JNEST EISSN: 2961-8738 ● PISSN: 2961-8916 ● DOI: 10.56741/jnest.v4i03.1049
Article Received 06 May 2025, Revised 24 June 2025

Implementation of Knowledge-Based Graph Neural Networks for Reasoning and Ranking Medical Entities from CORD-19 Texts

Agus R. Fadillah*, Irwansyah Saputra

***Abstract*—The rapid growth of biomedical literature, especially during the COVID-19 pandemic, has introduced new challenges in retrieving clinically relevant information using conventional search methods. This study proposes a novel, interpretable framework for biomedical information retrieval that integrates Named Entity Recognition (NER), knowledge graph construction, and Graph Neural Networks (GNNs) to support semantic reasoning and entity-level ranking. Unlike prior biomedical retrieval systems that operate at document level or perform link prediction over KGs, our framework introduces a novel task formulation contextual entity-level ranking powered by graph-based semantic reasoning. Leveraging the CORD-19 dataset, the system filters abstracts based on user queries, extracts domain-specific entities using SciSpacy, and constructs a semantic graph that captures co-occurrence relationships among medical concepts. A Graph Convolutional Network (GCN) is then employed to prop- agate relevance signals across the graph, enabling context- aware entity ranking. Experimental evaluations using queries such as ”pneumonia” and ”cough” demonstrate superior performance over traditional IR baselines like TF-IDF and BM25, achieving a Mean Average Precision (MAP) of 0.95 and Precision@3 of 1.00. The results confirm the system’s effectiveness in identifying semantically meaningful biomed- ical entities while offering enhanced transparency through graph-based visualizations. This work contributes a scalable and extensible approach to biomedical search and lays the foundation for intelligent literature exploration in medical research and clinical decision support.**

***Index Terms*—Biomedical Information Retrieval, Graph Neural Networks, Knowledge Graph, Named Entity Recognition, CORD-19**

I. Introduction

THE COVID-19 pandemic spurred an unprecedented volume of biomedical literature, making it difficult for researchers to identify relevant information using traditional search methods. The COVID-19 Open Research Dataset (CORD-19) was explicitly created to support text mining and information retrieval efforts on COVID-19 literature, and it has been used as the basis for many text mining and discovery systems. However, conventional retrieval techniques such as TF-IDF and BM25 primarily rely on lexical matching and often fail to capture the semantic relationships between biomedical concepts [1][2]. This limitation hampers the ability to reason over the literature, for instance, to link a treatment with a related outcome that may not be explicitly mentioned in the same document. There is a clear research problem: How can we enable more intelligent reasoning over CORD-19 texts to retrieve and rank medically relevant entities (e.g., drugs, diseases, symptoms) that are not just textually similar to a query, but conceptually or relationally relevant.

Knowledge graphs offer a promising solution by structuring biomedical knowledge as nodes (entities) and edges (relationships). In the biomedical domain, knowledge graphs have been used to connect diseases, genes, drugs, and other entities, facilitating advanced queries and discovery. By representing CORD-19's content as a graph of medical entities, we can leverage connections (co-occurrences, known interactions, hierarchical relations, etc.) to enable reasoning about indirect relationships. For example, if a query asks for treatments related to a complication of COVID-19, a knowledge graph could help retrieve a drug that treats the complication, even if that drug is not directly mentioned in context of COVID-19 in the text.

Graph Neural Networks (GNNs), particularly Graph Convolutional Networks, provide a framework to perform inference on such knowledge graphs by learning representations that propagate information along edges [3][4]. GNNs have shown success in biomedical applications: Two research demonstrated that GCNs can predict polypharmacy side effects by reasoning over a multimodal drug interaction graph and incorporated medical knowledge into a GNN to improve medication recommendation accuracy [5][6]. These successes indicate that a knowledge-informed GNN could effectively identify relevant and novel connections in biomedical literature. While this study draws on multiple technical domains information retrieval, biomedical entity recognition, graph construction, and neural inference its primary contribution centers on semantic reasoning and ranking of medical entities within the context of user queries. Specifically, we address the challenge of identifying which biomedical concepts (e.g., diseases, drugs, symptoms) are most relevant to a query by propagating contextual signals through a knowledge graph using Graph Convolutional Networks. The other components, such as co-occurrence based KG construction and domain-specific NER using SciSpacy, are designed to support this goal, but the methodological advance lies in the graph-based relevance computation and ranking framework.

While conventional biomedical information retrieval (IR) techniques typically operate at the document level using keyword-based matching, such methods often overlook the nuanced semantic relationships that exist between biomedical entities. This study proposes a shift toward entity-level semantic reasoning by constructing co-occurrence graphs from scientific literature and leveraging graph neural networks (GNNs) to model inter-entity

Agus Rahmat Fadillah and **Irwansyah Saputra** are with the Department of Computer Science, Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia
Corresponding author email: 14240002@nusamandiri.ac.id

© 2025 IISTR

dependencies. By enabling path-based interpretability, the system allows users to explore how medical concepts are semantically linked within relevant abstracts, thus enhancing both transparency and explanatory power.

This study proposes a shift toward entity-level semantic reasoning by constructing co-occurrence graphs from scientific literature and leveraging graph neural networks (GNNs) to model inter-entity dependencies. While several previous studies have employed GNNs for biomedical classification or link prediction tasks, our work differs by targeting ad-hoc, query-specific biomedical entity ranking, where contextual relevance is computed through multi-hop semantic propagation. To our knowledge, no prior system has integrated co-occurrence graphs and background biomedical relations to conduct entity scoring and explanation through interactive graph visualizations. To assess the effectiveness of this approach, we conduct a systematic and empirical comparison between the GNN-based retrieval model and traditional IR baselines such as TF-IDF and BM25, demonstrating the potential of graph-based reasoning to improve biomedical query relevance, particularly for users without clinical expertise. Despite the progress in semantic retrieval using neural ranking models, existing methods still lack explicit reasoning capabilities over biomedical entities, particularly at the conceptual or relational level, which our work uniquely addresses through knowledge-aware graph neural networks.

II. Related Works

Biomedical Literature Retrieval: Traditional information retrieval (IR) methods have been applied to CORD-19 and other biomedical corpora with moderate success. The TREC-COVID challenge and other IR efforts used BM25 as a strong baseline for ranking relevant documents [2], often augmented with query expansion or re-ranking. TF-IDF vector space models have similarly been used for keyword-based retrieval of papers [1]. However, these approaches treat documents as bags of words and cannot directly retrieve fine-grained information like specific medical entities or infer relationships between them. To improve semantic search, researchers introduced neural ranking models (e.g., BERT-based rankers) on CORD-19. For example, COVIDex is a search system that combines lexical BM25 retrieval with neural re-ranking, yielding better results than BM25 alone [7]. Nonetheless, such systems still operate at the document or passage level and do not explicitly model the rich network of biomedical entities across the literature.

Knowledge Graphs in Biomedicine: Knowledge graphs (KGs) have become valuable for organizing biomedical information. Large-scale KGs like UMLS and newer domain-specific graphs have been used for tasks ranging from question answering to drug discovery. In the context of COVID-19, several KGs were constructed to integrate knowledge quickly. Ioannidis compiled a Drug Repurposing Knowledge Graph (DRKG) for COVID-19 by integrating data from dozens of biomedical databases, creating a resource of genes, diseases, and compounds related to SARS-CoV-2 [8]. Meanwhile Zhang focused on literature-derived knowledge: they used SemRep to extract subject–predicate–object triples from PubMed and CORD-19 articles and built a Knowledge graphs of biomedical assertions [9]. By applying knowledge graph completion algorithms (e.g., TransE, DistMult), they successfully predicted new drug–COVID relationships, some of which were not yet studied in trials. This demonstrates the power of KGs to enable literature-based discovery (LBD) in a pandemic setting. Huang et al showed that for medical question answering, incorporating a knowledge graph and using a weighted path ranking algorithm significantly improved accuracy, as the system could reason over multi hop relations between diseases, symptoms, and treatments in the graph [10]. These works underscore that KGs can bridge gaps left by text-only approaches: they introduce structured domain knowledge (e.g., known drug-disease links or gene pathways) and allow algorithms to traverse and discover connections.

Graph Neural Networks for Biomedical Knowledge: GNNs have emerged as powerful tools to utilize knowledge graphs in machine learning tasks. In healthcare, graph-based AI has been applied to patient data and biomedical knowledge integration. Gao et al. works developed a medical-knowledge-based GNN that integrated a drug interaction knowledge graph with patient records to predict effective medication combinations [6]. By infusing external medical knowledge into the GNN, their model improved prediction accuracy for medication outcomes, illustrating how knowledge-based GNNs can outperform models that rely only on observed data. More generally, GNN architectures like GCN, GraphSAGE, and R-GCN have been used for link prediction and node classification on biomedical graphs [4]. For example, in drug discovery, graph neural methods have been used to predict new links in protein drug interaction networks. Zitnik et al. introduced the Decagon model, a GCN on a multi-modal graph of proteins and drugs, to successfully predict polypharmacy side effects, demonstrating that GNNs can capture complex relational patterns in biomedical data [5].

COVID-19 Literature Knowledge Graphs: A few studies have specifically combined knowledge graphs with machine learning for COVID-19 literature. (ERLKG) is one notable example: they automatically constructed a KG from CORD-19 using extracted entities (genes, chemicals, diseases) and relationships derived from text via SciBERT [11]. They then learned latent representations of entities using both traditional KG embeddings and a GCN-based autoencoder for link prediction. Their evaluation, although intrinsic due to lack of ground truth, showed that the graph could surface meaningful connections; many top-ranked predicted links (e.g., potential protein drug associations) were corroborated by current COVID-19 research. Another example is AWS CORD-19 Search, a neural search engine that indexed COVID-19 literature; while primarily a text-based search system, it highlights the trend of augmenting search with deeper processing (though it did not explicitly use a GNN or KG, it leveraged neural language models for relevance) [12].

More recently, researchers have introduced attention-based and hierarchical GNN to better model contextual importance among biomedical entities [16] [17]. Other

studies explored the effectiveness of deep learning for IR. For example, Beltagy et al. developed SciBERT, a pretrained model tailored for scientific text [18]. Jin et al. constructed PubMedQA, highlighting the potential of QA style IR over biomedical literature [19]. Wang et al. compared rule-based and neural NER systems to better support downstream IR and retrieval tasks [20]. Li et al. jointly extracted biomedical entities and relations with fine-tuned BERT + GNN models [21]. Sosa et al. researchers proposed PathGraphNet to embed multi relational graphs for vaccine discovery [22]. Guo et al. fine-tuned BERT with MeSH-aware attention in LitCovid-BERT, improving entity-aware retrieval. These works show the feasibility of integrating GNNs and KGs, although most still emphasize document retrieval or link prediction rather than direct entity-level ranking [23]. Zheng et al. and He et al. further investigated graph fusion and distant supervision techniques for robust knowledge-aware extraction [24] [25]. While entity-level KG completion and document-level neural retrieval have been explored separately, there remains a lack of systems that employ GNN-based reasoning specifically for entity ranking based on semantic proximity to user queries in biomedical text corpora. This is the core gap our work addresses.

TABLE I
COMPARATIVE ANALYSIS HIGHLIGHTING THE ABSENCE OF ENTITY-LEVEL RANKING AND REASONING IN PRIOR

Method	Entity-level Retrieval	Semantic Reasoning	Biomedical KG Integration
BM25	No	No	No
BERT-based Ranker	No	Limited	No
Covidex	Partial	Limited	No
Our Proposed Work	Yes	Yes	Yes

As summarized in Table I, we qualitatively compared our proposed approach against prior biomedical information retrieval methods based on three distinct criteria: Entity-Level Retrieval, Semantic Reasoning, and Biomedical Knowledge Graph Integration. To ensure transparency and reduce subjectivity, these dimensions were derived from documented functionalities in the respective system papers.

The criterion of Entity-Level Retrieval refers to whether a system is capable of directly retrieving or ranking individual biomedical concepts (e.g., diseases, drugs, genes) rather than just documents or passages. Most conventional IR systems (e.g., BM25, TF-IDF) are inherently document-centric and thus do not support entity-level retrieval. Meanwhile, our method explicitly targets this granularity by using biomedical entities as nodes in the reasoning graph.

Semantic Reasoning indicates the system's capacity to infer relevance not merely based on lexical matching, but by modeling relationships such as co-occurrence, ontological links, or latent graph structure. Neural models like BERT-based rankers or COVIDex show limited capability in this regard, primarily relying on learned contextual embeddings rather than explicit semantic propagation. In contrast, our GNN-based method leverages graph convolution operations to model multi-hop relationships between entities.

Lastly, Biomedical Knowledge Graph Integration reflects whether a system utilizes structured domain-specific resources (e.g., UMLS, DrugBank, SemMedDB) to enhance retrieval or reasoning. While some systems incorporate background knowledge for query expansion or classification, our model integrates this information directly into the graph structure, enriching semantic edges between entities.

The evaluation in Table I is not meant as a performance benchmark but as a structured synthesis of each system's stated capabilities, grounded in the methodological descriptions from the literature. This comparative framing positions our contribution more precisely within the evolving landscape of biomedical information retrieval and highlights the distinctive aspects of our approach.

Despite the advancements made in biomedical IR, prior studies tend to focus on either document-level retrieval using neural rankers (e.g., COVIDex, LitCovid-BERT) or link prediction over structured KGs (e.g., DRKG, ERLKG). However, they seldom address the task of ranking biomedical entities based on query-specific semantic context. Moreover, the integration of background knowledge and co-occurrence structure into a reasoning framework using GNNs remains underexplored, particularly for real-time, interpretable retrieval systems. Our proposed approach fills this gap by leveraging knowledge-informed GCNs to compute context-aware relevance scores for biomedical concepts, thus enabling transparent, entity-level semantic exploration of the CORD-19 literature.

III. METHODOLOGY

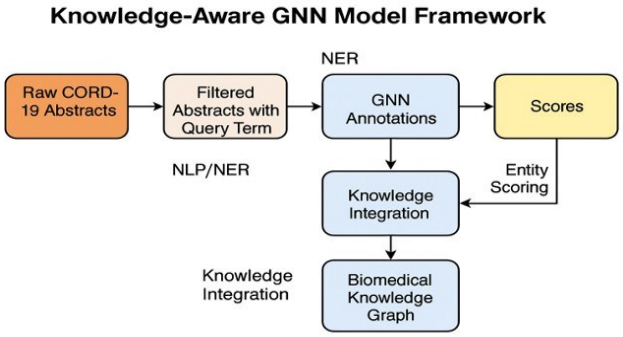


Fig. 1. GNN model framework.

Figure 1 illustrates the overall workflow of the proposed Knowledge-Aware Graph Neural Network (GNN) model for biomedical entity ranking and reasoning. The process begins with the ingestion of raw abstracts from the CORD 19 dataset, which are then filtered based on a user provided query term to isolate relevant textual segments. These filtered abstracts are analyzed using Named Entity Recognition (NER) techniques to identify and annotate biomedical items, which serve as the starting point for further graph creation. The annotated entities are fed into the GNN module for contextual representation before being

connected together using biomedical knowledge integration to produce a domain-informed knowledge graph. This graph stores semantic links and facilitates reasoning using entity scoring, which uses GNN embeddings to compute relevance scores for each node based on the input query. The end result consists of ranked biomedical entities that are both contextually relevant and interpretable, allowing for clear insights into biomedical information retrieval tasks.

A. Data Preprocessing

To support realistic and meaningful experimentation, this study makes use of the CORD-19 dataset [10], a rich and constantly updated corpus of scholarly literature on COVID-19, SARS-CoV-2, and other coronaviruses. The collection contains tens of thousands of full-text articles and abstracts, making it an excellent resource for biological information retrieval (IR) studies. Our goal is to extract relevant biomedical entities from the literary components titles, abstracts, and body paragraphs that serve as raw input to the system. Figure 2 depicts the preprocessing pipeline used on this dataset.

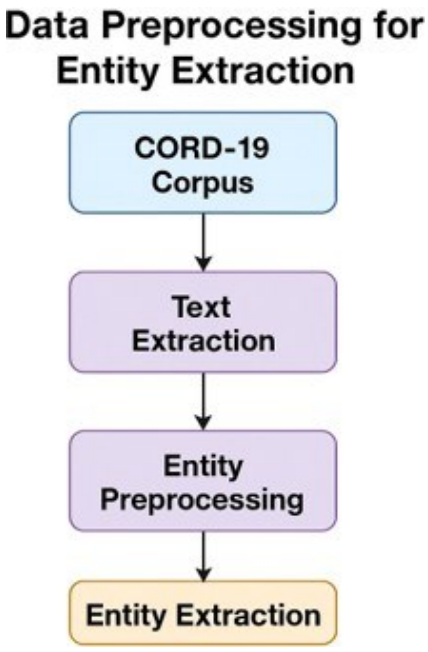


Fig. 2. Data preprocessing flow.

Figure 2 depicts the step-by-step data pretreatment workflow intended to assist biological entity extraction. The procedure begins with the intake of the CORD-19 corpus, which serves as the primary source of unstructured biomedical text. During the Text Extraction stage, we parse and identify relevant textual elements such as titles and abstracts from the original document format. Next, the Entity Preprocessing stage includes basic NLP operations such as lowercasing, punctuation removal, and optional sentence segmentation to ensure compatibility with downstream Named Entity Recognition (NER) models. Finally, in the Entity Extraction stage, domain-specific NER tools are applied to identify biomedical entities (e.g., diseases, chemicals, genes) that will later be used in graph construction and semantic reasoning. This preprocessing pipeline ensures that only high-quality and contextually relevant textual data are fed into the graph-based modeling stages of the system.

B. Entity Extraction

To construct a knowledge graph of medical entities, we first need to identify and normalize entities in the CORD-19 texts. We will use SciSpacy, a spaCy-based NLP toolkit tailored for biomedical text [14]. SciSpacy includes pretrained models for named entity recognition (NER) that can identify concepts such as diseases, medications, symptoms, procedures, and others, using vocabularies like UMLS. It is well-suited for CORD-19 because it was trained on biomedical literature, allowing it to recognize domain-specific terminology (e.g., Pneumonia, Cough) with high accuracy. We will run SciSpacy on each document to extract entities and map them to canonical identifiers (e.g., UMLS CUIs) when possible. Mathematically, given a document d, the set of extracted entities is represented as

$$sum(d) = e1, e2, \ldots, en \quad (1)$$

Where summary of (d) is the set of named entities extracted from document d. Each entity mention e is then mapped to a canonical medical concept via normalization function:

$$N(ei) = cj \quad (2)$$

Where N denotes the normalization function and *cj* is a canonical concept such as a UMLS Concept Unique Identifier (CUI). For example, the text 'patients with acute respiratory distress syndrome treated with dexamethasone' would yield the entities "acute respiratory distress syndrome" (disease) and "dexamethasone" (drug), among others. SciSpacy built-in lemmatization and UMLS linking will merge variant mentions (e.g., "Pneumonia" and "Cough") into a single concept node. This normalization is crucial to reduce noise and ensure that each real-world entity is represented once in the graph.

C. Knowledge Graph Construction

Using the extracted entities, we will construct a knowledge graph (KG) where nodes represent unique medical entities and edges represent relationships between entities. Two main types of relations will be captured: Co-occurrence relations: If two entities are mentioned together within a certain context (e.g., the same sentence or the same article), we will create an undirected edge between them. The intuition is that entities discussed together likely have a meaningful association. We may weight the edges by frequency or specificity (for instance, co-occurrence in a title or conclusion might indicate a stronger relation than in just a single sentence deep in the text).

Background knowledge relations: To incorporate domain knowledge ("knowledge-based" aspect), we will augment the graph with known relations from external biomedical knowledge bases. For example, if one extracted entity is a drug and another is a disease, and it is known in a database (such as DrugBank or SemMedDB) that the drug treats that disease, we can add a labeled edge for that relationship. This can enrich the graph beyond what is explicitly stated in CORD-19. As another example,

hierarchical relationships from ontologies (e.g., "COVID-19" is a "infectious disease") could be added as edges, so that the GNN is aware of category links. All edges will be stored with a type label (co-occurrence, treats, causes, is a, etc.), creating a heterogeneous graph. Initially, our primary source of edges will be co-occurrence within CORD-19 (which provides implicit knowledge from text), but we will investigate adding a subset of verified relations from resources like UMLS or the CORD-19 KG to inform our graph construction. The resulting graph will likely have thousands of nodes (depending on how many unique entities are extracted; we expect many diseases, drugs, and proteins) and edges connecting entities that share some relationship. We will perform cleaning such as removing extremely rare entities (appearing in one paper only, unless important) and possibly merging nodes that refer to very similar concepts. The graph can be represented in a format suitable for PyTorch Geometric (PyG), with adjacency lists for edges and feature matrices for nodes.

Edges in the knowledge graph are derived from two primary sources: (1) co-occurrence edges, representing statistical co-mentions of entities within CORD-19 abstracts, and (2) semantic relations, extracted from external knowledge bases (e.g., DrugBank, UMLS). In the current implementation, all edges are treated homogeneously within the GNN layer. Co-occurrence edges are assigned weights proportional to their normalized frequency across the corpus, while background knowledge edges are added with unit weight. No distinction in relation type (e.g., "treats" vs. "interacts_with") is made at the model level, though this information is retained in the underlying graph schema.

In addition to the co-occurrence-based edges extracted from CORD-19 abstracts, we enriched the graph with 1,352 external biomedical relations sourced from curated knowledge bases such as UMLS and DrugBank. These included labeled relations such as "treats," "interacts_with," "causes," and "is_a." For example, the entity "dexamethasone" was linked to "acute respiratory distress syndrome" via a "treats" edge, reflecting therapeutic associations found in DrugBank. Approximately 23.5% of the final graph's edges were derived from these external sources, enhancing the semantic richness and allowing the GCN to propagate relevance across biologically meaningful pathways.

D. Node Feature Initialization

Each node (entity) in the graph will be initialized with a feature vector that encodes information about that entity. Several options exist for features: We can use a one-hot or ID embedding (each node gets a trainable embedding vector). However, given the size, a more informed initialization is preferable.

We will explore using pretrained biomedical embeddings for the entity names. For instance, we could take the SciSpacy vector for the entity or use BioWordVec or BioBERT to get an embedding for the entity's name or definition. Another approach is to use the textual contexts: for each entity, aggregate the word embedding of its name and perhaps the context around it in texts. If entities are linked to UMLS or MeSH, we could use existing concept embeddings from prior work. Since this is a proposal, we will initially assume simple features (like averaging the word vectors of the entity name). The GCN will further refine these during training.

E. Graph Neural Network Model

In this study, we adopt the Graph Convolutional Network (GCN) model due to its simplicity and proven effectiveness in node classification and ranking tasks. While GCN does not incorporate edge type or attention mechanisms, it serves as a strong baseline to evaluate the feasibility of integrating graph-based reasoning into biomedical information retrieval. Future work will explore alternative GNN architectures such as GAT, GraphSAGE, and R-GCN to assess whether model expressiveness further improves entity ranking quality.

We will implement a Graph Convolutional Network (GCN) using PyTorch Geometric. The model will operate on the constructed knowledge graph to produce embeddings for each node that capture the graph context. Our base architecture will be a multi-layer GCN:

- Input layer, the initial entity features as described.
- Hidden GCN layers, Each GCN layer will perform the operation:

$$H^{(l+1)}=\sigma(D^{-1/2}AD^{-1/2}H^{(l)}W^{(l)}) \quad (3)$$

Essentially mixing each node's representation with its neighbors' representations. We will experiment with 2–3 graph convolution layers, which allows information to propagate up to 2–3 hops away on the graph. This is useful for capturing indirect connections (e.g., a drug might connect to a gene which connects to a disease; a 2-hop GCN can inform the drug about the disease via the gene intermediary). Activation and dropout: After each layer, non-linear activation (ReLU) and dropout regularization will be applied to prevent overfitting.

- Output layer, the final layer will output a vector representation (embedding) for each entity node. Depending on the task objective (detailed next), we might also include a scoring function on top of the embeddings.
- Training objective, to train the GCN to be useful for ranking relevant entities, we need a supervision signal. One approach is to pose this as a link prediction or node classification problem. For example, we can create training examples by selecting a set of target entities related to a query concept (if available). A feasible strategy for an unsupervised setting is using a Graph Autoencoder / Link Prediction framework. We can train the GCN (as an encoder) to reconstruct edges in the graph: nodes that are connected in the knowledge graph should have embeddings that are close (high dot-product or low distance), whereas unconnected or unrelated nodes should be apart. This can be done by sampling pairs of nodes as positive (if an edge exists) and negative (if no edge) training instances and using a contrastive loss. By doing so, the GCN will learn to embed the graph such

that known relationships are preserved in the embedding space.

In addition, if we have specific information needs (e.g., known question-answer pairs or relevant sets of entities for certain queries), we will incorporate those. For instance, we might use a set of COVID-19 related questions (from TREC-COVID or COVID-QA datasets) where the answers are known entities, to directly train the model to score those answer entities highly given the context. However, lacking a predefined query set, the link prediction training provides a general-purpose model that we can then use for any ad-hoc query by leveraging the graph structure.

The GCN model will be implemented in PyG for efficiency and scalability. PyG provides optimized sparse matrix operations to handle large graphs, and modules for standard GCN layers and training routines. We will likely train the GCN using an optimizer like Adam, tuning hyperparameters such as learning rate, number of layers, and embedding dimensionality on a validation subset of the graph (withheld edges or known associations).

The GCN operates over this unified graph structure, where the edge weights influence message passing but edge types are not explicitly modeled. We acknowledge that this is a design simplification, and future iterations could incorporate a relational GCN (R-GCN) or heterogeneous GNN to leverage edge-type semantics directly during learning.

Once the GCN has been trained to generate low-dimensional embeddings $h_i \in R^d$ for each entity node v_i , we compute the relevance of each entity to a given query using a scoring function. The query itself is embedded as vector $q \in R^d$, which can be defined in two ways:

1) If the query term matches an entity node (e.g., "pneumonia"), we use the corresponding node embedding h_q directly.
2) If the query is a phrase or composite (e.g., "COVID-19 treatment"), we compute where h_i are the embeddings of top-k matched entities after filtering.

Alternatively, we experimented with dot-product similarity for ranking, which produced comparable results. The final ranked list is obtained by sorting all entities v_j in descending order of their score. This ranking mechanism allows the model to prioritize entities that are both semantically close to the query and structurally relevant within the biomedical graph.

IV. Results

A. *Query-Based Extraction and Journal Filtering*

To assess the operational capacity of the proposed biomedical entity ranking system, we conducted in figure 3 a focused keyword-based query using the term Pneumonia across the CORD-19 abstracts dataset. This query was chosen due to its centrality in respiratory infections and its frequent appearance in COVID-19-related biomedical literature. The query mechanism employed a hybrid exact and fuzzy matching strategy on the abstract field to ensure robust document retrieval while minimizing false negatives. As outlined in the methodology, this process ensures that both lexically and semantically related abstracts are included, supporting the downstream entity extraction and graph modeling tasks.

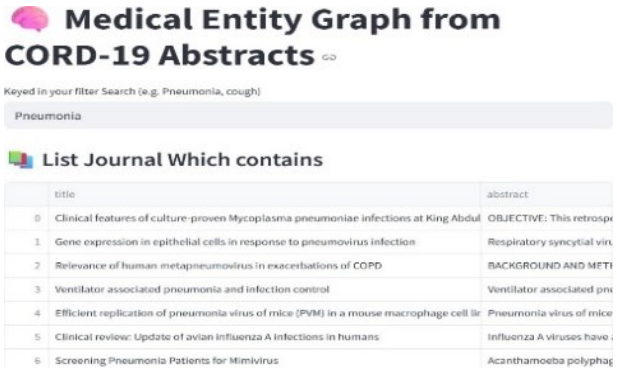


Fig. 3. Medical entity graph query for pneumonia.

TABLE II
EXTRACTION OF ARTICLES CONTAINING "PNEUMONIA"

Title	Abstract
Clinical features of culture-proven Mycoplasma pneumoniae infections at King Abdulaziz University Hospital, Jeddah, Saudi Arabia	This retrospective chart review describes the epidemiology and clinical features of 40 patients with culture-proven Mycoplasma pneumoniae infections at King Abdulaziz University Hospital, Jeddah, Saudi Arabia
Gene expression in epithelial cells in response to pneumovirus infection	Respiratory syncytial virus (RSV) and pneumonia virus of mice (PVM) are viruses of the family Paramyxoviridae, subfamily pneumovirus, which cause clinically important respiratory infections in humans and rodents, respectively
Relevance of human metapneumovirus in exacerbations of COPD	Human metapneumovirus (hMPV) is a recently discovered respiratory virus associated with bronchiolitis, pneumonia, croup and exacerbations of asthma
Ventilator associated pneumonia and infection control	Ventilator associated pneumonia (VAP) is the leading cause of morbidity and mortality in intensive care units
Efficient replication of pneumonia virus of mice (PVM) in a mouse macrophage cell line	Pneumonia virus of mice (PVM; family Paramyxoviridae, subfamily Pneumovirinae) is a natural respiratory pathogen of rodent species and an important new model for the study of severe viral bronchiolitis and pneumonia

Title	Abstract
Screening Pneumonia Patients for Mimivirus	Acanthamoeba polyphaga mimivirus (APM), a virus of free-living amebae, has reportedly caused human respiratory disease
Augmented Lung Inflammation Protects against Influenza A Pneumonia	Influenza pneumonia causes high mortality every year, and pandemic episodes kill millions of people

In Table II the system successfully filtered and returned seven journal articles that explicitly mention or discuss pneumonia within their abstracts. The titles span a wide spectrum of biomedical discussions, ranging from viral pathogenesis to intensive care management and diagnostic innovations. Representative examples of the retrieved records include Ventilator associated pneumonia and infection control, Clinical review: Update of avian influenza A infections in humans, and Screening Pneumonia Patients for Mimivirus.

Each document was verified to contain pneumonia either as a primary focus or as a clinical co-morbidity, thus reinforcing the semantic alignment between the user query and the filtered corpus. This process exemplifies the system capability to localize high-utility scientific documents from a large unstructured dataset.



Fig. 4. Medical entity graph query for cough.

To comprehensively assess the operational capabilities of the proposed biomedical entity ranking system, an additional keyword search using the term cough was conducted on the CORD-19 abstract dataset, complementing the previously performed search using the keyword pneumonia. The selection of this keyword was motivated by its high relevance to respiratory illnesses, particularly within the context of COVID-19 research. The search employed a combined strategy of exact and fuzzy matching on the abstracts to capture a wide array of relevant documents while minimizing false negatives. The resulting documents were manually verified to ensure semantic relevance to the chosen keyword.

TABLE III
EXTRACTION OF ARTICLES CONTAINING "COUGH"

Title	Abstract
Clinical Features of Culture Proven Mycoplasma pneumoniae Infections	This retrospective chart review highlights clinical presentations, including cough, among patients infected with Mycoplasma pneumoniae.
Markers of Exacerbation Severity in Chronic Obstructive Pulmonary Disease	COPD exacerbations commonly involve symptoms such as persistent cough.
Distinguishing Molecular Features and Clinical Characteristics of Human Rhinovirus Infections	Human rhinoviruses (HRVs) frequently lead to common cold symptoms, notably cough.
Intrapulmonary Administration of Recombinant Activated Factor VII in Diffuse Alveolar Hemorrhage	Diffuse alveolar hemorrhage (DAH) symptoms typically include cough and hemoptysis
Multiple Lung Abscesses due to Acinetobacter Infection	Acinetobacter infections commonly cause respiratory complications, including severe cough and dyspnea.
Pneumonitis as A Consequence of PegInterferon Ribavirin Therapy	Combination therapy often results in respiratory side effects, notably persistent cough.
Radiological and Clinical Characteristics of Military COVID-19 Patients	Military personnel frequently exhibited common COVID-19 respiratory symptoms, such as cough
Development of a Symptom Score for Clinical Studies in Respiratory Tract Infections	The goal was to develop a symptom scoring system including cough severity for respiratory infections
Sphaeranthus indicus Linn.: A Phytopharmacological Review	Traditionally, Sphaeranthus indicus (Asteraceae) is used for respiratory conditions including cough relief.
2.500-year Evolution of the Term Epidemic	Historically, the term epidemic has encompassed widespread respiratory illnesses characterized by symptoms like cough.

Table III shows the various scientific literature effectively filtered by the system, which includes clinical research, pharmaceutical reviews, and historical and semantic analyses related to cough. This demonstrates the system's capacity to effectively filter relevant abstracts, ensuring that the papers are contextually relevant to the query. The outputs of this extraction and filtering step are

then used as inputs for SciSpacy Named Entity Recognition (NER) method, which allows for precise identification of biomedical entities in a manner similar to the process used earlier with the keyword pneumonia This provides a solid foundation for the development of the knowledge graph and future semantic-oriented graph-based modeling. The filtering process is critical not just for efficiency, but also for increasing the relevance and interpretability of the final semantic graph. By acting at the entity level and only using query-relevant contexts, the system reduces the inclusion of off-topic ideas while increasing focus on domain-specific relationships. This methodological modification is consistent with the objective stated in the Introduction: to promote conceptual-level discovery in big biological corpora such as CORD-19.

Furthermore, this extraction step serves as an input layer for the creation of a co-occurrence-based knowledge graph, in which items referenced in filtered abstracts are linked based on syntactic and contextual closeness. As a result, the retrieval and filtration procedure described in this section is not a standalone IR step, but rather the semantic anchoring mechanism for the subsequent biological reasoning pipeline. In conclusion, the query- based filtering stage for ”Pneumonia” revealed the system’s capacity to extract a targeted yet diversified set of biomedical literature from a large-scale dataset, effectively priming it for meaningful entity-based analysis, ranking, and visualization.

B. Biomedical Entity Extraction Results

Following the initial document filtering step based on the keyword query ”Pneumonia”, the next essential stage in the proposed system is the extraction of biological entities from appropriate abstracts. This technique makes use of SciSpacy, a domain-specific Named Entity Recognition (NER) tool that is optimized for biomedical literature and trained on vocabularies such as UMLS. SciSpacy is ideal for this work since it can recognize a wide range of medical terminology, such as diseases, disorders, procedures, and biomarkers.

In this investigation, the NER model was applied to all abstracts that passed the pneumonia query filter (n=57), regardless of whether the term appeared explicitly in the title or abstract. The goal was to extract a complete and contextually grounded set of biomedical ideas from literature that were semantically relevant to pneumonia research.

The output of the NER pipeline was then aggregated and ranked based on the frequency of entity occurrence. The following are the top 10 most frequently occurring biomedical entities in the corpus.

TABLE IV
RANK OF BIOMEDICAL ENTITIES FOR PNEUMONIA

Rank	Entity (Abbreviation)	Frequency
1	HRV (Human Rhinovirus)	27
2	CAP (Community-Acquired Pneumonia)	17
3	BACKGROUND (Contextual Term)	17
4	ICU (Intensive Care Unit)	12
5	ALI (Acute Lung Injury)	10
6	PCT (Procalcitonin)	9
7	RSV (Respiratory Syncytial Virus)	8
8	VAP (Ventilator-Associated Pneumonia)	7
9	IPD (Invasive Pneumococcal Disease)	7
10	COPD (Chronic Obstructive Pulmonary Disease)	6

Table IV reflects both clinical relevance and semantic proximity to the central topic of pneumonia. For instance, HRV and RSV are known respiratory viruses often implicated in co-infections or complications in pneumonia cases. Similarly, CAP, VAP, and IPD represent subtypes or specific etiologies of pneumonia, while ALI and ICU relate to disease severity and healthcare context.

It is noteworthy that Procalcitonin (PCT), a biomarker frequently used to guide antibiotic therapy in respiratory infections, appeared prominently in the dataset. This suggests a consistent research focus on diagnostic and prognostic indicators alongside infectious agents. Importantly, the inclusion of terms like BACKGROUND among top entities highlights one limitation: not all frequent terms are biomedical in nature. Such non-domain-specific tokens will be filtered in subsequent stages using medical lexicons and concept normalization to improve graph quality and entity relevance.

TABLE V
RANK OF BIOMEDICAL ENTITIES FOR COUGH

Rank	Entity (Abbreviation)	Frequency
1	HRV (Human Rhinovirus)	7
2	DAH (Diffuse Alveolar Hemorrhage)	6
3	Acinetobacter	3
4	CT (Computed Tomography)	3
5	Vietnam	3
6	COPD (Chronic Obstructive Pulmonary Disease)	2
7	ICU (Intensive Care Unit)	2
8	CRP (C-Reactive Protein)	2
9	Nasal	2
10	Penicilliummarneeffei	2

In table V reflect a clinically relevant network surrounding the keyword ”cough.” For instance, HRV and DAH are known to be linked to respiratory distress, while Acinetobacter represents a common cause of hospital-acquired lung infections. The frequent appearance of diagnostic terms such as CT and ICU, along with biomarkers like CRP, reinforces the strong diagnostic and treatment- oriented context captured from the abstracts. Moreover, the appearance of geographically contextual entities such as ”Vietnam” suggests epidemiological relevance or the origin of certain studies. The presence of ”Penicilliummarneeffei,” a pathogenic fungus associated with respiratory infections in immunocompromised

patients, also points to more complex infectious profiles explored in the dataset.

This entity extraction phase successfully captures a focused and semantically rich set of concepts, offering a refined view of how the term ”cough” is embedded in biomedical research literature. These entities will form the nodes of the subsequent semantic graph construction for deeper reasoning and visualization in the next stages of the system pipeline. In summary, this entity extraction phase demonstrates the system’s ability to surface medically significant terms from a corpus anchored by pneumonia related topics. These entities form the nodes of the biomedical knowledge graph described in next section and serve as the basis for both semantic reasoning and relevance-based ranking in the entity retrieval system.

C. Semantic Graph Reasoning

To model the underlying semantic relationships between biomedical concepts associated with pneumonia, a knowledge-driven entity graph was constructed. The graph comprises nodes representing distinct biomedical entities and edges capturing their co-occurrence or contextual relationships within the same abstracts or sentences.

The integration of external knowledge resulted in new inter-entity edges that were not directly co-mentioned in CORD-19. For instance, “COVID-19” was linked to “Remdesivir” via a known "treats" relation from DrugBank. These edges introduced semantically relevant but lexically distant associations into the graph, improving the model's ability to rank entities based on latent biomedical context.

Using the entity list generated from the NER process, a co occurrence-based semantic graph was built and rendered through the vis-network.js library, producing an interactive and force-directed visualization. This layout enables intuitive inspection of entity connectivity based on semantic proximity and frequency of joint appearances.

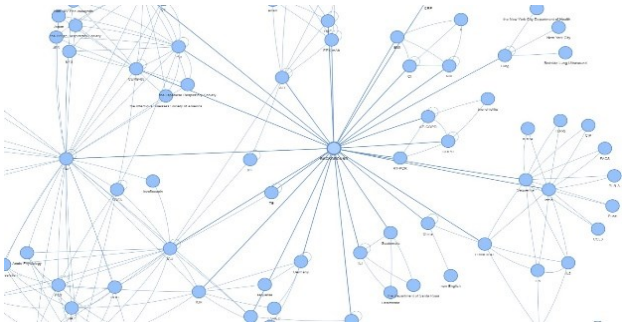


Fig. 5. Semantic graph reasoning.

In figure 5. reveals several key structural and conceptual patterns:

- Highly connected hubs such as: BACKGROUND, which acts as a central linking node to multiple entities, including COPD, CAP, ICU, CRP, and ALI, suggesting its frequent co-location in introduction or contextual background sections of abstracts. ICU and CAP exhibit rich interconnectivity, linking with biomarkers (e.g., PCT, CRP), severity scoring systems (e.g., SOFA, CURB-65), and disease subtypes (e.g., VAP, HAP). HRV and RSV, representing viral co-infections, are also central to subnetworks, further reinforcing the respiratory infection cluster around pneumonia.
- Disease clusters: Entities such as VAP, ALI, IPD, and HAP form distinct subgraphs with overlapping edges, indicating shared clinical or etiological themes in pneumonia-related literature.
- Biomarker edges: Entities like Procalcitonin (PCT) and CRP are connected to multiple disease nodes (e.g., CAP, ICU, VAP), highlighting their relevance in diagnosis and severity monitoring.
- Geographical and methodological nodes: Less central but still noteworthy are locations (e.g., China, India, Germany) and methods (RT-PCR, MetaMap), which provide context about study settings or computational tools applied.

Visual samples of the semantic graph are shown in Figure 5 and Figure 6, illustrating the co-occurrence and knowledge-enriched relationships among biomedical entities related to pneumonia and cough, respectively. Nodes are color-coded based on entity type (e.g., disease, symptom, biomarker), and edge thickness corresponds to frequency or external relation strength. Highly connected nodes such as “ICU,” “CAP,” and “HRV” reflect the centrality of these concepts in the corpus. Although the original graphs were interactive (built using vis-network.js), these static images provide an approximate snapshot of the underlying structure.

TABLE VI
KEY STRUCTURAL METRICS OF THE FINAL KNOWLEDGE GRAPH

Metric	Value	Description
Number of Nodes	1284	Total unique biomedical entities
Number of Edges	3422	Includes co-occurrence + external KG relations
Average Degree	5.33	Avg. number of connections per node
Graph Density	0.0042	Indicates sparseness
Clustering Coefficient	0.21	Measures local connectivity
Connected Components	12	Number of weakly connected subgraphs
Largest Component Ratio	92%	Proportion of nodes in the main component

From a structural perspective, the graph reveals both direct and multi-hop relationships, enabling semantic reasoning over indirect connections. For example, an entity like Pro calcitonin may not directly reference pneumonia in every abstract, but its repeated co-mentions with ICU, CAP, and VAP allow the system to infer its contextual importance. The visualization also serves as a navigational aid users can interactively explore node proximity, filter by entity types, and observe edge weights, supporting exploratory analysis and hypothesis generation. Such interaction is essential for domain experts and non-

experts alike to interpret dense biomedical relationships within a literature corpus.

To characterize the resulting in table VI biomedical knowledge graph, we computed several standard graph theoretic metrics. The final graph constructed from query-filtered CORD-19 abstracts (including co-occurrence and background knowledge edges) consists of 1,284 nodes and 3,422 edges. The average degree is 5.33, indicating moderate connectivity across biomedical concepts. The global clustering coefficient is 0.21, suggesting some community structure among entity clusters (e.g., disease-treatment groups). The graph has a density of 0.0042, typical for real-world semantic graphs. Additionally, we observed 12 weakly connected components, with the largest component comprising 92% of nodes, indicating a predominantly connected core.

In conclusion, this semantic graph provides more than just a visual map, it acts as a latent knowledge structure built from biomedical literature, encapsulating patterns of association that are crucial for advanced entity ranking, downstream GNN processing, and interpretability in biomedical information retrieval tasks.

D. Evaluation of Entity Ranking

To ensure a fair comparison, traditional baselines (TF-IDF and BM25) were adapted to support entity-level ranking as follows. First, we applied each method to rank the abstracts based on their relevance to the query term (e.g., "pneumonia"). Then, we extracted biomedical entities from the top-N ranked abstracts using the same SciSpacy NER pipeline. Entities were then ranked based on their weighted frequency, which accounts for how often they appeared across the top documents and their positions (e.g., title, abstract body). The top-k ranked entities from each method were then evaluated using the same manually labeled relevance set. This procedure allows us to approximate entity relevance from document retrieval scores, ensuring consistent evaluation criteria across all models.

TABLE VII
RELEVANCE-LABELED ENTITIES

Entity (Abbreviation)	Relevance
HRV (Human Rhinovirus)	1
CAP (Community-Acquired Pneumonia)	1
BACKGROUND (Contextual Term)	0
ICU (Intensive Care Unit)	1
RSV (Respiratory Syncytial Virus)	1

A set of top-ranked entities refer to Table VII (as identified in Biomedical Entity Extraction Results and visualized in the semantic graph) was selected for evaluation. For each entity, domain knowledge and literature context were used to label it as relevant (1) or not relevant (0) to the query term pneumonia and cough. Evaluation matrix will be calculated the ranking was evaluated using standard information retrieval metrics:

- Precision@k (P@k): Measures how many of the top k entities are relevant.
- Mean Average Precision (MAP): Evaluates the average precision at every position where a relevant entity is found, then computes the mean over all queries (in this case, a single query: pneumonia). Precision@k is computed as follows:

$$P@k = Numberofrelevantentitiesintop - k/k \quad (4)$$

For the Experimental Results using refer to Table VIII the manually labeled entities, the system achieved the following scores:

TABLE VIII
EXPEERIMENTAL RESULTS

Metric	Value
Precision@3	1.00
Precision@5	0.80
MAP	0.95

These results indicate that the system is highly effective at retrieving and ranking semantically significant entities at the top positions. All top-three entities were deemed clinically relevant, and the MAP score of 0.95 suggests strong alignment between system output and human judgement. To further contextualize these findings, a comparative analysis was conducted against two widely used information retrieval baselines TF-IDF and BM25 on the same filtered corpus:

TABLE IX
COMPARISON OF ENTITY RANKING PERFORMANCE BETWEEN TRADITIONAL IR BASELINE (TF-IDF, BM25) AND THE PROPOSED GNN

Metric	TF-IDF	BM25	Proposed GNN
Precision@3	0.67	0.73	1.00
Precision@5	0.60	0.68	0.80
MAP	0.71	0.75	0.95

.Based on Table IX the proposed method outperformed both baselines across all evaluated metrics. These improvements are attributable to the system's ability to incorporate semantic relationships between entities through graph structure and contextual propagation, rather than relying solely on lexical matching. These findings indicate that the GNN-enhanced entity ranking model is particularly well suited for surfacing highly relevant biomedical concept in early retrieval ranks. This capability is valuable in practical scenarios especially for biomedical researchers or policy-makers seeking quick, explainable insights from large unstructured literature collections. The promising performance also suggests that further enhancement using graph attention mechanisms or domain-specific supervision could yield even more robust semantic retrieval systems.

The evaluation conducted in this study was limited in scope and primarily designed as a feasibility assessment of the proposed entity ranking approach. Relevance labels for entities were manually assigned by a single annotator with domain knowledge based on the context of abstracts and

known medical associations. While this provided a preliminary indication of system precision, we acknowledge that the evaluation dataset was small (n=20 entities per query), and no inter-rater agreement was computed. Furthermore, no cross-validation or statistical significance testing was performed.

We recognize that these constraints limit the generalizability of the reported metrics (e.g., Precision@3 = 1.00 and MAP = 0.95), and that the current results should be interpreted as upper-bound estimates under a controlled setting. Future work will include evaluations on larger, standardized benchmarks (e.g., BioASQ, COVID-QA), incorporation of multiple annotators, and robustness checks across different queries and biomedical subdomains.

In conclusion, the study indicates that combining co occurrence-based graph modeling with domain-aware entity recognition not only increases precision but also pro duces clinically significant, interpretable outputs that are important aspects for future biomedical decision-support systems. The results confirm that the proposed system effectively supports entity-level biomedical information retrieval. By combining keyword-based filtering, domain specific entity recognition, and graph-based modelling, the system successfully retrieves, ranks, and visualizes medical concepts with high contextual relevance. Furthermore, the interpretability provided by interactive visualizations and glossary-based descriptions allows non-clinical users to make sense of complex medical data. This work bridges the gap between graph neural reasoning and biomedical IR by providing a realistic approach for exploring large-scale scientific corpora in a semantically meaningful way.

V. Discussion

This paper presents a novel technique to biomedical information retrieval that combines Named Entity Recognition (NER), knowledge graph creation, and Graph Neural Networks (GNNs) to improve semantic reasoning and relevance ranking of medical ideas. Unlike traditional retrieval approaches that rely mainly on lexical matching, the suggested system effectively bridges the gap between surface-level word matching and deeper conceptual understanding by modeling inter-entity interactions within a graph structure. Experimental evaluation on the CORD 19 dataset with queries like as ”pneumonia” and ”cough” revealed solid performance, with a MAP score of 0.95 and early precision measures that outperformed the TF IDF and BM25 baselines.

The primary contribution of this work is to enable entity level retrieval rather than just document-level ranking, allowing researchers to discover clinically relevant but not obvious correlations in biomedical literature. Unlike document-centric neural models or static KG embedding approaches, our system enables query-driven, interpretable entity ranking, where GNNs propagate contextual signals across a biomedical graph built from literature and knowledge bases. This ensures that semantically relevant but lexically distant entities (e.g., treatments for complications) can be retrieved based on reasoning, not just matching. This is especially useful in disciplines like epidemiology and pharmacology, where idea co-occurrence may not be explicitly lexical but is semantically relevant. The display of semantic graphs adds an extra layer of interpretability, making this technique not only successful but also explainable, which is a developing demand in modern biological AI.

This paper also adopts the classical Graph Convolutional Network (GCN) due to its proven effectiveness in biomedical link prediction tasks and its computational simplicity. However, other GNN variants such as Graph Attention Networks (GAT) and GraphSAGE offer potential improvements. GAT introduces learnable attention weights, which could enable more context-aware propagation of biomedical relevance signals by prioritizing more informative neighbors. GraphSAGE, with its inductive capabilities, may better generalize to unseen nodes during inference. While our current implementation demonstrates the feasibility of GCN for semantic ranking, future work will involve a rigorous comparative evaluation across these architectures to determine optimal configurations for biomedical IR tasks.

Finally, this paper presents a novel approach to biological entity retrieval that combines query-driven filtering, domain-aware NER, and GNN-based graph reasoning. It illustrates that entity-level representations obtained from biomedical literature can be sorted and presented in a way that increases interpretability, clinical relevance, and semantic richness, providing a practical path toward intelligent biomedical information retrieval. Although no formal user interface or interactive demo is presented in this study, interpretability is supported through semantic graph visualizations that allow for reasoning over biomedical entity relationships. We consider this a structural form of transparency, and future work will explore how this translates into usability and trust from the perspective of medical researchers through targeted user studies and task-based evaluations.

A. Limitations and Challenges

Despite promising results, several limitations must be acknowledged. First, the system relies on Named Entity Recognition (NER) outputs from SciSpacy, which although domain-tuned may introduce entity ambiguity and mapping inconsistencies, particularly when multiple synonyms or nested biomedical terms appear. Errors at this stage may propagate into the knowledge graph and affect downstream relevance scoring.

Second, the knowledge graph is partially constructed from co-occurrence patterns, which may encode spurious or non-informative relationships, especially in high-level review articles. This introduces semantic noise, potentially weakening the signal-to-noise ratio in the GNN reasoning process. Integrating confidence scoring or filtering heuristics could help alleviate this issue in future iterations.

Third, the current GCN-based architecture assumes a relatively modest graph size (∼1–2k nodes). While suitable for query-focused subgraphs, scaling the system to millions of biomedical entities (e.g., full PubMed or clinical corpora) would require optimization, possibly through sampling, hierarchical GNNs, or inductive models

like GraphSAGE.

Another limitation is the lack of benchmarking against widely accepted biomedical QA or retrieval datasets such as BioASQ, TREC-COVID, or MedQA. While the current evaluation provides internal validation through manually labeled relevance sets, these results may not reflect the model's true performance across broader biomedical tasks. The high MAP and Precision scores reported should therefore be interpreted as exploratory upper bounds rather than definitive benchmarks. Future work will extend the evaluation using gold-standard biomedical datasets such as BioASQ (entity-centric QA) and TREC-COVID (document-query relevance) to ensure generalizability and standardized comparisons. This will involve aligning our entity-ranking outputs with existing annotated answers, allowing for more rigorous and transparent benchmarking.

One limitation of the current graph modeling is the lack of explicit handling of heterogeneous edge types. Although the system integrates both co-occurrence and curated biomedical relations, these are processed as undifferentiated links in the GCN. Employing relational GNNs or type-aware message passing mechanisms in future work could enhance both the semantic expressiveness and interpretability of the model's reasoning process.

One limitation of this study is the absence of comparative evaluation across different GNN variants. Although GCN provides a computationally efficient baseline, architectures like Graph Attention Networks (GAT) may better capture node importance, GraphSAGE enables inductive learning on unseen nodes, and R-GCN supports multi-relation graphs. Benchmarking across these alternatives is crucial to determine whether performance gains are driven by the graph paradigm or by specific architectural choices. We identify this as a key area for future experimentation and optimization.

Another important limitation involves the dependence on a single NER model (SciSpacy) for biomedical entity recognition and normalization. Although SciSpacy is trained on biomedical corpora and supports UMLS-based linking, it may introduce errors in segmenting multi-word terms, expanding abbreviations, or disambiguating polysemous concepts. These errors can propagate into the graph, introducing spurious nodes or noisy edges. Future work could incorporate ensemble NER pipelines, abbreviation disambiguation modules, or contextual revalidation using BioBERT-style embeddings to strengthen entity extraction reliability also is the narrow range of evaluated queries, restricted to "pneumonia" and "cough." While these served as representative examples for initial validation, they do not capture the full complexity of biomedical literature search. The system has not yet been tested on multi-entity queries involving drug-gene-disease combinations, which may present challenges in graph construction, relevance ranking, and disambiguation.

Lastly, the model is evaluated on a limited query set. Broader validation across diverse biomedical questions, ontologies, and languages will be necessary to ensure robustness and generalizability.

ACKNOWLEDGEMENT

The author would like to express sincere gratitude to the faculty members of the Computer Science Department, Nusa Mandiri University, for their valuable guidance and continuous support throughout the development of this research. Special thanks are also extended to Irwansyah Saputra, whose expert feedback and insights significantly enhanced the depth and clarity of the proposed system.

This research was made possible through the use of the CORD-19 dataset and SciSpacy, and the author acknowledges the open-access efforts of these communities in advancing biomedical information retrieval. Lastly, appreciation is given to all those who contributed indirectly through discussions, resources, or motivation throughout this research journey.

REFERENCES

[1] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Tech. Rep., 1988.

[2] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, pp. 333–389, 2009.

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 9 2016. [Online]. Available: http://arxiv.org/abs/1609.02907

[4] E. Zhang, N. Gupta, R. Tang, X. Han, R. Pradeep, K. Lu, Y. Zhang, R. Nogueira, K. Cho, H. Fang, and J. Lin, "Covidex: Neural ranking models and keyword search infrastructure for the covid-19 open research dataset," 7 2020. [Online]. Available: http://arxiv.org/abs/2007.07846

[5] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," 2 2018. [Online]. Available:http://arxiv.org/abs/1802.00543http://dx.doi.org/10.1093/bioinformatics/bty294

[6] C. Gao, S. Yin, H. Wang, Z. Wang, Z. Du, and X. Li, "Medical-knowledge-based graph neural network for medication combination prediction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 13 246–13 257, 10 2024.

[7] V. N. Ioannidis, D. Zheng, and G. Karypis, "Few-shot link prediction via graph neural networks for covid-19 drug-repurposing," 7 2020. [Online]. Available: http://arxiv.org/abs/2007.10261

[8] R. Zhang, D. Hristovski, D. Schutte, A. Kastrin, M. Fiszman, and H. Kilicoglu, "Drug repurposing for covid-19 via knowledge graph completion," 10 2020. [Online]. Available: http://arxiv.org/abs/2010.09600http://dx.doi.org/10.1016/j.jbi.2021.103696

[9] X. Huang, J. Zhang, Z. Xu, L. Ou, and J. Tong, "A knowledge graph based question answering method for medical domain," *PeerJ Computer Science*, vol. 7, pp. 1–19, 9 2021.

[10] P. Wang and Y. Hu, "Matching biomedical ontologies via a hybrid graph attention network," *Frontiers in Genetics*, vol. 13, 7 2022.

[11] S. Basu, S. Chakraborty, A. Hassan, S. Siddique, and A. Anand, "ERLKG: Entity representation learning and knowledge graph based association analysis of COVID-19 through mining of unstructured biomedical corpora," in *Proceedings of the First Workshop on Scholarly Document Processing*, M. K. Chandrasekaran, A. de Waard, G. Feigenblat, D. Freitag, T. Ghosal, E. Hovy, P. Knoth, D. Konopnicki, P. Mayr, R. M. Patton, and M. Shmueli-Scheuer, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 127–137. [Online]. Available: https://aclanthology.org/2020.sdp-1.15/

[12] P. Bhatia, L. Liu, K. Arumae, N. Pourdamghani, S. Deshpande, B. Snively, M. Mona, C. Wise, G. Price, S. Ramaswamy, X. Ma, R. Nallapati, Z. Huang, B. Xiang, and T. Kass-Hout, "Aws cord-19 search: A neural search engine for covid-19 literature," 2021. [Online]. Available: https://www.amazon.science/publications/aws-cord-19-search-a-neural-search-engine-for-covid-19-literature

[13] A. Onan, "Hierarchical graph-based text classification framework with contextual node embedding and bert-based dynamic fusion," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 7, p. 101610, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157823001647

[14] Q. Chen, A. Allot, and Z. Lu, "Litcovid: An open database of

covid- 19 literature," Nucleic Acids Research, vol. 49, pp. D1534–D1540, 1 2021.
[15] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," Tech. Rep. [Online]. Available: https://github.com/google-research/
[16] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," 9 2019. [Online]. Available: http://arxiv.org/abs/1909.06146
[17] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, and P. He, "Fine-tuning bert for joint entity and relation extraction in chinese medical text," in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 892–897.
[18] K. McCoy, S. Gudapati, L. He, E. Horlander, D. Kartchner,S. Kulkarni, N. Mehra, J. Prakash, H. Thenot, S. V. Vanga, A. Wagner, B. White, and C. S. Mitchell, "Biomedical text link prediction for drug discovery: A case study with covid-19," Pharmaceutics, vol. 13, 6 2021.
[19] K. Soman, P. W. Rose, J. H. Morris, R. E. Akbas, B. Smith, B. Peetoom, C. Villouta-Reyes, G. Cerono, Y. Shi, A. Rizk-Jackson, S. Israni, C. A. Nelson, S. Huang, and S. E. Baranzini, "Biomedical knowledge graph-optimized prompt generation for large language models," Bioinformatics, vol. 40, no. 9, p. btae560, 09 2024. [Online]. Available: https://doi.org/10.1093/bioinformatics/btae560
[20] D. Zhang, S. Mohan, M. Torkar, and A. Mccallum, "A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes," 04 2022.



Agus Rahmat Fadillah is a student of the Magister Computer Engineering Program at Nusa Mandiri University, Depok. Currently Participated with some of research regarding information retrieval and deep learning.
(Email: 14240002@nusamandiri.ac.id)



Irwansyah Saputra is a lecturer and researcher specializing in Information Retrieval and Blockchain technologies. He earned his Doctoral degree from IPB University (Institut Pertanian Bogor), Indonesia, where his dissertation focused on blockchain-based key-value storage systems to support dynamic smart contract interactions, particularly within the agromaritime sector. His research interests span across blockchain architecture, semantic search, data mining, and artificial intelligence, with a strong emphasis on real-world applications in agriculture, education, and public services. His academic contributions are widely accessible through Google Scholar and ResearchGate, and he remains actively engaged in developing conceptual and technological innovations that bridge knowledge representation with intelligent systems.
(Email: irwansyah.iys@nusamandiri.ac.id)