

HiCo-EL: Encoding Ontology Structure into Bi-Encoder Entity Linking

Albert Lee Prasanth Bathala Puru Malhotra*

Christophe Ye Cassie S. Mitchell

Georgia Institute of Technology

{albert.jb.lee, pbathala3, cye73}@gatech.edu,
purumalhotra99@gmail.com, cassie.mitchell@bme.gatech.edu

Abstract

Bi-encoder architectures for biomedical entity linking learn flat embedding spaces that ignore ontology structure. HiCo-EL encodes hierarchical knowledge into bi-encoder representations through hierarchical contrastive pretraining. The core idea is *within-space* hierarchical multi-label contrastive learning (HMLC): contrastive losses applied independently within each encoder space, preserving cross-space compatibility for retrieval. Our formulation drops the monotonicity constraints of prior HMLC work, handles variable-depth hierarchy paths, and adds depth-weighted alignment between encoder spaces. Treating each entity as a leaf node unifies coarse type structure with fine-grained entity discrimination. On Med-Mentions and NCBI Disease, HiCo-EL outperforms ArboEL by +3.7% and +1.8% Recall@1 respectively. The gains reveal a key insight: *breadth amplifies, depth generalizes*—broad hierarchies help common entities most (+6.3% R@1), while deep hierarchies benefit unseen entities (+9.6% R@1 zero-shot).

1 Introduction

Mapping entity mentions in biomedical text to canonical knowledge base entries is central to clinical NLP (Bodenreider, 2004). Applications spanning clinical decision support, pharmacovigilance, and literature mining depend on accurate entity resolution (Sung et al., 2020).

Bi-encoder architectures now dominate this space (Wu et al., 2020; Humeau et al., 2020), encoding mentions and entities independently to enable sub-linear retrieval via approximate nearest neighbor search. Recent advances have pushed accuracy further through improved training objectives and inference procedures (Agarwal et al., 2022; Liu et al., 2021).

A limitation persists: the embedding space has no structure beyond what emerges from the linking objective. All entities compete equally regardless of semantic relationships. Yet biomedical knowledge bases are inherently hierarchical. UMLS organizes concepts into 127 semantic types in a hierarchy of maximum depth 8 (Figure 1). “Aspirin” falls under Pharmacologic Substance, not Anatomical Structure—the embedding geometry should reflect this.

A natural tool exists: hierarchical multi-label contrastive learning (HMLC; Zhang et al., 2022b) organizes an embedding space around a tree of labels. But HMLC was designed for a single embedding space, while bi-encoder retrieval depends on two separate spaces whose dot products remain meaningful. Imposing hierarchy across the two spaces conflates type structure with mention-entity matching; imposing it within each space alone leaves nothing to hold the two spaces together. The central question this paper answers is how to adapt hierarchical contrastive learning to the dual-encoder retrieval setting.

HiCo-EL’s answer is *within-space* hierarchical contrastive pretraining: we jointly train both encoders with two complementary losses—the HMLC loss organizes embeddings by type hierarchy within each encoder space, while an alignment loss maintains compatibility between the spaces. Without alignment, the encoders drift into incompatible representations—retrieval accuracy falls below the baseline (Table 5). The pre-trained representations provide structured initialization for downstream linking; we demonstrate this with MST-based clustering (Agarwal et al., 2022), though the approach is model-agnostic.

An important design choice: we extend the type hierarchy by appending each CUI as a leaf node. This allows HMLC to enforce type coherence at upper levels and entity discrimination at the deepest level, unifying coarse-to-fine learning in a sin-

* Work done while at Georgia Institute of Technology; now at Apple.

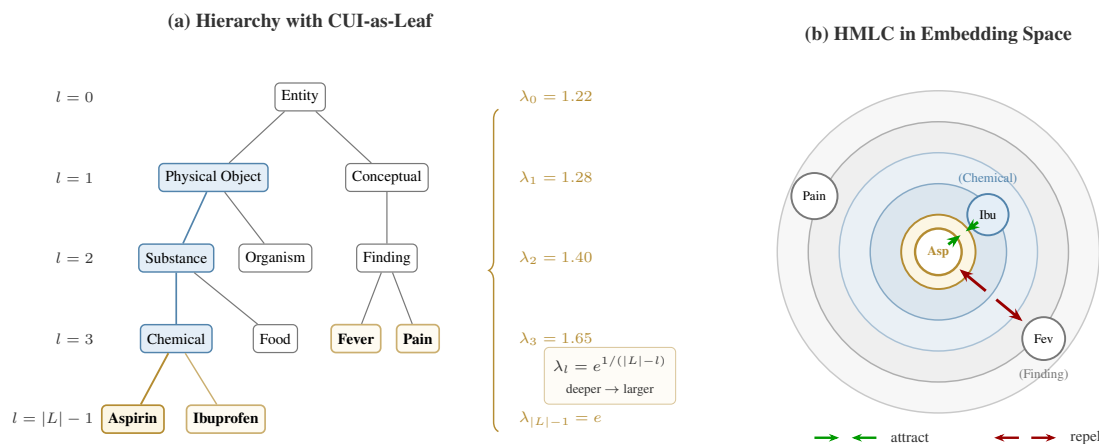


Figure 1: **Unified ontology-to-entity learning** via CUI-as-leaf. **(a)** The type hierarchy is extended by appending each entity (e.g., Aspirin) as a leaf node. Depth-dependent weights $\lambda_l = e^{1/(|L|-l)}$ penalize violations more heavily at deeper levels, prioritizing fine-grained entity discrimination. **(b)** HMLC shapes embedding geometry: entities sharing the same type path (Aspirin–Ibuprofen, both Chemicals) are pulled together, while entities from different branches (Aspirin–Fever) are pushed apart.

gle loss.

Our contributions center on this adaptation:

- **Hierarchy within, alignment across:** We decouple hierarchical structuring from cross-space retrieval. HMLC (Zhang et al., 2022b) is applied *independently within* each encoder space to organize embeddings by type, while a separate alignment loss couples the spaces *across* encoders for retrieval. This avoids conflating type structure with mention-entity matching.
- **Unified ontology-to-entity learning:** By appending each CUI as a leaf node, HMLC enforces type coherence at coarse levels and entity discrimination at the finest level—all within a single hierarchical loss. No separate type classifier or entity-level objective is needed.
- **Dual-space coupling for retrieval:** Without explicit alignment, independently-trained encoders drift into incompatible coordinate systems and retrieval degrades below baseline. Our alignment loss couples the two spaces, enabling effective cross-space retrieval while preserving within-space hierarchical structure.
- **Breadth amplifies, depth generalizes:** +3.7% R@1 on MedMentions and +1.8% on NCBI Disease over ArboEL (Agarwal et al., 2022). Broad hierarchies amplify

learning for common entities (+6.3%); deep hierarchies generalize to unseen entities (+9.6%).

2 Related Work

2.1 Biomedical Entity Linking

The field evolved from rule-based systems and feature-engineered classifiers (Leaman et al., 2013; Doğan et al., 2014) to neural ranking models (Li et al., 2017). Domain-specific language models—BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021)—brought substantial gains. BioSyn (Sung et al., 2020) combines sparse TF-IDF with dense BioBERT representations. SapBERT (Liu et al., 2021) pre-trains on UMLS synonym pairs to learn that different surface forms refer to the same concept.

Bi-encoder architectures (Wu et al., 2020) are now standard, allowing offline entity encoding and fast approximate nearest neighbor retrieval. Agarwal et al. (2022) propose MST-based clustering over mention-entity similarities to jointly resolve coreferent mentions. GenBioEL (Yuan et al., 2022b) reformulates linking as autoregressive generation; BERGAMOT (Sakhovskiy et al., 2024) incorporates graph neural networks for inter-concept relationships; BELHD (Garda and Leser, 2024) addresses homonym disambiguation.

A persistent challenge is the long-tail distribution; kNN-BioEL (Lin et al., 2024) addresses it with retrieval-enhanced learning, while CYCLE (Zhang et al., 2024) tackles temporal degrada-

tion of evolving knowledge graphs with cross-year contrastive learning. More recently, large language models have been applied after candidate generation for disambiguation and re-ranking (e.g., Ye and Mitchell, 2025). Such methods operate on the outputs of a candidate generator, whereas HiCo-EL improves the generator’s representations themselves; the two directions are complementary rather than competing. Recent methods also differ in benchmark suites and knowledge-base preprocessing (e.g., BELHD), which complicates comparison of published numbers; our experiments therefore use the BioEL framework (Bathala et al., 2025), which provides a common preprocessing and evaluation protocol across models.

2.2 Hierarchical Contrastive Learning

Standard contrastive objectives (Chen et al., 2020; He et al., 2020) treat all negatives uniformly. Hierarchical Multi-label Contrastive learning (Zhang et al., 2022b) relaxes this by applying losses at multiple hierarchy levels, with depth-dependent penalties weighting finer distinctions more heavily. The approach extends supervised contrastive loss (Khosla et al., 2020) to settings where labels form a tree.

Related techniques appear in hierarchical image classification (Bertinetto et al., 2020) and document categorization (Wang et al., 2022). Entity linking has not been addressed with HMLC despite the obvious hierarchical structure of biomedical ontologies. Prior work uses UMLS relations as training signal—SapBERT (Liu et al., 2021) for synonyms, CODER (Yuan et al., 2022c) for relation triplets—but none applies level-wise hierarchical contrastive losses with depth-dependent weighting. We adapt HMLC to the bi-encoder setting with modifications for cross-space alignment and variable-depth hierarchies.

2.3 Type-Aware Entity Linking

Prior work uses types to constrain or mediate the linking decision. DeepType (Raiman and Raiman, 2018) uses type constraints to prune candidates; Onoe and Durrett (2020) link entities through predicted fine-grained types. Some systems build separate indexes per semantic type (Agarwal et al., 2022).

Our approach differs: types shape the embedding space during training, not retrieval at inference. The goal is representations that inherently

respect type boundaries.

3 Method

Figure 2 gives an overview. The core principle is *hierarchy within, alignment across*: Stage 1 applies HMLC independently within each encoder space (hierarchy within), while alignment loss couples them for retrieval (alignment across). We further introduce *unified ontology-to-entity learning* by extending the type hierarchy with CUI-as-leaf. Stage 2 fine-tunes with MST-based clustering.

3.1 Problem Formulation

Let m denote a mention in context and $\mathcal{E} = \{e_1, \dots, e_N\}$ a knowledge base of entities. Entity linking identifies the entity $e^* \in \mathcal{E}$ that m refers to. Each entity has semantic type t_e from a hierarchy $\mathcal{T}$ of depth D .

The bi-encoder architecture encodes mentions and entities independently:

$$\mathbf{h}_m = f_\theta(m) \in \mathbb{R}^d \quad (1)$$

$$\mathbf{h}_e = g_\phi(e) \in \mathbb{R}^d \quad (2)$$

where f_θ is the context encoder and g_ϕ the entity encoder. Linking scores are dot products: $s(m, e) = \mathbf{h}_m^\top \mathbf{h}_e$.

3.2 Stage 1: Hierarchical Contrastive Pretraining

3.2.1 Hierarchy Sources

We leverage different hierarchical knowledge sources depending on the dataset:

UMLS Semantic Type Hierarchy For MedMentions (diverse entity types), we use the UMLS semantic type hierarchy. UMLS organizes concepts into 127 semantic types in a hierarchy of maximum depth 8. For MedMentions ST21PV, we use the 21 designated types plus their ancestors. This provides a *broad* hierarchy with multiple top-level branches.

CTD MEDIC Hierarchy For NCBI Disease, we use the CTD MEDIC hierarchy with 13,388 disease concepts organized in a *deep* parent-child structure. Unlike UMLS semantic types, CTD MEDIC captures fine-grained disease relationships (e.g., “chronic kidney disease” as child of “kidney disease”), providing structural priors for rare and unseen entities.

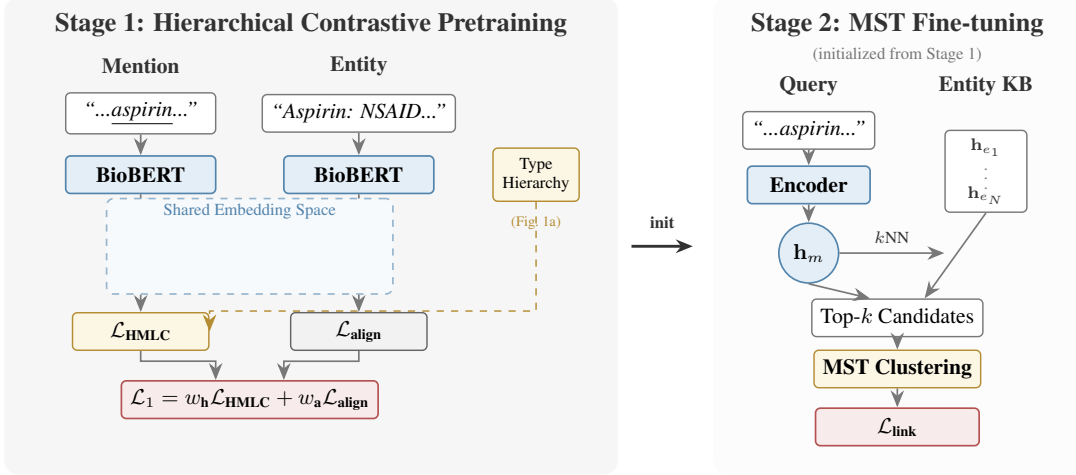


Figure 2: HiCo-EL architecture demonstrating *hierarchy within, alignment across*. **Stage 1:** HMLC loss structures each encoder space independently by type hierarchy, while alignment loss maintains *dual-space coupling* for retrieval. **Stage 2:** MST-based clustering (Agarwal et al., 2022) adds entity-level discrimination atop the hierarchically structured initialization.

For each entity e , we extract its root-to-leaf path:

$$\text{path}(e) = [t_e^{(1)}, t_e^{(2)}, \dots, t_e^{(D)}] \quad (3)$$

where $t_e^{(d)}$ is the ancestor of e 's type at depth d .

Entity-Level Hierarchy Extension Semantic types alone are too coarse—all chemicals share the “Chemical” path. We extend the hierarchy by treating each entity identifier as a leaf node (CUIs for MedMentions; MEDIC concept identifiers for NCBI Disease), creating a unique path per entity while preserving shared structure at higher levels. With $|L|$ total levels (0-indexed), HMLC learns type-level structure (levels 0 to $|L| - 2$) and entity identity (level $|L| - 1$) within a unified framework. At the CUI level, positive pairs are mentions linking to the same entity; samples with no same-CUI partner in the batch (singletons) are excluded from the CUI-level loss computation but still contribute to coarser levels.

We represent each entity’s hierarchical label as a multi-hot vector $\mathbf{y}_e \in \{0, 1\}^{|\mathcal{T}|+|\mathcal{E}|}$ where $y_{e,t} = 1$ if type t is in $\text{path}(e)$ and $y_{e,e} = 1$ for the entity’s own CUI.

3.2.2 Within-Space Hierarchical Multi-label Contrastive Loss

Standard hierarchical contrastive methods (Zhang et al., 2022b) operate on a single embedding space, but bi-encoder retrieval requires two separate spaces that remain compatible for cross-space dot products. Applying hierarchical contrastive

loss *across* spaces conflates type structure with retrieval alignment, producing suboptimal representations.

Our solution is **within-space HMLC**: apply hierarchical contrastive loss independently within each encoder space. For a batch of pairs $\{(m_i, e_i)\}_{i=1}^B$, the normalized embeddings are:

$$\mathbf{z}_m = \frac{f_\theta(m)}{\|f_\theta(m)\|_2} \quad (4)$$

$$\mathbf{z}_e = \frac{g_\phi(e)}{\|g_\phi(e)\|_2} \quad (5)$$

The HMLC loss is computed *within* each space separately: once for mentions ($\mathbf{z}_m$ vs. $\mathbf{z}_m$) and once for entities ($\mathbf{z}_e$ vs. $\mathbf{z}_e$). This organizes each embedding space by type hierarchy while keeping cross-space retrieval decoupled (handled by alignment loss, Section 3.2.3).

At each hierarchy level l , we define positive pairs as samples sharing the same ancestor at level l . The contrastive loss at level l for anchor i is:

$$\mathcal{L}_i^{(l)} = \frac{-1}{|P_l(i)|} \sum_{p \in P_l(i)} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p / \tau_l)}{\sum_{a \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}_a / \tau_l)} \quad (6)$$

where $P_l(i)$ is the set of positives for sample i at level l , and the denominator sums over all samples $a \neq i$ (both positives and negatives), following the supervised contrastive loss formulation (Khosla et al., 2020).

The full HMLC loss aggregates across levels

with depth-dependent weights:

$$\mathcal{L}^{\text{HMLC}} = \sum_{l=0}^{|L|-1} \lambda_l \cdot \frac{1}{B} \sum_{i=1}^B \mathcal{L}_i^{(l)} \quad (7)$$

where $|L|$ is the total number of hierarchy levels and $\lambda_l = e^{1/(|L|-l)}$ assigns higher weight to deeper (more specific) levels, following the HMLC formulation (Zhang et al., 2022b): fine-grained distinctions receive stronger supervision. At the root ($l = 0$), $\lambda_0 \approx 1.22$; at the CUI level ($l = |L| - 1$), $\lambda_{|L|-1} = e \approx 2.72$.

Unlike prior HMLC methods (Zhang et al., 2022b), we drop the monotonicity constraint and sum level losses independently, and handle variable-depth paths natively (losses aggregate only over levels present). This accommodates real-world ontologies such as CTD MEDIC, in which root-to-leaf paths differ in length across entities. We use adaptive temperature $\tau_l = \tau_{\text{base}} \cdot \exp(-\gamma \cdot l/|L|)$ so deeper levels produce sharper distributions ($\tau_{\text{base}} = 0.1$, $\gamma = 1.0$).

The final within-space HMLC loss averages over both encoder spaces:

$$\mathcal{L}_{\text{total}}^{\text{HMLC}} = \frac{1}{2} (\mathcal{L}_{\text{mention}}^{\text{HMLC}} + \mathcal{L}_{\text{entity}}^{\text{HMLC}}) \quad (8)$$

3.2.3 Dual-Space Coupling via Alignment

Bi-encoder retrieval requires mention and entity embeddings to share a common space where dot products are meaningful. The alignment loss achieves *dual-space coupling* by pulling gold pairs together. To incorporate hierarchy structure, we weight by type path depth: entities deeper in the hierarchy (more specific types) receive stronger alignment supervision. Given multi-hot type labels $\mathbf{y}_e$ encoding all ancestors, alignment weights are:

$$w(m_i, e_i) = 1 + \frac{\|\mathbf{y}_{e_i}\|^2}{\max_j \|\mathbf{y}_{e_j}\|^2} \quad (9)$$

The weighted alignment loss is (writing $w_i = w(m_i, e_i)$):

$$\mathcal{L}^{\text{align}} = -\frac{1}{B} \sum_{i=1}^B w_i \log \frac{\exp(\mathbf{z}_{m_i}^\top \mathbf{z}_{e_i} / \tau)}{\sum_j \exp(\mathbf{z}_{m_i}^\top \mathbf{z}_{e_j} / \tau)} \quad (10)$$

where the sum is over all entities in the batch and $\tau = 0.1$. Entities with longer type paths (e.g., “Aspirin” at depth 5) receive higher weight than shallow ones, encouraging the model to learn precise representations for specific entity types.

3.2.4 Pretraining Objective

The Stage 1 loss combines both terms:

$$\mathcal{L}_{\text{Stage 1}} = w_h \cdot \mathcal{L}_{\text{total}}^{\text{HMLC}} + w_a \cdot \mathcal{L}^{\text{align}} \quad (11)$$

The losses operate at different scales: raw HMLC loss is typically smaller (1–2) than alignment loss (5–6) at initialization. Setting $w_h = 4.0$ and $w_a = 1.0$ ensures HMLC receives sufficient gradient signal. Alignment still learns at lower relative weight since cross-space compatibility is reinforced in Stage 2.

Hierarchical Batch Sampling HMLC requires type-similar samples within each batch. Random sampling yields few such pairs—most batches contain samples from different type subtrees with no shared ancestors beyond the root. A hierarchical sampler groups samples by type path, so batches contain multiple entities from similar categories. This increases informative positive pairs per batch, particularly at deeper levels where positives would otherwise be rare.

3.3 Stage 2: Entity Linking with MST-based Clustering

Stage 1 representations provide a structured initialization for entity linking. To isolate hierarchical pretraining’s contribution, Stage 2 uses the MST-based clustering of Agarwal et al. (2022) without modification. This constructs a minimum spanning tree over mention-entity similarities to jointly resolve coreferent mentions. Keeping Stage 2 identical to the baseline means performance differences come from hierarchical initialization. The stages are complementary: HMLC organizes embeddings by type; MST fine-tuning adds entity-level discrimination within type clusters.

3.3.1 Graph Construction and Hard Negative Mining

Following Agarwal et al. (2022), we construct a graph connecting each mention to its k -nearest entity and mention neighbors. Gold clusters group mentions linking to the same entity. A minimum spanning tree over each cluster selects positive edges; negatives come from nearest neighbors outside the cluster.

3.3.2 Contrastive Linking Loss

The linking objective pulls gold pairs together while pushing apart hard negatives. For gold en-

Dataset	Train	Test	KB Size
MedMentions	122,241	39,038	2,369,483
NCBI Disease	5,065	960	13,388

Table 1: Dataset statistics. Train/Test: number of mentions. KB Size: number of candidate entities in the knowledge base.

tity e^+ and hard negatives $\mathcal{N}^-$:

$$\mathcal{L}_{\text{link}} = -\log \frac{\exp(s^+)}{\exp(s^+) + \sum_{e^- \in \mathcal{N}^-} \exp(s^-)} \quad (12)$$

where $s^+ = s(m, e^+)$ and $s^- = s(m, e^-)$. The full objective combines entity-based and mention-based negatives:

$$\mathcal{L}_{\text{Stage 2}} = \mathcal{L}_{\text{link}}^{\text{entity}} + \mathcal{L}_{\text{link}}^{\text{mention}} \quad (13)$$

4 Experiments

4.1 Datasets

We evaluate on two biomedical entity linking benchmarks with contrasting hierarchy characteristics:

MedMentions (Mohan and Li, 2019) The largest publicly available biomedical entity linking benchmark: 4,392 PubMed abstracts with mentions linked to UMLS concepts. Following prior work (Agarwal et al., 2022), we use the ST21PV subset restricted to 21 semantic types. This dataset contains diverse entity types including diseases, chemicals, procedures, and anatomical structures. The **broad hierarchy** (21 types, shallow depth) tests whether hierarchical pretraining can improve disambiguation among common, well-supervised entities.

NCBI Disease (Doğan et al., 2014) A disease-focused corpus with 793 PubMed abstracts. We use the full MEDIC vocabulary with 13,388 disease concepts from CTD. Unlike MedMentions, this dataset contains only disease entities but with a **deep domain hierarchy** from CTD MEDIC (average path depth 6.7; Table 4) that captures fine-grained parent-child relationships. This tests whether hierarchical structure can provide priors for unseen long-tail entities.

Table 1 summarizes the dataset statistics.

4.2 Baselines

We compare against representative methods evaluated using the BioEL framework (Bathala et al.,

Method	MedMentions		NCBI Disease	
	R@1	R@5	R@1	R@5
<i>Evaluated using BioEL framework (Bathala et al., 2025)</i>				
SapBERT	59.2	77.3	75.1	89.7
KRISSBERT	54.6	68.7	74.3	79.5
scispaCy	57.5	75.3	69.5	83.6
BioBART	56.7	78.6	72.6	83.2
GenBioEL	56.3	76.2	73.2	85.3
ArboEL	68.4	83.0	76.3	82.8
<i>Ours</i>				
HiCo-EL	72.1	87.0	78.1	85.8

Table 2: Entity linking results on two biomedical benchmarks with contrasting hierarchy types. Best results in **bold**. ArboEL (Agarwal et al., 2022) uses MST-based clustering.

2025) for fair comparison: **SapBERT** (Liu et al., 2021), self-alignment pretraining on UMLS synonyms; **KRISSBERT** (Zhang et al., 2022a), a knowledge-enhanced bi-encoder; **scispaCy** (Neumann et al., 2019), a spaCy-based biomedical NLP pipeline with entity linking; **BioBART** (Yuan et al., 2022a), a biomedical BART model fine-tuned for entity linking; **GenBioEL** (Yuan et al., 2022b), generative entity linking with knowledge base guidance; **ArboEL** (Agarwal et al., 2022), bi-encoder with MST-based clustering (our baseline without hierarchical pretraining).

4.3 Implementation Details

Both encoders initialize from BioBERT-base-cased-v1.2 (Lee et al., 2020). Stage 1 pretrains with HMLC loss ($w_h = 4.0$) and alignment loss ($w_a = 1.0$) using temperature $\tau = 0.1$. Stage 2 fine-tunes with MST-based contrastive learning (Agarwal et al., 2022). All experiments run on a single NVIDIA A40 GPU; training takes 21–53 hours total depending on dataset. Full hyperparameters are in Appendix A.

4.4 Evaluation Metrics

Standard metrics: **Recall@k** (fraction where the correct entity appears in the top- k candidates). We report R@1 and R@5 to enable direct comparison with BioEL baselines.

4.5 Main Results

Table 2 reports test set performance. HiCo-EL achieves the best R@1 on both benchmarks and outperforms ArboEL (Agarwal et al., 2022) across all metrics.

Category	MedMentions		NCBI Disease	
	$\Delta R@1$	n	$\Delta R@1$	n
Unseen (0)	+0.3	8,716	+9.6	166
Rare (1)	+0.7	1,833	-2.1	48
Low (2–5)	+1.5	4,477	0.0	86
Medium (6–10)	+1.2	2,992	-1.4	74
Common (>10)	+6.3	21,020	+0.7	586
All	+3.7	39,038	+1.8	960

Table 3: R@1 improvement (Δ) stratified by entity frequency. MedMentions shows largest gains for common entities; NCBI Disease shows dramatic gains for unseen entities (+9.6%).

MedMentions (Broad Hierarchy: breadth amplifies) HiCo-EL improves +3.7% R@1 over ArboEL (68.4% $\rightarrow$ 72.1%) and +4.0% R@5. The ArboEL baseline already substantially outperforms prior methods in BioEL (Bathala et al., 2025), including SapBERT (59.2%) and GenBioEL (56.3%)—MST-based clustering is a strong baseline. The broad 21-type hierarchy *amplifies* learning for well-supervised entities.

NCBI Disease (Deep Hierarchy: depth generalizes) Gains are +1.8% R@1 (76.3% $\rightarrow$ 78.1%) and +3.0% R@5. The improvement concentrates on unseen entities (+9.6% R@1)—the deep CTD MEDIC hierarchy *generalizes* to rare diseases even without direct supervision.

4.6 Analysis

4.6.1 Frequency-Stratified Performance

To understand where hierarchical pretraining helps most, we stratify mentions by how often their gold entity appeared in training. Table 3 reports Recall@1 for each category across datasets.

The results show contrasting patterns across the two hierarchy types:

MedMentions: Common entities benefit most. Gains scale with frequency: unseen (+0.3%), rare (+0.7%), common (+6.3%). The UMLS semantic type hierarchy provides *broad categorical boundaries*—21 types separating fundamentally different concepts (Chemical $\neq$ Disease $\neq$ Procedure). For unseen entities, this type-level prior is too coarse: knowing something is a “Chemical” provides weak guidance about *which* chemical. The broad hierarchy *amplifies supervised learning* rather than providing independent priors; common entities with many training examples benefit most because type structure reinforces what direct supervision already teaches.

Level	NCBI Disease	MedMentions
0 (Root)	100%	100%
1	100%	$\sim 5\%$
2	87%	$< 5\%$
3+	$< 7\%$	$< 3\%$
Avg depth	6.7	~ 4

Table 4: Positive pair rate per hierarchy level during HMLC training (per-GPU batch size 32, before gradient accumulation). NCBI Disease has high positive rates at early levels (100% at L0-1, 87% at L2), enabling transfer to unseen entities. MedMentions has low rates due to 21-way type branching, limiting zero-shot generalization but amplifying supervised learning.

NCBI Disease: Unseen entities benefit most.

The CTD MEDIC hierarchy provides *deep domain-specific structure*—fine-grained parent-child relationships within a single semantic type. For unseen diseases, hierarchical position serves as the prior: a disease under “kidney diseases” embeds near other kidney diseases, even without direct supervision (+9.6% for unseen). For example, “nail dystrophy” (unseen in training) is correctly linked by HiCo-EL but missed by the baseline—the hierarchy places it near related nail/skin disorders. Rare entities (frequency 1) show slight regression (-2.1%): HMLC pulls embeddings toward hierarchical neighbors, which can hurt when the single training example was atypical.

Breadth amplifies, depth generalizes. *Broad hierarchies* (many types, shallow depth) amplify supervised learning for common entities; *deep hierarchies* (fine-grained structure) provide priors that generalize to unseen entities. Section 4.6.3 quantifies this via positive pair rates.

4.6.2 Embedding Space Visualization

Figure 3 visualizes entity embeddings via t-SNE, comparing BioBERT baseline with HiCo-EL at two hierarchy levels. Hierarchical contrastive pretraining improves type-aware clustering at both coarse and fine granularities.

4.6.3 Hierarchy Structure Analysis

Why do broad and deep hierarchies produce different benefit patterns? Table 4 shows positive pair rates at each hierarchy level during HMLC training.

This quantifies why *breadth amplifies*, *depth generalizes*: NCBI Disease has 100% positive pairs at levels 0–1 and 87% at level 2, enabling

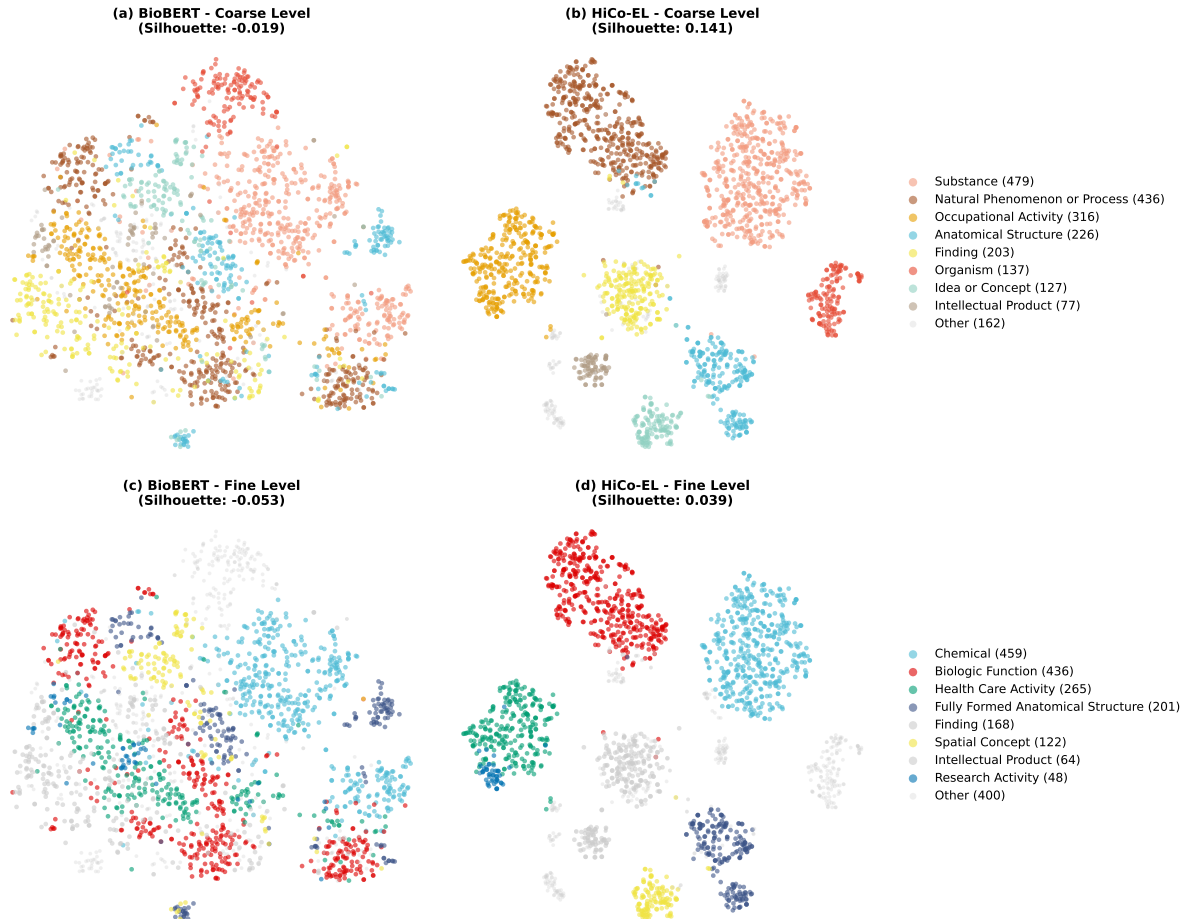


Figure 3: t-SNE visualization of entity embeddings at multiple UMLS hierarchy levels. **Top row:** Coarse-level clustering (Substance, Organism, Anatomical Structure, etc.). **Bottom row:** Fine-grained clustering (Chemical, Biologic Function, Health Care Activity, etc.). Silhouette scores quantify clustering quality: HiCo-EL improves from negative (poor separation) to positive (type-aware clusters) at both levels. Coarse: $-0.019 \rightarrow 0.141$ (+0.160); Fine: $-0.053 \rightarrow 0.039$ (+0.092).

HMLC to learn structure that generalizes to unseen entities (+9.6%). MedMentions’ 21-way branching yields only $\sim 5\%$ at level 1—the hierarchy amplifies supervised learning, benefiting common entities most (+6.3%). The contrast is structural: CTD MEDIC’s deep paths (average depth 6.7) sustain informative positive pairs across multiple levels, whereas MedMentions’ shallower hierarchy (average depth ~ 4) concentrates them at the root.

4.7 Ablation Study

Table 5 ablates each component on MedMentions.

Dual-space coupling is essential. Removing alignment causes the largest drop (-5.7%), falling below baseline—without *dual-space coupling*, the encoders develop incompatible coordinate systems. Using alignment only (no HMLC) matches

Configuration	R@1	R@5	$\Delta R@1$
HiCo-EL (full)	72.1	87.0	—
– Hierarchy (flat contrastive)	68.4	81.7	−3.7
– CUI-as-leaf	68.6	82.9	−3.5
– Alignment (HMLC only)	66.4	81.1	−5.7
– HMLC (alignment only)	68.0	81.4	−4.1
Bi-enc + MST (baseline)	68.4	83.0	−3.7

Table 5: Ablation study on MedMentions. All components contribute to the full model’s performance. Removing alignment causes the largest drop (-5.7%), falling below the baseline.

baseline (68.0%) but lacks the benefits of *hierarchy within*. Removing *ontology-to-entity unification* (CUI-as-leaf, -3.5%) or hierarchy entirely (-3.7%) confirms both contribute.

5 Conclusion

Bi-encoder entity linking benefits from embedding spaces that reflect ontology structure. HiCo-EL adapts hierarchical contrastive learning to this setting through three key ideas: *hierarchy within*, *alignment across* for decoupled structuring; *unified ontology-to-entity learning* via CUI-as-leaf; and *dual-space coupling* for retrieval compatibility.

Results on MedMentions (+3.7% R@1) and NCBI Disease (+1.8% R@1) over ArboEL reveal a core insight: **breadth amplifies, depth generalizes**. Broad hierarchies amplify learning for common entities (+6.3% R@1); deep hierarchies generalize to unseen entities (+9.6% R@1 zero-shot).

Ontological hierarchies are pervasive in biomedical NLP. Future work might extend these ideas to cross-lingual settings or multi-ontology scenarios with overlapping type systems.

Limitations

Computational Cost. The two-stage training pipeline increases computational overhead compared to single-stage approaches. Stage 1 pre-training requires 20 hours for MedMentions and 1 hour for NCBI Disease on a single A40 GPU, followed by Stage 2 fine-tuning (33 hours and 3 hours respectively). This may limit applicability in resource-constrained settings.

Hierarchy Quality Dependence. The quality of learned representations depends directly on the underlying type hierarchy. Noisy, incomplete, or inconsistent type annotations in the knowledge base will propagate to the embedding space. Our approach assumes a well-structured hierarchy is available, which may not hold for all biomedical ontologies or domains.

Language and Domain Scope. All experiments use English biomedical text. Cross-lingual generalization is untested, and performance on non-English clinical text or other domains (e.g., legal, scientific literature outside biomedicine) remains unknown. The method’s reliance on domain-specific hierarchies (UMLS, CTD MEDIC) may also limit direct transfer to other fields.

Benchmark Coverage. Our evaluation focuses on two benchmarks (MedMentions ST21PV and NCBI Disease) under the BioEL framework (Bathala et al., 2025). Evaluating on a broader set

of standardized benchmarks, and combining hierarchical pretraining with post-generation disambiguation methods, are directions for future work.

Ethics Statement

This work uses publicly available biomedical datasets and pre-trained language models. Entity linking systems deployed in clinical settings require careful validation before use, as incorrect linking could propagate errors to downstream applications with patient safety implications.

Acknowledgments

We used generative AI assistants (Claude, ChatGPT) for grammar checking and language polishing; all original content, scientific claims, experimental design, and conclusions are the authors’ own. Funding support provided by National Science Foundation CAREER grant 1944247, National Institutes of Health grant R35GM152245, and Chan Zuckerberg Initiative grant 253558 to C.S.M.

References

- Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4644–4658.
- Prasanth Bathala, Christophe Ye, Batuhan Nursal, Shubham Lohiya, David Kartchner, and Cassie S. Mitchell. 2025. BioEL: A comprehensive Python package for biomedical entity linking. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1709–1721.
- Luca Bertinetto, Romain Mueller, Konstantinos Terzikas, Sina Samangooei, and Nicholas A Lord. 2020. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12506–12515.
- Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

- Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.
- Samuele Garda and Ulf Leser. 2024. BELHD: Improving biomedical entity linking with homonym disambiguation. *Bioinformatics*, 40(8):btac474.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in neural information processing systems*, volume 33, pages 18661–18673.
- Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNORM: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. CNN-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18(Suppl 11):385.
- Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024. Improving biomedical entity linking with retrieval-enhanced learning. In *ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 11461–11465.
- Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.
- Sunil Mohan and Donghui Li. 2019. MedMentions: A large biomedical corpus annotated with UMLS concepts. In *Proceedings of the 2019 Conference on Automated Knowledge Base Construction*.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.
- Yasumasa Onoe and Greg Durrett. 2020. Fine-grained entity typing for domain independent entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8576–8583.
- Jonathan Raiman and Olivier Raiman. 2018. DeepType: multilingual entity linking by neural type system evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 5406–5413.
- Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2024. Biomedical entity representation with graph-augmented multi-objective transformer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4626–4643.
- Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650.
- Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7109–7119.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407.
- Christophe Ye and Cassie S. Mitchell. 2025. [LLM as entity disambiguator for biomedical entity-linking](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–312, Vienna, Austria. Association for Computational Linguistics.
- Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. BioBART: Pre-training and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.
- Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022b. Generative biomedical entity linking via knowledge

base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022c. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983.

Pengyu Zhang, Congfeng Cao, Klim Zaporozhets, and Paul Groth. 2024. CYCLE: Cross-year contrastive learning in entity-linking. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 3197–3206.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022a. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. 2022b. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669.

A Hyperparameters

Table 6 lists all hyperparameters used in our experiments.

Hyperparameter	MedMentions	NCBI Disease
<i>Stage 1: Hierarchical Contrastive Pretraining</i>		
Base model	BioBERT-base-cased-v1.2	
Batch size	32	32
Gradient accumulation	4	4
Effective batch size	128	128
Learning rate	2×10^{-5}	5×10^{-6}
Max epochs	30	30
Early stopping patience	5	5
Temperature τ	0.1	0.1
HMLC weight w_h	4.0	4.0
Alignment weight w_a	1.0	1.0
Training time	20 hours	1 hour
<i>Stage 2: MST-based Fine-tuning</i>		
Learning rate	5×10^{-6}	5×10^{-6}
Epochs	20	20
k -NN for graph	10	10
Training time	33 hours	3 hours

Table 6: Hyperparameters for all experiments. All runs use a single NVIDIA A40 GPU with FAISS for nearest neighbor search.