Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 2, 2025 (New Series), pp 267–290
http://www.ssca.org.in/journal

A Framework for Building a Novel Causal Proximity Driven GNN from Biomedical Text

Samridhi Dev and Aditi Sharan

School of Computer and Systems Sciences
Jawaharal Nehru University, New Delhi 110067

Received: 11 February 2025; Revised: 25 March 2025; Accepted: 02 April 2025

Abstract

In the digital era, the abundance of biomedical text, along with the advanced Graph Neural Network (GNN) based algorithm, provides a forum for unrevealing significant information hidden in the text. In this context, researchers have tried to predict Adverse Drug Reactions (ADRs) in cancer treatments using GNN. However, ADR prediction in cancer is still a challenging research problem. Current GNN based methodologies face critical limitations, including over-smoothing, equal weighting of neighboring nodes, and inability to capture intersentential, causal, and high-granular indirect information. Additionally, the field is hindered by limited annotated data and the absence of automated methods for corpus construction, making it challenging to scale the models for effective ADR prediction. To overcome these barriers, we propose a two-model framework. The first is a linguistics-driven, rule-based approach designed to automate corpus construction and causal relation extraction process, thereby generating a cancer-specific ADR corpus with annotated entity types and causal relationships. This automated corpus construction leverages sophisticated linguistic rules to ensure accurate and consistent causality annotation, significantly enhancing available resources. Building on this, the second neural network-based model, termed the Causality and Proximity-based Relational Multi-head Attention Model (CPRMAM), is trained on the constructed corpus, integrating both causal semantics and proximity considerations. Together, these models address key limitations in existing GNN frameworks, advancing ADR prediction by enabling the extraction of complex causal structures and enhancing interpretability within oncological contexts.

Key words: Graph neural network; Adverse drug reactions; Aggregation function; CPRMAM; Knowledge graph completion; Data mining; Healthcare; Neural networks.

1. Introduction

With the digitization of health data, a significant amount of textual information is being generated from biomedical journals, research articles, and case reports. However, much

Corresponding Author: Aditi Sharan
Email: aditisharan@gmail.com

of this data remains untapped due to its unstructured nature. Graph neural network based models offer a systematic approach to extract valuable insights from this information, transforming unstructured medical/biomedical text into structured graphs, leading to meaningful findings. These models assist in maximizing the potential of previously underutilized medical data. By improving data-driven inferences, the model assist healthcare professionals in making informed clinical decisions.

Data-driven graph neural networks have been studied widely to infer meaningful insights for complicated diseases such as cancer. Cancer is a complicated disease mainly because it is caused by changes in genes that control functioning of multiple cells and involve complex progression pathways. Inspite of a lot of progress in detecting and treating cancer, cancer-related Adverse Drug Reactions (ADRs) remain a significant challenge in healthcare and biomedical research. To address this, we have chosen to construct a cancer-related corpus specifically for ADR prediction, enabling a more structured and data-driven approach to analyze adverse drug effects in cancer treatment. Effective extraction of cancer-related entities and relations from textual data is crucial for understanding the complexities of drug interactions in oncology. Considering various types of relations that can be extracted from textual data, causal relations extraction is an open research problem. However, the success of such extraction processes hinges on accurate and comprehensive annotation of biomedical data, which is inherently challenging due to its complexity, terminologies, and ambiguity. In the context of ADR prediction, earlier deep learning and machine learning models have shown limitations, especially in capturing complex, intersentential, and indirect relationships within biomedical knowledge graphs. Graph Neural Networks (GNNs) have a better potential of capturing these complicated relations.

This paper proposes a framework to build causal proximity based GNN. The study introduces two models; the first model is a novel causal linguistic driven framework for complex entity and causal relation annotation, leveraging linguistic and grammatical properties alongside a rule-based approach. The framework consists of key subtasks: entity extraction, coreference resolution, rule generation, relation extraction, and the formation of triplets in a novel format that surpasses the limitations of state of art triplet representations. By combining semantic and linguistic dependencies, this approach ensures robust and efficient relation extraction in biomedical texts.

Our second proposed model is the Causality and Proximity-based Relational Multi-head Attention Model (CPRMAM). It is trained on a cancer-specific ADR knowledge graph constructed using the first model. The model integrates causal reasoning and proximity for better knowledge discovery, addressing critical gaps in current methods and improving the robustness and accuracy of predictions. The paper is structured as follows: it begins with the automatic construction of the required corpus, where entities and causal relations are systematically annotated. Afterward, we design and implement the GNN model, leveraging the constructed dataset to predict ADRs in oncology.

2. Related work

This section briefly reviews and summarizes the recent studies related to our work. Adverse drug reaction prediction has been approached using various methodologies, encompassing statistical, machine learning (ML), deep learning (DL), graph-based, and similarity-

based techniques. In statistical-based approaches, researchers have developed drug-pair protein interaction profiles using data from the STITCH and TWOSIDES databases, employing the Laplacian-corrected estimator to predict drug-induced effects [Pauwels *et al.* (2011)]. Another group of researchers modeled ADR-drug relationships with a three-layer hierarchical Bayesian model [Bate *et al.* (1998)], utilizing latent Dirichlet allocation to uncover biochemical mechanisms linking ADRs to drug structures[Liu *et al.* (2017)]. Additional statistical methods include using proportional reporting ratio, reporting odds ratio, and empirical Bayes geometric mean algorithms to identify drug-associated adverse event [Kwak *et al.* (2020)], and extracting biomedical knowledge via MetaMap and SemRep to support reasoning about drug-effect relationships [Yildirim *et al.* (2014)]. Moreover, researchers collected data from seven databases and implemented methods such as Bayesian confidence propagation neural network, gamma Poisson shrinker, proportional reporting ratio, and reporting odds ratio to detect ADRs [Szarfman *et al.* (2002) and Schuemie *et al.* (2012)]. Extended LRT methods based on Poisson models were used to identify ADR signals with disproportionately high reporting rates [Zhao *et al.* (2018)], while other studies leveraged EudraVigilance data [Monaco *et al.* (2017)], tree-based scan statistics [Wang *et al.* (2018)], and disproportionate methods to detect unknown causal associations [Lerch *et al.* (2015)]. In the realm of ML/DL-based approaches, researchers focused on integrating data from DrugBank, DrugCentral, CTD, and TWOSIDES databases, proposing the HCNS-ADR machine learning method to predict ADRs from combined medication [Xiao *et al.* (2017)]. Another team developed a structure-enhanced line graph convolutional network to learn comprehensive representations of drug-disease pairs, transforming the task into a node classification problem using the SEAL architecture for link prediction [Zheng *et al.* (2018) and Shang *et al.* (2014)]. A convolutional framework was also proposed to construct chemical fingerprint features and assess their associations with ADRs [Mantripragada *et al.* (2021)]. Graph-based approaches have seen significant advancements, with authors proposing a GNN model trained on clinical data using SIDER database labels to predict side effect signals between drug-disease pairs [Gao *et al.* (2023)]. Another innovative method, the contextualized graph embedding model (CGEM), captures cause-effect relations for ADR detection by combining contextualized embeddings, convolutional GNNs, and BertGCN for classification [Liu *et al.* (2024) and Dey *et al.* (2018)]. Additional graph-based techniques include analyzing non-directional heterogeneous healthcare data for triad prediction [Zhang *et al.* (2021)], combining deep learning with a biomedical tripartite network to predict drug-ADR associations [Xue *et al.* (2020)], and inferring new associations through drug-ADR network topological features [Manzi and Reis (2011)]. An external link concept was also proposed to infer associations in a heterogeneous network by connecting drugs sharing common ADRs [Ji *et al.* (2011) and Lin *et al.* (2013)]. Similarity-based approaches involve predicting drug side effects by combining canonical correlation analysis with network-based diffusion [Atias and Sharan (2011)], identifying significant correlations between chemical fragments and side effects for ADR prediction [Pauwels *et al.* (2011)], and using a naive Bayesian model to infer drug-ADR associations based on known drug-protein and drug-ADR associations [Xiang *et al.* (2015)]. These diverse methodologies illustrate the multifaceted nature of ADR prediction, leveraging a broad spectrum of data sources and analytical techniques to enhance the safety and efficacy of pharmaceutical treatments.

3. Proposed work

We propose two models; the first model is to construct an automatically annotated relational corpus by extracting entities of interest and the causal relationship between them. In this model, we propose an algorithm and an extended version of the triplet structure. The second proposed model is a causal proximity based multi-head attention relational graph neural network that enhances biomedical knowledge discovery. This model is trained on the constructed corpus. The proposed work section is divided into three parts: the construction of the corpus, the development of the GNN-based CPRMAM knowledge discovery model, and the training of the CPRMAM model with the corresponding joint probability density function.

3.1. Automatic annotation model

This sub-section provides a brief overview of the workflow illustrated in Figure 1, that outlines the process for automatic construction of the relational corpus and the steps taken to annotate the same. The proposed algorithm begins by focusing on coreference resolution within raw textual data derived from case reports that have been pre-processed. Once coreference resolution is achieved, entities are carefully extracted and annotated, followed by a normalization process.

Using the extracted entities along with grammatical dependencies and linguistic properties, rules for identifying complex and causal relationships are systematically developed. Subsequently, these relationships are extracted and organized into a novel triplet structure based on the established rules. The generation of novel triplets involves assigning relevant properties to the entities and establishing connections that link triplets that occur in close contextual proximity.

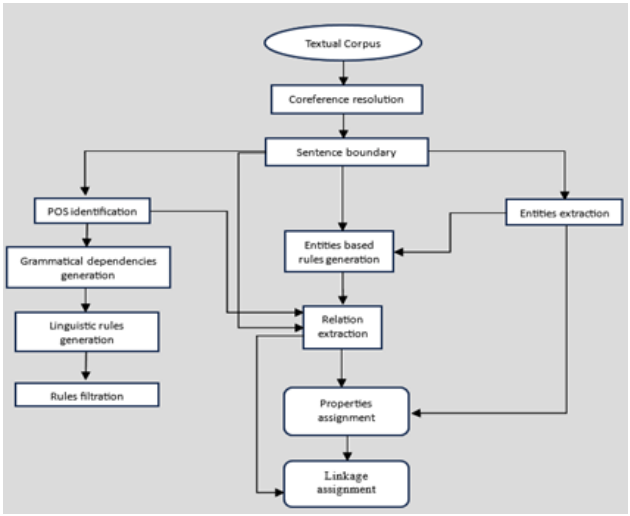

Figure 1: Workflow of first model

3.1.1. Raw corpus collection

Case reports, acknowledged for their ability to capture realistic and practical knowledge, serve as the primary data source for the constructed corpus. The curation process

involved the selection and gathering of case reports from PubMed by executing the query, with the overarching objective of semantically capturing pertinent and realistic medical data [Dev and Sharan (2023b)]. PubMed executed query is responsible for collecting only relevant case reports (cancer specific case reports).

3.1.2. Coreference resolution

The objective of Coreference Resolution is to effectively cluster expressions denoting the same entity within a document, thereby mitigating textual ambiguity. To achieve this, SpanBERT, a self-supervised method meticulously designed to optimize the prediction of text spans, is employed for coreference resolution. SpanBERT adopts a training approach where a single contiguous segment of text is sampled for each training example. Figure 3 elucidates the architectural representation of SpanBERT. Figure 2a and 2b exemplify an instance of both raw and coreferenced corpus, explaining the imperative need for such a resolution in comprehensive linguistic analysis. Figures 2c and 2d, respectively, depict the extracted pattern with and without including the SpanBERT algorithm. In the absence of coreference resolution using SpanBE, the extraction of relations from a paragraph would follow the pattern depicted in Figure 2c. Highlighting the significance of coreference, once coreference resolution is applied, it becomes evident that without this process, the relationship highlighted in Figure 2d would remain uncaptured. Coreference resolution plays a pivotal role in ensuring complete latent semantics encapsulation.

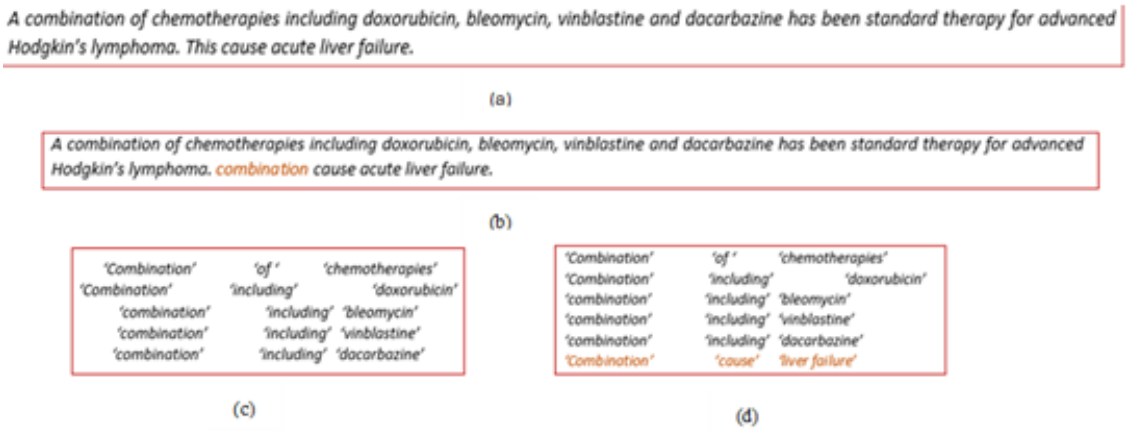

Figure 2: Instance of coreference resolution

3.1.3. Entity extraction

Entity extraction procedure is compartmentalized into two segments: the extraction of nine primary entities and the seventeen secondary entities delineated in Table 1. Primary entities represent the key entities considered as nodes (subjects, or objects) while secondary entities encompass properties assigned to the nodes and relationships. Primary entities include the following categories: drugs, adverse drug reactions, cancer, and dysn. These entities were extracted using METAMAP and the variant settlement algorithm [Dev and Sharan (2023a)]. The extraction of the remaining entities was accomplished using clinical BERT [Alsentzer *et al.* (2019)], a model trained through the integration of three datasets: BC2GM, Jnlpba, and Maccrobat. Earlier, BERT models did not produce the desired results

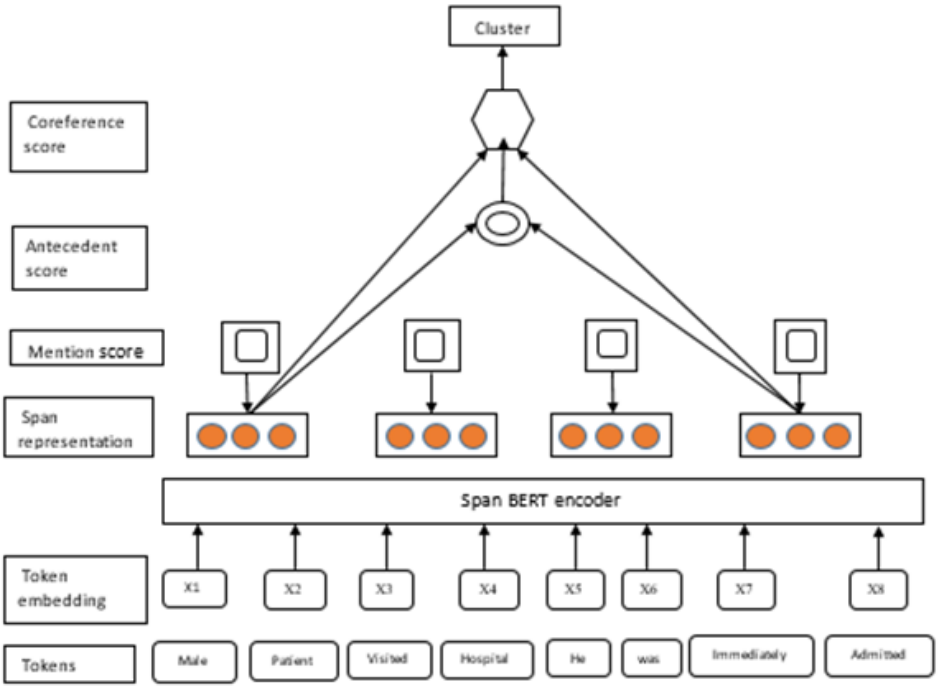

Figure 3: Span BERT architecture

in the medical domain because they were not pre-trained on medical data. This led to the realization that a model needed to be trained on a medical corpus, resulting in the development of medical domain specific BERT models such as MedBERT and ClinicalBERT. Clinical BERT is pre-trained on clinical reports and electronic health records making it suitable for extraction of secondary entities. Clinical BERT enhances the accuracy of secondary entity extraction, as clinical reports contain complex sentences filled with medical jargon and rare terms.

In Clinical BERT, input sentence is represented by vector $X = \{x_1, x_2, \ldots, x_N\}$, where x_i is the i-th word and N represents the length of the sentence. The Clinical BERT model aims to classify word $x_i \in X$ and assign it a corresponding label $y \in Y$ by fine-tuning the pre-trained embeddings. The model is tested using test data. The testing accuracy obtained was 0.951. Following the testing of the model, it is deployed to extract entities from the case report corpus, subsequently mapping them to their respective entity types as illustrated in Table 1. These identified entities are then employed in subsequent tasks related to relation extraction. The clinical BERT architecture is depicted in Figure 4.

3.1.4. Relation extraction

An annotated document is represented as $D = [d_1, ...d_m]$, words set of each document as $W = [w_1, ..., w_n]$, set of tagged entities as E = $[e_1, \ldots .e_n]$ and $R(w_i, w_j)$ describes the relation between pair of entities w_i and w_j. Extraction of relation is grounded in dependency relations and grammatical properties to delineate specific relation patterns. Dependencies between the words in a document required for relation extraction were extracted using BiLSTM based deep bi-affine neural network dependency parser. Resultant relations are subse-

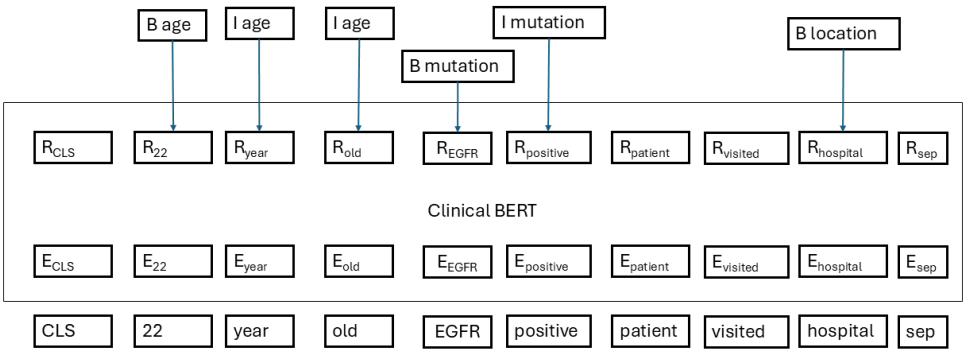

Figure 4: Workflow of first model

Table 1: Primary and secondary entities

Category	Entities
Primary Entities	Drug, Cancer type, Adverse drug reaction, Dysn, Protein, Gene, Gene mutation, Cell type
Secondary Entities	Age, Date, Detailed description, Family history, Dosage, Duration, Height, History, Non-biological location, Lab value, Biological structure, Activity, Personal background, Gender, Severity, Clinical event, Area

quently transformed into triplets, a format conveying semantics in text. It includes subject, object, and predicate. They can be either three words or three phrases. The assignment of triplet elements is contingent on their respective grammatical roles and dependencies.

This baseline triplet structure proves insufficient in capturing the semantic nuances inherent in complex sentences. To address this limitation, we introduce a novel extended triplet structure, as illustrated in Figure 5. This structure builds upon the base triplet format by incorporating properties for each triplet component and proximity links between them. By augmenting triplet elements with properties, we encapsulate detailed and vital information. Concurrently, the inclusion of proximity links establishes connectivity among proximate triplets, facilitating the extraction of comprehensive and meaningful relational data.

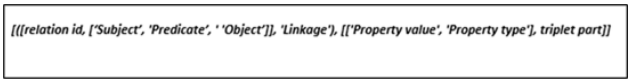

Figure 5: Proposed triplet format

For example, consider a sentence of a case report. '*Severe rashes were caused due to bleomycin during cycle 1 treatment of metastasis breast cancer*'. In this sentence, an extracted relation triplet was

['Rashes', 'caused due to', 'bleomycin'].

But in this triplet format, most of the information is lost, such as:

- Rashes are severe
- Bleomycin was given during breast cancer cycle 1 treatment
- Breast cancer was metastasis

The introduced novel extended triplet format serves as a robust solution to circumvent this constraint. We showcase in the given example how the relational data was extracted and significantly enriched through the utilization of our innovative triplet structure.

(1, ['rashes', 'caused due to', 'bleomycin'], [2, 3], [severe, severity, subject])
(2, ['bleomycin', 'during', 'treatment'], [3], [cycle 2, process, object])
(3, ['treatment', 'of', 'breast cancer'], [Metastasis, characteristic, object])

For instance, within Sentence 1, the individual triplets, each denoted by IDs 1, 2, and 3, may initially appear devoid of substantial significance. However, when considered collectively and augmented with incorporated properties and proximity links, they synergistically yield a profound, comprehensive, and contextually rich array of relational information.

3.1.5. Relation categorization

Relations extracted in the form of triplets can be classified into two distinct types: regular relations and causal relations. Figure 6 visually represents the categorization of the extracted relation types.

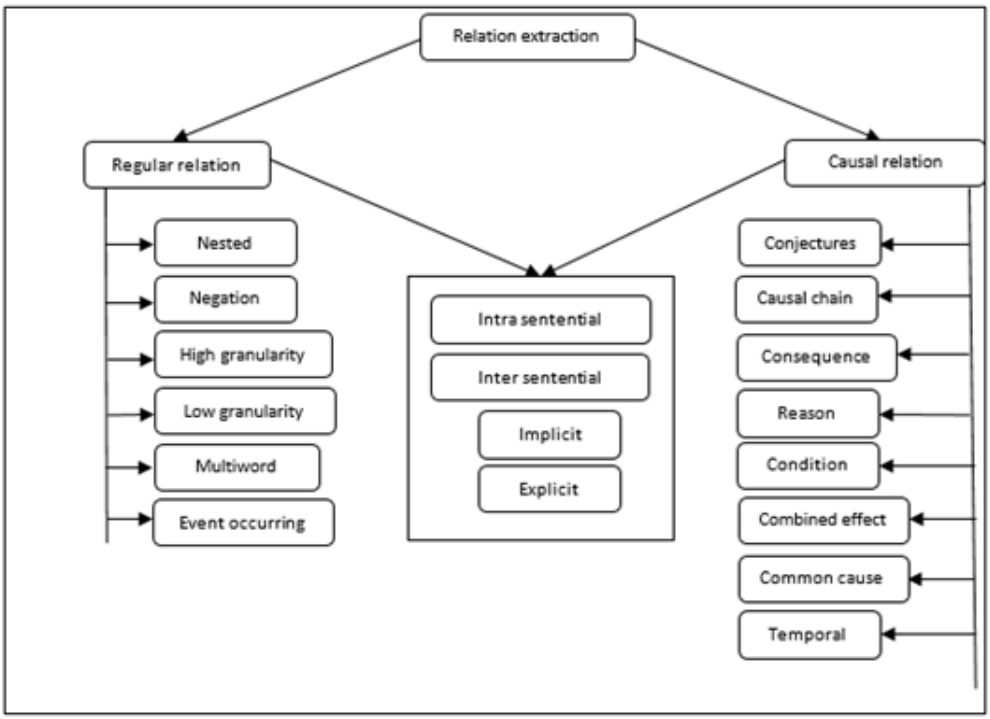

Figure 6: Relations classification

Regular relation: Regular relations are typically based on various semantic associations between entities. These associations can be diversely based on biomedical information.

These relations can be further divided into binary, complex relations, negation, conjectures, implicit relations, explicit relations, event-occurring relations, high granularity relations, nested relations, and multiword relations.

Causal relations: Causal relations are typically more complex and explicitly concerned with cause-and-effect connections and focus on understanding the factors or events that lead to a specific outcome or result. Causality or causation is a concept that expresses how one variable contributes to the creation of another variable [Szarfman *et al.* (2002)]. Causal relations were further classified into intra-sentential relations, inter-sentential relations, causal chains, consequence, reason, condition, common cause, combined effect, common effect, and action. Extracting these causal relations from medical literature is integral for constructing a knowledge graph. Such a graph serves to assist healthcare professionals in swiftly identifying causality patterns, such as diseases causing symptoms, diseases leading to complications, treatments improving conditions, and ultimately tailoring personalized treatment plans.

3.2. CPRMAM model development

This section outlines the development of the second proposed model, the Causal Proximity-based Multihead Attention Relational Graph Neural Network (CPRMAM), that is trained on triplets in the proposed triplet structure extracted by the first model. CPRMAM is designed to predict adverse drug reactions (ADRs) specific to cancer treatment drugs as a case study. To enhance message passing and mitigate over-smoothing, we introduce a novel aggregation function incorporating a proximity vector. Our approach aims to improve GNN performance in heterogeneous graph learning by effectively capturing comprehensive node roles and relationships. Each subsection in this section defines a step in the process, presented sequentially, to provide a clear and structured overview of the model development workflow. The proposed model is elucidated in Figure 7 and an algorithm for the proposed model is described in Algorithm 1.

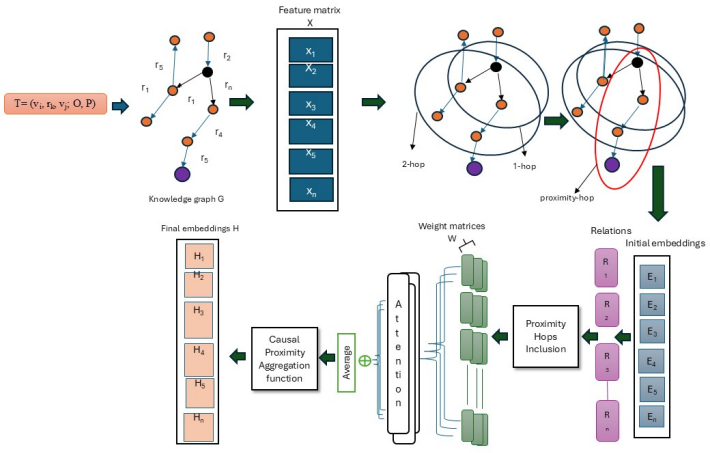

Figure 7: Proposed CPRMAM model

Figure 7 illustrates the Causal Proximity-based Multihead Attention Relational Graph Neural Network (CPRMAM) framework, designed for ADR prediction in oncology by lever-

aging relational information and proximity-aware aggregation. The process begins with a heterogeneous knowledge graph representation of triplets, where nodes (such as drugs, ADRs, and genes) are connected through different relationships. Mathematically, given a set of triplets T, the feature matrix is generated as: $X^{|n|\times d}$ where d is the feature dimension and n is the node size. To enhance information propagation, proximity hops were added. A weight matrix W is applied to encode relational dependencies, followed by the multihead attention mechanism. The final embeddings were generated using the proposed aggregation function.

3.2.1. Initial node embedding generation

The efficacy of graph machine learning tasks, such as link prediction, hinges upon the acquisition of a useful feature representation for nodes within the graph. Conventionally, node embedding generation methods rooted in structural equivalence have predominated in literature. However, these methods suffer from two primary drawbacks. Firstly, they are inherently transductive, thereby impeding their generalization to unseen nodes. Secondly, there exists no standardized approach for seamlessly integrating node attributes into the network representation. Consequently, in these methods, proximity between nodes fails to guarantee semantic similarity.

To circumvent these challenges, we propose an extension to the role2vec [Ahmed *et al.* (2018)] algorithm, designed to learn a function capable of capturing the semantic role and behavior of nodes within a graph. The basic idea of role2vec is to introduce the notion of attributed random walks that are not tied to structural similarity but tied to a function that maps a node to a role in a graph.

The goal of the proposed algorithm is to generate the embedding of a node based on role that is plays in a graph. The role of a node is defined by its features that are described in Table 2. Notion is that the nodes having similar roles must have similar embedding. It includes structural properties as well as semantic properties. Role2vec framework integrates vertex mapping and attributed random walks.

Consider a knowledge graph $G = (V, R, T)$
Where $V =$ *set of entities*, $R =$ *set of relations*, $T =$ *set of triplets in extended format*
Now V_N *and* R_L *can be represented as follows*
$V_N = \{v_1, v_2, v_3, \ldots, v_n\}$
$R_L = \{r_1, r_2, r_3, \ldots, r_l\}$
Now extended triplet t can be represented as follows
Extended triplet $t = \{(v_i, r_k, v_j; O, P); v_i, v_j \in V_N \text{ and } r_k \in R_L\}$
Where O is the element properties set and P is the proximity link set.

$$a_{ijr_k} \in A^{N\times N} \tag{1}$$

Consider $A^{N\times N}$ as the relational adjacency matrix for G, where a_{ijr_k} defines the existence of a relation type r_k between entities v_i and v_j. Let X be the set of node attributes defining the semantic role of each node. Let $F^{N\times P}$ be the feature matrix for the set of

Algorithm 1 CPRMAM Model

Input: Extended triplets $t = (v_i, r_k, v_j; O, P)$ where $v_i, v_j \in V$; $r_k \in R$; O = element properties set and P = proximity link set
Output: Missing link triplets $M = (v_i^m, r_k^m, v_j^m)$

1. Initialize feature set for a node v
2. Generate feature matrix $F^{N \times P}$ for V

$$X_V = \{x_1, x_2, x_3, \ldots, x_{N_x}\}$$

3. Calculate initial node embeddings e for each node v

$$E_d = \{e_1, e_2, e_3, \ldots, e_{N_e}\}$$

4. Set initial relational weight matrix W
5. Initialize Proximity nodes and embedding set Y
6. For each $r_k \in R$ do
   (a) For each $v_n \in V$ do
      i. Initialize causal proximity aggregation function
      ii. Calculate Multi-head attention vector for different heads
      iii. Calculate final embedding for v based on (7) and (8)
7. Initialize Decoder (scoring function)
8. Initialize final embeddings for V
9. Generate negative embeddings for V using Negative sampling
10. Compute triplet loss and update weight W
11. Train relational model
12. For $t \in T$ do
   (a) Generate corrupted triple set T^c
   (b) Initialize final embeddings for V
   (c) Initialize scoring function $f_{r_k}(h_i, h_j)$
   (d) Calculate scores for testing triplet t and T^c
   (e) Generate rank for t against T^c
13. Evaluate model
14. Predict missing links

Table 2: Node semantic features

Feature	Description
Entity_name	Name of the entity
Entity_class	Class or type of the entity
Entity_CUI	Concept Unique Identifier (CUI) of the entity
Entity_property	Properties associated with the entity
Entity_Synonyms	Synonyms of the entity
Entity_normalized version	Normalized version of the entity

entities having N entities and P features. Additionally, e_{id} denotes the initial embedding of a node v_i, generated by the $\phi(v_i)$ function using the extended Role2Vec algorithm.

$$E_d = \{e_1, e_2, e_3, \ldots, e_n\} \quad \text{and} \quad \{x_1, x_2, x_3, \ldots, x_k\} \in X \tag{2}$$

$$\phi(v_i) \rightarrow e_i v_i \tag{3}$$

$$\phi(v_i) = x_1^i \circ x_2^i \circ x_3^i \circ \ldots \circ x_k^i \quad \textit{where} \quad \circ \ \textit{denotes concatenation.} \tag{4}$$

Extended Role2Vec focuses on minimizing the cosine distance d or maximizing cosine similarity s that relates a node to the nodes having similar roles using the mapping function ¥. For v_i and $v_j \in V$, which are two nodes having similar semantic roles, the goal should be:

$$¥(e_i^v | e_j^v) \rightarrow Max(S(e_i^v | e_j^v)) \quad or \quad Min(d(e_i^v | e_j^v)) \tag{5}$$

$$S(v_i | v_j) = \frac{e_i^v \cdot e_j^v}{\|e_i^v\| \|e_j^v\|} \tag{6}$$

$$d(v_i | v_j) = 1 - \frac{e_i^v \cdot e_j^v}{\|e_i^v\| \|e_j^v\|} \tag{7}$$

3.2.2. Attention vector calculation

Attention weight α_{uv} is the result of the attention mechanism, which defines how much attention or weightage must be given to a neighbor u of node v. In graph convolutional methods, it is defined as $\frac{1}{|N(v)|}$, where a weighing factor for all nodes is the same in GCN. The basic graph attention network is given by:

$$h_u^{(L)} = \sigma \left(\sum_{u \in N(v)} \alpha_{uv} W^{(L)} h_u^{(L-1)} \right) \tag{8}$$

$$a_{cv_u} = Q(W^{(L)} h_u^{(L-1)}, W^{(L)} h_v^{(L-1)}) \tag{9}$$

$$\alpha_{uv} = \frac{e^{a_{cv_u}}}{\sum_{k \in N(v)} e^{a_{cv_k}}} \tag{10}$$

Multiple α_{uv} values with different parameters for each attention head are used in multi-head attention. These are calculated as:

$$h_u^{r(L)}[1] = \sum_{u \in N(v)} \alpha_{uv1} W^{(L)r} h_u^{(L-1)} \tag{11}$$

$$h_u^{r(L)}[2] = \sum_{u \in N(v)} \alpha_{uv2} W^{(L)r} h_u^{(L-1)} \tag{12}$$

$$h_u^{r(L)}[n] = \sum_{u \in N(v)} \alpha_{uvn} W^{(L)r} h_u^{(L-1)} \tag{13}$$

Finally, the embedding of node u after multi-head attention is obtained by aggregating the results from all attention heads:

$$H_{u,r}^{(MH^L)} = \text{AGG}(h_u^L[1], h_u^L[2], \ldots, h_u^L[n]) \tag{14}$$

3.2.3. Proposed aggregation function

After the initial embedding generation, we introduce a novel aggregation function tailored for message passing and final node generation. This pioneering aggregation function is specifically designed to integrate a proximity vector, a crucial addition to address the over-smoothing dilemma pervasive in Graph Neural Networks (GNNs). Typically, GNN architectures limit the number of layers to 2-5 hops to mitigate over-smoothing. However, this restricted depth fails to capture the intricate semantics and contextual intricacies inherent in complex relations.

To overcome this limitation, we propose the inclusion of a proximity vector, which aggregates all nodes interconnected and proximate to a given relation. These proximity relations, derived from the training set produced by our novel triplet format, enrich the model's understanding by incorporating contextual information from neighboring nodes. In general, the Graph Neural Network (GNN) framework consists of layers, where each layer L includes a message function M and an aggregation function η. The message function for a node u is defined as:

$$M_L^u = mf^L(h_u^{(L-1)}) \tag{15}$$

where

$$mf^L(h_u^{(L-1)}) = W^{(L)} h_u^{(L-1)} \tag{16}$$

The proposed aggregation function is defined as:

$$\eta = H^L = ReLU\left(\sum_{r\in R}\sum_{u\in N(v)}\frac{1}{c(v,r)}H_{u,r}^{MH^L} + w_0^L h_v^{(L-1)} + \Delta\sum_{a\in Y}P_a\right) \tag{17}$$

where for v_i,

$$Z_i = H_i^1 + H_i^2 + \cdots + H_i^L \tag{18}$$

Here, H^L represents the updated embedding at layer L, and Z is the final embedding. The relational node degree is given by:

$$c_{v,r} = |N_{v,r}|, \quad N_{v,r} = \textit{relational node degree} \tag{19}$$

$$P = \sum_{Y} Z, \quad Y = \textit{set of proximity triplets} \tag{20}$$

The proximity function is defined as:

$$P_a = ReLU\left(\sum_{r\in R}\sum_{b\in N(a)}\frac{1}{c(a,r)}\alpha_{uv}w_r^L h_b^{(L-1)} + w_0^L h_a^{(L-1)}\right) \tag{21}$$

The message function is given by:

$$M = mf^L(h_u^{(L-1)}) \tag{22}$$

which can also be written as:

$$M = mf^L(H_u^L, \theta_G) \tag{23}$$

where θ_G represents the shared network parameters. The normalization function is denoted by:

$$\Delta = \textit{Normalization function} \tag{24}$$

Each relation has L relational weight matrices. For a given relation r, the relational weight matrices are:

$$W_r^1, W_r^2, \ldots, W_r^L \tag{25}$$

3.3. CPRMAM model training

3.3.1. Positive sampling training

We trained the model using the proposed aggregation function on the constructed ADR corpus. We set the number of hidden layers to two (k=2). The final embedding for each node has been encoded. We trained four distinct translational and deep learning graph

models on our cancer-specific triplet corpus. Our proposed model, the Causal Proximity Relational Multihead Attention Model (CPRMAM), was compared against four existing models: TransE, DistMult, ComplexE, and RAGAT. CPRMAM produced the best results because its encoders are trained to handle complex and indirect relationships, thereby enhancing the efficiency of knowledge discovery to generate accurate and precise results. The training data contains 132184 triplets in the proposed novel format.

3.3.2. Negative sampling training

Negative sampling is also important to reduce false positive cases. We opted 1:N negative sampling strategy [Qian *et al.* (2021)].

3.3.3. Loss computation and weight updation

We trained our CPRMAM model on two different loss functions; cross entropy loss and pairwise hinge loss. It was seen that model performed better with the cross entropy loss function. Then we applied gradient descent to update the weights.

$$Cross\ Entropy\ Loss = -log(ReLU(f_{r_k}(h_i, h_j))) - log(1 - ReLU(f_{r_k}(h_i, h_j))) \tag{26}$$

3.3.4. Selection of most suitable parameters

Efficient model training depends on the selection of correct parameters and avoiding overfitting and underfitting. To make sure that the model gives the best results, we used grid search to find the suitable embedding size z, learning rate lr, number of corruptions that have to be made for each triplet *eta*, and maximum epochs to run.

3.3.5. Triplet scoring

After training the encoder part and generation of final embeddings, the decoder plays an important role in scoring the sample triplets and generates non-existent knowledge. We used four different scoring functions with each different encoder. Table 4 shows the possible combinations of encoders and decoders, or we can say aggregation function and scoring functions. It was noticed that the proposed aggregation function as encoder and ConvE as decoder/ scoring function gave the best results.

3.3.6. Corrupt triplet generation

To test the model against false positives, we generated corrupted triplets in test data using the filter strategy. In the filter strategy, only one element of a triplet will be corrupted by replacing that element with random entities and relations. Corruption parameter eta defines the total number of corruptions that should be generated for each triplet. It was found that the most suitable value for *eta* is 20. Practically generating corruption for every triplet is not feasible. Therefore, we corrupted triplets that contain the relations among entities of interest: drug, ADR, cancer types, food, genes, protein, and allele.

$$t = (v_i, r_k, v_j), \textit{create corrupted triplets} \quad t_c = (v_i, r_k, v_j^c) \tag{27}$$

3.3.7. Scoring of triplets

After generating corrupted triplets in test data, every triplet will now be scored based on scoring functions/decoder. We used four available scoring functions: TransE, DistMult, Complex, and ConvE. Score tells how likely two nodes relate to a particular relation. Each scoring function mentioned have been used in possible combination with the available algorithms and proposed model. It was noticed that ConvE performed better as shown in Table 5.

Compute the score for the triplet t using the function f that takes the relationally weighted final layer embeddings h and the transformation matrix $\bigodot$.

$$f_{r_k}(h_i, h_j) = f_{r_k}(W_{r_k}, h_i, W_{r_k}, h_j) \tag{28}$$

$$f_{r_k}(h_i, h_j) = ReLU\Big(vec\Big(ReLU\Big([H_i; e_r] * \bigodot\Big) W_r\Big) H_j\Big) \tag{29}$$

4. Triplet ranking and model evaluation

Since there is no benchmark dataset available for model validation, the performance of the model has been evaluated using several metrics: Mean Reciprocal Rank (MRR) [Craswell (2009)], Hits@1, Hits@3, Hits@10, and Hits@100 [Khan *et al.* (2024)], as shown in Tables 3 and 4. Each triplet in the test data is ranked against its corrupted triplet variants based on the generated score. In Table 3, the left side presents the available and proposed encoders for final embedding generation, while the top side lists the available decoders used to score triplets based on the translation method. These encoders and decoders are combined, and the model's performance is evaluated for each possible combination using the mentioned metrics. Table 4 provides the evaluation metrics for the GNN-based model and the proposed model when combined with translational and convolutional decoders. The proposed model, CPRMAM, achieved the best results when paired with the ConvE scoring function. We have used dense ranking strategy for triplet ranking. These metrics assess the overall quality of the model in ranking the correct triplet for ADR prediction, which is particularly crucial in high-stakes domains like oncology, where incorrect ranking can have severe consequences.

$$rank = COUNT(\textit{corruption score} \geq \textit{hypothesis score}) + 1 \tag{30}$$

The MRR (equation 31) measures the ranking quality by computing the average reciprocal rank of the first correct triplet among the predicted scores. It is defined as:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{31}$$

where Q is the set of all queries, and rank_i represents the position of the first correct triplet for the i-th query. A higher MRR value indicates that the model ranks correct predictions closer to the top, which is critical in clinical settings where lower-ranked correct predictions may be overlooked.

The Hits@K metric (equation 32) evaluates the proportion of test samples where the correct entity appears within the top K ranked predictions. It is formulated as:

$$Hits@K = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathbf{H}(rank_i \leq K) \tag{32}$$

where $\mathbf{H}$ is an indicator function that returns 1 if the condition inside holds, otherwise it will be 0. A higher Hits@K value signifies that the model can reliably place correct ADR predictions among the top K results, ensuring that critical drug interactions are not missed.

These metrics are particularly significant in the domain of oncology, where precise ADR prediction is essential for patient safety. A high MRR ensures that correct predictions appear early in ranked lists, making them more accessible for decision-making. Similarly, a high Hits@K ensures that crucial ADR predictions are consistently included in the top-ranked results, reducing the risk of missing life-threatening interactions. By optimizing these metrics, models can improve clinical decision support systems, leading to safer and more effective treatment strategies.

Table 3: Evaluation metrics for translational-based models

Decoder	Encoder	MRR	Hits@1	Hits@3	Hits@10
TransE	TransE	0.236	0.004	0.087	0.125
	DistMult	0.284	0.005	0.051	0.150
	ComplEx	0.297	0.008	0.101	0.189
DistMult	TransE	0.284	0.005	0.471	0.145
	DistMult	0.290	0.007	0.018	0.124
	ComplEx	0.295	0.052	0.907	0.105
ComplEx	TransE	0.297	0.008	0.101	0.189
	DistMult	0.295	0.007	0.018	0.128
	ComplEx	0.307	0.052	0.087	0.287

5. ADR related knowledge discovery

After evaluating the CPRMAM model on testing data, we try to predict all possible adverse drug reactions for all cancer related drugs that were not earlier mentioned in the training text data by predicting the missing edges in the knowledge graph. Our model predicted 200 missing links that have been ranked based on the score generated by the scoring function for each triplet. The following triplets are generated through the knowledge discovery process and were not present earlier in the corpus.

(colon cancer, undergo, galantamine)

Table 4: Evaluation metrics for GNN-based models

Decoder	Encoder	MRR	Hits@1	Hits@3	Hits@10
RAGAT	TransE	0.244	0.144	0.344	0.414
	DistMult	0.244	0.144	0.344	0.414
	ComplEx	0.234	0.154	0.354	0.398
	ConvE	0.498	0.301	0.341	0.398
CPRMAM	TransE	0.348	0.258	0.548	0.878
	DistMult	0.348	0.258	0.548	0.877
	ComplEx	0.375	0.281	0.581	0.818
	ConvE	0.658	0.538	0.621	0.695

(Ipilimumab, adverse effects, fever)
(Ipilimumab, adverse effects, increased orthostatic hypotensive activities)

Table 6 provides a knowledge discovery analysis of ADRs identified by CPRMAM and the baseline models (TransE, DistMult, and ComplEx). While all models detect common ADRs, CPRMAM uniquely incorporates critical medical contexts, such as affected body organs, age-related factors, and dose dependency key aspects for clinical relevance.

For example, baseline models identify "peeling" as an ADR of 5-Fluorouracil, but only CPRMAM specifies "peeling at palms and soles," enhancing clinical interpretability. Likewise, CPRMAM uniquely captures age-related nausea and constipation (more prevalent in patients aged 65+) and dose-dependent seizures, improving ADR specificity. These refinements enhance real-world applicability in oncology pharmacovigilance.

6. Model comparison and validation

6.1. Validation of constructed ADR corpus

To ensure the accuracy and quality of the constructed ADR corpus, we conducted validation at four stages.

First, we validated the extracted entities by normalizing and mapping them to standard medical terminologies using SNOMED CT. In the second stage, we validated the relationships between drugs and the adverse drug reactions associated with cancer and its various types, using established pharmaceutical knowledge databases: SIDER and DrugBank. Third, any relationships not found in these databases were confirmed by a domain expert. Finally, in the fourth stage, we validated the missing links generated by the CPRMAM model, again in consultation with medical experts.

6.2. Model comparison

The results obtained through the proposed model 1 were systematically compared with two state-of-the-art models, namely ScispaCy and BERT, for relation extraction. Notably, the proposed approach demonstrated a superior capability in incorporating additional semantics, evidenced by its extraction of a greater number of relations compared to the aforementioned methods. Figure 8 visually presents the relations extracted by the proposed

approach and also provides an illustration of the relations extracted by the existing methods. Proposed model 2 CPRMAM addresses the existing limitations in the literature and performs better than existing models. Different GNN algorithms were implemented in combination with different scoring functions. We applied different scoring function combinations and found proposed aggregation function with ConvE scoring function gave the best results in evaluation. Our model achieved 0.658 MRR, 0.538 Hits@1, 0.621 hits@3, 0.695 hits@10 and 0.877 Hits@100. The proposed model was compared with the state-of-the-art model; RAGAT. It was noticed that the proposed model performed better. Comparison can be seen in Table 5.

Table 5: Model comparison based on MRR and Hits@K metrics

Model	MRR	Hits@1	Hits@3	Hits@10	Hits@100
RAGAT	0.498	0.301	0.341	0.398	0.414
CPRMAM	0.658	0.538	0.621	0.695	0.877

Figure 8: Results comparison

Table 6: Analysis of adverse drug reaction related knowledge discovery with CPRMAM and other models

Drug	Adverse drug event (ADE)	CPRMAM	TransE	DistMult	ComplEx	Clinical Relevance
Erlotinib	*Hepatorenal syndrome*	No	No	No	No	Rare ADR
	Internal bleeding	No	No	No	No	Severe ADR
5-Fluorouracil	*Swelling*	No	No	No	No	Common ADR
	Palmar-plantar erythrodysesthesia	No	No	No	No	Drug-specific ADR
	Redness	No	No	No	No	Common ADR
	Peeling (palms and soles)	Yes	No	No	No	Body location-enhanced ADR (unique to CPRMAM)
	Blisters (skin)	Yes	No	No	No	Body location-enhanced ADR (unique to CPRMAM)
Cabazitaxel	*Nausea (increased in 65+ years)*	No	No	No	No	Age-related ADR
	Constipation (increased in 65+ years)	No	No	No	No	Age-related ADR
Prednisone	*Weight gain (belly)*	Yes	No	No	No	Body location-enhanced ADR (unique to CPRMAM)
Doxorubicin	*Pain (belly)*	Yes	No	No	No	Location-enhanced ADR (unique to CPRMAM)
Fludarabine	*Seizures (with high doses)*	No	No	No	No	Dose-dependent ADR
	Stroke	No	No	No	No	Severe ADR
	Heart attack	No	No	No	No	Severe ADR
	Blindness	No	No	No	No	Rare ADR

7. Conclusion

In conclusion, this study presents two models designed to enhance adverse drug reaction (ADR) prediction in oncology, addressing the limitations in existing approaches. Our first model focuses on automated corpus construction, employing rule-based entity extraction, coreference resolution, and relation extraction to annotate entities of interest from cancer-specific case reports. By introducing a novel triplet format, it captures regular and causal relations with improved semantic clarity. The second model, CPRMAM, leverages this annotated corpus and integrates causal and proximity-based information into Graph Neural Networks (GNNs). By overcoming challenges like over-smoothing and equal neighbor weightage, CPRMAM captures complex, indirect, and nested relations, significantly enhancing the quality of knowledge graphs and message passing in GNNs. Together, these models offer a robust framework for ADR prediction and knowledge discovery, with potential applications in broader biomedical domains. Future work will focus on expanding the models' scope, incorporating additional data sources, and further enhancing their generalizability across diverse biomedical contexts.

Acknowledgements

We would like to express our heartfelt gratitude to Dr. Neera Samar, Professor of General Medicine, RNT Medical College and Hospital, Udaipur, for her invaluable expertise and validation of our findings. Her insightful guidance and expert validation have significantly contributed to the rigor and relevance of this research. We would also like to acknowledge the University Grants Commission (UGC), India, for providing financial support through the Junior Research Fellowship (JRF) under the UGC-JRF scheme. Finally, we thank the reviewers for their constructive feedback and suggestions, which have greatly enhanced the quality of this manuscript.

Conflict of interest

No financial or non-financial conflicts of interest are associated with this research work included in this article.

References

Ahmed, N. K., Rossi, R., Lee, J. B., Willke, T. L., Zhou, R., Kong, X., and Eldardiry, H. (2018). Learning role-based graph embeddings. In *ArXiv*. `/abs/1802.02896`.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Atias, N. and Sharan, R. (2011). An algorithmic framework for predicting side effects of drugs. *Journal of Computational Biology*, **18**, 207–218.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., and DeFreitas (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, **54**, 315–321.

Craswell, N. (2009). *Mean Reciprocal Rank*, pages 1703–1703. Springer US, Boston, MA.

Dev, S. and Sharan, A. (2023a). Annotated and normalized causal relation extraction corpus for improving health informatics. In *Proceedings of ICON 2023, 20th International Conference on Natural Language Processing*. Presented at ICON 2023.

Dev, S. and Sharan, A. (2023b). Automatic construction of named entity corpus for adverse drug reaction prediction. In *Lecture Notes in Computer Science: Innovations in Data Analytics (ICIDA)*, volume 1442. Springer.

Dey, S., Luo, H., Fokoue, A., Hu, J., and Zhang, P. (2018). Predicting adverse drug reactions through interpretable deep learning framework. *BMC Bioinformatics*, **19 Suppl 21**, 476.

Gao, Y., Ji, S., Zhang, T., Tiwari, P., and Marttinen, P. (2023). Contextualized graph embeddings for adverse drug event detection. In *Lecture Notes in Computer Science: Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 13714. Springer.

Ji, Y., Ying, H., Dews, P., Mansour, A., Tran, J., Miller, R. E., and Massanari, R. M. (2011). A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine*, **15**, 366–372.

Khan, M., Mello, G. B. M., Habib, L., Engelstad, P., and Yazidi, A. (2024). Hits-based propagation paradigm for graph neural networks. *ACM Trans. Knowl. Discov. Data*, **18**.

Kwak, H., Lee, M., Yoon, S., Chang, J., Park, S., and Jung, K. (2020). Drug-disease graph: Predicting adverse drug reaction signals via graph neural network with clinical data. In *Advances in Knowledge Discovery and Data Mining*, volume 12085, pages 633–644. Springer.

Lerch, M., Nowicki, P., Manlik, K., and Wirsching, G. (2015). Statistical signal detection as a routine pharmacovigilance practice: effects of periodicity and resignalling criteria on quality and workload. *Drug Safety*, **38**, 1219–1231.

Lin, J., Kuang, Q., Li, Y., Zhang, Y., Sun, J., Ding, Z., and Li, M. (2013). Prediction of adverse drug reactions by a network based external link prediction method. *Analytical Chemistry*, **5**, 6120.

Liu, B.-M., Gao, Y.-L., Li, F., Zheng, C.-H., and Liu, J.-X. (2024). SLGCN: Structure-enhanced line graph convolutional network for predicting drug–disease associations. *Knowledge-Based Systems*, **283**, 111187.

Liu, R., AbdulHameed, M. D. M., Kumar, K., Yuˆ, X., Wallqvist, A., and Reifman, J. (2017). Data-driven prediction of adverse drug reactions induced by drug-drug interactions. *BMC Pharmacology and Toxicology*, **18**, 44.

Mantripragada, A. S., Teja, S. P., Katasani, R. R., Joshi, P., V, M., and Ramesh, R. (2021). Prediction of adverse drug reactions using drug convolutional neural networks. *Journal of Bioinformatics and Computational Biology*, **19**, 2050046.

Manzi, S. and Reis, B. (2011). Predicting adverse drug events using pharmacological network models. *Science Translational Medicine*, **3**, 114ra127.

Monaco, L., Melis, M., Biagi, C., Donati, M., Sapigni, E., Vaccheri, A., and Motola, D. (2017). Signal detection activity on eudravigilance data: analysis of the procedure and findings from an italian regional centre for pharmacovigilance. *Expert Opinion on Drug Safety*, **16**, 271–275.

Pauwels, E., Stoven, V., and Yamanishi, Y. (2011). Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*, **12**, 169.

Qian, J., Li, G., Atkinson, K., and Yue, Y. (2021). Understanding negative sampling in knowledge graph embedding. *International Journal of Artificial Intelligence and Applications*, **12**, 71–81.

Schuemie, M. J., Coloma, P. M., Straatman, H., Herings, R. M., Trifirò, G., Matthews, J. N., Prieto-Merino, D., Molokhia, M., Pedersen, L., Gini, R., *et al.* (2012). Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Medical Care*, **50**, 890–897.

Shang, N., Xu, H., Rindflesch, T. C., and Cohen, T. (2014). Identifying plausible adverse drug reactions using knowledge extracted from the literature. *Journal of Biomedical Informatics*, **52**, 293–310.

Szarfman, A., Machado, S. G., and O'Neill, R. T. (2002). Use of screening algorithms and computer systems to efficiently signal higher than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Safety*, **25**, 381–392.

Wang, S. V., Maro, J. C., Baro, E., Izem, R., Dashevsky, I., Rogers, J. R., Nguyen, M., Gagne, J. J., Patorno, E., Huybrechts, K. F., *et al.* (2018). Data mining for adverse drug events with a propensity score-matched tree-based scan statistic. *Epidemiology*, **29**, 895–903.

Xiang, Y.-P., Liu, K., Cheng, X. Y., Cheng, C., Gong, F., Pan, J. B., and Ji, Z. L. (2015). Rapid assessment of adverse drug reactions by statistical solution of gene association network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **12**, 844–850.

Xiao, C., Zhang, P., Chaowalitwongse, W. A., Hu, J., and Wang, F. (2017). Adverse drug reaction prediction with symbolic latent dirichlet allocation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*, pages 1590–1596. AAAI Press.

Xue, R., Liao, J., Shao, X., Han, K., Long, J., Shao, L., Ai, N., and Fan, X. (2020). Prediction of adverse drug reactions by combining biomedical tripartite network and graph representation model. *Chemical Research in Toxicology*, **33**, 202–210.

Yildirim, P., Majnarić, L., Ekmekci, O., and Holzinger, A. (2014). Knowledge discovery of drug data on the example of adverse reaction prediction. *BMC Bioinformatics*, **15 Suppl 6**, S7.

Zhang, F., Sun, B., Diao, X., Zhao, W., and Shu, T. (2021). Prediction of adverse drug reactions based on knowledge graph embedding. *BMC Medical Informatics and Decision Making*, **21**, 38.

Zhao, Y., Yi, M., and Tiwari, R. C. (2018). Extended likelihood ratio test-based methods for signal detection in a drug class with application to FDA's adverse event reporting system database. *Statistical Methods in Medical Research*, **27**, 876–890.

Zheng, Y., Peng, H., Zhang, X., Zhao, Z., Yin, J., and Li, J. (2018). Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases. *BMC Bioinformatics*, **19 Suppl 19**, 517.