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

Mathew Satursky msatursky@renci.org
Hong Yi hongyi@renci.org
Chris Bizon cbizon@renci.org
Renaissance Computing Institute, 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. We introduce RELATE, a three-stage pipeline that maps LLM-extracted relations to standardized ontology predicates using ChemProt and 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 transforms 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 Protocol, 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/tree/multi-ontology

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 knowledge graph 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.

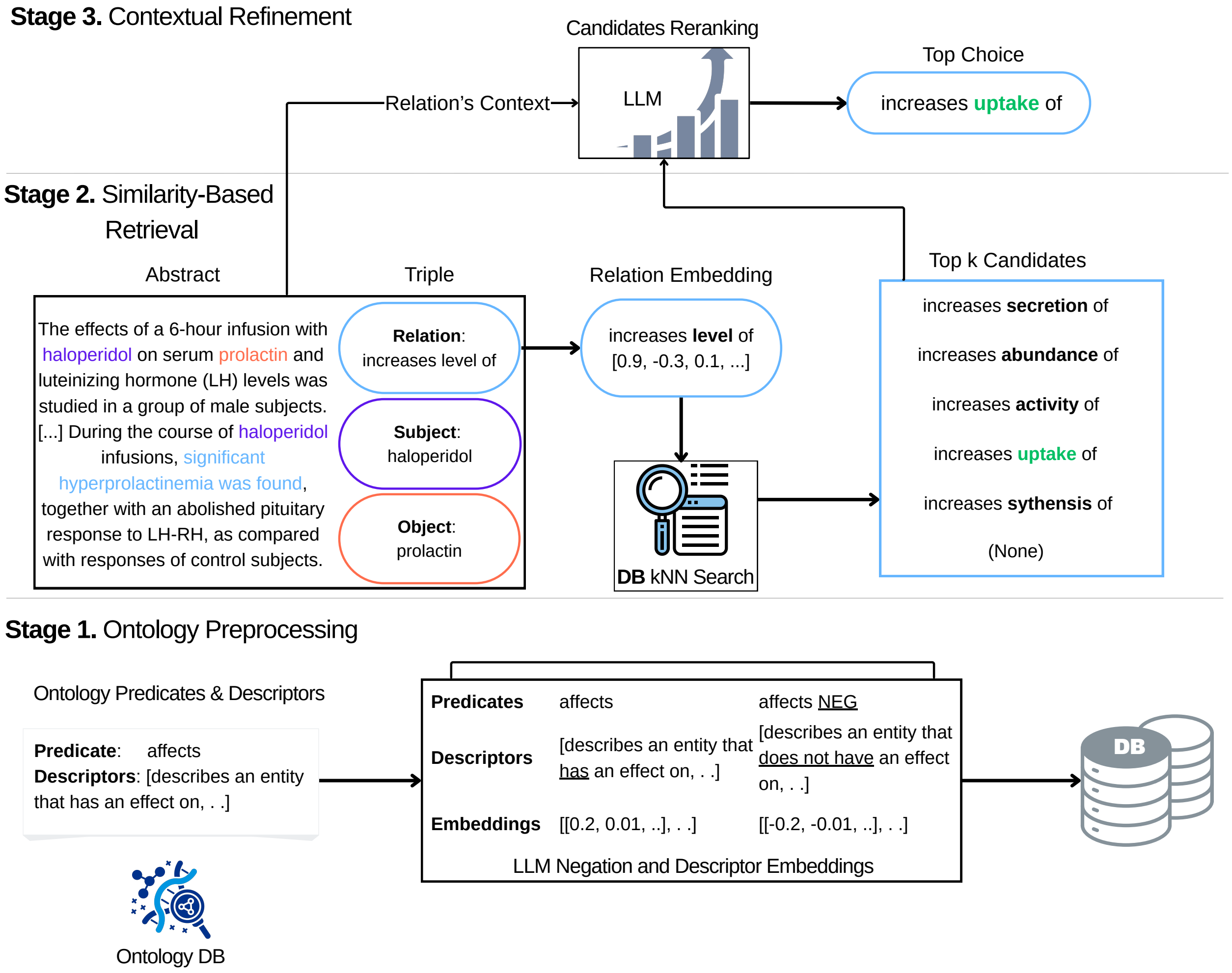
3. **Comprehensive evaluation:** We conduct dual-ontology assessment on ChemProt (Peng et al., 2019) benchmark and 2,400 real-world HEAL Project abstracts. Integration with an established biomedical KG (ROBOKOP) is also planned to further demonstrate its translational impact.

2. Related Work

No existing framework systematically bridges free-text biomedical 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., 2024a; 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.

Recent work in LLM-based relation extraction has explored various validation and standardization approaches. RelCheck (Ourekouch et al., 2025) addresses low-confidence predictions from pre-trained models using ontology-guided 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 such as RAG-enhanced LLMs for automated ontology extension (Georgakopoulos et al., 2025); structure-oriented RAG for KG-LLM co-learning (Yang et al., 2024); and integrated GNNs with LLM-generated support documents (Dong et al., 2024), and LLM-driven ontology enrichment pipelines (Kollapally et al., 2025).



The effects of a 6-hour infusion with haloperidol on serum prolactin and luteinizing hormone (LH) levels was studied in a group of male subjects. [...] During the course of haloperidol infusions, significant hyperprolactinemia was found, together with an abolished pituitary response to LH-RH, as compared with responses of control subjects.

Predicate: affects
Descriptors: [describes an entity that has an effect on, . .]

Predicates	affects	affects NEG
Descriptors	[describes an entity that has an effect on, . .]	[describes an entity that does not have an effect on, . .]
Embeddings	[[0.2, 0.01, ..], . .]	[[-0.2, -0.01, ..], . .]

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 triple with 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.

However, these approaches primarily target domain-specific extraction, confidence validation, or knowledge system extension 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) applied structured prompt interrogation for ontology-grounded annotations from scientific publications, yet their focus remained on entity linking rather than comprehensive relation standardization. Even enhancements to systems such as SemRep (Ming et al., 2024) that leverage contrastive learning operate within existing predicate frameworks rather than enabling mapping to broader ontological schemas.

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 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 Table 2 and Table 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 natural, semantically coherent negations using a structured prompt (Figure 2). For example, the earlier descriptor becomes "describes an entity that does not have 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* $\rightarrow$ *affects_NEG*) to maintain clear separation from positive instances.

Finally, every *descriptor* in D_r^+ and D_r^- is transformed into d-dimensional vector representations 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 (Liu et al., 2021), this work presents its first application to relation extraction through predicate embedding enhancement. Our hybrid approach leverages SapBERT's biomedical do-

```
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 3.1 for creating negative descriptors from positive descriptors.

main 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 extracted relations.

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, T could be “increases level of,” which does not exist in Biolink ontology in this format. First, free-form relation text T is embedded to 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.

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, focusing only on lexical and semantic similarity.

3.3. Stage 3: Contextual Refinement

```
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 for LLM-based predicate selection with explicit negation handling.

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 incorporate the biomedical context in which the relation was expressed. Stage 3 introduces contextual refinement, where an LLM evaluates and re-ranks 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 re-ranking 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 knowledge graph, 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.

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 re-ordered 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 will allow us to analyze the effect of LLM-based refinement in isolation as well.

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 application scenario where relationships extracted from literature 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 underscore the necessity of RELATE's rejection and negation-handling capabilities to ensure only semantically appropriate biomedical relationships enter the KG. As part of the integration plan, the abstract authors and Subject Matter Experts (SMEs) are being recruited to systematically evaluate the quality and reliability of the extracted triples.

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 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.

[4] https://heal.nih.gov

[5] Name hidden during double blind review process.

Table 1: RELATE performance on ChemProt. Exact Match: accuracy of RELATE; a@k: accuracy in top-k candidates *before* LLM-based refinement in Stage 3 of RELATE; MRR: Mean reciprocal rank.

LLM	Embedding	Exact Match	a@1	a@3	a@5	a@10	MRR
MedGemma 27B	nomic	0.464	**0.528**	**0.686**	0.744	0.755	**0.612**
MedGemma 27B	bge-m3	0.467	0.468	**0.686**	**0.788**	0.812	0.590
MedGemma 27B	nomic + SapBERT	**0.520**	0.100	0.328	0.688	0.914	0.300
MedGemma 27B	bge + SapBERT	0.516	0.100	0.321	0.691	**0.940**	0.302

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` (Zach et al.), a 768-dimensional embedding model, and `bge-m3`, a 1,024-dimensional multilingual embedding model for diverse textual contexts (Chen et al., 2024b). 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

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. RELATE resolves this through its LLM-based refinement stage, which reranks SapBERT's expanded candidate lists in context. This allows RELATE to recover early precision while preserving high recall, transforming broad retrieval into precise ontology-constrained mappings. In short, SapBERT enhances retrieval coverage, but RELATE's LLM-based refinement is essential to convert this coverage into accurate, ontology-aligned relations.

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%). 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%) indicates that the majority of extracted relationships represent valid biomedical associations while maintaining strict quality standards for edge cases requiring specialized handling.

Chemprot analysis also reveals potential false rejections where high similarity scores failed to produce mappings. Cases like "loperamide modifies corticotropin-releasing hormone actio" (0.661 score for "modulator") and "ammonium sulfate reduces CBG concentrations" (0.632 for "downregulator") were rejected despite appropriate candidates. While some discrepancies reflect stricter semantic criteria, others suggest overly conservative contextual reranking, highlighting evaluation challenges with traditional benchmarks.

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.

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 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.

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 reranking, RELATE systematically converts free-text relations into ontology-aligned, interoperable knowledge graph 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 identification of negated assertions and appropriate rejection of non-biomedical associations. Together, these results establish a foundation for scalable, ontology-constrained relation extraction and highlight a path forward for enriching biomedical knowledge graphs with high-quality, standardized information. Future work includes improving computational overhead, designing a scalable evaluation that combines expert judgment with automated consistency checks, and integration into the existing large biomedical KG (ROBOKOP).

Acknowledgments

We thank the LitCoin project partners for their immense contributions, the HEAL initiative consortium for providing access to opioid 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.

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.

J. Chen, H. Lin, X. Han, and L. Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024a.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, 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.

I. Jahan, M. T. R. Laskar, C. Peng, and J. X. 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 *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.

M. Krallinger, O. Rabal, and A. Lourenço. Overview of the biocreative vi chemical-protein interaction track. *Proceedings of the BioCreative VI 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.

F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. 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.

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.

Y. Peng, S. Yan, and Z. 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, Benjamin M Good, James P Balhoff, Tommaso Fontana, 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.

Y. Shang, Y. Guo, S. Hao, and R. Hong. Biomedical relation extraction via adaptive document-relation cross-mapping and concept unique identifier. *arXiv preprint arXiv:2501.05155*, 2025.

Y. Tao, Y. Wang, and L. Bai. Graphical reasoning: Llm-based semi-open relation extraction. *arXiv preprint arXiv:2402.06785*, 2024.

D. R. Unni, S. A. T. Moxon, M. Bada, M. Brush, R. Bruskiewich, J. H. Caufield, P. A. Clemons, V. Dancik, M. Dumontier, K. 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*, 2024.

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

H. Zhong, X. Wei, and H. Zhang. Leveraging llm for enhancing document-level relation extraction with correction and completion. *arXiv preprint arXiv:2404.03702*, 2024.

H. Zhou, M. Li, Y. Xiao, H. Yang, and R. Zhang. Leap: Llm instruction-example adaptive prompting framework for biomedical relation extraction. *Journal of Biomedical Informatics*, 143:104432, 2024.

Appendix A. Relation Extraction Pipeline Stages Algorithms

Algorithm 1: Ontology Preprocessing

```
Input: Biolink Model YAML $\mathcal{B}$, Biolink Toolkit $\mathcal{T}$, LLM $\mathcal{M}$, Embedding model $f_\theta$
Output: Embedding store $\mathcal{V}$, Positive descriptors $\mathcal{D}^+$, Negative descriptors $\mathcal{D}^-$
$\mathcal{R} \leftarrow$ ExtractPredicateMappings($\mathcal{B}, \mathcal{T}$);
$\mathcal{D}^+ \leftarrow \emptyset$;
for each predicate mapping $r \in \mathcal{R}$ do
    predicate ← $r$.predicate;
    descriptions ← GatherDescriptions(predicate);
    $\mathcal{D}^+$[predicate] ← descriptions;
end
$\mathcal{D}^- \leftarrow \emptyset$;
for each predicate $p \in \mathcal{D}^+$ do
    pos_descriptions ← $\mathcal{D}^+[p]$;
    neg_descriptions ← $\emptyset$;
    for each description $d \in$ pos_descriptions do
        prompt ← CreateNegationPrompt($d$);
        $d^{neg} \leftarrow \mathcal{M}$(prompt);
        neg_descriptions ← neg_descriptions $\cup \{d^{neg}\}$;
    end
    neg_key ← $p$ + "_NEG";
    $\mathcal{D}^-$[neg_key] ← neg_descriptions;
end
$\mathcal{V} \leftarrow \emptyset$;
for each predicate $p \in \mathcal{D}^+ \cup \mathcal{D}^-$ do
    descriptions ← $\mathcal{D}^+[p]$ if $p \in \mathcal{D}^+$ else $\mathcal{D}^-[p]$;
    combined_text ← CombineDescriptions(descriptions);
    $\mathbf{v}_p \leftarrow f_\theta$(combined_text);
    $\mathcal{V} \leftarrow \mathcal{V} \cup \{(\text{label} : p, \text{embedding} : \mathbf{v}_p)\}$;
end
return $\mathcal{V}, \mathcal{D}^+, \mathcal{D}^-$;
```

Algorithm 2: Similarity-Based Retrieval

```
Input: Loose triples $\mathcal{T} = \{(s_i, o_i, r_i, a_i)\}_{i=1}^n$, Predicate embeddings $\mathcal{V}$, Embedding model $f_\theta$, Top-k parameter $k$
Output: $\mathcal{C}$
// Relationship Embedding Phase
$\mathcal{R}_{embed} \leftarrow \emptyset$;
for each triple $(s_i, o_i, r_i, a_i) \in \mathcal{T}$ do
    $\mathbf{v}_{r_i} \leftarrow f_\theta(r_i)$;
    $\mathcal{R}_{embed} \leftarrow \mathcal{R}_{embed} \cup \{(i, \mathbf{v}_{r_i})\}$;
end
// Cosine Similarity Search and Top-k Candidate Retrieval
$\mathcal{C} \leftarrow \emptyset$;
for each $(i, \mathbf{v}_{r_i}) \in \mathcal{R}_{embed}$ do
    similarities ← $\emptyset$;
    for each $(label_j, \mathbf{v}_j) \in \mathcal{V}$ do
        $\text{sim}_j \leftarrow \cos(\mathbf{v}_{r_i}, \mathbf{v}_j) = \frac{\mathbf{v}_{r_i} \cdot \mathbf{v}_j}{\|\mathbf{v}_{r_i}\| \|\mathbf{v}_j\|}$;
        similarities ← similarities $\cup \{(\text{label}_j, \text{sim}_j)\}$;
    end
    sorted ← SortDescending(similarities);
    candidates$_i$ ← sorted[1 : $k$];
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{(i, \text{candidates}_i)\}$;
end
return $\mathcal{C}$;
```

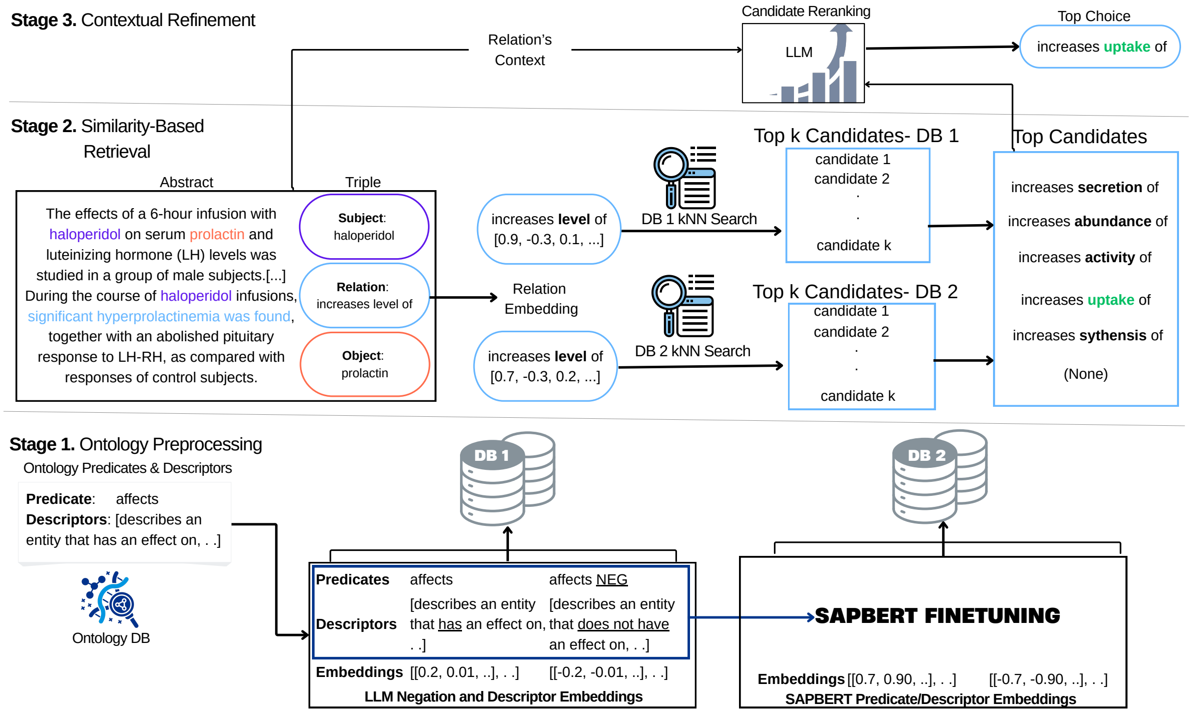

Stage 3. Contextual Refinement
Relation's Context
Candidate Reranking
LLM
Top Choice
increases uptake of
Stage 2. Similarity-Based Retrieval
Abstract
Triple
The effects of a 6-hour infusion with haloperidol on serum prolactin and luteinizing hormone (LH) levels was studied in a group of male subjects.[...] During the course of haloperidol infusions, significant hyperprolactinemia was found, together with an abolished pituitary response to LH-RH, as compared with responses of control subjects.
Subject: haloperidol
Relation: increases level of
Object: prolactin
Relation Embedding
increases level of [0.9, -0.3, 0.1, ...]
increases level of [0.7, -0.3, 0.2, ...]
DB 1 kNN Search
DB 2 kNN Search
Top k Candidates- DB 1
Top k Candidates- DB 2
candidate 1
candidate 2
candidate k
Top Candidates
increases secretion of
increases abundance of
increases activity of
increases uptake of
increases sythensis of
(None)
Stage 1. Ontology Preprocessing
Ontology Predicates & Descriptors
Predicate: affects
Descriptors: [describes an entity that has an effect on, . .]
Ontology DB
DB 1
DB 2
Predicates affects affects NEG
Descriptors [describes an entity that has an effect on, . .] [describes an entity that does not have an effect on, . .]
Embeddings [[0.2, 0.01, ..], . .] [[-0.2, -0.01, ..], . .]
LLM Negation and Descriptor Embeddings
SAPBERT FINETUNING
Embeddings [[0.7, 0.90, ..], . .] [[-0.7, -0.90, ..], . .]
SAPBERT Predicate/Descriptor Embeddings

Figure 4: SapBERT Augmented System workflow diagram showing the three-stage pipeline as in Figure 1: (1) Ontology Preprocessing, (2) Similarity-Based Retrieval, and (3) Contextual Refinement. Just as in Figure 1, stage 3.1 changes only if the ontology schema or the LLM and embedding Model changes. However, the one-stage preprocessing of raw ontology predicate data involves two embedding generations- LLM-generated embedding and SapBERT finetuned stored embedding. Also, given a raw input triple with abstract context, the frequently executed stage—similarity-based retrieval and contextual refinement—encompasses the input relation's dual embedding transformation, dual similarity search on the pair of stored embeddings, merger of top-k candidates from each search, and final contextual reranking.

Appendix B. SAPBERT Workflow Enhancement

Appendix C. Predicate and Descriptor Distribution

Table 2 and Table 3 showed the top 10 predicates that account for the majority of descriptors in the dataset. *substrate* and *agonist* together accounting for over one-third of all descriptors in Table 2 while Table 3 dominated by predicates such as *coexists with* and *related to*.

Algorithm 3: Contextual Refinement

```
Input: Loose triples $\mathcal{T} = \{(s_i, o_i, r_i, a_i)\}_{i=1}^{n}$,
       Candidate sets $\mathcal{C}$, LLM $\mathcal{M}$
Output: Mapped triples
        $\mathcal{T}^* = \{(s_i, o_i, r_i^*, a_i)\}_{i=1}^{n}$
$\mathcal{T}^* \leftarrow \emptyset$;
for each triple $(s_i, o_i, r_i, a_i) \in \mathcal{T}$ do
    candidates$_i \leftarrow \mathcal{C}[i]$;
    predicate_list ← ExtractPredicates(candidates$_i$);
    // Prompt Construction
    prompt ← ”Given the biomedical relationship context:”;
    prompt ← prompt + ” Subject: ” + $s_i$ + ” Object: ” + $o_i$;
    prompt ← prompt + ” Relationship: ” + $r_i$ + ” Abstract: ” + $a_i$;
    prompt ← prompt + ” Select best Biolink predicate or NONE:”;
    for $j = 1$ to |predicate_list| do
        prompt ← prompt + $j$ + ”. ” + predicate_list[$j$];
    end
    prompt ← prompt + (|predicate_list| + 1) + ”. NONE”;
    // LLM Selection and Validation
    response ← $\mathcal{M}$(prompt);
    $r_i^*$ ← ParseLLMSelection(response, predicate_list);
    if $r_i^* =$ ”NONE” or $r_i^* \notin$ predicate_list then
        $r_i^* \leftarrow \emptyset$;
    end
    $\mathcal{T}^* \leftarrow \mathcal{T}^* \cup \{(s_i, o_i, r_i^*, a_i)\}$;
end
return $\mathcal{T}^*$;
```

Table 2: Chemprot Dataset Top Predicates by Descriptor Count and Percentage Contribution

Predicate	**Descriptors**	**Percentage (%)**
substrate	6	20.00
agonist	5	16.67
upregulator	4	13.33
downregulator	4	13.33
regulator	3	10.00
part of	2	6.67
antagonist	2	6.67
modulator	2	6.67
cofactor	2	6.67

Table 3: Biolink Dataset Top Predicates by Descriptor Count and Percentage Contribution

Predicate	**Descriptors**	**Percentage (%)**
coexists with	178	39.04
related to	96	21.05
located in	40	8.77
part of	34	7.46
temporally related to	30	6.58
has part	21	4.61
precedes	15	3.29
causes	15	3.29
affects	14	3.07
has output	13	2.85