MANY-TO-MANY RELATIONSHIP MODELING VIA BI-ORDER HYPERBOLIC ENTAILMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperbolic entailment cones are effective for modeling hierarchical and multi-modal relationships and have been applied to diverse fields. However, existing multimodal formulations typically assign one modality to be universally more general than the other. This one-sided inductive bias is suitable for one-to-many relationships, but becomes suboptimal in many-to-many settings such as recommendation, protein–ligand binding, and protein function annotation, where neither modality is inherently more general. To alleviate this limitation, we propose Bi-Order Entailment (BiOE), a product-manifold framework that decomposes many-to-many relationships into directional latent orders modeled by specialized hyperbolic entailment factors. To handle noisy and incomplete data, BiOE further introduces confidence-weighted pseudo-order supervision from neighborhood and semantic-cluster coverage, together with a lightweight residual compatibility factor for non-hierarchical interactions. We theoretically and empirically show that single-entailment models cannot faithfully capture both compatibility and induced order structure, and that they introduce direction-dependent retrieval biases. Across recommender systems, protein–ligand binding, and protein function prediction, BiOE significantly improves over hyperbolic baselines and single-entailment variants. Our experiments also introduce new hyperbolic downstream pipelines, including the first hyperbolic Collaborative Filtering–LLM alignment method and one of the first hyperbolic gene ontology classification models.

1 INTRODUCTION

Recent work in machine learning has moved beyond Euclidean architectures toward models designed for non-Euclidean geometries (Mettes et al., 2023; Yang et al., 2022b; He et al., 2025b). Among these, hyperbolic neural networks have attracted substantial interest, as their exponentially growing volume suits them for representing hierarchical and tree-like structures (Ganea et al., 2018a; Shimizu et al., 2020; Chen et al., 2021b; He et al., 2025c), whereas Euclidean embeddings can struggle to represent such data (Krioukov et al., 2010; Sarkar, 2011). As a result, hyperbolic representation learning has been explored across domains, including natural language processing (Nickel & Kiela, 2017; 2018; Dhingra et al., 2018; Chen et al., 2021a), Transformers and large language models (Yang et al., 2024c; He et al., 2025a; Chen et al., 2021b), graph learning (Chami et al., 2019; Zhang et al., 2021; Yang et al., 2024c), recommender systems (Yang et al., 2022a; Sun et al., 2021), and image representation (Desai et al., 2023; Pal et al., 2025).

One technique used in recent hyperbolic models is the entailment cone (Desai et al., 2023; Ganea et al., 2018b), which has been applied to image-text representation (Desai et al., 2023; Pal et al., 2025; Yoshikawa & Matsubara, 2026), video retrieval (Li et al., 2025), mRNA language modeling (van Spengler et al., 2025), protein-ligand binding (Wang et al., 2026), recommender systems (Bai et al., 2021), and more (Hasan et al., 2026). Each point is associated with a uniquely defined entailment cone, and related embeddings are encouraged to lie within it. However, in multi-modal settings, hyperbolic entailment cones typically assign one modality as the more general entity (Desai et al., 2023), treating the other as more specific. Since the entailment objective further pushes the cone-owning modality closer to the origin, this assumption is amplified in hyperbolic models, where radius often encodes hierarchy levels (Desai et al., 2023; Yang et al., 2024a). While this assumption suits highly one-sided relationships, such as image-text modeling where one text description may correspond to millions of images but not vice versa (Desai et al., 2023; Pal et al., 2025), it becomes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

suboptimal for many-to-many relationships. For instance, in recommender systems, users can prefer many items and items can be preferred by many users. Thus, neither entity is inherently more general, and using a single-entailment cone endows the model with an incorrect inductive bias.

To address this limitation, we propose **Bi**-**O**rder **E**ntailment (BiOE), a product-manifold framework that decomposes many-to-many relationships into directional latent orders and models them with specialized hyperbolic entailment factors. As real-world data is often noisy and incomplete, we further introduce a soft pseudo-order entailment objective deriving confidence-weighted supervision from neighborhood and semantic-cluster coverage, together with a lightweight residual compatibility factor that captures non-hierarchical interactions. Our contributions can be summarized as follows:

(i) We formalize many-to-many relationships via directional compatibility operators and induced latent preorders, and show that single-sided entailment cones cannot faithfully capture both compatibility and order structure.

(ii) We theoretically and empirically show that single-entailment cones introduce direction-dependent retrieval biases, favoring different regions of the target distribution depending on which modality is assigned as the cone owner.

(iii) We demonstrate the generality of BiOE through extensive experiments on recommender systems, protein–ligand binding, and protein function prediction. Beyond improving over hyperbolic baselines and single-entailment variants, our experiments instantiate BiOE in several new downstream settings: we develop the first hyperbolic alignment technique of Collaborative Filtering to LLM semantics, introduce a simple yet effective hyperbolic protein–ligand retrieval model, and present one of the first hyperbolic approaches to protein function annotation and gene ontology classification.

2 Related Works

Hyperbolic Representation Learning. Hyperbolic representation learning has gained attention as a geometric prior for hierarchical and power-law structured data, due to the exponential volume growth of negative curvature (Peng et al., 2021; Mettes et al., 2023; Yang et al., 2022b; Krioukov et al., 2010). This has led to improved results across WordNet, graphs, social networks, recommendation systems, and vision tasks (Nickel & Kiela, 2017; 2018; Chami et al., 2019; Liu et al., 2019; Yang et al., 2022a; Khrulkov et al., 2020; Ermolov et al., 2022). Within deep learning, HNN and HNN++ introduced foundational hyperbolic layers and operations, while later works developed fully hyperbolic architectures (Ganea et al., 2018a; Shimizu et al., 2020; Chen et al., 2021b; Yang et al., 2024c; Bdeir et al., 2024; He et al., 2025a;d), including Transformers, graph Transformers, vision models, ResNets, and large language models. Recent work has also explored hyperbolic or mixed-curvature Transformers and hybrid hyperbolic components for pretrained Euclidean models (Gulcehre et al., 2019; Cho et al., 2023; Desai et al., 2023; Pal et al., 2025; Yang et al., 2024a).

Hyperbolic Models with Entailment Cones. Hyperbolic entailment cones were proposed (Ganea et al., 2018b; Desai et al., 2023) as a tool for modeling hierarchical relationships. The entailment cone is then applied to diverse fields, including visual language models (Pal et al., 2025; Yoshikawa & Matsubara, 2026; Li et al., 2025; Kim et al., 2024), biological applications (Wang et al., 2026; van Spengler et al., 2025), recommender systems (Bai et al., 2021; Du et al., 2022), semantic segmentation (Hasan et al., 2026), image classification (Fonio et al., 2025), and natural language processing (Hu et al., 2025; Tseng et al., 2023). However, these works largely use single directional entailment cones. Later works have also attempted to improve cone formulations, introducing shadow cones (Yu et al., 2024) and unified cones (Yang et al., 2024b). The closest work to ours is ConE (Bai et al., 2021), which models heterogeneous knowledge graph hierarchies using relation-specific hyperbolic cones in a product embedding space. Unlike ConE, which assumes explicit relation types and learns a separate order structure for each KG relation, our setting considers many-to-many multimodal relations where the relevant orders are latent and direction-dependent.

3 Preliminaries

Lorentz hyperbolic space. An n-dimensional Lorentz hyperbolic space, denoted as $\mathbb{H}^{n,\kappa}$, is a Riemannian manifold with constant negative curvature $\kappa < 0$. Each point $\mathbf{x} \in \mathbb{H}^{n,\kappa}$ can be

parametrized as $[x_t, \mathbf{x}_s]^T$, where $x_t \in \mathbb{R}$ and $\mathbf{x}_s \in \mathbb{R}^n$ are the *time* dimension and *space* dimensions, respectively. For points $\mathbf{x}, \mathbf{y} \in \mathbb{H}^{n,\kappa}$, their *Lorentz inner product* is given by $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = -x_t y_t + \mathbf{x}_s^T \mathbf{y}_s = \mathbf{x}^T \mathfrak{g}_n \mathbf{y}$, where $\mathfrak{g}_n = \mathrm{diag}(-1, 1, \ldots, 1)$. Formally, the space is $\mathbb{H}^{n,\kappa} = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = 1/\kappa,\ x_t > 0\}$. The origin $\mathbf{o} \in \mathbb{H}^{n,\kappa}$ is $[\sqrt{-1/\kappa}, 0, \ldots, 0]^T$.

Hyperbolic Entailment Cones. Hyperbolic entailment cones in Lorentz space are defined through the half-aperture of each $\mathbf{x} \in \mathbb{H}^{n,\kappa}$, given as $\mathrm{aper}(\mathbf{x}) = \sin^{-1}\left(2\gamma \left(\sqrt{-\kappa}\|\mathbf{x}_s\|\right)^{-1}\right)$, where $\gamma > 0$ controls the boundary behavior near the origin. We refer to $\mathbf{x}$ as the *cone owner*. Given the entailment cone at $\mathbf{x}$, the exterior angle between $\mathbf{x}$ and $\mathbf{y}$ is

$$\mathrm{ext}(\mathbf{x}, \mathbf{y}) = \cos^{-1}\left((y_t - x_t \kappa \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}) \left(\|\mathbf{x}_s\| \sqrt{(-\kappa \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}})^2 - 1} \right)^{-1} \right). \tag{1}$$

The entailment loss is $L_{\mathrm{ent}}(\mathbf{x}, \mathbf{y}) = \max(0, \mathrm{ext}(\mathbf{x}, \mathbf{y}) - \mathrm{aper}(\mathbf{x}))$, so that if $\mathrm{ext}(\mathbf{x}, \mathbf{y}) < \mathrm{aper}(\mathbf{x})$, then $\mathbf{y}$ is inside the cone at $\mathbf{x}$ and is not penalized. The half-aperture grows as $\mathbf{x}$ moves closer to the origin, so for $\mathbf{y}$ to lie inside the cone of $\mathbf{x}$, $\mathbf{x}$ must be closer to the origin than $\mathbf{y}$, so the cone owners are pushed toward the origin (Desai et al., 2023).

Hyperbolic Contrastive Learning. Hyperbolic contrastive models use negative hyperbolic distance, or squared distance, as the similarity score instead of cosine similarity. Given a batch of size B with positive pairs $(\mathbf{x}_j, \mathbf{y}_j)$, the *hyperbolic contrastive loss* is $L_{\mathrm{cont}}(\mathbf{x}_j, \mathbf{y}_j) = -\frac{1}{2} \log \frac{e^{-d_{\mathbb{H}}(\mathbf{x}_j, \mathbf{y}_j)/\tau}}{\sum_{i \neq j}^{B} e^{-d_{\mathbb{H}}(\mathbf{x}_j, \mathbf{y}_i)/\tau}} - \frac{1}{2} \log \frac{e^{-d_{\mathbb{H}}(\mathbf{x}_j, \mathbf{y}_j)/\tau}}{\sum_{i \neq j}^{B} e^{-d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{y}_j)/\tau}}$ for temperature τ. This also applies to hyperbolic recommender systems, whose triplet loss is given as $L_{\mathrm{tri}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \max(0, d^2_{\mathbb{H}}(\mathbf{x}, \mathbf{y}) - d^2_{\mathbb{H}}(\mathbf{x}, \mathbf{z}) + m)$, where $m \geq 0$ is the margin, and $\mathbf{y}, \mathbf{z}$ are the positive and negative samples, respectively. These hyperbolic models encode hierarchy through a sample's distance to the origin, with more general samples embedded closer to it (Yang et al., 2022a; Desai et al., 2023; Pal et al., 2025).

4 Method

4.1 Problem Statement and Theoretical Analysis

We formalize many-to-many relationships by decomposing them into directional compatibility operators and the latent orders they induce. Let U, V be two finite sets of entities and let $\mathcal{R} \subseteq U \times V$ be a bipartite relationship between them. For $u \in U$ and $v \in V$, let $N_V(u) = \{v \in V : (u, v) \in \mathcal{R}\}$ and $N_U(v) = \{u \in U : (u, v) \in \mathcal{R}\}$ denote their neighborhoods. The relationship $\mathcal{R}$ then induces two preorders, namely $\preceq_U, \preceq_V$, where $u_2 \preceq_U u_1 \iff N_V(u_1) \subseteq N_V(u_2)$ and $v_2 \preceq_V v_1 \iff N_U(v_1) \subseteq N_U(v_2)$. Under this convention, $u_2 \preceq_U u_1$ means that u_2 is at least as general as u_1 w.r.t. $\mathcal{R}$, and similarly for $v_2 \preceq_V v_1$. The binary relation can also be viewed as directional compatibility operators, $\rightsquigarrow_U$ and $\rightsquigarrow_V$, defined by $u \rightsquigarrow_U v$ and $v \rightsquigarrow_V u \iff (u, v) \in \mathcal{R}$. These operators encode the same relation but are directional w.r.t. the query side: $u \rightsquigarrow_U v$ represents compatibility when querying from U, and $v \rightsquigarrow_V u$ from V. The goal is to learn an embedding map $f : U \cup V \to \mathcal{X}$ together with directional compatibility maps such that the induced latent preorders agree with $\preceq_U, \preceq_V$ and the latent compatibility maps agree with $\rightsquigarrow_U, \rightsquigarrow_V$.

We now show that single entailment cones or a global ordering score cannot faithfully model such many-to-many relationships. We define a **one-sided monotone order model** to be a transitive order space $(\mathcal{X}, \leq_{\mathcal{X}})$ together with an embedding map $f : U \cup V \to \mathcal{X}$ such that $(u, v) \in \mathcal{R} \iff f(u) \leq_{\mathcal{X}} f(v)$. In the case of hyperbolic entailment cones, $\mathcal{X} = \mathbb{H}^{n,\kappa}$ is an n-dimensional hyperbolic space with curvature κ, and for $\mathbf{x}, \mathbf{y} \in \mathbb{H}^{n,\kappa}$, $\mathbf{x} \leq_{\mathcal{X}} \mathbf{y} \iff \mathbf{y}$ lies inside the cone entailed by $\mathbf{x}$.

Proposition 4.1. *Let U, V be finite sets and let $\mathcal{R} \subseteq U \times V$ be a bipartite relationship.*

- *(i) There exist choices of $(U, V, \mathcal{R})$ such that no global scalar order score $g : U \cup V \to \mathbb{R}$ satisfies $(u, v) \in \mathcal{R} \iff g(u) \leq g(v)$.*
- *(ii) Let f be a one-sided monotone order model. Suppose the same order is also required for the target-side latent preorder, in the sense that $v_a \preceq_V v_b \implies f(v_a) \leq_{\mathcal{X}} f(v_b)$. Then every comparable pair $v_a \preceq_V v_b$ must have identical neighborhoods, i.e. $N_U(v_a) = N_U(v_b)$.*

By (i), a single global notion of generality, such as a scalar radius or ordering score, cannot represent arbitrary many-to-many relationships. By (ii), a one-sided monotone order model, including single-direction entailment cones, cannot generally represent both the cross-modal compatibility relation and a nontrivial target-side preorder with the same order. As a result, while single entailment cones are well suited for one-to-many relationships where one side can naturally act as the general entity, they impose an overly restrictive inductive bias in many-to-many settings.

Biases of Single Entailment Cones. Next, we analyze how single-sided entailment cones introduce direction-dependent biases during inference in hyperbolic models that use negative hyperbolic distance as the similarity score, which is the standard for current hyperbolic multi-modal models. Without loss of generality, assume the task is to retrieve $v \in V$ given a query $u \in U$. Hyperbolic models often encode generality through radius, with more general entities embedded closer to the origin. This is also the embedding assumption of entailment models, as radius directly controls cone size (Desai et al., 2023). Let $b : V \to \mathbb{R}_{\geq 0}$ denote a target-side generality score, e.g. $b(v) = |N_U(v)|$, and let $r(\mathbf{x}) = d_{\mathbb{H}}(\mathbf{o}, \mathbf{x})$ be the radius of $\mathbf{x}$. Here we assume a perfect hyperbolic model, i.e. $b(v_1) \geq b(v_2) \Longrightarrow r(v_1) \leq r(v_2)$. Let $V_H = \{v \in V : b(v) \geq c\}$ and $V_T = \{v \in V : b(v) < c\}$ be a generality partition of V for some $c > 0$. For any $S \subseteq V$, define $\bar{d}(r, S) = \mathbb{E}_{u_r, v \sim S}[d_{\mathbb{H}}(u_r, v)]$, where u_r has radius r and rotationally invariant angular distribution.

Proposition 4.2. *Suppose similarity is given by $-d_{\mathbb{H}}$ and an ideal hyperbolic model. Assume that $r(v_H) \leq r(v_T)$ almost surely for $v_H \sim V_H$ and $v_T \sim V_T$.*

(i) If $u \in U$ are cone owners, then $F(r) = \bar{d}(r, V_T) - \bar{d}(r, V_H)$ is non-increasing in r.

(ii) If $v \in V$ are cone owners, assume the angular distribution of U is rotationally invariant conditional on radius. For $0 < \delta < r(v)$, let $T_\delta(v)$ denote the point obtained by moving v a distance δ along the geodesic toward $\mathbf{o}$, and define $\Delta(v, \delta) = \mathbb{E}_{u \sim U}[d_{\mathbb{H}}(u, v) - d_{\mathbb{H}}(u, T_\delta(v))]$. Then $\Delta(v, \delta)$ depends only on $r(v)$ and is non-decreasing in $r(v)$.

In the $U \to V$ entailment direction, the query-side embeddings are encouraged to move toward the origin. Since $F(r)$ increases as r decreases, head elements in V_H become relatively closer than tail elements in V_T on average, biasing retrieval toward more general targets. In the opposite $V \to U$ direction, moving a target toward the origin produces a larger average distance reduction for larger-radius targets. This gives larger score gains to less general or tail targets. We empirically validate these opposite effects in Section 5.4.

4.2 Proposed Method

We introduce BiOE, a product-manifold framework for modeling many-to-many relationships by using multiple hyperbolic factors, each specializing in one directional latent order.

Bi-order Entailment Factors. Since $\mathrm{ext}(\mathbf{x}, \mathbf{y}) < \mathrm{aper}(\mathbf{x})$ only when $y_t > x_t$, two hyperbolic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{H}^{n,\kappa}$ cannot simultaneously entail each other. Thus, naive two-sided entailment in a single hyperbolic space is not possible. To address this, we use a product of hyperbolic factors $\mathbb{H}^{\mathcal{K},\ell,n} = \prod_{i=1}^{\ell} \mathbb{H}^{n_i,\kappa_i}$, where $\mathcal{K} = \{\kappa_1, \ldots, \kappa_\ell\}$ is the set of curvatures and $\sum_{i=1}^{\ell} n_i = n$. Each factor is assigned to one entailment direction induced by the observed relation. For the i-th factor, we denote the entailment loss from U to V as $L_{\mathrm{ent}}^{U \to V}(u_j^{(i)}, v_j^{(i)}) = L_{\mathrm{ent}}(u_j^{(i)}, v_j^{(i)})$, and the entailment loss from V to U as $L_{\mathrm{ent}}^{V \to U}(u_j^{(i)}, v_j^{(i)}) = L_{\mathrm{ent}}(v_j^{(i)}, u_j^{(i)})$. These losses encourage each factor to specialize toward a distinct directional latent order.

Residual Compatibility. Since not all relationships are strictly hierarchical, the full model also uses a small residual compatibility factor to capture interactions not captured by hyperbolic entailment factors alone. The final embedding space is $\mathcal{X} = \mathbb{H}^{\mathcal{K},\ell,n} \times \mathcal{Z}_{\mathrm{res}}$, where $\dim(\mathcal{Z}_{\mathrm{res}}) \ll n$. For the hyperbolic factors, we use the ℓ_1-product metric on $\mathbb{H}^{\mathcal{K},\ell,n}$, which has been shown to improve compositionality modeling (Yoshikawa & Matsubara, 2026), given as $d_{\mathbb{H}}^{\ell_1}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\ell} \alpha_i d_{\mathbb{H}}^{\kappa_i}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, where $\mathbf{x}^{(i)}, \mathbf{y}^{(i)}$ are the i-th components and $\alpha_i > 0$ are weights. The similarity score is then $s(u, v) = -(d_{\mathbb{H}}^{\ell_1}(u, v))^2 + \gamma s_{\mathrm{res}}(u, v)$, where s_{res} denotes the similarity score in $\mathcal{Z}_{\mathrm{res}}$.

Soft Entailment Induced Pseudo Order. Real-world data is often noisy and incomplete, making exact neighborhood inclusion too brittle for constructing latent orders. We therefore construct soft

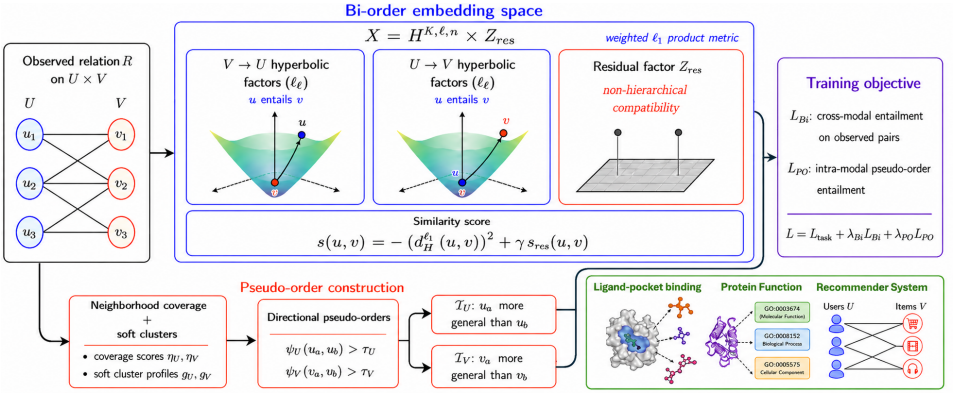


Figure 1: Overview of BiOE. Given an observed relation $\mathcal{R} \subseteq U \times V$, entities are embedded into a bi-order product space with separate hyperbolic factors for the $U \to V$ and $V \to U$ entailment directions, together with an optional residual factor for non-hierarchical compatibility. Observed pairs supervise the cross-modal bi-order entailment loss L_{Bi}, while weighted neighborhood coverage and soft cluster profiles construct confidence-weighted pseudo-orders for the intra-modal pseudo-order loss L_{PO}. Final predictions use the weighted ℓ_1 product hyperbolic distance and residual similarity.

pseudo-orders from weighted neighborhood coverage:

$$\eta_U(u_1,u_2) = \frac{\sum_{v\in N_V(u_1)\cap N_V(u_2)} w_v}{\sum_{v\in N_V(u_2)} w_v}, \qquad \eta_V(v_1,v_2) = \frac{\sum_{u\in N_U(v_1)\cap N_U(v_2)} w_u}{\sum_{u\in N_U(v_2)} w_u}, \tag{2}$$

where w_v, w_u are generality weights computed from empirical neighborhoods. Thus, $\eta_U(u_1,u_2)$ measures how much of u_2's neighborhood is covered by u_1, and similarly for $\eta_V(v_1,v_2)$. To make the pseudo-orders more robust, we additionally compute cluster-level coverage scores $\eta_U^{\mathrm{cl}}, \eta_V^{\mathrm{cl}}$ using soft semantic cluster assignments for entities in U and V. These cluster-level scores measure coverage over semantic neighborhoods rather than exact observed neighbors, allowing the pseudo-order supervision to account for missing or noisy relations. The final coverage scores are

$$\bar{\eta}_U(u_1,u_2) = (1-\lambda)\eta_U(u_1,u_2) + \lambda\eta_U^{\mathrm{cl}}(u_1,u_2), \quad \bar{\eta}_V(v_1,v_2) = (1-\lambda)\eta_V(v_1,v_2) + \lambda\eta_V^{\mathrm{cl}}(v_1,v_2),$$

where $\lambda \in [0,1]$ controls the strength of the soft-cluster signal. We then obtain directional pseudo-order scores by antisymmetrizing the coverage scores:

$$q_U(u_1,u_2) = \bar{\eta}_U(u_1,u_2) - \bar{\eta}_U(u_2,u_1), \qquad q_V(v_1,v_2) = \bar{\eta}_V(v_1,v_2) - \bar{\eta}_V(v_2,v_1). \tag{3}$$

A positive value $q_U(u_1,u_2) > 0$ suggests that u_1 is more general than u_2, since u_1 covers most of u_2's neighborhoods but not vice versa, with the magnitude serving as a confidence score. The same interpretation applies to q_V. We then form pseudo-order sets

$$\mathcal{T}_U = \{(u_a,u_b,q) : q = q_U(u_a,u_b) > \tau_U\}, \qquad \mathcal{T}_V = \{(v_a,v_b,q) : q = q_V(v_a,v_b) > \tau_V\}, \tag{4}$$

where $\tau_U, \tau_V > 0$ are confidence thresholds. These pseudo-orders provide confidence-weighted entailment supervision within the corresponding bi-order factors. Full details of the cluster-level construction are deferred to Appendix B.1.

Overall Framework. The overall framework is shown in Figure 1. Each pair $x \in U \cup V$ is embedded into the product space $\mathbb{H}^{\mathcal{K},\ell,n}$, where the first ℓ_U factors model the $U \to V$ direction and the remaining $\ell_V = \ell - \ell_U$ factors model the $V \to U$ direction. The bi-order entailment loss is

$$L_{\mathrm{Bi}}(u_j,v_j) = \sum_{i=1}^{\ell_U} L_{\mathrm{ent}}^{U\to V}(u_j^{(i)}, v_j^{(i)}) + \sum_{i=\ell_U+1}^{\ell} L_{\mathrm{ent}}^{V\to U}(u_j^{(i)}, v_j^{(i)}). \tag{5}$$

This defines the base variant $\mathrm{BiOE}_{\mathrm{Bi}}$, which uses the product hyperbolic factors with objective $L = L_{\mathrm{task}} + \lambda_{\mathrm{Bi}} L_{\mathrm{Bi}}$ and similarity score $s(u,v) = -(d_{\mathbb{H}}^{\ell_1}(u,v))^2$. Using the pseudo-order sets $\mathcal{T}_U, \mathcal{T}_V$ constructed above, we define the pseudo-order entailment loss as

$$L_{\mathrm{PO}} = \sum_{(u_a,u_b,q)\in\mathcal{T}_U} q \sum_{i=1}^{\ell_U} L_{\mathrm{ent}}(u_a^{(i)}, u_b^{(i)}) + \sum_{(v_a,v_b,q)\in\mathcal{T}_V} q \sum_{i=\ell_U+1}^{\ell} L_{\mathrm{ent}}(v_a^{(i)}, v_b^{(i)}). \tag{6}$$

Table 1: Results on Amazon-book, Yelp, and Steam. Each base row reports for the CF models and the indented row reports those for LLM-aligned models. Best values in each column are bolded. Both BiOE variants achieve significantly improved performance as compared to the base HICF model without entailment losses and the single-entailment models. Results are averaged over 3 runs, standard deviations are largely below 0.005, and are omitted for presentation clarity and consistency with prior work; full results with standard deviation on Amazon-book are provided in Appendix B.

Model	**Amazon-book**				**Yelp**				**Steam**			
	R@5	R@20	N@5	N@20	R@5	R@20	N@5	N@20	R@5	R@20	N@5	N@20
LightGCN	0.0681	0.1566	0.0676	0.0976	0.0417	0.1165	0.0491	0.0740	0.0537	0.1355	0.0586	0.0864
+ LLM Alignment	0.0720	0.1640	0.0714	0.1027	0.0449	0.1213	0.0522	0.0774	0.0560	0.1405	0.0608	0.0896
HICF	0.0754	0.1675	0.0754	0.1060	0.0431	0.1212	0.0489	0.0750	0.0554	0.1431	0.0608	0.0907
+ LLM Alignment	0.0775	0.1717	0.0784	0.1094	0.0439	0.1223	0.0500	0.0758	0.0590	0.1481	0.0644	0.0946
HICF + $U \to I$ Ent	0.0764	0.1722	0.0770	0.1086	0.0448	0.1239	0.0516	0.0777	0.0572	0.1441	0.0633	0.0921
+ LLM Alignment	0.0809	0.1759	0.0804	0.1116	0.0469	0.1256	0.0537	0.0794	0.0600	0.1486	0.0660	0.0956
HICF + $I \to U$ Ent	0.0665	0.1465	0.0669	0.0933	0.0402	0.1205	0.0432	0.0704	0.0352	0.1085	0.0355	0.0621
+ LLM Alignment	0.0773	0.1704	0.0779	0.1086	0.0461	0.1256	0.0528	0.0788	0.0404	0.1218	0.0401	0.0696
BiOE$_{Bi}$	0.0810	0.1761	0.0825	0.1135	0.0467	0.1317	0.0548	0.0824	0.0603	0.1484	0.0636	0.0996
+ LLM Alignment	0.0833	0.1819	0.0852	0.1170	0.0491	0.1347	0.0570	0.0850	0.0651	0.1536	0.0696	0.1026
BiOE$_{PO}$	0.0841	0.1788	0.0828	0.1145	0.0491	0.1330	0.0559	0.0837	0.0631	0.1495	0.0651	0.1010
+ LLM Alignment	**0.0862**	**0.1849**	**0.0859**	**0.1183**	**0.0503**	**0.1366**	**0.0581**	**0.0870**	**0.0688**	**0.1551**	**0.0722**	**0.1047**

The full variant BiOE$_{\mathrm{PO}}$ adds this pseudo-order supervision together with the residual compatibility factor, using embedding space $\mathcal{X} = \mathbb{H}^{\mathcal{K},\ell,n} \times \mathcal{Z}_{\mathrm{res}}$ and score $s(u,v) = -(d^{\ell_1}_{\mathbb{H}}(u,v))^2 + \gamma s_{\mathrm{res}}(u,v)$. In our experiments, we choose $\mathcal{Z}_{\mathrm{res}} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|^2 = 1\}$ and use cosine similarity for s_{res}. The final objective for BiOE$_{\mathrm{PO}}$ is $L = L_{\mathrm{task}} + \lambda_{\mathrm{Bi}} L_{\mathrm{Bi}} + \lambda_{\mathrm{PO}} L_{\mathrm{PO}}$.

Theoretical Guarantees for BiOE. We now demonstrate that BiOE addresses the limitations in Propositions 4.1 and 4.2 to recover the desired order structure and improve retrieval. We provide intuitive discussion here and defer the full formal statements and proofs to Appendix A:

1. **Expressivity and Bias Cancellation.** Proposition A.3 proves that, by separating the two entailment directions into different hyperbolic factors, ***BiOE can represent both induced orders simultaneously*** without forcing comparable entities to share neighborhoods, resolving the expressivity restriction of Proposition 4.1, and the residual factor can capture compatibility not explained by either order alone. Moreover, the specialized $U \to V$ and $V \to U$ factors enable BiOE to learn a balanced representation where the bias vanishes when the two factors are symmetric, so BiOE learns to balance the two directions from data instead of inheriting a fixed bias.
2. **Retrieval Learning Guarantees.** Corollary A.4 shows that since BiOE recovers each single-entailment model exactly as a special case, its best achievable expected retrieval loss is never worse than that of either single-entailment model.

5 Experiments

We evaluate BiOE across recommender systems, protein-ligand binding, and protein function prediction. For each task, we compare the same backbone under no entailment, single-direction entailment, and our bi-order variants, isolating the effect of entailment using recent SOTA Euclidean or hyperbolic backbones. Across all tasks, BiOE improves over base models and single-entailment models, while mitigating the direction-dependent retrieval biases we show in single-entailment models. Our experiments further introduce several hyperbolic downstream pipelines, including the first Collaborative-Filtering-LLM alignment pipeline and hyperbolic protein function annotation model. Additional details are provided in Appendix B.

5.1 Application to Hyperbolic Recommender Systems

We test BiOE on recommendation tasks as user-item interactions are naturally many-to-many. We build on HICF (Yang et al., 2022a), a hyperbolic collaborative filtering model that processes user and item embeddings through tangent-space GNNs and uses the HAML and HINS losses as the task loss.

Experimental Setup and Baselines. We compare five HICF variants under the same backbone: (1) the base HICF model with only task loss; (2) HICF with $U \to V$ entailment; (3) HICF with

Table 2: Protein-ligand binding retrieval performance on the DUD-E and PCBA virtual screening benchmarks. We compare BiOE variants against Euclidean LigUnity, Hyperbolic LigUnity, single-direction entailment variants. HypSeek introduces additional model complexity and modality, and is shown for reference. Best results excluding HypSeek are shown in bold.

Model	DUD-E			PCBA		
	AUROC	BEDROC (80.5)	EF (1%)	AUROC	BEDROC (80.5)	EF (1%)
HypSeek	0.9435	0.7892	51.44	0.6210	0.1196	6.81
LigUnity (Euclidean)	0.8922	0.6526	41.63	0.5985	**0.1133**	6.47
Hyperbolic LigUnity	0.9248	0.7008	45.41	0.6044	0.0726	5.99
$L \to P$ entailment	0.9163	0.6812	43.12	0.5988	0.0806	5.91
$P \to L$ entailment	0.9172	0.6938	44.61	0.6089	0.0771	6.75
BiOE$_{Bi}$	**0.9278**	**0.7038**	**45.71**	0.6132	0.0916	8.46
BiOE$_{PO}$	0.9250	0.7011	45.66	**0.6188**	0.0966	**8.67**

$V \to U$ entailment; (4) BiOE$_{\mathrm{Bi}}$, which adds L_{Bi}; and (5) BiOE$_{\mathrm{PO}}$, which further adds pseudo-order entailment and the residual compatibility factor. We also include LightGCN (He et al., 2020), the Euclidean counterpart of HICF, where results are taken from FACE (Wang et al., 2025). For fairness, all models use the same number of parameters, and the single-entailment baselines and BiOE variants all use the same product of 2 hyperbolic factors. We follow the FACE setup and evaluate on Amazon-Book, Yelp, and Steam. Additional details are provided in Appendix Section B.

Hyperbolic CF–LLM Alignment. We further introduce, to our knowledge, the first method for aligning hyperbolic CF models with semantic LLM embeddings. We lift the FACE (Wang et al., 2025) pipeline to hyperbolic space by treating LLM token embeddings as tangent vectors, mapping them through exponential maps, and aligning them with hyperbolic CF embeddings using hyperbolic contrastive learning. The CF embeddings are disentangled and quantized with Möbius operations, while the residual factor in BiOE$_{\mathrm{PO}}$ remains Euclidean. We defer details to Appendix B and apply the same alignment pipeline to all HICF variants.

Performance Results. The results are shown in Table 1. Both BiOE $_{\mathrm{Bi}}$ and BiOE $_{\mathrm{PO}}$ outperform HICF and the single-entailment variants, with BiOE $_{\mathrm{PO}}$ achieving the best performance in all cases. Notably, single-entailment cones do not always improve over the base model, suggesting that the incorrect directional inductive bias can hurt performance. In contrast, bi-order entailment and LLM alignment further consistently improves performance. These results demonstrate both the effectiveness of BiOE and our hyperbolic CF–LLM alignment pipeline.

5.2 Application to Protein-Ligand Binding

We evaluate BiOE on the task of retrieving the most likely ligands for a protein binding pocket. This task is also many-to-many, where a pocket can bind multiple molecules, and a molecule may interact with multiple binding sites. We follow LigUnity (Feng et al., 2025), a Euclidean CLIP model, using the same datasets and training configuration. Ligands and pockets are encoded with pretrained Uni-Mol (Zhou et al., 2023) encoders, projected with an MLP into hyperbolic space, and trained with hyperbolic contrastive and entailment losses. We evaluate on DUD-E (Mysinger et al., 2012) and PCBA (Wu et al., 2018) benchmarks and measure AUROC, BEDROC, and EF at 1%.

Baselines and BiOE Variants. We consider the following hyperbolic baselines and BiOE variants: (1) Hyperbolic LigUnity; (2) Hyperbolic LigUnity with ligand-to-protein entailment; (3) Hyperbolic LigUnity with protein-to-ligand entailment; (4) BiOE$_{Bi}$; and (5) BiOE$_{PO}$. All hyperbolic baselines and BiOE variants use a product of 4 hyperbolic manifolds. HypSeek (Wang et al., 2026) is a recent work that also builds hyperbolic protein-ligand retrieval models. However, they introduce a third modality through a protein language model and use seven loss terms in total. To better isolate the effect of entailment losses and keep a fair comparison with LigUnity, we therefore use our Hyperbolic LigUnity backbone instead of building on HypSeek, and keep Hypseek as an additional reference.

Results. The results are shown in Table 2. Both BiOE variants improve over the hyperbolic baselines. The larger improvement is for the more difficult PCBA benchmark, where the BiOE variants also outperform HypSeek (Wang et al., 2026) significantly in EF(1%), despite using a simpler model. On PCBA, BiOE has lower BEDROC than LigUnity despite improving EF(1%). This gap is shared by all hyperbolic baselines rather than specific to BiOE, and BiOE$_{PO}$ still improves BEDROC by

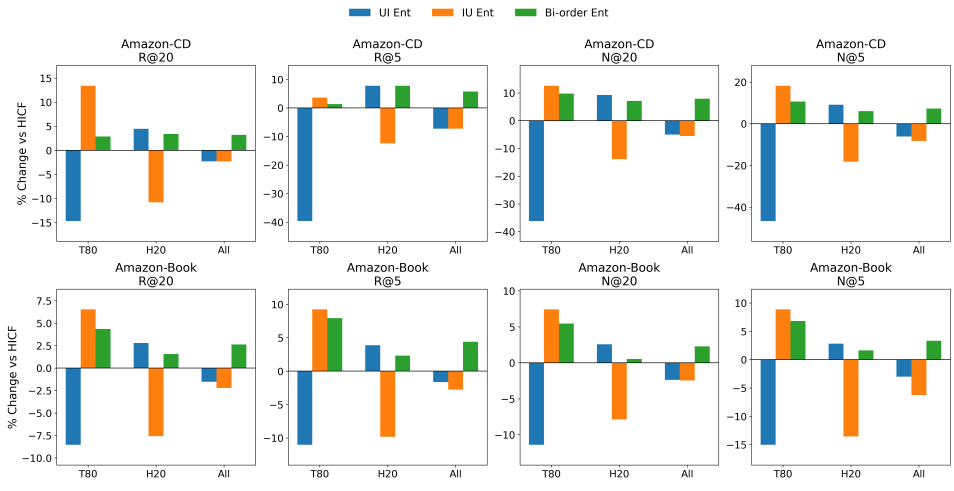


Figure 2: Popularity-based bias analysis. Items are split into tail items ($T80$) and head items ($H20$), and each bar reports the percent change relative to HICF without entailment. Single entailment models show clear biases, while bi-order entailment improves performance across both groups.

19.8% over the strongest hyperbolic baseline. Since EF(1%) emphasizes retrieving top candidates and BEDROC emphasizes ordering the full candidate set, this suggests that hyperbolic geometry is more consistent at retrieving top candidates than at ordering all candidates. Nevertheless, these results show that adding entailment objectives can improve protein-ligand retrieval.

5.3 Application to Protein Function Annotation

We further apply BiOE to protein function annotation, where hierarchical GO terms and protein annotations form a many-to-many relation between proteins and functional labels. We follow the MLP variant of StarGO (Akça et al., 2025), a recent SOTA model for GO term retrieval. The original model concatenates mean-pooled protein embeddings with GO embeddings and predicts logits with an MLP. We adapt this model to hyperbolic contrastive learning by lifting mean-pooled protein and GO embeddings into hyperbolic space, and transform them with two hyperbolic linear layers. We then use the resulting embeddings for contrastive learning and, where applicable, the entailment losses, before being projected to logits. We benchmark on the Molecular Function (MF) subontology. Full details of model architecture can be found in Appendix B.

Table 3: Protein function prediction results.

Model	Micro AUPR	Macro AUPR	AUC	Fmax
StarGO (MLP)	0.623±0.010	0.564±0.006	**0.995**±0.000	0.650±0.019
HyperGO	0.627±0.003	0.581±0.002	0.994±0.000	0.676±0.004
HyCoGO	0.641±0.006	0.600±0.006	0.993±0.001	0.683±0.009
$G \to P$ Ent.	0.599±0.011	0.523±0.008	0.994±0.000	0.622±0.009
$P \to G$ Ent.	0.589±0.005	0.467±0.006	**0.995**±0.000	0.625±0.012
BiOE_{Bi}	0.656±0.006	0.601±0.008	**0.995**±0.001	0.700±0.007
BiOE_{PO}	**0.671**±0.002	**0.637**±0.009	0.994±0.000	**0.719**±0.003

Baselines and BiOE Variants. We compare against several the following: (1) HyperGO, a hyperbolic adaptation of StarGO without contrastive learning; (2) HyCoGO, a hyperbolic contrastive adaptation of StarGO; (3) HyCoGO with protein-to-GO entailment; (4) HyCoGO with GO-to-protein entailment; (5) BiOE_{Bi}; and (6) BiOE_{PO}. HyperGO is included to measure the effect of adding contrastive learning. We also compare with StarGO-MLP to evaluate the benefit of moving from Euclidean to hyperbolic space. All hyperbolic baselines and BiOE use a product of 2 hyperbolic manifolds.

Results. The results are shown in Table 3. Both BiOE variants outperform all baselines, with BiOE_{PO} achieving the best overall performance. Compared to HyCoGO, BiOE_{PO} improves all metrics, showing that bi-order entailment and pseudo-order supervision are effective for protein-GO annotation. Single-entailment variants underperform HyCoGO, further demonstrating that a fixed entailment direction imposes an incorrect inductive bias. Additionally, HyCoGO and HyperGO improve over StarGO-MLP, showing that hyperbolic contrastive learning is beneficial for the task.

5.4 Additional Analysis, Ablation Studies, and Visualizations

Single-Entailment Biases. To examine single-entailment biases, we analyze model performance by item popularity. Following HICF (Yang et al., 2022a), we split items into head (top 20% most popular, denoted $H20$) and tail (remaining 80%, denoted $T80$) items. Results are in Figure 2, showing the percent change of each entailment model relative to the base HICF model. The user-to-item (UI) entailment model consistently improves performance on $H20$ items across metrics and datasets, but degrades performance on $T80$ items. The item-to-user (IU) entailment model shows the opposite pattern, where it improves performance on $T80$ items but harms performance on $H20$ items. These results support Proposition 4.2, showing that single-entailment cones introduce clear directional biases. On the other hand, the bi-order entailment model improves performance on both $H20$ and $T80$ items, indicating that it mitigates the popularity biases observed in single-entailment models.

Embedding Visualization. We visualize BiOE_{PO}'s learned embedding distributions for protein function annotation Figure 3, plotting the radius $r(x) = \|x_s\|$. Since hyperbolic entailment cones encode generality through radius, the cone-owning side should move closer to the origin in each directional factor. Indeed, in the $G \to P$ factor, GO terms concentrate near the origin while proteins sit farther away; in the $P \to G$ factor, the pattern is reversed. Thus, the two product factors specialize into opposite directional orders. However, in the product space, the embedding distributions are much more overlapped than prior methods that use single-entailment models (Desai et al., 2023; Pal et al., 2025). We further group GO and protein embeddings based on popularity. The head/tail split shows that popular proteins and GO terms generally occupy smaller-radius regions than their tail counterparts, indicating that BiOE_{PO} also captures meaningful intra-modal generality induced by the pseudo-order supervision.

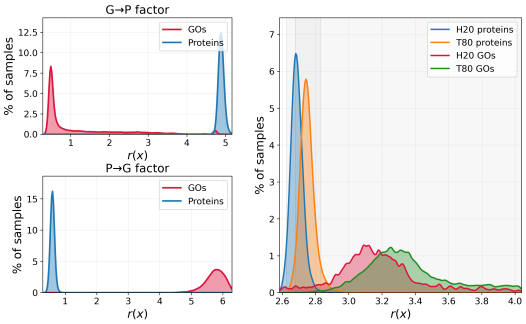


Figure 3: Embedding distributions of GO and Protein for BiOE_{PO}. Left: radii in the two bi-order factors. Right: head/tail radius distributions, where H20 denotes the top 20% most popular entities and T80 denotes the remaining 80%.

Ablation Studies. We perform an extensive ablation study of BiOE's components for the CF model with LLM alignment on Amazon-Book, shown in Table 4. For BiOE_{Bi}, removing L_{Bi} or keeping only one entailment direction consistently degrades performance, showing that the gains come from modeling both directional orders rather than adding a single entailment constraint or simply using a product metric. For BiOE_{PO}, the full model achieves the best results across all metrics. Removing the residual factor, cluster coverage, neighborhood coverage, or using only one direction of the pseudo order all hurts performance, demonstrating that each component contributes to the final model. Overall, our ablations show that each component of BiOE is vital for the final performance.

Table 4: Amazon-Book ablation after LLM alignment. Best results are bolded.

Model	R@5	R@20	N@5	N@20
BiOE_{Bi}	**0.0833**	**0.1819**	**0.0852**	**0.1170**
Only UI Ent	0.0798	0.1731	0.0791	0.1110
Only IU Ent	0.0728	0.1698	0.0732	0.1079
without L_{Bi}	0.0801	0.1739	0.0802	0.1112
BiOE_{PO}	**0.0862**	**0.1849**	**0.0859**	**0.1183**
without $\mathcal{Z}_{\text{res}}$	0.0826	0.1811	0.0837	0.1158
without cluster coverage	0.0824	0.1787	0.0827	0.1146
without neighborhood coverage	0.0847	0.1805	0.0849	0.1167
Only UI Pseudo Ent	0.0824	0.1784	0.0833	0.1146
Only IU Pseudo Ent	0.0782	0.1705	0.0781	0.1090

6 Conclusion

We proposed BiOE, a bi-order hyperbolic entailment framework for modeling many-to-many relationships. By decomposing a relationship into directional latent orders, BiOE avoids the one-sided inductive bias of single-entailment cones while retaining the hierarchical modeling benefits of hyperbolic geometry. We further introduced soft pseudo-order supervision and residual compatibility to handle noisy, incomplete, and non-hierarchical interactions. Across recommender systems, protein–ligand binding, and protein function annotation, BiOE consistently improves over hyperbolic baselines and single-entailment variants.

AI USE STATEMENT

In this work, we used generative AI tools for formatting the plots for better presentation, checking the correctness of mathematical proofs after the proofs were written, and code debugging. We have not used generative AI tools for the rest of the required disclosure tasks, such as generating ideas, interpreting result, designing methodology, implementing or executing methods, etc. Additionally, we used generative AI tools for polishing the writing (word choices and grammar checks) and searching for addition literature. We have reviewed all AI-assisted work, where we manually checked all figure generation code, check all mathematical proofs manually for regardless of whether a proof was also checked with AI, and checked any code debugged by an AI manually. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Full proofs and assumptions for all theoretical results are given in Appendix A. Appendix B provides all implementation, dataset, and hyperparameter details for all three experiments, including the pseudo-order construction (Section B.1), the hyperbolic CF–LLM alignment pipeline (Section B.2), and the hyperbolic protein function annotation model (Section B.3), with dataset preprocessing and hyperparameters given in Section B.4.1, Section B.4.2, and Section B.4.3.

REFERENCES

Mehmet Efe Akça, Gökçe Uludoğan, Arzucan Özgür, and İnci M. Baytaş. Star-go: Improving protein function prediction by learning to hierarchically integrate ontology-informed semantic embeddings. *arXiv preprint arXiv:2512.05245*, 2025.

Yushi Bai, Zhitao Ying, Hongyu Ren, and Jure Leskovec. Modeling heterogeneous hierarchies with relation-specific hyperbolic cones. *NeurIPS*, 34:12316–12327, 2021.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *ICLR*, 2024.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *NeurIPS*, pp. 4868–4879, 2019.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. Probing bert in hyperbolic spaces. *arXiv preprint arXiv:2104.03869*, 2021a.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. *arXiv preprint arXiv:2105.14686*, 2021b.

Sungjun Cho, Seunghyuk Cho, Sungwoo Park, Hankook Lee, Honglak Lee, and Moontae Lee. Curve your attention: Mixed-curvature transformers for graph representation learning. *arXiv preprint arXiv:2309.04082*, 2023.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pp. 7694–7731. PMLR, 2023.

Bhuwan Dhingra, Christopher J Shallue, Mohammad Norouzi, Andrew M Dai, and George E Dahl. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*, 2018.

Yuntao Du, Xinjun Zhu, Lu Chen, Baihua Zheng, and Yunjun Gao. Hakg: Hierarchy-aware knowledge gated network for recommendation. In *SIGIR*, 2022.

Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *CVPR*, pp. 7409–7419, 2022.

Bin Feng, Zijing Liu, Hao Li, Mingjun Yang, Junjie Zou, He Cao, Yu Li, Lei Zhang, and Sheng Wang. Hierarchical affinity landscape navigation through learning a shared pocket-ligand space. *Patterns*, 6(10):101371, 2025.

Samuele Fonio, Roberto Esposito, and Marco Aldinucci. Hyperbolic prototypical entailment cones for image classification. In *AISTATS*, 2025.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *NeurIPS*, pp. 5345–5355, 2018a.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, pp. 1646–1655. PMLR, 2018b.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. In *ICLR*, 2019.

Zahid Hasan, Masud Ahmed, and Nirmalya Roy. Lorentz entailment cone for semantic segmentation. In *WACV*. IEEE, 2026.

Neil He, Rishabh Anand, Hiren Madhu, Ali Maatouk, Smita Krishnaswamy, Leandros Tassiulas, Menglin Yang, and Rex Ying. Helm: Hyperbolic large language models via mixture-of-curvature experts. In *NeurIPS*, 2025a.

Neil He, Jiahong Liu, Buze Zhang, Ngoc Bui, Ali Maatouk, Menglin Yang, Irwin King, Melanie Weber, and Rex Ying. Position: Beyond euclidean–foundation models should embrace non-euclidean geometries. *arXiv preprint arXiv:2504.08896*, 2025b.

Neil He, Hiren Madhu, Ngoc Bui, Menglin Yang, and Rex Ying. Hyperbolic deep learning for foundation models: A survey. In *SIGKDD*, 2025c.

Neil He, Menglin Yang, and Rex Ying. Lorentzian residual neural networks. In *KDD*, 2025d.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, pp. 639–648, 2020.

Biao Hu, Zhen Huang, Minghao Hu, Pinglv Yang, Peng Qiao, Yong Dou, and Zhilin Wang. Partial order-centered hyperbolic representation learning for few-shot relation extraction. In *COLING*, 2025.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, pp. 6418–6428, 2020.

Wonjae Kim, Sanghyuk Chun, Taekyung Kim, Dongyoon Han, and Sangdoo Yun. Hype: Hyperbolic entailment filtering for underspecified images and texts. In *ECCV*, 2024.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

Jun Li, Jinpeng Wang, Chaolei Tan, Niu Lian, Long Chen, Yaowei Wang, Min Zhang, Shu-Tao Xia, and Bin Chen. Hlformer: Enhancing partially relevant video retrieval with hyperbolic learning. In *ICCV*, 2025.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *NeurIPS*, pp. 8230–8241, 2019.

Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *arXiv preprint arXiv:2305.06611*, 2023.

Michael M. Mysinger, Michael Carchia, John J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, pp. 6338–6347, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, pp. 3779–3788, 2018.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *ICLR*, 2025.

Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *TPAMI*, 2021.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366. Springer, 2011.

Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *ICLR*, 2020.

Jianing Sun, Zhaoyue Cheng, Saba Zuberi, Felipe Pérez, and Maksims Volkovs. HGCF: Hyperbolic graph convolution networks for collaborative filtering. In *WebConf*, pp. 593–601, 2021.

Albert Tseng, Tao Yu, Toni J. B. Liu, and Christopher De Sa. Coneheads: Hierarchy aware attention. In *NeurIPS*, 2023.

Max van Spengler, Artem Moskalev, Tommaso Mansi, Mangal Prakash, and Rui Liao. Hyperhelm: Hyperbolic hierarchy encoding for mrna language modeling. *arXiv preprint arXiv:2509.24655*, 2025.

Chao Wang, Yixin Song, Jinhui Ye, Chuan Qin, Dazhong Shen, Lingfeng Liu, Xiang Wang, and Yanyong Zhang. Face: A general framework for mapping collaborative filtering embeddings into llm tokens. In *NeurIPS*, 2025.

Jianhui Wang, Wenyu Zhu, Bowen Gao, Xin Hong, Ya-Qin Zhang, Wei-Ying Ma, and Yanyan Lan. Learning protein-ligand binding in hyperbolic space. *AAAI*, 2026.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, 2020.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Menglin Yang, Zhihao Li, Min Zhou, Jiahong Liu, and Irwin King. Hicf: Hyperbolic informative collaborative filtering. In *KDD*, pp. 2212–2221, 2022a.

Menglin Yang, Min Zhou, Zhihao Li, Jiahong Liu, Lujia Pan, Hui Xiong, and Irwin King. Hyperbolic graph neural networks: a review of methods and applications. *arXiv preprint arXiv:2202.13852*, 2022b.

Menglin Yang, Aosong Feng, Bo Xiong, Jihong Liu, Irwin King, and Rex Ying. Hyperbolic fine-tuning for large language models. *ICML LLM Cognition Workshop*, 2024a.

Menglin Yang, Jiahong Liu, Irwin King, and Rex Ying. Uhcone: Universal hyperbolic cone for implicit hierarchical learning. In *ICML 2024 Workshop GRaM*, 2024b.

Menglin Yang, Harshit Verma, Delvin Ce Zhang, Jiahong Liu, Irwin King, and Rex Ying. Hypformer: Exploring efficient transformer fully in hyperbolic space. In *KDD*, pp. 3770–3781, 2024c.

Daiki Yoshikawa and Takashi Matsubara. PHyclip: ℓ_1-product of hyperbolic factors unifies hierarchy and compositionality in vision-language representation learning. In *ICLR*, 2026.

Tao Yu, Toni J. B. Liu, Albert Tseng, and Christopher De Sa. Shadow cones: A generalized framework for partial order embeddings. In *ICLR*, 2024.

Yiding Zhang, Xiao Wang, Chuan Shi, Xunqiang Jiang, and Yanfang Fanny Ye. Hyperbolic graph attention network. *TBD*, 2021.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *ICLR*, 2023.

A Proof for Theoretical Results

A.1 Proof for Proposition 4.1

Proposition Statement.

Proposition A.1. *Let U, V be finite sets and let $\mathcal{R} \subseteq U \times V$ be a bipartite relationship.*

(i) There exist choices of $(U, V, \mathcal{R})$ such that no global scalar order score $g : U \cup V \to \mathbb{R}$ satisfies $(u, v) \in \mathcal{R} \iff g(u) \leq g(v)$.

(ii) Let f be a one-sided monotone order model. Suppose the same order is also required for the target-side latent preorder, in the sense that $v_a \preceq_V v_b \implies f(v_a) \leq_{\mathcal{X}} f(v_b)$. Then every comparable pair $v_a \preceq_V v_b$ must have identical neighborhoods, i.e. $N_U(v_a) = N_U(v_b)$.

Proof. For (i), let $U = \{u_1, u_2\}, V = \{v_1, v_2\}$ and let $\mathcal{R} = \{(u_1, v_1), (u_2, v_2)\}$. Then suppose such a g exists. Then by the assumption of g, we get $g(u_1) \leq g(v_1) < g(u_2) \leq g(v_2) < g(u_1)$, which is impossible, completing the proof.

For (ii), fix some $v_a \preceq_V v_b$. Then by definition, $N_U(v_b) \subseteq N_U(v_b)$. Then fix $u \in N_U(v_a)$ so that $(u, v_a) \in \mathcal{R}$. Then we have $f(u) \leq_{\mathcal{X}} f(v_a)$. Since $v_a \preceq_V v_b$, we have $f(u) \leq_{\mathcal{X}} f(v_a) \leq_{\mathcal{X}} f(v_b)$ Then by definition of f, we have $(u, v_b) \in \mathcal{R}$. Thus v_a, v_b have identical neighborhoods.

□

A.2 Proof for Proposition 4.2

Proposition Statement.

Proposition A.2. *Suppose similarity is given by $-d_{\mathbb{H}}$ and an ideal hyperbolic model. Assume that $r(v_H) \leq r(v_T)$ almost surely for $v_H \sim V_H$ and $v_T \sim V_T$.*

(i) If $u \in U$ are cone owners, then $F(r) = \bar{d}(r, V_T) - \bar{d}(r, V_H)$ is non-increasing in r.

(ii) If $v \in V$ are cone owners, assume the angular distribution of U is rotationally invariant conditional on radius. For $0 < \delta < r(v)$, let $T_\delta(v)$ denote the point obtained by moving v a distance δ along the geodesic toward $\mathbf{o}$, and define $\Delta(v, \delta) = \mathbb{E}_{u \sim U}[d_{\mathbb{H}}(u, v) - d_{\mathbb{H}}(u, T_\delta(v))]$. Then $\Delta(v, \delta)$ depends only on $r(v)$ and is non-decreasing in $r(v)$.

Proof. For simplicity we prove the result for curvature $\kappa = -1$. The general case then follows by scaling by a positive constant. Fix $x, y \in \mathbb{H}^{1,n}$ have radii $r = r(x)$ and $\rho = r(y)$. let the angle between their radial directions be θ, let $c = \cos\theta$ so that $d(r, \rho, c) := d_H(x, y) = \operatorname{arcosh}(\cosh r \cosh \rho - \sinh r \sinh \rho\, c)$. Now let $A(r, \rho, c) = \cosh r \cosh \rho - \sinh r \sinh \rho\, c$ so that $d(r, \rho, c) = \operatorname{arcosh} A(r, \rho, c)$. Since $c \in (-1, 1)$, $d(r, \rho, c)$ is smooth, and differentiation gives

$$\partial_r d = \frac{\sinh r \cosh \rho - \cosh r \sinh \rho\, c}{\sinh d}.$$

Differentiating again w.r.t. ρ, and using $A = \cosh d$, gives

$$\partial_\rho \partial_r d = \frac{(\partial_\rho \partial_r A)(A^2 - 1) - (\partial_r A)A(\partial_\rho A)}{(A^2 - 1)^{3/2}} = \frac{(c^2 - 1)\sinh r \sinh \rho}{\sinh^3 d} \leq 0.$$

Thus, $\partial_r d(r, \rho, c)$ is non-increasing in ρ for $c \in (-1, 1)$. So $0 \leq \rho_1 \leq \rho_2 \implies$ the map $r \mapsto d(r, \rho_2, c) - d(r, \rho_1, c)$ is non-increasing. The boundary cases $c = -1$ and $c = 1$ follow by continuity.

Now for (i), since u_r has rotationally invariant distribution, for any fixed v with radius $\rho = r(v)$, there is a random variable $C \in [-1, 1]$, independent of ρ, such that $\mathbb{E}_{u_r}[d_H(u_r, v)] = \mathbb{E}_C[d(r, \rho, C)]$. Let $h(r, \rho) = \mathbb{E}_C[d(r, \rho, C)]$. From the above, if $0 \leq \rho_1 \leq \rho_2$, then $r \mapsto h(r, \rho_2) - h(r, \rho_1)$ is non-increasing. Then since $r(v_H) \leq r(v_T)$ almost surely, $h(r, r(v_T)) - h(r, r(v_H))$ is non-increasing in r Then since $F(r) = \bar{d}(r, V_T) - \bar{d}(r, V_H) = \mathbb{E}[h(r, R_T) - h(r, R_H)]$ and taking expectation preserves monotonicity, so $F(r)$ is non-increasing.

Now for (ii), let $s = r(u)$. and since U has rotationally invariant distribution conditioned on radius, the conditional distribution of the angle between u and v depends only on s and not on the direction of v. Therefore, if $\rho = r(v)$, then $\Delta(v, \delta) = \mathbb{E}_s \mathbb{E}_C[d(s, \rho, C) - d(s, \rho - \delta, C)]$ depends only on $r(v)$. Now for fixed s and $c \in (-1, 1)$, differentiating $d(s, \rho, c)$ gives

$$\partial_\rho^2 d(s, \rho, c) = \frac{(c^2 - 1)(c \sinh s \sinh \rho - \cosh s \cosh \rho) \sinh^2 s}{\sinh^3 d(s, \rho, c)}.$$

Since $c \leq 1$, we have

$$c \sinh s \sinh \rho - \cosh s \cosh \rho \leq \sinh s \sinh \rho - \cosh s \cosh \rho = -\cosh(s - \rho) < 0.$$

Also $c^2 - 1 \leq 0$. Hence $\partial_\rho^2 d(s, \rho, c) \geq 0$, so $\rho \mapsto d(s, \rho, c)$ is convex. The boundary cases $c = -1$ and $c = 1$ follow directly from $d(s, \rho, -1) = s + \rho$ and $d(s, \rho, 1) = |s - \rho|$, both of which are convex in ρ. Therefore, for every fixed s and c, $\rho \mapsto d(s, \rho, c) - d(s, \rho - \delta, c)$ is non-decreasing for $\rho > \delta$, since we are increasing a convex function over an interval of length δ. Taking expectation over C and over $s = r(u)$ preserves this, completing the proof □

A.3 Proof for Proposition A.3

For a factor with entailment direction $U \to V$, we write $\mathbf{x} \leq_{U \to V} \mathbf{y}$ for the cone order induced by that factor, i.e. $\mathbf{x} \leq_{U \to V} \mathbf{y} \iff \mathbf{y}$ lies inside the entailment cone of $\mathbf{x}$ in that factor, and define $\leq_{V \to U}$ analogously for the $V \to U$ factor.

Proposition A.3. *Let U, V be finite sets and $\mathcal{R} \subseteq U \times V$ be a bipartite relationship with induced preorders $\preceq_U, \preceq_V$. Consider the embedding space $\mathcal{X} = \mathbb{H}^{n_U, \kappa_U} \times \mathbb{H}^{n_V, \kappa_V} \times \mathcal{Z}_{\text{res}}$, where the first hyperbolic factor models the $U \to V$ entailment direction and the second factor models the $V \to U$ entailment direction, and let $f(x) = (f_{U \to V}(x), f_{V \to U}(x), z_x)$ be an embedding of $U \cup V$ into $\mathcal{X}$.*

(i) For every $\mathcal{R}$, there exists an embedding f satisfying $(u, v) \in \mathcal{R} \implies f_{U \to V}(u) \leq_{U \to V} f_{U \to V}(v)$ and $f_{V \to U}(v) \leq_{V \to U} f_{V \to U}(u)$, as well as $u_a \preceq_U u_b \implies f_{U \to V}(u_a) \leq_{U \to V} f_{U \to V}(u_b)$ and $v_a \preceq_V v_b \implies f_{V \to U}(v_a) \leq_{V \to U} f_{V \to U}(v_b)$. These constraints are jointly feasible even when comparable entities have distinct neighborhoods. Moreover, if $\mathcal{Z}_{\text{res}} = \{z \in \mathbb{R}^m : \|z\|^2 = 1\}$ with $m = |U|$, then there exist residual embeddings z_x, a weight $\gamma > 0$, and a threshold $t \in \mathbb{R}$ such that $(u, v) \in \mathcal{R} \iff s(u, v) > t$, where $s(u, v) = -\left(\alpha_U d_{\mathbb{H}}(f_{U \to V}(u), f_{U \to V}(v)) + \alpha_V d_{\mathbb{H}}(f_{V \to U}(u), f_{V \to U}(v))\right)^2 + \gamma \langle z_u, z_v \rangle$.

(ii) Suppose the assumptions of Proposition 4.2 hold separately in the $U \to V$ and $V \to U$ factors, with similarity given by the unsquared weighted sum of the two factors' hyperbolic distances, and let $\bar{d}_U, \Delta_V$ denote the quantities $\bar{d}, \Delta$ of Proposition 4.2 computed using distances in the $U \to V$ and $V \to U$ factors respectively. Let $F_U(r) = \bar{d}_U(r, V_T) - \bar{d}_U(r, V_H)$ and, for inward displacements $\delta_U, \delta_V > 0$ representing the training-induced radial shift of the cone owner in each factor, define $B_U = F_U(r - \delta_U) - F_U(r)$ and $B_V = \mathbb{E}_{v_T \sim V_T}[\Delta_V(v_T, \delta_V)] - \mathbb{E}_{v_H \sim V_H}[\Delta_V(v_H, \delta_V)]$. Then $B_U, B_V \geq 0$, and the sign of $B_{\text{Bi}} = \alpha_U B_U - \alpha_V B_V$ determines whether BiOE is biased toward head or tail targets.

Proof. (i) We construct $f_{U \to V}$ explicitly and $f_{V \to U}$ follows from a symmetrical argument. Let G be a directed graph on $U \cup V$ with an edge $u \to u'$ whenever $u \preceq_U u'$ for $u, u' \in U$, and an edge $u \to v$ whenever $(u, v) \in \mathcal{R}$. Then contracting each equivalence class of U under $\preceq_U$ gives a strict partial order on the quotient, and since edges into V are sinks, G is a DAG. Let $\rho(z)$ be the length of the longest directed path in G ending at z. Then $\rho(u') \geq \rho(u)$ whenever $u \preceq_U u'$ and $\rho(v) > \rho(u)$ whenever $(u, v) \in \mathcal{R}$. Fix a ray from the origin in $\mathbb{H}^{n_U, \kappa_U}$, for r_0 at or above the minimum radius at which the half-aperture is defined, and embed every $z \in U \cup V$ at radius $r_0 + \rho(z)$ along this ray. Then $\text{ext}(\mathbf{x}, \mathbf{y}) = 0$ whenever $\mathbf{x}, \mathbf{y}$ share the same angular direction in the space-like dimension and $r(\mathbf{y}) \geq r(\mathbf{x})$, so we get $f_{U \to V}(z) \leq_{U \to V} f_{U \to V}(z')$ whenever $\rho(z) \leq \rho(z')$, giving both implications.

For the residual embeddings, enumerate $U = \{u_1, \ldots, u_m\}$. Assume WLOG every $v \in V$ has $N_U(v) \neq \emptyset$ (an isolated v needs one extra dimension). Set $z_{u_i} = e_i$ and $z_v = \tilde{z}_v / \|\tilde{z}_v\|$ where $\tilde{z}_v = \sum_{i:(u_i, v) \in \mathcal{R}} e_i$. Then $\langle z_{u_i}, z_v \rangle = 1/\sqrt{|N_U(v)|}$ if $(u_i, v) \in \mathcal{R}$ and 0 otherwise. Since U, V are finite, the squared distance term in s is bounded by some $D_{\max} < \infty$,

and $\delta_{\min} = \min_{v \in V} 1/\sqrt{|N_U(v)|} > 0$. Choose $\gamma > D_{\max}/\delta_{\min}$ and $t = 0$. Then $s(u, v) \geq -D_{\max} + \gamma\delta_{\min} > 0$ whenever $(u, v) \in \mathcal{R}$, and $s(u, v) \leq -D_{\max} \leq 0$ otherwise. Thus $(u, v) \in \mathcal{R} \iff s(u, v) > t$, as desired.

(ii) Applying Proposition 4.2(i) in the $U \to V$ factor, we get that $F_U(r)$ is non-increasing, so $B_U = F_U(r - \delta_U) - F_U(r) \geq 0$. Applying Proposition 4.2(ii) in the $V \to U$ factor, we get that $\Delta_V(v, \delta_V)$ is non-decreasing in $r(v)$. Then since $r(v_T) \geq r(v_H)$ almost surely, we have $B_V \geq 0$. Since the combined similarity is the unsquared weighted sum, the head-tail gap decomposes additively as $\alpha_U F_U + \alpha_V F_V$. Then the net change under the joint shift is $\alpha_U B_U - \alpha_V B_V = B_{\text{Bi}}$, and its sign gives the bias direction, as desired.

□

A.4 Proof for Corollary A.4

Corollary A.4. *Let V_H, V_T each contain a single element, and let $u \in U$ be a query whose ground-truth target lies in V_H with probability p and in V_T with probability $1 - p$. For a retrieval method that returns the element of V_H with probability q, consider the expected retrieval loss $L_p(q) = -p \log q - (1 - p) \log(1 - q)$. Let $q_U, q_V \in [0, 1]$ denote the probability that the $U \to V$ only and $V \to U$ only single-entailment models retrieve the element of V_H. Then $\min_{\alpha_U, \alpha_V} L_p(q_{\text{BiOE}}(\alpha_U, \alpha_V)) \leq \min(L_p(q_U), L_p(q_V))$ where α_U, α_V are weights as in Proposition A.3.*

Proof. Setting $\alpha_V = 0$ in the similarity score of Proposition A.3(i) removes the contribution of the $V \to U$ factor, so BiOE reduces exactly to the $U \to V$-only single-entailment model, achieving retrieval probability $q_{\text{BiOE}}(\alpha_U, 0) = q_U$; symmetrically, $\alpha_U = 0$ gives $q_{\text{BiOE}}(0, \alpha_V) = q_V$. Hence $q_U, q_V \in \{q_{\text{BiOE}}(\alpha_U, \alpha_V) : \alpha_U, \alpha_V \geq 0\}$, and therefore $\min_{\alpha_U, \alpha_V} L_p(q_{\text{BiOE}}(\alpha_U, \alpha_V)) \leq \min(L_p(q_U), L_p(q_V))$, completing the proof. □

B Implementation and Experimental Details

B.1 Soft Entailment Induced Pseudo Order

Real-world data is often noisy and incomplete, resulting in exact neighborhood inclusion being too brittle. As a result, we construct soft pseudo-orders from weighted neighborhood coverage. For each $v \in V$ and $u \in U$, let w_v, w_u be generality scores computed from empirical neighborhoods. Then for pairs $(u_1, u_2) \in U^2$ and $(v_1, v_2) \in V^2$, we define the coverage scores as

$$\eta_U(u_1, u_2) = \frac{\sum_{v \in N_V(u_1) \cap N_V(u_2)} w_v}{\sum_{v \in N_V(u_2)} w_v}, \qquad \eta_V(v_1, v_2) = \frac{\sum_{u \in N_U(v_1) \cap N_U(v_2)} w_u}{\sum_{u \in N_U(v_2)} w_u}. \tag{7}$$

Intuitively, $\eta_U(u_1, u_2)$ measures how much of u_2's neighborhood is covered by the neighborhood of u_1, weighted by the generality score of each $v \in N_V(u_2)$. Similarly, $\eta_V(v_1, v_2)$ measures the coverage of v_2's neighborhood by v_1's neighborhood. In implementation, denominators are stabilized with a small constant when needed.

To make the induced preorders more robust, we additionally lift neighborhoods to a soft semantic cluster space and use the resulting cluster-level coverage relation as additional pseudo-order supervision. Let Δ^{m-1} denote the $(m-1)$-dimensional probability simplex. For each $v \in V$, let $\pi_V(v) \in \Delta^{M_V - 1}$ be a soft cluster assignment over M_V clusters, and similarly let $\pi_U(u) \in \Delta^{M_U - 1}$ for each $u \in U$. These assignments may be obtained from metadata, pretrained semantic encoders, or learned side information, and allow each entity to belong to multiple latent groups with different strengths.

Using these soft assignments, we define the cluster-level neighborhood profiles of $u \in U$ and $v \in V$ as

$$g_U(u) = \sum_{v \in N_V(u)} a_{uv} \pi_V(v) \in \mathbb{R}^{M_V}_{\geq 0}, \qquad g_V(v) = \sum_{u \in N_U(v)} b_{uv} \pi_U(u) \in \mathbb{R}^{M_U}_{\geq 0}, \tag{8}$$

where $a_{uv} \geq 0$ and $b_{uv} \geq 0$ are nonnegative weights on the relation (u, v), which may be binary or may incorporate denoising or debiasing terms. Intuitively, $g_U(u)$ summarizes the distribution of

semantic target-side clusters covered by u, while $g_V(v)$ summarizes the distribution of semantic source-side clusters covered by v.

Coverage scores over clusters can then be computed from g_U and g_V. Let $\rho_k^V > 0$ and $\rho_r^U > 0$ be positive coefficients for target-side and source-side clusters, respectively. The cluster coverage scores are defined as

$$\eta_U^{\mathrm{cl}}(u_1, u_2) = \frac{\sum_{k=1}^{M_V} \rho_k^V \min(g_U(u_1)_k, g_U(u_2)_k)}{\sum_{k=1}^{M_V} \rho_k^V g_U(u_2)_k}, \eta_V^{\mathrm{cl}}(v_1, v_2) = \frac{\sum_{r=1}^{M_U} \rho_r^U \min(g_V(v_1)_r, g_V(v_2)_r)}{\sum_{r=1}^{M_U} \rho_r^U g_V(v_2)_r}. \tag{9}$$

These scores measure how much of the semantic neighborhood profile of the second entity is covered by that of the first. The final coverage scores combine exact neighborhood coverage and cluster-level coverage:

$$\bar{\eta}_U(u_1, u_2) = (1 - \lambda_{\mathrm{cl}})\eta_U(u_1, u_2) + \lambda_{\mathrm{cl}}\eta_U^{\mathrm{cl}}(u_1, u_2), \tag{10}$$

$$\bar{\eta}_V(v_1, v_2) = (1 - \lambda_{\mathrm{cl}})\eta_V(v_1, v_2) + \lambda_{\mathrm{cl}}\eta_V^{\mathrm{cl}}(v_1, v_2), \tag{11}$$

where $\lambda_{\mathrm{cl}} \in [0, 1]$ controls the strength of the soft-cluster signal.

The directional pseudo-order supervision is obtained by antisymmetrizing the coverage scores:

$$q_U(u_1, u_2) = \bar{\eta}_U(u_1, u_2) - \bar{\eta}_U(u_2, u_1), \qquad q_V(v_1, v_2) = \bar{\eta}_V(v_1, v_2) - \bar{\eta}_V(v_2, v_1). \tag{12}$$

Thus, $q_U(u_1, u_2) > 0$ suggests that u_1 covers more of the neighborhood of u_2 than vice versa, and can therefore be treated as more general than u_2 on the source side. Similarly, $q_V(v_1, v_2) > 0$ suggests that v_1 is more general than v_2 on the target side. The magnitude of q_U and q_V serves as a confidence score for the pseudo-order relation.

We then form pseudo-order sets

$$\mathcal{T}_U = \{(u_a, u_b, q) : q = q_U(u_a, u_b) > \tau_U\}, \qquad \mathcal{T}_V = \{(v_a, v_b, q) : q = q_V(v_a, v_b) > \tau_V\}, \tag{13}$$

where $\tau_U, \tau_V > 0$ are confidence thresholds. These sets are used to impose confidence-weighted entailment supervision within the corresponding source-side and target-side order factors.

B.2 Alignment of Hyperbolic CF Models to LLM Semantic Embeddings

We adapt FACE (Wang et al., 2025) to align hyperbolic collaborative filtering representations with semantic representations produced by a frozen language model encoder. Unlike the original Euclidean formulation, we perform the mapping, quantization, reconstruction, and alignment procedure through hyperbolic geometry. Let $\mathcal{E} = U \cup V$ denote the set of users and items, and let $h_e = (h_e^{(1)}, \ldots, h_e^{(\ell)}) \in \prod_{i=1}^{\ell} \mathbb{H}^{n_i, \kappa_i}$ be the product hyperbolic CF embedding of entity $e \in \mathcal{E}$. Since LLM token embeddings are Euclidean vectors, we first use the tangent space at the origin as an interface between hyperbolic CF representations and the language-model token space. Specifically, we define

$$z_e = \left[\log_{o_1}^{\kappa_1}(h_e^{(1)})_s \| \cdots \| \log_{o_\ell}^{\kappa_\ell}(h_e^{(\ell)})_s\right] \in \mathbb{R}^d,$$

where $(\cdot)_s$ denotes the spatial coordinates after applying the logarithmic map. We then decompose z_e into M latent concept vectors $Z_e = (z_{e,1}, \ldots, z_{e,M}) \in \mathbb{R}^{M \times d_w}$ using a learned linear mapping followed by a shallow Transformer encoder. Each concept vector is then lifted to the Lorentz hyperboloid by $\tilde{z}_{e,m} = \exp_o^{\kappa}([0, z_{e,m}])$. We construct a discrete semantic codebook from the frozen LLM vocabulary. For each candidate token t with LLM embedding a_t, we learn a linear projection W_c into the concept dimension and lift the projected token embedding to hyperbolic space: $c_t = \exp_o^{\kappa}([0, W_c a_t])$. The m-th concept vector is quantized by nearest-neighbor search as $q_{e,m} = \arg\min_t d_{\mathbb{H}}(\tilde{z}_{e,m}, c_t)$. Once a token is selected for one concept position, it is masked from the remaining positions for the same entity. This encourages the resulting token set to cover diverse semantic attributes rather than repeatedly selecting the same descriptor.

To make the quantization differentiable, we use a hyperbolic straight-through estimator. Let $c_{q_{e,m}}$ denote the selected codebook point. We compute the stopped hyperbolic displacement from $\tilde{z}_{e,m}$ to $c_{q_{e,m}}$ using Möbius subtraction and add it back to $\tilde{z}_{e,m}$:

$$\tilde{q}^{\mathrm{st}}_{e,m} = \tilde{z}_{e,m} \oplus_\kappa \mathrm{sg}\big(c_{q_{e,m}} \ominus_\kappa \tilde{z}_{e,m}\big)\,,$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The quantized representation used by the decoder is obtained by mapping $\tilde{q}^{\mathrm{st}}_{e,m}$ back to the tangent space at the origin. The vector-quantization loss is

$$\mathcal{L}_{\mathrm{vq}} = \frac{1}{|\mathcal{B}|M} \sum_{e\in\mathcal{B}} \sum_{m=1}^{M} \left[\beta_1 d^2_{\mathbb{H}}(c_{q_{e,m}}, \mathrm{sg}(\tilde{z}_{e,m})) + \beta_2 d^2_{\mathbb{H}}(\tilde{z}_{e,m}, \mathrm{sg}(c_{q_{e,m}}))\right],$$

where the first term updates the codebook projection and the second term is the commitment loss. We note that empirically, we found the ***fully hyperbolic quantization to be crucial for the model's success***. In fact, performing the quantization in the tangent space using the previous tangent vectors without lifting them back to hyperbolic space, while using Euclidean quantization methods, lead to model collapse and results in a functioning CF model to output random predictions.

We further require the discrete token bottleneck to preserve the original hyperbolic CF geometry. The quantized concepts are passed through a Transformer decoder and a reverse linear mapping to produce $\hat{z}_e \in \mathbb{R}^d$. We split $\hat{z}_e$ according to the product factors and lift each part back to the corresponding Lorentz factor, giving $\hat{h}^{(i)}_e = \exp^{\kappa_i}_{o_i}([0, \hat{z}^{(i)}_e])$. The reconstruction loss is

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\mathcal{B}|\ell} \sum_{e\in\mathcal{B}} \sum_{i=1}^{\ell} d^2_{\mathbb{H}}(\hat{h}^{(i)}_e, h^{(i)}_e).$$

Finally, the quantized concept tokens are used to construct a semantic prompt for each entity, see FACE (Wang et al., 2025) for full details of the prompts. The selected token embeddings are inserted after the prompt and encoded by the frozen LLM encoder, producing a collaborative semantic representation $\hat{s}_e$. Let s_e denote the semantic representation of the same entity obtained directly from its text-side profile using the same frozen encoder. We lift both representations to hyperbolic space and align them with a contrastive objective:

$$\mathcal{L}_{\mathrm{align}} = -\frac{1}{|\mathcal{B}|} \sum_{e\in\mathcal{B}} \log \frac{\exp(-d_H(\hat{s}_e, s_e)/\tau)}{\sum_{e'\in\mathcal{B}} \exp(-d_H(\hat{s}_e, s_{e'})/\tau)},$$

where $\mathrm{sim}_{\mathbb{H}}$ is implemented as negative hyperbolic distance or an equivalent Lorentz inner-product based similarity. This objective encourages the tokens decoded from collaborative behavior to recover the semantic description of the same entity while separating them from other entities in the batch.

The full alignment objective is $\mathcal{L}_{\mathrm{FACE-H}} = \lambda_{\mathrm{vq}}\mathcal{L}_{\mathrm{vq}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}$. We train the alignment module in two stages, following FACE. In the mapping stage, the model learns the hyperbolic token bottleneck using the vector-quantization and reconstruction losses. In the alignment stage, we additionally generate collaborative semantic representations through the frozen LLM encoder and optimize the hyperbolic contrastive alignment loss. The same alignment pipeline is applied to all hyperbolic CF variants, allowing us to isolate the effect of CF representation geometry while keeping the LLM semantic alignment procedure fixed.

B.3 Hyperbolic Contrastive Protein Function Annotation Models

We adapt the MLP variant of StarGO (Akça et al., 2025) to a product hyperbolic contrastive model. Let $\mathcal{P}$ be the set of proteins, $\mathcal{G}$ be the set of GO terms. For each protein p, we are given per-residue embeddings $\mathbf{H}_p \in \mathbb{R}^{L_p \times d_p}$ and first compute a masked mean representation $\bar{\mathbf{h}}_p$. For each GO term g, we use its pretrained embedding $\mathbf{a}_g \in \mathbb{R}^{d_g}$. Both protein and GO embeddings are lifted into a product of two Lorentz factors. The first factor is assigned to the $G \to P$ order and the second factor is assigned to the $P \to G$ order. For $i \in \{1, 2\}$, $\mathbf{g}^{(i)}_g = \mathrm{MLP}^{\mathbb{H}}_{G,i}\left(\Pi_i(\mathbf{a}_g)\right), \mathbf{p}^{(i)}_p = \mathrm{MLP}^{\mathbb{H}}_{P,i}\left(\Pi_i(\bar{\mathbf{h}}_p)\right)$, where $\Pi_i(\mathbf{x}) = [\sqrt{c_i + \|\mathbf{x}\|^2_2}, \mathbf{x}]^T$ denotes projection to Lorentz space and $\mathrm{MLP}^{\mathbb{H}}$ denotes the hyperbolic MLP used in the projection heads (Yang et al., 2024c). The classifier follows the MLP structure of StarGO, but replaces the Euclidean MLP with a hyperbolic MLR (Bdeir et al., 2024). We

concatenate the space dimension of the GO and protein factors, lift the concatenated vector to the Lorentz space, and apply a hyperbolic MLP followed by a Lorentz MLR layer:

$$\ell_{pg} = \mathrm{MLR}^{\mathbb{H}}\left(\mathrm{MLP}^{\mathbb{H}}\left(\Pi\left([\mathbf{g}_{g,s}^{(1)} \| \mathbf{g}_{g,s}^{(2)} \| \mathbf{p}_{p,s}^{(1)} \| \mathbf{p}_{p,s}^{(2)}]\right)\right)\right). \tag{14}$$

The task loss L_{BCE} is binary cross entropy over all proteins in the batch and all GO terms.

For $\mathrm{BiOE}_{\mathrm{PO}}$, we additionally use a residual compatibility factor. Protein and GO embeddings are passed through separate two-layer Euclidean MLPs and normalized to the unit sphere, giving residual embeddings $\mathbf{r}_p^P$ and $\mathbf{r}_g^G$. If a residual factor is used, the logits are obtained via

$$\ell_{pg} = \mathrm{MLR}^{\mathbb{H}}\left(\mathrm{MLP}^{\mathbb{H}}\left(\Pi\left([\mathbf{g}_{g,s}^{(1)} \| \mathbf{g}_{g,s}^{(2)} \| \mathbf{p}_{p,s}^{(1)} \| \mathbf{p}_{p,s}^{(2)} \| \mathbf{r}_p^P \| \mathbf{r}_g^G]\right)\right)\right). \tag{15}$$

For contrastive learning, the score is computed using the hidden embeddings $\mathbf{g}_g, \mathbf{p}, \mathbf{r}_p, \mathbf{r}_g$. The bi-order entailment loss is applied only to positive protein–GO annotations. The entailment loss is computed on the hideen embeddings as well:

$$L_{\mathrm{Bi}} = \frac{1}{|\mathcal{R}_{\mathcal{B}}|} \sum_{(p,g)\in\mathcal{R}_{\mathcal{B}}} \frac{1}{2}\left[L_{\mathrm{ent}}(\mathbf{g}_g^{(1)}, \mathbf{p}_p^{(1)}) + L_{\mathrm{ent}}(\mathbf{p}_p^{(2)}, \mathbf{g}_g^{(2)})\right]. \tag{16}$$

The single-entailment baselines are obtained by using the same entailment direction in both factors. For $\mathrm{BiOE}_{\mathrm{PO}}$, we further impose pseudo-order entailment within each modality. Protein–protein pseudo-orders are computed from annotation-neighborhood coverage between rows of Y, while GO–GO pseudo-orders are computed from coverage between columns of Y, where $Y \in \{0,1\}^{|\mathcal{P}|\times|\mathcal{G}|}$ is the observed annotation matrix. The L_{PO} loss is computed on the hidden embeddings as well.

B.4 Experimental Details

B.4.1 Experiments on Recommender Systems

Here we discuss in detail our experimental setups. Our experiments were conduct on a cluster with 10 NVIDIA RTX A6000 GPUs, and at most 3 were used at a time. Total training time took several weeks across all experiments and models.

Dataset Details. We follow the FACE (Wang et al., 2025) setup and evaluate on three public recommendation datasets: Amazon-book, Yelp, and Steam. Amazon-book is derived from the Amazon Product Review corpus and contains user–book interactions together with review text and product descriptions. Yelp contains user–business interactions with categorical attributes and consumer feedback. Steam contains user–game engagements and gameplay reviews. Following FACE, each dataset is split into training, validation, and test sets with a $3:1:1$ ratio. We evaluate recommendation performance using an all-ranking protocol, where all non-interacted items are treated as candidate items for each user.

Table 5: Statistics of the recommender system datasets.

Dataset	# Users	# Items	# Interactions	Density
Amazon-book	11,000	9,332	120,464	1.2×10^{-3}
Yelp	11,091	11,010	166,620	1.4×10^{-3}
Steam	23,310	5,237	316,190	2.6×10^{-3}

Implementation Details. We closely follow the experimental setup for FACE (Wang et al., 2025). For HICF, we use Lorentz hyperbolic embeddings with curvature $c = 1.0$, i.e. $\kappa = -1$, which we fix across all factors and all three experiments to isolate the effect of how entailment directions specialize each hyperbolic factor from the effect of curvature itself. For product-manifold variants, the total hyperbolic dimension 256 is split across 2 Lorentz factors. We use the HICF task loss, where for each positive pair we sample 12 candidate negatives and select the hardest negative according to the current hyperbolic representation. For all hyperbolic models, we use margins $m = 1.0$ for Amazon-book and Yelp, and $m = 2.0$ for Steam. For the entailment models, we use an entailment warmup of 300 optimization steps, where we found that without warmup steps the entailment models

might face early collapse. On Amazon-book and Yelp, we use $\lambda_{ent} = 0.2$ for the single entailment models and $\lambda_{ent} = 0.4$ for BiOE variants, as we found using $\lambda_{ent} = 0.4$ also lead to model collapse for single entailment models. For Steam, we halve the entailment weights for all entailment models. For BiOE_{Bi}, we use $\lambda_{\text{Bi}} = 0.1$.

For BiOE_{PO}, we additionally use soft pseudo-order supervision and the residual compatibility factor. Before training, we precompute user–user and item–item pseudo-order pairs from the training interaction matrix. We construct both exact neighborhood coverage and semantic-cluster coverage, where semantic clusters are obtained by applying KMeans to MiniLM (Wang et al., 2020) embeedings of user and item descriptions. We use empirical-CDF inverse degree weights to reduce the dominance of highly popular entities, combine exact and cluster coverage with weight $\lambda_{\text{cl}} = 0.7$, keep pseudo-orders with confidence greater than 0.1, and retain the top 64 pseudo-order pairs per anchor. During training, we sample pseudo-order pairs in each minibatch. The residual compatibility factor is implemented small LightGCN model and normalized to the unit sphere. We use $\gamma = 0.3$ and the total embedding dimensions are also split $[0.7, 0.3]$ among the hyperbolic and Euclidean components.

For LLM-alignment, we use the exact same $\lambda_{vq}, \lambda_{rec}, \lambda align$ as FACE.

The mdoels are trained with the largely the same setup as FACE as well. All models are trained with Adam using learning rate 1e-3 and batch size 2048 early stopping. We use embedding dimension 256, regularization weight 1e-6, and 3 graph layers.

Results with Variance. Table 1 reports results averaged over 3 runs without standard deviation for presentation clarity, following prior work. Table 6 reports the full results with standard deviation on Amazon-book for all LLM-aligned HICF variants; all improvements of BiOE over the strongest baseline are statistically significant.

Table 6: Amazon-book results with standard deviation, all models LLM-aligned. Best values in each column are bolded.

Model	R@5	R@20	N@5	N@20
HICF	0.0775±0.0013	0.1717±0.0031	0.0784±0.0029	0.1094±0.0030
$U \to I$ Ent.	0.0809±0.0011	0.1759±0.0033	0.0804±0.0012	0.1116±0.0037
$I \to U$ Ent.	0.0773±0.0023	0.1704±0.0049	0.0779±0.0017	0.1086±0.0048
BiOE_{Bi}	0.0833±0.0017	0.1819±0.0031	0.0852±0.0014	0.1170±0.0034
BiOE_{PO}	**0.0862**±0.0015	**0.1849**±0.0041	**0.0859**±0.0016	**0.1183**±0.0029

B.4.2 Experiments on Protein Ligand Binding

Dataset Details. We follow the training data construction of LigUnity. The initial pre-training data are collected from ChEMBL and BindingDB, covering 2,196 unique proteins, 26,748 assays, 428,767 unique ligands, and 36,662 unique pockets. Each assay contains, on average, 30.29 ligands and 10.96 pockets from PDB. To further improve protein and pocket coverage, the training data are augmented with protein-ligand complexes from PDBBind (Wang et al., 2004), which contribute 3,889 unique proteins. After curation, the final pre-training data contain 4,847 unique proteins and 53,406 unique pockets.

Evaluation was performed on DUD-E (Mysinger et al., 2012) and LIT-PCBA (Wu et al., 2018). DUD-E contains 22,886 active compounds with binding affinities across 102 protein targets (about 224 ligands per target), and for each active ligand provides around 50 decoys that match physicochemical properties but differ in 2D topology. The dataset is mainly used to test how well docking methods can distinguish true binders from non-binders. The LIT-PCBA dataset is constructed from dose-response PubChem bioassays to mitigate the target and decoy selection biases found in other benchmarks. It comprises 15 protein targets with 7,844 experimentally confirmed actives and 407,381 inactive, which reflects realistic hit rates in high-throughput screening.

Implementation Details. We closely follow the setup for LigUnity. In particular, we train on 2 GPUs with batch size 24, learning rate 1e-4 with 6% warmup for a total of 50 epochs, trained with the AdamW optimizer. The encoder setup is identical to LigUnity as well. As in our other experiments,

we fix curvature $\kappa = -1$ for all hyperbolic factors. We use $\lambda_{Bi} = 0.5$ and $\lambda_{PO} = 0.25$ for BiOE, and for single entailment model we set the weight for entailment as 0.5.

B.4.3 Experiments on Protein Function Annotation

Dataset Details.

For the protein function annotation experiment, we used the Molecular Function (MF) subset of the STAR-GO/PFresGO benchmark. This benchmark is derived from the DeepFRI protein function prediction dataset and contains 36,641 protein sequences annotated with 2,752 Gene Ontology (GO) terms across the three GO subontologies. We restrict our experiment to the MF subontology, which contains 489 GO terms. Each GO term in the benchmark is associated with at least 50 non-redundant Protein Data Bank (PDB) chains. The dataset follows the standard STAR-GO split, with 29,902 training proteins, 3,323 validation proteins, and 3,416 test proteins. The test set contains proteins with experimentally validated annotations, and protein sequences are limited to at most 1,000 residues. To reduce train-test redundancy, the original benchmark applies CD-HIT at a 95% sequence-identity threshold. Following STAR-GO, annotations are evaluated under the Gene Ontology true path rule using the June 2020 go-basic release.

Implementation Details. Our hyperparameters were not tuned extensively as we found that BiOE was able to achieve strong results with default settings, where our training setting largely follows StarGO (Akça et al., 2025). All hyperbolic GO models use a product of 2 Lorentz factors with curvature parameter $c = 1.0$, i.e. $\kappa = -1$, fixed as in our other experiments to isolate the effect of the entailment directions from the effect of curvature. The GO embedding dimension and the protein embedding dimension is 128. The hyperbolic MLP projection to MLR uses hidden dimensions [512,256,128]. The total hyperbolic projection dimension is 128, split evenly across product factors. We use GeLU for activation and 0.1 for dropout. For BiOE_{PO}, the residual compatibility factor uses a two-layer Euclidean MLP with hidden dimension 256 and output dimension 64, followed by ℓ_2 normalization. We train with AdamW using learning rate 3e-4, weight decay 0.01, batch size 32, and maximum number of 100 epochs. We use a cosine learning-rate schedule with warmup ratio 0.05 and minimum learning-rate ratio 3e-5. The contrastive temperature is $\tau = 1.0$. The loss weights chosen so that the loses are of similar magnitude. As the BCE loss is at a much smaller scale than both the contrastive and entailment losses, the weights are chosen as $\lambda_{\text{CL}} = 0.03$, $\lambda_{\text{Bi}} = 0.03$, and $\lambda_{\text{PO}} = 0.03$. For pseudo-order construction, we use confidence threshold 0.1 and keep the top 64 pseudo-order pairs per anchor. The residual compatibility weight is $\gamma = 1.0$.

Training Dynamics. Table 7 reports the training loss progression of BiOE_{PO} on protein function annotation, where L_{cl} is the contrastive loss introduced in HyCoGO. The BiOE losses $L_{\text{Bi}}, L_{\text{PO}}$ decrease steadily alongside the task losses $L_{\text{task}}, L_{\text{cl}}$ throughout training, indicating that the BiOE losses supplement rather than compete with the task objective.

Table 7: Training loss progression for BiOE_{PO} on protein function annotation.

	L_{task}	L_{cl}	L_{Bi}	L_{PO}
epoch 1	0.00286	2.615	2.075	1.284
epoch 20	0.00071	0.738	0.597	0.382
epoch 50	0.00016	0.139	0.394	0.297
epoch 75	0.00011	0.089	0.267	0.228
epoch 100	0.00008	0.037	0.158	0.128