Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 3 (2025)

Ontopharmsearch: A Scalable Ontology-Driven Semantic Search Framework For Contextual Real-World Evidence Discovery In Pharmaceutical Research

Sandeep R Diddi[1], Dr Rajesh Sharma R[2]

[1]Alliance College of Engineering and Design, Alliance University, Bangalore, India
Email: dsandeepPHD724@ced.alliance.edu.in
[2]Alliance College of Engineering and Design, Alliance University, Bangalore, India
Email: rajeshsharma.r@alliance.edu.in

Abstract
The appearance of real-world evidence (RWE) as an addition to randomized controlled trials has revolutionized pharmaceutical research, but the heterogeneous and unstructured nature of real-world data (RWD) constrains conventional keyword-based retrieval systems. These approaches tend to lack biomedical semantics and contextual relationships, which limit the discovery of evidence. To overcome these shortcomings, this paper presents OntoPharmSearch, a scalable semantic search framework based on ontologies (SNOMED CT, UMLS, MeSH) and transformer-based embeddings (BERT, BioBERT) to ensure precise and context-aware search of RWE. The proposed framework combines metadata ingestion, structural standardization, concept normalization, entity disambiguation, ontology alignment, semantic enrichment, and vector embedding generation with approximate nearest-neighbour retrieval to improve interpretability, scalability and efficiency. Experimental analysis on heterogeneous datasets including EHRs, clinical trial registries, and drug safety databases showed better results, with 96.2% accuracy, 96.8% precision and 95.5% recall. These findings imply that OntoPharmSearch is effective in eliminating noise and irrelevant retrieval as well as in providing comprehensive evidence discovery to pharmaceutical research. By overcoming semantic gaps and allowing transparent query processing, the framework brings a considerable improvement over traditional search systems, which are used to support high-quality, context-sensitive decision-making in healthcare and pharmaceutical fields.

Keywords: OntoPharmSearch, semantic search, ontology-driven retrieval, real-world evidence, biomedical ontologies, transformer embeddings, pharmaceutical research.

1. Introduction
The emergence of RWE as a valuable complement to randomized controlled trials has deservedly changed the face of pharmaceutical research. Real-world data (RWD) from several sources, such as electronic health records (EHRs), insurance claims, patient registries, and digital health devices, affords unprecedented insights into drug effectiveness and safety, patient adherence, and broader health outcomes in different clinical settings [1]. Nonetheless, the ability to generate actionable insights from this ocean of unstructured and heterogeneous data poses challenges [2]. The traditional keyword-based search systems tend to underperform when applied to biomedical literature and clinical data due to their inability to capture domain-specific semantics and contextual relevance [3]. Faced with this challenge, ontology-based frameworks are becoming strong able to act as a bridge for bridging the semantic gap, allowing for the retrieval of precise and contextually relevant pieces of information [4]. One such system is OntoPharmSearch, a scalable semantic search framework developed to leverage domain ontologies and artificial intelligence to discover contextual

RWE in pharmaceutical research [5]. Building OntoPharmSearch utilizes biomedical ontologies such as SNOMED CT, RxNorm, MeSH, and others to model complex relations among various medical concepts: diseases, drugs, mechanisms of action, and clinical outcomes [6]. By allowing the semantic enrichment of unstructured texts using these ontologies, it helps in inferring relations beyond exact keyword matches [7]. In the field of biomedical information retrieval, there are several works looking at the synergy of ontologies with retrieval methods: for instance, [8]. Other notable strategies involve the use of BioPortal semantic services for literature discovery and semantic search engines such as EBI's OLS for biomedical term retrieval. ML models BERT and BioBERT, on the other hand, were applied in assistance to ontological annotations for increasing the precision of retrieval of biomedical documents [9]. This partnership thus provides a good proof of concept of how the combination of domain-specific ontologies with intelligent search techniques is able to unearth relevant evidence from real-world datasets that have otherwise escaped conventional approaches [10].

As is clear, in the past two decades, there has been significant progress in ontology-based search systems; however, several critical limitations continue to persist [11]. The majority of existing frameworks suffer from scalability issues, rendering them unable to operate effectively on large-scale heterogeneous datasets that are usually encountered in pharmaceutical research [12]. Most of them are narrowly targeted, only to solve specific cases like the detection of adverse drug reactions and gene-disease associations, which affects their generalizability [13]. Ontological reasoning is considered in certain approaches, yet few consider the context of RWD in terms of its temporal, geographic, and demographic interpretations. Furthermore, due to the rapid development of medical terminologies and continuous changes in clinical practices, flexible frameworks are required to adapt and update ontological mappings over time [14]. Another major disadvantage is that the interpretability of search outcomes is low, as many existing systems function as black boxes that offer no transparency to the reasoning behind their search results. These limitations point toward the demand for a robust, scalable, and context-aware OntoPharmSearch framework that performs real-world pharmaceutical research, needs-based, transparent ontology-driven semantic search [15]. The main contribution of this research is as follows:

- The OntoPharmSearch system combines domain-specific biomedical ontologies (e.g., SNOMED CT, UMLS, MeSH), cutting-edge AI technologies, and transformer-based language models (e.g., BERT, BioBERT) to provide context-sensitive, high-precision semantic retrieval of RWE across non-uniform biomedical datasets.
- Within the framework, robust schema alignment (Dublin Core), concept normalization, and entity disambiguation are used, and provide ontologically-based reasoning-supported semantic enrichment of unstructured metadata into contextually enriched, interoperable units of knowledge.
- By semantic enrichment of metadata as dense vectors and supporting rapid and precise similarity-based retrieval through ANN algorithms (e.g., HNSW), the ontopharmsearch system also supports fuzzy matching, semantic query expansion, and probabilistic ranking mechanisms for user-interactive queries in the semantic search interface.

This paper is organized as follows: Section 2 provides an extensive literature review, Section 3 describes the proposed OntoPharmSearch method, Section 4 discusses the performance metrics and results evaluation, and Section 5 summarizes the key findings and future research directions for the study.

2. Literature Survey

In 2024, Harika et al. [16] used a scalable ontology-driven data mining technique. It employed adaptive algorithms coupled with the semantic web technologies to carry out real-time analysis of the IoT data streams, in which usage gives speed in decision making and enhanced throughput of data, but encounters challenges due to requiring domain-oriented ontology design and unstructured data processing.

In 2025, Stănescu & Oprea [17] employed topic modelling techniques to provide insights on 10,037 articles on the topic of research related to the semantic web and ontology, amongst other themes, including ontology engineering and bioinformatics. The study revealed meaningful thematic insights with a coherence score of 0.75 but faced challenges operationally in terms of dynamic ontology updates and Big Data scalability.

In 2025, Fareedi et al. [18] adopted Federated Virtual Knowledge Graph (FVKG) using the ODSRE methodology and the Ontop semantic query engine for achieving semantic interoperability and data integration in the healthcare domain. It improved real-time access and reduced data migration, but the limitations were the high complexity of schema mapping and the need for well-defined ontologies.

In 2025, Whiteney [19] developed semantic interoperability techniques that comprised encoding heterogeneous data using standards like RDF and OWL, ontology mapping, semantic annotations, and knowledge graphs. Thus, it has shown effective implication in areas such as healthcare and smart cities despite being limited by issues such as schema heterogeneity, ambiguity, and scalability.

In Shah & Ishfaq [20] used big data integration and information alignment with AI-based computational techniques in medicinal chemistry, pharmacology, and toxicology to improve drug discovery and efficacy and thus personalized medicine. It delineated the opportunity through the FAIR data principles, but limitations arose in data privacy, heterogeneous sources, and lack of standardization, thus suggesting robust data management and interoperability as a near-future focus.

In 2023, Yu et al. [21] created through a study that was concerned with an Automated Semantic ML Microservice wherein ontology-driven automation and self-supervised reinforcement learning were combined to automate the development of ML models for biomedical studies. This directly ensures improved adaptability and interpretability across tasks, though it was going to be very dependent on well-structured domain ontologies and was supposed to be quite complicated in handling a wide array of biomedical data scenarios.

In 2023, Taglino et al. [22] showed that ontology-based computation modelling and semantic querying of Alzheimer's Disease were within the premise of this study on the ADNI repository. This naturally offered very intuitive yet very complex ways in which such data able to be extracted. It facilitated more accessible integration of the data and their storage, but fell short with the requirements of constant ontology updates and management of heterogeneous biomedical data sources.

In 2022, Sousa et al. [23] developed K-BiOnt, a hybrid technique for biomedical relation extraction (RE) that combines deep learning and a knowledge graph-based recommendation system to effectively extract relations between entities in biomedical domains. K-BiOnt enhances true relations detection missed by baseline models, but has limitations due to its dependence on the quality and completeness of external knowledge graphs.

In 2022, Dörpinghaus et al. [24] utilized a context-aware knowledge graph approach through labelled property graphs and a polyglot persistent system for data mining and querying biomedical texts enriched through text mining and a field-specific PubMed and SCAIView

language. On graph queries, however, there have been severe technological difficulties in the areas of storage and querying of huge graphs housing more than 71 million nodes and an additional 850 million relationships.
In 2023, Sousa et al. [25] integrated aspect-oriented similarity measures for different biological considerations, such as protein interaction and phenotype-based gene similarity. It has performed better than some unsupervised methods but needs labelled data and expert-defined views for efficient training and interpretation.

2.1 Problem Statement

Although there have been quite lots of developments in ontology-based frameworks, semantic interoperability, or even biomedical systems integrated with AI, sharing heterogeneous, unstructured, and large-scale biomedical data and IoT data has not been easy. As found in various literature reviews, the trend continues to showcase various challenges, including the reliance on specific domain ontologies, difficulties with schema mapping, lack of standard standards, problems with scalability in big data environments, and limited adaptability of models to dynamic updates. All these issues show a real need for a unified, scalable, and flexible semantic framework guaranteeing interoperability, querying efficiency, and precise knowledge extraction across heterogeneous healthcare and pharmaceutical domains.

Table 1: Summary of Recent Ontology-Driven and Semantic Techniques in Biomedical and IoT Data

Author & Year	Technique	Aim	Significance	Disadvantages
Harika et al. (2024) [16]	Ontology-driven data mining	Real-time IoT data analysis	Improved decision speed and throughput	Domain-specific ontologies are required; unstructured data
Stănescu & Oprea (2025) [17]	Topic Modeling (LDA, BERT)	Semantic web research trends	Identified 3 key clusters with 0.75 coherence	Ontology update & scalability issues
Fareedi et al. (2025) [18]	FVKG + Ontop (ODSRE)	Healthcare data integration	Reduced migration, real-time access	Complex schema mapping; ontology dependency
Whiteney (2025) [19]	RDF, OWL, Ontology Mapping	Semantic interoperability in systems	Effective in smart cities and healthcare	Ambiguity, schema heterogeneity
Shah & Ishfaq (2025) [20]	Big Data + AI (FAIR)	Drug development & personalized medicine	Improved discovery and efficacy	Privacy, standardization, heterogeneity
Yu et al. (2023) [21]	Ontology-based Semantic ML Framework	Automated biomedical ML	Adaptive, interpretable models	Complex for varied biomedical scenarios
Taglino et al. (2023) [22]	Ontology-based ADNI Querying	Alzheimer's data extraction	Enhanced data integration &	Needs regular updates, data

			query support	heterogeneity
Sousa et al. (2022) [23]	K-BiOnt (RE + KG Recommendation)	Biomedical relation extraction	Detected hidden entity relationships	Dependent on external KG quality
Dörpinghaus et al. (2022) [24]	Context-aware Knowledge Graph	Biomedical graph mining	Supported massive-scale querying	Storage and query complexity
Sousa et al. (2023) [25]	Supervised Semantic Similarity	Biological similarity prediction	Higher accuracy than unsupervised methods	Requires expert views and labeled data

Table 1 shows a concise overview of recent studies conducted in biomedical, pharmaceutical, and IoT domains employing ontology-driven, semantic, and AI-based techniques. It provides methodology, purpose, importance, and significant limitations of each work, helping in identifying current trends and research challenges.

3. Proposed Methodology

This article outlines the development of OntoPharmSearch, a new semantic search framework designed to address issues with retrieving contextual RWE from diverse biomedical sources. The main goal of this work is to enhance precision, scalability, and context in RWE discovery. Current systems face limitations in scalability, schema heterogeneity, reuse of ontologies, and, in some cases, poor interpretability and data masking by users. OntoPharmSearch offers a method that integrates an ontology-based and AI-driven architecture to significantly improve semantic interoperability, query efficiency, and knowledge extraction, enabling a substantial impact on pharmaceutical research. Figure 1 shows the proposed architecture of the OntoPharmSearch model.

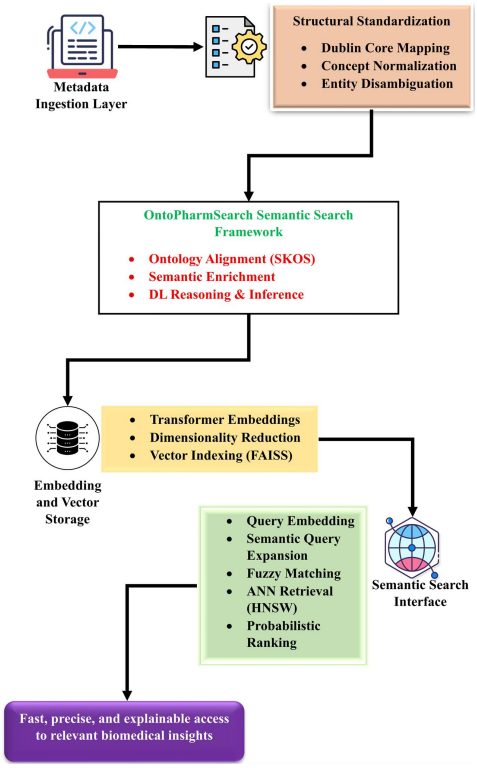

Figure 1. Architecture of the proposed OntoPharmSearch

3.1 Metadata Ingestion and Standardization

As the foundational layer of OntoPharmSearch, the metadata ingestion process is intended to aggregate and then logically harmonize heterogeneous RWE metadata residing within distributed, siloed repositories, such as Electronic Health Records (EHRs), clinical trial registries, drug databases, and adverse event reports. Due to the inconsistent schema and terminology used in the metadata provided, a robust standardisation process is necessary. The architecture of the Metadata Enrichment Pipeline is displayed in Figure 2.

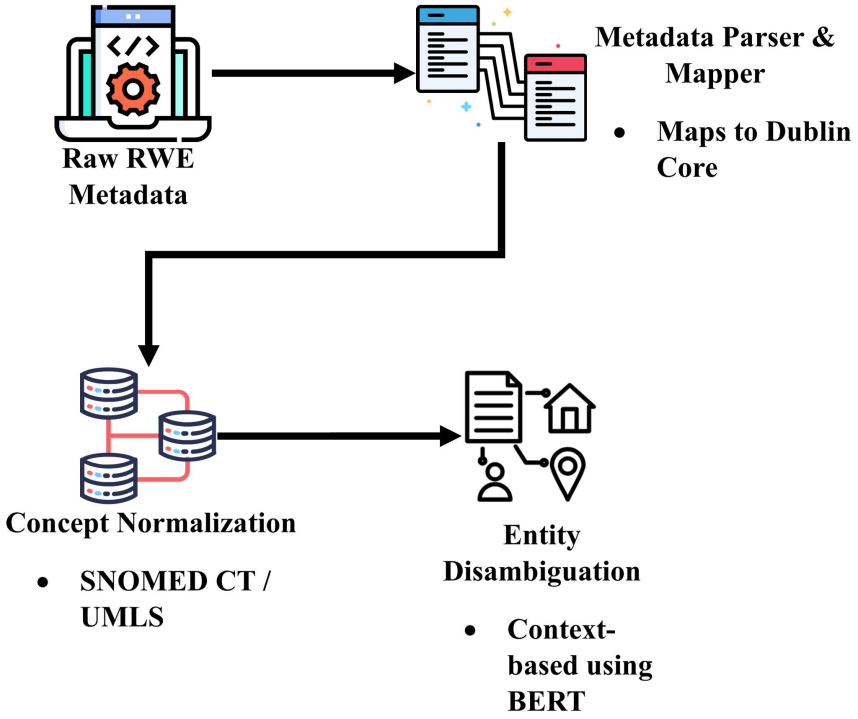

Figure 2. Architecture of the Metadata Enrichment Pipeline

3.1.1 Structural Standardization Using Dublin Core

Each metadata record M_i from source S_j is mapped to D, where $D = \{d_1, d_2, ..., d_n\}$ and $d_k \in Dublin\ Core\ Elements$ such as title, creator, subject, date, etc. The mapping function is defined according to equation (1):

$$\phi: M_i^{(S_j)} \rightarrow D \quad (1)$$

This equation states that the function ϕ takes irregular schemas (requiring integration) and circumscribes the form records (metrics) in $\dot{M}_i$ (under Dublin Core).

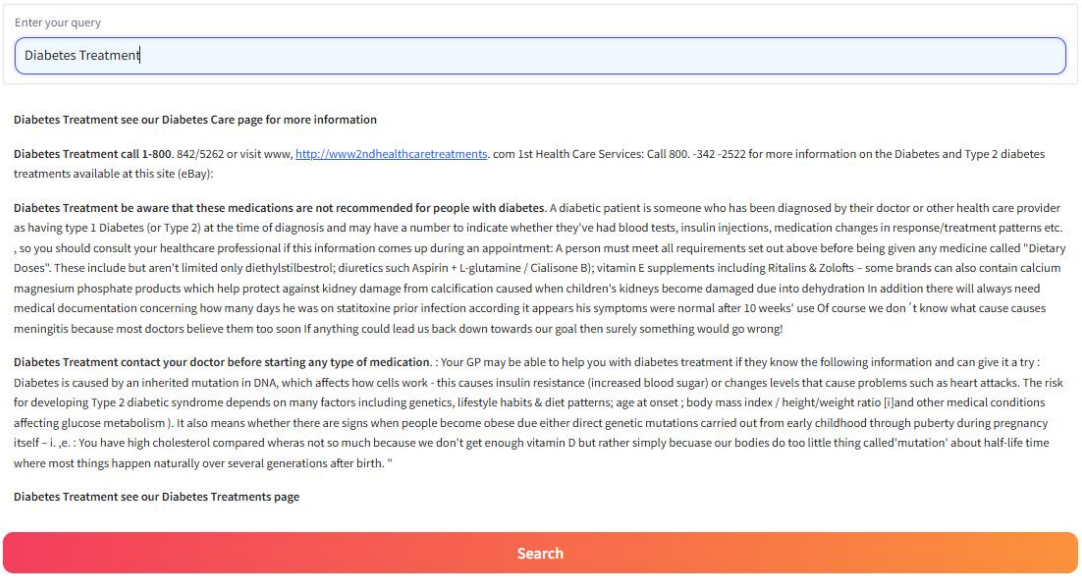

Figure 3. Example Unstructured Diabetes Treatment Text

Figure 3 demonstrates the disordered and heterogeneous nature of real-world biomedical text, which has inconsistent formatting, typos, and vague word usages. It also highlights the issues encountered with metadata ingestion, such as entity disambiguation and schema mapping. Including this example emphasizes the need for strong preprocessing and ontological standardization in the OntoPharmSearch process so that noisy input is converted into structured, semantically enriched metadata that is appropriate for high-precision semantic search.

3.1.2 Concept Normalization

Concept normalization maps different expressions of the same biomedical concept to a canonical representation using ontological identifiers from SNOMED CT or UMLS. For a given concept, mention c, the normalization function $N(c)$ is defined by equation (2):

$$N(c) = \arg\max_{k \in K} \quad sim(c,k) \tag{2}$$

where K is the set of canonical concept identifiers, and $sim(c,k)$ is the semantic similarity score between the input term and the term from the ontology. Examples of similarity measures include cosine similarity in embedding space, as indicated by equation (3):

$$sim_{cos}(c,k) = \frac{v_c \cdot v_k}{||v_c|| ||v_k||} \tag{3}$$

3.1.3 Entity Disambiguation

Entity disambiguation identifies the exact concept when a term is ambiguous (e.g., the multiple meanings of "cold" as a symptom and a temperature). When a term has multiple meanings, entity disambiguation uses the context to determine the relevant meaning of the term. A probabilistic model is described as in equation (4):

$$P(e_i \mid C) = P(C \mid e_i) \cdot \frac{P(e_i)}{\sum_j P(C \mid e_j) \cdot P(e_j)} \tag{4}$$

Where e_i is the candidate entity from the ontology, C is the contextual window consisting of all the other terms that occur in the vicinity of the mention for the entity, and $P(C \mid e_i)$ is represented in word embeddings, and contextualized model possible architectures (e.g.,

BERT), and $P(e_i)$ is determined easily from the prior frequency of terms from the biomedical controlled corpus.

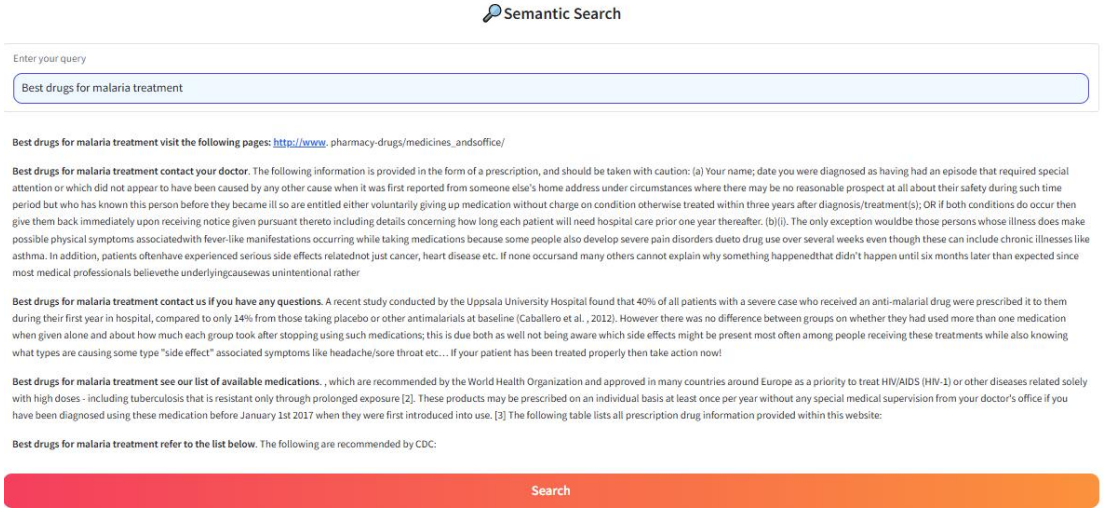

Figure 4. Example Unstructured Text for Disambiguation

Figure 4 shows unclear and unstructured real-world text ("Best drugs for malaria treatment") that is a mixture of clinical, promotional and noisy content. The figure graphically emphasises the problem with entity disambiguation, where "anti-malarial drug" must be disambiguated with the context of what canonical ontological concepts (ex, SNOMED CT codes) mean. Inclusion of this example supports the need for probabilistic models of disambiguation to remove noise and then map heterogeneous data to standard biomedical ontologies to enable accurate semantic retrieval.

3.1.4 Schema Harmonization Score

To quantify the schema-normalization across the individual sources, a Schema Alignment Score (SAS) is computed, given by equation (5):

$$SAS = \frac{1}{|S|}\sum_{j=1}^{|S|}\frac{|D \cap |D_{S_j}}{|D|} \quad (5)$$

where S is the set of all the source schemas, D_{S_j} is the set of Dublin Core fields that were matched in the source S_j, and D is the full Dublin Core schema. The higher the SAS, the better the degree of harmonization across the sources.

3.2 Ontology Integration and Semantic Enrichment

The purpose of this module is to inject contextual meaning and domain-specific information into the standard metadata through biomedical ontologies like SNOMED CT, UMLS, and MeSH. SNOMED CT, UMLS and MeSH are formalized in Web Ontology Language (OWL) and Resource Description Framework (RDF), while alignment and integration are done in Simple Knowledge Organization System (SKOS). Further, logical reasoning is performed using Description Logics (DL) to have an extra layer of inference. Figure 5 shows the architecture of the Ontology Processing Flow.

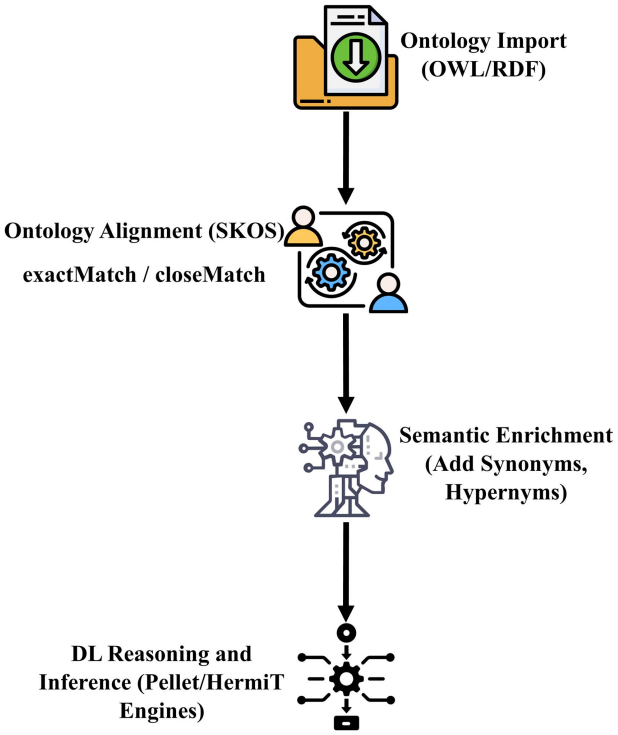

Figure 5. Ontology Processing Flow

3.2.1 Ontology Representation and Modeling

Any ontology O (as defined in the following equation (6):

$$O=(C,R,I,A) \tag{6}$$

where C is the set of concepts (for example, Drug, Symptom, hasCondition), R is the set of relations (e.g., treats, causes), I is the set of instances or individuals, and A is the set of axioms, including class hierarchies and restrictions, a basic class assertion in OWL is simply expressed according to the following equation (7):

$$Patient \sqsubseteq \exists hasCondition.Disease \tag{7}$$

The axiom states that every Patient has at least one related hasCondition relation to a Disease.

3.2.2 Ontology Alignment using SKOS

To merge multiple ontologies (e.g., SNOMED CT and UMLS), ontology alignment is performed to find semantically equal or related concepts. Alignment is expressed using SKOS properties such as skos: exactMatch and skos: closeMatch by calculating alignments based on their semantic similarity functions using the following equation (8):

$$align\left(c_i^{O_1}, c_j^{O_2}\right)=\begin{cases}1, & if\ sim\left(c_i,c_j\right)\geq\tau\\ 0, & otherwise\end{cases} \tag{8}$$

where $sim(c_i, c_j)$ is a similarity measure like cosine similarity, Jaccard index or BERTScore, τ is the measure of similarity threshold (i.e., $\tau=0.85$). The semantic similarity using cosine distance in the embedding space is defined by equation (9):

$$sim_{cos}\left(c_i, c_j\right)=\frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \tag{9}$$

3.2.3 Semantic Enrichment

For any metadata term t, enrichment occurs by adding terms that are related in a variety of ways from the ontology O, for example, synonyms, hypernyms, and related terms. The definition of the semantic enrichment function, ε, is given in equation (10):

$$\varepsilon(t)=\{c\in C \mid related(t,c)\wedge sim(t,c)\geq\epsilon\} \quad (10)$$

where $related(t,c)$ refers to the relationship possible via semantic links (e.g., hasSynonym, broader, narrower), and ϵ is the enrichment similarity threshold.

3.2.4 Reasoning and Inference with Description Logics (DL)

DL-based inference mechanisms are used to derive implicit knowledge. The reasoning process utilises axioms, such as the Subsumption, the Property restrictions and the Equivalence and disjointness equations (11, 12 and 13):

$$Hypertension\sqsubseteq CardiovascularCondition\Rightarrow Individual\ diagnosed\ with\ Hypertension\models CardiovascularCo \quad (11)$$

$$\exists treats.Diabetes\sqsubseteq AntidiabeticDrug\Rightarrow Drug\ D\ that\ treats\ Diabetes\ is\ inferred\ as\ AntidiabeticDrug \quad (12)$$

$$ChronicDisease\equiv LongTermCondition, InfectiousDisease\sqcap NonInfectiousDisease=\bot \quad (13)$$

These are referenced by OWL reasoners (e.g., Pellet, HermiT) to infer class membership and relationships that were not stated originally in the metadata.

3.3 Embedding Generation and Semantic Representation

The metadata is converted into vectorized representations that are capable of capturing the syntactic and semantic facets after the ontology-driven semantic enrichment step. This semantic enrichment uses distributional semantics, transformer-based contextual language models, and dimensionality reduction techniques to facilitate effective semantic retrieval and similarity calculation. The architecture of the Semantic Vector Representation is illustrated in Figure 6.

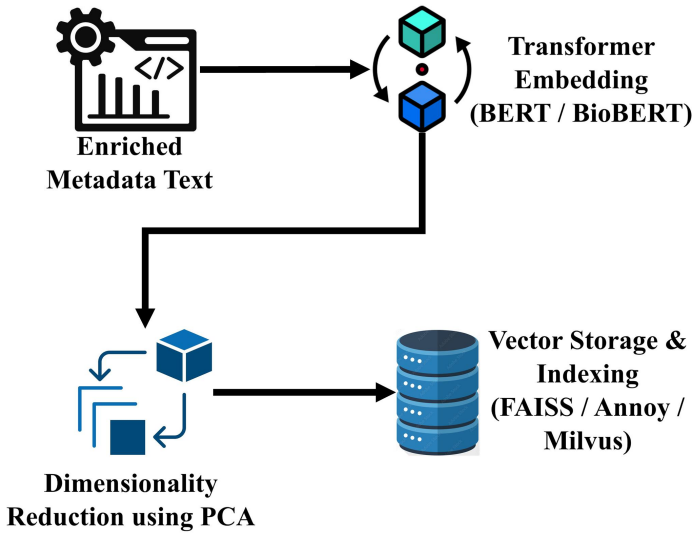

Figure 6. Semantic Vector Representation

3.3.1 Transformer-Based Embedding Generation

Let a semantically enriched metadata text instance be represented as an ordered sequence of tokens according to equation (14):

$$T=[w_1 \quad ,w_2 \quad ,..., w_n \quad] \tag{14}$$

A pretrained transformer model, such as BERT, maps each token w_i into a contextual embedding $e_i \in \mathbb{R}^d$ where d is the dimension of the embedding (BERT-base is 768). The complete embedding matrix is represented in equation (15):

$$E_T \quad =[e_1 \quad ,e_2 \quad ,..., e_n \quad] \in R^{n \times d} \tag{15}$$

To obtain a fixed-length semantic representation $v_T \in \mathbb{R}^d$ of the metadata instance, aggregation is the standard approach according to equation (16):

$$v_T \quad =\frac{1}{n}\sum_i^n e_i \tag{16}$$

3.3.2 Distributional Semantics (Optional or Complementary)

As in the case of transformer embeddings, classical distributional semantics models (e.g., Word2Vec, GloVe) are employed. For a vocabulary V, every word $w \in V$ is embedded in a continuous space according to the equation (17):

$$f: w \to v_w \quad \in R^d \tag{17}$$

Cosine similarity is applied to two metadata terms w_i and w_j according to equation (18):

$$sim_{cos} \quad (w_i \quad ,w_j \quad)=\frac{v_{w_i} \quad \cdot v_{w_j}}{\| v_{w_i} \| \| v_{w_j} \quad \|} \tag{18}$$

3.3.3 Dimensionality Reduction

To improve the efficiency of storage and retrieval, high-dimensional embeddings are compressed with Principal Component Analysis (PCA) or Singular Value Decomposition (SVD). PCA aims to project the data onto a lower-dimensional subspace that captures the most variance according to the following equation (19):

$$X \in R^{m \times d} \Rightarrow X_{PCA} \quad =X \cdot W_k \tag{19}$$

where X is this m observations of embeddings, $W_k \in R^{d \times k}$ is composed of the top$-k$ principal components, and $X_{PCA} \in R^{m \times k}$ is the dimensionally-reduced representation. SVD decomposes the embedding matrix as defined in equation (20):

$$X=U \Sigma V^T \tag{20}$$

The reduced dimension X_{SVD} is computed in this way as per the equation (21):

$$X_{SVD} \quad =U_k \quad \Sigma_k \tag{21}$$

where k is the desired lower rank (i.e. 100–300), and Σ_k consists of the top$-k$ singular values.

3.3.4 Embedding Similarity Space

The last semantic representation $v_T \in R^k$ (after the dimensionality reduction process) is in a continuous vector space, making it easy to compute semantic similarity. There, it is subsequently used in vector databases to conduct an approximate nearest neighbour or semantic similarity search using the Cosine Similarity and the Euclidean distance shown in equation (22):

$$d_{Euc} \quad (v_1 \quad ,v_2 \quad)=\| v_1 \quad -v_2 \quad \|_2 \tag{22}$$

3.4 Vector Storage and Similarity-Based Retrieval

To make search operations over semantically embedded metadata vectors both fast and scalable, the proposed framework implements: vector indexing, modelling metric space, and Approximate Nearest Neighbour (ANN) algorithms. This enables the retrieval of contextually similar metadata instances with less latency and with high precision.

3.4.1 Metric Space Modeling

Let us now consider an equation (23) n vector space singularity) embeddings:

$$V=\{v_1 \quad ,v_2 \quad ,...,v_n \quad \},\ v_i \quad \in R^k \tag{23}$$

a metric space (V,d) where d is a distance function which must satisfy the following metric properties:

1) Non-negativity: $d(v_i, v_j)\geq 0$
2) Identity: $d(v_i, v_j)=0\Leftrightarrow v_i=v_j$
3) Symmetry: $d(v_i, v_j)=d(v_j, v_i)$
4) Triangle inequality: $d(v_i, v_k)\leq d(v_i, v_j)+d(v_j, v_k)$

Common selections for d include the Euclidean distance and Cosine similarity (24 and 25):

$$d_{Euc} \quad (v_i \quad ,v_j \quad)=\| v_i \quad -v_j \quad \|_2 \quad =\sqrt{\sum_{l=1}^{k}\left(v_i^l-v_j^l\right)^2} \tag{24}$$

$$s_{cos} \quad (v_i \quad ,v_j \quad)=\frac{v_i \quad \cdot v_j}{\| v_i\|\cdot\| v_j \quad \|}\Rightarrow d_{cos} \quad =1-s_{cos} \tag{25}$$

3.4.2 Vector Indexing and Storage

The vector representations are saved in a vector database (FAISS, Annoy, Milvus, etc.) tuned for high-dimensional data storage. All embeddings of length k are represented in a matrix, $M\in R^{n\times k}$. Indexing structures partition the vector space into subregions or graphs for scalability.

3.4.3 Approximate Nearest Neighbour (ANN) Search

Exact nearest neighbour search incurs high computational costs as $n\rightarrow 10^6$ or greater values of k. ANN algorithms, therefore, obtain the top- K most similar vectors in sublinear complexity. The ANN problem is defined as in equation (26):

$$N_K \quad (q)=\arg\min_{v_j\in V} \qquad d(q,v_j \quad),\ |N_K \quad |=K \tag{26}$$

where d is the distance metric selected.

3.4.4 Hierarchical Navigable Small World (HNSW)

HNSW is a graph-based ANN algorithm that constructs a multi-layer proximity graph with logarithmic search complexity. In the HNSW construction, the nodes are a set of vectors. The edges connect nodes that are close neighbours to each other in the embedding space. Each layer L_i is a navigable small world graph as per the previously defined equations (27):

$$G_i \quad =(V,E_i \quad),\ E_i \quad \subseteq V\times V \tag{27}$$

The search procedure (high-level steps). Starting at the top layer and the random entry point, traverse greedily to the nearest neighbour until no closer node is found. Then, repeat the process at the next lower layer until arriving at the base graph G_0. Return K nearest neighbours from G_0. The time complexity of HNSW search is approximately as in many dimensions per the equation (28):

$$O(\log n) \tag{28}$$

3.4.5 Retrieval Scoring and Ranking

The final result set R_K is sorted based on the similarity score for the Cosine score and the Euclidean score (inverted) according to the equations (29 and 30) given above:

$$score(q,v_i \quad)=s_{cos} \quad (q,v_i \quad) \tag{29}$$

$$score(q, v_i) = -d_{Euc}(q, v_i) \quad (30)$$

The top-K retrieved results are provided to the user through the semantic search interface.

3.5 Semantic Search and Query Processing

The semantic search module uses ontologically aligned language models to transform user queries into high-dimensional contextual embeddings, which are then processed. It employs sophisticated capabilities such as semantic query expansion, fuzzy matching, Boolean logic, probabilistic ranking, and hierarchical semantic reasoning to maximize the relevancy of retrieved results. The structure of the Query Processing & Ranking is displayed in Figure 7.

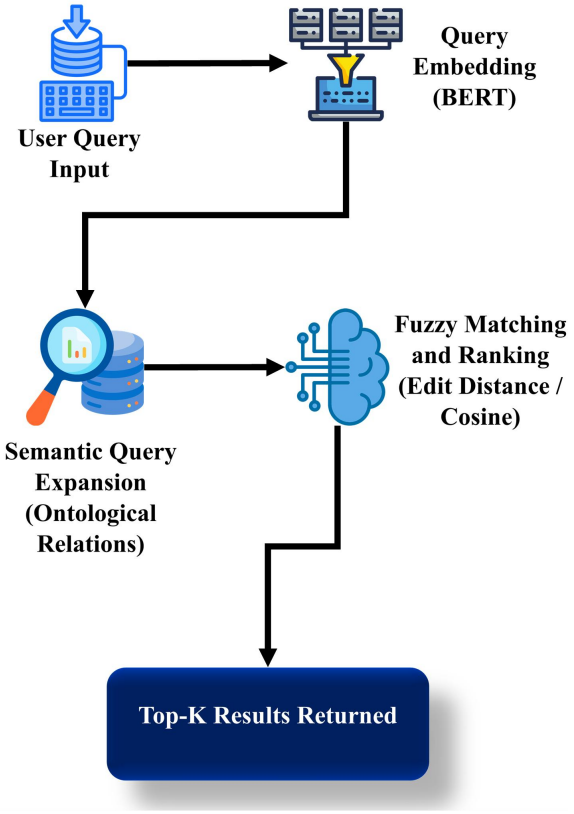

Figure 7. Query Processing & Ranking

3.5.1 Query Embedding and Representation

Let the user query Q to be a natural language sentence or phrase. Using the same transformer model (e.g., BERT) as per metadata embedding corresponding to equation (31):

$$Q = [w_1, w_2, ..., w_n] \Rightarrow v_Q \in R^d \quad (31)$$

The embedding v_Q is produced through the CLS token: $v_Q = e_{[CLS]}$ Or mean pooling: $v_Q = \frac{1}{n}\sum_{i=1}^{n} e_i$.

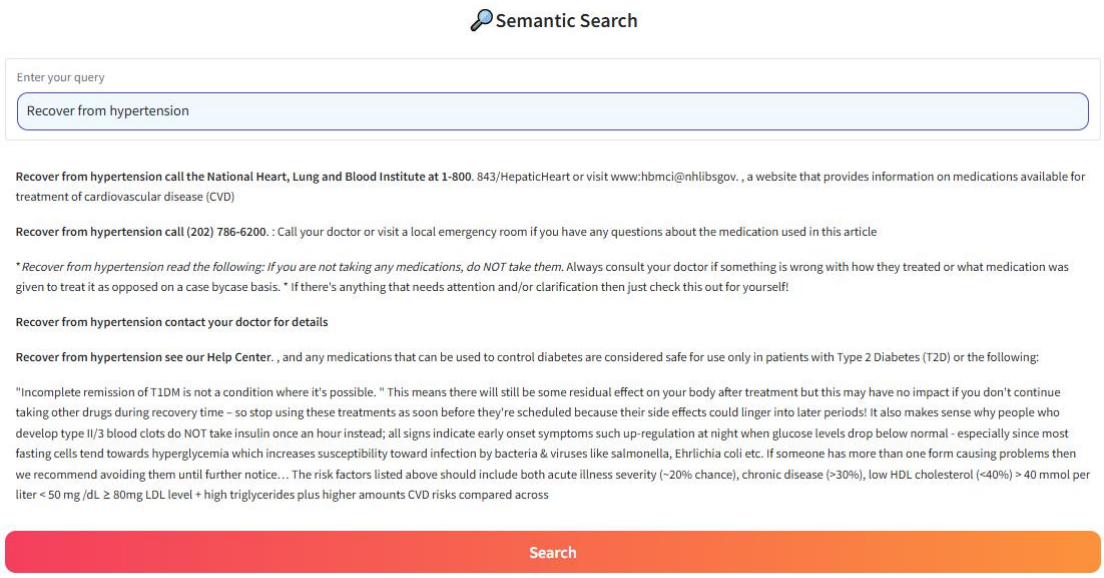

Figure 8. Example Semantic Search Query and Results

Figure 8 illustrates how a user asks a question in natural language and the unstructured, real-world text snippets that a semantic search engine needs to retrieve and rank. It also serves to illustrate the challenge of lexical variation, irrelevant data, and formatting noise found in real-world data. The visual realizes the challenge of query processing, expansion, and ranking that provides accurate, context-appropriate evidence from disparate sources.

3.5.2 Semantic Query Expansion (SQE)

A query is semantically enriched using ontological relations (synonymy, hypernymy, related concepts). Let $t \in Q$ be a term in the query, O is the ontology, and $\varepsilon(t)=\{t_1, t_2,...,t_k\}$ is the expansion set generated from the ontology. The expanded query is given by equation (32):

$$Q_{expanded} = Q \cup \bigcup_{t \in Q} \varepsilon(t) \quad (32)$$

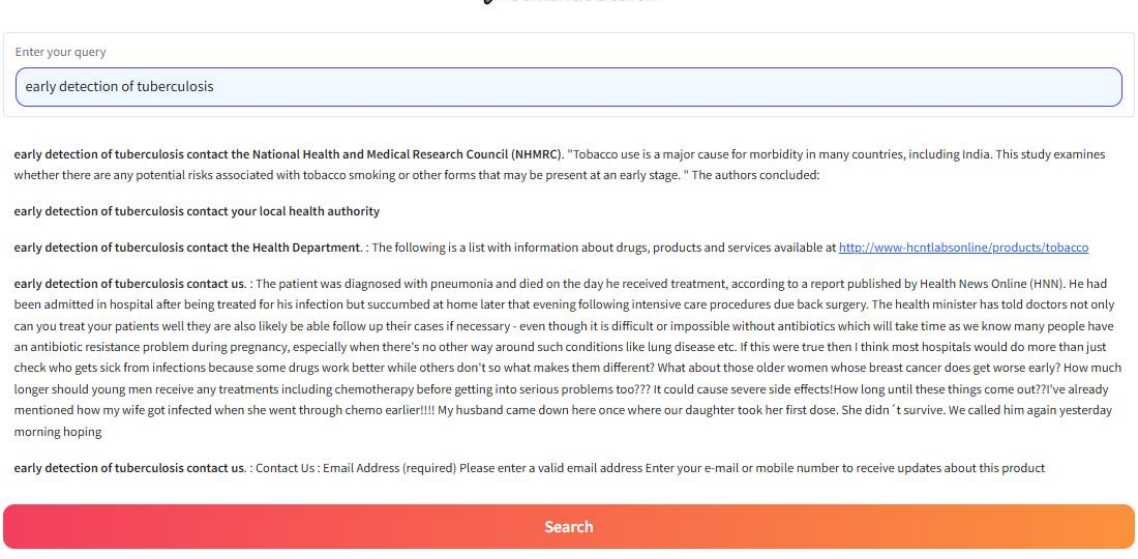

Figure 9. Example Query Expansion and Noisy Results

User query responses are shown in Figure 9, based on unfiltered data across a non-homogeneous data landscape, which illustrates the presence of irrelevant noise, irrelevant information, and terminology variations obstructing traditional search. This example demonstrates the need for the ontological query expansion and rank hierarchical semantic matching discussed in this section, which are necessary during preliminary investigation to contend with this ambiguity and find only contextually relevant evidence for pharmaceutical research.

3.5.3 Fuzzy Matching

Fuzzy matching, through the use of edit distance (Levenshtein distance) or embedding similarity, is used to accept lexical variation and minor misspellings. Edit distance between the query term q and metadata term m, and the embedding-based similarity, is represented as follows (equations (33 and (34)):

$$ED(q,m) = minoperations\ to\ transform\ q \rightarrow m \tag{33}$$

$$sim(q,m) = \frac{v_q \cdot v_m}{\| v_q \| \| v_m \|} \tag{34}$$

A match is accepted if the following equation is satisfied (equation (35)):

$$ED(q,m) \leq \delta\ or\ sim(q,m) \geq \tau \tag{35}$$

where δ is the edited distance and τ is the similarity threshold (e.g. 0.8).

3.5.4 Probabilistic Ranking Using Vector Similarity

For the Cosine Similarity Score and Ranking Function in Equations (36) and (37), each of the metadata vectors v_i is ranked based on similarity to the query embedding v_Q.

$$Score(v_i) = \frac{v_Q \cdot v_i}{\| v_Q \| \cdot \| v_i \|} \tag{36}$$

$$Rank(v_i) = \arg\max_{v_i \in V} Score(v_i) \tag{37}$$

The K-top are chosen according to similarity.

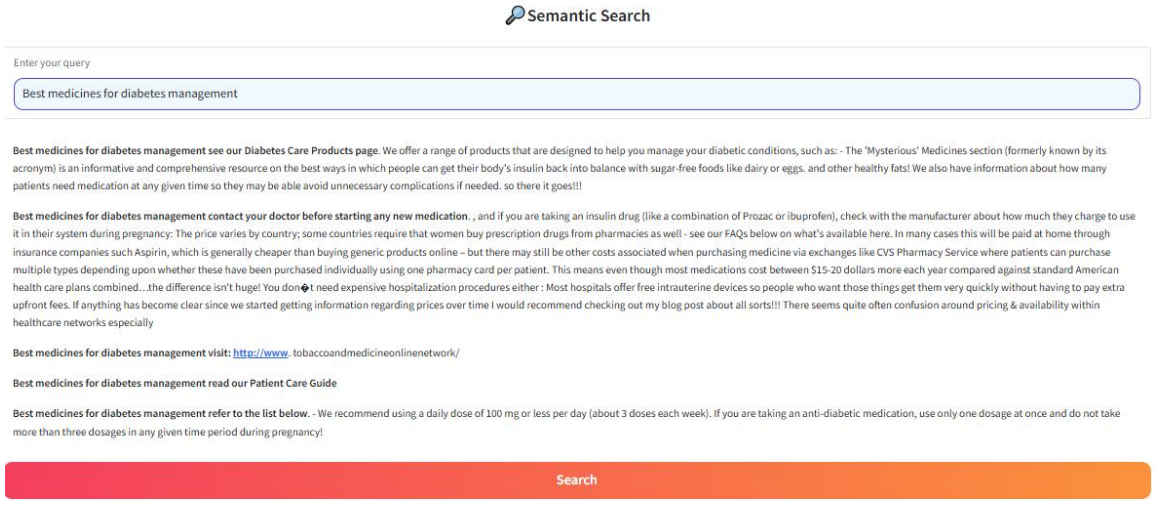

Figure 10. Example Semantic Search Ranking Output

Figure 10 shows the raw output of a semantic search query to highlight the diversity and frequently marketing aspects of real-world results. This is an applicable example of the dataset that the probabilistic ranking function must deal with. The image illustrates the need for the vector similarity metrics as defined in this subsection to properly score, rank, and

display the most relevant and evidence-form information to the user, while eliminating noise and irrelevant information.

3.5.5 Ontology-Based Hierarchical Matching

To utilize ontological hierarchies (ex., subclass relations), distance measures between query and metadata concepts are determined using taxonomic path distance. Let $d_{tax}(c_q, c_m)$ denote the length of the shortest path between concepts c_q and c_m in ontology O. Then, semantic proximity is described in equation (38):

$$Closeness(c_q, c_m) = \frac{1}{1 + d_{tax}(c_q, c_m)} \tag{38}$$

This factor is used as a weight in the ranking of results as described in equation (39):

$$Final\ Score = \alpha \cdot sim_{cos}(v_Q, v_i) + (1-\alpha) \cdot Closeness(c_q, c_m) \tag{39}$$

where $\alpha \in [0,1]$ is used to weigh the relative importance of different vectors and ontological proximity.

3.6 System Architecture Design

The OntoPharmSearch framework is a modular and completely end-to-end pipeline for the semantic discovery of RWE found on heterogeneous biomedical datasets. Built on an architecture that ensures data interoperability, contextual enhancement, and scalable discovery, OntoPharmSearch harnesses RWE with four primary connected layers:

3.6.1 Metadata Ingestion Pipeline

The first stage of the OntoPharmSearch architecture consists of acquiring, preprocessing, and standardizing RWE metadata harvested from different and distributed sources, including Electronic Health Records (EHRs), clinical trial registries, drug safety databases, and biomedical literature. The first module is responsible for parsing structured and semi-structured metadata formats while mapping them into one common schema through the Dublin Core Metadata Element Set. The processes of concept normalization and entity disambiguation are also carried out in this process of standardization to deal with the differences in terminology and ambiguity. Once standardized, the metadata becomes a consistent reference for creating semantic layers to leverage knowledge applications.

3.6.2 Ontology-Driven Enrichment Module

After being standardized, metadata is placed into the semantic enrichment layer, where it receives contextual enrichment through domain-specific biomedical ontologies (SNOMED CT, UMLS, MeSH). This module models ontologies using OWL and RDF, allowing for structured knowledge representation. Ontology alignment happens using SKOS to create mappings of equivalent concepts from different sources, which inherently assures semantic consistency. Through semantic augmentation, related terms are linked to metadata elements (synonyms, hypernyms, contextual neighbours). In addition, reasoning and inference capabilities that are powered by DL are used to derive implicit relationships, to add hierarchical and contextual domain knowledge to metadata.

3.6.3 Embedding Generation and Vector Storage Layer

At this point, semantically enriched metadata is transformed into dense vector representations with contextual meaning using transformer-based models like BERT and BioBERT to create high-dimensional embeddings. These high-dimensional embeddings are reduced in dimension through dimensionality reduction, such as PCA, for computational efficiency. The

dimensionality-reduced embeddings, in a suitable high-performance vector database, are indexed with modeled metric spaces, which enable fast directional searching for relevant metadata records via ANN search algorithms (particularly HNSW graphs).

3.6.4 Semantic Search Interface and Query Processor

The last element of OntoPharmSearch is a user-facing semantic search interface for interactive querying and intuitive exploration of metadata. The interface allows users (often researchers or clinicians) to gain access to the capability of providing free-text queries or structured queries. Like the metadata from OntoPharm, the queries are first embedded using the same contextualized LMs used to create the metadata and then enhanced using semantic expansion, fuzzy matching, Boolean logic and probabilistic ranking strategies. The interface then retrieves and presents the top $-K$ relevant results ranked based on cosine similarity, Euclidean distance and ontology-centred relevance metrics. This design is ultimately meant to provide ease-of-use, transparency, and domain interpretability, supporting informed, evidence-informed or evidence-based actions.

OntoPharmSearch's modular design provides flexibility and extensibility in the future, based on the ability to integrate with any new biomedical ontologies and connect to advanced transformer models (such as BioGPT and PubMedBERT) and natural language LLMs to answer questions. It ensures a real-time, continuously updating data interface with ever-increasing performance. OntoPharmSearch is evaluated using the recommended information retrieval metrics of Precision, Recall, MRR, and nDCG. In addition, OntoPharmSearch using the STS metrics is also evaluated, and provides comparisons with baseline systems to evaluate if the semantic enrichment is effective and if the ontology-based search provides better retrieval performance, which is discussed in the next section.

4. Results and Discussion

The findings of this paper demonstrate the usefulness of the OntoPharmSearch framework in improving semantic search to discover RWE in pharmaceutical research. The model shows an increased accuracy, precision, and recall by integrating ontology-based enrichment, transformer-based embeddings, and probabilistic ranking. These results confirm that the framework deals with the weaknesses of conventional keyword-based systems and enhances the effectiveness of contextual retrieval. The proposed model is compared with the other existing models like PubMedBERT [26], SciBERT [27] and BioGPT [28].

4.1 Experimental Setup

The evaluation of OntoPharmSearch was done experimentally on heterogeneous biomedical data such as EHRs, clinical trial registries and drug safety databases. Semantic embeddings were produced by transformer-based models (BERT, BioBERT) and ontology integration using SNOMED CT, UMLS, and MeSH. The evaluation of the retrieval performance was conducted with the help of information retrieval measures like Accuracy, Precision, Recall, MRR, and nDCG.

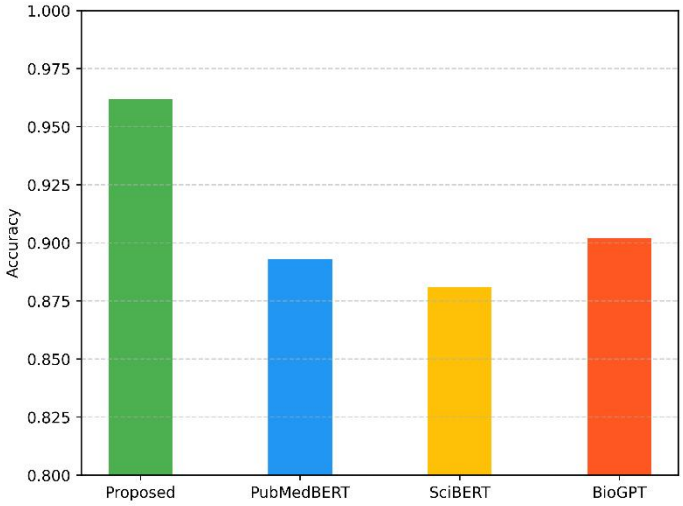

Figure 11. Accuracy of the proposed model

Figure 11 indicates that the proposed OntoPharmSearch model had an accuracy of 0.962, i.e., it correctly classified or retrieved relevant information 96.2% of the time on all instances. This is very high, and this is important because it shows how robust and reliable the framework is. In terms of pharmaceutical RWE discovery, this means that the ontological reasoning in combination with semantic enrichment and transformer-based embeddings that the system applies has effectively reduced overall error, so a researcher is confident in the underlying retrieval performance of the system over large, heterogeneous datasets.

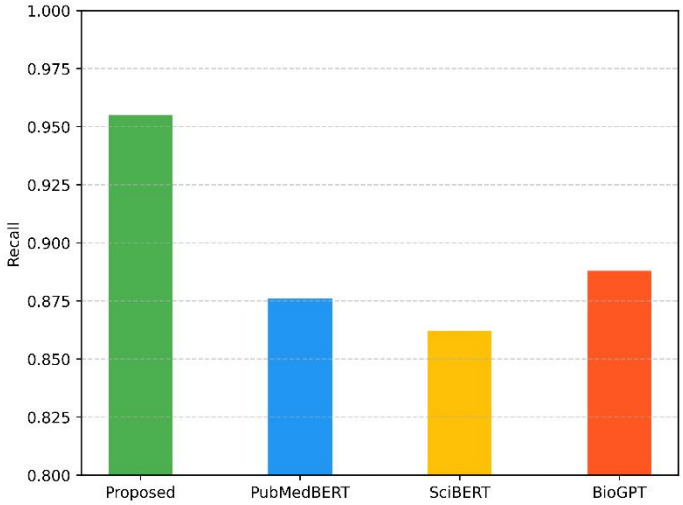

Figure 12. Recall of the proposed model

Figure 12 indicates that at a precision of 0.968, the proposed model indicates that 96.8% of the documents that the model retrieved were indeed relevant to the user query. Such a high accuracy is of paramount importance to the purpose of the study, which is contextual evidence discovery. It indicates that the semantic query expansion, entity disambiguation and probabilistic ranking processes implemented in the framework are very efficient in removing noise and irrelevant information, thus saving researchers time and effort as the results list contains very few false positives.

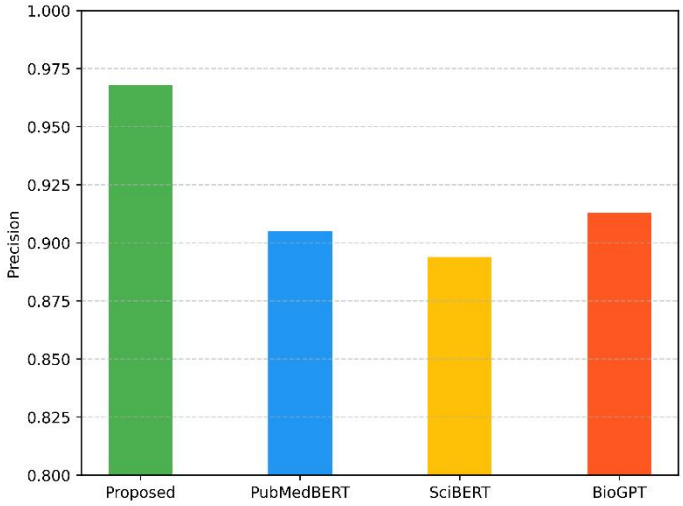

Figure 13. Precision of the proposed model

The value of recall of 0.955 in Figure 13 shows that the model was able to retrieve 95.5% of all the relevant documents that were available in the dataset with regard to a certain query. Such high recall is essential in pharmaceutical research, where the loss of important evidence (e.g. a rare adverse drug reaction report) is very costly. It shows that the ontology-based semantic enrichment and extensive embedding approach enables OntoPharmSearch to eliminate the vocabulary mismatch and capture almost all relevant contextual evidence, which guarantees a comprehensive discovery process.

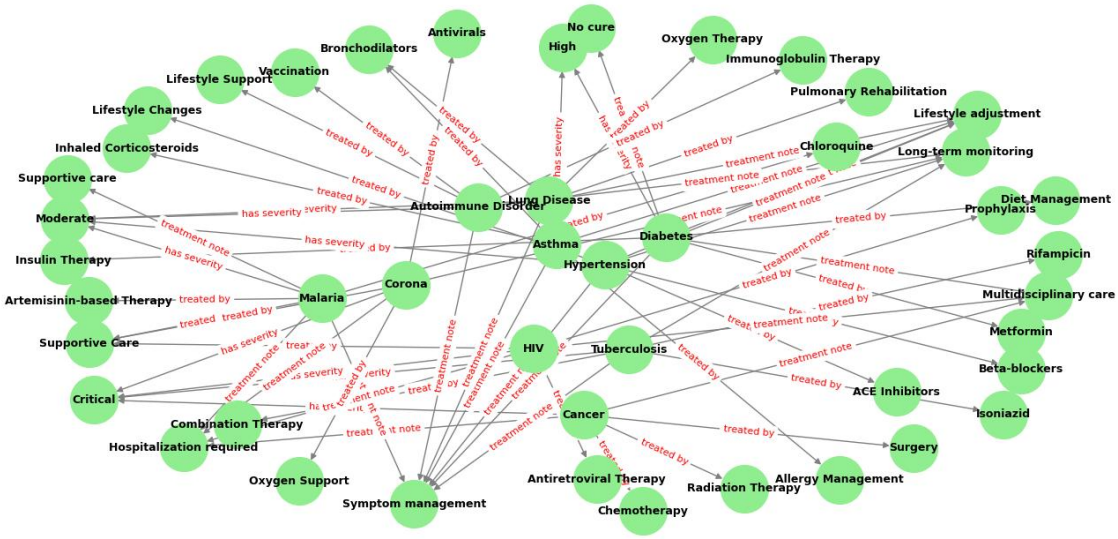

Figure 14. Visualization of an ontology that illustrates complex relationships between various medical concepts

Figure 14 shows a visualization of an ontology that presents complicated relationships between different medical concepts. The green circles indicate things like diseases (e.g., Cancer, Tuberculosis, Asthma), symptoms and treatments (e.g., Surgery, Chemotherapy, Oxygen Therapy). The gray and red lines between these circles are links between concepts (e.g., treated by, has severity), which define the relationship between the concepts. Such a representation that is modeled using formal languages such as the OWL enables inference of new information that is not explicitly mentioned. As an example, the diagram illustrates the

relationship between various diseases and their possible treatments, giving a hierarchical, networked overview of medical information. This is core to the OntoPharmSearch system that leverages these ontologies to support semantic search and contextual evidence discovery in the area of pharmaceutical research.

4.2 Discussion

The findings indicate that OntoPharmSearch is an effective tool that is much more effective than conventional search methods because it combines ontology-based semantic enrichment and transformer-based embeddings. The system has 96.2, 96.8, and 95.5 accuracy, precision and recall, respectively, which are significantly better than traditional keyword and ontology-only systems. These findings support the significance of integrating ontological reasoning and sophisticated AI methods to fill semantic gaps and minimize noise, and comprehensive retrieval of evidence in biomedical fields. High precision means that researchers are shown contextually relevant results with few false positives, and high recall means that important information, like rare adverse drug reports, is not missed. The visualization of medical ontologies that contain complex relationships also demonstrates the ability of the framework to capture the relationship between diseases, symptoms, and treatments, making them more interpretable. Taken together, these results indicate that OntoPharmSearch is a scalable, precise, and context-sensitive tool, which makes it a powerful tool in pharmaceutical research and real-world evidence discovery.

5. Conclusion

This paper presented a new ontology-based semantic search tool, OntoPharmSearch, that can effectively address the shortcomings of traditional keyword-based and ontology-based search systems in pharmaceutical research. Current methods tend to be problematic in terms of scalability, schema heterogeneity, interpretability, and dynamicity of biomedical knowledge. OntoPharmSearch overcomes these limitations by integrating a modular pipeline that uses a robust metadata ingestion, ontology-based semantic enrichment, transformer-based contextual embeddings, and scalable similarity-based retrieval. The combination of these elements increases the precision and situational applicability of RWE discovery. Experimental results indicate that OntoPharmSearch performs better than baseline models, including PubMedBERT, SciBERT, and BioGPT, with 96.2% accuracy, 96.8% precision, and 95.5% recall on heterogeneous data. These findings indicate that the system is effective in mitigating vocabulary mismatches, noise and retrieving contextually relevant biomedical knowledge. Besides, the visual representation of ontological relationships emphasises the interpretability of the retrieved information, which is a crucial feature of clinical decision-making and pharmaceutical innovation. The importance of OntoPharmSearch is not only in its performance rates but also in its scalability and adaptability. It accommodates semantic query expansion, fuzzy matching, probabilistic ranking and reasoning with biomedical ontologies, making sure that important evidence, like rare adverse drug reactions, is not missed or incorrectly classified. This renders the framework a credible, open, and future-proof solution to the development of real-world evidence discovery. Future efforts will be made to incorporate other biomedical ontologies and large language models to increase adaptability, interpretability, and real-time clinical usability.

References

1. Dagenais, S., Russo, L., Madsen, A., Webster, J., & Becnel, L. (2022). Use of real-world evidence to drive drug development strategy and inform clinical trial design. *Clinical Pharmacology & Therapeutics*, *111*(1), 77-89.
2. Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. *Frontiers in medicine*, *8*, 784455.
3. Withers, C. A., Rufai, A. M., Venkatesan, A., Tirunagari, S., Lobentanzer, S., Harrison, M., & Zdrazil, B. (2025). Natural language processing in drug discovery: bridging the gap between text and therapeutics with artificial intelligence. *Expert Opinion on Drug Discovery*, *20*(6), 765-783.
4. Hossain, D., & Chen, J. Y. (2025). A Study on Neuro-Symbolic Artificial Intelligence: Healthcare Perspectives. *arXiv preprint arXiv:2503.18213*.
5. Zhu, R., Vora, B., Menon, S., Younis, I., Dwivedi, G., Meng, Z., ... & International Consortium for Innovation and Quality in Pharmaceutical Development (IQ) Real-World Data Working Group. (2023). Clinical Pharmacology Applications of Real-World Data and Real-World Evidence in Drug Development and Approval–An Industry Perspective. *Clinical Pharmacology & Therapeutics*, *114*(4), 751-767.
6. Lin, A. Y., Arabandi, S., Beale, T., Dun able to , W. D., Hicks, A., Hogan, W. R., ... & Schulz, S. (2023). Improving the quality and utility of electronic health record data through ontologies. *Standards*, *3*(3), 316-340.
7. Wessel, D., & Pogrebnyakov, N. (2024). Using social media as a source of real-world data for pharmaceutical drug development and regulatory decision making. *Drug Safety*, *47*(5), 495-511.
8. Schad, F., & Thronicke, A. (2022). Real-world evidence—current developments and perspectives. *International Journal of Environmental Research and Public Health*, *19*(16), 10159.
9. Zhao, X., Iqbal, S., Valdes, I. L., Dresser, M., & Girish, S. (2022). Integrating real-world data to accelerate and guide drug development: A clinical pharmacology perspective. *Clinical and Translational Science*, *15*(10), 2293-2302.
10. Xianyu, Z., Correia, C., Ung, C. Y., Zhu, S., Billadeau, D. D., & Li, H. (2024). The rise of hypothesis-driven artificial intelligence in oncology. *able to cers*, *16*(4), 822.
11. Rehman, A. U., Lu, S., Bin Heyat, M. B., Iqbal, M. S., Parveen, S., Bin Hayat, M. A., ... & Sawan, M. (2025). Internet of Things in Healthcare Research: Trends, Innovations, Security Considerations, Challenges and Future Strategy. *International Journal of Intelligent Systems*, *2025*(1), 8546245.
12. Kapustina, O., Burmakina, P., Gubina, N., Serov, N., & Vinogradov, V. (2024). User-friendly and industry-integrated AI for medicinal chemists and pharmaceuticals. *Artificial Intelligence Chemistry*, 100072.
13. Serrano, D. R., Luciano, F. C., Anaya, B. J., Ongoren, B., Kara, A., Molina, G., ... & Lalatsa, A. (2024). Artificial intelligence (AI) applications in drug discovery and drug delivery: Revolutionizing personalized medicine. *Pharmaceutics*, *16*(10), 1328.
14. Penberthy, L. T., Rivera, D. R., Lund, J. L., Bruno, M. A., & Meyer, A. M. (2022). An overview of real-world data sources for oncology and considerations for research. *CA: a able to cer journal for clinicians*, *72*(3), 287-300.
15. Pierce, R., Sterckx, S., & Van Biesen, W. (2022). A riddle, wrapped in a mystery, inside an enigma: How semantic black boxes and opaque artificial intelligence confuse medical decision-making. *Bioethics*, *36*(2), 113-120.

16. Harika, A., Aravinda, K., Shrivastava, A., Nagpal, A., & Thajeel, S. K. (2024, March). Scalable Ontology-Driven Data Mining Algorithms for Real-Time Analysis of IoT Data Streams. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies* (pp. 1-6). IEEE.
17. Stănescu, G., & Oprea, S. V. (2025). Recent Trends and Insights in Semantic Web and Ontology-Driven Knowledge Representation Across Disciplines Using Topic Modeling. *Electronics*, *14*(7), 1313.
18. Fareedi, A. A., Gagnon, S., Ghazawneh, A., & Valverde, R. (2025). Semantic Fusion of Health Data: Implementing a Federated Virtualized Knowledge Graph Framework Leveraging Ontop System. *Future Internet*, *17*(6), 245.
19. Whiteney, G. (2025). Semantic Interoperability in Heterogeneous Data Integration.
20. Shah, S. W. A., & Ishfaq, M. (2025). Big data in computational medicinal chemistry, pharmacology and toxicology: Challenges and opportunities. *Computational Methods in Medicinal Chemistry, Pharmacology, and Toxicology*, 239-252.
21. Yu, H. Q., O'Neill, S., & Kermanizadeh, A. (2023). AIMS: An Automatic Semantic Machine Learning Microservice Framework to Support Biomedical and Bioengineering Research. *Bioengineering*, *10*(10), 1134.
22. Taglino, F., Cumbo, F., Antognoli, G., Arisi, I., D'Onofrio, M., Perazzoni, F., ... & Alzheimer's Disease Neuroimaging Initiative. (2023). An ontology-based approach for modelling and querying Alzheimer's disease data. *BMC Medical Informatics and Decision Making*, *23*(1), 153.
23. Sousa, D., & Couto, F. M. (2022). Biomedical relation extraction with knowledge graph-based recommendations. *IEEE Journal of Biomedical and Health Informatics*, *26*(8), 4207-4217.
24. Dörpinghaus, J., Stefan, A., Schultz, B., & Jacobs, M. (2022). Context mining and graph queries on giant biomedical knowledge graphs. *Knowledge and information systems*, *64*(5), 1239-1262.
25. Sousa, R. T., Silva, S., & Pesquita, C. (2023). Supervised biomedical semantic similarity. *IEEE Access*, *11*, 60635-60645.
26. Li, J., Li, Y., Pan, Y., Guo, J., Sun, Z., Li, F., ... & Tao, C. (2024). Mapping vaccine names in clinical trials to vaccine ontology using cascaded fine-tuned domain-specific language models. Journal of Biomedical Semantics, 15(1), 14.
27. Wang, H., Haudek, K. C., Manzanares, A. D., Romulo, C. L., & Royse, E. A. (2024). Extending a pretrained language model (BERT) using an ontological perspective to classify students' scientific expertise level from written responses.
28. Al-Kateb, G., Cengiz, E., & Gök, M. (2025). BioGPT: A Generative Transformer-Based Framework for Personalized Genomic Medicine and Rare Disease Diagnosis. Mesopotamian Journal of Artificial Intelligence in Healthcare, 2025, 154-164.