Learning to Rank Complex Biomedical Hypotheses for Accelerating Scientific Discovery

Juncheng Ding[1], Shailesh Dahal[2], Bijaya Adhikari[2] and Kishlay Jha[2]
[1] *University of North Texas, Denton, TX, USA*
[2] *University of Iowa, Iowa City, IA, USA*
{shailesh-dahal, bijaya-adhikari, kishlay-jha}@uiowa.edu, junchengding@my.unt.edu

2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) | 979-8-3503-8373-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICHI61247.2024.00044

***Abstract*—Hypothesis generation (HG) is a fundamental problem in biomedical text mining that uncovers plausible implicit links (*B* terms) between two disjoint concepts of interest (*A* and *C* terms). Over the past decade, many HG approaches based on distributional statistics, graph-theoretic measures, and supervised machine learning methods have been proposed. Despite significant advances made, the existing approaches have two major limitations. First, they mainly focus on enumerating hypotheses and often neglect to *rank* them in a semantically meaningful way. This leads to wasted time and resources as researchers may focus on hypotheses that are ultimately not supported by experimental evidence. Second, the existing approaches are designed to rank hypotheses with only one intermediate or evidence term (referred as simple hypotheses), and thus are unable to handle hypotheses with multiple intermediate terms (referred as complex hypotheses). This is limiting because recent research has shown that the complex hypotheses could be of greater practical value than simple ones, especially in the early stages of scientific discovery.**

To address these issues, we propose a new HG ranking approach that leverages upon the expressive power of Graph Neural Networks (GNN) coupled with a domain-knowledge guided Noise-Contrastive Estimation (NCE) strategy to effectively rank both simple and complex biomedical hypotheses. Specifically, the message passing capabilities of GNN allows our approach to capture the rich interactions between biomedical entities and succinctly handle the complex hypotheses with variable intermediate terms. Moreover, the proposed domain knowledge guided NCE strategy enables the ranking of complex hypotheses based on their coherence with the established biomedical knowledge. Extensive experiment results on five recognized biomedical datasets show that the proposed approach consistently outperforms the existing baselines and prioritizes hypotheses worthy of potential clinical trials.

***Index Terms*—biomedical text mining, hypothesis generation, graph neural networks, self-supervised learning**

I. Introduction

Hypothesis generation is a crucial step in developing testable propositions (or predictions) that after undergoing rigorous testing and evaluation leads to scientific discoveries [1]. Traditionally, scientists rely on their intuition, creativity, and prior knowledge acquired by selectively reading numerous articles to form hypotheses. However, in modern data-intensive era, keeping up with the vast amount of relevant literature is impractical for individual researchers or teams. As an illustration, consider a January 2024 PubMed (the most comprehensive life-sciences search engine [2]) search on the topic of "Alzheimer's disease" that resulted in over 200,000 citations. If a research team were to read 20 papers per day, it would take them approximately 28 years, by which time millions more articles would have emerged [3]. This overwhelming amount of literature can create major bottlenecks in generating novel hypotheses, as researchers cannot efficiently explore the vast space of continually growing biomedical information landscape.

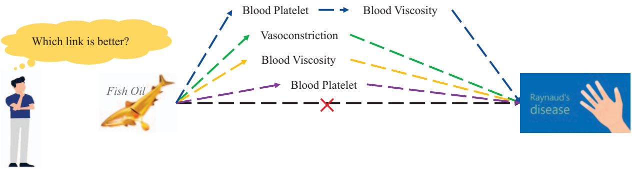


Fig. 1: Example of simple and complex hypothesis between Fish oils and Raynaud's disease. The simple hypothesis contains only one intermediate concept (e.g., Blood Viscosity), and the complex ones contain multiple concepts (e.g., Blood Platelet → Blood Viscosity) with different connectivity patterns.

To address this challenge, researchers in the biomedical text mining communities have shown increased interest in developing computational models that can mine the scientific literature to automatically suggest hypotheses that are new, interesting, testable, and likely to be true. As a result, automated hypothesis generation (HG) has established itself as a fundamental problem within biomedical text-mining [4] that aims to discover cross-silo connections (also referred to as undiscovered public knowledge [5]) by stitching together the already known and established scientific facts that remain dispersed across the literature. In other words, given an input concept of interest (e.g., disease or gene), HG attempts to find implicit links (e.g., potential drug target or novel indicator of disease's mechanism) that connects them in a previously unknown but semantically meaningful way. Different from standard link prediction problems [6] that solely focuses on predicting links between entities, HG aims to provides a *rationale* (or evidence) in the form of connecting terms for a particular hypothesis. Figure 1 shows an example of a possible hypothesis between "Raynaud's disease" and "Fish Oils" via intermediate (or evidence) terms such as "Blood Viscosity" and "Blood Platelets".

Over the past few years, many HG approaches [7]–[10]

2575-2634/24/$31.00 ©2024 IEEE
DOI 10.1109/ICHI61247.2024.00044

have been proposed in the literature. Broadly, they can be categorized into three major groups: a) distributional approaches [11], [12], b) embedding-based methods [7], [8], and c) supervised machine learning based approaches [13]. While these prior studies made significant advances, a common limitation lies in their emphasis on enumerating as many hypotheses as possible, without necessarily ranking them in a informative way. This lack of informative ranking can lead to wasted time and resources, as researchers may focus on hypotheses that are ultimately not supported by experimental evidence [14]. To overcome this issue, it is imperative to develop an informative HG ranking approach that identifies the most promising hypotheses from tens of thousands of unranked hypotheses for possible *in-vitro* clinical trials. This is the core objective of the proposed research in this paper.

A conventional way to rank the generated hypotheses is by using information retrieval metrics such as term frequency-inverse document frequency (TF-IDF), BM25, and query likelihood model [15]. More recently, neural network based biomedical embedding models [16]–[18] have been proposed and obtained significant improvement in performance. Despite significant advances, the existing approaches are coarse and thus unable to rank concepts with paucity of training data in the scientific corpus (e.g., domain-specific or rare concepts). In other words, the existing ranking methods do not take guidance from the known biological knowledge to rank the candidate hypotheses. Moreover, almost all of the existing ranking methods are designed to rank hypotheses with only one intermediate term (i.e., simple hypotheses). It is unclear on how the existing approaches would rank hypotheses with multiple intermediate terms (i.e., complex hypotheses).

To address these issues, we design a novel HG ranking approach that is guided by the known biological knowledge and accurately ranks both simple and complex biomedical hypotheses. Specifically, we propose to model the relationships between biomedical entities as a graph structure and develop a HG tailored Graph Neural Network (GNN) that can effectively capture the intricate connectivity patterns crucial for ranking the hypotheses accurately. Moreover, the proposed GNN based ranking model effectively captures the rich interactions between intermediate terms of variable length, making them adept at ranking complex hypotheses accurately. Further, we propose a domain-knowledge guided Noise-Contrastive Estimation (NCE) based strategy that enables identification of statistically significant hypotheses and assists in ranking of complex hypotheses based on their coherence with the established biomedical knowledge.

Altogether, the proposed approach effectively addresses the unique challenges of hypothesis ranking in the biomedical domain and facilitates the identification of top-ranked hypotheses for possible clinical trials. Finally, the proposed approach is designed to work as an *add-on* enhancer module to the existing HG approaches. Given the fact that there are multiple competing HG approaches [8]–[10], [14], it is desirable to develop approaches that do not jeopardize the HG training process and flexibly enables the users to utilize the proposed ranking module as a pluggable module for obtaining ranked hypotheses.

In this research, our contributions can be summarized as:

- We propose a new HG ranking approach that accurately ranks both simple and complex hypotheses. Central to our approach is the emphasis on handling *complex* hypotheses that is both novel and has immediate practical benefits in applications such as drug-repurposing and precision medicine.
- The proposed research designs a new domain-knowledge guided NCE strategy that learns to distinguish between well-supported hypotheses and spurious ones. Moreover, the domain knowledge guidance ensures that the hypotheses are aligned with the biological factors or prior knowledge.
- The experimental results corroborate the efficacy of the proposed HG ranking approach - we obtain a significant improvement over baselines in terms of Spearman's correlation @Top-*K*. Qualitative evaluation of results demonstrate that the top ranked hypotheses are plausible and worthy of further investigations.

II. Preliminaries

A. PubMed

We will use PubMed [2] as our training corpus. PubMed consists of a vast collection of articles encompassing the fields of life sciences and biomedicine. With over 32 million documents, this corpus encompasses a diverse range of study types. Each article entry in PubMed includes essential data elements such as the title, abstract, medical subject headings (MeSH) terms, author names, affiliations, publication date, journal information, and citation details.

B. Medical Subject Headings (MeSH)

In this paper, MeSH will be used as the source of domain knowledge. MeSH is controlled vocabulary curated and maintained by subject matter experts at the National Library of Medicine (NLM). Since these are curated by domain experts, they are highly precise and widely used as a resource for domain knowledge in the biomedical domain. Every article in the PubMed (specifically indexed by MEDLINE) are assigned MeSH terms. These terms encapsulate the conceptual meaning of the article. On average, every article in PubMed is assigned 12 MeSH terms. Moreover, MeSH terms are organized in an hierarchical fashion (i.e., *ISA* tree). The distance between concepts in the tree indicates the degree of semantic proximity between them. The depth of a concept in the tree indicates its level of specificity.

III. Approach

A. Overview of Proposed HG Ranking Model

For clear explanation of our proposed HG ranking model, we will use PubMed as the corpus and MeSH terms as the unit of analysis (i.e., hypotheses) throughout this paper. However, we note that the proposed method is entirely general and can

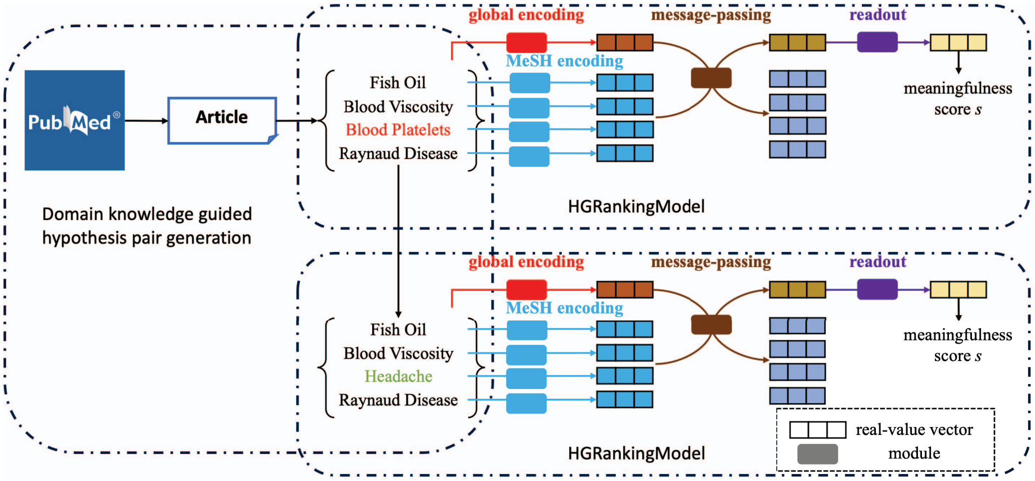


Fig. 2: The architecture of proposed HG ranking model. It takes a set (i.e., a link or hypothesis) as the input and outputs the meaningfulness score.

be readily adapted with other choices of corpora and units of analysis.

Our goal is to rank both simple and complex hypotheses generated from the biomedical corpus. Specifically, a hypothesis in this paper is a MeSH term set $\{M_1, \cdots, M_L\}$ that connects two disjoint input MeSH terms with a variable number of intermediate (or evidence) MeSH terms in a meaningful way. As such, we use a bag (unordered set) of MeSH terms to represent an article in PubMed as $\{M_1, \cdots, M_L\}$, where M_i is a MeSH term and L can be different for different articles. The input to our proposed HG ranking model is a hypothesis set as defined above, and the output is a real-number score indicating the hypothesis's meaningfulness. We propose to use these scores to compare and rank hypotheses. To obtain these scores, we propose to first model the relational dependencies between MeSH term using a GNN. This is because GNN are a natural choice to handle graph sets (i.e., MeSH term set in current problem setting) with variable length (different intermediate terms). Moreover, we propose to train the proposed ranking module in a self-supervised way by contrasting the pairs between meaningful hypotheses (i.e., published articles in PubMed) and random/meaningless hypotheses (i.e., articles generated using a random MeSH terms). Notably, this self-supervised strategy overcomes the necessity of a labeled dataset for hypotheses ranking that are both time-consuming and monetarily expensive to obtain. Fig. 2 show overview of our proposed model.

B. HG Ranking Model Structure

The general idea of proposed model is to encode a set of L MeSH terms (i.e., a link or a hypothesis) as a vector in latent space. The meaningfulness of such a MeSH set is indicated by the vector's $norm$. Since each input to the model may have variable numbers of MeSH terms, or L, an effective HG ranking model should be able to handle input sets with variable length. As such, we propose a HG tailored GNN framework that has as an *encoding*, a *message-passing*, and a *readout* module. *Encoding* encodes each of the input MeSH terms and the set into the latent space as the MeSH term vectors and the set global vector, respectively. These two groups of vectors feed further into *message-passing*, which "passes the message" between MeSH term vectors and the set global vector, to ensure that they both contain the set's context information. *readout* decodes the set global vector as the model's output whose $norm$ indicates the set's meaningfulness.

MeSH encoding $\boldsymbol{\mathcal{E}_\mathbf{e}}$**.** This module encodes MeSH terms $M_i, i = 1 \cdots L$ into the latent space as $\mathbf{v}_{i,\text{enc}}, i = 1 \cdots L$. It uses an *embedding* layer ($\mathcal{E}_{\text{emb}}$) to map MeSH terms into embeddings (i.e., a learnable lookup table or matrix), and a two-layer *Multiple-Layer Perceptron* (MLP) as in Equation 1 to encode the embeddings into the latent space. Note that Equation 1 is a general description of MLP used throughout the whole proposed model and $\mathbf{A}_p$ is the weight for different components in different parts of the whole model. Equation 2 is our detailed implementation of the MeSH encoding module. Note that the $\mathcal{E}_\text{e}$ are identical for all input MeSH terms.

$$\mathcal{E}_{\text{MLP}}(\mathbf{v}) = \mathcal{L}_2\left(\mathcal{L}_1(\mathbf{v})\right), \mathcal{L}_t(\mathbf{v}) = \text{ReLU}(\mathbf{A}_p\mathbf{v}) \tag{1}$$

$$\mathbf{v}_{i,\text{enc}} = \mathcal{E}_\text{e}\left(M_i\right) = \mathcal{E}_{\text{MLP}}\left(\mathcal{E}_{\text{emb}}\left(M_i\right)\right), i = 1 \cdots L \tag{2}$$

where $\mathbf{A}_p$ are the learnable parameters of the respective MLPs. In the MeSH encoding module, we use the same MLP with the same weight for different MeSH terms. Note that throughout the proposed model, MLP structure remains the same whereas the weights $\mathbf{A}_p$ are different in different parts of the model.

Global encoding $\boldsymbol{\mathcal{E}_\mathbf{g}}$**.** The proposed ranking model generates a default global vector $\mathbf{v}_{\text{g,init}} = [0.5]$ for the input set and uses an MLP in Equation 1 (with a different shape and set of weights) to encode the global vector into $\mathbf{v}_{\text{g,enc}}$, in the same latent space as MeSH term vectors, as $\mathbf{v}_{\text{g,enc}} = \mathcal{E}_{\text{MLP}}(\mathbf{v}_{\text{g,init}})$. The set global vector will carry the set's context information after *message-passing*.

***Message-passing* MP.** This module ensures the set global vector and the MeSH term vectors contain the set's context information by "passing" messages between them, i.e., updating them according to each other alternatively. It has two steps: *Set Update* (SU) and *MeSH Update* (MU). SU, as in Equation 4, updates the set global vector according both to terms' and its previous one. MU, as in Equation 5, updates each term's vector independently according to its previous one and the updated global vector after SU. MP runs N times using the identical network and Equation 3 presents the n^{th} round MP. Note that MP does not exchanges the information between MeSH term vectors (nodes) via edges as in typical GNNs [19] because the set has no edges as [20].

$$\begin{aligned} \mathbf{v}_{\text{g},n+1,\text{mp}}, \mathbf{V}_{n+1,\text{mp}} &= \text{MP}\left(\mathbf{v}_{\text{g},n,\text{mp}}, \mathbf{V}_{n,\text{mp}}\right) \\ &= \text{MU}\left(\text{SU}\left(\mathbf{v}_{\text{g},n,\text{mp}}, \mathbf{V}_{n,\text{mp}}\right), \mathbf{V}_{n,\text{mp}}\right) \\ \mathbf{V}_{n,\text{mp}} &= \left[\mathbf{v}_{1,n,\text{mp}}; \cdots ; \mathbf{v}_{L,n,\text{mp}}\right] \end{aligned} \tag{3}$$

$$\begin{aligned} \mathbf{v}_{\text{g},n+1,\text{mp}} &= \text{SU}\left(\mathbf{v}_{\text{g},n,\text{mp}}, \mathbf{V}_{n,\text{mp}}\right) \\ &= \mathcal{E}_{\text{MLP}}\left(\left[\mathbf{v}_{\text{g},n,\text{mp}}; \text{AGGR}_{\text{SU}}(\mathbf{V}_{n,\text{mp}})\right]\right) \\ \text{AGGR}_{\text{SU}}(\mathbf{V}_{n,\text{mp}}) &= \sum_{i=1}^{L} \mathbf{v}_{i,n,\text{mp}} \end{aligned} \tag{4}$$

$$\begin{aligned} \mathbf{V}_{n+1,\text{mp}} &= \text{MU}\left(\mathbf{v}_{\text{g},n+1,\text{mp}}, \mathbf{V}_{n,\text{mp}}\right), \text{or} \\ \mathbf{v}_{i,n+1,\text{mp}} &= \mathcal{E}_{\text{MLP}}\left(\left[\mathbf{v}_{\text{g},n+1,\text{mp}}; \mathbf{v}_{i,n,\text{mp}}\right]\right), i = 1 \cdots L \end{aligned} \tag{5}$$

where $[\mathbf{x}; \mathbf{y}]$ is the concatenation of $\mathbf{x}, \mathbf{y}$ vectors. $\mathcal{E}_{\text{MLP}}$ in the above equations have the same structure as in Equation 1. However, they have different learnable parameters. AGGR_{SU} is the aggregation function. As suggested in previous studies [20], we use the sum aggregation as default. However, we will compare different aggregation functions in our experiments. The input of the MP_0 is $\mathbf{v}_{\text{g,enc}}$ and $\mathbf{V}_{enc} = [\mathbf{v}_{1,\text{enc}}; \cdots ; \mathbf{v}_{L,\text{enc}}]$, and it outputs $\mathbf{v}_{\text{g},N,\text{mp}}$ and $\mathbf{V}_{N,\text{mp}}$ in the last round. Note that the MP module remains exactly the same but is applied to each MeSH in the set. This is central reason for our proposed model to employ GNN at its core - our model can address any set with different number of MeSH terms without necessarily relying on its order or cardinality.

***readout* $\mathcal{R}_{\mathbf{g}}$.** After N rounds of MP, the updated set global vector $\mathbf{v}_{\text{g},N,\text{mp}}$ contains the set's context information. $\mathcal{R}_{\text{g}}$ further decodes $\mathbf{v}_{\text{g},N,\text{mp}}$ into the output vector $\mathbf{v}_{\text{g,output}}$ as Equation 6. The output score is the norm of the final output vector as $s = |\mathbf{v}_{\text{g,output}}|$ indicating the set's meaningfulness.

$$\mathbf{v}_{\text{g,output}} = \mathcal{R}_{\text{g}}\left(\mathbf{v}_{\text{g},N,\text{mp}}\right) = \text{Tanh}\left(\mathbf{A}_2\left(\text{ReLU}\left(\mathbf{A}_1\left(\mathbf{v}_{\text{g},N,\text{mp}}\right)\right)\right)\right) \tag{6}$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are learnable parameters. We choose Tanh as the final activation to constrain the output vector into a fixed range.

In summary, the proposed ranking model scores a set (i.e., a link or a hypothesis) as $s_{\text{set}} = \text{HGRankingModel}(\{M_1, \cdots, M_L\})$ on its meaningfulness. We use 128 as the latent size and 3 as the number of "message-passing" to balance the performance and efficiency in our implementation. These hyperparameters are recommended either by the existing literature [19], [20] or Python package implementation as defaults for similar design, and we will also test other options in the ablation studies.

C. Self-supervised Learning for Proposed HG Ranking Model

One of the central challenges [12] in HG studies the availability of ground truth results (i.e., clinically validated hypotheses). This is because conducting *in-vitro* clinical trials is both time-consuming and costly. At the same time, biomedical domain has bibliographic repositories that are curated and maintained by domain experts. Thus, it is imperative to design a method that trains our ranking model with the guidance of highly specialized domain knowledge automatically. We propose a self-supervised learning algorithm to train the proposed ranking model using only the existing literature instead of labeled hypotheses. Similar to NCE [21], the proposed algorithm generates noise and forces the model to differentiate the noise and real data in the training process. The noise generation in our algorithm is different from that in NCE because we have no noise (false hypotheses) distribution. Instead, we propose a domain knowledge-guided positive-negative pair generation algorithm that generates the training data. Another key feature of our proposed self-supervised learning algorithm is the loss function. To ensure that our model can compare and rank hypotheses of different types (numbers of concepts), we also propose a new loss function for our model training algorithm. Generally speaking, the central idea of NCE is to train a neural network via enforcing it to differentiate positive data samples and negative data samples. Following this idea, in our training algorithm, we propose to generate positive and negative samples guided by the domain knowledge. In HGRankingModel, a positive sample is a valid hypotheses and should be ranked high or get a high meaningfulness score. On the other hand, a negative sample should be less valid than the positive sample and get a lower meaningfulness score. We will introduce how we get the two kinds of sample in the following.

PubMed indexes each published article with a few concepts (MeSH terms) without any specific ordering (i.e., bag of words). Since all the concepts are annotated by domain experts, we argue that the concepts in one article are describing a certain topic with respect to the article and are highly coherent. In other words, we can consider each article's concepts set as a meaningful set or a positive data sample in the context of NCE-based training algorithm. Now, to generate useful negative samples, we propose to generate less meaningful articles (negative samples) $\mathcal{D}_{\text{worse}} = \{M_1, \cdots, M_{k-1}, M_{k+1}, \cdots, M_L, M_{L+1}\}$ from existing ones $\mathcal{D}_{\text{original}} = \{M_1, \cdots, M_k, \cdots, M_L\}$, by randomly choosing M_k and changing it to a random MeSH term M_{L+1} which has never co-occurred with other concepts in the set. After generation positive-negative pairs, we feed each $(\mathcal{D}_{\text{worse}}, \mathcal{D}_{\text{original}})$

pair into two identical HGRankingModel and score each of them.

In the ranking model training, we sample each pair independently from each $\mathcal{D}_{\text{original}}$, via selecting a MeSH term M_k from $\mathcal{D}_{\text{original}}$ following a uniform distribution over the MeSH terms in $\mathcal{D}_{\text{original}}$ and sampling an M_{L+1} following a uniform distribution over the MeSH term vocabulary to replace M_k, in each training epoch. Note that the sampled term M should have never co-occurred with any term in any existing article $\mathcal{D}_{\text{original}}$ during the sampling process. We propose to train the model by constraining it to score $\mathcal{D}_{\text{original}}$ higher (with a margin τ) than its respective $\mathcal{D}_{\text{worse}}$.

The training objective is to minimize the margin-based loss as in Equation 7. We tailored the loss function like this because even though a negative sample is mostly less valid than its respective positive sample, it could be more valid than other positive samples in other pairs.

$$\mathcal{L} = \max\left(0, s_{\text{worse}} - s_{\text{original}} + \tau\right) \tag{7}$$

where τ is the margin hyperparameter, s_{worse} and s_{original} are the two scores for a $(\mathcal{D}_{\text{worse}}, \mathcal{D}_{\text{original}})$ pair by proposed ranking model, respectively. In the training process, we employ the ADAM optimizer with a default learning rate of 0.001. The margin τ is 1.0 as recommended. The training stops when the training loss converges. For consistency, the stop criterion is that the training loss first drops below 0.3 for all datasets and settings.

IV. Experiments

In this section, we evaluate the proposed HG ranking model to answer three questions: 1) can it address both simple and complex hypotheses (Sec. IV-B)? 2) can it boost the performance of current approaches, especially embedding-based ones (Sec. IV-C)? 3) how its performance relies on its parameters choices (Sec. IV-F) and is graph neural network necessary? Before experiments, we will introduce the datasets and preprocessing. The source code is publicly available at https://github.com/JunchengDing/HGRanking.

A. Datasets and Preprocessing

Following previous work [8], [12], we organize our experiments to evaluate proposed HG ranking model in two groups: the qualitative study (question 1) and the quantitative study (question 2 and 3). The qualitative group investigates the most recognized pair "Fish Oil" and "Raynaud Disease" by checking whether the top-ranked simple and complex hypotheses are meaningful. The quantitative group follows the standard evaluation that cuts off PubMed into two parts via a cutoff date, ranks hypotheses according to the pre-cutoff data, and evaluates how the ranking is consistent with that from the post-cutoff data ("future" observation) [12]. We use the five standard "golden" datasets listed below for comparisons between different approaches. Each dataset cuts the PubMed into two parts: a) articles before the cutoff date to find and rank hypotheses; and b) articles after the cutoff date as the ground truth. To ensure uniformity, we run proposed HG ranking model with the same setting on these datasets and probe for the results.

1) Fish Oil (FO) & Raynaud Disease (RD) (1985)
2) Magnesium (MG) & Migraine Disorder (MIG) (1988)
3) Somatomedin C (IGF1) & Arginine (ARG) (1994)
4) Indomethacin (INN) & Alzheimer Disease (AD) (1989)
5) Schizophrenia (SZ) & Calcium - Independent Phospholipase A2 (CI, PA2) (1997)

Table I presents the details of the five datasets. In Table I, $N_{\text{doc}}^{\text{before}}$ and $N_{\text{doc}}^{\text{after}}$ are the number of documents before and after the cutoff date. $\overline{N}_{\text{MeSH terms}}^{\text{before}}$ and $\overline{N}_{\text{MeSH terms}}^{\text{after}}$ are the averaged number of MeSH terms in each document before and after the cutoff date.

We select articles to build the training corpus for the proposed HG ranking model. The process preserves articles published before the cutoff date and containing either one of the input terms. It can significantly ease model training without limiting performance. Besides, TF-IDF, the standard approach to distinguish between documents, is added as a baseline in the experiments to verify that the corpus building does not introduce bias.

The following experiments generate candidate hypotheses as "{input term 1, *B*, input term 2}" before comparing and ranking them. The candidate hypotheses generation process selects a limited number of MeSH terms, rather than using all in the vocabulary, as *B* terms. The plausible *B* terms are terms co-occurring with either one of the input terms in the pre-cutoff documents.

B. Qualitative Evaluation

This experiment aims to validate whether proposed HG ranking model can handle both simple and complex hypotheses. Specifically, we use proposed HG ranking model to find the top-ranked hypotheses with one to three intermediate concepts (MeSH terms) linking the "Fish Oil" and "Raynaud Disease" pair and evaluate whether they are meaningful. Figure 3 shows the top-ranked three hypotheses of each category and their respective evidence (the PMID of the paper that contains the association). We observe from Fig. 3 that all the top-ranked hypotheses are valid with direct evidence from PubMed. The high-ranking "Cryoglobulins" rarely appear in previous literature, but it is meaningful because "Fish Oil" can treat "Cryoglobulins" (PMID: 7842531) which is associated with "Raynaud Disease" (PMID: 11455056). The remaining concepts are well recognized evidence [22].

We can also note from Fig. 3 that the top-ranked complex hypotheses have higher meaningfulness scores than simple ones. The reason is that these more complex hypotheses are combinations of simpler valid ones, and they are thus more meaningful than its components. Previous approaches cannot score such complicated hypotheses properly. The observation demonstrates proposed HG ranking model's ability to address both simple and complex hypotheses, which solves the challenges 1 in this paper. Besides, one may wonder if proposed HG ranking model favors (scores higher) more

TABLE I: The details of the five datasets.

	FO-RD	MIG-MG	INN-AD	IGF-ARG	SZ-CI,PA2
$N_{\text{doc}}^{\text{before}}$	7,048,184	8,092,246	10,447,830	8,473,640	11,718,234
$N_{\text{doc}}^{\text{after}}$	17,398,057	16,353,995	13,998,411	15,972,601	12,728,007
$\overline{N}_{\text{MeSH terms}}^{\text{before}}$	8.20	8.37	8.84	8.43	9.12
$\overline{N}_{\text{MeSH terms}}^{\text{after}}$	11.48	11.60	11.79	11.65	11.83

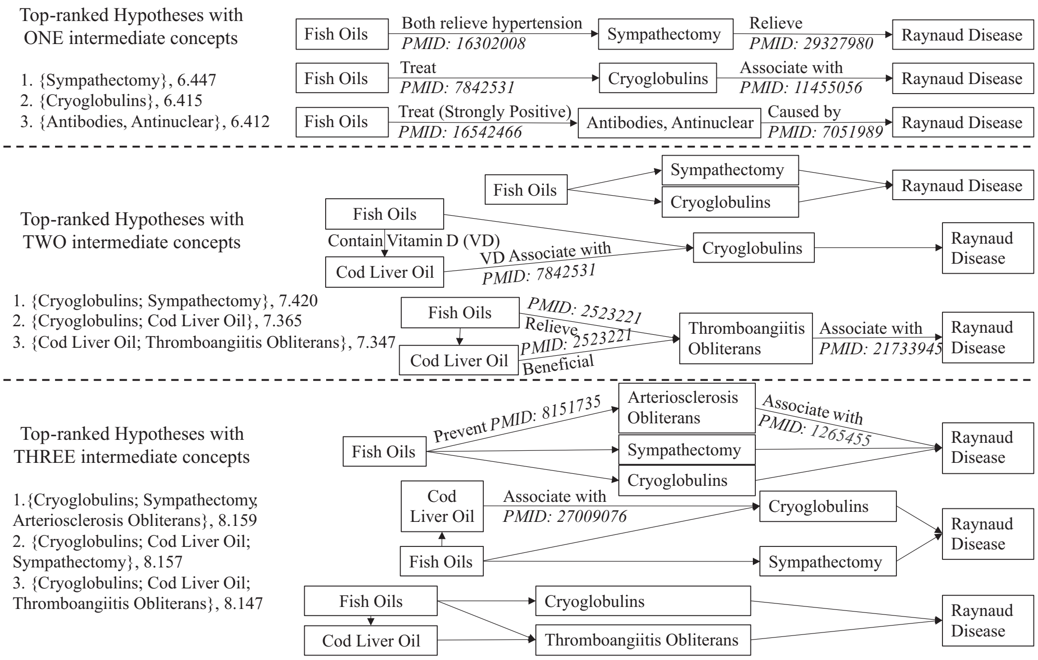


Fig. 3: Top-Ranked Hypotheses with one, two, and three intermediate concepts on "Fish Oil" and "Raynaud Diesease" and their scores by proposed model. The left are the top-ranked hypotheses (omitting "Fish Oil" and "Raynaud Diesease" for brevity) with their scores. The right are their respective evidence with our manually found ground truth.

complex hypotheses. To answer this question, we fed 5,000 randomly generated hypotheses with one, two, and three intermediate concepts into proposed HG ranking model. The averaged scores with one standard deviation are 4.387 ± 0.847, 2.900 ± 0.825, and 1.900 ± 0.701 respectively, which reveals that proposed HG ranking model does not favor complex hypotheses. The observation can be explained by the fact that a smaller portion of complex hypotheses is meaningful than that of simple ones. Moreover, the concepts (i.e., platelet aggregation, blood viscosity, vasoconstriction) identified by the pioneers [5] as meaningful evidence are all in our top-20 hypotheses (18^{th}, 9^{th}, 11^{th} respectively) with one intermediate concept, verifying the results in another aspect. To summarize, this qualitative study shows that proposed HG ranking model can address both simple and complex hypotheses.

C. Quantitative Evaluation

In this experiment, we rank hypotheses using different approaches and compare the rankings' consistency with the ground truth ranking quantitatively. Specifically, the evaluation cuts PubMed into two parts via a cutoff date and evaluate a hypothesis by checking the number of documents mentioning it after cutoff. Given that a hypothesis $\{M_{\text{A}}, M_{\text{B}}, M_{\text{C}}\}$ is more meaningful if more documents discuss it in the future, we define plausible hypotheses' gt($\{M_{\text{A}}, M_{\text{B}}, M_{\text{C}}\}$) scores and use the score to rank them as the ground truth ranking. Specifically, the score is defined as gt($\{M_{\text{A}}, M_{\text{B}}, M_{\text{C}}\}$) = #($M_{\text{A}}$,$M_{\text{B}}$)+#($M_{\text{B}}$,$M_{\text{C}}$), where M_{A} and M_{C} are the two input MeSH terms, M_{B} is the intermediate term in the hypothesis, and $\#(M_i, M_j)$ is the number of documents containing both M_i and M_j in the post-cutoff data. Higher "gt()" scores indicate that a larger number of documents discussing the hypotheses in the "future" and the respective hypotheses are therefore more worth investigation (semantically meaningful) "now" at the cutoff year.

D. Baseline Algorithms

- *TF-IDF* [10], [11]: TF-IDF ranks candidate hypotheses using the intermediate terms' TF-IDF in the training data.
- *BITOLA* [23]: BITOLA is a popular HG system that generates hypotheses based on semantics information and is a graph-based approach.
- *Jaccard* and *Preferential attachment (PA)* [10]: Jaccard and PA are two commonly used and recent link prediction

TABLE II: Spearman's Correlation at k (FO-RD)

Algorithm	k=300	k=600	k=900	k=1200
TF-IDF	-0.041	-0.177	-0.338	-0.421
BITOLA	-0.097	0.039	0.131	0.145
Jaccard	0.101	0.011	0.035	-0.006
PA	0.129	0.065	-0.029	-0.074
Emb&Sim	0.234	0.252	0.253	0.163
BioBERT	0.013	0.051	-0.018	-0.002
PubMedBERT	0.031	0.020	0.060	0.060
Proposed	**0.484**	**0.477**	**0.433**	**0.347**

TABLE III: Spearman's Correlation at k (MIG-MG)

Algorithm	k=300	k=600	k=900	k=1200
TF-IDF	-0.183	-0.206	-0.287	-0.299
BITOLA	-0.304	-0.025	0.000	0.080
Jaccard	0.068	0.020	-0.002	0.022
PA	0.045	-0.025	-0.076	-0.130
Emb&Sim	0.350	0.288	0.313	0.326
BioBERT	0.019	0.001	-0.024	-0.025
PubMedBERT	0.174	0.184	0.152	0.165
Proposed	**0.474**	**0.512**	**0.488**	**0.529**

TABLE IV: Spearman's Correlation at k (INN-AD)

Algorithm	k=300	k=600	k=900	k=1200
TF-IDF	-0.284	-0.286	-0.292	-0.363
BITOLA	-0.056	-0.030	-0.217	-0.135
Jaccard	0.108	0.046	-0.028	-0.050
PA	-0.033	-0.002	-0.003	0.005
Emb&Sim	0.251	0.162	0.183	0.119
BioBERT	0.034	0.054	0.035	0.017
PubMedBERT	0.222	0.213	0.146	0.179
Proposed	**0.442**	**0.436**	**0.469**	**0.504**

TABLE V: Spearman's Correlation at k (IGF-ARG)

Algorithm	k=300	k=600	k=900	k=1200
TF-IDF	-0.166	-0.268	-0.317	-0.363
BITOLA	-0.028	-0.058	0.126	0.289
Jaccard	-0.018	0.055	-0.013	-0.017
PA	-0.071	0.015	0.044	0.054
Emb&Sim	0.337	0.341	0.269	0.271
BioBERT	0.070	0.036	0.065	0.076
PubMedBERT	0.177	0.210	0.240	0.230
Proposed	**0.367**	**0.370**	**0.431**	**0.482**

techniques in graph-based approaches of HG.

- *Embedding & Similarity Measure (Emb&Sim)* [8]: Embedding-based approaches have achieved the SOTA performance in HG research. To make a fair comparison, this baseline uses CBOW [24] to learn the MeSH term embeddings and the embedding's cosine similarity to score the hypotheses. The parameters are identical to those in proposed model.
- *BioBERT* [16]: BioBERT is a pretrained language model trained on PubMed abstracts and full-text PMC articles.
- *PubMedBERT* [25]: PubMedBERT is another pretrained biomedical language model trained on PubMed abstracts. Similar to embedding-based approaches language models such as these have achieved the SOTA performance in HG research.

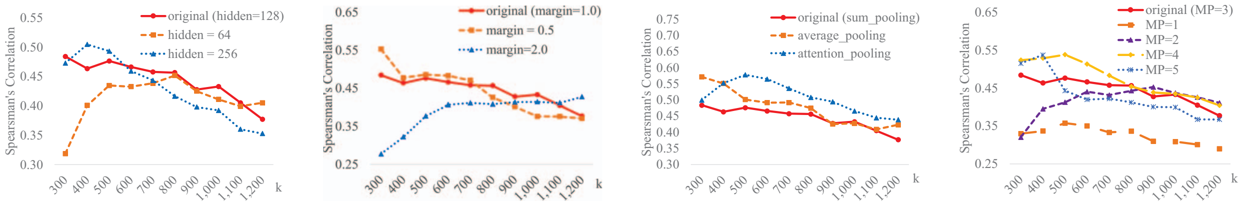
Notably, we are not comparing with the supervised machine learning approaches [13] because our proposed proposed HG ranking model is a self-supervised learning solution in terms of the training data used.

E. Results

We compare each ranking with its ground truth ranking using Spearman's correlation scores at k, and employ different k to make reliable comparisons. Tables II, III, IV, V, VI list the scores on the five datasets and by different baselines, and in those tables, our proposed model is using the default hyperparameters as in Section III-B. Higher scores mean better rankings, and bold numbers mark the group's best performance. Tables II, III, IV, V, VI show that all scores are above -0.5, showing all approaches are producing meaningful hypotheses. The scores by TF-IDF are low, confirming that our data preparation does not introduce bias. BITOLA, Jaccard, and PA perform better than TF-IDF but show no significantly different performance between each other. The result is because their predefined metrics work for different single aspects (e.g., statistics, semantics) of data. Emb&Sim can capture latent semantics as well as statistics and thus always perform better than the previous three. Our proposed model achieves higher scores than Emb&Sim. The reason is that our unified model can ensure both optimal embeddings and scoring mechanism in contrast to Emb&Sim that learn the embeddings and devise the scoring independently (i.e., Emb&Sim's training process is not optimizing the scoring mechanism). Similarly, the proposed model also outperforms models such as BioBERT [16] and PubMedBERT [25]. The reason for this result is that these language models take MeSH terms as a string and tokenize each of them into multiple tokens, and such split will may lead to loss of information that could be crucial to HG. To conclude, proposed HG ranking model outperforms the baselines on all five datasets, verifying our proposed unified model can boost performance. Moreover, we can see the scores' tendencies as k increases are different by different approaches in Tables II, III, IV, V, VI, e.g., TF-IDF scores always decrease while BITOLA scores always grow when k increases. The observation indicates that the rankings by different approaches perform differently with different measurements (k) and verifies that it is necessary to compare different rankings using different ks in Spearman's Correlation. Our proposed proposed HG ranking model outperforms the baselines in all settings, showing its reliability.

F. Parameters Sensitivity

This experiment evaluates how proposed HG ranking model's performance relies on its parameters, i.e., whether the improvement in performance comes from the structure or the parameter selection. We evaluate proposed HG ranking model with different parameters using the "Fish Oil" and "Raynaud Disease" pair. Specifically, we evaluate the latent dimensions [64, 256], margins [0.5, 2.0], aggregation functions [$average\_pooling$, $attention\_pooling$], and numbers of message-passing [1, 2, 4, 5] other than the parameters in our proposed HG ranking model model. The respective scores are in Figs. 4a, 4b, 4c, 4d. These scores show that proposed HG ranking model performs stable (and importantly, always better



(a) hidden size (b) margin τ (c) aggregation function (d) no. of Message-Passing

Fig. 4: HG ranking model's performance given different settings. They all outperform baselines in Table II.

TABLE VI: Spearman's Correlation at k (SZ-CI,PA2)

Algorithm	k=300	k=600	k=900	k=1200
TF-IDF	-0.091	-0.203	-0.228	-0.284
BITOLA	-0.218	-0.211	-0.035	0.002
Jaccard	0.092	0.014	0.031	-0.009
PA	0.048	0.126	0.074	0.106
Emb&Sim	0.139	0.057	0.076	0.104
BioBERT	0.066	0.046	0.0723	0.061
PubMedBERT	0.001	0.124	0.112	0.133
Proposed	**0.407**	**0.444**	**0.552**	**0.555**

than the five baselines as in Table II) with different parameters. Moreover, the performance of proposed HG ranking model decreases greatly when the number of message-passing is 1. The reason is that proposed HG ranking model becomes a simple fully-connected neural network when the number of message-passing is 1 (there is no message-passing). The result shows the *necessity of message-passing (or adapting graph neural network) in proposed HG ranking model*. The observations are similar in the other four datasets, so we omit them for brevity. To conclude, the message-passing mechanism in proposed HG ranking model is necessary, and parameter choices impact little on proposed HG ranking model's performance. Namely, it is the structure of proposed HG ranking model that leads to its superior performance.

V. Related Work

Hypothesis generation is a core task in biomedical text mining [11], [26]–[29] with applications to a variety of tasks such as drug-repurposing, precision medicine, and biomarker discovery. For a recent survey on this topic, please refer [30]. The initial approaches [8], [11], [12], [31] generated hypothesis by using standard information retrieval metrics such as term-frequency, document frequency, and term-frequency-inverse document frequency [5], [32]–[37]. While these purely statistics-based approaches made great advances, their reliance on corpus based co-occurrence information affected performance for concepts that had a paucity of training data. To overcome these issues, subsequent studies employed neural network inspired word embedding models [38], [39] that are better able to quantify the strength of associations between implicit links. These embedding-based approaches use the concept embedding similarity to evaluate the hypotheses meaningfulness. Compared to purely statistical approaches, the embedding based approaches have obtained significant improvement in results [11]. Approaches such as [35] proposed to build complex graphs that facilitate identification of both simple and complex links that are of practical value. However, these approaches rely on the graph's pre-defined schema and cannot find links that have strong implicit association but are not directly connected in the graph. Moreover, these approaches learn the embeddings and evaluate the links in two independent steps. The independence leads to sub-optimal modules in both steps and compromise their performance. We solve these problems by modeling the links with variable numbers of concepts as term set and building a unified model (end-to-end gradient-based training) to score the set directly. More recently, studies such as [9], [31] use supervised machine learning to identify meaningful hypotheses. While these approaches have shown promising results, they require a comprehensive expert-labeled training set that is too costly to build.

Our work is motivated from the recent developments made in the research areas of Graph Neural Network (GNN) and Noise-Contrastive Estimation (NCE). Specifically, the approach proposes a HG tailored GNN [19], that can take graphs with variable numbers of nodes as inputs. NCE [21], which estimates model parameters using only positive samples, motivates our self-supervised algorithm to train the model. NCE generates negative samples from assumed distributions and trains a model by contrasting the existing positive and the generated negative samples. Our proposed approach also uses only positive samples (existing literature), but it differs from NCE because we create negative samples based on the nature of biomedical literature other than assumed distributions.

VI. Conclusion

This paper proposes a new HG ranking model to compare and rank both simple and complex hypothesis. To the best of our knowledge, it is among the first works that focuses on the ranking on hypotheses rather than their enumeration. One unique aspect of proposed approach is to train the designed algorithms using only the existing documents, overcoming the lack of labeled hypotheses. We conducted extensive experiments to justify those advantages. In the future, we will explore more advanced models and to incorporate more domain knowledge into our model to further improve the performance.

VII. Acknowledgment

The authors thank anonymous reviewers for their valuable comments and helpful suggestions. K. Jha's work is supported in part by the US National Institute of Health (NIH) and National Science Foundation (NSF) under grants R01LM014012-01A1 and ITE-2333740. B. Adhikari's work is supported in part by NSF OAC 2320980 and CDC MInD Healthcare group under cooperative agreement U01-CK000594 and its associated COVID-19 supplemental funding. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of funding agencies.

References

[1] A. F. Syafiandini, G. Song, Y. Ahn, H. Kim, and M. Song, "An automatic hypothesis generation for plausible linkage between xanthium and diabetes," *Scientific Reports*, vol. 12, no. 1, p. 17547, 2022.

[2] Z. Lu, "Pubmed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, p. baq036, 2011.

[3] S. Spangler, A. D. Wilkins, B. J. Bachman, M. Nagarajan, T. Dayaram, P. Haas, S. Regenbogen, C. R. Pickering, A. Comer, J. N. Myers *et al.*, "Automated hypothesis generation based on mining scientific literature," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1877–1886.

[4] S. Zhao, C. Su, Z. Lu, and F. Wang, "Recent advances in biomedical literature mining," *Briefings in Bioinformatics*, vol. 22, no. 3, p. bbaa057, 05 2020. [Online]. Available: https://doi.org/10.1093/bib/bbaa057

[5] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspect Biol Med.*, vol. 30, no. 1, pp. 7–18, 1986.

[6] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: statistical mechanics and its applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[7] K. Jha, G. Xun, Y. Wang, V. Gopalakrishnan, and A. Zhang, "Concepts-bridges: Uncovering conceptual bridges based on biomedical concept evolution," in *KDD'18*, 2018, pp. 1599–1607.

[8] K. Jha, G. Xun, Y. Wang, and A. Zhang, "Hypothesis generation from text based on co-evolution of biomedical concepts," in *KDD'19*, 2019, pp. 843–851.

[9] J. Sybrandt, I. Tyagin, M. Shtutman, and I. Safro, "AGATHA: Automatic graph mining and transformer based hypothesis generation approach," in *CIKM'20*, 2020, pp. 2757–2764.

[10] S. Pyysalo, S. Baker, I. Ali, S. Haselwimmer, T. Shah, A. Young, Y. Guo, J. Högberg, U. Stenius, M. Narita *et al.*, "Lion lbd: a literature-based discovery system for cancer biology," *Bioinformatics*, vol. 35, no. 9, pp. 1553–1561, 2019.

[11] P. Srinivasan, "Text mining: generating hypotheses from MEDLINE," *J. Assoc. Inf. Sci. Technol*, vol. 55, no. 5, pp. 396–413, 2004.

[12] M. Yetisgen-Yildiz and W. Pratt, "A new evaluation methodology for literature-based discovery systems," *J Biomed Inform.*, vol. 42, no. 4, pp. 633–643, 2009.

[13] S. Sang, Z. Yang, Z. Li, and H. Lin, "Supervised learning based hypothesis generation from biomedical literature," *BioMed research international*, vol. 2015, 2015.

[14] J. Sybrandt, M. Shtutman, and I. Safro, "Moliere: Automatic biomedical hypothesis generation system," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1633–1642.

[15] S. Zhuang, H. Li, and G. Zuccon, "Deep query likelihood model for information retrieval," in *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 2021, pp. 463–470.

[16] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[17] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "Biowordvec, improving biomedical word embeddings with subword information and mesh," *Scientific data*, vol. 6, no. 1, p. 52, 2019.

[18] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo, "How to train good word embeddings for biomedical nlp," in *Proceedings of the 15th workshop on biomedical natural language processing*, 2016, pp. 166–174.

[19] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst*, 2020.

[20] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep Sets," in *NeurIPS'17*, 2017, pp. 3391–3401.

[21] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *AISTATS'10*, 2010, pp. 297–304.

[22] D. R. Swanson, N. R. Smalheiser, and V. I. Torvik, "Ranking indirect connections in literature-based discovery: The role of medical subject headings," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1427–1439, 2006.

[23] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin, "Exploiting semantic relations for literature-based discovery," in *AMIA'06*, 2006, pp. 349–353.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[25] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[26] M. Weeber, H. Klein, L. de Jong-van den Berg, R. Vos *et al.*, "Using concepts in literature-based discovery: Simulating swanson's Raynaud–Fish oil and Migraine–magnesium discoveries," *J. Assoc. Inf. Sci. Technol.*, vol. 52, no. 7, pp. 548–57, 2001.

[27] W. Pratt and M. Yetisgen-Yildiz, "Litlinker: capturing connections across the biomedical literature," in *Proceedings of the 2nd international conference on Knowledge capture*, 2003, pp. 105–12.

[28] K. Jha and A. Zhang, "Continual knowledge infusion into pre-trained biomedical language models," *Bioinformatics*, vol. 38, no. 2, pp. 494–502, 2022.

[29] K. Jha, G. Xun, V. Gopalakrishnan, and A. Zhang, "Dwe-med: Dynamic word embeddings for medical domain," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 13, no. 2, pp. 1–21, 2019.

[30] V. Gopalakrishnan, K. Jha, W. Jin, and A. Zhang, "A survey on literature based discovery approaches in biomedical domain," *Journal of biomedical informatics*, vol. 93, p. 103141, 2019.

[31] U. Akujuobi, M. Spranger, S. K. Palaniappan, and X. Zhang, "T-PAIR: Temporal node-pair embedding for automatic biomedical hypothesis generation," *IEEE Trans. Knowl Data Eng.*, 2020.

[32] V. Gopalakrishnan, K. Jha, A. Zhang, and W. Jin, "Generating hypothesis: Using global and local features in graph to discover new knowledge from medical literature," in *Proceedings of the 8th International Conference on Bioinformatics and Computational Biology, BICOB 2016*, 2016, pp. 23–30.

[33] D. R. Swanson and N. R. Smalheiser, "An interactive system for finding complementary literatures: a stimulus to scientific discovery," *Artif. Intell.*, vol. 91, no. 2, pp. 183–203, 1997.

[34] C. B. Ahlers, D. Hristovski, H. Kilicoglu, and T. C. Rindflesch, "Using the literature-based discovery paradigm to investigate drug mechanisms," *AMIA Annu Symp Proc*, pp. 6–10, Oct 2007.

[35] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, and O. Bodenreider, "Context-driven automatic subgraph creation for literature-based discovery," *J. Biomed. Inform.*, vol. 54, pp. 141–157, 2015.

[36] G. Xun, K. Jha, V. Gopalakrishnan, Y. Li, and A. Zhang, "Generating medical hypotheses based on evolutionary medical concepts," in *IEEE 17th International Conference on Data Mining, ICDM 2017, December 18-21, 2017, New Orleans, USA*, 2017.

[37] R. K. Lindsay and M. D. Gordon, "Literature-based discovery by lexical statistics," *Journal of the Association for Information Science and Technology*, vol. 50, no. 7, p. 574, 1999.

[38] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[39] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.