED CT, and UMLS encode broad-to-specific relations among diseases, interventions, organisms, and molecular concepts. MeSH is especially relevant for PubMed evidence because it indexes literature with descriptors arranged in hierarchical trees [26, 27]. Hyperbolic space is a natural fit for such branching structure because its volume grows exponentially with radius [13, 28, 29].

We treat biomedical evidence retrieval as a geometry-sensitive foundation model interface. MeSH-HyRerank extracts question MeSH concepts, uses document MeSH descriptors, embeds MeSH hierarchy in a Poincare ball, and reranks a frozen retriever's top-K candidates before evidence is exposed to a downstream reader. This framing is important because the retriever is not merely a search component: it defines the set of facts available to the foundation model at inference time. An ontology-aware interface can therefore improve factual grounding without changing either the retriever encoder or the generative reader. Our contribution is threefold:

- We identify ontology-geometric mismatch as a key bottleneck in biomedical foundation model retrieval and formalize evidence selection as alignment between frozen retrievers and hierarchical medical knowledge.
- We propose MeSH-HyRerank, a plug-in hyperbolic ontology adapter that converts curated MeSH paths into a calibrated rank-correction signal without updating the retriever or reader.
- We provide controlled evidence across lexical, dense, and biomedical pre-trained retrievers, showing that retrieval gains track MeSH-neighborhood alignment and transfer to a frozen Qwen2.5-7B reader when missing evidence is recovered.

2 Related Work

Hyperbolic representation learning. Hyperbolic space supports low-distortion representations of tree-like hierarchies because negative curvature allocates exponentially increasing volume away from the origin [15, 28–30]. This property has been used for text, entailment, graph, and knowledge-graph representations [2, 6, 8, 10, 12, 13]. We use hyperbolic geometry not as a new neural backbone, but as an inference-time metric correction over a frozen biomedical retriever.

Foundation model retrieval interfaces. RAG systems expose external knowledge to large pre-trained models through retrievers and rerankers [14, 17, 20, 21, 23]. Recent work argues that non-Euclidean geometry can better reflect hierarchical or relational data accessed by foundation models [5, 7, 25, 33, 34]. Our setting keeps the retriever and reader fixed and changes only the geometry at the evidence interface.

Biomedical retrieval and ontology priors. MeSH, UMLS, ICD, and SNOMED CT encode curated biomedical relations used for PubMed indexing, semantic search, and concept normalization [1, 4, 11, 24, 26, 27, 31]. Biomedical language models and retrievers such

*Corresponding author.

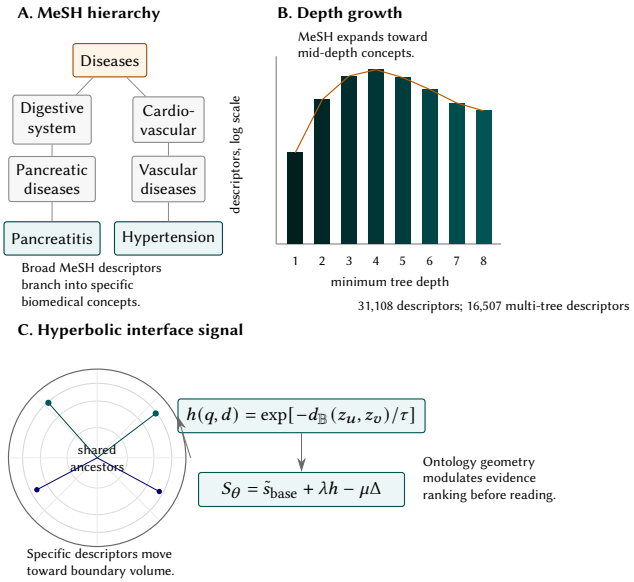

Figure 1: MeSH hierarchy as geometry. MeSH descriptors form broad-to-specific ontological paths; hyperbolic space matches this branching growth by placing shared ancestors near the center and specific concepts toward the boundary.

as SciBERT, BioBERT, PubMedBERT, SPECTER, and MedCPT provide strong evidence-access backbones [3, 9, 16, 19, 22]. MeSH-HyRerank treats MeSH as a metric space for reranking evidence, rather than as metadata or query-expansion text.

3 Method

MeSH-HyRerank is an inference-time ontology adapter at the evidence interface. It leaves the retriever and reader fixed and only changes how top-K retrieved candidates are ordered.

3.1 Evidence Interface

Let $\mathcal{D}$ be a biomedical abstract corpus, d_q^+ the gold evidence for question q, and $s_{\text{base}}(q,d)$ the score from a frozen retriever. We model biomedical RAG as a frozen reader p_ψ whose context is determined by an evidence ordering:

$$\begin{aligned} R_K(q) &= \text{TopK}_{d\in\mathcal{D}}\ s_{\text{base}}(q,d), \\ \pi_\theta(q) &= \text{argsort}_{d\in R_K(q)}\ S_\theta(q,d), \\ E_k(q;\theta) &= (d_{\pi_\theta(1)}, \ldots, d_{\pi_\theta(k)}), \\ p_\psi(y \mid q;\theta) &= p_\psi(y \mid q, E_k(q;\theta)). \end{aligned} \tag{1}$$

The retriever and reader parameters are fixed; MeSH-HyRerank only learns the interface score S_θ that selects which evidence reaches the reader. Question and document concept extractors return $C_q \subset V$ and $C_d \subset V$, where V is the MeSH descriptor vocabulary. Document concepts are PubMed MeSH descriptors; question concepts are normalized matches to MeSH names and synonyms, with generic demographic and study-design descriptors filtered. Figure 1 summarizes the hierarchy prior; Eq. (1) formalizes the adapter pipeline.

3.2 Hyperbolic Ontology Adapter

MeSH descriptors form broad-to-specific tree-number paths and can appear in multiple branches. Let $G_T = (V, E_T)$ be the induced MeSH tree-number graph and $\mathbb{B}_c^p = \{\mathbf{z} \in \mathbb{R}^p : c\|\mathbf{z}\|^2 < 1\}$ the Poincare ball of curvature $-c$. For $\mathbf{x}, \mathbf{y} \in \mathbb{B}_c^p$, the distance is

$$d_{\mathbb{B}}^c(\mathbf{x},\mathbf{y}) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\left(1 + \frac{2c\|\mathbf{x}-\mathbf{y}\|^2}{(1-c\|\mathbf{x}\|^2)(1-c\|\mathbf{y}\|^2)}\right). \tag{2}$$

We build the PubMedQA-relevant MeSH subgraph with ancestors and train 20-dimensional embeddings $\mathbf{z}_u \in \mathbb{B}_c^p$ from parent-child edges:

$$\mathcal{L}_{\text{hyp}} = -\sum_{(u,v)\in E_T} \log \frac{\exp[-d_{\mathbb{B}}^c(\mathbf{z}_u,\mathbf{z}_v)]}{\sum_{v'\in\mathcal{N}(u)} \exp[-d_{\mathbb{B}}^c(\mathbf{z}_u,\mathbf{z}_{v'})]}, \tag{3}$$

where $\mathcal{N}(u)$ contains the true child and sampled negatives, following Nickel and Kiela [28]. For a question–document pair, let $T_m(q,d)$ be the $M = \min(m, |C_q||C_d|)$ concept pairs in $C_q \times C_d$ with smallest hyperbolic distance. The ontology affinity is a top-m partial match:

$$h_{\tau,m}(q,d) = \frac{1}{M} \sum_{(u,v)\in T_m(q,d)} \exp[-d_{\mathbb{B}}^c(\mathbf{z}_u,\mathbf{z}_v)/\tau], \tag{4}$$

with $h_{\tau,m} = 0$ when no concept pair is available. Let $\ell(u)$ be the minimum MeSH tree depth of descriptor u, $\ell_{\max}$ the maximum depth in the induced graph, and $\bar{\ell}(C) = |C|^{-1}\sum_{u\in C}\ell(u)$. We min-max normalize the base score within $R_K(q)$ and penalize normalized depth mismatch:

$$\begin{aligned} \tilde{s}_{\text{base}}(q,d) &= \frac{s_{\text{base}}(q,d) - \min_{d'\in R_K(q)} s_{\text{base}}(q,d')}{\max_{d'\in R_K(q)} s_{\text{base}}(q,d') - \min_{d'\in R_K(q)} s_{\text{base}}(q,d') + \varepsilon}, \\ \Delta(q,d) &= \frac{|\bar{\ell}(C_q) - \bar{\ell}(C_d)|}{\ell_{\max}}, \\ S_\theta(q,d) &= \tilde{s}_{\text{base}}(q,d) + \lambda h_{\tau,m}(q,d) - \mu\Delta(q,d). \end{aligned} \tag{5}$$

For candidates with no document MeSH descriptors, we set $h_{\tau,m}(q,d) = \Delta(q,d) = 0$; if $C_q = \emptyset$, the base ranking is preserved. We tune $\theta = (\lambda, \mu)$ on development queries by

$$\theta^\star = \arg\max_{\lambda,\mu\geq 0} \frac{1}{|Q_{\text{dev}}|} \sum_{q\in Q_{\text{dev}}} \text{RR}_{10}(\pi_\theta(q), d_q^+), \tag{6}$$

and sort candidates by $S_{\theta^\star}$. For a base distractor d^- above the gold evidence, HyRerank corrects the local inversion when

$$\begin{aligned} \lambda\left[h_{\tau,m}(q,d_q^+) - h_{\tau,m}(q,d^-)\right] > \tilde{s}_{\text{base}}(q,d^-) - \tilde{s}_{\text{base}}(q,d_q^+) \\ + \mu\left[\Delta(q,d_q^+) - \Delta(q,d^-)\right]. \end{aligned} \tag{7}$$

Equation (7) is the rank-correction view of the adapter: ontology geometry must overcome the base score gap, while depth alignment lowers the required margin. If no question MeSH concept is found, the base ranking is preserved. MeSHOverlap, RandomHyRerank, and TreeRerank instantiate counterfactual ontology scores:

$$\begin{aligned} h_J(q,d) &= \frac{|C_q \cap C_d|}{|C_q \cup C_d|}, \\ h_R(q,d) &= h_{\tau,m}(C_q, \rho(C_d)), \\ h_T(q,d) &= \frac{1}{M_T} \sum_{(u,v)\in T_m^T(q,d)} \exp[-d_T(u,v)/\tau_T], \end{aligned} \tag{8}$$

where ρ randomly permutes document MeSH assignments and T_m^T selects top pairs by symbolic tree distance d_T. We use $\tau = 2.0$, $m = 5$, and $K = 50$.

3.3 Complexity and Interface Scope

MeSH-HyRerank is applied only after the frozen retriever has produced a shortlist. Let $\bar{c}_d = K^{-1} \sum_{d \in R_K(q)} |C_d|$ be the average number of document descriptors in the shortlist. The adapter cost per query is

$$T_{\text{adapt}}(q) = O\big(K\,|C_q|\,\bar{c}_d\,p + K \log K\big)\,, \tag{9}$$

where the first term scores hyperbolic concept pairs and the second sorts the reranked candidates. This cost is independent of the full corpus size after $R_K(q)$ is available and requires no gradient updates to the retriever or reader. Thus the architecture is a lightweight geometric layer around a foundation model retrieval interface: the base retriever proposes evidence, the ontology adapter changes the ordering, and the frozen reader consumes the resulting top-k contexts. Because the adapter acts after retrieval but before reading, it is compatible with frozen foundation model deployments where retraining the retriever or reader is expensive, unavailable, or undesirable. The learned hyperbolic score can be interpreted as an ontology prior over the evidence list: it promotes candidates that are close to the question in MeSH geometry while preserving the base retriever whenever concept evidence is absent.

4 Experiments

The experiments evaluate MeSH-HyRerank as an evidence-interface adapter rather than as a new retriever trained end-to-end. We test retrieval quality, randomized and lexical controls, behavior on a strong frozen biomedical retriever, hierarchy alignment, and a frozen-reader RAG probe.

4.1 Dataset and Evaluation

We use the 1,000 expert-labeled examples in PubMedQA [18]. Each example contains a question, a PubMed context document, an answer label, and MeSH descriptors. We formulate a controlled closed-corpus evidence retrieval task: for each question, retrieve its gold PubMed context from the 1,000 PubMedQA documents. This isolates the retrieval interface of a biomedical QA/RAG system before generation.

We evaluate Recall@1, Recall@5, Recall@10, MRR@10, and nDCG@10. Each split uses 200 development queries for calibration and 800 held-out test queries for evaluation. We report mean and standard deviation over five random development/test splits with seeds 13, 17, 23, 29, and 31. Reranking coefficients are calibrated on the development split and evaluated on held-out test queries. We use $K = 50$ for reranking.

For downstream RAG, a frozen Qwen2.5-7B-Instruct reader receives the top three retrieved contexts and predicts yes, no, or maybe. The prompt uses only the question and retrieved abstracts; it does not expose the PubMedQA long answer or final decision. We report answer accuracy, macro-F1, and gold-in-top-3 coverage on the full corpus, an LSA hard subset, and a rescue subset where HyRerank recovers evidence missed by LSA:

$$\begin{aligned} \mathcal{H}_B &= \{q : 1 < \text{rank}_B(d_q^+) \le K\}, \\ \mathcal{R}_k &= \{q : d_q^+ \notin E_k(q;B),\ d_q^+ \in E_k(q;\theta^\star)\}, \\ \text{Acc}_{\mathcal{S}}(\theta) &= \frac{1}{|\mathcal{S}|} \sum_{q \in \mathcal{S}} \mathbf{1}\{\hat{y}_\psi(q;\theta) = y_q\}. \end{aligned} \tag{10}$$

Here B is the base retriever and $\hat{y}_\psi$ is the frozen reader prediction from Eq. (1).

Table 1: PubMedQA evidence retrieval. Bold marks MeSH-HyRerank (ours); TreeRerank is a symbolic diagnostic.

Method	R@1	R@10	MRR@10
BM25	.949±.004	.985±.001	.964±.003
BM25 + **MeSH-HyRerank (ours)**	**.950±.006**	**.986±.002**	**.965±.004**
LSA	.706±.006	.945±.003	.785±.003
LSA + MeSHOverlap	.748±.005	.956±.003	.816±.003
LSA + RandomHyRerank	.703±.007	.946±.004	.784±.003
LSA + **MeSH-HyRerank (ours)**	**.756±.006**	**.961±.003**	**.826±.004**
LSA + TreeRerank	.776±.006	.964±.003	.840±.003
MedCPT	.962±.002	.997±.001	.976±.002
MedCPT + MeSHOverlap	.961±.002	.998±.001	.975±.002
MedCPT + RandomHyRerank	.961±.003	.997±.001	.975±.002
MedCPT + **MeSH-HyRerank (ours)**	**.964±.004**	**.998±.001**	**.977±.003**
MedCPT + TreeRerank	.966±.002	.998±.001	.978±.002

4.2 Retrieval Backbones and Controls

We compare three retrieval backbones. BM25 is a lexical baseline. LSA is TF-IDF followed by 128-dimensional truncated SVD and cosine similarity, used as a controlled dense retrieval setting that exposes the ontology signal. MedCPT is the frozen biomedical pre-trained retrieval backbone in our study, with separate query and article encoders trained from PubMed search logs [19]. We use the public NCBI MedCPT query and article encoders as frozen first-stage scoring functions. The controls follow the definitions in the method section: MeSHOverlap tests whether lexical MeSH overlap is sufficient, RandomHyRerank tests whether document-ontology alignment matters, and TreeRerank is a symbolic tree-distance reference for the learned hyperbolic surrogate.

4.3 Main Results

Table 1 shows that the geometric reranker substantially improves the controlled dense retriever, increasing Recall@1 from .706 to .756 and MRR@10 from .785 to .826. The randomized hierarchy control stays at the LSA level, indicating that the gains are not due to adding arbitrary reranking features. Among deployable ontology adapters, MeSH-HyRerank is the strongest method under both the controlled dense LSA backbone and the frozen biomedical MedCPT backbone. TreeRerank acts as a symbolic hierarchy reference, showing that the curated MeSH tree contains strong signal and that learned Poincare geometry captures a useful differentiable surrogate. Figure 2 visualizes the LSA retrieval gain together with the hierarchy diagnostic.

For MedCPT, the full corpus represents a strong-backbone regime: the frozen biomedical pre-trained retriever already reaches .997 Recall@10. Table 2 therefore isolates correctable hard cases where MedCPT ranks the gold evidence in positions 2–50. Upward denotes the fraction of hard queries whose gold rank improves, and

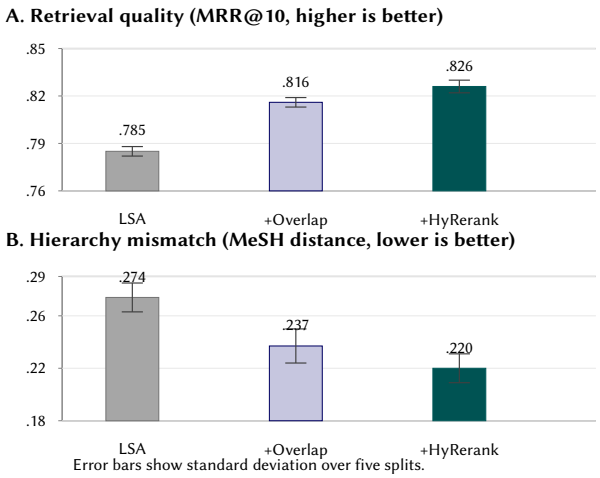

Figure 2: Retrieval and hierarchy diagnostics. MeSH-HyRerank improves MRR@10 while reducing top-1 MeSH distance among deployable adapters.

Gain denotes mean gold-rank improvement. MeSH-HyRerank improves hard-subset MRR@10 from .372 to .434, exceeding both MeSHOverlap and RandomHyRerank; TreeRerank remains the symbolic hierarchy reference.

4.4 Foundation Model Reader Probe

Table 3 and Figure 3 connect evidence reranking to a frozen reader. On the full corpus, MeSH-HyRerank improves gold-in-top-3 exposure from .844 to .874, showing that the adapter changes the evidence seen by the reader. On the evidence-gap rescue subset, where the adapter recovers missing gold evidence, Qwen2.5-7B accuracy increases from .257 to .571 and macro-F1 from .217 to .525. This gain is paired-significant under a two-sided sign test, with 12 improved cases, one regressed case, and $p = .0034$. A select-then-answer rescue ablation further separates ontology geometry from simpler

Table 2: MedCPT hard subset. Bold marks MeSH-HyRerank (ours); TreeRerank is diagnostic.

Method	Upward	Gain	MRR@10
MeSHOverlap	.156±.085	.167±.104	.396±.018
RandomHy	.044±.072	.022±.036	.384±.021
MeSH-HyRerank (ours)	**.244±.146**	**.322±.208**	**.434±.042**
TreeRerank	.344±.025	.428±.087	.461±.019

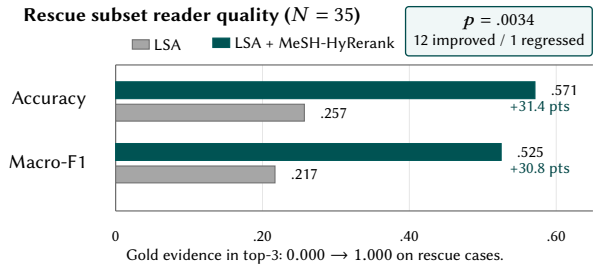

Figure 3: Frozen-reader rescue cases. When MeSH-HyRerank recovers missing evidence, Qwen2.5-7B accuracy and macro-F1 increase sharply.

Table 3: Frozen Qwen2.5-7B RAG probe. Rescue means LSA misses top-3 but MeSH-HyRerank (ours) recovers the gold evidence.

Setting	Method	N	Acc.	F1	Gold@3
Full	LSA	1000	.547	.463	.844
	+HyRerank (ours)	1000	.542	.453	**.874**
Rescue	LSA	35	.257	.217	.000
	+HyRerank (ours)	35	**.571**	**.525**	**1.000**
Hard	LSA	239	.414	.380	.586
	+HyRerank (ours)	239	**.431**	**.383**	.703
	+Tree	239	.431	.370	.736

controls: MeSH-HyRerank reaches .543 accuracy and .416 macro-F1, compared with .486/.354 for MeSHOverlap and .486/.333 for TreeRerank. Thus answer-level gains appear when ontology geometry changes the evidence actually consumed by the foundation model reader.

4.5 Hierarchy Diagnostics

Table 4: Top-1 hierarchy diagnostics. Lower is better.

Method	MeSH Dist.	Depth Gap
LSA	.274±.011	.165±.005
LSA + MeSHOverlap	.237±.013	.148±.004
LSA + RandomHyRerank	.274±.008	.164±.004
LSA + MeSH-HyRerank	.220±.011	.132±.004
LSA + TreeRerank	.195±.011	.131±.004
MedCPT	.021±.004	.021±.002
MedCPT + MeSHOverlap	.017±.006	.022±.002
MedCPT + RandomHyRerank	.022±.005	.021±.002
MedCPT + MeSH-HyRerank	.018±.004	.021±.002
MedCPT + TreeRerank	.018±.005	.021±.002

Table 4 examines whether the reranker retrieves documents that are closer to the gold evidence in MeSH hierarchy. For top-1 retrieved document $\hat{d}_q$, the diagnostics are

$$D_{\text{MeSH}}(q) = \frac{1}{\ell_{\max}} \min_{u \in C_{\hat{d}_q},\, v \in C_{d_q^+}} d_T(u, v),$$
$$G_{\text{depth}}(q) = \frac{|\bar{\ell}(C_{\hat{d}_q}) - \bar{\ell}(C_{d_q^+})|}{\ell_{\max}}. \tag{11}$$

These diagnostics directly test the geometry claim: a useful ontology adapter should not only increase retrieval metrics, but also move the evidence interface toward the gold document's biomedical neighborhood. For LSA, MeSH-HyRerank reduces the top-1 gold-retrieved MeSH distance from .274 to .220 and the depth gap from .165 to .132. The randomized control leaves both metrics essentially unchanged. For MedCPT, the base retriever already returns evidence from a tightly aligned MeSH neighborhood; RandomHyRerank leaves this diagnostic unchanged while the aligned hierarchy variants improve it.

4.6 Evidence Synthesis

Taken together, the experiments cover the three claims required by the paper: ranking improves, retrieved evidence moves closer in MeSH geometry, and a frozen reader benefits when the evidence interface changes. LSA provides headroom signal, MedCPT hard cases show targeted correction, and the controls rule out random or lexical-only explanations. The diagnostic pattern is consistent across abstraction levels: lexical overlap is helpful but incomplete, random hierarchy

assignments do not help, and the learned hyperbolic surrogate gives a deployable adapter for frozen retrieval pipelines.

5 Conclusion and Future Work

We presented MeSH-HyRerank, a hyperbolic ontology adapter for frozen biomedical foundation model retrieval interfaces. Across retrieval metrics, controls, hierarchy diagnostics, MedCPT hard cases, and a frozen Qwen2.5-7B reader probe, curated MeSH geometry improves evidence ranking and hierarchy alignment without retraining the retriever or reader. Future work will scale the evidence interface to open-corpus biomedical retrieval.

References

[1] Alan R. Aronson and Francois-Michel Lang. 2010. An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association* 17, 3 (2010), 229–236.

[2] Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. Multi-Relational Poincare Graph Embeddings. In *Advances in Neural Information Processing Systems*.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[4] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research* 32, Database issue (2004), D267–D270.

[5] Linxiao Cao, Ruitao Wang, Jindong Li, Zhipeng Zhou, and Menglin Yang. 2025. HyperbolicRAG: Enhancing Retrieval-Augmented Generation with Hyperbolic Representations. arXiv:2511.18808

[6] Ines Chami, Zhitao Ying, Christopher Re, and Jure Leskovec. 2019. Hyperbolic Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*.

[7] Weize Chen et al. 2024. Hypformer: Exploring Efficient Hyperbolic Transformer Fully in Hyperbolic Space. arXiv:2407.01290

[8] Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. Fully Hyperbolic Neural Networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

[9] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[10] Bhuwan Dhingra, Christopher J. Shallue, Mohammad Norouzi, Andrew M. Dai, and George E. Dahl. 2018. Embedding Text in Hyperbolic Spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing*. 59–69.

[11] Kevin Donnelly. 2006. SNOMED-CT: The Advanced Terminology and Coding System for eHealth. *Studies in Health Technology and Informatics* 121 (2006), 279–290.

[12] Octavian-Eugen Ganea, Gary Becigneul, and Thomas Hofmann. 2018. Hyperbolic Entailment Cones for Learning Hierarchical Embeddings. In *Proceedings of the 35th International Conference on Machine Learning*.

[13] Octavian-Eugen Ganea, Gary Becigneul, and Thomas Hofmann. 2018. Hyperbolic Neural Networks. In *Advances in Neural Information Processing Systems*.

[14] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2023).

[15] Mikhail Gromov. 1987. *Hyperbolic Groups*. Springer. 75–263 pages.

[16] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. In *ACM Transactions on Computing for Healthcare*, Vol. 3.

[17] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

[18] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2567–2577.

[19] Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with Large-scale PubMed Search Logs for Zero-shot Biomedical Information Retrieval. *Bioinformatics* 39, 11 (2023).

[20] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

[21] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[22] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. In *Bioinformatics*, Vol. 36. 1234–1240.

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*.

[24] Zhiyong Lu, Won Kim, and W. John Wilbur. 2009. Evaluation of Query Expansion Using MeSH in PubMed. *Information Retrieval* 12, 1 (2009), 69–80.

[25] Hiren Madhu, Ngoc Bui, Ali Maatouk, Leandros Tassiulas, Smita Krishnaswamy, Menglin Yang, Sukanta Ganguly, Kiran Srinivasan, and Rex Ying. 2026. HypRAG: Hyperbolic Dense Retrieval for Retrieval Augmented Generation. arXiv:2602.07739

[26] National Library of Medicine. 2026. Medical Subject Headings Home Page. https://www.nlm.nih.gov/mesh/meshhome.html. Accessed: 2026-05-30.

[27] National Library of Medicine. 2026. MeSH Tree Structures. https://www.nlm.nih.gov/mesh/intro_trees.html. Accessed: 2026-05-30.

[28] Maximilian Nickel and Douwe Kiela. 2017. Poincare Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems*.

[29] Maximilian Nickel and Douwe Kiela. 2018. Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. In *Proceedings of the 35th International Conference on Machine Learning*.

[30] Rik Sarkar. 2011. Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane. In *Graph Drawing*.

[31] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator Central: Automated Concept Annotation for Biomedical Full Text Articles. *Nucleic Acids Research* 47, W1 (2019), W587–W593.

[32] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*.

[33] Menglin Yang et al. 2024. Helm: Hyperbolic Large Language Models via Mixture-of-Curvature Experts. arXiv:2410.13842

[34] Menglin Yang, Yifei Zhang, Jialin Chen, Melanie Weber, and Rex Ying. 2025. Towards Non-Euclidean Foundation Models: Advancing AI Beyond Euclidean Frameworks. arXiv:2505.14417 doi:10.1145/3701716.3717806