RELATE: Relation Extraction in Biomedical Abstracts with LLMs and Ontology Constraints

Olawumi Olasunkanmi OLAWUMI@CS.UNC.EDU
Department of Computer Science, University of North Carolina, Chapel Hill, United States

Matthew Satusky MSATUSKY@RENCI.ORG
Hong Yi HONGYI@RENCI.ORG
Chris Bizon CBIZON@RENCI.ORG
Renaissance Computing Institute, Chapel Hill, United States

Harlin Lee HARLIN@UNC.EDU
Stanley Ahalt AHALT@UNC.EDU
School of Data Science and Society, University of North Carolina, Chapel Hill, United States

Abstract

Biomedical knowledge graphs (KGs) are vital for drug discovery and clinical decision support but remain incomplete. Large language models (LLMs) excel at extracting biomedical relations, yet their outputs lack standardization and alignment with ontologies, limiting KG integration with free texts. We introduce RELATE, a three-stage pipeline that maps LLM-extracted relations to standardized ontology predicates, e.g., the Biolink Model. The pipeline includes: (1) ontology preprocessing with predicate embeddings, (2) similarity-based retrieval enhanced with SapBERT, and (3) LLM-based reranking with explicit negation handling. This approach performs relation extraction from free-text outputs to structured, ontology-constrained representations. On the ChemProt benchmark, RELATE achieves 52% exact match and 94% accuracy@10, and in 2,400 HEAL Project abstracts, it effectively rejects irrelevant associations (0.4%) and identifies negated assertions. RELATE captures nuanced biomedical relationships while ensuring quality for KG augmentation. By combining vector search with contextual LLM reasoning, RELATE provides a scalable, semantically accurate framework for converting unstructured biomedical literature into standardized KGs.

Keywords: Relation Extraction, Large Language Models, Ontology, Biolink Model, Biomedical Knowledge Graphs

Data and Code Availability Data used in this project are publicly available ChemProt dataset[1] and PubMed abstracts available through the NIH Helping to End Addiction Long-term (HEAL) Initiative[2]. The code is available in a public Git repository[3].

Institutional Review Board (IRB) This research analyzes public biomedical publication texts on PubMed and does not require IRB approval.

1. Introduction

Knowledge graphs (KGs) serve as structured repositories of facts, capturing entities and their relationships to support applications such as semantic search, recommendation systems, and drug repurposing (Olasunkanmi et al., 2024; Bizon et al., 2019). However, real-world KGs are often incomplete, with missing facts that limit their utility.

Unstructured biomedical text offers a vast, underutilized source for KG completion. Natural language processing techniques, particularly *relation extraction (RE)*, play a central role in mining relationships between entities mentioned in the text. While Named Entity Recognition (NER) identifies entities (e.g., drugs, proteins, diseases), RE captures the *predicates* describing their relationships (e.g., treats, inhibits), which provide structure and meaning to extracted knowledge. Predicates thus connect entities into an interconnected biomedical KGs. Biomedical texts,

1. https://huggingface.co/datasets/bigbio/chemprot
2. https://heal.nih.gov/
3. https://github.com/RENCI-NER/pred-mapping/releases/tag/v1.0

© 2025 O. Olasunkanmi, M. Satusky, H. Yi, C. Bizon, H. Lee & S. Ahalt.

however, are challenging to analyze due to domain-specific jargon, abbreviations, and subtle contextual nuances.

Recent advancements in large language models (LLMs) have demonstrated remarkable capabilities in biomedical text understanding (Jahan et al., 2024). Nevertheless, current LLM-based RE methods often lack integration with domain ontologies, leading to high false-positive rates and inconsistent relationship representations. Standardized *ontology* frameworks such as the Biolink Model (Unni et al., 2022) define well-structured predicates for biomedical relationships (Bizon et al., 2019; Joachimiak et al., 2021; Reese et al., 2021; all, 2022). Yet, mapping free-text relations to ontology predicates remains a difficult task due to the semantic gap between natural language expressions and formal ontological representations.

To address these challenges, we propose **RELATE**, a three-stage pipeline that maps free-form biomedical relations to standardized ontology predicates. The key contributions of this work are:

1. **Ontology-constrained relation standardization:** RELATE introduces the first systematic framework for mapping LLM-extracted biomedical relations to established ontologies (ChemProt and Biolink), moving beyond relation classification toward ontology-grounded standardization. This design ensures semantic interoperability with biomedical KGs. While ChemProt has been used for supervised relation classification (Peng et al., 2019; Warikoo et al., 2021), and Biolink for KG schema definition (Joachimiak et al., 2021; Reese et al., 2021), our work represents the first application of these ontologies as *standardization targets* for free-text relation extraction, bridging the gap between unstructured text mining and structured knowledge representation.

2. **Hybrid retrieval and efficient refinement:** We develop a three-stage pipeline that combines ontology-driven preprocessing, SapBERT-enhanced hybrid vector retrieval, and LLM-based contextual reranking. RELATE leverages SapBERT for relation semantics for the first time, and applies computationally intensive reasoning only to small candidate sets ($k = 10$), achieving both scalability and accuracy.

3. **Comprehensive evaluation:** We conduct dual-ontology assessment on ChemProt (Peng et al., 2019) benchmark and 2,400 real-world HEAL Project abstracts. Integration with an established, large-scale biomedical KG (ROBOKOP) is also planned to further demonstrate its translational impact.

These capabilities enable downstream applications in literature-based discovery, biomedical KG construction, and multi-source knowledge integration; see Appendix A.

2. Related Work

No existing framework systematically bridges free-text biomedical relation extraction (RE) with standardized ontological predicates. While zero-shot LLMs (Li et al., 2025) can outperform fine-tuned models on smaller datasets, they struggle when confronted with complex, domain-specific tasks that demand specialized knowledge (Chen et al., 2024b; Jahan et al., 2024). Further research has explored adaptive instruction-rich prompting (Zhou et al., 2024; Tao et al., 2024) and soft prompt-based learning approaches that allow models to automatically optimize prompts rather than relying on hand-crafted instructions (Peng et al., 2024). These methods show promise but have not achieved systematic biomedical ontology integration.

Fine-tuning approaches have demonstrated strong performance on domain-specific benchmarks. Techniques such as adaptive document mapping, ensemble learning with attention mechanisms, and specialized training procedures (Shang et al., 2025; Zhong et al., 2024; Orlova and Orlov, 2025) improve classification accuracy. However, they require extensive retraining and do not address the core challenge of standardizing predicates extracted from free-text relations. Chandak et al. (2023) emphasized the importance of standardized schemas and quality control in its PrimeKG framework by integrating heterogeneous data sources. RELATE applies these principles to the RE layer, enabling automated standardization of literature-derived relations for seamless integration into KGs such as PrimeKG, with ensured schema compliance and semantic interoperability.

Recent work in LLM-based RE has explored various validation approaches. RelCheck (Ourekouch et al., 2025) addresses low-confidence predictions from pretrained models using ontology-guided

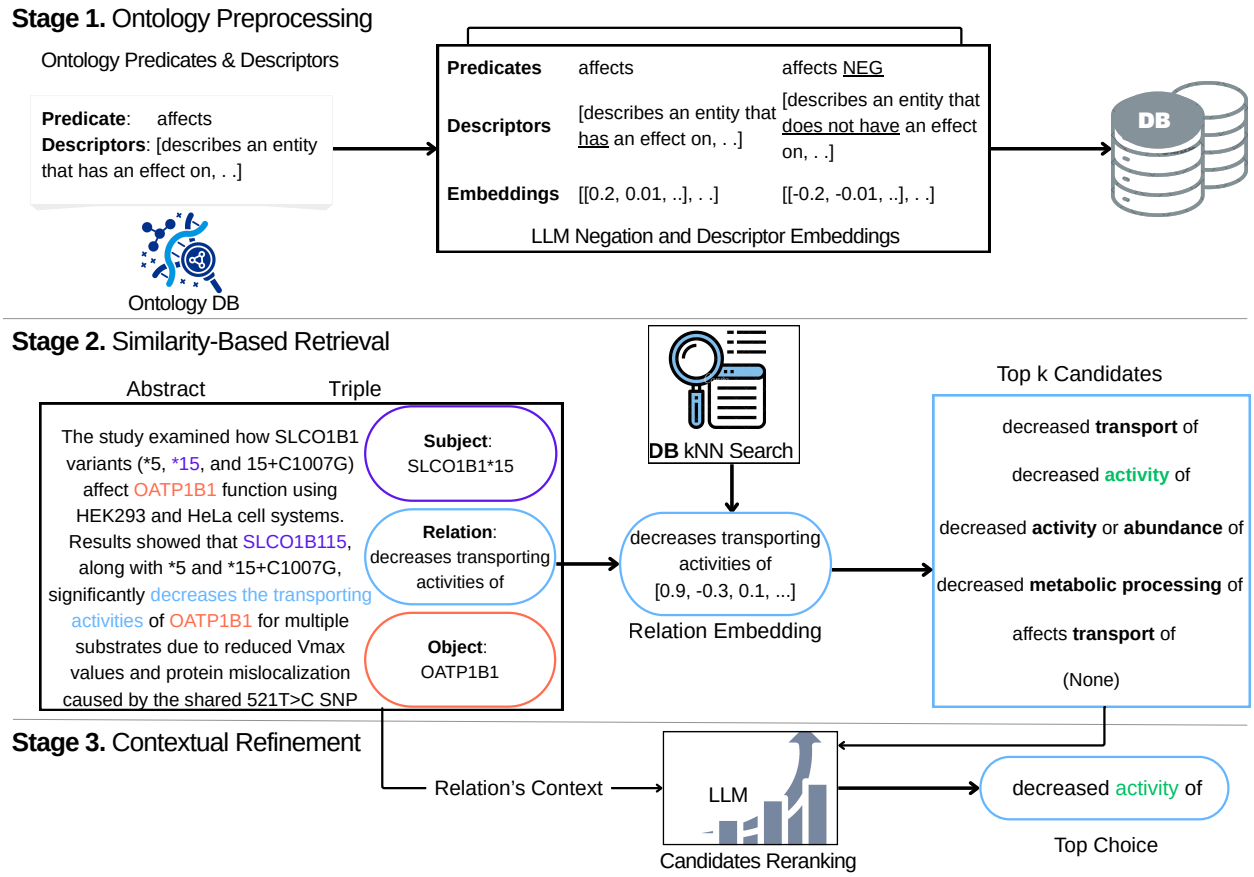

Figure 1: The three-stage RELATE pipeline: (1) ontology preprocessing, (2) similarity-based retrieval, and (3) contextual refinement. Ontology preprocessing generates embeddings for predicates and their negated variants. It updates only if the ontology schema or embedding models change. Given an input quadruple—subject, object, relation text, and abstract context—RELATE performs similarity-based retrieval by embedding the relation and retrieving top-k ontology candidates. In the final stage, contextual refinement reranks these candidates with an LLM using the full abstract context, producing the most semantically appropriate ontology predicate.

LLM validation, though it focuses on confidence rather than standardization. SPIREX (Cavalleri et al., 2024) applies schema-constrained prompts with graph machine learning validation but restricts itself to RNA-specific domains. Graph-augmented approaches include RAG-enhanced LLMs for automated ontology extension (Georgakopoulos et al., 2025), structure-oriented RAG for KG-LLM co-learning (Yang et al., 2024), integrated GNNs with LLM-generated support documents (Dong et al., 2024), and LLM-driven ontology enrichment pipelines (Kollapally et al., 2025). However, these approaches primarily target domain-specific extraction or confidence validation rather than systematic standardization of extracted relationships to established ontological frameworks. Multi-channel neural architectures combine textual and graph structures but focus primarily on structural features rather than systematic predicate mapping. Caufield et al. (2024) used structured prompts for ontology-grounded annotations, focusing on entity linking rather than full relation standardization, while enhancements to SemRep (Ming et al., 2024) remain within existing predicate frameworks.

While Cao et al. (2024) built an end-to-end system for extracting knowledge about rare diseases, their work addresses only 6 relation types in a narrow domain. More fundamentally, no existing framework systematically bridges free-text biomedical RE with standardized ontological predicates. RELATE addresses this critical gap by mapping free-text relations to standardized ontology predicates, handling both narrow (ChemProt's 9 predicates) and broad (Biolink's 284 predicates) ontologies, ensuring semantic interoperability and enabling seamless integration into biomedical KGs.

3. RELATE Methodology

Biomedical relation extraction (RE) systems face a critical challenge: while LLMs excel at identifying subject–object relationships in scientific text, the extracted relations often lack standardization and fail to align with biomedical ontologies. This disconnect limits their value for KG construction and augmentation.

To address this limitation, we propose a three-stage pipeline (Figure 1) that bridges free-text RE with standardized predicates through an ontology-driven protocol. The pipeline consists of: (1) **Ontology Preprocessing**, which generates positive and negative descriptors for predicates and embeds them into a searchable space; (2) **Similarity-Based Retrieval**, which efficiently narrows candidate predicates using vector search enhanced with biomedical embeddings; and (3) **Contextual Refinement**, which leverages LLM reasoning and abstract-level context to rerank candidates and select the most appropriate predicate.

3.1. Stage 1: Ontology Preprocessing

The first stage of our pipeline (Alg. 1) extracts ontology predicates with positive and negative descriptors to encode domain knowledge, and transforms them into a searchable embedding database suitable for relation mapping.

We begin by collecting all predicate definitions from biomedical ontologies such as the Biolink Model toolkit (Unni et al., 2022) or ChemProt (Krallinger et al., 2017), thereby constructing a base vocabulary of biomedical relationships. We will denote the set of all ontology predicates as $\mathcal{R}$. Each predicate $r \in \mathcal{R}$ represents a distinct biological or medical relationship (e.g., affects, inhibits) that can exist between entities in the KG.

The descriptor dataset $\mathcal{D}_r^+$ is constructed by gathering existing textual descriptions for each predicate r. For example, one descriptor for the predicate *affects* may be "describes an entity that has an effect on the state or quality of another existing entity." These descriptions are sourced from multiple authoritative biomedical sources. In the Biolink example, this includes official documentation and standardized vocabularies, while in the ChemProt case, online relation category descriptions suffice. The distributions of the predicate descriptors are presented in Tables 2 and 3. These aggregated descriptions form the positive dataset $\mathcal{D}_r^+$ and capture the semantic diversity with which biomedical relationships appear in scientific literature.

To complement the positive dataset, we introduce a novel strategy for generating negative descriptors. For each descriptor in D_r^+, an LLM produces a natural, semantically coherent negation using a structured prompt (Figure 2). For example, the earlier descriptor becomes "describes an entity that <u>does not have</u> an effect on the state or quality of another existing entity." These negated descriptors are stored in a separate dictionary D_r^- to serve as contrastive examples that help the embedding model distinguish valid from invalid predicate mappings. D_r^- is labeled by r

+ "_NEG" (i.e. *affects* → *affects_NEG*) to maintain clear separation from positive instances.

Finally, every *descriptor* in D_r^+ and D_r^- is transformed into a d-dimensional vector representation using the embedding model f_θ. Instead of aggregating these descriptor embeddings to a single predicate embedding, we store all descriptor embeddings, $\mathcal{V}_r$, in a searchable database for the next stage.

To address limitations of general-purpose embedding models (e.g., `nomic-text-embed` or `bge-m3`) in biomedical domains, we introduce an (optional) hybrid retrieval enhancement based on SapBERT (Liu et al., 2021), a model specialized in biomedical entity linking. In this setup (Figure 4), SapBERT was fine-tuned on D_r^+ and D_r^- to generate a second set of embeddings $\mathcal{V}_r'$. While SapBERT has demonstrated effectiveness in biomedical entity alignment, this work presents its first application to RE through descriptor embedding enhancement. Our hybrid approach leverages SapBERT's biomedical domain knowledge to improve relation semantic similarity beyond general-purpose embedding models.

3.2. Stage 2: Similarity-Based Retrieval

For every predicate $r \in \mathcal{R}$, Stage 1 has produced positive and negative descriptor dictionaries $\mathcal{D}_r^+$, $\mathcal{D}_r^-$ and corresponding descriptor embedding vectors $\mathcal{V}_r$. We now leverage these data to retrieve candidate ontology predicates for incoming relations extracted by an upstream RE system, which could be any general-purpose LLM.

The extracted relations come in a quadruple (s, o, T, a): s subject entity, o object entity, T the free-form relation text, and a the abstract context from which the relation was identified. For example, in Figure 1, T is "decreases transporting activities of," which does not exist in Biolink ontology $\mathcal{R}$ in this format.

First, free-form relation text T is embedded to a query vector $q = f_\theta(T)$. We use the same model f_θ from Stage 1, ensuring semantic consistency between query relations and stored ontology predicates.

Once embedded, the query vector q is compared against all predicate descriptor embeddings in $\mathcal{V}_r, \ \forall r \in \mathcal{R}$. This similarity search retrieves the top-k candidates (typically $k = 10$) using cosine similarity. If multiple descriptors of the same predicate are in the top-k results, they are collapsed into a single predicate candidate. When SapBERT enhancement is on, the top-k candidates in $\mathcal{V}_r', \ \forall r \in \mathcal{R}$ are retrieved in parallel. Then, the two sets of candidates are merged and de-duplicated to identify (up to) $2k$ candidate predicates.

```
You are a biomedical researcher extracting
negations of ontological predicates.

Your Task:
Given a description, return its natural negation.

Rules:
1.  Preserve the meaning but negate the entire
description.
3.  Do not summarize or change the structure of
the descriptor text.
4.  If there is not enough information to create
a negation, your response should be "NOT ENOUGH
INFORMATION"
4.  Only return the negation|no explanations or
extra text.

Examples:
- "has effect" → "does not have effect"
- "during which ends" → "during which does not
ends"
- "happens during" → "does not happen during"

Input:  "{descriptor_text}"

Output:  A JSON object with these exact keys
and format:
{{"negation_of_the_descriptor_text":  "negated
version" or "NOT ENOUGH INFORMATION"}}
```

Figure 2: Negation generation prompt used in Stage 1 (Section 3.1) for creating negative descriptors from positive descriptors.

By narrowing the search space from hundreds of ontology predicates to a small candidate set, this stage provides an efficient yet semantically informed mechanism for predicate selection. At this point, however, the search operates without the benefit of contextual information from the original biomedical abstract a, focusing only on lexical and semantic similarity.

3.3. Stage 3: Contextual Refinement

The final stage addresses the key limitation of similarity-based retrieval: its reliance on context-free embeddings. While Stage 2 efficiently narrows the predicate space to a small candidate set, it does not

```
You are an expert in biomedical relationships.
Based on the text below:

Subject: {subject}
Object: {object}
Original Relationship: {relationship}
Abstract: {abstract}

Candidate Predicates:
{choices_str}

Instructions:
- Choose the best predicate from the list that
matches the intended meaning and direction.
- If the original relationship implies negation
(e.g., "does not cause"), select the matching
base predicate, but set "negated" to "True".
- If no match exists, return '"mapped_predicate":
"none"'.

Respond with ONLY this JSON object:
{{"mapped_predicate": "one of the predicate keys
or 'none'", "negated": "True" or "False"}}
```

Figure 3: Contextual reranking prompt used in Stage 3 (Section 3.3) for LLM-based predicate selection with explicit negation handling.

incorporate the biomedical context in which the relation was expressed. Stage 3 introduces contextual refinement, where an LLM evaluates and reranks the top-k candidates against the full relation context.

The LLM is given structured prompt (Figure 3) along with the following information: the top candidate predicates retrieved by similarity-based search and the full RE quadruple (s, o, T, a) including the abstract context. By explicitly grounding the candidate predicates in this richer context, the LLM can better assess which ontology predicate is most appropriate for the specific biomedical scenario.

The LLM then performs reranking of the candidate set, selecting the single predicate that best aligns with the domain context. Importantly, the system includes a `NONE` option, allowing the model to reject all candidates if none provide a semantically valid match. This safeguard prevents the forced assignment of predicates in cases where the extracted relation is either too domain-specific, erroneous, or incompatible with the ontology.

The output of Stage 3 is a contextually validated predicate r^* that integrates both semantic similarity and biomedical context. This refinement ensures that only relationships consistent with both the ontology and the source text are admitted into the KG, thereby improving accuracy and reducing false positives.

4. Experiments

We assess RELATE on two independent datasets with different ontology models, as well as four different embedding model configurations. See Appendix D for example descriptors in both datasets.

4.1. Datasets

The ChemProt dataset (Krallinger et al., 2017) provides annotated chemical-protein relations representing diverse pharmacological and biochemical relationship types, including regulators, modulators, agonists, antagonists, and substrate interactions. We treat the information in ChemProt as ground truth abstract-predicate pairs for evaluation purposes. Standard metrics in information retrieval are used, such as exact match accuracy, accuracy@k, and Mean Reciprocal Rank (MRR). Recall from Figure 3 that Stage 3 of RELATE only outputs its top choice predicate r^* or `NONE`, and does not return a fully reranked or reordered list of k candidates. Therefore, exact match accuracy is calculated with respect to the final output of RELATE (r^*), but accuracy@k and MRR are calculated using the top-k candidates after Stage 2, before LLM-based Stage 3. This allows us to analyze the impact of LLM-based refinement in isolation.

Next, we apply RELATE to 2,400 PubMed abstracts from the HEAL Project[4] targeting opioid use disorder research. This real-world dataset reflects the practical scenario in which extracted relationships must be mapped to standardized Biolink predicates for integration into a large biomedical KG[5]. Unlike ChemProt, the HEAL abstracts lack ground-truth annotations, requiring qualitative assessment of RELATE's outputs.

Importantly, this corpus captures the complexity and noisiness of real biomedical literature. In contrast to benchmark datasets, these abstracts often contain methodological statements or negative scientific findings (e.g., "drug does not affect condition," "gene not associated with disease"). Such cases

4. https://heal.nih.gov
5. https://robokop.renci.org/

underscore the necessity of rejection and negation-handling to ensure only semantically appropriate biomedical relationships enter the KG.

4.2. LLM and Embedding Models

We experimented with MedGemma (4B), LLaMA2 (7B), LLaMA3 (8B), LLaMA3 (70B), MedGemma (27B), and OpenAI GPT-4o-mini (unknown number of parameters). The last two demonstrated greater consistency and fewer hallucinations compared to the other models. Since those two models had very similar performances, we only report results using MedGemma (27B) in this work. However, note that any other LLM could be substituted to reproduce the pipeline. MedGemma (27B) is Google's medical language model (Sellergren et al., 2025), trained exclusively on medical text and optimized for efficient inference through architectural improvements in performance and computational efficiency.

To evaluate RELATE's pipeline contribution beyond raw LLM capability, we implemented a direct classification baseline. This baseline uses MedGemma (27B) to classify ChemProt relations by providing the same input quadruple (subject, object, relation text, abstract) along with the list of 9 ChemProt predicates. This represents a common approach in LLM-based relation extraction and provides a fair comparison since it uses the same LLM and has access to the same predicate definitions.

As with the LLMs, we also experimented with several embedding approaches. While we prioritize accessibility through `ollama`-based open-access models, we tested OpenAI's `text-embed-large` models and others with 8,192-token context windows that provide general-purpose semantic understanding. These include `nomic-embed-text` (Nussbaum et al., 2025), a 768-dimensional embedding model, and `bge-m3`, a 1,024-dimensional multilingual embedding model for diverse textual contexts (Chen et al., 2024a). Since `nomic-embed-text` and `bge-m3` achieved comparable performance, we report only the results obtained with the 768-dimensional `nomic-embed-text` in this work.

For (optional) hybrid retrieval, we experimented with adding fine-tuned SapBERT (Liu et al., 2021). We initialized the model with `BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext` and optimized it using a contrastive learning framework. Specifically, we employed multi-similarity loss with hard negative mining, which encourages the model to better distinguish between semantically close but incorrect pairs. To keep training efficient while still capturing relevant semantic features, we limited input sequences to 25 tokens and represented each instance using the `[CLS]` embedding.

Training was carried out for 10 epochs with a batch size of 256 and a learning rate of 2×10^{-5}. We used mixed-precision training to reduce computational overhead and set a fixed random seed for reproducibility.

5. Results and Discussions

RELATE Pipeline is Necessary. To validate RELATE's pipeline design, we compared against direct LLM classification where MedGemma (27B) selects from the 9 ChemProt predicates without retrieval-refinement. The baseline achieves only 22.0% exact match (Table 1), compared to RELATE's 52.0% (136% improvement). This demonstrates that structured retrieval outperforms asking an LLM to select directly from a predicate list. The baseline's poor performance reflects the difficulty of simultaneous predicate understanding and selection without semantic guidance from embedding-based retrieval.

The exact match accuracy of 52% reflects the difficulty of mapping to fine-grained ontology predicates. However, this metric may understate practical utility. Manual analysis of the misclassifications show semantically acceptable near-misses, such as mapping to *agonist* instead of *modulator*. When allowing such semantically close alternatives, accuracy increases to approximately 70-80%.

LLM-based Refinement is Necessary. Table 1 illustrates how embedding choice shapes RELATE's performance on the ChemProt benchmark. Baseline configurations with `nomic` and `bge-m3` produce stronger top-rank accuracy (a@1 up to 0.528, MRR $\approx$ 0.6). In contrast, SapBERT-enhanced embeddings yield slightly higher Exact Match (0.520) and dramatically improve retrieval coverage, with a@10 reaching 0.940. This indicates that SapBERT ensures the correct ontology predicate is almost always retrieved, though often ranked lower in the list (a@1 drops to 0.1, MRR $\approx$ 0.3).

These results highlight a trade-off: baseline embeddings rank the correct answer more accurately at the top, while SapBERT embeddings broaden coverage and guarantee inclusion within candidate sets.

Table 1: RELATE performance on ChemProt. **Stage 2 (Candidate Retrieval)**: a@k = accuracy in top-k candidates *before* LLM-based refinement; MRR = mean reciprocal rank. **Post-Refinement**: Exact Match = final accuracy *after* LLM-based refinement; *Baseline* = direct LLM classification without RELATE's retrieval-refinement pipeline. LLM used is MedGemma (27B) for all.

Embedding Model	Stage 2 (Candidate Retrieval)					Post-Refinement
	a@1	a@3	a@5	a@10	MRR	Exact Match (a@1)
`nomic`	**0.528**	**0.686**	0.744	0.755	**0.612**	0.464
`bge-m3`	0.468	**0.686**	**0.788**	0.812	0.590	0.467
`nomic` + SapBERT	0.100	0.328	0.688	0.914	0.300	**0.520**
`bge-m3` + SapBERT	0.100	0.321	0.691	**0.940**	0.302	0.516
Baseline (no embedding)	-	-	-	-	-	0.220

While selecting the top-ranked candidate from embeddings without LLM refinement would yield higher exact match on this benchmark, such an approach assumes embedding similarity reliably ranks correct predicates first across all datasets and relation types. This assumption cannot be guaranteed a priori. RELATE's design prioritizes robustness through SapBERT's comprehensive retrieval combined with LLM-based contextual reranking, which serves as a buffer against cases where embedding similarity alone performs poorly. This allows RELATE to recover early precision while preserving high recall, transforming broad retrieval into precise ontology-constrained mappings regardless of how initial embeddings behave on new corpora.

When Does RELATE Reject Candidates? RELATE's rejection capability (`NONE`) represents a critical quality control mechanism. Out of 2,400 PubMed abstracts from the HEAL project, 10 relationships were appropriately rejected as unsuitable for Biolink predicate mapping (rejection rate: 0.4%). Manual analysis of these rejections reveals contextual understanding that distinguishes between legitimate biomedical relationships and other types of associations commonly found in scientific literature.

The rejected relationships fell into distinct categories reflecting non-biomedical associations: genetic nomenclature specifications (3 cases), clinical procedures and documentation (4 cases, including "Clinicians prescribe Opioid" and "Women on BUP had documented Opioid misuse"), methodological relationships (2 cases, such as "Anisomycin used to detect GNPTG"), and specialized reporting contexts (1 case). These rejections demonstrate the pipeline's ability to recognize when extracted relationships represent procedural, methodological, or administrative associations rather than biological or medical relationships suitable for KG representation.

These rejections occurred despite high embedding similarity scores, demonstrating the reranker's comprehensive contextual analysis. The low rejection rate (0.4%) suggests that the majority of extracted relationships represent valid biomedical associations while maintaining strict quality standards for edge cases requiring specialized handling. All 10 rejections were deemed appropriate (100% precision).

On ChemProt, there were 5 correct rejections out of 18 total rejections. These false rejections occur where, despite high similarity candidate scores, RELATE failed to produce mappings. Cases like "loperamide modifies corticotropin-releasing hormone action" (0.661 score for *modulator*) and "ammonium sulfate reduces CBG concentrations" (0.632 for *down-regulator*) were rejected despite appropriate candidates. While some discrepancies reflect stricter semantic criteria, others suggest overly conservative contextual reranking, highlighting evaluation challenges with traditional benchmarks. These contrasting patterns reflect differences in semantic alignment between upstream extraction and ontology predicates across datasets. See more details in Appendix C.1.

When Does RELATE Choose Negated Candidates? Out of 2,400 PubMed abstracts from the HEAL project, 77 relationships were correctly flagged as negated assertions, revealing how well the system processes the nuanced ways of reporting biomedical negative findings. Most prevalent was "not associated with", which accounted for nearly half of the

negated cases (32 instances). Beyond these straightforward cases, we found the pipeline successfully navigating more complex terrain. Other negations like "does not affect" appeared in 8 cases, while direct causal negations ("does not cause") showed up 7 times. The remaining quarter of cases presented genuine linguistic challenges. Consider phrases like "remained impaired despite treatment" or "not necessary for acute insulin-induced uptake." These require contextual understanding that goes well beyond pattern matching. The system had to recognize that "fails to affect" and "unable to deliver" carry negative semantic weight, even when the word "not" never appears. See more details in Appendix C.2.

Perhaps most importantly, RELATE maintained conceptual accuracy throughout this process. When it encountered "Drug X does not treat Disease Y," it correctly identified the underlying relationship as therapeutic (mapping to *treats*) while flagging the negation. This dual parsing of both the relationship type and its polarity proved consistent across our dataset. The most frequent base relationships were association patterns (28 cases) and effect relationships (18 cases), followed by causal relationships (8 cases). The other application capability of negation handling includes findings on use, such as which treatments fail, which genetic variants are not linked to diseases, and which interventions prove ineffective.

SAPBERT Enhancement Helps. To evaluate SapBERT's effect, we compared 2,139 predicate assignments between `bge-m3` and `bge-m3 + SapBERT`. Their disagreement patterns reveal four key improvements by SAPBERT aligned with biomedical expert judgment. Clinical terminology refinement was most prominent, with 28 cases shifting from *associated with: increased likelihood of* to *predisposes to condition*. Domain specificity enhancement appeared in 12 cases where generic *associated with* became *gene associated with condition* for relationships like "DRD4 associated with Heroin Addiction," recognizing that genetic contexts require more informative predicates.

Evidence-based caution moderated 12 strong causal claims, changing *causes* to *contributes to* for relationships like "MDMA causes Depression," reflecting the complex multifactorial causation typical in biomedicine, where direct causality claims often overstate evidence strength. Mechanistic precision enhanced molecular relationship descriptions, with 10 cases shifting from *affects: increased abundance* to *affects: increased activity or abundance* for relationships like "Betaine enhances phosphorylation of STAT3," capturing both quantitative and functional effects in biochemical interactions. These systematic patterns demonstrate that SapBERT's biomedical entity embeddings guide predicate selection toward clinically appropriate, evidence-aware terminology that better reflects expert knowledge of biomedical relationship complexity.

Limitations. RELATE's performance depends on the quality of upstream entity and relation extraction. While we do not assume perfect extraction, the input quadruples must be reasonably accurate. Errors in entity boundaries, entity type misclassifications, or poorly extracted relation phrases will propagate through the standardization pipeline. In production KG construction, additional methods address these entity errors. RELATE's modular design allows it to benefit from improvements in upstream extraction systems without requiring system redesign.

6. Conclusion

We introduced RELATE, a three-stage pipeline for automated ontology-constrained predicate mapping in biomedical relation extraction. By combining ontology-driven preprocessing, efficient vector search, and context-aware LLM-based reranking, RELATE systematically converts free-text relations into ontology-aligned and interoperable KG edges. Experiments on ChemProt demonstrate the trade-offs between early precision and retrieval depth, while application to 2,400 HEAL Project abstracts illustrates RELATE's ability to handle noisy real-world biomedical literature, including detecting negated assertions and appropriately rejecting non-biomedical associations. Together, these results establish a foundation for scalable, ontology-constrained relation extraction and highlight a path forward for enriching biomedical KGs with high-quality, standardized information. Future work includes improving computational overhead, incorporating further top-k reordering and evaluation in the refinement stage, designing a scalable evaluation that combines expert judgment with automated consistency checks, and integration into the existing large biomedical KG (ROBOKOP). As part of the integration plan, the HEAL abstract authors and domain experts are being recruited to systematically evaluate the quality and reliability of RELATE and the LLM-extracted quadruples that serve as its inputs.

Acknowledgments

We thank the LitCoin project partners for their immense contributions, and the HEAL initiative consortium for providing access to opioid-related research abstracts. Also, we acknowledge the RENCI's Data Analytics group and Knowledge Graph Working Group for constructive reviews. This research was supported by NIH #75N95023C00032 (LITCOIN).

References

Harmonizing model organism data in the alliance of genome resources. *Genetics*, 220(4):iyac022, 2022.

Chris Bizon, Steven Cox, James Balhoff, Yaphet Kebede, Patrick Wang, Kenneth Morton, Karamarie Fecho, and Alexander Tropsha. ROBOKOP KG and KGB: integrated knowledge graphs from federated sources. *Journal of Chemical Information and Modeling*, 59(12):4968–4973, 2019.

Lang Cao, Jimeng Sun, Adam Cross, et al. An automatic and end-to-end system for rare disease knowledge graph construction based on ontology-enhanced large language models: Development study. *JMIR Medical Informatics*, 12(1):e60665, 2024.

J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, et al. Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40 (3):btae104, 2024.

Emanuele Cavalleri, Mauricio Soto Gomez, Ali Pashaeibarough, Dario Malchiodi, JH Caufield, JT Reese, C Mungall, Peter N Robinson, Elena Casiraghi, Giorgio Valentini, et al. SPIREX: Improving LLM-based relation extraction from RNA-focused scientific literature using graph machine learning. In *Proceedings of Workshops at the 50th International Conference on Very Large Data Bases*, pages 1–11. VLDB. org, 2024.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024b.

Vicky Dong, Hao Yu, and Yao Chen. Graph-augmented relation extraction model with LLMs-generated support document. *arXiv preprint arXiv:2410.23452*, 2024.

Antonios Georgakopoulos, Jacco Van Ossenbruggen, and Lise Stork. From text to knowledge: Leveraging LLMs and RAG for relationship extraction in ontologies and thesauri. In *Joint of Posters, Demos, Workshops, and Tutorials of the 24th International Conference on Knowledge Engineering and Knowledge Management, EKAW-PDWT 2024*, pages 1–16. CEUR-WS, 2025.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189, 2024.

Marcin P Joachimiak, Harshad Hegde, William D Duncan, Justin T Reese, Luca Cappelletti, Anne E Thessen, and Christopher J Mungall. KG-Microbe: A reference knowledge-graph and platform for harmonized microbial information. In *International Conference on Biomedical Ontologies (ICBO)*, pages 131–133, 2021.

Navya Martin Kollapally, James Geller, Vipina Kuttichi Keloth, Zhe He, and Julia Xu. Ontology enrichment using a large language model: Applying lexical, semantic, and knowledge network-based similarity for concept placement. *Journal of Biomedical Informatics*, page 104865, 2025.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, et al. Overview of the BioCreative VI

chemical-protein interaction track. *Proceedings of the sixth BioCreative challenge evaluation workshop*, 141-146, 2017.

Changjian Li, Yang Song, and Aiping Li. GRAG-ZRE: Graph retrieval-augmented generation for zero-shot relation extraction in domain-sensitive scenarios. In *International Conference on Intelligent Computing*, pages 273–285. Springer, 2025.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 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, June 2021.

Shufan Ming, Rui Zhang, and Halil Kilicoglu. Enhancing the coverage of SemRep using a relation classification approach. *Journal of Biomedical Informatics*, 155:104658, 2024.

Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. Nomic embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Olawumi Olasunkanmi, Evan Morris, Yaphet Kebede, Harlin Lee, Stanley C Ahalt, Alexander Tropsha, and Chris Bizon. Explainable enrichment-driven graph reasoner (edgar) for large knowledge graphs with applications in drug repurposing. In *2024 IEEE International Conference on Big Data (BigData)*, pages 777–782. IEEE, 2024.

Nina G Orlova and Yuriy L Orlov. Challenges in bioinformatics education courses organization. *Biophysical Reviews*, pages 1–9, 2025.

Mounir Ourekouch, Mohammed-Amine Koulali, and Mohammed Erradi. RelCheck: Improving relation extraction with ontology-guided and LLM-based validation. In *European Semantic Web Conference*, pages 441–459. Springer, 2025.

Cheng Peng, Xi Yang, Kaleb E Smith, Zehao Yu, Aokun Chen, Jiang Bian, and Yonghui Wu. Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics*, 153: 104630, 2024.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, 2019.

Justin T Reese, Deepak Unni, Tiffany J Callahan, Luca Cappelletti, Vida Ravanmehr, Seth Carbon, Kent A Shefchek, et al. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns*, 2(1), 2021.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Yufei Shang, Yanrong Guo, Shijie Hao, and Richang Hong. Biomedical relation extraction via adaptive document-relation cross-mapping and concept unique identifier. *arXiv preprint arXiv:2501.05155*, 2025.

Yicheng Tao, Yiqun Wang, and Longju Bai. Graphical reasoning: LLM-based semi-open relation extraction. *arXiv preprint arXiv:2402.06785*, 2024.

Deepak R Unni, Sierra AT Moxon, Michael Bada, Matthew Brush, Richard Bruskiewich, J Harry Caufield, Paul A Clemons, Vlado Dancik, Michel Dumontier, Karamarie Fecho, et al. Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and Translational Science*, 15(8):1848–1855, 2022.

Neha Warikoo, Yung-Chun Chang, and Wen-Lian Hsu. LBERT: Lexically aware transformer-based bidirectional encoder representation model for learning universal bio-entity relations. *Bioinformatics*, 37(3):404–412, 2021.

Carl Yang, Ran Xu, Linhao Luo, and Shirui Pan. Knowledge graph and large language model co-learning via structure-oriented retrieval augmented generation. *IEEE Data Engineering Bulletin*, 48 (4):9–46, 2024.

Huageng Zhong, Xiao Wei, and Huiran Zhang. Leveraging LLM for enhancing document-level relation extraction with correction and completion. *arXiv preprint arXiv:2404.03702*, 2024.

Huixue Zhou, Mingchen Li, Yongkang Xiao, Han Yang, and Rui Zhang. LEAP: LLM instruction-example adaptive prompting framework for biomedical relation extraction. *Journal of Biomedical Informatics*, 143:104432, 2024.

Appendix A. Clinical and Research Applications

A.1. RELATE's Role and Significance

RELATE is a relation standardization component that ensures relations extracted from biomedical text are accurately mapped to standardized ontology predicates. This standardization is foundational for downstream knowledge graph applications.

In literature-based discovery, standardized predicates enable systematic querying across diverse sources. A researcher investigating drug repurposing can query for all compounds that upregulate proteins downregulated in a specific disease, confident that the *upregulator* predicate maintains consistent semantics across thousands of papers. Without standardization, the same relationship might appear as "increases expression," "upregulates," or "enhances activity," requiring separate queries and potentially missing relevant findings.

Biomedical knowledge graph systems such as ROBOKOP require ontology-aligned relations for reliable reasoning and inference. RELATE ensures that literature-derived evidence integrates seamlessly with structured knowledge bases, enabling complex queries that traverse multiple relationship types and data sources.

Our evaluation on HEAL demonstrates how relations from distinct domains can be mapped to common Biolink predicates, while ChemProt enables unified knowledge graphs that integrate chemical interactions.

A.2. Comparison with End-to-End Systems

Unlike end-to-end systems such as AutoRD, which perform extraction, standardization, and graph construction, RELATE focuses specifically on standardization. This provides modularity, allowing RELATE to work with any upstream extraction system, and scalability, handling 284 Biolink predicates compared to 6 relation types in AutoRD.

A.3. Broader Applicability

While evaluated on biomedical corpora, RELATE generalizes to any domains with structured predicate vocabularies for computing embeddings.

Appendix B. RELATE Stages Algorithms

Algorithms 1, 2, and 3 summarize the three stages of RELATE. When the optional SAPBERT workflow enhancement is on, RELATE follows the modified pipeline described in Figure 4.

Algorithm 1: Ontology Preprocessing

Input: Biolink Model YAML $\mathcal{B}$, Biolink Toolkit $\mathcal{T}$, Embedding model f_θ, LLM

```
R ← ExtractPredicateMappings(B, T);
D+ ← ∅;
for each predicate r ∈ R do
    descriptions ← GatherDescriptions(r);
    D+[r] ← descriptions;
end
D− ← ∅;
for each predicate p ∈ D+ do
    neg_descriptions ← ∅;
    for each description d ∈ D+[p] do
        prompt ←
          CreateNegationPrompt(d);
        d− ← LLM(prompt);
        neg_descriptions ←
          neg_descriptions ∪ {d−};
    end
    neg_key ← p + "_NEG";
    D−[neg_key] ← neg_descriptions;
end
V ← ∅;
for each predicate p ∈ D+ ∪ D− do
    descriptions ← D+[p] if p ∈ D+ else D−[p];
    combined_text ←
      CombineDescriptions(descriptions);
    v ← f_θ(combined_text);
    V ← V ∪ {(label : p, embedding : v)};
end
return Embedding store V, Positive descriptors
  D+, Negative descriptors D−;
```

Algorithm 2: Similarity-Based Retrieval

Input: Loose quadruples $\{(s_i, o_i, T_i, a_i)\}$, Predicate embeddings $\mathcal{V}$, Embedding model f_θ, Parameter k

```
// Relationship Embedding Phase
```

$\mathcal{Q} \leftarrow \emptyset$;
for *each quadruples* (s_i, o_i, T_i, a_i) **do**
  $q_i \leftarrow f_\theta(T_i)$;
  $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{(i, q_i)\}$;
end

```
// Cosine Similarity Search and Top-k
   Candidate Retrieval
```

$\mathcal{C} \leftarrow \emptyset$;
for *each* $(i, q_i) \in \mathcal{Q}$ **do**
  similarities $\leftarrow \emptyset$;
  for *each* $(p, v) \in \mathcal{V}$ **do**
    sim $\leftarrow \cos(q_i, v)$;
    similarities $\leftarrow$ similarities $\cup \{(p, \text{sim})\}$;
  end
  $\mathcal{C}[i] \leftarrow$ Top-k(similarities);
end
return Candidate sets $\mathcal{C}$;

Algorithm 3: Contextual Refinement

Input: Loose quadruples $\{(s_i, o_i, T_i, a_i)\}$, Candidate sets $\mathcal{C}$, LLM

$\mathcal{A} \leftarrow \emptyset$;
for *each quadruple* (s_i, o_i, T_i, a_i) **do**
  predicate_list $\leftarrow$ ExtractPredicates($\mathcal{C}[i]$);
  prompt $\leftarrow$ CreateRerankingPrompt((s_i, o_i, r_i, a_i), predicate_list)

```
// LLM Selection and Validation
```

  $r_i^* \leftarrow LLM$(prompt);
  if $r_i^* =$ *NONE or* $r_i^* \notin$ *predicate_list* **then**
    $r_i^* \leftarrow \emptyset$;
  end
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{(s_i, o_i, r_i^*, a_i)\}$;
end
return Mapped quadruples $\mathcal{A} = \{(s_i, o_i, r_i^*, a_i)\}$;

Appendix C. Additional Analysis

C.1. Rejected Relations

RELATE includes a rejection mechanism (`NONE` option, Section 3.3) to filter relations that cannot be reliably mapped to ontology predicates. Here, we analyze rejection patterns across both datasets.

C.1.1. HEAL Dataset Rejections

In the HEAL dataset of 2,400 abstracts, RELATE rejected 10 relationships (rejection rate: 0.4%). Manual review confirmed that all 10 rejections were appropriate (100% precision), filtering non-biomedical associations.

The rejected relationships fell into distinct categories:

1. Genetic nomenclature specifications (3 cases): Relations describing genetic naming conventions rather than biological relationships (e.g., gene symbol designations)

2. Clinical procedures and documentation (4 cases): Administrative or procedural relationships such as "Clinicians prescribe Opioid" and "Women on BUP had documented Opioid misuse."

3. Methodological relationships (2 cases): Experimental procedures such as "Anisomycin used to detect GNPTG."

4. Specialized reporting contexts (1 case): Context-specific descriptions not representing biological relationships.

These rejections demonstrate RELATE's ability to distinguish between legitimate biomedical relationships and other types of associations commonly found in scientific literature. All rejections were semantically distant from Biolink predicates, making the `NONE` decision appropriate.

C.1.2. ChemProt Dataset Rejections

Analysis of the ChemProt test set (695 cases with ground truth labels) revealed 18 instances where RELATE selected `NONE`. Of these, only 5 were correct rejections, yielding a rejection precision of 27.8% (72% false rejection rate among rejected cases).

The 13 incorrect rejections occurred despite appropriate candidate predicates being retrieved with moderate-to-high similarity scores during Stage 2 (Section 3.2). Examples include:

1. "loperamide modifies corticotropin-releasing hormone action" retrieved *modulator* with similarity score 0.661, yet was rejected in Stage 3.

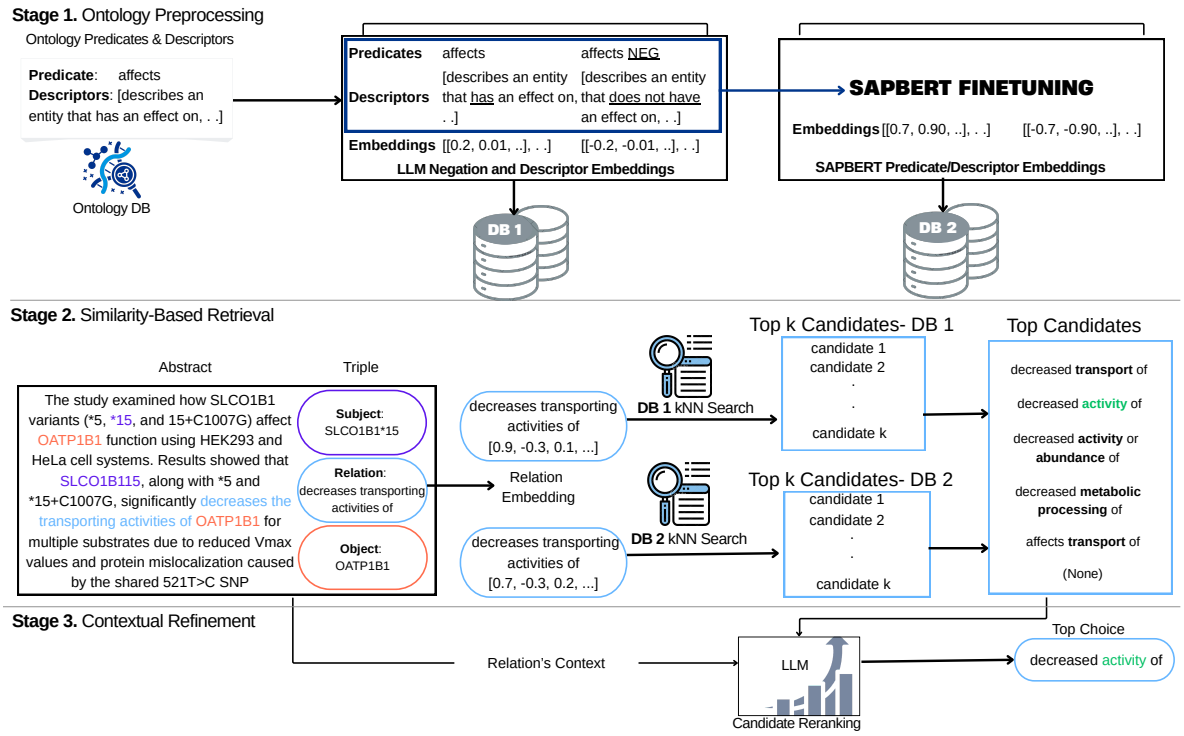

Figure 4: SapBERT-augmented RELATE pipeline as in Figure 1. The ontology preprocessing now involves both LLM-generated embeddings and SapBERT finetuned embeddings. Additional changes include dual embedding of the relation text, dual similarity searches on the stored embeddings, and merging of the two top-k candidates.

2. "ammonium sulfate reduces CBG concentrations" retrieved *downregulator* with similarity score 0.632, yet was rejected in Stage 3.

These patterns indicate that incorrect rejections stem from overly conservative contextual evaluation during LLM-based refinement in Stage 3 rather than retrieval failure in Stage 2. The system prioritizes precision by applying strict semantic criteria when assessing predicate-context fit, which reduces false positives but increases false rejections in ambiguous cases.

C.2. Negation Handling

We conducted manual evaluation of negation detection on the HEAL dataset. Of 2,411 relationship extractions from 2,400 abstracts, we identified 123 (5.1%) negated relation texts through manual review.

RELATE correctly flagged 74 as negated assertions, yielding:

1. Precision: 96.1% (74 true positives / 77 flagged as negated)
2. Recall: 60.2% (74 true positives / 123 actual negations)
3. F1-score: 74.2%

The high precision indicates that when RELATE identifies a negation, it is almost always correct (only 3 false positives). The moderate recall indicates that approximately 40% of negated statements are missed (49 false negatives), typically involving:

1. Implicit negations (e.g., "failed to affect", "unable to deliver")
2. Comparative negations (e.g., "no significant difference")

3. Complex syntactic constructions not captured by pattern matching

The negation categories in the 123 manually identified cases were:

1. Explicit negations ("not", "no", "does not"): 90 cases (73%)
2. Implicit negations ("failed to", "unable to"): 21 cases (17%)
3. Comparative negations ("no significant difference"): 12 cases (10%)

RELATE's negation handling maintains conceptual accuracy by identifying the underlying relationship type while flagging the negation polarity, enabling accurate representation of negative findings in KGs.

Appendix D. Predicate and Descriptor Distributions

ChemProt defines 10 classes for chemical-protein interactions: 9 positive relationship types and 1 negative class (*NOT/no interaction*). We utilize all 9 positive relationship types, yielding 30 positive descriptors (Table 2). ChemProt's NOT class, which indicates no interaction exists, is handled through RELATE's explicit rejection (`NONE`) mechanism rather than treating *NOT/no interaction* as a mappable predicate. This design ensures that only valid biomedical relationships are standardized to ontology predicates while filtering non-relationships before KG integration.

After the preprocessing stage (Section 3.1), the ontology database stores both positive predicates and descriptors and their LLM-negated versions. For ChemProt, this yields 18 stored predicates (9 positive + 9 NEG variants) with 60 total descriptor texts (30 positive + 30 negated). For Biolink, this yields 568 stored predicates (284 positive + 284 NEG variants) with 2,208 total descriptor texts (1,104 positive + 1,104 negated).

Tables 2 and 3 show the descriptor distribution for the most frequent predicates. In ChemProt (Table 2), *substrate* and *agonist* together account for over one-third of all positive descriptors (36.67%). In Biolink (Table 3), the top 10 predicates account for 41.30% of all positive descriptors, with *coexists with* (16.12%) and *related to* (8.70%) being the most prevalent. The remaining 274 Biolink predicates account for 58.70% of descriptors, demonstrating the diversity of biomedical relationship types captured in the ontology.

Table 2: ChemProt Dataset. All 9 positive relationship predicates ordered by descriptor count. ChemProt's 10th class (NOT/no interaction) is handled via RELATE's rejection mechanism (`NONE` option, Stage 3) rather than predicate mapping.

Predicate	**Sample Descriptors**	**Descriptor count (%)**
substrate	Product of, Substrate product of	6 (10.00%)
agonist	Agonist-activator, Agonist-inhibitor	5 (8.33%)
upregulator	Activator, Indirect upregulator	4 (6.67%)
downregulator	Inhibitor, Indirect downregulator	4 (6.67%)
regulator	Direct regulator, Indirect regulator	3 (5.00%)
part of	Indicates that the chemical is a structural component or subunit of the protein complex	2 (3.33%)
antagonist	Chemical substance that binds to and blocks the activation of certain receptors on cells, preventing a biological response	2 (3.33%)
modulator	Modulator-activator, Modulator-inhibitor	2 (3.33%)
cofactor	A non-protein chemical compound that is required for an enzyme's biological activity	2 (3.33%)

Table 3: Biolink Dataset. Top 10 of 284 predicates by descriptor count. Percentages calculated from 1,104 total positive descriptors across all 284 Biolink predicates used in this study.

Predicate	**Sample Descriptors**	**Descriptor count (%)**
coexists with	holds between two entities that are co-located in the same aggregate object, process, or spatio-temporal region	178 (8.06%)
related to	holds between two entities	96 (4.39%)
located in	holds between a material entity and a material entity or site within which it is located (but of which it is not considered a part)	40 (1.81%)
part of	holds between parts and wholes (material entities or processes)	34 (1.49%)
temporally related to	holds between two entities with a temporal relationship	30 (1.36%)
has part	holds between wholes and their parts (material entities or processes)	21 (0.95%)
precedes	holds between two processes, where one completes before the other begins	15 (0.68%)
causes	holds between two entities where the occurrence, existence, or activity of one causes the occurrence or generation of the other	15 (0.68%)
affects	Describes an entity that has an effect on the state or quality of another existing entity	14 (0.63%)
has output	holds between a process and a continuant, where the continuant is an output of the process	13 (0.63%)