# Comprehensive Evaluation Metrics for Hierarchical Embeddings in Hyperbolic Space

Hierarchical embeddings in hyperbolic space require specialized evaluation that captures both geometric properties and hierarchical structure preservation. arXiv This report synthesizes classical and recent (2020-2025) metrics from NeurIPS, ICML, ICLR, and domain-specific venues, providing practical guidance for evaluating taxonomy embeddings like NAICS codes embedded using graph neural networks.

## INTRINSIC METRICS

### Distance and Structure Preservation

**Cophenetic Correlation Coefficient (CPCC)** measures correlation between embedding distances and tree distances (cophenetic distances - the height of the lowest common ancestor). Wikipedia Computing for All **Compute by** calculating Pearson correlation between all pairwise embedding distances $\varrho(vi, vj)$ and tree cophenetic distances $dT(vi, vj)$. Values range -1 to 1; **scores >0.9 indicate excellent** hierarchy preservation. Wikipedia +2 Hyperbolic space naturally achieves **CPCC >0.99 for balanced trees** even in 2D, while Euclidean space suffers fundamental distortion limitations requiring high dimensions. Bounded Rationality +2 This metric directly evaluates whether embeddings preserve the tree metric structure, making it foundational for taxonomy evaluation. NeurIPS

**Embedding Distortion** quantifies multiplicative or additive error when mapping hierarchical graphs to continuous space. **Multiplicative distortion** = $\max(dH(f(u),f(v))/dG(u,v)) \times \max(dG(u,v)/dH(f(u),f(v)))$; **additive distortion** = $\max|dH(f(u),f(v)) - dG(u,v)|$. arxiv +3 **Good scores approach 1.0** for multiplicative (typically <1.5) and **approach 0 for additive** (<0.1 × max_distance). Trees can embed in hyperbolic space with arbitrarily low distortion arXiv PubMed Central due to exponential volume growth, while Euclidean requires $\Omega(\log n)$ dimensions for bounded distortion on n-node trees. arXiv +9 This fundamental difference makes distortion the theoretical gold standard for comparing embedding quality across geometric spaces.

**Gromov δ-Hyperbolicity** measures tree-likeness by quantifying how "slim" geodesic triangles are. For any four points, **compute the three distance sums** $S1=d(a,b)+d(c,d)$, $S2=d(a,c)+d(b,d)$, $S3=d(a,d)+d(b,c)$, sort them to find $S1\geq S2\geq S3$, then $\delta=(S1-S2)/2$. ScienceDirect SageMath Use scale-invariant $\delta rel=2\delta/diam(X)$. Perfect trees have **δ=0; values <0.2 indicate strong** tree-like structure suitable for hyperbolic embeddings. Wikipedia Semantic Scholar Hyperbolic embeddings naturally achieve δrel of 0.09-0.15, while Euclidean shows 0.17-0.24. Nature TheCVF This metric serves as a **predictor of when hyperbolic geometry will outperform Euclidean** - datasets with δ<5 consistently show 20-100% improvement in hyperbolic space. ResearchGate

### Hierarchical Relationship Metrics

**Mean Average Precision (MAP) for Tree Reconstruction** evaluates hypernym-hyponym relationship preservation through ranking retrieval quality. arXiv For each node u with hypernyms, **rank all potential hypernyms by embedding distance**, calculate precision at each position where a true hypernym appears, average these precision values to get Average Precision, then compute MAP across all nodes. arxiv Values range 0-1; **MAP >0.9 is excellent** for strict hierarchies. MathWorks State-of-the-art hyperbolic methods achieve **MAP 0.989 in 2D on WordNet** versus 0.87 in 200D Euclidean - representing >10× dimensional efficiency. arXiv +2 This metric captures both recall and ranking quality, making it the standard for taxonomy embeddings in NLP applications.

**Mean Rank** measures the average rank position of true neighbors when all nodes are sorted by embedding distance. ACM Digital Library For each node u with neighbors N(u), **rank all other nodes by distance from u**, record ranks of true neighbors, and average across all edges. Gensim arxiv **Mean Rank <5 is excellent** for large graphs; state-of-the-art hyperbolic methods achieve MR≈1.3 for WordNet versus 3.0-8.0 for Euclidean. Lower values indicate better local neighborhood preservation, crucial for maintaining fine-grained hierarchical relationships. This metric complements MAP by focusing on absolute position rather than precision curves.

**Parent-Child Distance Consistency** evaluates whether parent nodes are consistently closer to the origin than their children, testing radial hierarchy encoding. **For each parent-child pair (p,c)**, measure norms $\|p\|$ and $\|c\|$ from origin, calculate the percentage where $\|p\| < \|c\|$, or compute Spearman correlation between tree depth and norm. arXiv **Consistency >95%** indicates proper hierarchical organization. In Poincaré ball models, this exploits the natural center-to-boundary progression with depth - a unique property of hyperbolic space where depth automatically corresponds to distance from origin. Bounded Rationality +2 Euclidean embeddings lack this natural radial structure and require explicit regularization.

## Level Structure Evaluation

**Coherence Score** measures whether nodes at the same hierarchical level cluster appropriately. For each node ci at level L, **find k-nearest neighbors**, count how many are also at level L, compute Coherence(ci, k) = |neighbors_at_same_level|/k, then average across nodes. wright **Scores >0.7 for k=10** or >0.8 for k=20 indicate good level clustering. wright This captures horizontal relationships in addition to vertical ancestor-descendant relationships. Both hyperbolic and Euclidean can achieve good coherence with proper training, but hyperbolic space provides better natural separation between levels due to exponential volume scaling.

**Categorization Score** tests whether parent nodes are positioned as natural cluster centers for their subtrees. For each concept ck, **compute average embedding of all instances** V̄_ck = (1/n)Σ V_ei, calculate similarity between concept embedding and instance centroid using cosine similarity or hyperbolic distance. wright **Values 0.6-0.9 are good**; higher values indicate concepts are well-positioned centroids. Fine-grained concepts typically score higher than coarse concepts. In hyperbolic space, use Einstein midpoint rather than arithmetic mean for proper centroid computation on the manifold.

## Geometry-Specific Metrics for Hyperbolic Space

**Greedy Routing Success Rate** measures the fraction of source-destination pairs where greedy forwarding (always moving to neighbor closest to destination) successfully reaches the target. [arXiv] **Simulate greedy routing** for random (source, target) pairs by repeatedly hopping to the neighbor with minimum hyperbolic distance to target, detecting cycles and successes. [Nature +2] **Success rate approaching 1.0 is ideal**; good hyperbolic embeddings achieve $ps>0.95$ for scale-free networks, $ps>0.90$ for social networks, and $ps\approx1.0$ for tree structures. [ResearchGate] This metric directly tests whether geometry supports efficient navigation - a fundamental motivation for hyperbolic embeddings. Euclidean embeddings typically achieve only $ps\approx0.6$-$0.7$ for the same graphs.

**Greedy Routing Stretch** measures path efficiency for successful routes by computing the ratio of greedy path length to true shortest path length. **For successful greedy paths**, divide hop count by graph shortest path distance and average. [Cut] [Semantic Scholar] **Stretch approaching 1.0 is perfect**; hyperbolic embeddings typically achieve $\bar{s}\approx1.06$ for Internet AS graphs and 1.1-1.2 for social networks, while Euclidean stretch often exceeds 1.5-2.0 when routing succeeds at all. [ACM Digital Library] This evaluates whether hyperbolic geodesics approximate graph shortest paths due to negative curvature.

**Radial Norm as Hierarchy Indicator** exploits the unique property of Poincaré ball embeddings where $\|\theta\|$ indicates hierarchical depth. Points near center ($\|\theta\|\approx0$) are high in hierarchy; points near boundary ($\|\theta\|\to1$) are leaves. [arXiv +2] **Validate by computing Spearman correlation** between tree depth and embedding norm; **correlation >0.9** (or <-0.9 negative) indicates strong hierarchical encoding. This automatic emergence of hierarchy without explicit supervision is unique to hyperbolic embeddings - Euclidean spaces have no canonical center. [arxiv] [arXiv] Recent work (2024) shows this norm also encodes prediction uncertainty, with familiar samples near boundary and uncertain samples near center. [TheCVF] [arXiv]

**Poincaré Ball Boundary Proximity** analyzes the distribution of norms $\|\theta\|$ within the unit ball to assess space utilization. **Compute mean and standard deviation** of norms across all embeddings, analyzing separately by hierarchy level. **Good embeddings show mean 0.4-0.7** with clear hierarchy where root≈0 and leaves>0.8. Poor embeddings cluster near origin (underutilized space) or boundary (numerical instability). This diagnostic metric identifies numerical issues and capacity problems.

**Learnable Curvature Impact** evaluates performance gain from learning curvature per relation/layer versus fixed curvature $K=-1$. **Compare metrics** (MRR, MAP) between learned and fixed curvature models: $\Delta\text{MRR} = \text{MRR(learned)} - \text{MRR(fixed)}$. Recent work shows **6-8% MRR improvement** on YAGO3-10 with learnable curvature. Different relations and hierarchy depths may require different curvatures; this metric quantifies whether adaptive geometry provides benefit beyond fixed hyperbolic space.

# EXTRINSIC METRICS

## Hierarchical Classification

**Hierarchical Precision, Recall, and F1** provide partial credit when predictions are hierarchically close rather than treating all misclassifications equally. **Expand predicted and true labels** to include all ancestors up to root, then compute $\text{hPrecision} = |\text{predicted\_ancestors} \cap \text{true\_ancestors}|/|\text{predicted\_ancestors}|$ and $\text{hRecall} = |\text{predicted\_ancestors} \cap \text{true\_ancestors}|/|\text{true\_ancestors}|$. **Values 0-1, higher better**; these scores account for hierarchical distance of errors. Misclassifying "poodle" as "terrier" (both dogs) receives partial credit versus "bicycle". For hyperbolic embeddings, add distance-based penalty: $\text{score} = -d(u,v) \times (1 + \alpha(\|v\|-\|u\|))$ where penalty increases for predictions lower in hierarchy. arxiv

**Average Taxonomy Distance (ATD)** weights classification error by rank difference rather than binary correct/incorrect. **Compute Taxonomy Distance** $\text{TD} = (\text{different ranks})/\text{max\_depth\_of\_labels}$ for each prediction, calculate ATD per taxon as mean TD across predictions for that taxon, then compute ATD_by_Taxa as mean across all taxa. **$\text{TD} \in [0,1]$, lower better**; TD=0.11 indicates average half-rank error. biomedcentral Hyperbolic embeddings achieve $\text{ATD} \approx 0.11$ on taxonomic data versus 0.3+ for Euclidean, reflecting superior hierarchical structure preservation.

## Advanced Clustering for Hierarchical Data

**Dendrogram Purity** evaluates hierarchical clustering quality by measuring the extent to which clustering produces pure subtrees. **Construct dendrogram** from embeddings, find dominant class for each cluster at each level, compute $\text{DP} = \Sigma(\text{max\_class\_count\_per\_cluster})/\text{total\_points}$. PubMed Central +2 **DP=1 is perfect; <0.5 is poor**. Hyperbolic space naturally supports hierarchical clustering with distance from origin providing implicit level structure, achieving DP>0.9 on tree-structured data versus 0.7-0.8 for Euclidean.

**Cophenetic Correlation for Clustering** validates distance preservation in hierarchical structure by correlating original pairwise distances with dendrogram cophenetic distances (merge heights). Wikipedia Computing for All **Build dendrogram, extract cophenetic distances**, calculate Pearson correlation. PubMed Central Revoledu **Values >0.7 good, >0.9 excellent**. MathWorks PubMed Central Poincaré embeddings typically achieve $\varrho>0.9$ on hierarchical data due to natural tree structure encoding in hyperbolic geometry.

Beyond basic Silhouette Score and Davies-Bouldin Index, consider **Subtree Cohesion** which measures whether descendants of each internal node cluster together. **For each internal node n**, identify descendants D(n), compute ratio of mean within-subtree distances to mean cross-subtree distances. **Ratios <0.5 indicate good cohesion** (within-subtree points much closer than cross-subtree). Hyperbolic space provides natural nested clustering, achieving 30-50% better cohesion ratios than Euclidean empirically.

## Link Prediction for Hierarchical Graphs

**Parent-Child Link Prediction** tests ability to predict direct edges in the hierarchy using standard ranking metrics. **Hold out parent-child edges**, rank potential parents by distance for each child, compute Mean Rank (MR), Mean Reciprocal Rank (MRR), Hits@K, and MAP. Gensim arXiv **MR<5 is excellent; MAP>0.8 shows strong** discrimination. Poincaré achieves MAP 0.85-0.87 on WordNet versus 0.49 Euclidean at same dimension, exploiting the property that parents are naturally closer to origin than children. arXiv This tests fine-grained local hierarchical structure.

**Sibling Prediction** identifies nodes at the same hierarchy level (shared parent) to test horizontal structure. **For each node**, rank all others by distance, calculate Precision@K and Recall@K for siblings in top-K results. **High precision** (>0.7) means embeddings successfully group siblings. This is often harder than parent-child prediction. In hyperbolic space, siblings have similar norms (depth); using angular distance weighting improves sibling detection.

**Transitive Closure Prediction** evaluates all ancestor relationships (not just immediate parents) to test long-range hierarchical reasoning. **Predict all ancestors** for each node, rank by distance weighted by tree distance, compute hierarchical precision/recall considering depth. **Perfect transitive closure prediction** means the embedding fully captures hierarchy. Performance degradation with depth indicates limited reasoning capacity. Tree structure is inherently transitive; Poincaré distance respects transitivity better than Euclidean.

**Standard Link Prediction Metrics** apply broadly: **Mean Reciprocal Rank (MRR)** = average 1/rank emphasizing top results; **Hits@K** = fraction of true edges in top-K; **Area Under ROC (AUC)** for binary edge classification. Hyperbolic embeddings show **AUC >0.90** for hierarchical networks, with 5-15% improvements over Euclidean on power-law graphs. TheCVF Medium

## Reconstruction Tasks

**Graph Reconstruction** tests whether the original graph can be recovered from embeddings alone - the ultimate capacity test. **Learn embeddings from full graph**, predict edges for all pairs using distance threshold, compute Precision, Recall, F1, MAP on reconstructed edges. Gensim arxiv **F1>0.9 indicates excellent** capacity; lower values suggest insufficient dimensionality. Remarkably, 2D hyperbolic space can perfectly embed trees: Poincaré achieves MAP 0.989 on WordNet with d=2 versus 0.168 for 200D Euclidean. arXiv +2

**Hierarchy Level Recovery** evaluates explicit hierarchy encoding by extracting levels from embeddings. **Extract depth indicator** (norm in hyperbolic), compare to ground truth levels, calculate Spearman correlation between embedding feature and true depth. **Correlation >0.9 indicates strong encoding**. Hyperbolic embeddings achieve $\varrho \approx -0.96$ (negative because norm increases with depth) since norm directly encodes depth with origin=root, boundary=leaves. arXiv Visualize with scatter plot of norm versus depth to verify monotonic relationship.

**Subtree Recovery** tests fine-grained local structure by identifying and separating subtrees correctly. **For each internal node**, identify its subtree, cluster descendants using embeddings, calculate purity, NMI, ARI per subtree, average across subtrees. **High purity** (>0.8) indicates cohesive subtrees in embedding space. Subtrees form natural geometric regions in hyperbolic space, making visual inspection easier in 2D Poincaré projections.

## Semantic Coherence

**Level-wise Semantic Similarity** validates horizontal organization quality by measuring semantic consistency among same-level nodes. **Group nodes by hierarchy level**, calculate within-level pairwise similarity, compute cohesion ratio = within-level similarity / across-level similarity. **Ratio >1** indicates coherent levels; higher values show stronger organization. Use angular distance at fixed radius for level comparisons in hyperbolic space.

**Hierarchy Reflection Score (HRS)** tests whether distances monotonically reflect hierarchical specificity. **For each concept**, compare distances to specific versus generic concepts, check that d(concept, specific) < d(concept, generic), compute HRS = satisfied_comparisons / total_comparisons. **Score approaching 1.0** indicates perfect reflection of specificity. arXiv Hyperbolic embeddings should achieve near-perfect HRS since norm difference naturally captures specificity gradients from root to leaves.

## Zero-Shot and Few-Shot Learning

**Zero-Shot Classification** tests classifying novel classes never seen during training using hierarchical transfer. **Split classes** into seen (training) and unseen (test), train on seen classes, classify into unseen using embedding similarity. Report **Flat Top-K accuracy** (within unseen), **Hierarchical accuracy** (partial credit for close predictions), and **Mean class accuracy**. **Flat Top-1 >0.4 is good**; hierarchical accuracy is more forgiving. Hyperbolic embeddings improve zero-shot by 5-10% over Euclidean since unseen classes inherit properties from nearest ancestors. Springer On WordNet/ImageNet, MAP improves from 0.3 to 0.5+ with hyperbolic space.

**Few-Shot Learning** evaluates classification with K examples per class (K=1,5,10) while leveraging hierarchical structure. **Sample K examples** per novel class, adapt embeddings or learn classifiers, evaluate on remaining examples, compare with/without hierarchy. TheCVF +2 **Performance exceeding non-hierarchical baselines** validates the hierarchical inductive bias. Hyperbolic space shows 3-5% improvement over Euclidean, with larger gains for smaller K. TheCVF Use prototype networks in hyperbolic space with hierarchy as regularizer.

**Hierarchical Zero-Shot Protocols** include multiple evaluation approaches: Flat ZSL (ignore hierarchy), Hierarchical ZSL (consider 2-hops, 3-hops), Generalized ZSL (both seen and unseen classes), and Hit@K at different hierarchy levels. TheCVF Key metrics: **Hierarchical Distance@K** (tree distance of top-K from truth), **Ancestor Accuracy** (correct high-level category?), **Level-wise Accuracy** (stratified by hierarchy level).

## Entailment and Hypernymy Detection

**Lexical Entailment (Binary)** predicts whether X is-a Y, testing core hierarchical relationships. **For each pair (X,Y)**, predict entailment using distance threshold or classifier, compute Precision, Recall, F1, Accuracy. Aclweb Standard benchmarks include WBLESS and BLESS. **F1 >0.7 indicates good** hypernymy detection. For hyperbolic space, use asymmetric scoring: $\text{score}(X \rightarrow Y) = -(1 + \alpha(\|Y\|-\|X\|)) \times d(X,Y)$ where penalty increases when Y is lower (higher norm) than X. Achieves **0.86 accuracy on WBLESS**. arxiv

**Graded Lexical Entailment** evaluates degree of entailment on continuous scales rather than binary. **Score**

**Graded Lexical Entailment** evaluates degree of entailment on continuous scales rather than binary. **Score pairs** using distance/asymmetric measures, correlate with human ratings, compute Spearman ϱ or Pearson r. The HyperLex benchmark contains 2,163 noun pairs with ratings [0-10]. arxiv **Spearman >0.5 shows strong** correlation with human judgment. State-of-the-art hyperbolic methods achieve **ϱ=0.512 versus 0.389 Euclidean**, representing 32% improvement. arxiv

**Transitivity and Logical Consistency** checks whether hierarchies maintain logical transitivity: Is-a(A,B) ∧ Is-a(B,C) → Is-a(A,C). **Detect hypernymy for all pairs**, check transitivity violations, compute Consistency = valid_triplets / all_triplets. **Consistency >0.95 indicates good** logical structure. Tree structure is inherently transitive; Poincaré distance respects this property better than Euclidean alternatives.

## Taxonomy Completion

**Taxonomy Expansion** evaluates inserting new concepts at correct positions in existing taxonomies. **Hold out concepts**, predict insertion point (parent node), compute Accuracy@K (correct parent in top-K?), Mean Rank (average rank of correct parent), Position Error (tree distance from predicted to correct), Subtree Purity (coherence after insertion). **Accuracy@1 >0.7 is excellent**; Mean Rank <3 is strong. Hyperbolic methods achieve **96-99% accuracy on WordNet/Gene Ontology** by embedding query concepts and finding nearest nodes as parents. arXiv

**Structure Preservation after Completion** ensures completed taxonomies maintain structural properties. **Monitor tree diameter** (should not increase dramatically), **branching factor distribution** (should match original), **depth distribution** (new concepts at appropriate depths), and **clustering coefficient** (local structure preserved). Significant changes indicate poor additions; should maintain scale-free and hierarchical properties.

## Retrieval and Ranking

**Normalized Discounted Cumulative Gain (NDCG)** measures ranking quality with position-based discounting and graded relevance. **Compute DCG@K** = Σ(rel_i / $\log_2$(i+1)) for i=1 to K, compute IDCG@K with perfect ranking, then NDCG@K = DCG@K / IDCG@K. Wikipedia Milvus **NDCG >0.8 is excellent**, >0.6 acceptable. Milvus Evaluate at multiple K values (1,5,10,20). Use Poincaré distance for relevance scores; natural separation in hyperbolic space produces good NDCG.

**Hierarchical NDCG** extends standard NDCG with hierarchical relevance grading for partial credit. **Define hierarchical relevance** rel_h = f(tree_distance), assign higher relevance for same-subtree items, calculate NDCG with hierarchical relevance weighted by level. **More forgiving than flat NDCG**, accounting for "close enough" retrievals. Particularly valuable for deep taxonomies where fine-grained distinctions matter less than coarse category accuracy.

**Level-Stratified Metrics** partition performance by hierarchy level to test balance across the hierarchy. **Partition by level**, calculate Precision/Recall per level, plot performance versus level, compute weighted average. **Flat performance across levels** indicates good balance; degradation at depth is common but problematic. Hyperbolic embeddings typically maintain more consistent performance across depths due to exponential volume growth matching exponential tree growth.

## Comparison: Hyperbolic vs Euclidean

**Fundamental Capacity Differences:** For trees with n nodes, hyperbolic space achieves arbitrarily low distortion in 2D, while Euclidean requires $\Omega(\log n)$ dimensions. [arXiv +7] This translates to 10-100× dimensional efficiency in practice. WordNet experiments show Poincaré embeddings achieve MAP 0.989 in 2D versus 0.87 for Euclidean in 200D [PubMed Central] - representing 100× compression with 13% quality improvement. [arXiv +2]

**When Hyperbolic Excels:** Data with Gromov $\delta$-hyperbolicity $<5$, particularly tree-like structures ($\delta<1$); [Github] taxonomies and ontologies; scale-free networks with power-law distributions; [ResearchGate] [Wikipedia] low-dimensional regimes ($d<64$); [Iclr] link prediction and hierarchy reconstruction tasks. [arXiv +2] Typical improvements: 20-100% better distortion, 10-30% better MAP/MRR, 5-15% better zero-shot accuracy.

**When Euclidean Remains Competitive:** High-dimensional embeddings ($d>200$); non-hierarchical data (cycles, lattices, grid structures); [ResearchGate] dense graphs with high clustering; situations requiring maximum numerical stability. Euclidean has no coordinate singularities and is more stable for optimization.

**Precision-Dimension Tradeoffs:** Low-dimensional hyperbolic embeddings ($d=2$) achieve excellent metrics but require ~500 bits precision to maintain accuracy near boundary. [arXiv] Higher dimensions ($d=10$) need only 32 bits for similar quality. [arxiv +4] This creates a fundamental choice: favor dimensionality reduction (2-5D with high precision) or computational efficiency (16-128D with standard precision).

## Practical Implementation Considerations

**Metric Selection by Application:** For taxonomy learning, prioritize MAP, Hierarchical F1, CPCC, and Gromov $\delta$. For knowledge graph completion, use MRR, Hits@K, and relation-specific metrics. For GNN applications, focus on task-specific metrics (node classification F1, link prediction AUC) plus smoothness metrics for deep models. Always include diagnostic metrics (distortion, $\delta$-hyperbolicity) to validate geometric fit.

**Essential Evaluation Protocol:** Always report Gromov hyperbolicity of datasets to justify hyperbolic geometry choice [Github] ($\delta<5$ strongly indicates benefit). [ResearchGate] [TheCVF] Evaluate across multiple dimensions (2, 8, 16, 32, 64, 128) to find optimal tradeoff. Compare both fixed curvature $K=-1$ and learnable curvature approaches. Include Euclidean baselines at matched capacity (same dimension) and higher capacity. Visualize embeddings using t-SNE or direct 2D Poincaré disk plots colored by hierarchy level. [arXiv] [wright] Report numerical precision used (critical for points with $\|\theta\|>0.8$ near boundary). Provide error bars or confidence intervals across multiple runs.

**Hyperbolic Distance Computation:** Always use proper Poincaré distance $d(u,v) = \text{arcosh}(1 + 2\|u-v\|^2/((1-\|u\|^2)(1-\|v\|^2)))$ rather than Euclidean distance on the coordinates. arxiv +2 For centroids and averaging, use

Einstein midpoint (Möbius addition) not arithmetic mean. Monitor gradient magnitudes during training; gradients $>10^4$ indicate numerical instability. Project out-of-bounds points back to manifold using clipping: if $\|\theta\|\geq1$, set $\theta \leftarrow \theta/\|\theta\| \times (1-\varepsilon)$ for small $\varepsilon$. PubMed Central

**Computational Complexity:** Most metrics scale $O(n^2)$ for n nodes since they require pairwise distances. For large graphs (>10K nodes), use sampling approaches: sample random node pairs for distortion metrics; use mini-batches during training for neighborhood metrics; stratified sampling by hierarchy level for level-wise metrics. Hyperbolic distance computation is more expensive than Euclidean (transcendental functions), but the dimensional reduction typically compensates in total compute time.

**Recent Benchmarks (2020-2025):** WordNet remains canonical for taxonomies (82K nouns). arxiv arXiv Gene Ontology provides biomedical hierarchies (27K terms). arXiv ResearchGate Genomic Benchmarks collection (2023) standardizes genomic sequence classification. HyperLex benchmark tests graded entailment (2,163 pairs). arxiv MARBLE (NeurIPS 2023) provides music hierarchy with 4 levels. Temporal knowledge graphs (ICEWS, GDELT) test evolving hierarchies. Critical gap: industrial taxonomies like NAICS lack standardized evaluation frameworks compared to academic benchmarks.

**Software and Tools:** GeomStats library provides Riemannian optimization and evaluation tools. Geoopt (PyTorch) offers standardized hyperbolic operations and metrics for reproducible evaluation. PyTorch Geometric supports hyperbolic graph neural networks. These enable consistent curvature, model, and metric implementations across studies.

## Summary and Recommendations

For comprehensive evaluation of hierarchical embeddings in hyperbolic space, employ a **multi-faceted approach combining intrinsic and extrinsic metrics**. Core intrinsic metrics should include MAP (hierarchy reconstruction), distortion (distance preservation), Gromov δ-hyperbolicity (geometric fit), and CPCC (tree distance correlation). Essential extrinsic metrics include hierarchical F1 (classification quality), MRR/Hits@K (link prediction), zero-shot accuracy (generalization), and NDCG (retrieval quality).

**Hyperbolic embeddings consistently outperform Euclidean by 10-100% on hierarchical tasks**, with gains most pronounced in low dimensions Iclr (2-20D) and for tree-like data ($\delta<5$). ACM Digital Library +2 The natural properties of hyperbolic geometry - exponential volume growth, radial hierarchy encoding, tree-like geodesics - align fundamentally with hierarchical structure. Bounded Rationality +7 This mathematical alignment manifests empirically: 2D Poincaré embeddings match or exceed 200D Euclidean on WordNet link prediction, arXiv PubMed Central achieve 0.512 correlation on HyperLex entailment versus 0.389 Euclidean, PubMed Central +2 and improve zero-shot learning by 5-10%. PubMed Central arXiv

**Critical considerations for NAICS codes and industrial taxonomies:** These deep multi-level hierarchies

**Critical considerations for NAICS codes and industrial taxonomies:** These deep multi-level hierarchies (NAICS has 6 levels) particularly benefit from hyperbolic space since distortion grows with depth in Euclidean but remains bounded in hyperbolic. Focus evaluation on level-stratified metrics to ensure consistent performance across all hierarchy depths. Use taxonomy expansion metrics to test whether embeddings can correctly place new codes as industries evolve. Monitor norm-depth correlation to verify automatic hierarchy encoding. The lack of standardized benchmarks for industrial taxonomies represents a gap; adapt protocols from WordNet and Gene Ontology research.

**The field has matured significantly from 2020-2025** with standardized core metrics (distortion, MAP, MRR, Gromov δ), domain-specific protocols for biomedical and taxonomic applications, and diagnostic tools (hyperbolicity measurement, radius analysis). Future directions include integration with foundation models, efficiency improvements through quantization and mixed-curvature spaces, and development of standardized evaluation suites comparable to MTEB for hierarchical embeddings. The convergence on complementary metrics enables rigorous multi-dimensional assessment of representation quality, downstream performance, and semantic coherence - essential for deploying hierarchical embeddings in production systems.