Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

Structure-Aware Contrastive Learning for Biomedical Embeddings: Bridging the Gap Between HPO and Clinical Literature

Jose Luis Mellina Andreu[1] , **Alejandro Cisterna García**[2] , **Juan Botía**[1]
[1]Department of Information and Communication Engineering, University of Murcia
[2]Department of Health Sciences and Technology (DHEST), ETH Zurich
{joseluis.mellinaa, juanbot}@um.es, acisterna@ethz.ch

Abstract

Large Language Models (LLMs) are extensively used at biomedical text processing but often fail to capture the complex, functional relationships encoded in expert knowledge graphs like the Human Phenotype Ontology (HPO). This "semantic gap" limits their utility in precision medicine tasks such as rare disease diagnosis, where distinguishing overlapping clinical presentations requires understanding underlying pathophysiological connections rather than just surface-level textual similarity. In this work, we propose a Neuro-Symbolic Alignment Framework that bridges this separation by integrating literature-mined specialized phenotypical descriptions with the ontological structure used as reference. Specifically, we augment phenotype representations with automatically selected text fragments from massive corpus of descriptions mined from scientific literature (PubMed), overcoming the typical data scarcity of standard ontology definitions. We define a new embedding adaptation procedure whose fine-tuning approach is guided by a novel "Disease-Overlap" similarity measure, which prioritizes clinical co-occurrence of phenotypes over taxonomic distance, and optimizes the embedding space using AnglE Loss to mitigate gradient saturation. Extensive evaluations show that our approach significantly outperforms state-of-the-art baselines, including SapBERT, on both intrinsic semantic correlation and practical downstream tasks, including synthetic patient disease ranking and solving real cases stored in Phenopacket, where our model achieves $\times 4$ top-1 accuracy than the previous best model.

1 Introduction

Accurate characterisation of patient phenotypes is crucial for diagnosing rare genetic diseases. After a clinical evaluation, the physician produces a medical report in which the phenotypes observed during that evaluation are noted and used as the basis to identify and suggest one or more possible genetic diseases (which is called differential diagnosis). A definitive genetic diagnosis is only obtained when a pathogenic variant associated with the corresponding disease and its related clinical phenotypes is identified in the genome. With more than 10,000 distinct rare disorders identified, clinicians rely heavily on standardized vocabularies, such as the *Human Phenotype Ontology* (HPO) [Robinson *et al.*, 2008; Köhler and others, 2020], to bridge the gap between clinical observations and the genetic diagnostic [Genetic Alliance and others, 2010]. Therefore, the application of LLM-related technologies to biomedical tasks in general, and to genetic diagnostic in particular, must be sustained in a precise knowledge and understanding of the terminology linked to phenotypes and diseases.

Large Language Models (LLMs) pre-trained on biomedical corpora, such as BioBERT [Lee *et al.*, 2019] and PubMedBERT [Gu *et al.*, 2021], are the state-of-the-art models when applied to biomedical information retrieval tasks. However, a significant *semantic gap* remains when applying these models to deep phenotyping, i.e., the task of describing the disease as completely and precisely as possible. While LLMs excel at capturing distributional semantics from unstructured text, they often fail to comprehend the strict, biologically grounded hierarchical relationships encoded in expert ontologies [Hao *et al.*, 2020]. For instance, a general-purpose biomedical LLM may recognize "cataract" and "glaucoma" as related eye diseases based on lexical co-occurrence, but it lacks the structural knowledge that links "iris coloboma" and "renal failure", two apparently unrelated phenotypes, as phenotypically associated manifestations of *Lowe Syndrome* [Lowe *et al.*, 1952].

While existing biomedical LLMs largely focus on entity linking (recognizing synonyms), our work specifically targets the optimization of embeddings for phenotype-disease association relationships, essential for differential diagnosis. Current approaches to bridge this gap typically rely on fine-tuned embeddings using the ontology's graph structure [Garnot and Landrieu, 2020] or phenotype definition-based contrastive learning [Wen *et al.*, 2025]. Yet, these methods face two critical limitations. First, ontologies suffer from *data scarcity* in terms of natural language: HPO terms are often defined by short, dictionary-like descriptions that fail to capture the lexical variability of real-world clinical reports. Second, standard topological similarity measures between pairs of terms in the ontology (e.g., Resnik [Resnik, 1995], Lin [Lin, 1998]) prioritize ontology graph distance over clinical co-occurrence

information on observed phenotypes, misaligning the embedding space for diagnostic tasks where functional association is potentially more critical than taxonomic proximity.

In this work, we propose a novel **Neuro-Symbolic Alignment Framework** that injects both ontological structure and rich clinical context into the latent space of LLMs. Our approach diverges from previous work in three key aspects:

1. **Literature-Mined Data Augmentation:** Instead of relying solely on static HPO definitions, we automatically generate a large training corpus of abundant and diverse phenotype descriptions mined from scientific literature (PubMed), exposing the model to the pragmatic and real usage of terms.
2. **Annotation-Based Supervision:** We guide the contrastive learning process using a "Disease-Overlap" similarity metric (see section 3.2) derived from the *Relative Best Pair* strategy [Gong *et al.*, 2018]. This forces the model to learn relationships between phenotypes based on their shared pathophysiology rather than mere graph distance.
3. **Rank-Aware Optimization:** We employ AnglE Loss [Li and Li, 2024] to optimize the vector space geometry, ensuring that phenotypes sharing a disease cause are pulled mathematically closer than those that merely sound alike, effectively translating clinical relevance into vector distance.

We evaluate our method against strong baselines, including SapBERT [Liu and others, 2021], on intrinsic semantic correlation tasks and extrinsic downstream applications such as Gene-Disease association prediction (GSC+) and literature-based phenotype retrieval (JAX). All the code is available in a GitHub repository and our best model is available in Hugging Face.

2 Related Work

2.1 Biomedical Language Models

The adaptation of BERT-based architectures to the biomedical domain has become the standard for clinical NLP. *BioBERT* [Lee *et al.*, 2019] demonstrated that pre-training on PubMed abstracts significantly improves performance on named entity recognition (NER) and relation extraction. *PubMedBERT* [Gu *et al.*, 2021] further advanced this by pre-training from scratch on biomedical text, overcoming the vocabulary mismatch inherent in models initialized on general-domain corpora. More recently, *SapBERT* [Liu and others, 2021] introduced a self-alignment pre-training scheme using the UMLS metathesaurus to cluster synonymous medical terms. While SapBERT successfully performs at entity linking (grouping synonyms), it does not explicitly optimize for the hierarchical or disease-association relationships required for differential diagnosis.

2.2 Semantic Similarity in Ontologies

Traditionally, semantic similarity between ontology terms has been quantified using Information Content (IC) approaches. *Resnik* [Resnik, 1995] defined similarity based on the IC of the Most Informative Common Ancestor (MICA). *Lin* [Lin, 1998] refines Resknik by applying a normalization to the measure. Although robust for graph analysis, these topological metrics ignore the *extensional* meaning of phenotypes, i.e., the set of diseases they annotate. Alternative approaches which account for this, such as those used in *Phenomizer* [Köhler and others, 2009] or *LIRICAL* [Robinson and others, 2020], utilize Bayesian frameworks or pre-computed statistical scores for diagnosis but do not generate embeddings that can be used in downstream tasks, e.g., model building pipelines. Our work bridges this traditionally independent approaches by using annotation-based similarity as a supervision signal for neural embeddings.

2.3 Knowledge-Enhanced Representation Learning

Recent efforts in "Neuro-Symbolic AI" seek to integrate Knowledge Graphs (KGs) into deep learning. Methods like *TransE* [Bordes and others, 2013] and *Node2Vec* [Grover and Leskovec, 2016] generate graph embeddings but discard the rich textual information of nodes. Conversely, text-based approaches like *CODER* [Yuan *et al.*, 2020] use contrastive learning to align medical terms with KG relations. However, CODER and similar models [Cui *et al.*, 2020] typically use the KG structure (parent-child) as the sole ground truth. Our approach is distinct in that we utilize *literature-mined contexts* as a data augmentation strategy, inspired by distant supervision techniques in phenotype recognition [Luo *et al.*, 2021], and optimize for both clinical outcome (disease overlap) and taxonomy.

3 Methodology

Our approach consists of three main components: (1) generating a literature-based corpus to enrich phenotype contexts, (2) transforming to a pairs dataset with anchor-based sampling and disease-overlap similarity metric, and (3) fine-tuning the embedding model using a rank-based loss function. This pipeline is shown in Fig. 1.

3.1 Literature-Based Corpus Generation

Standard approaches to generate a corpus based on HPO typically use the single, short definition provided by the HPO as the basic element to compose the training dataset. We will demonstrate that this basic approach for learning robust representations can be improved. For such purpose, we developed a corpus generation pipeline using literature included in PubMed with the focus of providing abundant quality context for all terms in the ontology. The algorithm is shown in Alg. 1.

For each term $t \in V$ (the set of HPO terms), we queried the PubMed database to retrieve scientific papers associated with t. Sentences containing the term or its synonyms provided by HPO were extracted. To ensure high-quality supervision, we applied a **Dynamic Context Windowing** strategy:

- **Filtering:** We discarded sentences containing explicit negations (e.g., "no signs of...") detected via dependency parsing, as well as enumeration lists containing more than three distinct phenotypes to avoid false co-occurrences.

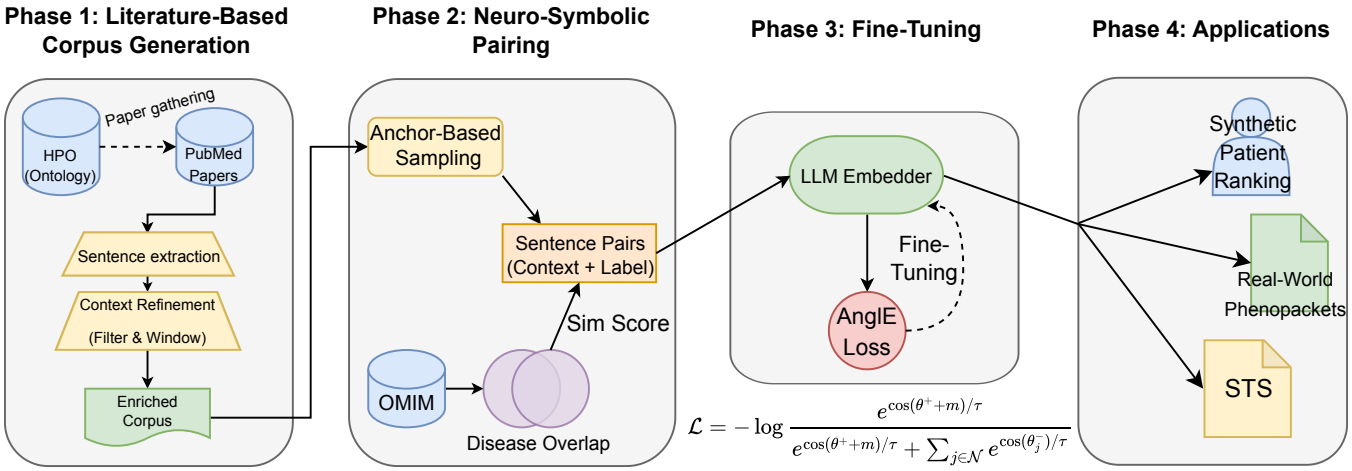

Figure 1: Overview of the framework. The pipeline consists of four stages: (A) Data Gathering & Augmentation, where HPO terms are enriched with diverse linguistic contexts mined from PubMed using dynamic windowing and negation filtering. (B) Pair Generation, where an Anchor-Based Sampling strategy creates training pairs supervised by our novel Disease-Overlap Similarity metric, which calculates ground-truth semantic distance based on shared disease annotations rather than graph topology. (C) Model Fine-Tuning, where a PubMedBERT bi-encoder is optimized using AnglE Loss and partial freezing. (D) Applications in diverse scenarios.

- **Windowing:** For sentences exceeding 50 words, we extracted a window of ± 25 words around the target term to maintain local semantic focus.

This results in a distribution of linguistic contexts s, for each phenotype t, denoted as $S_t = \{s_1, s_2, ..., s_n\}$, rather than a single definition.

Algorithm 1 Context Extraction and Refinement

Require: Set of HPO terms V, PubMed Papers $\mathcal{P}$
Ensure: Training Corpus $\mathcal{C} = \{(s, t) \mid s \text{ is sentence}, t \in V\}$

for all paper $p \in \mathcal{P}$ **do**
  $\mathcal{S}_{raw} \leftarrow \text{SentenceSplitter}(p)$
  for all sentence $s \in \mathcal{S}_{raw}$ **do**
    Clean references and brackets from s
    IdentifiedTerms $\leftarrow$ FindAllHPOTerms(s, V)
    if |IdentifiedTerms| > 3 **then**
      continue {Discard enumeration lists}
    end if
    for all term $t \in$ IdentifiedTerms **do**
      if DependencyParsing(s, t) indicates Negation **then**
        continue {Discard negated contexts}
      end if
      if Length(s) > 50 words **then**
        $s' \leftarrow$ ExtractWindow(s, center=t, window=25)
      else
        $s' \leftarrow s$
      end if
      Add (s', t) to $\mathcal{C}$
    end for
  end for
end for

3.2 Neuro-Symbolic Pairing

Semantic Similarity Labeling (Ground Truth)

To supervise the model, we require a "gold standard" similarity score between phenotype pairs. Instead of relying solely on topological measures (like graph distance), we employed an **annotation-based similarity measure** that aims towards a better recapitulation of clinical knowledge, through capturing the real extensional meaning of phenotypes by shared association with diseases. This similarity is based on the RelativeBestPair (RBP) similarity function [Gong *et al.*, 2018].

Let $\mathcal{D}(t)$ be the set of diseases annotated to a phenotype term t or any of its descendants (sourced from the **OMIM** database):

$$\mathcal{D}(t) = \{d \mid d \in \text{annotations}(t'), \forall t' \in \text{descendants}(t) \cup \{t\}\} \quad (1)$$

This definition integrates the ontology's topological structure via the *True Path Rule*: a term inherently carries the clinical semantics of its descendants. By basing similarity on these propagated sets, our metric implicitly captures the hierarchical information (similar to Resnik or Lin) but grounds it in extensional disease overlap. Thus, a separate topological term is not required, as the graph structure defines the composition of $\mathcal{D}(t)$.

Given a source phenotype t_i and a target t_j, we first compute a *directional similarity score* $\delta(t_i, t_j)$. This score measures the coverage of the target's disease set by the source, weighted by the specificity of the target:

$$\delta(t_i, t_j) = \underbrace{\frac{|\mathcal{D}(t_i) \cap \mathcal{D}(t_j)|}{|\mathcal{D}(t_i)|}}_{\text{Coverage}} \times \underbrace{\min\left(\alpha, \frac{1}{|\mathcal{D}(t_j)|}\right)}_{\text{Specificity Weight}} \quad (2)$$

Here, α is a saturation hyperparameter (set to 0.01). The term $1/|\mathcal{D}(t_j)|$ acts as an Information Content weighting. The final symmetric semantic similarity $Sim(t_i, t_j)$, used as

the label y_{sem}, is the normalized mean of the bidirectional scores:

$$Sim(t_i, t_j) = \frac{1}{2\alpha} \left(\delta(t_i, t_j) + \delta(t_j, t_i) \right) \tag{3}$$

Anchor-Based Hard Sampling

Given the sparse topology of the HPO graph, uniform random sampling of pairs from the $|V| \approx 18,000$ terms results in a distribution heavily skewed towards negligible similarity. Since most random pairs share only the root or highly generic ancestors (e.g., 'Phenotypic abnormality'), they constitute "easy negatives" that provide minimal gradient signal for ranking optimization. To enable efficient learning, we employed an **Anchor-Based Hard Sampling** strategy. For each anchor term t, we generated pairs in three categories:

1. **Positive Pairs (33%):** (t, t) pairs of the same phenotype using distinct sentences from S_t. This enforces semantic invariance across different linguistic contexts.
2. **Hard Negatives (33%):** Pairs (t, t_{hard}) where t_{hard} is a sibling or cousin node with high semantic similarity ($0.3 < Sim(t, t_{hard}) < 0.7$). We empirically capped the upper threshold at 0.7 to avoid "False Negatives": pairs with similarity > 0.7 often represent near-synonyms or clinically indistinguishable phenotypes. Treating such highly similar terms as negatives would force the model to artificially separate concepts that should naturally cluster together, potentially destabilizing the training.
3. **Random Negatives (33%):** Pairs (t, t_{rand}) with low similarity to preserve global structure.

3.3 Fine-Tuning

We employ a bi-encoder architecture optimized using **AnglE Loss** (Angle-optimized Text Embeddings) [Li and Li, 2024]. Unlike traditional cosine-based losses (e.g., CoSENT [Huang *et al.*, 2024]) which can result in vanishing gradients when the similarities approach 1, a frequent occurrence in ontology sub-branches, AnglE introduces an angular optimization in complex space.

We formulate the objective as a contrastive loss with an angular margin m, enhancing the discriminative power of the embeddings:

$$\mathcal{L} = -\log \frac{e^{\cos(\theta^+ + m)/\tau}}{e^{\cos(\theta^+ + m)/\tau} + \sum_{j \in \mathcal{N}} e^{\cos(\theta_j^-)/\tau}} \tag{4}$$

where θ^+ is the angle between the positive pair (determined by high annotation similarity), θ^- corresponds to negative pairs, and τ is the temperature parameter. We employ a **Partial Freezing** strategy.

4 Experiments

4.1 Experimental Setup

To ensure a comprehensive evaluation, we selected three distinct base models representing different paradigms: (1) **PubMedBERT** [Gu *et al.*, 2021], a domain-specific model pretrained from scratch on biomedical text; (2) **SapBERT** [Liu and others, 2021], the current state-of-the-art for biomedical entity alignment initialized from PubMedBERT; (3) **all-miniLM-L6-v2** [Wang *et al.*, 2020] and **all-mpnet-base-v2**, general-purpose efficient sentence transformers, included to assess domain transfer capabilities. For comparative analysis, we evaluated both the off-the-shelf versions (zero-shot) and versions fine-tuned using our proposed methodology.

Hyperparameter Optimization. We performed a Bayesian hyperparameter search using the Optuna framework [Akiba and others, 2019] to maximize the Spearman correlation on a held-out validation set of HPO pairs. The optimal configuration identified and used for all subsequent experiments was: 4 training epochs with a batch size of 64 and a maximum sequence length of 256. We utilized the AdamW optimizer with a weight decay of 0.05, a maximum learning rate of 7.87×10^{-5}, and a linear warmup over the first 6% of steps. We applied a Partial Freezing strategy for the bottom 6 layers of the encoder. Training was accelerated using Automatic Mixed Precision (AMP).

4.2 Components Study

To quantify the contribution of each component, we evaluated four configurations aside from the base models:

- **Baseline:** Topological Fine-Tuning (Defs Only + Lin Similarity).
- **+Context:** Using PubMed Corpus + Lin Similarity.
- **+Metric:** Using Defs Only + Disease-Overlap Similarity.
- **Full Method:** Using PubMed Corpus + Disease-Overlap Similarity.

4.3 Evaluation Tasks

Task 1: Intrinsic Semantic Correlation. How well the LLM recapitulates HPO structure? We computed the Spearman rank correlation between the cosine similarity of the model's embeddings and the ground-truth semantic similarity (both Lin and Disease-Overlap) on a held-out test set of HPO pairs.

Task 2: Extrinsic Phenotype Association (GSC+ and JAX). How well the LLM associates HPO phenotypes to biomedical texts? We utilized the GSC+ corpus [Lobo *et al.*, 2017], which consists of 228 manually annotated PubMed abstracts, with a total of 1933 annotations that cover 497 unique HPO concepts. We formulated this as a retrieval task: for each abstract, we ranked HPO terms and measured Recall@1, Recall@5 and Mean Reciprocal Rank (MRR). Additionally, we use the JAX corpus [Luo *et al.*, 2021], which provides document-level annotations of HPO terms in 131 PMC full-text articles, with a total of 988 unique HPO concepts.

Task 3: Synthetic Patient Disease Ranking. How does the LLM perform in modeling patients based on combination of phenotypes? To evaluate diagnostic reasoning, we simulated patients using the HPO annotation database. For a given disease D, we randomly sampled $k \in \{5, 10\}$ associated phenotypes. We generated a "Patient Vector" by averaging the embeddings of the corresponding PubMed sentences extracted

	Intrinsic Evaluation (STS)		Extrinsic (GSC+)			Extrinsic (JAX)		
Model Configuration	**Spearman**	**Pearson**	**Top-1**	**Top-5**	**MRR**	**Top-1**	**Top-5**	**MRR**
Base (PubMedBERT)	0.77	0.87	0.13	0.32	0.21	0.003	0.011	0.010
Defs + Lin (Topological)	0.78	0.87	0.16	0.30	0.23	0.005	0.015	0.013
Defs + RBP (Metric)	0.77	0.79	0.18	0.31	0.25	0.006	0.015	0.014
PubMed + Lin (Context)	0.75	0.87	0.10	0.26	0.17	0.005	0.010	0.008
PubMed + RBP (Ours)	**0.84**	**0.94**	**0.26**	**0.39**	**0.32**	**0.012**	**0.030**	**0.023**

Table 1: Component Study Results. **STS** indicates intrinsic semantic similarity on the test set. **GSC+** and **JAX** indicate the metrics in the phenotype-retrieval datasets. The **PubMed + RBP** configuration (Ours) demonstrates superior performance across all metrics.

for each phenotype. The model was tasked with retrieving the correct disease D from the set of all OMIM diseases. We report Top-1 and Top-5 Accuracy and MRR for 1000 generated synthetic patients, where diseases are represented with a "Disease Vector" averaging the embeddings of the PubMed sentences extracted for all the phenotypes related to the disease.

Task 4: Real-World Phenopacket Prioritization. Asking the same question again as in Task 3, we evaluated performance on real-world clinical cases from the **Phenopacket Store** [Danis and others, 2025]. This repository offers cohorts of GA4GH Phenopackets that represent individuals with Mendelian diseases. We selected a subset of cases with confirmed genetic diagnoses. For each case, we extracted the observed phenotypes (excluding negated terms), converted them to embeddings using our model, and ranked candidate diseases the same way as the synthetic patients. This represents the most challenging and realistic evaluation setting. We report Top-1 and Top-5 Accuracy and MRR for a total of 6556 patients with at least 3 phenotypes in its description and a confirmed Omim disease within the HPO dataset.

Task 5. General Semantic Capabilities. Does our model present catastrophic forgetting? We evaluated the performance on three standard biomedical and STS benchmarks to evaluate the degradation of performance in general semantic tasks: **BIOSSES** (Biomedical) [Soğancıoğlu *et al.*, 2017], **STS-B** (General) [Cer and others, 2017], and **SICK-R** (General) [Enevoldsen and others, 2025; Marelli and others, 2014; Muennighoff *et al.*, 2022].

5 Results

5.1 Components Study

To disentangle the contributions of data augmentation (PubMed contexts) and the supervision signal (Semantic Metric), we compared four fine-tuning configurations (see section 4.2) against the pre-trained PubMedBERT baseline. Table 1 summarizes the performance on intrinsic semantic correlation (Spearman ρ on held-out HPO pairs) and extrinsic gene-disease association retrieval (GSC+).

Comparing the models trained on standard definitions (*Defs*), we observe that shifting from topological similarity (*Lin*) to annotation-based similarity (*RBP*) improves the **Top-5 Accuracy** on GSC+ from 0.30 to 0.31 and the MRR on JAX from 0.013 to 0.014. This suggests that even with limited textual context, forcing the model to respect disease co-occurrence patterns aligns the latent space better with downstream diagnostic tasks.

A crucial finding is the behavior of the *PubMed + Lin* configuration. Surprisingly, augmenting the model with rich literature contexts while supervising with a graph-based metric (Lin) degrades performance below the baseline (GSC+ Top-1 drops to 0.10). We hypothesize that this represents a **neuro-symbolic mismatch**: the rich functional descriptions in PubMed papers often imply clinical connections that contradict the rigid, graph-based distance of Lin similarity, introducing noise during training [Gong *et al.*, 2018].

In contrast, our proposed **PubMed + RBP** configuration achieves a synergistic effect. By aligning rich clinical language (PubMed) with a clinically grounded metric (RBP), we achieve a remarkable improvement: GSC+ **Top-1 Accuracy** doubles compared to the baseline ($0.13 \rightarrow 0.26$), and intrinsic Spearman correlation reaches 0.84. This confirms that data augmentation requires a commensurate increase in the sophistication of the supervision signal.

5.2 Comparative Analysis

We benchmarked our proposed fine-tuning strategy (using PubMed contexts and RBP metric) across four distinct model architectures: **PubMedBERT** (domain-specific base), **SapBERT** (state-of-the-art entity alignment), **MiniLM** (efficient distilled model) and **Mpnet** (best general model). Table 2 details the performance shifts from baseline to fine-tuned states.

Performance vs. Architecture

Both **PubMedBERT** and **SapBERT** demonstrated significant gains, validating the universality of our method. Notably, our fine-tuned PubMedBERT achieves a GSC+ MRR of 0.320, effectively bridging the gap with the base SapBERT (MRR 0.215) and surpassing it significantly. This indicates that our *neuro-symbolic alignment* can instill deep clinical knowledge into a standard biomedical model without requiring the massive entity-level pre-training of SapBERT.

When applying our method to **SapBERT** (FT-SapBERT), we achieved the highest overall performance across all tasks (GSC+ MRR 0.340). This suggests that our contribution is additive: SapBERT provides robust synonym resolution (entity linking), while our method injects hierarchical and pathophysiological reasoning. The combination yields a model that

	Intrinsic	Extrinsic (GSC+)			Extrinsic (JAX)		
Model	**Spearman ρ**	**Top-1**	**Top-5**	**MRR**	**Top-1**	**Top-5**	**MRR**
Baselines (Pre-trained)							
PubMedBERT (Base)	0.770	0.131	0.322	0.209	0.003	0.011	0.010
SapBERT (Base)	0.761	0.137	0.323	0.215	0.003	0.009	0.008
MiniLM (Base)	0.775	0.142	0.29	0.205	0.004	0.010	0.010
Mpnet (Base)	0.855	0.112	0.29	0.185	0.003	0.010	0.008
Fine-Tuned (Ours: PubMed Contexts + RBP Metric)							
FT-PubMedBERT	0.839	0.261	0.39	0.320	**0.012**	**0.03**	**0.023**
FT-SapBERT	0.830	**0.269**	**0.42**	**0.340**	0.009	0.026	0.021
FT-MiniLM	0.748	0.087	0.20	0.145	0.003	0.010	0.008
FT-Mpnet	**0.855**	0.112	0.29	0.185	0.005	0.010	0.009

Table 2: Comparative performance of base vs. fine-tuned models. **STS**: Spearman correlation (ρ) on HPO test pairs. **GSC+**: Recall@1, Recall@5 and Mean Reciprocal Rank (MRR) on gene-disease associations. **JAX**: Recall@1, Recall@5 and MRR on literature-based phenotype retrieval.

is superior at both recognizing terms and understanding their clinical implications.

Conversely, the distilled **MiniLM** model suffered a performance collapse upon fine-tuning (Spearman dropped to 0.748, GSC+ Top-1 halved to 0.087). We hypothesize that the reduced parameter space of MiniLM lacks the *capacity* to accommodate the drastic re-organization of the embedding space required by our loss function. Something similar occurs with the **Mpnet** model, where the performance does not upgrade with our fine-tuning. This model has a really strong spearman correlation (0.855) but the metrics are way worse than the fine-tuned models in the rest of experiments.

Analysis of Downstream Tasks

GSC+ (Gene-Disease): The improvement in Top-1 Accuracy is the most striking result. FT-PubMedBERT retrieves the correct phenotype for a gene description in the first position 26.1% of the time, compared to just 13.1% for the baseline. This doubling of precision significantly enhances the model's utility for practical applications. For instance, in automated curation (extracting structured data from literature) and variant prioritization (ranking candidate genes in genomic analysis), a high Top-1 accuracy directly minimizes the need for manual verification by experts.

JAX (Literature Retrieval): While absolute MRR scores remain low due to the extreme difficulty of the task (matching short phenotype queries against full-text noise), our method achieves a relative improvement of over **130%** compared to baselines. This confirms that the model is learning a robust signal that persists even in high-noise environments, although specialized retrieval architectures (e.g., re-rankers) would be needed for practical deployment on full texts.

5.3 Diagnostic Reasoning

To rigorously evaluate the clinical utility of our aligned embeddings, we conducted a experiment ranging from controlled simulations to large-scale real-world patient data.

We generated 1000 synthetic patients by sampling $N = \{5, 10\}$ phenotypes from disease definitions in HPO to generate a synthetic clinical case for the patient with a sentence per phenotype. Phenotypes were represented by random sentences mined from PubMed rather than canonical definitions. Additionally, we evaluated the models on the Phenopacket store dataset, comprising 6,556 curated clinical cases extracted from medical literature with and average of 8.56 phenotypes per patient. Unlike synthetic patients, these cases contain noise, incomplete phenotyping, and heterogeneous reporting styles. The phenotypes were also represented by random sentences from PubMed related to the phenotypes.

As shown in Table 3, baselines perform relatively well in controlled synthetic settings, with MiniLM even showing strong keyword-matching capabilities (Top-5 24.9% for N=5). However, this performance collapses in the real-world Phenopackets task, where base models achieve a Top-1 accuracy of only $\sim$3-4% and a Top-5 around 11%. They fail to bridge the gap between complex clinical narratives and structured disease entities. The performance of Mpnet is really similar, so we can intuit that general models do not have the necessary knowledge for this type of task, even with specific fine-tuning.

In contrast, our fine-tuned models exhibit remarkable robustness. FT-PubMedBERT achieves a Top-1 Accuracy of 17.5% and a Top-5 Accuracy of 34.1%, representing a 4-fold improvement in the Top-1 metric and a 3-fold improvement in the Top-5 retrieval zone compared to the baseline (11.4%). This means that in a real-world scenario, our model successfully places the correct diagnosis within the top 5 candidates for more than one-third of the complex cases, drastically reducing the search space for clinicians.

5.4 The Cost of Specialization

We evaluated performance on three standard biomedical and general STS benchmarks: **BIOSSES** (Biomedical), **STS-B** (General), and **SICK-R**. Table 4 presents the results.

We observe a divergent behavior depending on the training data source. When fine-tuning with HPO Definitions (*FT Defs*), the model retains or even improves general biomedical similarity (BIOSSES $\rho = 0.85$). This is likely because definitions preserve standard lexicographic structures.

However, our best diagnostic model (*FT PubMed+RBP*)

	Synthetic (N=5)			Synthetic (N=10)			Real-World (Phenopackets)		
Model	**Top-1**	**Top-5**	**MRR**	**Top-1**	**Top-5**	**MRR**	**Top-1**	**Top-5**	**MRR**
Baselines									
PubMedBERT (Base)	10.1%	21.7%	0.165	14.5%	31.3%	0.233	4.2%	11.4%	0.087
SapBERT (Base)	6.5%	13.8%	0.111	8.7%	17.3%	0.138	3.6%	11.6%	0.083
MiniLM (Base)	11.5%	24.9%	0.183	11.3%	25.8%	0.188	3.1%	9.9%	0.076
Mpnet (Base)	8.7%	18.4%	0.145	10.3%	23.53%	0.171	0.040	0.120	0.087
Fine-Tuned (Ours: PubMed + RBP + AnglE)									
FT-PubMedBERT	**11.7%**	**23.1%**	**0.181**	**18.8%**	**34.9%**	**0.270**	**17.5%**	**34.1%**	**0.256**
FT-SapBERT	7.7%	17.9%	0.135	12.5%	26.2%	0.197	15.2%	30.8%	0.229
FT-MiniLM	6.3%	14.3%	0.109	7.4%	15.1%	0.121	1.2%	4.5%	0.037
FT-Mpnet	8.6%	18.3%	0.144	10.1%	23.69%	0.172	0.040	0.120	0.087

Table 3: **Diagnostic Performance Comparison.** We report **Top-1** and **Top-5** Accuracy and Mean Reciprocal Rank (**MRR**). **Synthetic Cohorts** ($N = 5, 10$) represent controlled environments. **Phenopackets** ($N = 6556$) represents real-world clinical performance. Note the massive performance gap ($\sim$4x) our method achieves in the real-world setting compared to baselines.

Model	**BIOSSES**	**STS-B**	**SICK-R**
Base (PubMedBERT)	0.83	0.76	0.66
FT (Defs + RBP)	**0.85**	0.73	0.66
FT (PubMed + Lin)	0.83	0.74	0.65
FT (PubMed + RBP)	0.57	0.63	0.59

Table 4: Transfer evaluation on general and biomedical semantic similarity tasks (Spearman ρ). A significant drop in general benchmarks indicates a shift from lexical to clinical alignment.

exhibits a significant performance drop on BIOSSES ($\rho = 0.83 \rightarrow 0.57$). Rather than simple catastrophic forgetting, we interpret this as a necessary semantic shift. Standard benchmarks like BIOSSES reward lexical and topical similarity (e.g., "Eye pain" $\approx$ "Eye irritation"). In contrast, our model is optimized for *functional* similarity (e.g., "Eye pain" $\approx$ "Kidney failure" in the context of Lowe Syndrome).

This degradation in general benchmarks is the trade-off for the massive gains in Phenopacket retrieval (4x improvement). It confirms that the model has specialized effectively: it has ceased to be a general-purpose text encoder and has become a specialized diagnostic retriever.

6 Conclusion

In this work, we proposed a Neuro-Symbolic Alignment Framework to bridge the semantic gap between biomedical Large Language Models and expert knowledge graphs. By augmenting the Human Phenotype Ontology with literature-mined contexts and supervising the embedding space with a disease-overlap metric, we successfully instilled clinical reasoning capabilities into standard transformer architectures.

Our extensive experimental evaluation leads to three critical conclusions. The most significant finding is the dramatic performance gap in real-world scenarios. While standard baselines like PubMedBERT and SapBERT collapse when facing the noise and heterogeneity of real clinical reports (Phenopackets), our fine-tuned models maintain robust performance, achieving a **4-fold improvement** in diagnostic retrieval (Top-1 Accuracy 17.5% vs 4.2%). This confirms that our method effectively aligns the latent space with pathophysiology, allowing the model to look beyond lexical surface forms.

We demonstrated that initializing with state-of-the-art weights (SapBERT) further boosts performance, but applying our methodology to a standard domain model (PubMedBERT) yields comparable or superior results to the original SOTA in diagnostic tasks. This suggests that *how* we align models (using disease logic) is as important as the data scale used for pre-training.

The decline of MiniLM highlight a boundary condition: deep semantic restructuring requires sufficient parameter capacity. Efficient models, while useful for keyword matching, appear too rigid to accommodate the complex, non-taxonomic relationships inherent in medical diagnosis.

While our neuro-symbolic alignment yields a 4-fold improvement over baselines in real-world scenarios, an absolute Top-1 accuracy of 17.5% (and Top-5 of 34.1%) on Phenopackets indicates that the problem of automated rare disease diagnosis via text is far from solved. The complexity of clinical variability, combined with the noise inherent in unstructured case reports, presents a significant barrier. Our results suggest that while embedding alignment is a necessary step, achieving clinical-grade accuracy likely requires integrating multimodal data (e.g., genomic variants) and perhaps reasoning capabilities beyond vector similarity. Consequently, current LLM-based phenotyping tools should be deployed as "human-in-the-loop" prioritization assistants rather than autonomous diagnostic agents.

Future work will explore extending this framework to other biomedical ontologies, such as the Gene Ontology (GO), and integrating these aligned embeddings into retrieval-augmented generation (RAG) pipelines for explainable rare disease diagnosis.

Acknowledgments

This work was supported by the Fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia (Spain)

(22308/FPI/23) and the Spanish Council of Science and Innovation through the Grant PID2022-136306OB-100 funded by MICIU/AEI/10.13039/501100011033 and FEDER/UE.

Contribution Statement

The authors Alejandro Cisterna-García and Juan A. Botía contributed equally.

References

[Akiba and others, 2019] Takuya Akiba et al. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery.

[Bordes and others, 2013] Antoine Bordes et al. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.

[Cer and others, 2017] Daniel Cer et al. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[Cui *et al.*, 2020] Baiyun Cui, Yingming Li, and Zhongfei Zhang. BERT-enhanced relational sentence ordering network. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6310–6320, Online, November 2020. Association for Computational Linguistics.

[Danis and others, 2025] Daniel Danis et al. A corpus of GA4GH phenopackets: Case-level phenotyping for genomic diagnostics and discovery. *HGG Adv.*, 6(1):100371, January 2025.

[Enevoldsen and others, 2025] Kenneth Enevoldsen et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

[Garnot and Landrieu, 2020] Vivien Sainte Fare Garnot and Loïc Landrieu. Metric-guided prototype learning. *CoRR*, abs/2007.03047, 2020.

[Genetic Alliance and others, 2010] Genetic Alliance et al. *Understanding genetics: A New England guide for patients and health professionals*. Genetic Alliance, Washington (DC), February 2010.

[Gong *et al.*, 2018] Xiaofeng Gong, Jianping Jiang, Zhongqu Duan, and Hui Lu. A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology. *BMC Bioinformatics*, 19(S4), May 2018.

[Grover and Leskovec, 2016] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA, 2016. Association for Computing Machinery.

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

[Hao *et al.*, 2020] Junheng Hao, Chelsea J.-T Ju, Muhao Chen, Yizhou Sun, Carlo Zaniolo, and Wei Wang. Biojoie: Joint representation learning of biological knowledge bases. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '20, New York, NY, USA, 2020. Association for Computing Machinery.

[Huang *et al.*, 2024] Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S. Yu. Cosent: Consistent sentence embedding via similarity ranking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2800–2813, 2024.

[Köhler and others, 2009] Sebastian Köhler et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464, 2009.

[Köhler and others, 2020] Sebastian Köhler et al. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, 12 2020.

[Lee *et al.*, 2019] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[Li and Li, 2024] Xianming Li and Jing Li. AoE: Angle-optimized embeddings for semantic textual similarity. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Lin, 1998] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[Liu and others, 2021] Fangyu Liu et al. SapBERT: Self-alignment pretraining for biomedical entity representation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4234–4243. Association for Computational Linguistics, 2021.

[Lobo *et al.*, 2017] Manuel Lobo, Andre Lamurias, and Francisco M Couto. Identifying human phenotype terms by combining machine learning and validation rules. *Biomed Res. Int.*, 2017:1–8, 2017.

[Lowe *et al.*, 1952] Charles Lowe, Mary Terrey, and E. A. MacLachlan. Organic-aciduria, decreased renal ammonia production, hydrophthalmos, and mental retardation: A clinical entity. *A.M.A. American Journal of Diseases of Children*, 83(2):164–184, 02 1952.

[Luo *et al.*, 2021] Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Eyasu Goshu, Tiange Li, Anna A Morgan, Chunhua Weng, and Zhiyong Lu. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, 37(13):1884–1890, 2021.

[Marelli and others, 2014] Marco Marelli et al. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[Muennighoff *et al.*, 2022] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[Resnik, 1995] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, page 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[Robinson and others, 2020] Peter N. Robinson et al. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3):403–417, 2020.

[Robinson *et al.*, 2008] Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008.

[Soğancıoğlu *et al.*, 2017] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, 07 2017.

[Wang *et al.*, 2020] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788, 2020.

[Wen *et al.*, 2025] Baole Wen, Sheng Shi, Yi Long, Yanan Dang, and Weidong Tian. PhenoDP: leveraging deep learning for phenotype-based case reporting, disease ranking, and symptom recommendation. *Genome Med.*, 17(1):67, June 2025.

[Yuan *et al.*, 2020] Zheng Yuan, Zhengyun Zhao, and Sheng Yu. CODER: knowledge infused cross-lingual medical term embedding for term normalization. *CoRR*, abs/2011.02947, 2020.