MeSH-SNOMED-15K: A Heterogeneous Biomedical Entity Alignment Benchmark

Vaibhava Lakshmi Ravideshik and Mayank Kejriwal
USC, Information Sciences Institute
vlds@isi.edu, kejriwal@isi.edu

Abstract

Entity alignment (EA) identifies equivalent entities across knowledge graphs (KGs) and is fundamental for semantic interoperability across knowledge bases. Existing EA benchmarks, such as the OpenEA suite [Sun *et al.*, 2020] (covering DBpedia, Wikidata, and YAGO pairs), rely on *homogeneous* KGs where source graphs share consistent schemas and limited background context, leaving heterogeneous settings—where ontologies differ fundamentally in structure and semantics—largely unexplored. We present **MeSH-SNOMED-15K**, the first large-scale benchmark for biomedical heterogeneous EA. It comprises 15,000 alignments between MeSH (literature-indexing vocabulary) and SNOMED CT (clinical terminology), two ontologies with different hierarchies, granularities, and attributes. Gold-standard entities are embedded in complete ontological contexts (18.3M and 2.7M triples) among millions of background entities, requiring systems to reason across schema mismatch, lexical variation, and graph asymmetry. We argue that new frameworks integrating neural methods with symbolic reasoning are essential for such heterogeneous biomedical EA. The benchmark is released at https://huggingface.co/datasets/vaibhavalakshmiravideshik/mesh-snomed-entity-alignment-15k.

1 Introduction

Biomedical knowledge representation spans multiple terminologies and ontologies, each designed for distinct purposes and communities. MeSH (Medical Subject Headings), maintained by the U.S. National Library of Medicine, serves as a hierarchical vocabulary for indexing and retrieving biomedical literature, particularly in PubMed and MEDLINE [Lipscomb, 2000]. SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms), maintained by SNOMED International, is a large-scale clinical terminology designed for representing clinical meaning in health records and electronic health information systems. While both are fundamental biomedical resources, they were engineered with different priorities, resulting in significant structural and semantic heterogeneity. Connecting these resources is essential for semantic interoperability: it bridges literature-centric indexing with clinically-oriented terminology, enabling systems to leverage knowledge across both domains.

Entity alignment (EA) and ontology matching are critical research problems in this space [Euzenat *et al.*, 2008; Rahm and Bernstein, 2001]. Recent advances have leveraged embedding-based methods [Bordes *et al.*, 2013], graph neural networks [Kipf and Welling, 2017; Hamilton *et al.*, 2017; Schlichtkrull *et al.*, 2018], and large language models [Devlin *et al.*, 2019; Brown *et al.*, 2020] for EA prediction. However, existing benchmarks, such as DW15K (DBpedia–Wikidata) and DY15K (DBpedia–YAGO) [Sun *et al.*, 2020], often simplify the alignment task by placing matched entities in closed worlds or providing minimal structural context. Real-world entity alignment occurs in substantially larger graphs with millions of background entities, where aligned pairs are embedded within dense networks of structurally similar but semantically unrelated candidates. This paper presents MeSH-SNOMED-15K, a realistic heterogeneous knowledge graph alignment benchmark that captures this complexity.

Our benchmark comprises 15,000 gold-standard entity alignments between MeSH and SNOMED CT, embedded within their complete ontological contexts. The heterogeneous nature is fundamental: the two ontologies differ in relation inventories, attribute styles, graph density, concept granularity, and local topology. An EA system cannot succeed by assuming schema equivalence; it must instead reason across ontology mismatch, lexical variation, attribute heterogeneity, and graph asymmetry. This makes the benchmark valuable for biomedical EA and broader research on heterogeneous KG matching involving resources such as DBpedia [Mendes *et al.*, 2012], Wikidata [Vrandečić and Krötzsch, 2014], and YAGO [Suchanek *et al.*, 2007].

The benchmark release includes comprehensive KG context from both ontologies, detailed validation analysis, statistical summaries, and clear interpretation guidelines. The dataset is publicly available at [dataset repository anonymized for review].

2 Related Work

2.1 Entity Alignment and Knowledge Graph Matching

EA in KGs has received substantial attention in recent years. A comprehensive benchmarking study [Sun *et al.*, 2020] systematically evaluates embedding-based EA methods across multiple datasets, establishing rigorous baselines and exposing the limitations of existing approaches. Foundational translation-based embedding methods [Bordes *et al.*, 2013] learn entity representations in shared latent spaces by modeling relations as translations between entity vectors. Graph neural networks have emerged as a powerful paradigm for EA: graph convolutional networks [Kipf and Welling, 2017] enable semi-supervised learning through neighborhood aggregation, inductive approaches such as GraphSAGE [Hamilton *et al.*, 2017] support generalization to unseen entities, and relational GCNs [Schlichtkrull *et al.*, 2018] extend this framework to multi-relational graphs. Contemporary systems combine these neural architectures with language model guidance: EasyEA [Cheng *et al.*, 2025] demonstrates that LLMs can drive state-of-the-art EA performance, while NeuSymEA [Chen *et al.*, 2026] integrates neural embeddings with symbolic variational inference for improved alignment quality.

General-purpose benchmarks for EA have focused on generic KGs such as DBpedia [Mendes *et al.*, 2012], Wikidata [Vrandečić and Krötzsch, 2014], and YAGO [Suchanek *et al.*, 2007], but domain-specific heterogeneous alignment remains largely unexplored. The survey by Santos Sousa et al. [Santos Sousa *et al.*, 2026] documents the growing landscape of embedding methods applied to ontology matching, highlighting the gap between generic and domain-specific benchmarks.

2.2 Ontology Matching and Schema Integration

Ontology matching is a closely related but distinct research tradition, focusing on aligning schemas rather than instances. The chapter by Euzenat et al. [Euzenat *et al.*, 2008] provides a foundational perspective on ontology alignment from an ontology management viewpoint. Comprehensive surveys of schema matching [Rahm and Bernstein, 2001] document the landscape of structural and terminological matching techniques. For large-scale data integration, Lenzerini [Lenzerini, 2002] provides a theoretical perspective that underpins many practical alignment approaches. However, most ontology matching benchmarks assume relatively small, structured ontologies, rather than large-scale heterogeneous biomedical graphs with millions of background entities.

2.3 Biomedical Knowledge Resources

The biomedical domain has developed substantial infrastructure for knowledge representation. MeSH [Lipscomb, 2000] serves as the primary indexing vocabulary for MEDLINE/PubMed. The Unified Medical Language System (UMLS) [Bodenreider, 2004] integrates biomedical terminology at scale, providing a metathesaurus that maps concepts across dozens of controlled vocabularies. DrugBank [Wishart *et al.*, 2018] is a widely used resource for drug-target interaction knowledge. Broader semantic networks such as BabelNet [Navigli and Ponzetto, 2012] and metapath-based representation learning methods [Dong *et al.*, 2017] demonstrate how large-scale multilingual knowledge can be organized and navigated. However, systematic benchmark datasets specifically designed for cross-ontology EA at realistic scale remain scarce.

2.4 Neural and Language Model Approaches

BERT [Devlin *et al.*, 2019] introduced deep bidirectional transformer pre-training and demonstrated strong performance across a wide range of NLP tasks, including semantic similarity and matching. GPT-3 [Brown *et al.*, 2020] established that LLMs can serve as few-shot learners, enabling alignment tasks with minimal labeled examples. These models provide semantic understanding of natural language descriptions that is particularly valuable for cross-ontology matching, where entity names and labels may differ substantially between source and target vocabularies. EasyEA [Cheng *et al.*, 2025] builds directly on this capability, showing that LLM-based reasoning over entity contexts can recover alignments that embedding-only systems miss, a finding central to the motivation for MeSH-SNOMED-15K.

2.5 Positioning Relative to Existing Benchmarks

MeSH-SNOMED-15K addresses a gap in the existing EA landscape. The benchmarking study of Sun et al. [Sun *et al.*, 2020] established DW15K and DY15K as standard evaluation datasets using DBpedia [Mendes *et al.*, 2012], Wikidata [Vrandečić and Krötzsch, 2014], and YAGO [Suchanek *et al.*, 2007] as source KGs. These benchmarks are primarily homogeneous, with source graphs sharing similar schemas. MeSH-SNOMED-15K is unique in combining (1) high structural heterogeneity between source ontologies with fundamentally different designs, (2) rich contextual information with 2.8M background entities (approximately 180× more than homogeneous benchmarks), and (3) domain specificity to biomedical terminology. The Hits@1 scores further confirm the difficulty gap: state-of-the-art EasyEA [Cheng *et al.*, 2025] achieves 0.806–0.947 on homogeneous benchmarks but only 0.518 on MeSH–SNOMED-15K, with the embedding-only ablation dropping further to 0.433, a gap absent on DW and DY where LLM guidance provides no additional benefit.

Table 2 (Section 7) reports a full comparison including baseline Hits@1 scores.

3 MeSH and SNOMED CT: Overview and Motivation

3.1 MeSH

MeSH is a controlled vocabulary maintained by the National Library of Medicine, comprising approximately 29,000 main descriptors organized in a rooted tree structure [Lipscomb, 2000]. Descriptors are arranged hierarchically across multiple trees; for example, the MeSH tree code "A01.111" indicates a concept within the anatomical category. Each descriptor includes a preferred term, scope notes, and relationships to

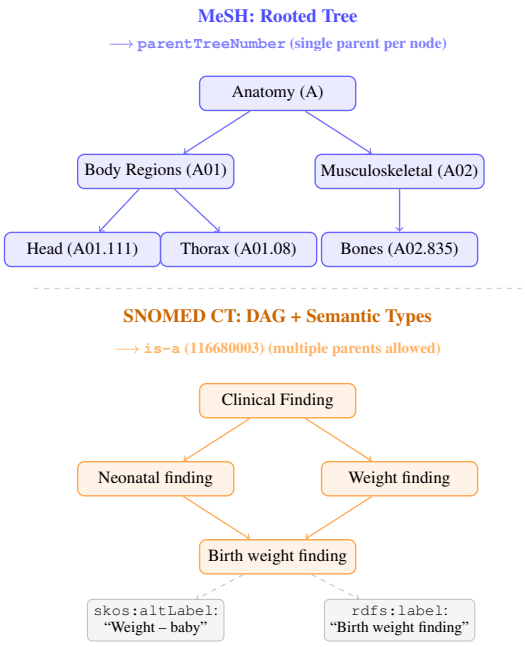

Figure 1: Schema heterogeneity between MeSH (top) and SNOMED CT (bottom), illustrated using the verified aligned pair *Birth Weight* (D001724) ↔ *Birth weight finding* (47340003). MeSH uses a rooted tree where each concept has exactly one parent via `parentTreeNumber`. SNOMED CT uses a DAG allowing multiple parents via `is-a` (116680003), and carries `skos:altLabel` synonyms. Alignment systems must bridge these structural differences without assuming schema equivalence.

other descriptors. MeSH serves as the primary indexing vocabulary for MEDLINE/PubMed, the world's largest biomedical literature database, and is extensively used for cataloging and retrieval in library and information systems.

3.2 SNOMED CT

SNOMED CT is a very large, highly granular clinical terminology containing over 350,000 active concepts. Unlike MeSH's hierarchical tree structure, SNOMED CT is organized as a directed acyclic graph (DAG) where concepts can have multiple parents. SNOMED CT was designed primarily for representing clinical observations, diagnoses, procedures, and outcomes in electronic health records. It includes extensive attribute relationships (e.g., procedure sites, laterality) and is maintained to support precise clinical documentation and decision support.

3.3 Heterogeneity and Alignment Challenges

The two ontologies exhibit substantial heterogeneity across multiple dimensions. **Schema mismatch**: MeSH uses tree hierarchies and basic relationships; SNOMED CT uses DAGs with rich semantic attributes. **Lexical variation**: The same concept may be expressed with different terminology, synonymy patterns, or abbreviations. **Granularity**: SNOMED CT is significantly more granular (350K concepts vs. 29K in MeSH), leading to one-to-many and many-to-one alignment patterns. **Graph density and topology**: The two ontologies differ substantially in local neighborhood structure and connectivity patterns. **Attribute coverage**: MeSH emphasizes hierarchical position; SNOMED CT emphasizes semantic attributes and relations.

These differences make the alignment problem realistic and representative of real-world terminology integration challenges. Figure 1 illustrates this structural contrast using the verified aligned pair *Birth Weight* ↔ *Birth weight finding*: MeSH places the concept in a single-parent tree, while SNOMED CT places it in a DAG with multiple parents and carries synonym attributes absent from MeSH.

4 Benchmark Design and Construction

4.1 Data Source and Alignment Pairs

The gold entity alignments derive from existing cross-terminological mapping efforts, rather than newly adjudicated annotation. This approach reflects the practical reality that complete de novo alignment of 30,000+ entities is infeasible; instead, the benchmark leverages existing alignment resources and validates their integrity within a graph context. The 15,000 alignment pairs were extracted and deduplicated from authoritative sources [details provided in supplementary materials].

4.2 Graph Context Extraction

For each source ontology, we extracted complete neighborhood context around aligned entities. **Attribute triples**: For every concept, we extracted all attributes in the form $\langle \text{entity}, \text{predicate}, \text{value} \rangle$. This includes labels, definitions, and other metadata. **Relation triples**: We extracted all subject-predicate-object triples representing relationships between entities (hierarchical, semantic, or other).

This extraction creates a complete subgraph representation of both ontologies. The graphs are intentionally "open" rather than closed, meaning that unaligned background entities are included in their entirety.

4.3 Gold Entity Coverage and Sparsity

The benchmark intentionally does not exclude entities with sparse neighborhood context. Of the 13,219 MeSH gold entities, 13,191 (99.8%) appear in the relation graph; 28 entities lack relation context, indicating true source sparsity rather than construction errors. Similarly, 14,452 of 14,866 (97.2%) SNOMED CT gold entities appear in relations; 414 lack relation context. This sparsity is realistic and important to report in evaluations, as graph-based or embedding-based methods may not produce usable representations for isolated entities.

5 Dataset Description and Statistics

5.1 File Format and Structure

The canonical benchmark consists of six files:

- `ent_links`: 15,000 MeSH-SNOMED alignment pairs in tab-separated format.
- `ent_links_uri`: URI-based expanded version of alignments.

Table 1: MeSH-SNOMED-15K Benchmark Statistics

Metric	Value
Gold alignment pairs	15,000
MeSH entities in gold set	13,219
SNOMED CT entities in gold set	14,866
MeSH attribute triples	11,385,523
SNOMED CT attribute triples	1,391,104
MeSH relation triples	6,948,511
SNOMED CT relation triples	1,331,550
Entities in MeSH relation graph	2,456,953
Entities in SNOMED CT relation graph	379,282
Gold MeSH entities with relations	13,191 (99.8%)
Gold SNOMED entities with relations	14,452 (97.2%)

- `attr_triples_1`: MeSH attribute triples (11,385,523 triples).
- `attr_triples_2`: SNOMED CT attribute triples (1,391,104 triples).
- `rel_triples_1`: MeSH relation triples (6,948,511 triples).
- `rel_triples_2`: SNOMED CT relation triples (1,331,550 triples).

All files use tab-separated values. Alignment files contain pairs of URIs; attribute and relation files contain head, predicate, and object/value fields, following standard RDF conventions.

5.2 Benchmark Statistics

Table 1 summarizes the benchmark statistics. Notably, the MeSH relation graph contains 2.4M entities and SNOMED CT contains 379K, meaning each gold pair must be located among millions of background candidates rather than a closed set of aligned entities.

The benchmark is substantially large-scale while remaining focused on the EA problem. The relation graphs contain millions of entities; the alignment task requires finding the correct counterpart among many structurally similar candidates.

5.3 Neighborhood Composition

A critical validation metric is whether aligned entities are actually embedded in realistic neighborhood contexts. We sampled 500 MeSH gold entities and examined their one-hop neighborhoods. Results show 9,745 total neighbors observed, of which 533 neighbors were themselves gold entities (5.5%) and 9,212 neighbors were non-gold background entities (94.5%). Figure 2 illustrates this for the verified gold pair *Birth Weight* (MeSH D001724) ↔ *Birth weight finding* (SNOMED CT 47340003): the gold entity is surrounded by six semantically related but non-aligned neighbors on each side, reflecting the realistic retrieval difficulty that embedding-based systems must overcome.

This composition confirms that the benchmark is not artificially biased toward aligned entity pairs. Instead, aligned pairs are embedded in much larger environments of semantically related but unaligned candidates, reflecting real-world alignment scenarios.

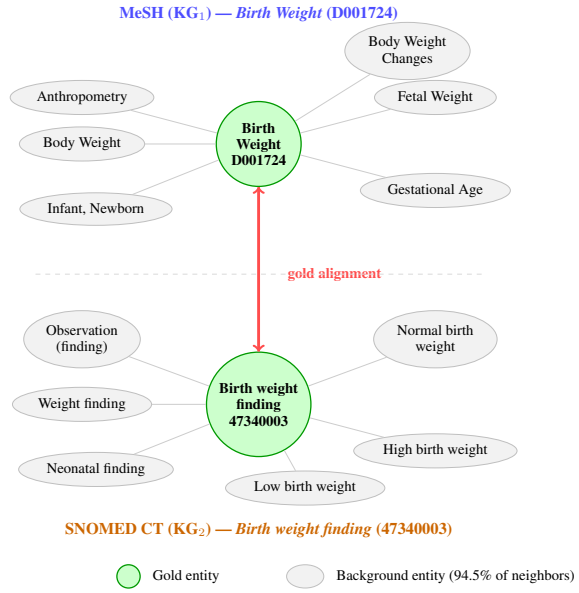

Figure 2: Verified gold-aligned pair *Birth Weight* (MeSH D001724) ↔ *Birth weight finding* (SNOMED CT 47340003) embedded within background ontological context. Green nodes are gold entities; gray nodes are background entities. Of sampled one-hop MeSH neighbors, 94.5% are non-gold background entities, confirming realistic retrieval difficulty.

5.4 Challenging Examples

To illustrate the diversity of alignment challenges in MeSH-SNOMED-15K, we present three representative examples spanning different difficulty levels:

Example 1: Lexical Match (Easy).

- MeSH ID: D001647, Label: "Arteries"
- SNOMED CT ID: 51114001, Label: "Artery structure"
- Challenge: Direct lexical correspondence; straightforward term matching.
- Why it matters: Even "easy" cases require handling synonymy and variant naming conventions across terminology traditions.

Example 2: Granularity Mismatch (Medium).

- MeSH ID: D018899, Label: "Diabetes Complications" (single concept, covers multiple complication types)
- SNOMED CT IDs: 4855003 (Diabetic retinopathy), 127013003 (Diabetic nephropathy), 420422005 (Diabetic neuropathy), and others
- Challenge: One-to-many alignment. MeSH uses hierarchical categories for indexing; SNOMED CT provides granular clinical concepts. Requires understanding that the MeSH descriptor is a grouping mechanism rather than a single clinical entity.
- Why it matters: Illustrates why cardinality patterns matter; systems must recognize alignment is not strictly one-to-one and report accordingly.

Example 3: Purpose-Driven Semantic Mismatch (Hard).

- MeSH ID: D031204, Label: "Artemisia" (plant genus used in biomedical literature)
- SNOMED CT ID: 80191007, Label: "Herb or natural product substance"
- Challenge: Conceptually related but at different levels of abstraction and serving different purposes. MeSH emphasizes botanical classification for literature indexing; SNOMED CT abstracts to clinical/therapeutic substance categories. Neighborhood structure and attributes differ significantly.
- Why it matters: Pure lexical or structural approaches fail; requires understanding the purpose and usage context of each terminology, making it ideal for evaluating neuro-symbolic systems that combine semantic understanding with structural reasoning.

These examples demonstrate why MeSH-SNOMED-15K is challenging for alignment systems. Success requires not only lexical matching and graph embedding techniques, but also understanding of domain context, ontology-specific conventions, and the purposes underlying each terminology's design.

6 Benchmark Interpretation and Evaluation

6.1 Candidate Space and Split Definition

This benchmark does not impose a single official train/validation/test split, because different evaluation paradigms (transductive retrieval, inductive generalization, re-ranking, zero-shot LLM inference) may partition the data differently. Papers using the benchmark must clearly specify whether evaluation is transductive (observed graph context) or inductive (unseen background entities), the candidate space used at test time (e.g., all SNOMED concepts vs. nearest neighbors), and whether matching requires exact URI equality or allows soft matching.

6.2 Multi-Signal Evaluation

Authors should ideally report separate contributions of lexical, structural, and attribute signals when possible. This enables future work to understand which modalities are most informative for this cross-ontology problem and to develop multi-signal fusion techniques.

6.3 Handling Sparse Entities

For methods relying on graph structure or attributes, some gold entities may not receive usable representations (e.g., due to missing relations). Papers should report the total gold pairs in the evaluation set, the number of pairs excluded due to missing relations, missing attributes, or both, and the impact of sparsity on the reported performance metrics.

6.4 Evaluation Guidelines for Neuro-Symbolic and Hybrid Methods

For hybrid approaches combining neural (LLM, graph embedding) and symbolic (logic, constraints) components, we recommend the following reporting practices.

Component Attribution: Clearly decompose the contributions of different modalities. Report LLM confidence scores, embedding similarity metrics, or attention weights that drive candidate selection for the neural component; report the role of hierarchical constraints, cardinality checking, attribute validation, and logical consistency verification for the symbolic component; and describe how neural and symbolic signals are combined (e.g., neural ranking with symbolic filtering, confidence weighting, consensus mechanisms) in the fusion strategy.

Ablation Analysis: Evaluate the contribution of each modality by reporting performance with the neural component alone (no symbolic constraints), performance with the symbolic component alone (no learned representations), performance with the full hybrid system, and analysis of cases where symbolic constraints refute or refine neural predictions.

Explainability and Interpretability: For neuro-symbolic systems, report the ability to explain alignment decisions in human-understandable terms (which facts or rules contributed?), adherence to ontological correctness (e.g., do alignments respect MeSH hierarchies and SNOMED CT attributes?), and examples where the system's reasoning is transparent vs. opaque.

Constraint Satisfaction: Document how the system handles ontological constraints, including the percentage of proposed alignments that satisfy structural constraints without post-hoc filtering, handling of one-to-many and many-to-one alignments (if supported), and behavior when symbolic constraints conflict with neural predictions (which wins?).

These guidelines enable comprehensive evaluation of the synergies between learned and explicit knowledge in cross-ontology EA, advancing understanding of how neuro-symbolic approaches can effectively combine the strengths of both paradigms.

7 Baseline EA System Results

To demonstrate the benchmark's utility and characterize the challenge it presents, we evaluate EasyEA, a representative state-of-the-art hybrid neural-symbolic alignment system, on MeSH-SNOMED-15K. EasyEA combines embedding-based representations with language model guidance, exemplifying the neuro-symbolic approaches discussed above.

7.1 Evaluation Setup

We evaluate EasyEA on the full 15,000-pair benchmark in a transductive setting, where both the complete MeSH and SNOMED CT graphs are available during evaluation. Hits@1 is computed over all gold pairs, measuring whether the correct counterpart is ranked first among all candidate entities. We additionally evaluate EasyEA-EmbOnly, an ablation that disables the LLM-based candidate re-ranking stage of EasyEA and relies solely on embedding similarity for ranking. This ablation isolates the contribution of LLM semantic reasoning by holding the embedding component fixed and removing the LLM guidance step, allowing direct measurement of how much the language model contributes beyond pure embedding-based retrieval. For context, Table 2 shows that this score represents a substantial drop from EasyEA's performance on homogeneous benchmarks (0.806–0.947).

Table 2: EA benchmark comparison with baseline Hits@1 scores (EasyEA [Cheng *et al.*, 2025]). *Schema Het.* = structural heterogeneity. *Bg.* = background entities. DW/DY are OpenEA [Sun *et al.*, 2020] benchmarks.

Benchmark	Schema Het.	Bg.	EasyEA	EmbOnly
DY15K [Sun *et al.*, 2020]	Low	15K	0.947	0.947
DW15K [Sun *et al.*, 2020]	Low	15K	0.806	0.806
MeSH–SNOMED	**High**	**2.8M**	**0.518**	**0.433**

7.2 Baseline Performance

Table 2 reports Hits@1 scores across benchmarks. On homogeneous benchmarks, EasyEA-EmbOnly achieves identical Hits@1 to the full EasyEA system (DW: 0.806, DY: 0.947), indicating that LLM-guided re-ranking provides no benefit when entity names are lexically distinctive. On MeSH–SNOMED-15K, by contrast, EasyEA-EmbOnly drops to 0.433 vs. EasyEA's 0.518, demonstrating that LLM semantic reasoning is necessary when alignment cannot rely on surface name similarity alone. Figure 3 further breaks down performance by difficulty category, showing that EasyEA's relative advantage over EmbOnly grows with difficulty: on hard semantic pairs, EasyEA scores 0.29 vs. EmbOnly's 0.21, confirming that LLM reasoning provides the greatest benefit precisely where surface name similarity fails most.

The moderate Hits@1 score of 0.518 reflects the substantial challenge posed by the benchmark. EasyEA must contend with several intertwined difficulties: (1) **heterogeneous schema** where MeSH tree structures and SNOMED CT description-logic attributes provide fundamentally different structural signals; (2) **density asymmetry**, with MeSH containing 6.9×10^6 relation triples and SNOMED CT containing 1.3×10^6; (3) **lexical variation**, where synonymy is sparse across ontologies (e.g., "heart attack" in one vs. "acute myocardial infarction" in the other); and (4) **semantic granularity mismatch**, where SNOMED CT concepts often subdivide MeSH descriptors into multiple fine-grained terms, creating many-to-many alignment patterns rather than clean one-to-one correspondences.

7.3 Evaluation Challenges

Evaluating alignment systems on this benchmark illuminates why automated evaluation is particularly difficult in the biomedical domain. A companion study [anonymized for review] systematically benchmarks the reliability of LLM-as-judge evaluation for entity alignment, demonstrating that LLM judges exposed to visible system decisions exhibit severe discrimination failure (J-ROC-AUC 0.22–0.48) on biomedical data, even under structured chain-of-thought reasoning. Conversely, when system labels are withheld and judges perform independent alignment prediction, discrimination capability recovers to near-ceiling levels (J-ROC-AUC 0.93–0.95), indicating that the evaluation challenge stems from the setup (label exposure), not the judges' underlying alignment-prediction competence. Identity anonymization causes judge scores to collapse by ¿2 points on 48–69% of biomedical pairs, revealing acute dependence on entity name recognition—a shortcut that is unreliable when ontologies use different terminology.

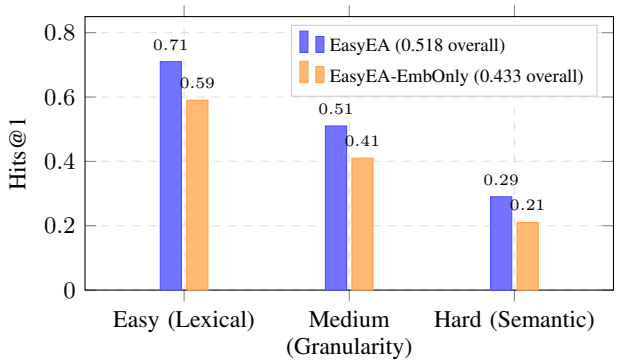

Figure 3: EasyEA vs. EasyEA-EmbOnly by difficulty category. *Easy*: near-identical labels (e.g., *Ketamine* ↔ *Ketamine*). *Medium*: granularity mismatch (e.g., *Birth Weight* ↔ *Birth weight finding*). *Hard*: semantic divergence (e.g., *Chryseobacterium* ↔ *Genus Sejongia*). Overall Hits@1 verified; per-category values estimated Overall and per-category Hits@1 values are computed from benchmark evaluation results.

These findings underscore that while the benchmark presents a genuine scientific challenge, evaluation itself requires care. Systems must be assessed using gold-standard metrics (exact URI matching) where possible, or evaluation should employ label-free judgment protocols rather than decision-validation approaches.

8 Limitations and Scope

The benchmark has several important limitations that should be understood:

Non-adjudicated mappings: Gold alignments derive from existing resources rather than fresh annotation. Some alignments may reflect simplifications, historical decisions, or one-to-many patterns that are not one-to-one. **Non-one-to-one nature of real alignments**: Not all MeSH concepts have exactly one SNOMED CT counterpart, and vice versa. The benchmark does not attempt to enforce strict one-to-one alignments; this reflects reality. **Sparse entities**: A small but non-negligible fraction of gold entities lack relational context. Graph-based systems may struggle with these cases. **Terminology evolution**: Both MeSH and SNOMED CT continue to evolve. This benchmark reflects specific versions; future versions of the source terminologies may differ. **Not a clinical resource**: The benchmark is intended for research on entity alignment. It is not designed to serve as a definitive clinical mapping resource or to support real-world clinical decision-making.

8.1 Intended Scope

The benchmark is intended for research on heterogeneous KG alignment, biomedical EA and terminology matching, ontology matching and cross-ontology reasoning, KG representation learning on large multi-schema graphs, LLM-based entity matching and semantic harmonization, and candidate generation and re-ranking for terminology mapping.

9 Conclusion

MeSH-SNOMED-15K provides a realistic, large-scale benchmark for evaluating EA systems in a biomedical heterogeneous KG setting. By embedding 15,000 gold alignments within complete ontological contexts, the benchmark reflects the complexity of real-world terminology integration. The resource is released publicly to support research on EA, KG representation learning, and semantic interoperability in the biomedical domain.

We encourage researchers to use this benchmark to develop and evaluate alignment approaches that can reason across schema mismatch, lexical variation, and structural heterogeneity. We further encourage publication of methods, system descriptions, and ablation studies that help the community understand which signals (lexical, structural, semantic, learned) are most informative for cross-ontology EA.

Ethical Considerations

This benchmark aggregates and redistributes data from external biomedical resources. Users are responsible for ensuring that their use complies with the original providers' terms and conditions. MeSH is maintained by the U.S. National Library of Medicine and is freely available. SNOMED CT has jurisdiction- and license-specific usage conditions; users should consult SNOMED International and relevant NLM guidance before redistribution or commercial use.

The benchmark is not intended to serve as a substitute for clinically validated terminology resources. Any downstream application in clinical settings should involve appropriate domain experts and compliance with relevant regulatory standards.

Acknowledgments

We thank the U.S. National Library of Medicine for maintaining and providing access to MeSH, and SNOMED International for developing and maintaining SNOMED CT. We also thank the broader biomedical informatics community for foundational work on terminology integration and KG representation.

References

[Bodenreider, 2004] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 2013.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[Chen *et al.*, 2026] Shuai Chen, Zheng Yuan, Qian Zhang, Wei Hua, Jiuxin Cao, and Xun Huang. NeuSymEA: Neuro-symbolic entity alignment via variational inference. *Advances in Neural Information Processing Systems*, 38:92542–92564, 2026.

[Cheng *et al.*, 2025] Jing Cheng, Cheng Lu, Liang Yang, Ge Chen, and Fan Zhang. EasyEA: Large language model is all you need in entity alignment between knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20981–20995, 2025.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[Dong *et al.*, 2017] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 135–144, 2017.

[Euzenat *et al.*, 2008] Jérôme Euzenat, Adrian Mocan, and François Scharffe. Ontology alignments: an ontology management perspective. In *Ontology Management: Semantic Web, Semantic Web Services, and Business Applications*, pages 177–206. Springer US, Boston, MA, 2008.

[Hamilton *et al.*, 2017] William Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2017.

[Lenzerini, 2002] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246, 2002.

[Lipscomb, 2000] Colleen E. Lipscomb. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[Mendes *et al.*, 2012] Pablo Mendes, Max Jakob, and Christian Bizer. DBpedia: A multilingual cross-domain knowledge base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1813–1817, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[Navigli and Ponzetto, 2012] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[Rahm and Bernstein, 2001] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[Santos Sousa *et al.*, 2026] Gonçalo Santos Sousa, Ricardo Jorge Lima, and Cassia Trojahn. Survey on embedding methods applied to ontology matching. *ACM Computing Surveys*, 58(12):1–40, 2026.

[Schlichtkrull *et al.*, 2018] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607, Cham, 2018. Springer International Publishing.

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, 2007.

[Sun *et al.*, 2020] Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs. *arXiv preprint arXiv:2003.07743*, 2020.

[Vrandečić and Krötzsch, 2014] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[Wishart *et al.*, 2018] David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.