MedPath: Multi-Domain Cross-Vocabulary Hierarchical Paths for Biomedical Entity Linking

Nishant Mishra[1,2,*] **Wilker Aziz**[3] **Iacer Calixto**[1,2]
[1]Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, The Netherlands
[2]Amsterdam Public Health, Methodology, Amsterdam, The Netherlands
[3]ILLC, University of Amsterdam, The Netherlands
{n.mishra, i.coimbra}@amsterdamumc.nl w.aziz@uva.nl

Abstract

Progress in biomedical Named Entity Recognition (NER) and Entity Linking (EL) is currently hindered by a fragmented data landscape, a lack of resources for building explainable models, and the limitations of semantically-blind evaluation metrics. To address these challenges, we present MedPath, a large-scale and multi-domain biomedical EL dataset that builds upon nine existing expert-annotated EL datasets. In MedPath, all entities are 1) normalized using the latest version of the Unified Medical Language System (UMLS), 2) augmented with mappings to 62 other biomedical vocabularies and, crucially, 3) enriched with full ontological paths—i.e., from general to specific—in up to 11 biomedical vocabularies. MedPath directly enables new research frontiers in biomedical NLP, facilitating training and evaluation of semantic-rich and interpretable EL systems, and the development of the next generation of interoperable and explainable clinical NLP models.

1 Introduction

Health-related textual narratives abound in the healthcare domain and can be found for example in a patient's electronic health record (Johnson et al., 2023), in scientific research papers (PubMed, 2025), or in social media posts (Basaldella et al., 2020). Making sense of and integrating the medical concepts in these narratives is a complex task that requires in-depth domain knowledge. Named Entity Recognition (NER) and Entity Linking (EL) are two foundational tasks in clinical NLP whose main goal is *structuring the unstructured* (Röder et al., 2018). In NER, we wish to identify all mention spans of clinically relevant entities in some input text (e.g., all drug mentions in a clinical progress note; Nadeau and Sekine 2007). In EL, we go a step further and wish to link these mention spans against a *biomedical knowledge graph* (BioKG), where medical knowledge is structured and systematised (Kartchner et al., 2023a).

While there are datasets available to train and benchmark clinical and biomedical EL models, e.g., SNOMED-CT EL challenge for clinical notes (Davidson et al., 2025), BC5CDR for chemical–disease literature (Li et al., 2016), or COMETA for social media posts (Basaldella et al., 2020), existing datasets suffer from three critical issues. **Semantic fragmentation:** Datasets are anchored to a single BioKG (e.g., SNOMED-CT) or include texts from a single domain (e.g., clinical notes). This creates information siloes leading to models that will not generalise beyond their biomedical vocabulary/domain. **Explainability:** "Black-box" models increasingly face regulatory push-back in safety-critical domains (Huang et al., 2024; Ullah et al., 2024). There is a distinct lack of ground-truth data to train and evaluate interpretable clinical NLP models, especially for EL and similar tasks. **Superficial evaluation**: The performance of EL models is typically measured using "flat" metrics like precision, recall, and F1-score. While useful insofar, these metrics treat all errors as equal, e.g., incorrectly linking *congestive heart failure* to *myocardial infarction* (both types of heart disease) is penalized identically to linking it to *influenza* (a completely unrelated viral disease). In other words, such metrics fail to capture the semantic nuance of prediction errors and do not distinguish models that make more plausible mistakes (Falis et al., 2021; Amigó and Delgado, 2022; Plaud et al., 2024).

In this work, we introduce MedPath, a large-scale Entity Linking dataset that addresses all the above issues. MedPath's main features include:

- **Integration**: We harmonise and integrate *nine expert-annotated, curated datasets* covering clinical notes (ShARe/CLEF 2013, Suominen et al. 2013a; SNOMED-CT EL Challenge, Davidson et al. 2025), biomedical lit-

* Corresponding author

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2959–2978

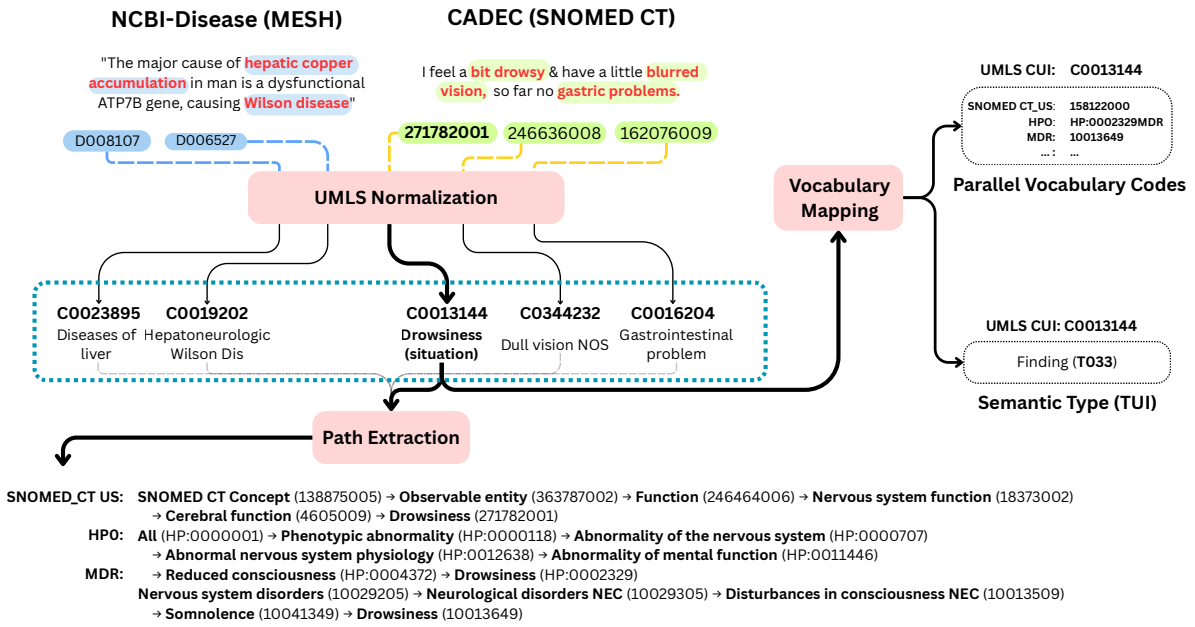

Figure 1: MedPath creation process. For illustration purposes, we show one example from two different datasets, and vocabulary mappings and path annotations for only one of the concepts, e.g., **C0013144 Drowsiness (situation).**

erature (BC5CDR, Li et al. 2016; NCBI Disease, Doğan et al. 2014; MedMentions, Mohan and Li 2019), drug-label prose (TAC 2017 ADR, Roberts et al. 2017a), and social media (CADEC, Karimi et al. 2015; COMETA, Basaldella et al. 2020), totalling 500,000+ mentions and 45,000 unique concepts.

- **Vocabulary normalization**: We normalise all entities, grounded in different BioKGs, to a canonical UMLS CUI (2025 AA; Bodenreider, 2004). We also map each code in one vocabulary to corresponding codes in all other covered vocabularies (up to 62 vocabularies in total).[1]
- **Hierarchical multi-vocabulary paths**: We annotate each concept with full hierarchical paths (i.e., from coarser to finer concepts) for the vocabularies that expose a usable API or where the full hierarchy is publicly available for download (11 biomedical vocabularies in total).

In Figure 1, we show the step-by-step process consisting of UMLS normalisation, vocabulary mapping, and hierarchical annotation generation. We show two running examples from the NCBI-Disease and CADEC datasets, respectively, clearly illustrating how our dataset addresses the semantic fragmentation and hierarchical annotation gaps.

In Sections 5 and 6, we introduce hierarchy-aware evaluation metrics (exact, ancestor-based, and hierarchy-based) and show initial experiments using MedPath on vocabulary-agnostic EL.

Finally, we release the codebase to reproduce MedPath under a permissive open-source licence at `https://github.com/mnishant2/MedPath`.

2 Related work

NER and EL are among the most important tasks in clinical NLP. NER helps us identify all mention spans of clinically relevant entities in some input text. In EL, the goal is to link these mention spans against a specific structured biomedical knowledge graph, e.g. SNOMED-CT.

Biomedical EL systems have evolved from lexical matchers like MetaMap (Aronson and Lang, 2010) and TaggerOne (Leaman and Lu, 2016), to embedding-based retrievers such as SapBERT (Liu et al., 2020) and BioSyn (Sung et al., 2020), and finally to generative architectures (De Cao et al., 2020; Xiao et al., 2023; Yuan et al., 2022).

Recent work has emphasized the critical need for interoperable biomedical NLP systems that are robust to vocabulary fragmentation. For instance, Neumann et al. (2019) and Beltagy et al. (2019) highlight the challenge of deploying models across datasets grounded in different BioKGs and domains. Similarly, Wadden et al. (2019) and Fries et al. (2022) underscore the brittleness of vocabulary-specific pipelines, which limit general-

[1]We use the terms *biomedical knowledge graph*, *controlled clinical vocabulary*, and *vocabulary* interchangeably.

Dataset / Benchmark	Release	NER	EL	#items	Unit	#datasets	#tasks	#vocabs	Vocab I&S	Path ann.	TUI ann.
GERBIL platform	2015	✗	✓	—	—	32	1 (EL eval)	5+	✗	✗	✗
MedMentions	2019	✓	✓	352k	mentions	1	2 (NER, EL)	1	✗	✗	✗
BLUE	2019	✓	✗	—	—	10	5 (NER, RE, QA)	0	n/a	✗	✗
CrossNER	2020	✓	✗	5,318	paragraphs	1	1 (NER)	0	n/a	✗	✗
Few-NERD	2021	✓	✗	491k	entities	1	1 (NER)	0	n/a	✗	✗
BLURB	2021	✓	✗	—	—	13	6 (LU tasks)	0	n/a	✗	✗
BigBio	2022	✓	✓	∼24M	examples	126+	13 categories	5+	P	✗	✗
BELB	2023	✗	✓	347k	mentions	11	1 (EL)	7	P	✗	✗
MedInst	2024	✗	✗	7M	instructions	133	133 (instr.)	8+	n/a	✗	✗
BRIDGE	2025	✓	✗	1.4M	samples	87	8 (clin. NLP)	0	n/a	✗	✗
MedPath (ours)	2025	✓	✓	∼512K	mentions	9	3†	62‡	✓	✓	✓

Table 1: Comparison of multi-dataset / benchmark resources. Symbols: ✓ = yes; ✗ = no; P = partial/limited support. #items counts gold-annotated units; "Unit" clarifies their type. **Vocab I&S**: cross-ontology vocabulary *integration & standardization*. **Path ann.**: ancestor / hierarchy paths provided. **TUI ann.**: UMLS Semantic Type identifiers attached. †: NER, EL and hierarchical EL. ‡: flat codes, and 11 vocabularies with full hierarchy

ization across real-world settings. Zu et al. (2024) proposes a collective entity linking method based on relationship paths, Moussallem et al. (2017), and Zwicklbauer et al. (2016) demonstrate a knowledge base agnostic entity linking system, while Jannet et al. (2014) introduced a novel metric for evaluation of hierarchical NER.

2.1 Task-specific corpora for NER and EL

Early biomedical NER efforts relied on single-vocabulary, single-domain corpora. Some notable examples include i2b2/VA 2010 linked against SNOMED CT (Uzuner et al., 2011), BC5CDR chemical–disease abstracts (Li et al., 2016) linked against MeSH, ShARe/CLEF 2013 clinical notes (Suominen et al., 2013b) linked against UMLS, and MedMentions (Mohan and Li, 2019) densely annotated PubMed abstracts against UMLS.

Later corpora widened the source spectrum, e.g. TAC2017 (Roberts et al., 2017b) that included drug-label prose (MedDRA) and COMETA (Basaldella et al., 2020) that had social-media posts (SNOMED CT). Individually, they cover a diverse range of data sources and formats, annotation guidelines, entity types, and native controlled clinical vocabularies. However, since they are anchored to a single vocabulary and a bespoke span guideline, it is difficult to harmonize model implementation and/or benchmarking across them, leading to a lack of interoperability.

2.2 Large-Scale Biomedical Benchmarks

Recently, efforts have been made to consolidate individual datasets into larger task-based benchmark suites, which aim to homogenize fragmented datasets in terms of volume and diversity (He et al., 2023; Rouhizadeh et al., 2024). **BLURB** (Gu et al., 2021a) is a composite dataset that bundles 13 biomedical language understanding tasks (sentence similarity, NER, QA, etc.), but it does not target entity linking, provides no unified concept identifiers, and focuses almost exclusively on literature. **BigBio** (Fries et al., 2022) is a wrapper that brings together over 120 different biomedical datasets, including NER and NED based corpora, and streamlines them into a common Huggingface schema, but keeps original IDs and offers no cross-vocabulary mapping. **BELB** (Garda et al., 2023) is another collated large-scale dataset that is specifically concerned with Entity Linking. It includes 11 different EL datasets with 7 knowledge bases into one shared leaderboard with thorough benchmarking, still evaluating "flat" CUI accuracy within the native KB of each corpus. **MedInst** (Han et al., 2024), the newest, LLM-focused entrant among large BioMedical benchmarks, repurposes 130+ datasets for LLM instruction-tuning, again without normalising concept identifiers or exposing vocabulary structure. **GERBIL** (Usbeck et al., 2015), while not a dataset, deserves a mention when talking about clinical NER and EL. It is a web-based benchmarking framework that provides a web API for EL evaluation but ships no harmonised data or hierarchy metadata.

In Table 1, we compare MedPath to well-known benchmarks across a breadth of capabilities. No existing benchmark simultaneously provides: (i) *cross-vocabulary integration and standardization* (from UMLS CUIs to codes in up to 62 KBs) and (ii) *explicit hierarchical paths in up to 11 vocabularies* for explainable model training and/or evaluation beyond flat metrics such as F1 score.

2.3 Hierarchy-aware Entity Linking

Entity linking and multi-label classification over BioKG-based label spaces require evaluation metrics that provide partial credit for semantically similar predictions rather than treating near-miss predictions as complete failures. The foundational work by Kosmopoulos et al. (2015) provides a comprehensive, unified framework for hierarchical evaluation, introducing LCA-based (Lowest Common Ancestor) metrics that construct minimal graphs connecting predicted and true labels via their LCAs, thereby avoiding over-penalization at deeper hierarchy levels. Earlier approaches (Kiritchenko et al., 2005, 2006) augment predictions with all ancestor classes, while the CoPHE metric (Falis et al., 2021) preserves count information during ancestor propagation, enabling detection of over- and under-prediction within label families. The H-loss framework (Cesa-Bianchi et al., 2006) charges loss only for the first classification mistake along prediction paths, capturing the intuition that coarse-grained errors subsume fine-grained mistakes, though limited to tree-structured hierarchies.

Despite the rich hierarchical structures in biomedical terminologies such as UMLS (127 semantic types) and SNOMED-CT (364K concepts in DAG structure), hierarchical evaluation remains notably absent from entity linking assessment. Kartchner et al. (2023b) demonstrates that major biomedical entity linking datasets rely exclusively on flat metrics, basic accuracy, relaxed matching, and strict matching, with no use of hierarchical partial credit, despite vocabularies providing is-a, part-of, and definitional relationships that could inform evaluation. Pesquita et al. (2009) survey content-based semantic similarity measures extensively used in Gene Ontology applications, yet these remain underutilized in entity linking evaluation. Kosmopoulos et al. (2015) report that metric selection can fundamentally alter system rankings and reveal distinct error patterns such as over-/under-specialization and sibling confusion, which flat metrics treat identically but have vastly different downstream consequences in clinical decision support applications. MedPath addresses this evaluation gap by providing hierarchical path annotations for 11 vocabularies, in addition to flat single-concept identifiers, explicitly capturing the ancestor lineage from root to leaf concepts. This design facilitates granular analysis of hierarchical evaluation metrics and informed modeling choices. In our initial experiments, we employ basic hierarchical metrics, including ancestor and descendant accuracy, to illustrate the utility of MedPath for path-based evaluation.

3 Dataset Construction

3.1 Curation rationale and source datasets

We first conducted a comprehensive survey of biomedical and clinical corpora available from institutional, shared task, and open-source repositories. Our primary selection criterion was the presence of high-quality, expert-validated ground-truth annotations suitable for EL. To ensure the final resource would be a challenging testbed for model generalization, we also prioritized datasets that collectively offered maximum diversity in textual domains and semantic types.

The nine corpora selected through this curation process are detailed in Table 2. While not an exhaustive representation of the ever-evolving biomedical field, this collection constitutes a large-scale and domain-diverse resource. It spans a wide spectrum, from formal scientific literature and clinical notes to product labels and informal social media content. Overall, the unified corpus comprises over 5 million tokens, more than 500,000 entity mentions, and 45,000 unique concepts, all drawn from source datasets with permissive licenses for research use.

3.2 Annotation

MedPath was created with a four-stage automated pipeline. This process integrates the individual datasets with fragmented annotations into a single, cohesive, and multi-vocabulary benchmark.

Stage 1: Unification and Standardization The nine source corpora are published in disparate formats, including BRAT, PubTator, XML, and TSV. Our first step was to develop dataset-specific parsers to ingest these formats and convert them into a standardized JSON schema. This process also involved light text cleaning to remove artifacts (e.g., de-identification remnants, stray characters) while preserving the original annotations.

Stage 2: Canonicalization via UMLS Mapping To resolve semantic fragmentation, we normalized all concept IDs from their native BioKG to the latest Unified Medical Language System (UMLS 2025AA) release. For each mention, we first attempt to map from its vocabulary native ID directly

Dataset	Year	Domain / Source (docs)	Entity types	Ontology	Licence
SNOMED CT EL Challenge (Davidson et al., 2025)	2023	MIMIC-IV ICU discharge notes (300)	Disorder, Procedure, Drug, Device	SNOMED CT	PhysioNet-R-A
ShARe/CLEF 2013 (Suominen et al., 2013a)	2013	Hospital discharge notes (199)	Disorder, Procedure, Medication, Device	UMLS	PhysioNet DUA
Mantra GSC (English) (Kors et al., 2015)	2015	Patents, drug labels, abstracts (1 050)	16 UMLS semantic groups	UMLS subset	CC-BY-SA 4.0
BC5CDR (Li et al., 2016)	2016	PubMed abstracts (1 500)	Chemical, Disease	MeSH	CC-BY 3.0
NCBI Disease (Doğan et al., 2014)	2014	PubMed abstracts (793)	Disease	MeSH, OMIM	CC-BY 4.0
MedMentions (Mohan and Li, 2019)	2019	PubMed abstracts (4 392)	Any UMLS concept	UMLS	CC-BY 4.0
TAC ADR 2017 (Roberts et al., 2017a)	2017	FDA Structured-Product Labels (200)	ADR, Drug	MedDRA, RxNorm	Public domain
CADEC (Karimi et al., 2015)	2015	Patient forum posts (1 250)	ADE, Drug	MedDRA, SNOMED CT	Ask-A-Patient T&C
COMETA (Basaldella et al., 2020)	2020	Reddit + Twitter posts (20 000)	Symptom	SNOMED CT	CC-BY-NC

Table 2: **MedPath core datasets**. Detailed statistics (mentions, CUIs, path depth) appear in Table 3.

to a UMLS Concept Unique Identifier (CUI) using a dictionary created from the UMLS database. If this fails, we fallback first to an exact match and then to a semantic containment heuristics.[2] This fallback strategy could introduce noise in the annotations, but the information loss from discarding these examples is a decrease in 2.5% in the number of unique mentions (2.13% from exact match, 0.37% from semantic containment) and 1.15% in the number of unique concepts (0.81% exact match, 0.35% semantic containment). In absolute numbers, we have 513,218 mentions / 44,259 unique concepts (CUIs) including examples mapped via exact match and semantic containment, and 500,384 mentions / 43,396 unique concepts after excluding these examples.

Stage 3: Multi-level Semantic Enrichment With a canonical CUI for each mention, we added two further layers of semantic information. First, we extracted the corresponding Semantic Type (TUI) for each unique CUI, providing a high-level categorization for every entity that was used in our initial experiments (see Section 5). Second, to enable interoperable, vocabulary-agnostic research, we mapped each CUI to its parallel concept identifiers in other major biomedical vocabularies, leveraging the atom-level information within UMLS.

Stage 4: Hierarchical Path Extraction The final and arguably most important stage of our pipeline was the extraction of full hierarchical paths, whereby we provide a data structure encoding rich hierarchical information to enable novel applications. This was a multi-step process. First, we identified the top 25 most frequently represented vocabularies in the datasets we use (Table 3) and determined which of them possessed both *a formal hierarchical structure* and *an accessible native taxonomy* (via public API or downloadable files). This process yielded 11 target vocabularies for path extraction. All the vocabularies we surveyed are listed in the Appendix D.

Next, we developed a custom path extraction method for each of these 11 vocabularies. This involved creating bespoke extractor modules that respect the unique structure of each vocabulary (e.g., interpreting different relationship types such as `is-a` relations in SNOMED CT or tree numbers in MeSH). For each concept, the extractor iteratively traverses the hierarchy from the entity to its more general/parent terms until either a root node is reached or no parent node can be found. The process was designed to be exhaustive, capturing and storing all possible paths for concepts that exist in multiple inheritance structures, i.e., when a vocabulary allows for more than one parent per node. Wherever presented with a choice, the latest version of the vocabulary was selected during this step. To ensure scalability and robustness, the implementation included several technical optimizations, such as result caching, filtering of inactive or obsolete codes, and robust API callback handling. This final stage produced a total of 573,786 distinct hierarchical paths for the 44,259 unique concepts.

3.3 Data Schema

An illustrative example of our multi-layered annotation schema is presented in Figure 1, with the

[2]Exact match: we string match between the mention text against each concept name in the UMLS term dictionary. If no exact match is found, we check for bidirectional substring containment (i.e., `X in Y` or `Y in X`). We use all concept names and synonyms available for each CUI, and choose the closest match based on token overlap and length similarity.

full JSON specifications detailed in Appendix A.2.

3.4 Data Quality Validation

We had a strict requirement to select only datasets with clinical expert oversight involved in the data curation. Moreover, we implemented multiple layers of checks and validations at each stage of the automated workflow to ensure data quality. The only stage where automatic heuristics may introduce noise is in mapping from source vocabulary codes to UMLS CUIs, and only when no mapping of native ID-to-CUI exists. We explain how we address this in Section 3.2, Stage 2. In short: possibly noisy examples mapped via exact match and semantic containment are clearly labelled in MedPath, meaning that users can either filter them out and have a noise-free dataset, or use them in case their use-case allows.

3.5 Availability and Update Strategy

In the interest of reproducibility, we release our complete annotation pipeline, data processing scripts, and evaluation code under a permissive open-source license. However, several of the constituent datasets and source vocabularies that form MedPath are protected by their own licenses or data usage agreements (DUAs) and cannot be redistributed directly. In such cases we provide detailed instructions and scripts that allow researchers who have obtained the necessary permissions from the original data providers to apply our pipeline and fully reconstruct MedPath. Keeping in mind that controlled clinical vocabularies are living resources which undergo frequent updates, we implemented the data preprocessing and annotation in a way that MedPath's scripts are easily compatible with any version of the various resources used (e.g., UMLS, SNOMED, MedDRA, etc). For example, we currently use *UMLS 2025 AA*, *SNOMED CT May 2025*, and *MedDRA 27.1*. However, one can easily adjust these versions by changing a single parameter/argument in the code to generate MedPath with future versions of these vocabularies.

4 Analysis

To characterize the properties of MedPath, we conducted a detailed statistical analysis. The following sections quantify the dataset's scale, its conceptual breadth, and the richness of its semantic and hierarchical annotations. More detailed analysis can be found in Appendix A.

Table 3: Biomedical entity linking datasets. Domain codes: SA =Scientific Abstracts; CN =Clinical Notes; SM =Social Media; DP =Drug Patents; MX =Mixed.

Dataset	Docs	Mentions	CUIs	TUIs	Domain
MedMentions	4392	352496	34631	126	SA
MIMIC-IV-EL	204	51574	5258	52	CN
TAC 2017 ADR	200	32585	3098	94	DP
BC5CDR	1500	29076	2487	69	SA
COMETA	20015	20015	3864	82	SM
CADEC	1186	9842	1256	68	SM
ShaRe/CLEF	291	8676	1372	34	CN
NCBI-Disease	792	7026	741	35	SA
Mantra-GSC	526	1928	1276	92	MX
Overall	**29,106**	**513,218**	**44,259**	**126**	**4**

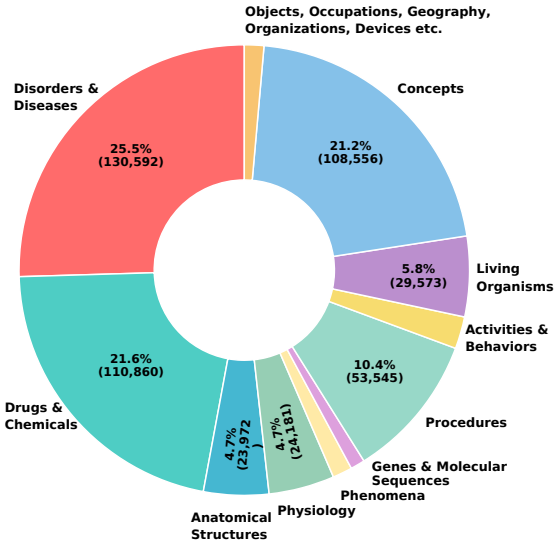

Figure 2: Semantic type distribution in MedPath.

4.1 Size and Genre Balance

The final harmonized corpus comprises over 5 million tokens and 513k expert-annotated mentions (Table 3). While MedMentions easily dominates the mentions count, the fact that it is itself a great mix of biomedical text is crucial for the diversity of our dataset. Social media posts add 20% of documents but only 6% of mentions, illustrating their short-form nature and motivating cross-length generalisation. For a more granular analysis, we categorized the source corpora into four primary domains based on their source as detailed in Table 2. Scientific literature (MedMentions, BC5CDR, NCBI-Disease) constitutes the largest portion, a reflection of its relative accessibility and textual density. The other three domains—clinical notes (MIMIC-IV-EL, ShARe/CLEF), social media (COMETA, CADEC), and drug labels and patents (ADR)—are well-balanced and contribute significant domain-specific richness. The Mantra-

GSC corpus, which contains text from Medline abstracts, drug labels, and patent claims, was classified as a mixed-domain dataset.

4.2 Concept Breadth and Semantic Diversity

Across datasets we observe a rich mix of concepts. Whereas all datasets combined have a mapped UMLS concept count of $\sim 54,000$, the unique mapped CUIs are $44,259$. This indicates a concept overlap of only about 20% across the datasets, underscoring the value of harmonization for creating a comprehensive benchmark that moves beyond the semantic scope of any single source. The semantic diversity of the corpus is equally broad. The annotated mentions span 126 of the 127 possible high-level UMLS Semantic Types (STYs). The most prominent semantic groups are *Disorders and Diseases* (25.5%), *Drugs and Chemicals* (21.6%), and *concepts* (21.2%). See Figure 2 for details.

4.3 Vocabulary Coverage and Hierarchy Insights

Figure 3 illustrates the extensive cross-vocabulary coverage of the resource, displaying the distribution of mentions from each source dataset across the 15 most frequent vocabularies. This visualization highlights the degree of interoperability achieved through our normalization pipeline. A key finding is the centrality of SNOMED-CT; seven of the nine datasets map over 80% of their mentions to SNOMED-CT concepts, even those with different native knowledge bases, demonstrating its comprehensive integration within UMLS.

Beyond simple coverage, we analyzed the structural properties of the 11 vocabularies for which we extracted full hierarchical paths. SNOMED-CT is the most information dense and structurally complex; 84% of its mapped concepts feature multiple inheritance paths, with an average of over 20 distinct paths per concept. The distribution of path depths, shown in Figure 4, reveals significant diversity across ontologies. SNOMED-CT exhibits a wide spread of path lengths, NCBI contains the deepest hierarchies on average, while others like ICD-9, ICD-10, and MedDRA have more concentrated path lengths of 3–5 levels, consistent with their defined structures. This variety in granularity, from complex directed acyclic graphs, like SNOMED CT, to simpler tree structures, confirms that MedPath is well-suited for developing and evaluating coarse-to-fine, hierarchy-aware models.

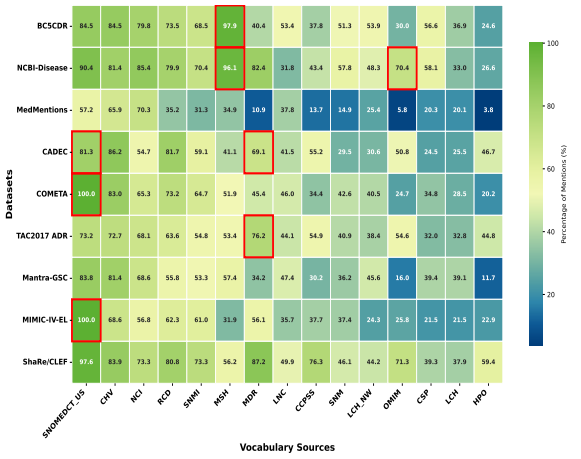

Figure 3: Vocabulary overlap heat map. Datasets' annotations using UMLS are not shown.

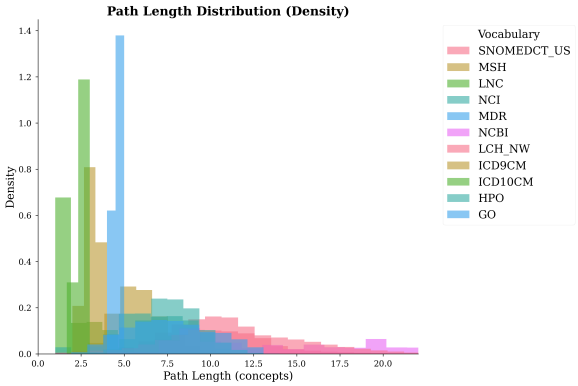

Figure 4: Histogram of lengths of entity hierarchical paths across different vocabularies.

5 Preliminary Experiments

We now provide initial experiments showcasing performance gains obtained using MedPath compared to training models on individual datasets and on datasets from a single domain. While MedPath can also be used for NER, our primary focus is on biomedical EL. We thus present EL experiments in this section and, for completeness, report on preliminary NER experiments in Appendix C.

5.1 Biomedical Entity Linking

We implement and benchmark a two-stage EL model adapted from the X-MEN library (Borchert et al., 2024).[3]

Retrieval We adopt two lightweight, dictionary-based retrieval methods implemented using: (i) TF-IDF-vectorizer operating over character 3-grams based retrieval, and (ii) embedding-based retrieval

[3] https://github.com/hpi-dhc/xmen/tree/main

with SapBERT (Liu et al., 2021). Both methods index a unified dictionary built from UMLS CUIs and their associated *names*, *synonyms*, and *lexical variants*. We include UMLS CUIs linked to any example from any source dataset in MedPath. Each test mention is treated as a query to retrieve the top-k most similar CUIs from this index.

Reranking The retrieved candidates from both TF-IDF and SapBERT are then passed to a cross-encoder model that performs reranking to identify the most relevant entity. Cross-encoders were trained on the top-32 generated candidates plus the gold entity (in case the gold entity is not in the top-32). We use a categorical cross-entropy loss function with regularization to optimize for ranking performance. The model can be initialized from various pretrained BERT encoders; we used `cambridgeltl/SapBERT-from-PubMedBERT-fulltext`. The model is trained to maximize top-1 accuracy, with the checkpoint that achieves the highest validation accuracy being selected for final inference.

Evaluation We use a test set consisting of all unique mentions with ground-truth CUIs across datasets. We report **accuracy@**k ($k = 1, 5, 32$) and **mean reciprocal rank** (MRR). To show the importance of semantically aware metrics for entity linking, we compute hierarchically-aware metrics that help assess both coarse and fine-grained performance of the models. For each mention whose gold CUI maps to one of 11 vocabularies with extracted hierarchies, which approximately covered all the mentions (98.7%), we evaluate whether any of the top-k predicted CUIs are (i) ancestors (**Ancestor@k**), (ii) descendants (**Descendant@k**), or (iii) part of a hierarchy in any way (**Hierarchy@k**) within the same BioKG, which could mean entities having common ancestors, or any hierarchy overlap with the test gold CUI, skipping top-3 levels from the root node of the vocabulary so we don't consider too general hierarchy match.

Vocabulary-Agnostic Entity Linking In our first experiment, we benchmark TF-IDF and SapBERT-based retrievers across the full test set using only surface form matching against the UMLS-derived CUI dictionary. This vocabulary-agnostic retrieval simulates realistic scenarios where the mention surface form may originate from different vocabularies or domains.

Table 4: Overall Entity linking performance. "CG" = candidate generator, "+RR" = with reranker. Per row: best score in CG underlined; best **score in CG+RR** bolded.

	CG		CG+RR	
Metric	**TF-IDF**	**SapBERT**	**TF-IDF**	**SapBERT**
Standard Metrics				
Acc@1	51.46%	48.12%	**80.84%**	79.02%
Acc@5	64.85%	65.44%	91.22%	**92.60%**
Acc@32	72.01%	73.68%	96.36%	**98.76%**
MRR@32	0.5756	0.5594	0.857	**0.861**
Hierarchical Metrics				
Hierarchy@1	68.60%	61.39%	85.40%	**86.24%**
Hierarchy@5	82.56%	80.66%	95.31%	**96.30%**
Ancestor@1	20.73%	18.74%	**24.80%**	24.68%
Ancestor@5	27.58%	25.16%	32.38%	**34.02%**
Descendant@1	20.10%	18.46%	23.45%	**23.74%**
Descendant@5	29.42%	25.39%	**32.98%**	32.75%

Ablations In our ablation, we systematically compare training strategies for EL and NER reranking under three regimes: **in-dataset**, whereby train/test come from disjoint splits from a same dataset; **in-domain**, whereby we train on all but one dataset within a domain and test on that held-out dataset; and **overall** whereby we train on the union of all datasets across all domains, i.e. MedPath. The unified UMLS mapping enables consistent label semantics across datasets, letting us 1) pool supervision in the overall setting, 2) measure generalization across datasets within a domain, and 3) compute comparable per-type macro summaries.

6 Preliminary Results and Discussion

6.1 EL Experiment Results

Table 4 shows the performance of the two EL candidate generation methods using both standard and hierarchical metrics. Overall, TF-IDF outperforms SapBERT in accuracy (Acc@1 = 51.5% vs. 48.1%) and MRR. This suggests that lexical overlap remains a strong signal in biomedical entity linking, and high coverage of the dictionary built for linking. SapBERT surpasses TF-IDF at k = 5 and k = 32 (Acc@5 = 65.4%, Acc@32 = 73.7%), indicating its strength in retrieving semantically similar or morphologically varied candidates not captured by character n-grams. Adding a reranker on top drastically improves all metrics for both candidate generators, with the SapBERT generator plus the SapBERT reranker outperforming TF-IDF plus SapBERT reranker on most metrics.

Hierarchy-aware evaluation shows that our

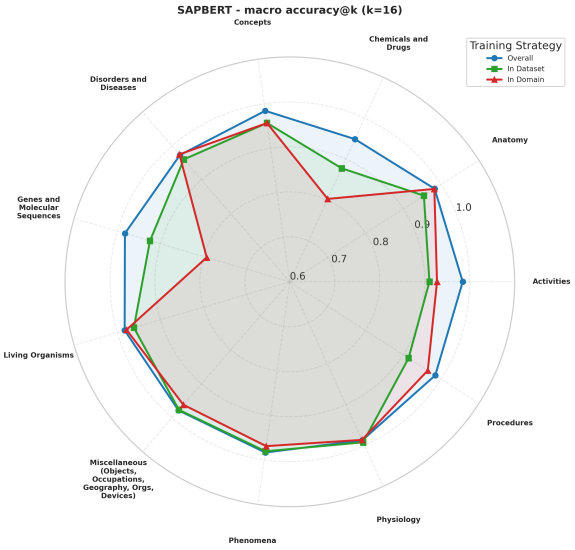

Figure 5: Figure showing EL performance in the three data settings

dataset enables a much richer analysis than exact CUI accuracy. While TF-IDF attains Acc@1 = 51%, its Hierarchy@1 jumps to 68.6%, indicating that an additional ∼17% of mentions retrieve a concept that is semantically related (i.e., a sibling, cousin, or ancestor). SapBERT exhibits a similar 13% gain (48→61%). Roughly 20% of errors are over-general (ancestor) and 20% are over-specific (descendant), highlighting granularity ambiguity rather than synonym mismatch.

In Figure 5, we see the reranker models performance when trained on a single dataset (in-dataset), on a single domain (in-domain), or on all datasets in MedPath (overall), in terms of macro-averaged acc@16 per semantic class. We observe that the model trained on MedPath consistently outperforms the other two settings, and sometimes by a large margin, thus validating the utility and information gain from collation and canonicalization in MedPath.

7 Conclusions and Future Work

In this work, we introduced MedPath, a large-scale resource for training and evaluating biomedical EL models that addresses three main limitations: semantic fragmentation, lack of explainability, and use of semantically-blind evaluation metrics. We integrate and harmonise nine diverse and expert-curated datasets across 4 domains with 513k mentions and 45,000 unique entities. We normalize all entity mentions to an up-to-date, canonical UMLS backbone, which means MedPath directly tackles the problem of data siloes.

MedPath includes mappings across up to 62 controlled clinical vocabularies and ∼ 575k hierarchical path annotations in 11 prominent clinical and biomedical knowledge graphs. It enables the training and evaluation of inherently explainable NER and EL models, and facilitates the development of truly diverse systems in terms of BioKG, a vital step towards achieving the interoperability required for real-world clinical deployment compatible with state-of-the-art generative AI methods. We release MedPath publicly at https://github.com/mnishant2/MedPath and hope to accelerate the development of biomedical NER and EL models that are more robust, trustworthy, and semantically aware.

Future work We believe MedPath opens research avenues in many directions. 1) We can go *beyond post-hoc explanation techniques and build inherently explainable models*. We envision using MedPath's hierarchical paths for training generative models that predict not only an entity's ID, but a mention's entire hierarchical path as a means to shed light on the model prediction process. 2) MedPath's vocabulary mapping across BioKGs allows for the construction of unified models that are *fluent in several medical vocabularies*. Future work may investigate multi-task learning setups where a single model is trained to make predictions across all 11 vocabularies. An EL system like this would be able to map a mention to its equivalent concepts in SNOMED-CT, MeSH, and ICD-10 all at once and would be a big step forward for model interoperability. 3) The hierarchical path annotations across 11 controlled clinical vocabularies provide a test-bed for the community to design and validate more sophisticated hierarchical evaluation metrics that can measure errors that encode domain-specific semantics within and across BioKGs. 4) Furthermore, hierarchical paths allow for fine-grained error analysis including answering questions like *'Do models more frequently confuse sibling concepts more than distant ones?'* or *'At what depth of the hierarchy do models begin to fail?'* 5) Additional applications of MedPath could include knowledge graph generation, and pre-training and fine-tuning LLMs so that LLMs are more factually grounded in established medical knowledge.

Limitations

While MedPath represents a significant step towards more diverse resources for biomedical NLP, we highlight a few limitations that users and researchers should be aware of.

Diversity Although we merged nine corpora to achieve broad domain diversity, this collection is not exhaustive and does not represent the universe of biomedical and clinical text. Models trained on MedPath may not generalize well to text from under-represent sources, e.g., clinical notes from other Electronic Health Record (EHR) systems or from specialized sub-disciplines. MedPath also only covers English datasets and does not address the needs of multilingual research in this domain.

Annotation issues MedPath's scale necessitated a largely automated pipeline for path extraction and entity normalization. While we employed state-of-the-art tools and devised a stringent methodology, with validation and statistical analysis, we did not perform expert, clinical validation of the new layers of annotation and mappings added. Moreover, there may be errors inherited from the original datasets, e.g., incorrect entity links for highly ambiguous mentions, missing nested mentions, missing or misaligned mappings, incorrect/outdated codes. Manual verification of mentions was not feasible. While we make the script with the annotation pipeline available to ensure transparency, users must be aware that the annotations reflect these limitations.

Versioning and updates Another potential area of concern and a well-known challenge is the ever-evolving nature of medical knowledge bases. Our annotations—UMLS CUIs, semantic types, and especially hierarchical paths—are tied to particular versions of the underlying ontologies, e.g., UMLS 2025 AA, MedDRA 27.0, SNOMED CT US March 2025. Biomedical knowledge is, however, not static; these terminologies are continually updated, with concepts being added, deprecated, or redefined. To this end, MedPath should be considered a high-fidelity snapshot at a particular point in time. As time passes, some paths will become outdated, and new concepts will not be represented, which may affect the resource's utility in the long term without periodic updates. Although MedPath's codebase makes it very easy to use a future version of a BioKG already in our pipeline, changes that break backward compatibility can still be an issue. Moreover, adding novel BioKGs would require researchers and other users to contribute to MedPath's codebase.

Ethical Considerations

Data Licensing and Access Some datasets (like MIMIC-IV EL Challenge and ShARe/CLEF) have protective DUAs that do not allow redistribution. For these datasets, we provide annotation and mapping scripts, which can be run locally, under the assumption that the user has lawfully obtained the requisite raw data.

Privacy and De-identification Discharge summaries, clinical notes as well as social media posts were the only patient-facing corpora utilized in this study. They were released publicly in a de-identified format and cannot be re-identified through our processing methods.

Ontology licensing and distribution Controlled vocabularies subject to license restrictions are not redistributed, and thus scripts are provided which extract relevant paths and metadata, provided there exists a local install or relevant ontology sources.

Potential Biases The corpus inherits biases from constituent datasets, including geographic bias from US-centric hospital systems and linguistic bias from the predominantly English corpus. These factors should be taken under consideration when interpreting model performance and deploying systems built from derived models.

References

Enrique Amigó and Agustín Daniel Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Annual Meeting of the Association for Computational Linguistics*.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American medical informatics association*, 17(3):229–236.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

Florian Borchert, Ignacio Llorca, Roland Roller, Bert Arnrich, and Matthieu-P Schapranow. 2024. xmen: a modular toolkit for cross-lingual medical entity normalization. *JAMIA Open*, 8(1):ooae147.

Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. 2006. Incremental algorithms for hierarchical classification. *Journal of Machine Learning Research*, 7:31–54.

Rory Davidson, Will Hardman, Guy Amit, Yonatan Bilu, Vincenzo Della Mea, Aleksandr Galaida, Irena Girshovitz, Mikhail Kulyabin, Mihai Horia Popescu, Kevin Roitero, Gleb Sokolov, and Chen Yanover. 2025. Snomed ct entity linking challenge. *Journal of the American Medical Informatics Association : JAMIA*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inform*, 47:1–10.

Matúš Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. 2021. Cophe: A count-preserving hierarchical evaluation metric in large-scale multi-label text classification. *arXiv preprint arXiv:2109.04853*.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, and 1 others. 2022. Bigbio: A framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35:25792–25806.

Samuele Garda, Leon Weber-Genzel, Robert Martin, and Ulf Leser. 2023. Belb: a biomedical entity linking benchmark. *Bioinformatics*, 39(11):btad698.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021a. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021b. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Wenhan Han, Meng Fang, Zihan Zhang, Yu Yin, Zirui Song, Ling Chen, Mykola Pechenizkiy, and Qingyu Chen. 2024. Medinst: Meta dataset of biomedical instructions. *arXiv preprint arXiv:2410.13458*.

Zexue He, Yu Wang, An Yan, Yao Liu, Eric Y Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2023. Medeval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. *arXiv preprint arXiv:2310.14088*.

Guangming Huang, Yingya Li, Shoaib Jameel, Yunfei Long, and Giorgos Papanastasiou. 2024. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *Computational and structural biotechnology journal*, 24:362–373.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Mohamed Ben Jannet, Martine Adda-Decker, Olivier Galibert, Juliette Kahn, and Sophie Rosset. 2014. Eter: a new metric for the evaluation of hierarchical named entity recognition. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3987–3994.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *J Biomed Inform*, 55:73–81.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie S. Mitchell. 2023a. A comprehensive evaluation of biomedical entity linking models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2023:14462–14478.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie S. Mitchell. 2023b. A comprehensive evaluation of biomedical entity linking models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14462–14478, Singapore. Association for Computational Linguistics.

Svetlana Kiritchenko, Stan Matwin, and A. Fazel Famili. 2005. Functional annotation of genes using hierarchical text categorization. In *Proceedings of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics (BioLINK)*.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A. Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to

text categorization. In *Advances in Artificial Intelligence, 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006*, pages 395–406, Québec City, Québec, Canada. Springer.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. 2015. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865.

Robert Leaman and Zhiyong Lu. 2016. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32 18:2839–46.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. _eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/48983216/bioinformatics_36_4_-1234.pdf.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

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. *Preprint*, arXiv:1902.09476.

Diego Moussallem, Ricardo Usbeck, Michael Röeder, and Axel-Cyrille Ngonga Ngomo. 2017. Mag: A multilingual, knowledge-base agnostic and deterministic entity linking approach. In *Proceedings of the 9th Knowledge Capture Conference*, K-CAP '17, New York, NY, USA. Association for Computing Machinery.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Catia Pesquita, Daniel Faria, André O Falcão, Phillip Lord, and Francisco M Couto. 2009. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):e1000443.

Roman Plaud, Matthieu Labeau, Antoine Saillenfest, and Thomas Bonald. 2024. Revisiting hierarchical text classification: Inference and metrics. *ArXiv*, abs/2410.01305.

PubMed. 2025. Pubmed. [Online; accessed 28-July-2025].

Kirk Roberts, Dina Demner-Fushman, and Joseph M Tonning. 2017a. Overview of the tac 2017 adverse reaction extraction from drug labels track. In *TAC*.

Kirk Roberts, Dina Demner-Fushman, and Joseph M Tonning. 2017b. Overview of the tac 2017 adverse reaction extraction from drug labels track. In *TAC*.

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. Gerbil–benchmarking named entity recognition and linking consistently. *Semantic Web*, 9(5):605–625.

H. Rouhizadeh, Irina Nikishina, A. Yazdani, A. Bornet, Boya Zhang, Julien Ehrsam, C. Gaudet-Blavignac, Nona Naderi, and Douglas Teodoro. 2024. A dataset for evaluating contextualized representation of biomedical concepts in language models. *Scientific Data*, 11.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. *arXiv preprint arXiv:2005.00239*.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. 2013a. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hanna Suominen, Sanna Salantera, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Angus Roberts, Liadh Kelly, Lorraine Goeuriot, Diego Martinez, Guido Zuccon, and 1 others. 2013b. Overview of the share/clef ehealth evaluation lab 2013. In *CLEF (Working Notes)*.

Ehtesham Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajeev Singh. 2024. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology: a recent scoping review. *Diagnostic Pathology*, 19(1):43.

Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, and 1 others. 2015. Gerbil: general entity annotator benchmarking framework. In *Proceedings of the 24th international conference on World Wide Web*, pages 1133–1143.

Ozlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. i2b2/va 2010 nlp challenge. In *AMIA Annual Symposium Proceedings*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, Jian Pei, and Daxin Jiang. 2023. Instructed language models with retrievers are powerful entity linkers. *arXiv preprint arXiv:2311.03250*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, and 1 others. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition. *arXiv preprint arXiv:2504.00676*.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. *ArXiv*, abs/2204.05164.

Lizheng Zu, Lin Lin, Song Fu, Jie Liu, Shiwei Suo, Wenhui He, Jinlei Wu, and Yancheng Lv. 2024. Pathel: A novel collective entity linking method based on relationship paths in heterogeneous information networks. *Information Systems*, 126:102433.

Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016. Doser - a knowledge-base-agnostic framework for entity disambiguation using semantic embeddings. pages 182–198.

A Source Corpora and Schema

A.1 Source Corpora Details

Below, we provide details about the nine expert-annotated corpora that constitute MedPath.

MIMIC-IV SNOMED EL Challenge 2023 (Davidson et al., 2025) Includes 300 de-identified ICU discharge summaries richly annotated by two clinical experts with SNOMED CT disorder, procedure, drug, and device codes. Provides the largest publicly available gold-standard clinical EL dataset.

ShAReCLEF 2013 (Suominen et al., 2013b) Part of an eHealth evaluation shared task, contains 199 hospital notes from Beth Israel hospital, double-annotated by clinical trainees and adjudicated by a senior MD for disorders, procedures, medications, and devices with UMLS CUIs.

Mantra GSC English (Kors et al., 2015) A multilingual dataset with 1,050 snippets drawn from patents, EU drug labels, and PubMed abstracts, covering 16 UMLS semantic groups annotated by three biomedical linguists.

BC5CDR (Li et al., 2016) A popular BioNLP benchmark dataset, it contains 1,500 PubMed abstracts with exhaustive Chemical and Disease spans normalised to MeSH, manually annotated by a team of 3 biocurators.

NCBI Disease (Doğan et al., 2014) A relatively smaller dataset with 793 abstracts focusing exclusively on diseases, mapped to MeSH 2012 tree numbers. It was annotated by three biology graduate students.

MedMentions (Mohan and Li, 2019) The biggest dataset in MedPath in terms of scale, contains 4,392 PubMed abstracts with mentions linked to any of 3.2 million UMLS 2017AB concepts with no type restrictions. As the broadest coverage literature corpus, it provides scalability and long-tail concept retrieval. It was triple-annotated by seven life science graduate students.

TAC 2017 ADR (Roberts et al., 2017b) One of the most mention-dense and token-rich datasets in MedPath, it has 200 FDA Structured-Product Labels annotated for adverse-reaction spans (MedDRA linked) and drug names (RxNorm linked). It was also manually annotated by two pharma-safety scientists and reviewed by NIST.

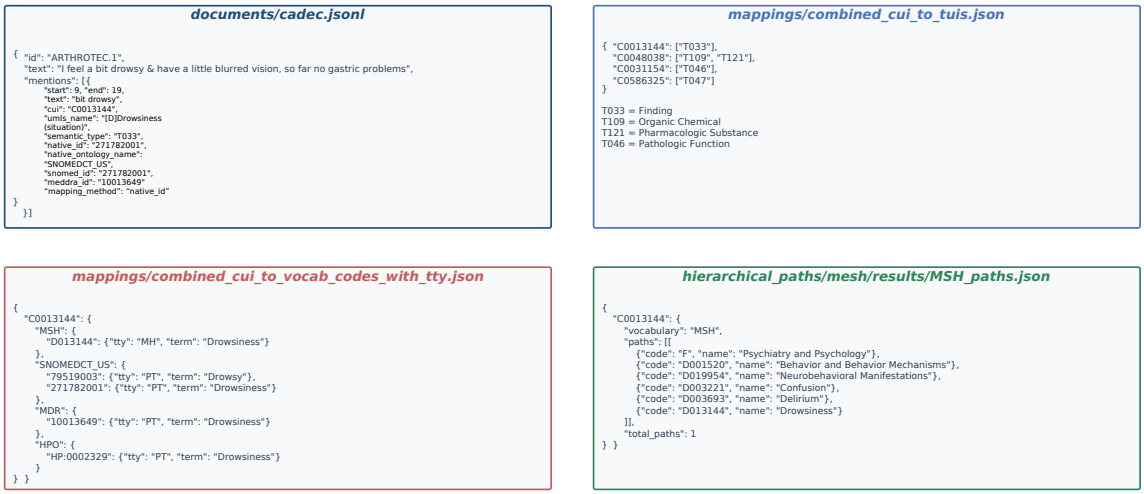

Figure 6: An example showing the schema of the proposed dataset, which shows the four components (clockwise): (i) the preprocessed document data with annotations and CUI mappings, (ii) the semantic type mapping, (iii) the cross-vocabulary mappings, and (iv) the hierarchical ontological paths.

CADEC (Karimi et al., 2015) One of the datasets from the social media/free text domain. It contains 1,250 Ask-a-Patient forum posts, labelled for patient-reported ADEs and drugs, and normalised to MedDRA, annotated in two steps by nurses and a biomedical ontologist.

COMETA (Basaldella et al., 2020) Social media-based dataset with 20,000 Reddit/Twitter posts with symptom spans mapped to SNOMED CT 2019 version. It was annotated by five trained crowdsourced annotators and adjudicated by an MD.

A.2 Dataset Schema

To ensure both ease of use and computational efficiency, MedPath is distributed across several files, each with a distinct purpose. The primary data is provided in a standardized JSON format, containing the source documents and a list of all annotated mentions with their character offsets, original concept IDs, and canonical UMLS CUIs. Supplementary annotations are provided in separate, optimized formats. Mappings from each unique CUI to its corresponding Semantic Type (TUI) and its parallel codes in other vocabularies are stored in simple key-value files. The core hierarchical annotations are structured as per-vocabulary lists of linear *Root* $\rightarrow$ *Leaf* chains, each containing the codes and names for each concept along the path. Figure 6 demonstrates the final schema of the dataset.

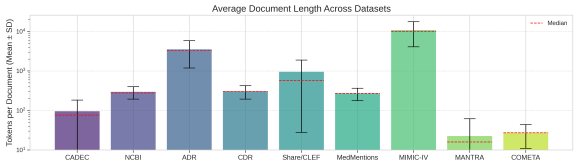

Figure 7: The mean and median document length for each dataset, shown in terms of BERT tokens.

B Additional Data Analysis

In this section, we present further analyses.

B.1 Document length

As illustrated in Figure 7, the document lengths vary considerably across the source corpora. Clinical notes (MIMIC-IV) and drug labels (ADR) feature the longest documents, whereas snippets from patents and abstracts (Mantra-GSC) and social media posts (COMETA) are significantly shorter.

B.2 Semantic type distribution

To visualize the contribution of each source dataset to the overall semantic diversity, we present a Sankey diagram in Figure 8. This plot depicts the flow of mentions from each source dataset to the 15 most frequent semantic type categories, confirming that MedMentions provides the broadest coverage across all categories.

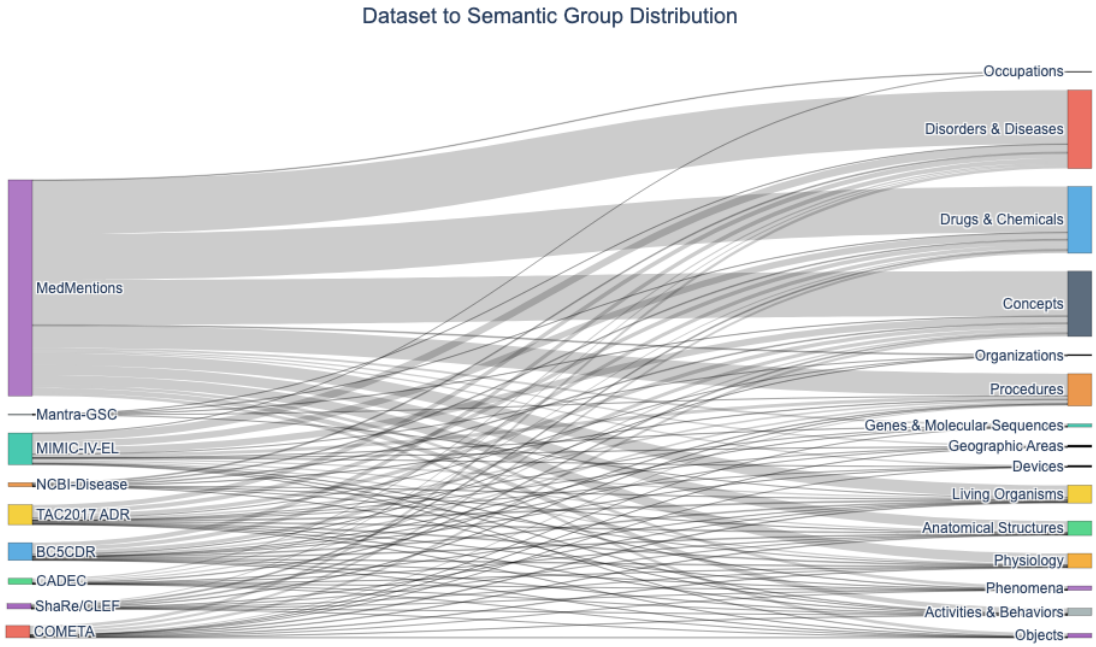

Figure 8: Contribution of each source dataset to concepts belonging to 15 major semantic type categories.

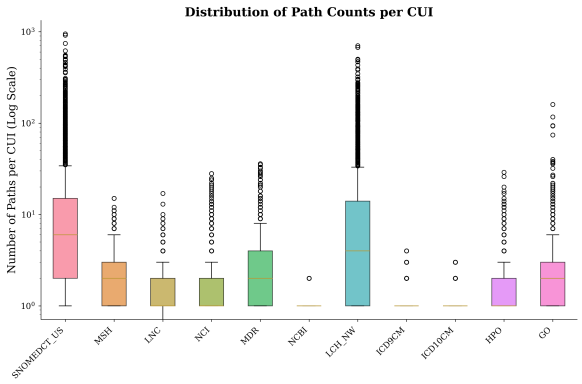

Figure 9: Number of paths per CUI in each vocabulary.

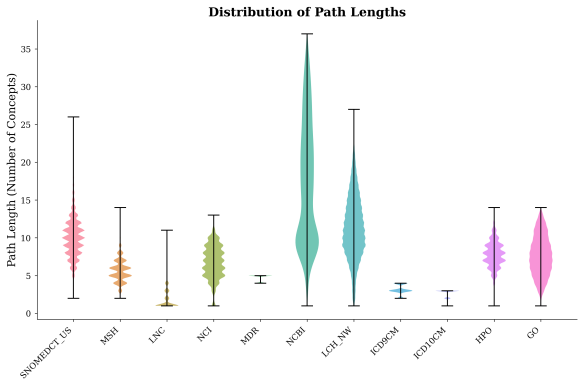

Figure 10: Path length distribution per vocabulary.

B.3 Ontology Path Statistics

Number of Paths The number of hierarchical paths per concept differs significantly across vocabularies, as shown in the boxplots in Figure 9. SNOMED CT and LCH_NW exhibit the largest variance, whereas vocabularies such as NCBI and ICD are largely monohierarchical.

Path Length Distribution The violin plot in Figure 10 illustrates the distribution of path lengths. On average, NCBI has the deepest paths, while MedDRA shows the least variance, consistent with its well-defined five-level hierarchy.

C Benchmark and Baseline Experiments

This section provides additional details on our experimental setup and preliminary results.

C.1 Named Entity Recognition (NER) Setup

We cast NER as token-level sequence labeling with a BIO scheme over chunks of our documents. Mentions are judged correct only if span boundaries and types match exactly.

Metrics We calculate strict and lenient micro-F1 scores per class and overall. We also calculate span-detection performance regardless of the predicted entity type.

Data Preprocessing First, a document was segmented into chunks of 512 characters with a 128-character sliding window. Mention offsets were recalculated relative to each chunk. Source domain and dataset information were preserved to facilitate ablation studies. An example of the final JSON format is shown below:


```
{
  "chunk_id": "227508_0",
  "source_dataset": "cdr",
  "source_domain": "abstracts",
  "text": "Naloxone reverses the
  antihypertensive effect of clonidine...",
  "entities": [
    {
      "start": 0,
      "end": 8,
      "label": "CHEM",
      "text": "Naloxone",
      "cui": "C0027358",
      "original_start": 0,
      "original_end": 8
    },
  ],
  "entity_types": ["CHEM", "DISO", "MISC"],
  "chunk_start": 0,
  "chunk_end": 512,
  "doc_length": 1135
}
```


Data Split and Labels For datasets with predefined splits, we retained them. For those without, we created a 50/10/40 train/dev/test split. If only a train/test split existed, a 10% dev set was carved out from the training data. Models were trained on 11 high-level semantic type classes derived from UMLS Semantic Groups (see Figure 2).

Experiment Paradigms To demonstrate robustness, get insights into the dataset composition, and highlight the value of cross-domain unification, we run four types of experiments.

- **Full-Mix**: Train on the union of all training splits; evaluate on the union of all test splits.
- **In-Domain**: Build train/dev/test splits per single dataset.
- **Leave-One-Dataset-Out (LODatO)**: Hold out one dataset for testing, and train on all other datasets.
- **Leave-One-Domain-Out (LODomO)**: Hold out all datasets from one domain for testing (clinical / literature / social / label), training on all datasets from the remaining domains.

Models For the **Full-Mix** setting, we fine-tune and evaluate five biomedical PLMs pretrained on different domains. These models are listed below:

- **GatorTron-base**: An encoder-only transformer model pre-trained on a large corpus of over 82 billion words from de-identified clinical notes and clinical trial publications, developed by the University of Florida (Yang et al., 2022).
- **ClinicalBERT**: A BERT model pre-trained on the MIMIC-III dataset, which contains de-identified health records, making it highly specialized for tasks on clinical notes (Huang et al., 2019).
- **PubMedBERT**: A BERT model pre-trained from scratch exclusively on biomedical literature, specifically 21GB of text from PubMed abstracts and full-text articles (Gu et al., 2021b).
- **BioBERT**: One of the first domain-specific BERT models, initialized from Google's BERT and continually pre-trained on a large-scale biomedical corpus including PubMed abstracts and PMC full-text articles (Lee et al., 2019).
- **BlueBERT**: BERT model pre-trained on a combination of biomedical (PubMed abstracts) and clinical data (MIMIC-III notes), designed to perform well on a diverse range of biomedical and clinical NLP tasks (Peng et al., 2019).
- **GliNER-BioMed** We also evaluated GliNER-BioMed (Yazdani et al., 2025), a task-specific NER model in a zero-shot setting. GliNER is a generative encoder-decoder model that takes natural-language class labels, along with the input sentence, and outputs spans of mentions belonging to those classes. For our evaluation, we used all variations of our semantic classes in natural language as potential class names to pass to GliNER. E.g., for Disorder (DISO), we passed {'disease', 'disorder', 'condition', 'syndrome', 'pathology', 'findings'} along with sentences, and then mapped the extracted entities to our class names for consistent comparison.

For all **ablation studies** (LODatO, LODomO), we use **PubMedBERT** because it is consistently strong across domains and classes, yet faster and lighter than GatorTron. This keeps computation manageable and isolates the effect of the data splits from that of the model size.

Hyperparameter Tuning For all the experiments that involved fine-tuning models: full mix, and the various ablations as described in C.1, we implemented a thorough hyperparameter tuning across a range of hyperparams (focal loss, crf layer, batch size, learning rate, weight decay, warmup ratio, early stopping) through randomized trials. The

Model	Strict F_1	Lenient F_1	EA F_1
Finetuned Models			
GatorTron-base	**0.663**	**0.746**	**0.760**
PubMedBERT	0.642	0.728	0.743
ClinicalBERT	0.615	0.713	0.726
BioBERT	0.623	0.717	0.730
BlueBERT	0.606	0.705	0.720
Zero-shot			
GLiNER-BioMed	0.365	0.482	0.524

Table 5: Micro-averaged NER test performance across all 11 semantic groups. **EA** F_1: Entity-agnostic F_1.

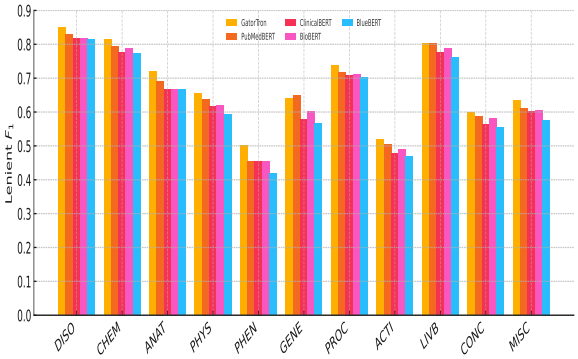

Figure 11: Performance of fine-tuned models in full mix setting across the 11 classes, shown using lenient F1.

best model in a full-mix setting, i.e., GatorTron-base, achieved the best performance with a linear-CRF layer, a base learning rate of 5e-5 with a CRF layer learning rate of 1e-5, a batch size of 32, 0.01 weight decay, and 0.1 warmup. Additionally, we used a 3x random oversampling to balance under-represented classes, along with a class-weighted loss function.

C.1.1 Preliminary NER Experiment Results and Discussion

Table 5 shows the overall performance of the NER models on our unified dataset. GatorTron performed best across all metrics, with PubmedBERT a close second. GliNER zero-shot performed poorly across most categories, except Disorders and Chemicals/Drugs. Comprehensive results across classes, domains, and datasets, along with observations from the ablation studies, are presented below.

Full Mix Figure 11 shows the per-class F1 performance for the five fine-tuned models in the Full-Mix setting.

Figure 12 breaks down the strict F1 performance by domain for all models, including the zero-shot GliNER-BioMed.

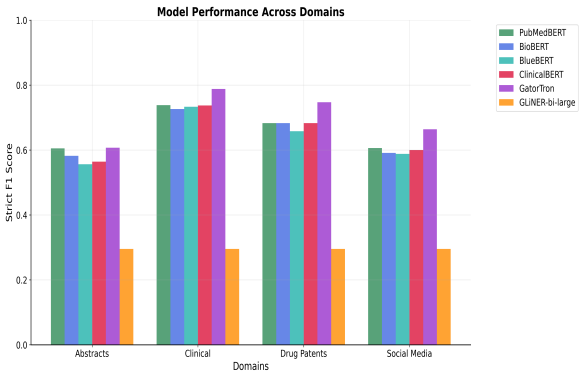

Figure 12: Performance of models across the four main domains, shown using strict F1.

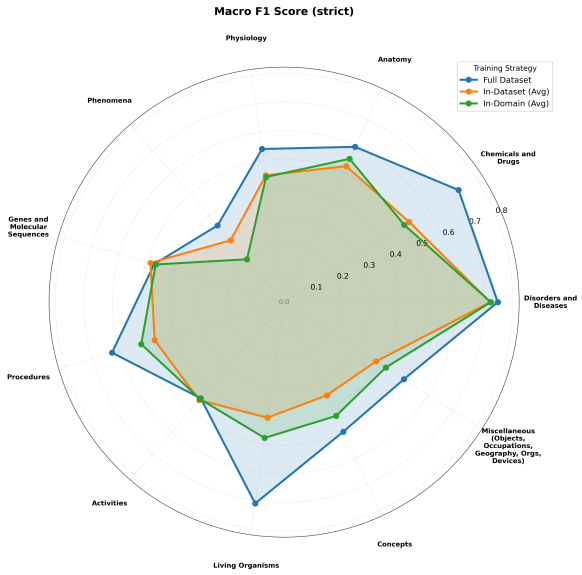

Figure 13: Figure showing the macro average performance over semantic types of the NER model in-domain, in-dataset, and a full mix setting

Ablations The radar chart in Figure 13 compares the average performance of models trained on the full mix versus those trained in-domain or in-dataset, demonstrating the clear benefit of MedPath. Figure 14 quantifies the performance impact of holding out each dataset and domain, with MedMentions and the abstracts domain showing the most significant impact due to their scale.

C.2 Additional Entity Linking Results

Here, we present the performance of the TF-IDF and SapBERT-based entity linkers across various verticals. Figure 15 shows the candidate-generation-based EL metrics for mentions across the nine datasets using the Acc@32 metric. We see

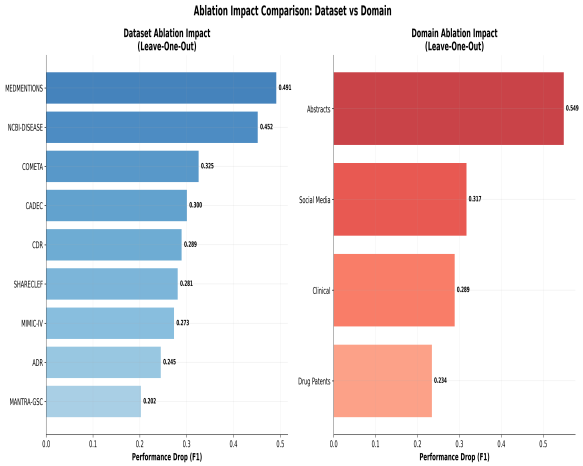

Figure 14: Figure showing the performance Δ in the LOO and LOODom ablation experiments

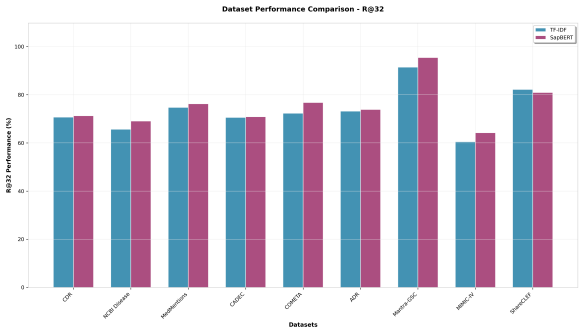

Figure 15: Performance of TF-IDF and SapBERT Candidate Generation across datasets

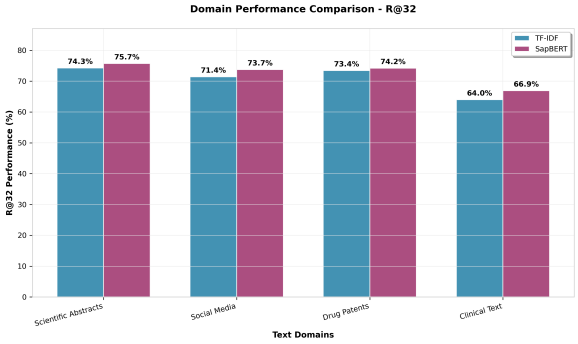

Figure 16: Performance of TF-IDF and SapBERT Candidate Generation across domains

Table 6: Candidate Generation performance per Entity Type by SapBERT and TF-IDF, **boldface** denotes highest R@32 and underline denotes highest R@1 per type

Entity Type	Count	TF-IDF		SapBERT	
		R@1	R@32	R@1	R@32
Disorder	61,054	55.9%	74.3%	52.9%	**76.5%**
Concept	32,094	54.9%	76.6%	51.9%	**77.7%**
Procedure	20,528	27.0%	58.2%	28.9%	**60.3%**
Chemical	19,216	55.4%	69.5%	49.2%	**70.2%**
Living Being	8,653	62.6%	81.0%	56.6%	**82.5%**
Anatomy	7,502	53.6%	73.8%	42.7%	**75.6%**
Physiology	7,175	51.9%	75.2%	44.6%	**75.7%**
Miscellaneous	5,045	41.3%	59.7%	38.3%	**62.3%**
Activity	3,751	56.4%	78.9%	59.1%	**79.4%**
Phenomenon	2,009	40.6%	63.3%	38.9%	**64.9%**
Gene	1,523	42.5%	64.5%	34.9%	**66.6%**

that SapBERT consistently outperforms TF-IDF by small margins. Mantra-GSC, ShaRE/CLEF, and MedMentions show the best performance, which follows logically from the fact that their original ground truth annotations were in UMLS (Table 2), and the dictionary we created utilized UMLS concepts.

Similarly, in Figure 16, we visualize the EL performance across the four domains in our dataset. The major takeaway is the underwhelming performance on clinical notes, suggesting these terms have multiple surface forms or lack vocabulary integration.

Table 6 details the performance of these two linkers over the various semantic groups. Living Beings and Activity mentions show the highest performance, while Procedure and MISC classes, including devices, occupation, and organization, show weak performance.

Figure 17 shows the final performance of the reranker trained on the candidate generators across three different ablation scenarios—In Dataset, In Domain, and Overall—is presented here. Performance was measured by calculating the macro-average of all metrics (Accuracy@k and MRR) for each semantic category's mentions. We consistently observe that the Overall performance surpasses both in-domain and in-dataset performance across various metrics and semantic types. These results provide strong evidence for the advantages of using a consolidated, canonicalized, uniform resource, such as MedPath, to enhance semantic richness and retrieval performance in Biomedical NER and EL.

D Vocabularies Glossary

Table 7 presents the top 25 most frequent BioKGs in MedPath, based on the unique CUIs that map to them. The table provides information on which of these BioKGs possess a hierarchical structure and whether their native hierarchy can be accessed through an API or download. As indicated by the highlighted rows, only 11 of these biokgs meet our criteria for extracting full hierarchical paths.

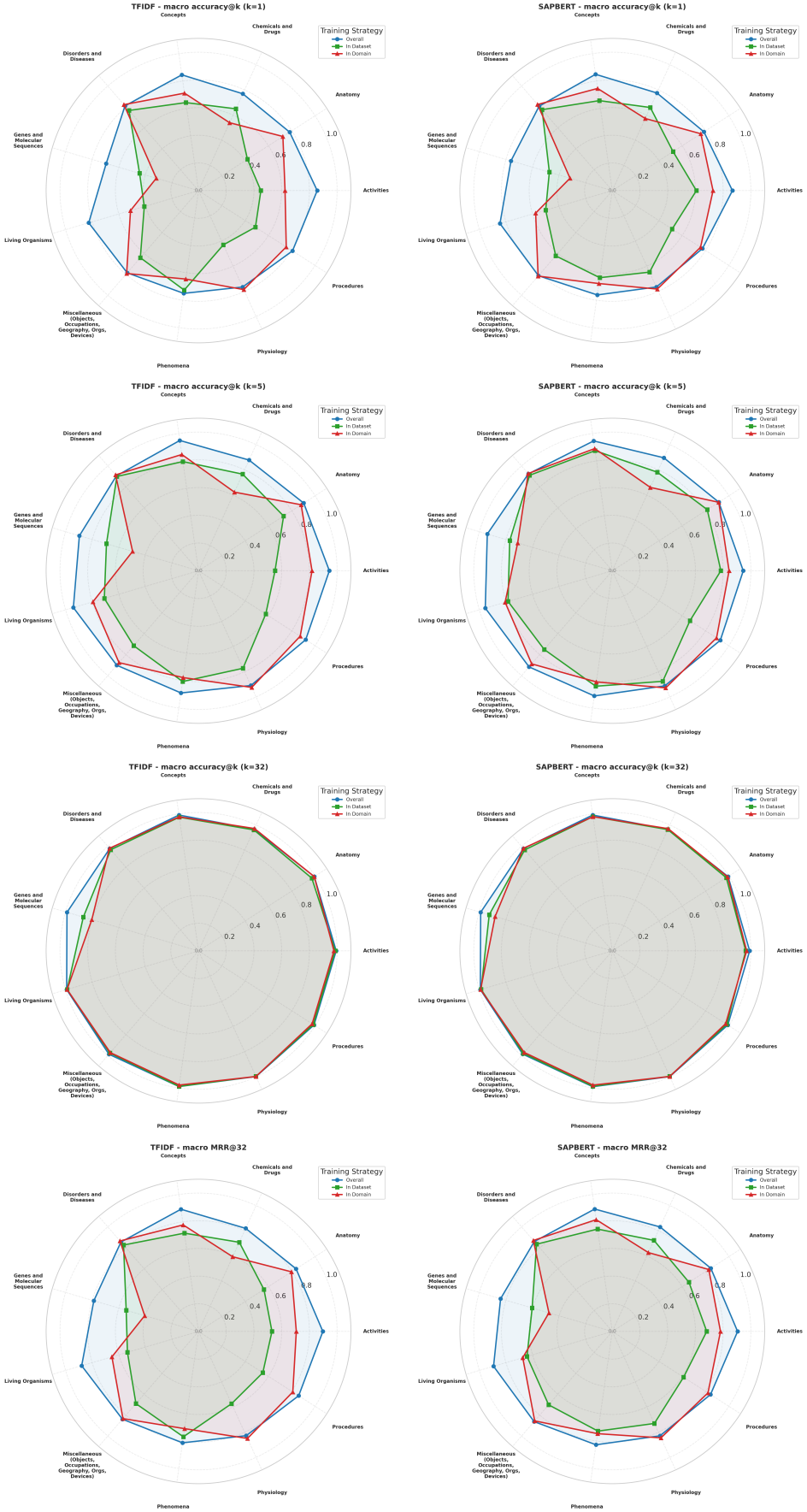

Figure 17: Entity linking performance across metrics (rows: Acc@1, Acc@5, Acc@32, MRR@32), averaged over all semantic-type , for all three training strategies. Left: TF-IDF+Reranker; Right: SapBERT+Reranker

Table 7: A comparative overview of biomedical and clinical vocabularies with full names. Counts and percentages are based on the final provided distribution across 41,619 unique CUIs. Vocabularies highlighted in green (prefixed with †) were used to extract hierarchical paths. **H.** = vocabulary has an internal Hierarchy; **H.A.** = Hierarchy Available/Accessible for path extraction..

Vocabulary	Unique CUIs	%	H.	H.A.	Notes
† **SNOMED CT** (Systematized Nomenclature of Medicine Clinical Terms)	23 261	55.89	✓	✓	Intl. clinical terminology. Requires free license for full use.
CHV (Consumer Health Vocabulary)	18 856	45.31	✗	✗	Maps consumer terms to professional terms. No standalone API.
† **NCI** (National Cancer Institute Thesaurus)	16 668	40.05	✓	✓	Comprehensive cancer ontology. Accessible via NCI's EVS and BioPortal.
† **MESH** (Medical Subject Headings)	13 812	33.19	✓	✓	Thesaurus for indexing literature. API and bulk data from NLM.
RCD (Read Codes)	13 435	32.28	✓	✗	UK primary care codes. Replaced by SNOMED CT.
SNMI (SNOMED International)	11 200	26.91	✓	✗	SNOMED Intl. v3.5. Superseded by SNOMED CT.
† **MDR** (Medical Dictionary for Regulatory Activities)	7 699	18.50	✓	✗	For adverse events reporting. Requires license.
† **LOINC** (Logical Observation Identifiers Names and Codes)	6 968	16.74	✓	✓	Laboratory/observation codes. Free with registration.
SNM (Systematized Nomenclature of Medicine 1982)	5 926	14.24	✓	✗	Obsolete SNOMED edition. Replaced by SNOMED CT.
† **LCH_NW** (Library of Congress Headings, NW Subset)	5 809	13.96	✓	✓	Northwestern Univ. subset for biomedical topics.
MEDCIN	5 653	—	✓	✗	Proprietary clinical terminology (Medicomp Systems).
CSP (CRISP Thesaurus)	4 868	11.70	✓	✗	Former NIH thesaurus. Now historical; available in UMLS.
OMIM (Online Mendelian Inheritance in Man)	3 694	8.88	✗	✓	No inherent taxonomy. API requires free registration.
LCH (Library of Congress Headings)	3 685	8.85	✓	✓	Broad multidisciplinary subject headings.
CCPSS (Canonical Clinical Problem Statement System)	3 662	8.80	✓	✗	Vanderbilt, 1999. Standard problem names; available via UMLS.
PSY (Thesaurus of Psychological Index Terms)	3 137	7.54	✓	✗	APA's thesaurus for PsycINFO indexing.
† **HPO** (Human Phenotype Ontology)	2 272	5.46	✓	✓	Open ontology for genetic phenotype annotation.
FMA (Foundational Model of Anatomy)	2 192	5.27	✓	✓	Extensive anatomy ontology with part-of hierarchy.
† **ICD-10-CM** (Intl. Classification of Diseases, 10th Rev, Clin. Mod.)	2 092	5.03	✓	✓	Clinical mod. of WHO's ICD-10. Maintained by CDC/NCHS.
RXNORM	2 042	4.91	✓	✗	Normalized drug nomenclature.
† **ICD-9-CM** (Intl. Classification of Diseases, 9th Rev, Clin. Mod.)	1 796	4.32	✓	✓	Legacy system, replaced by ICD-10-CM.
ICPC2ICD10ENG (ICPC-2 to ICD-10 Mapping)	1 726	3.96	✗	✗	Links primary care codes (ICPC-2) to ICD-10.
UWDA (Univ. of Washington Digital Anatomist)	1 608	3.86	✓	✗	Early anatomical ontology; superseded by FMA.
† **GO** (Gene Ontology)	1 378	3.31	✓	✓	Biological ontology (MF, BP, CC). Open access.
† **NCBI** (National Center for Biotechnology Information)	1 354	3.25	✓	✓	Suite of databases and tools via E-utilities API.