Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 20(5):8892–8932.
DOI: 10.3934/mbe.2023391
Received: 30 December 2022
Revised: 16 February 2023
Accepted: 20 February 2023
Published: 09 March 2023

Research article

Enhanced disease-disease association with information enriched disease representation

Karpaga Priyaa Kartheeswaran*, Arockia Xavier Annie Rayan and Geetha Thekkumpurath Varrieth

Department of Computer Science and Engineering, CEG Campus, Anna University, Chennai, Tamil Nadu, India

* **Correspondence:** Email: karpagapriyaa.phd@gmail.com.

Abstract: Objective: Quantification of disease-disease association (DDA) enables the understanding of disease relationships for discovering disease progression and finding comorbidity. For effective DDA strength calculation, there is a need to address the main challenge of integration of various biomedical aspects of DDA is to obtain an information rich disease representation. Materials and Methods: An enhanced and integrated DDA framework is developed that integrates enriched literature-based with concept-based DDA representation. The literature component of the proposed framework uses PubMed abstracts and consists of improved neural network model that classifies DDAs for an enhanced literature-based DDA representation. Similarly, an ontology-based joint multi-source association embedding model is proposed in the ontology component using Disease Ontology (DO), UMLS, claims insurance, clinical notes etc. Results and Discussion: The obtained information rich disease representation is evaluated on different aspects of DDA datasets such as Gene, Variant, Gene Ontology (GO) and a human rated benchmark dataset. The DDA scores calculated using the proposed method achieved a high correlation mainly in gene-based dataset. The quantified scores also shown better correlation of 0.821, when evaluated on human rated 213 disease pairs. In addition, the generated disease representation is proved to have substantial effect on correlation of DDA scores for different categories of disease pairs. Conclusion: The enhanced context and semantic DDA framework provides an enriched disease representation, resulting in high correlated results with different DDA datasets. We have also presented the biological interpretation of disease pairs. The developed framework can also be used for deriving the strength of other biomedical associations.

Keywords: disease representation; information integration; biomedical literature; ontology; DDA quantification

1. Introduction

DDA acts as a key factor to understanding disease relationships, such as comorbidity, which is essentially the co-occurrence of diseases among the same patients that plays an important role in health care for drug discovery [1] and better treatment plan. To meet the emerging need, several studies in biomedical domain for relating diseases have been carried out [1,2–4]. In the work of Suratanee and Plaimas [3], a network-based approach was employed to calculate DDA strength that achieves a performance of 0.71 area under curve (AUC). Zitnik et al. [1] has predicted DDA relationships and found about 66 disease classes have significant high relationships with p-value < 0.001. Another work in [4], a disease similarity database tool was developed that performs hypergeometric test of p-values for different pairs of diseases. On the other hand, the DDA relationships were analysed using disease causality network. Further, the sorted potential association strength were compared between top and bottom group of disease pairs and found 95% of disease pairs in upper group. Since one disease can multiply into another in any patient, treating associated diseases is a great challenge for modern medicine. Hence, exploring DDA helps in gaining better insight of disease relationships, which is helpful for clinicians in proper diagnosis and treatment.

For better understanding of DDA, it is important to know the various underlying aspects with which diseases are associated. One such aspect considers biological entities such as other diseases [5], genes [4,6], pathways [7], drugs [8], and phenotypes [9] as intermediate factors, facilitating indirect DD association. Another aspect, revolves around the vast established heterogeneous biomedical databases such as biomedical datasets including Protein-Protein Interaction Network [4,10], HumanNet [11] and biomedical ontologies like DO [12], GO [13], Human Phenotype Ontology [14], Unified Medical Language System (UMLS) [15], Medical Subject Headings (MeSH) [16]. On the other hand, connection between diseases can be inferred using biomedical text such as PubMed [17,18], MedLine [19], Clinical Notes, Claims Database and PubMed Central (PMC) [20], Electronic Health Records [21] and HealthMap Corpus [22]. In order to widen the range of components affecting disease associations, non-Biomedical Text such as Wikipedia [17,23] has also been considered.

In addition, measuring the strength of DDA helps to improve the clinical decision making. As a quantitative measurement, disease similarity is generally used to indicate the extent to which the diseases are associated, since similar diseases are usually caused by similar semantic aspects such as similar etiology, markers, mechanisms, patterns etc. In this regard, by involving a single biological source, the strength of disease associations is computed by IC-Based methods such as Wang et al. [24], Resnik [25] and Lin [26], accomplished solely based on semantic associations of ontologies such as MeSH, DO, HPO. Taking advantage of biological process terms, some statistical-based approaches are proposed. In the work of Mathur and Dinakarpandian [27] calculated the association strength by overlapping genes of diseases using GO. In another work, association of diseases is computed using both information content and co-occurrence of terms in ontology [28]. Recently some research employed neural network approach, word embedding model, to learn ontological node vector representations used in application of associating diseases through similarity values [29]. Apart from ontologies, DDAs can also be quantified by mining a large corpus of biomedical literature. In the context of text, O'Shea [18] used a network-based shortest path distance method to calculate the relatedness between diseases from occurrence frequency of disease terms. Alternatively, using neural network-based approach, Beam et al. [20] derived distributional vector representations from clinical notes, insurance claims, journal articles and projected the learned context-based concept vector representations to distributional space for relatedness computation. Therefore, in general either semantic aspects or concept-based aspects have been considered for the calculation of DDA strength. However, considering both the above aspects could lead to more effective strength calculation.

Some efforts have been put-forth to combine different biomedical knowledge from various sources to derive representations of biomedical concepts for measuring the relatedness of the concepts. There are works that fused various biomedical knowledge such as biomedical entities, biomedical datasets and ontologies [30,31]. On the other hand, with the growing biomedical literature, some work has attempted to compute relatedness of biomedical concepts, with an integrated vector representations mined from both literature and semantic ontological information [32,33]. However, the integrated vector encoded only limited aspects of contextual relations from literature and semantic relations from ontology. Hence, in this paper, an integrated vector is derived covering a wide range of both contextual and semantic relations for an effective DDA strength calculation.

The structure of the paper is organized as follows: Section 2 briefly reviews the state-of-the-art methods related to biomedical association classification and strength computation. Subsequently, a set of datasets used in this work and the proposed DDA framework is described in detail in Sections 3 and 4 respectively. Section 5 presents the experimental results that evaluates the quantified DDA scores obtained using the proposed framework. Finally, an outline of conclusion is drawn in Section 6.

2. Related works

2.1. Literature-based approaches

Biomedical literature contains associations linking diseases with other diseases. Given their significance in health-oriented applications, it is imperative to investigate these digitized data to extract the type of association using text mining approach. Given a sentence and disease pair appearing within the sentence, the DDA type can be of 3 types: positive association, where there exists an explicit mention of association with words like association, comorbidity factors, complicatin, risk factors, etc., negative association, in which a negative word explicitly conveying that no relation exists between the two disease mentions and neutral or null association that does not state about any association between the co-occurring diseases. Towards this end, a number of literature-based methods have been proposed for the extraction of associations between different biomedical entities [17,34–37].

The co-occurrence statistical technique, assumes that more the frequency of entities occurring together within abstract or sentence higher the chance of being positively associated [8]. Li et al. [38] employed the co-occurrence statistics to detect disease-related associations. Rosário-Ferreira et al. [47] considered diseases to be related if they are co-mentioned in the abstract text. However, entities occurring together may not be semantically connected, and thus result in low precision [39–41].

Some manually or automatically formulated rules finds its role in the association extraction task. Lee et al. [42] and Song et al. [43] drafted number of rules manually for PPI and disease-gene relation extraction respectively. In addition, Tari et al. [44] used automatically created rules to identify the biomedical relations from MEDLINE abstracts. The major limitation of rule-based system is that it is difficult to create rules entailing all types of associations and moreover a deeper insight into the biomedical knowledge for creation of such rules is required.

However, with the huge set of annotated training text available for biomedical associations, machine learning approach can overcome the above limitations by its ability to learn relation patterns of sentences which can then automatically detect the association type in unseen texts. Bhasuran and Natarajan [45] used a supervised machine learning method for gene-disease association extraction, which required a large training set and was time-consuming. Zhang and Lu [46] and Rosário-Ferreira et al. [47] eliminated this deficiency by using a semi-supervised method, that utilized a small training set which learns DDA patterns from PubMed abstracts. However, machine learning (ML)-based methods require enormous manual efforts in designing biomedical relation features for the association extraction task as ML methods lack automatic feature extraction.

These issues were addressed by employing deep neural networks for efficient feature engineering in text-mining for curating number of biomedical relation types, as it involved an automatic feature learning process [35,48–50]. One of the popular deep neural network models, Convolutional Neural Network (CNN), was widely used for classifying whether sentences contain positive, negative or null associations between biomedical entities using sentence representation, where different representations of various local-level features captured at sentence-level and global-level features captured at corpus-level were used for classification [17,34,37].

A Multi-Channel Dependency based CNN extracted PPIs into positive and negative associations, where the sentence representation covered word embeddings trained only on global-level features from PubMed and PMC [35]. Using additional embeddings from Wikipedia and MEDLINE, the Multi-Channel CNN (MCCNN) model classified DDI and PPI into positive associations such as effect, mechanism, etc and negative associations. An attempt was made to classify different biomedical associations such as gene-disease associations (GDAs) [34], using disease position as the only local-level feature, DDAs [17] using Parts-of-Speech (POS) as additional feature and spice-disease [37] using Parts-of-Speech (POS) and chunk tag as additional local-level features.

However, only a limited number of local-level and global-level features were used in sentence representation for the sentence-level classification of biomedical associations into positive, negative and null.

Similar research considering local and global text and video features have been carried out in the work of Wang et al. [51] for video-text retrieval. In the text part, they considered only the encoded full text representation as global text feature and the decoded global representation is extracted as local text feature. In neither case, no various local-level features nor the global-level features of each word in given text is embedded.

Moreover, most of the above work, only classified associations and did not attempt to calculate the association strength. An attempt was made to calculate only the strength of positively correlated pairs using statistical [18] and pattern-based approaches [52]. While literature-based approaches have mainly been used for the classification of biomedical associations, we need a concept-based approach for effective association strength calculation.

2.2. Concept-based approaches

Biomedical ontologies have integrated non-duplicative biomedical concept terms and medical data, providing a high coverage of biomedical concept terms which have been used to compute the semantic association strength between biomedical entities. Quantitative semantic association among diseases help clinicians gain a better knowledge of diseases, since semantically associated diseases reveal similar or common underlying attributes, that further help in proper treatment plan [31]. Therefore, discovering the quantitative semantic biomedical associations using biomedical ontologies plays a crucial role in biomedical field [11,31].

Some work has encoded conceptual sources for computing semantic associations. Wei et al. [53], Beam et al. [20] and Pakhomov et al. [54] used only unstructured corpora such as insurance claims, clinical notes, etc., to include the conceptual aspects into the association computation. While Wei et al. [53] exploited ontology only to retrieve disease concepts. With additional semantic relation types information, Yu et al. [33] attempted to associate biological entities with improved semantics. However, taxonomic relationships conveyed by ontologies are needed for an enhanced semantic association quantification.

Most of the ontology-based methods were node-based, edge/path-based and hybrid-based. The node-based approaches use properties of the node such as Information Content (IC) [25,55] and their variants [56–58] for computing semantic association between the concepts based on their lowest common ancestor. However, the IC values computation is based on the annotated corpus and hence is corpus dependant. On the other hand, the edge/path- based approach uses the edges count between the given concepts to measure the association. One such method proposed by Wu and Palmer [59], used the common path from root node to the least common ancestor node while Richardson et al. [60] used the edge weight technique based on node density, depth and connections between parent-child nodes for computing the conceptual associations. Further, Cheng et al. [61] proposed a weighted maximum common ancestor depth and Wu et al. [62] proposed a non-weighted maximum common ancestor depth to measure the semantic associations. Using the topology of DO, Wang et al. [63] calculated the strength of association by considering the semantic impact of ancestors on the entities involved in association. However, the problem with edge-based measure is that the concepts at same depth are not semantically well differentiated. As a hybrid measure, Mazandu and Mulder [64] used the topological positional characteristics of the GO for association strength calculation. Zhao and Wang [58] computed relatedness using the count of children nodes and topology of GO. Kamran and Naveed [65] also exploited the topology of GO along with common descendants to calculate the strength of associations. However, the computation of semantic relatedness using hybrid methods have not incorporated the semantic meaning of the concepts captured within the ontology.

Semantic associations based on semantic meaning of concepts can also be computed using vectors learnt from the ontological graph structure. Camacho-Collados et al. [66] used the graph-based vectors and computed the semantic association, where the vector representation is solely based on the structure of the graph. Guo et al. [67] and Zhong et al. [68], used graph embeddings which can capture the structural information connecting nodes in graph but no relationship information was considered. Smaili et al. [69] represented concepts by general corpus trained aggregated embeddings of all its annotated nodes including the ancestors, where there is no control on the amount of ancestorial information affecting the given concept. Hence, the problem with vector-based association is that representation of vectors has encoded only a limited ontological relationship information without any control of the contribution effect of the entities involved in the association.

2.3. Integrated approaches

Attempts have been made to measure association between diseases by integrating multiple data sources as well as fusing the details of various biological entities extracted from these biomedical sources. Su et al. [31] developed a joint association method combining biological entities such as genes, phenotypes and integrating ontological sources (DO, HPO), where semantic associations determine the disease associations. Similarly, Cheng et al. [30] spans different biomedical sources (DO, HumanNet) fusing functional and semantic associations for measuring the association strength. With the unprecedented growth of biomedical literature, there has been a significant gap between the increasing published scientific knowledge and the tailored biological data knowledge [70]. Hence, it is necessary to integrate the contextual knowledge obtained from biomedical literature with the semantic knowledge of biological data sources for the DDA task. Deng et al. [71] used the biological-process based approach, integrating both literature and ontology (GO) and proposed a combined score of semantic and contextual associations using symptoms, genes and their related functions. In addition, li et al. [72] proposed a relatedness method integrating contextual and functional associations mined from literature (MedLine) and biomedical network (PPI), respectively. Moreover, Jiang et al. [32] proposed a hybrid semantic embedding model incorporating both corpus-based distributional representation into multiple ontologies to gain a better similarity score of biomedical concepts. Similarly, Yu et al. [33] used neural network approach to induce the vector representation of biomedical concepts by retrofitting contextual information from literature (PubMed) using semantic information from ontology (UMLS) such that the resulting vectors can be utilized to measure the association strength. However, both Jiang et al. [32] and Yu et al. [33], generated the corpus-based representation for each concept independently without considering the different types of context (association) of the sentences. On the other hand, the ontological knowledge integrated by Jiang et al. [32], was only edge-based semantic similarity of concept pairs that did not incorporate semantic meaning of concepts as well as their ontological relationship connections. In addition, the existing methods associate the biomedical concepts (entities) using only a limited aspect of contextual and semantic relations, which results in low correlation with human judged association scores.

Thus, for the biomedical association quantification from literature, particularly DDA, the existing classification model has used only a limited number of local-level and global-level features that could capture only limited syntactic, semantic, and contextual features for sentence representation learning. Hence, in order to improve the classification performance, there is a need to include additional local and global-level features. The existing methods either not calculated or calculated only positive association strengths. However, it is important to quantify the strength of DDA pairs based on all types of DDA pairs positively, negatively, and null associated by sentence embeddings under different contexts.

Similarly, for concept-based quantification of DDA, existing methods embedded concepts by considering only the connectivity of concepts in ontology. The semantic meaning of concepts and the various ontological relationships affecting the associations not embedded. In addition, all ancestors are treated equally. However, controlling the impact of ancestorial embedding is important as each ancestor may either be closely or distantly related to each concept in the association.

The integrated approaches fusing literature and ontology, did not consider different context types of sentences from literature and did not incorporate multiple semantic meaning of concepts with ontological relationships. Moreover, the existing methods have fused only limited semantic type relations from ontology with limited contextual relations from literature. However, the association varies based on the taxonomical connection relationship type that exists in the ontology. Therefore, there is a need to integrate both contextual relations from literature and richer semantic relationships from ontologies for an enhanced DDA strength quantification.

Of significance, while there are existing association quantification methods that have fused semantic relations from ontology with contextual relations from literature, we improve the association quantification in this paper:

1) We enhanced literature-based DDA representation by considering all context types of association sentences such as positive, negative and null with improved sentence representation.

2) We also enhanced concept-based DDA representation by the proposed ontology-based joint multi-source association representation where semantic meaning of concepts and the various ontological relationship connections are incorporated for a better DDA quantification.

3) We present an enhanced and integrated DDA framework to widen the coverage of various relationship aspects of association components both contextually A) and conceptually (semantically) B) to build an information enriched disease vector representation.

3. Dataset description

3.1. Collection of unlabeled PubMed abstracts

We initially used the available and already annotated 521 abstracts dataset [17] for training of the proposed ESEC-CNN model. However, in order to achieve better modelling, we expanded this dataset. To assist the DDA dataset expansion, an initial set of approximately 3 million bio-concept annotated disease-related PubMed abstracts have been extracted using PubTator. PubTator, an automatic text-mining tool, recognize various biomedical entities such as genes/proteins, diseases, genetic variants, spices and chemicals in the titles, abstracts of PubMed articles [73]. To ensure sentence-based DDA, only 39,510 abstracts with at least a DDA sentence are retained for further processing.

3.2. Disease ontology

DO, a taxonomy of diseases, in which each disease term is linked to another in a hierarchical manner by a semantic type “is_a” association has been used [12]. DO mapping each disease term to its disease id DOID along with the term definition and the human disease related knowledge base is downloaded from http://purl.obolibrary.org/obo/doid/releases/2022-06-07/doid.owl_(accessed 7 June 2022). In this work, the conceptual linking of diseases for concept-based DDA has been established using various DO relationships. Approximately 8000 diseases out of 14,958 diseases from the enhanced dataset were mapped to DO, whose corresponding term definitions are further utilized in concept embedding.

3.3. Unified Medical Language System

The UMLS consists of three components, Metathesaurus, Semantic Network and Lexicon tools, that has concepts with concept ID (CUI), definitions and its linkage to other concepts with semantic relations such as CHD “Child”, SY “asserted synonymy”, RN “has a narrower relationship”, RO “has other relationship”, RQ “related and possibly synonymous”, etc. In this work, only Metathesaurus concepts file, containing the concept pairs relationships are used for concept embedding in concept-based DDA [15].

3.4. Datasets for evaluation

We evaluate the obtained DDA scores of our approach against the results of DisGeNET, that contains about 10,48,575 DD pairs from a curated DDA database. DisGeNET defines DDAs based on shared genes and variants among the available gene-disease associations [74]. This well-known database has been used for direct comparison of DDA strengths in both the perspectives. Nicia et al. [47] used DisGeNET to evaluate the results of DDAs obtained using SicknessMiner. The phenotypic similarity of diseases werealso evaluated using the DisGeNET scores for inborn errors of immunity [75]. Further, we created a standard dataset, to compute DDA strength using functional GO as an association criteria. The disease-related GOs are obtained from CTD. Some of the attributes of the datasets are disease1, disease2 and the Jaccard similarity scores using genes, variants and GOs. In this work, we have adopted DisGeNET as well as the created standard dataset for evaluating DDA strength.

In addition, the performance of the obtained DDA strength of our approach is also evaluated using the human rated DDA pairs. Hence, a combined standard DDA dataset with human assessed scores is created using 213 disease-disease pairs obtained from UMNSRS [54] and MayoSRS [76], by mapping the concept terms to disease terms using CTD disease vocabulary [77].

4. Integrated and enhanced DDA strength quantification framework

The proposed work effectively measures the association strength between different diseases by integrating various types of disease-disease linking contextual and conceptual relations. In this work, contextual relationships are obtained from biomedical literature such as the PubMed abstracts. Similarly, biomedical databases (DO [12], UMLS [15] and biomedical text (Clinical Notes, Insurance Claims Database, Journal Articles) [20] are utilized to obtain conceptual relations. Deriving DDAs through integration of multiple linking perspectives associating the given disease pair and computing the aggregated DDA strength are important.

Figure 1 describes the proposed framework. With the list of diseases as main input, collection of associated PubMed abstracts is the first step. In Section 4.1, the proposed deep neural network model, Enhanced Sentence Embedding with Context-Based CNN (ESEC-CNN) is trained on preprocessed and labelled (positive, negative and null DD pairs) 521 PubMed abstracts [17]. The built model is further exploited to classify a new set of PubMed abstracts collected iteratively. This dataset is used to improve the general performance of DDA prediction. This dataset is used to improve the general performance of DDA prediction. The set of classified DDAs and sentence embeddings obtained from the enhanced dataset are further utilized to construct literature-based DDA matrices. In addition, the enhanced list of diseases is also used for the construction of concept-based DDA matrix of DDA representations as described in Section 4.2. Using the biomedical text and biomedical databases, Ontology-based joint multi-source association embedding model is proposed to improve concept-based DDA. The integration of literature-based and concept-based DDAs for DD association enhancement is described in Section 4.3 using a modified vector-similarity fusion method [78] to improve the quality of integrated disease vector. Finally, the relatedness score between DDs is calculated using cosine similarity of the integrated disease vector [79].

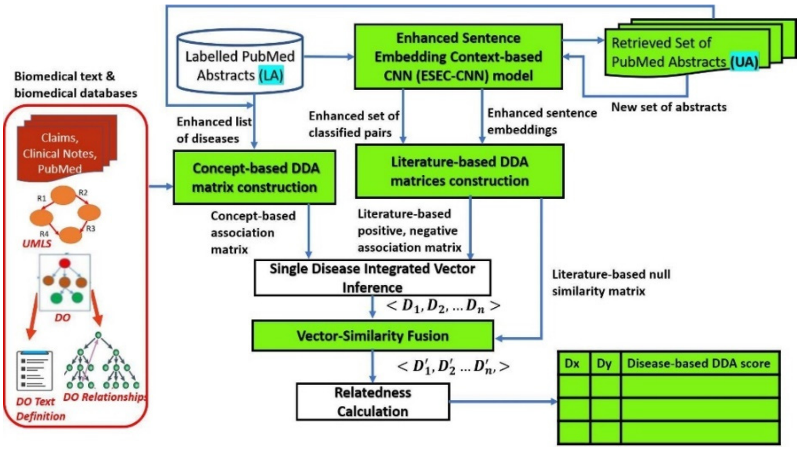


Figure 1. The proposed framework for calculating DDA.

4.1. Enhanced Literature-Based DDA

4.1.1. Enhanced sentence embedding context-based CNN

The DDA dataset derived from initial 521 labelled abstracts are used for construction of enhanced literature-based DDA matrices using sentences with disease pairs classified into positive, negative and null pairs. For this classification, we proposed a neural network architecture as illustrated in Figure 2. The network is designed to capture syntactic and semantic information for a given sentence with DD pairs from three different perspectives using

1) Sentence-based local-level features

At sentence-level, we have used Parts-of-Speech (POS) feature using one-hot encoding scheme represented by 11-bit binary vector [35] and two-dimensional disease distance feature [17]. For DDA, new additional features such as dependency relations [80] and chunk [81] are included and Named Disease Entity (NDE) feature is obtained, similar to the work of Peng and Lu [35]. The NDE feature is applied to each word in a sentence represented by a four dimensional encoding < D1, D2, D, O >, where D1 and D2, represents the disease pair under consideration. Other disease words and non-disease words are represented by D and O respectively.

2) Sentence-based global-level features

Using a popular embedding model word2vec [82], the embedding of each word in a sentence is learnt at corpus-level using both domain-specific context such as PubMed and PMC and general contexts including news, in addition to Wikipedia [83].

3) Document-level features

Similar to the work of Lai et al. [17], the traditional document features such as Bag-Of-Word, word-based Parts of Speech, NDE information and document-based information are represented using one-hot encoding.

Thus, in this work, an enhanced sentence embedding with additional features is framed that helps the proposed classification model in better classification of different types of association.

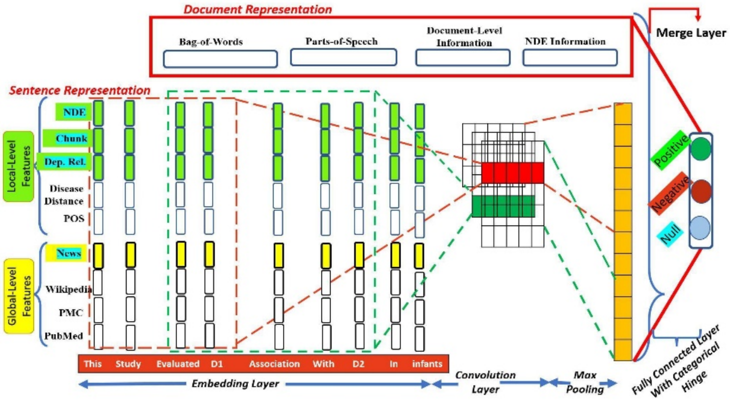


Figure 2. ESEC-CNN) architecture NDE-Named Disease Entity, POS- Parts-of-Speech, Dep. Rel.-dependency relation.

In Figure 2, the input to ESEC-CNN is the embedding layer representing the sentence followed by convolution and pooling layers outputting an n-dimensional enhanced sentence embedding vector. Similarly, the document representation [17] of m-dimension is merged with enhanced sentence embedding to create (n + m) dimensional final single vector. The fully connected layer with categorical hinge loss in activation function [84] is applied to the obtained merged vector. The combined vector is further passed on to three-dimensional output layer representing the probability of classes: positive, negative, null.

4.1.2. Enhanced construction of literature-based DDA matrices

The trained classifier model is effectively utilized in our work to classify the new set of extracted PubMed abstracts. In order to improve the performance of DDA strength calculation, it is essential to widen the range of positive, negative and null contexts of DD pairs, therein, aggregating the contextual information contribution to the DD strength during the construction of enhanced literature-Based DDA matrix. Further, the number of seed diseases is also increased, thus we attempt to measure the strength of association between a larger number of DD pairs. The dataset is constructed by an iterative technique with initial 213 seed DD pairs collected from a combined benchmark datasets including UMNSRS Similarity and Relatedness [54], MayoSRS and MiniMayoSRS between Medical term pairs [76], until we obtain 58,980 unique DD pairs.

In order to effectively quantify DDA strength using literature, considering positive, negative and null associations is important as each type conveys different degrees of association. Hence, the DDA classes (positive, negative and null) predicted by LC-CNN model along with improved sentence representations are further utilized to construct two literature-based DDA matrices namely, literature-based positive, negative DDA matrix of DDA representations and literature-based null similarity matrix.

1) Literature-based positive, negative DDA matrix

As discussed in Section 2.1, sentence-based biomedical associations are classified into only positive, negative [17,34–37] or only as negative [36]. While during the strength calculation, O'Shea [18] and Xu et al. [52] considered only positively correlated pairs. However, it is important to calculate the strength of association of pairs that occur in both positive and negative contexts and those that occur only in negative context. Considering the above aspects, cumulative association strength is calculated in Eq (1).

$$LV_{xy} = \begin{Bmatrix} n_{pos} * \sum_{i=1}^{l}(D_x - D_y)TDVector_i \\ (-) \\ n_{neg} * \sum_{j=1}^{m}(D_x - D_y)TDVector_j \end{Bmatrix} \tag{1}$$

where: LV_{xy} represents association vector of disease pair $D_x - D_y$, n_{pos} and n_{neg} is the number of positive contexts and negative contexts respectively. $TDVector_i$ and $TDVector_j$ denote enhanced sentence representations with two disease mentions vector in positive and negative cases respectively.

The association strength of disease pair $D_x - D_y$, is dealt differently if it falls in any of the three cases.

Case 1 $n_{pos} * \sum_{i=1}^{l}(D_x - D_y)TDVector_i$, strengths the DDA if $D_x - D_y$ occurs only in positive contexts.

Case 2 $n_{neg} * \sum_{j=1}^{m}(D_x - D_y)TDVector_j$, identifies negative association strength if $D_x - D_y$ occurs only in negative context.

Case 3 Eq (1) combines case 1 and case 2 using an association modification factor (-) that modifies association strength if $D_x - D_y$ occurs in both positive and negative contexts.

2) Literature-based null similarity matrix

Though Rakhi et al. [37] has classified sentence-based biomedical entity pairs as null, these associations were not considered while calculating the strength of association. However, null pairs with unmentioned associations may also be associated with some strength and hence needs to be taken into consideration. In addition, in this work, we have also extended the concept of null association within same sentence [17,34,37] to across different sentences having single disease mention and therefore, including corresponding embedding information also contributes to DDA strength computation. Accordingly, we have derived an equation Eq (2) representing a disease vector.

$$LV(D_x) = \left(\sum_{i=1}^{N} TDVector_{D_x - D_i}\right) + \left(\sum_{j=1}^{M} ODVector_{D_x - D_j}\right) \tag{2}$$

where: $LV(D_x)$ denote the disease vector representation of disease D_x , $TDVector_{D_x - D_i}$ and $ODVector_{D_xj}$ denote two-disease and single disease mention enhanced sentence representations.

The represented disease vector $LV(D_x)$, consists of 2 important components in the context of DDA as follows:

- $\sum_{i=1}^{N} TDVector_{D_x - D_i}$, accumulates enhanced sentence representations of D_x when it occurs in the same sentence with all other unmentioned or null associated diseases.
- $\sum_{j=1}^{M} ODVector_{D_xj}$, accumulates enhanced sentence representations of D_x when it occurs as single disease mention in sentences.

$LV(D_y)$ is calculated in the same way and $D_x - D_y$ strength is calculated using cosine similarity, $\cos\left(LV(D_x), LV(D_y)\right)$, that helps modify DDA with null associations and discover DDAs that are not directly conveyed by positive/negative associations.

Using Eqs (1) and (2) described in 1) and 2), we are able to construct an enhanced literature-based positive, negative DDA matrix and literature-based null similarity matrix shown in Figure 3. that is later used to calculate literature-based DDA strength.

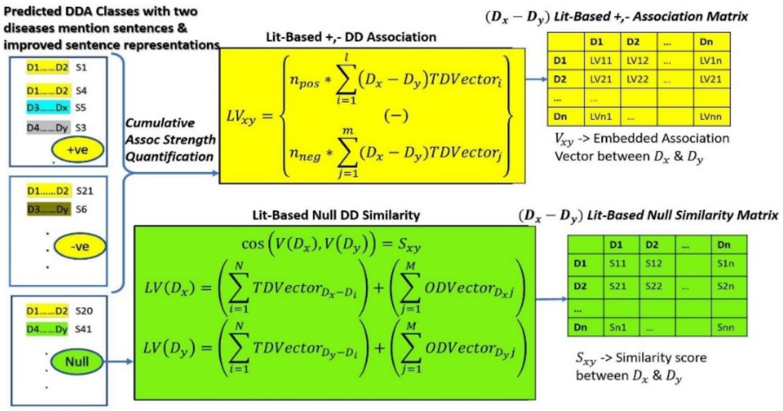

Predicted DDA Classes with two diseases mention sentences & improved sentence representations
D1......D2 S1
D1......D2 S4
D3......Dx S5
D4......Dy S3
+ve
D1......D2 S21
D3......Dy S6
-ve
D1......D2 S20
D4......Dy S41
Null
Cumulative Assoc Strength Quantification
Lit-Based +,- DD Association
$LV_{xy} = \left\{ n_{pos} * \sum_{i=1}^{l} (D_x - D_y) TDVector_i \; (-) \; n_{neg} * \sum_{j=1}^{m} (D_x - D_y) TDVector_j \right\}$
$(D_x - D_y)$ Lit-Based +,- Association Matrix
D1
D2
...
Dn
LV11
LV12
LV1n
LV21
LV22
LVn1
LVnn
V_{xy} -> Embedded Association Vector between D_x & D_y
Lit-Based Null DD Similarity
$\cos\left(V(D_x), V(D_y)\right) = S_{xy}$
$LV(D_x) = \left(\sum_{i=1}^{N} TDVector_{D_x - D_i}\right) + \left(\sum_{j=1}^{M} ODVector_{D_x j}\right)$
$LV(D_y) = \left(\sum_{i=1}^{N} TDVector_{D_y - D_i}\right) + \left(\sum_{j=1}^{M} ODVector_{D_y j}\right)$
$(D_x - D_y)$ Lit-Based Null Similarity Matrix
S11
S12
S1n
S21
S22
S2n
Sn1
Snn
S_{xy} -> Similarity score between D_x & D_y

Figure 3. Literature-based matrices with association vector LV_{xy} and similarity score S_{xy}.

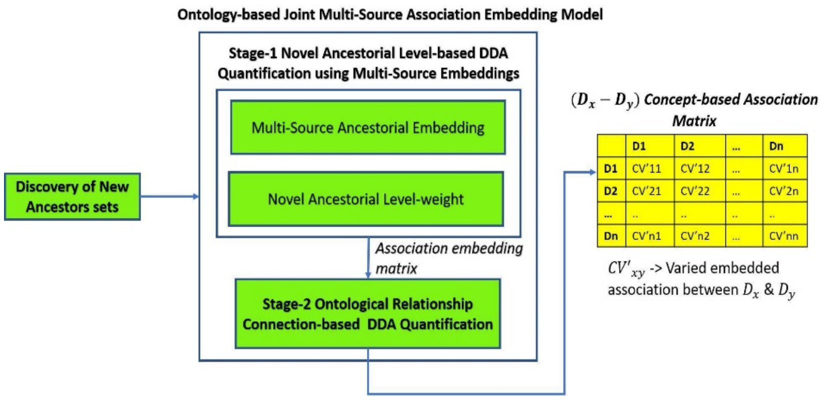
4.2. Concept-based DDA using ontology-based joint multi-source embedding model

In order to integrate conceptual aspects for DDA calculation, a detailed ontological mapping covering a wide range of taxonomic relationships, plays a vital role and contributes to the quantification of semantic associations between diseases. Some of the taxonomical ontological relationships include ancestorial parent-child relationship and other relationships like sibling and indirect relationships (uncle, cousin). Wang et al. [63] has not considered the semantic relationship in disease association measurement while only parent-child relationship is considered in the prediction of onset of diseases [85,86]. For DDA, in this work, we consider ancestorial and other closely related taxonomical relationships to derive a better degree of association linking diseases. Given DO as a DAG, having nodes corresponding to the ancestors and disease concepts $D_x \; and \; D_y$ involved in $D_x - D_y$ association, the ancestorial relationship and ontological relationship connection between D_x (disease concept1), D_y (disease-concept2) are used to learn the association representation.

For DDA measurement, when we embed each disease (concept), we need to do so in relation to a disease pair. For this, the connectedness of concepts [68] and semantic information of all ancestors are used [29,85,86]. However, discovering new ancestors sets New_Anc_Set, prior to association representation is important as not all ancestors contribute to the final association.

After discovering the ancestors sets, we introduce a 2-stage DDA quantification, ontology-based joint multi-source association representation, shown in Figure 4. In stage-1, we have included the association effect of the influential factors by infusing multi-source semantic (DO, UMLS) and contextual information (clinical notes, insurance claims, journal articles) of ancestors including the root ancestor node and leaf node. In addition, we add novel level-weight to the multi-source ancestorial representation, where the level-weight is based on new ancestors sets New_Anc_Set discovered initially, thus producing an association embedding matrix. In stage-2, we introduce ontological relationship connection-based DDA quantification that varies the embedded association strength between diseases based on their type of relation connection in the ontology, thus resulting in concept-based association matrix of DDA representations.

Thus, in this work, we try to improve the concept-based DDA by constructing a concept-based DDA matrix of DDA representations using ontology-based joint multi-source association embedding model as shown in Figure 4.


Ontology-based Joint Multi-Source Association Embedding Model
Stage-1 Novel Ancestorial Level-based DDA Quantification using Multi-Source Embeddings
Multi-Source Ancestorial Embedding
Novel Ancestorial Level-weight
Discovery of New Ancestors sets
Association embedding matrix
Stage-2 Ontological Relationship Connection-based DDA Quantification
$(D_x - D_y)$ Concept-based Association Matrix
CV'_{xy} -> Varied embedded association between D_x & D_y

	D1	D2	...	Dn
D1	CV'11	CV'12	...	CV'1n
D2	CV'21	CV'22	...	CV'2n
...	..	..	..	..
Dn	CV'n1	CV'n2	...	CV'nn

Figure 4. Pipeline of concept-based DDA using proposed ontology joint multi-source association representation.

4.2.1. Discovery of new ancestors sets

As discussed in Section 4.2, including all ancestors of given disease concept may cause semantic contribution of even the concepts that are not common between diseases in the disease pair and hence, embedding of disease under consideration may lead to incorrect association. In order to tackle this aspect, that is, rather than considering all ancestors of a particular node in the ontology, we consider only those ancestors that contribute to the association between diseases by defining new sets of ancestors $New_Anc_Set(D_x)$and$New_Anc_Set(D_y)$ for D_x and D_y respectively for $D_x - D_y$ association. Therefore, the derived ancestors set $Anc\_Set(D_x)$ of disease D_x in $D_x - D_y$ association is described in Eq (3), where only common ancestors $A_i's$ are considered since two diseases are associated by sharing of common diseases in the DO. In addition, the ancestors on the longest path $A'_{jx}s$ with respect to D_x is also considered to cover a broader etiology of the disease concept.

$$New_{Anc_{Set(D_x)}} = [A_i's \in common(D_x, D_y)] + A'_{jx}s \text{ on longest path from } LCS(D_x, D_y) \tag{3}$$

where $common(D_x, D_y)$ denotes the common ancestors of D_x and D_y.

Further, by utilizing the discovered ancestors sets, ontology-based joint multi-source association embedding model is proposed, consisting of 2 stages, described in sub-sections 4.2.2 and 4.2.3.

4.2.2. Ontology-based joint multi-source association embedding model

Stage-1 Novel-ancestorial level-based DDA quantification using multi-source embeddings

Figure 5 shows the derived embedded association representation, CV_{xy} for two disease nodes in the given DO, where the representation is divided into two components, *A) Multi-source ancestorial Embedding* and *B) Novel ancestorial level-weight* for each of the diseases D_x *and* D_y respectively, discussed in following sections.

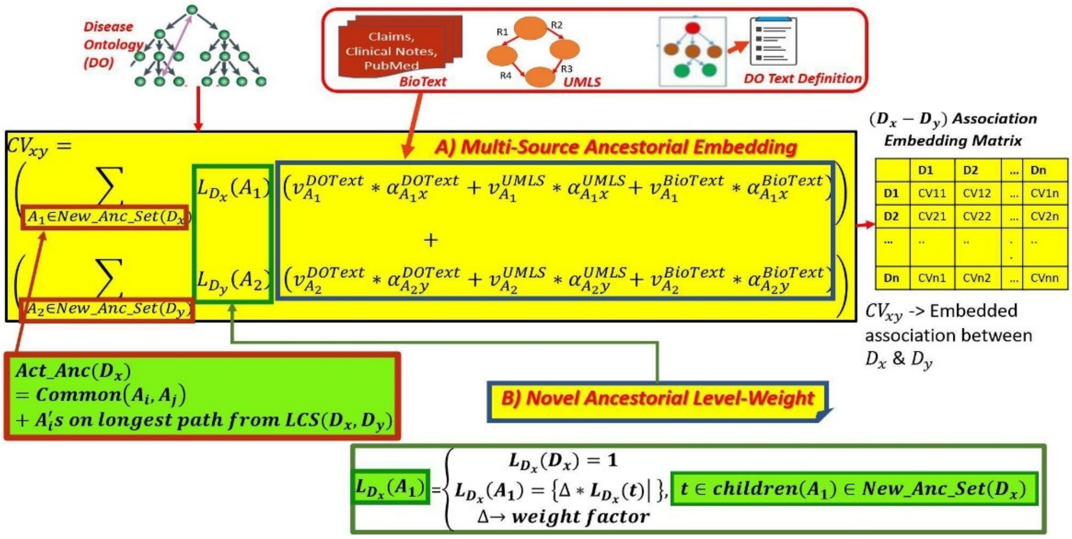

Disease Ontology (DO)
Claims, Clinical Notes, PubMed
BioText
R1
R2
R3
R4
UMLS
DO Text Definition
A) Multi-Source Ancestorial Embedding
$CV_{xy} =$
$\left(\sum_{A_1 \in New\_Anc\_Set(D_x)} L_{D_x}(A_1)\left(v_{A_1}^{DOText} * \alpha_{A_1x}^{DOText} + v_{A_1}^{UMLS} * \alpha_{A_1x}^{UMLS} + v_{A_1}^{BioText} * \alpha_{A_1x}^{BioText}\right)\right)$
$+$
$\left(\sum_{A_2 \in New\_Anc\_Set(D_y)} L_{D_y}(A_2)\left(v_{A_2}^{DOText} * \alpha_{A_2y}^{DOText} + v_{A_2}^{UMLS} * \alpha_{A_2y}^{UMLS} + v_{A_2}^{BioText} * \alpha_{A_2y}^{BioText}\right)\right)$
$(D_x - D_y)$ Association Embedding Matrix
D1
D2
...
Dn
CV11
CV12
CV1n
CV21
CV22
CV2n
CVn1
CVn2
CVnn
CV_{xy} -> Embedded association between D_x & D_y
$Act\_Anc(D_x) = Common(A_i, A_j) + A_i's$ on longest path from $LCS(D_x, D_y)$
B) Novel Ancestorial Level-Weight
$L_{D_x}(A_1) = \begin{cases} L_{D_x}(D_x) = 1 \\ L_{D_x}(A_1) = \{\Delta * L_{D_x}(t) \mid \}, \\ \Delta \rightarrow weight\ factor \end{cases}$
$t \in children(A_1) \in New\_Anc\_Set(D_x)$

Figure 5. CV_{xy}, an embedded association representation of $D_x - D_y$.

A) Multi-source ancestorial embedding

As discussed earlier in Section 4.2, Song et al. [86] considered all ancestors and included only semantic embeddings of ancestors excluding the root ancestor node and leaf node ($D_x\ in\ D_x - D_y$). However, we consider only new ancestors sets, $New\_Anc\_Set(D_x)\ and\ New\_Anc\_Set(D_y)$ as discussed in Section 4.2.1 and various conceptual knowledge of ancestors from multiple sources, since, $D_x - D_y$ $association may be influenced by several factors such as symptoms, biological entities (genes, proteins, etc.), other diseases, affected patient records, etc., which can be covered by infusing embeddings from different sources. In addition, considering multi-source information of root node and leaf node (D_x) is important in the context of DDA as root node is common to both $D_x$and D_y and leaf node D_x is involved in $D_x - D_y$ $association. As shown in Figure 5, the multi-source ancestorial embedding of $A_1 \in New_Anc_Set(D_x)$is given by the component A, in which we assign multi-source contextual embeddings $v_{A_1}^{DOText}$, $v_{A_1}^{BioText}$ from DO and biomedical text [3] and semantic embedding $v_{A_1}^{UMLS}$ from UMLS [33]. For embedding text definition from DO, in this work, we adopted the procedure used by Park et al. [23] to fill in the definition of diseases using the first lead paragraph from Wikipedia, applying an embedding method, Doc2Vec [87]. The combined semantic and contextual information is then infused into the deep neural network embedding model through attention mechanism [85,86]. The attention weights on multi-source embeddings with respect to D_x are denoted by $\alpha_{A_1x}^{DOText}, \alpha_{A_1x}^{UMLS}, \alpha_{A_1x}^{BioText}$. The weight computation for text definition embedding from DO for ancestor $A_1 \in New\_Anc\_Set(D_x)$ is computed using equation Eq (4.1) by SoftMax function as follows:

$$\alpha_{A_1x}^{DOText} = \frac{\exp\left(f_{DOText}\left(v_{A_1}^{DOText}, v_{D_x}^{DOText}\right)\right)}{w_{D_x}^{DOText}} \tag{4.1}$$

where $w_{D_x}^{DOText}$is given by Eq (4.2).

$$w_{D_x}^{DOText} = \sum_{A_k \in New\_Anc\_Set(D_x)} \left(\exp\left(f\left(v_{A_k}^{DOText}, v_{D_x}^{DOText}\right)\right) + \exp\left(f\left(v_{A_k}^{UMLS}, v_{D_x}^{DOText}\right)\right) + \exp\left(f\left(v_{A_k}^{BioText}, v_{D_x}^{DOText}\right)\right)\right) \tag{4.2}$$

where $f\left(v_{A_1}^{DOText}, v_{D_x}^{DOText}\right), f\left(v_{A_1}^{UMLS}, v_{D_x}^{DOText}\right), f\left(v_{A_1}^{BioText}, v_{D_x}^{DOText}\right)$ denotes the scalar score functions defined in Eq (4.2) to find the compatibility between text embedding of D_x from DO and multi-source ancestorial embeddings, which are computed using a single layer feed forward neural network using Eq (4.3).

$$f_{DOText}\left(v_{A_1}^{DOText}, v_{D_x}^{DOText}\right) = z^T \tanh\left(N \begin{bmatrix} v_{A_1}^{DOText} \\ v_{D_x}^{DOText} \end{bmatrix} + bias\right) \tag{4.3}$$

Z, N and bias are the learning parameters used by the neural network.
Similarly, other attention weights of ancestor A_1w.r.t D_x from other sources are calculated in similar manner. Similar kind of equations are adopted in case of ancestor A_2 w.r.t D_y.

B) Novel ancestorial level-weight

The next component of stage-1, controls the semantic and contextual contribution effect of each ancestor by adding level-weight to the aggregated multi-source embeddings obtained using component A. We used the ancestorial level-weights similar to Wang et al. [63] (relative positions in MeSH) and Kamran et al. [65]. Wang et al. [63] and Kamran and Naveed [65], calculated the ancestorial level-weight by choosing the maximum of level-weights among all children of ancestor with respect to each entity in association. This may lead to assigning level-weight of ancestor by children which may be neither common nor on the longest path to D_x and D_y, thus failing to include level-weights of nodes contributing to the association. Thus, selecting the level-weight contributed by children that are common ancestors and those that fall into longest path with respect to D_x and D_y, New_Anc_Set of D_x and D_y, reveals the actual semantic value or level-weight of ancestors. As a special case of computing level-weight of least common subsumer (LCS), Kamran et al. [65], calculated the semantic value or level-weight of LCS by considering only the level-weights of the ancestors on the longest path from root to LCS which included only the influential effect of ancestors of LCS. However, this will not help in identifying the true level weight of LCS with respect to each of the descendant entities in association. Therefore, for computing the level-weight of LCS, it is required to consider level-weights of children of LCS on deeper or longest path that connects LCS with each of its descendant entities in association as it reveals the actual semantic value of LCS. Therefore, in this work, a novel ancestorial level-weight contributing to the association strength is derived and is denoted by component B in Figure 5 and given in equation Eq (5) for ancestor A_1 w.r.t D_x.

Therefore, in this work, a novel ancestorial level-weight contributing to the association strength is derived and is denoted by component B in Figure 5 and given in equation Eq (5) for ancestor A_1 w.r.t D_x.

$$L_{D_x}(A_1) = \begin{cases} L_{D_x}(D_x) = 1 \\ L_{D_x}(A_1) = \left\{\Delta * L_{D_x}(t) \middle| t \in children(A_1) \in New\_Anc\_Set(D_x)\right\} \\ \Delta \rightarrow weight\ factor \end{cases} \tag{5}$$

where Δ is the weight factor of the edge linking A_1 with its child t. The weight factor helps reduce the contribution effect of ancestors that are distant from D_x, ranging from 0 to 1 and we found that $\Delta = 0.4$ gives better correlation with the standard DDA scores from DisGeNET. Similarly, level-weight of ancestor A_2 w.r.t D_y is derived.

Finally, the derived two components in Section 4.2.2 are then multiplied to get the final association representation, CV_{xy}, for $D_x - D_y$ association. With the derived $D_x - D_y$ association vector CV_{xy}, we further vary the association based on the connectedness ontological relationship between D_x and D_y, using an additional DDA quantification described in the following Section 4.2.3.

4.2.3. Stage-2 Ontological relationship connection-based DDA quantification

Given a disease pair $D_x - D_y$, whose association can be established through other diseases in the ontology using ancestorial relationship without considering the variation factor is discussed in Section 4.2.2. However, the type of ontological relationship connection between $D_x \, and \, D_y$, reveals the actual association. Hence, varying the association based on type of the relationship connection, provides a finer adjustment to the already derived association vector CV_{xy}. Therefore, in this work, we proposed an ontological relationship variation factor (ORVF) for the second level of DDA quantification.

As a diagrammatic illustration, ORVF values for different types of ontological relationship connections are shown in Figure 6.

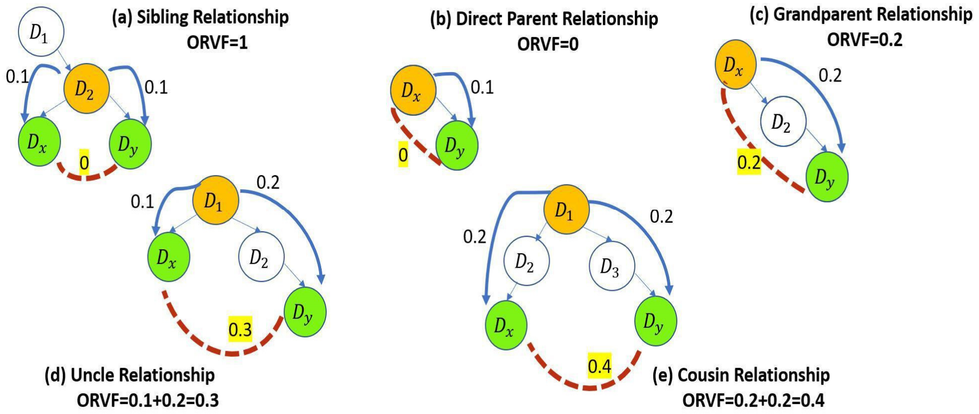


Figure 6. ORVF calculation for different types of ontological relationship connections between $D_x \, and \, D_y$.

In Figure 6(a), the ORVF is 0 when both $D_x \, and \, D_y$ are at same distances 0.1 or immediate children of D_2, considering the edge weight as 0.1. Similarly, in Figure 6(b), the ORVF is 0 as $D_x$is the direct parent of D_y, with a distance 0.1. Thus, ORVF 0, represents that there is no variation of association when $D_x \, and \, D_y$ are very closely related as a sibling and direct parent-child relationships. However, the variation occurs when $D_x \, and \, D_y$ are distantly related. For example, the ORVF values are calculated for the indirect relationships shown in Figure 6(c) (d) and (e). In Figure 6(c), $D_x$acts as grandparent of D_y, producing ORVF 0.2 as D_y is at a distance of 0.2 from D_x, while an uncle relationship connection in Figure 6 (d), calculated ORVF of 0.3as an aggregation of distances 0.1 and 0.2 with respect toD_x and $D_y$respectively, from LCS(D_x,D_y) (D_1). On the other hand, in Figure 6(e), D_x acts as a cousin of $D_y$resulting in ORVF of 0.4 as both D_x and D_y are at distance 0.2 from LCS (D_x,D_y) (D_1). Thus, ORVF helps in varying the extent of DD association by each Ds independent distance from LCS.

Algorithm 1 summarizes the procedure of adjusting the stage-1 association vector CV_{xy} by the proposed ORVF is as follows.

Algorithm 1 CV_{xy} adjustment by ORVF

```
$path(D_x) \leftarrow ordered\_New\_Anc\_Set(D_x)$, $path(D_y) \leftarrow ordered\_New\_Anc\_Set(D_y)$,
$LCS_{D_k} \leftarrow Least\ Common\ Subsumer(path(D_x), path(D_y))$
$Compute\ \{Level(LCS_{D_k-D_x}), Level\left(LCS_{D_k-D_y}\right)\}$
Compute ORVF:
Case 1: Direct parent/Siblings
If $Level(LCS_{D_k-D_x}) == Level\left(LCS_{D_k-D_y}\right) == 1$
    $CV'_{xy} \leftarrow CV_{xy}$, ORVF=0 i.e. No Variation
Case 2: Broader/Indirect Relationship
If $(Level(LCS_{D_k-D_x}), Level\left(LCS_{D_k-D_y}\right)) \neq 1$
    $CV'_{xy} \leftarrow CV_{xy} *$ ORVF, ORVF$\leftarrow \left[Level(LCS_{k-D_x}) + Level\left(LCS_{k-D_y}\right)\right]$
```

An illustration of the above algorithm is given in Figure 7(a) and (b) showing the ORVF calculations for sibling and cousin ontological relationships connecting D_x *and* $D_y$respectively.

Figure 7(a) and (b) follows the same procedure to compute ORVF. The first step gives the $LCS(D_x, D_y)$ denoted as LCS_k, by defining the $path(D_x)$ and $path(D_y)$ using the new ancestors sets of D_x and D_y respectively, where LCS_k is equal toD_2 and D_1 corresponding to Figure 7(a) and (b). The next step is to find the distance of LCS_k from D_x and D_y independently using $Level(LCS_{k-D_x})$ and $Level\left(LCS_{k-D_y}\right)$ and found to be 0.1 for sibling relationship in Figure 7(a) and found to be of different distances 0.1 and 0.2 for cousin relationship in Figure 7(b). Finally, with the calculated distances, the ORVF is computed for direct/sibling relationships in Figure 7(a) and for broader/indirect relationships in Figure 7(b). For direct/sibling relationship, the association embedding is not varied since ORVF is 0 whereas the association embedding is reduced by a factor of 0.3 which is the total distance of variation between D_x *and* D_y, through LCS_k . Hence, the association embedding CV_{xy} is the final association embedding CV'_{xy} in case of sibling relationship connection in Figure 7(a) whereas CV_{xy} is reduced by a factor of 0.3 contributed by 0.1 and 0.2 from LCS_k from D_x *and* D_y respectively.

Using CV_{xy}as shown in Figure 4 and the proposed ORVF, we are able to construct an enhanced concept-based DDA matrix of DDA representations CV'_{xy} that is later used for concept-based DDA strength.

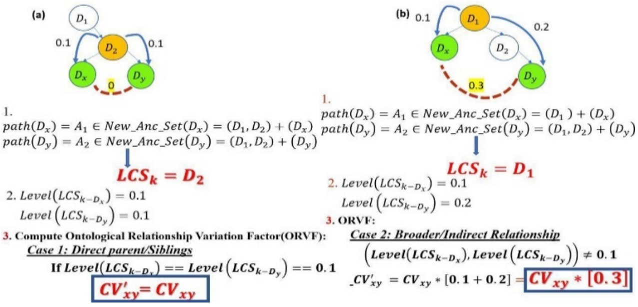
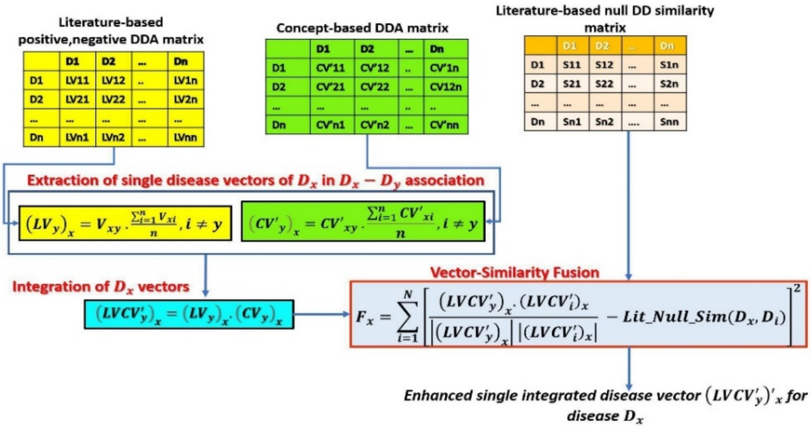

(a)
D_1
0.1
D_2
0.1
D_x
D_y
0
1.
$path(D_x) = A_1 \in New\_Anc\_Set(D_x) = (D_1, D_2) + (D_x)$
$path(D_y) = A_2 \in New\_Anc\_Set(D_y) = (D_1, D_2) + (D_y)$
$LCS_k = D_2$
2. $Level(LCS_{k-D_x}) = 0.1$
$Level\left(LCS_{k-D_y}\right) = 0.1$
3. Compute Ontological Relationship Variation Factor(ORVF):
Case 1: Direct parent/Siblings
If $Level(LCS_{k-D_x}) == Level\left(LCS_{k-D_y}\right) == 0.1$
$CV'_{xy} = CV_{xy}$
(b) 0.1
D_1
0.2
D_x
D_2
D_y
0.3
1.
$path(D_x) = A_1 \in New\_Anc\_Set(D_x) = (D_1\ ) + (D_x)$
$path(D_y) = A_2 \in New\_Anc\_Set(D_y) = (D_1, D_2) + (D_y)$
$LCS_k = D_1$
2. $Level(LCS_{k-D_x}) = 0.1$
$Level\left(LCS_{k-D_y}\right) = 0.2$
3. ORVF:
Case 2: Broader/Indirect Relationship
$\left(Level(LCS_{k-D_x}), Level\left(LCS_{k-D_y}\right)\right) \neq 0.1$
$\_CV'_{xy} = CV_{xy} * [0.1 + 0.2] = CV_{xy} * [0.3]$

Figure 7. Adjusting association vector CV_{xy} by the proposed ORVF for sibling relationship (left) and cousin relationship connection (right).

4.3. Integration and Enhancement of final disease vector representation

Finally, an information rich single disease vector of D_x in $D_x - D_y$ Association, can be obtained as shown in Figure 8, by the following steps. Extracting literature-based D_x vectors, from the constructed literature-based positive, negative DD association matrix of LV_{xy} and concept-based D_x vectors from concept-based DD association matrix of CV'_{xy} as discussed in Sections 4.1.2 and 4.2. Further, the extracted D_x vectors are integrated into single integrated disease vector. As an enhancement to final DDA strength, the integrated single disease vector is enhanced with additional contextual information obtained from literature-based null DD similarity matrix in Section 4.1.2, using vector-similarity fusion method, in order to obtain the final DDA strength.


Literature-based positive,negative DDA matrix
Concept-based DDA matrix
Literature-based null DD similarity matrix
Extraction of single disease vectors of D_x in $D_x - D_y$ association
Integration of D_x vectors
Vector-Similarity Fusion
Enhanced single integrated disease vector $(LVCV'_y)'_x$ for disease D_x

	D1	D2	...	Dn
D1	LV11	LV12	..	LV1n
D2	LV21	LV22	...	LV2n
...	...	...	...	...
Dn	LVn1	LVn2	...	LVnn

	D1	D2	...	Dn
D1	CV'11	CV'12	..	CV'1n
D2	CV'21	CV'22	...	CV12n
...	...	...	..	..
Dn	CV'n1	CV'n2	...	CV'nn

	D1	D2	...	Dn
D1	S11	S12	...	S1n
D2	S21	S22	...	S2n
...	...	...	...	...
Dn	Sn1	Sn2	....	Snn

$(LV_y)_x = V_{xy} \cdot \frac{\sum_{i=1}^{n} V_{xi}}{n}, i \neq y$

$(CV'_y)_x = CV'_{xy} \cdot \frac{\sum_{i=1}^{n} CV'_{xi}}{n}, i \neq y$

$(LVCV'_y)_x = (LV_y)_x \cdot (CV_y)_x$

$F_x = \sum_{i=1}^{N} \left[\frac{(LVCV'_y)_x \cdot (LVCV'_i)_x}{\left|(LVCV'_y)_x\right| \left|(LVCV'_i)_x\right|} - Lit\_Null\_Sim(D_x, D_i) \right]^2$

Figure 8. Integration and enhancement of final disease vector representation.

4.3.1. Extraction of single disease vector

For $D_x - D_y$ association, literature-based single disease vector $(LV_y)_x$ of D_x with respect to D_y, is extracted using association vectors obtained from literature-based positive, negative association matrix in Eq (1) by averaging the literature based DDA vectors $LV_{xi}'s$ of $D_x - D_i$ associations, where $i \leftarrow \{1,2,..,n\}$ and $i \neq y$ and finally concatenating the averaged component with association vector LV_{xy} of $D_x - D_y$ association as shown in Eq (6). For $D_x - D_y$ association, it is important to preserve the actual information component of D_y through concatenation while representing D_x vector. Similarly, single disease vector for D_y is extracted. A similar strategy is followed while extracting disease vector for D_x and D_y from concept-based DDA matrix, where $(CV_y)_x$ of D_x with respect to D_y is shown in Eq (7).

$$(LV_y)_x = LV_{xy} \cdot \frac{\sum_{i=1}^{n} LV_{xi}}{n}, i \neq y \tag{6}$$

$$(CV'_y)_x = CV'_{xy} \cdot \frac{\sum_{i=1}^{n} CV'_{xi}}{n}, i \neq y \tag{7}$$

where: LV_{xy} and CV_{xy} are literature-based and concept-based association vector of $D_x - D_y$ pair. $(LV_y)_x$ represents literature-based single disease vector of D_x with respect to D_y. Similarly, $(CV'_y)_x$ represents concept-based single disease vector of D_x with respect to D_y.

4.3.2. Integration of single disease vectors

For an information-enriched disease vector representation, the extracted literature-based and concept-based single disease vectors are integrated into a single information rich disease vector. However, for disease vector representation, only a narrow disease-disease linking relations were fused [32,33]. In order to achieve better association, in this work, the disease vector is represented by integrating vector representations on a wide range of disease-disease linking information from both literature and concept-based biomedical data sources.

Thus, for an information-enriched representation of diseases in $D_x - D_y$ association, the extracted literature-based and concept-based disease vector components in Eqs (6) and (7), respectively, are concatenated into a single integrated disease vector $\left(LVCV_y'\right)_x$ for D_x with respect to D_y as in Eq (8).

$$\left(LVCV_y'\right)_x = \left(LV_y\right)_x . \left(CV'_y\right)_x \tag{8}$$

where $\left(LVCV_y'\right)_x$ represents the single integrated disease vector D_x with respect to D_y. $\left(LV_y\right)_x$ represents literature-based single disease vector of D_x with respect to D_y. $\left(CV'_y\right)_x$ represents concept-based single disease vector of D_x with respect to D_y. Similarly, $(LVCV_x')_y$ for D_y with respect to D_x can be defined using Eq (8).

4.3.3. Enhancement to the integrated disease vector

In addition, the information-enriched integrated disease vector is enhanced with additional contextual relationship with all other diseases obtained from literature-based DD null similarity matrix derived earlier in as discussed in Section 4.1.2. Manchanda and Anand [78] enhanced the disease vector representation by updating the initial vector representation using only literature (PubMed) with the corresponding similarity information with all other diseases. Enhancing such a low informative disease vector with similarity is needed to produce a proper enhanced disease vector. Hence, in this work, we use the information-enriched integrated disease vector derived in Eq (8) as an initial vector for similarity updation using vector-similarity fusion method defined in Eq (9), that uses an objective function [rep learning paper], where the scalar component is replaced by the null similarity scores.

Thus, the enhanced integrated vector $\left(LVCV_y'\right)'_x$ for D_x with respect to D_y in $D_x - D_y$ association is obtained from $\left(LVCV_y'\right)_x$ in Eq (8) when updated if the objective function F_x is minimized as shown in Eq (9)

$$F_x = \sum_{i=1}^{N} \left[\frac{\left(LVCV_y'\right)_x . \left(LVCV_i'\right)_x}{\left|\left(LVCV_y'\right)_x\right| \left|\left(LVCV_i'\right)_x\right|} - Lit_{Null_{Sim(D_x,D_i)}} \right]^2 \tag{9}$$

where $\left(LVCV_y'\right)_x$ represents the integrated disease vector D_x with respect to D_y, $Lit_{Null_{Sim(D_x,D_i)}}$ denotes the literature-based null similarity scores between D_x and D_i, $\left|\left(LVCV_y'\right)_x\right|$ denote length of vector $\left(LVCV_y'\right)_x$, Similarly, the enhanced integrated vector $(LVCV_x')'_y$ for $D_y$with respect to D_x in $D_x - D_y$ association is updated when the objective function F_y is minimized.

Thus, a rich integrated and enhanced disease vector representation is derived that helps DDA both contextually and semantically, leading to a better quality of final DDA Strength.

4.4. Relatedness calculation

Finally, with the enhanced-integrated disease vector representations obtained in Section 4.3.3, a cosine similarity is applied to obtain the final score measuring the actual strength of association for the given disease pair as shown in Eq (10).

$$Assoc\_Score(D_x, D_y) = \cos\left(\left(LVCV_y'\right)'_x, (LVCV_x')'_y\right) \tag{10}$$

where $\left(LVCV_y'\right)'_x$ and $(LVCV_x')'_y$ represent enhanced integrated disease vector D_x with respect to D_y and $D_y$with respect to D_x respectively.

Therefore, in this section, instead of finding the embedding vector for a disease in isolation, we used a modified method similar to Manchanda and Anand [78], in which the disease embedding is discovered in relation with DDA. We used an integration of literature-based and concept-based conceptual and semantic multi-source embeddings and richer ontological embeddings to obtain and discover DD associations and derive their strengths.

5. Results and discussion

For evaluating the enhanced DDA framework, we first evaluate the performance of the proposed association classification model ESEC-CNN with improved sentence representation, which on training facilitated the construction of enhanced DDAE dataset. The classification model was evaluated by measuring the model's classification performance using Precision, Recall and F-measure. The correlation between the association scores obtained from the enhanced literature-based DDA representations and the association metrics Wang et al. [24], Resnik [25], Schlicker et al. [88] and Lin [26] is evaluated using spearman's rank correlation coefficient. Second, the enhanced concept-based DDA representations is evaluated on both established biomedical dataset DisGeNet and human-rated DDA datasets using spearman's rank correlation coefficient. Third, the evaluation of single disease vector representation is carried out using literature and concept-based approaches independently and using the integration of both in a similar manner. Finally, the quantification of DDA pairs obtained using the enhanced single disease vector representation is compared to the state-of-art methods and evaluated in different perspectives of DDA criteria. Additionally, we have also shown the biological effect of the DDA scores derived by integrated and enhanced disease vector representation for mostly associated disease pairs category-wise.

5.1. Literature-based DDA evaluation

5.1.1. Evaluation of improved sentence representation

We conducted experiments to show the effect of additional features in sentence representation using classification performance of various sentence classification models in Table 1 and also in Figure 9. DDA classification performance of the baseline models without (limited local and global-level features) and with (additional local and global-level features) improved sentence representation such as, LSTM [49], BiLSTM [89], CNN [90], BERT [91], BioBERT [92] and LC-CNN [17] are then evaluated on the available annotated DDA dataset, on a 5-fold cross validation. Implementation is carried out on a TensorFlow with hyperparameters of learning rate as 0.025, batch size of 8, epochs of 5,10,15 and layer size of 352.

On comparing with all classification models, CNN-based models are found to perform better as LSTM, BiLSTM are sequence-based and hence, CNN-based model shows better sentence classification performance.

The LC-CNN model with additional news embedding feature (global-level) has shown only less improvement of F-measure than that of LC-CNN with limited features. With the combined additional local-level embeddings of NDE, dependency relation, chunk tag along with other global-level embeddings including news, ESEC-CNN model (LC-CNN model with improved sentence representation) outperformed the other baseline models including LC-CNN model without improved sentence representation with F-measure of 85.54%.

A notable observation of F-measure in other baseline models show that models have achieved better F-measure when the sentence representation is improved with additional local and global level features. Hence, the effect of improved sentence representation has a major positive effect on other models also.

Table 1. Performance of improved sentence representation with different classification models.

Methods	Without improved sentence representation			With improved sentence representation		
	Performance measure					
	Precision (%)	Recall (%)	F-measure (%)	Precision (%)	Recall (%)	F-measure (%)
LSTM [49]	65.13	67.11	66.11	66.92	68.37	67.64
BiLSTM [89]	64.88	66.15	65.51	65.15	68.02	66.55
CNN [90]	74.13	71.27	72.67	75.20	72.64	73.90
BERT [91]	78.65	80.32	79.48	80.63	82.12	81.37
BioBERT [92]	81.54	82.01	81.77	82.69	83.88	83.28
LC-CNN [17]	82.16	84.89	83.50	–	–	–
ESEC-CNN*	–	–	–	83.06	86.54	84.76
ESEC-CNN**	–	–	–	84.03	87.12	85.54

*- partial improved sentence representation{(PubMed,PMC,Wiki,News),(POS,NE Dist)}

**- improved sentence representation {(PubMed,PMC,Wiki,News),(POS,position,Dep. Rel.,Chunk,NE)}

The better performing ESEC-CNN model (LC-CNN with improved sentence representation) is further utilized for DDA dataset expansion, where the size of the labelled PubMed abstracts is increased using an initial 213 seed DD pairs obtained from a combined benchmark similarity dataset as discussed in Section 4.1.2.

From PubTator, a set of abstracts are downloaded in BioCXML format from https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/PubTatorCentral_BioCXML/BioCXML.9.tar (accessed 12 July 2022), ensuring only abstracts that contain sentences with the given DD pairs are retrieved. At each iteration, a new unique set of DD pairs are produced from the retrieved set of abstracts. The number of newly produced DD pairs are found to increase at the initial few iterations and the drop in the count of new DD pairs acts as a stopping criterion for the abstracts retrieval process.

With the retrieved 39,510 abstracts, a total of 58,980 unique DD pairs are identified. However, for the construction of increased DDA extraction (DDAE) dataset, the LC-CNN model with improved sentence representation is trained on the available labelled abstracts [17] and then applied on to the created dataset. The trained model is able to identify a large number positive, negative and null pairs with only a minimum number of seed pairs. A statistical comparison of the enhanced constructed DDA extraction (DDAE) dataset starting with the available 521 labelled DDAE dataset [17] is tabulated in Table 2.

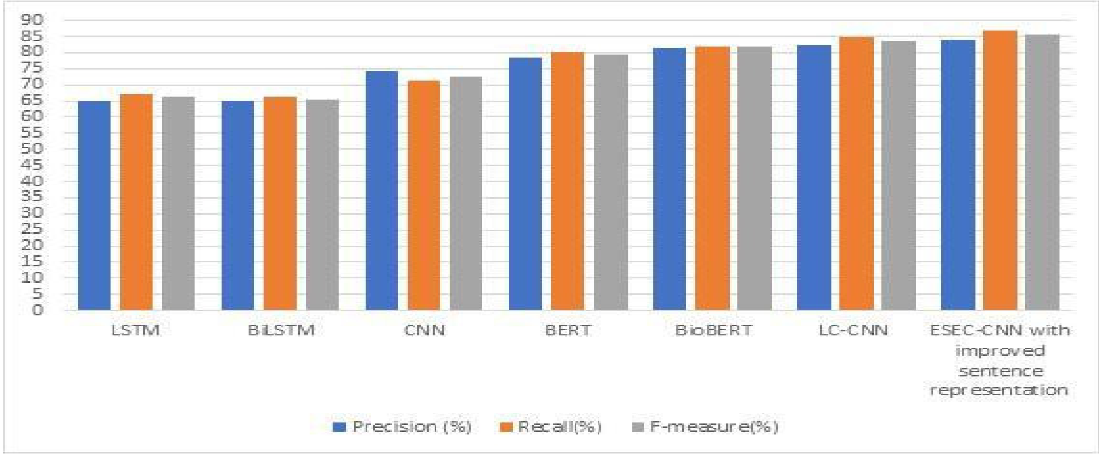


Figure 9. DDA classification performance of baseline models without improved sentence representation and proposed ESEC-CNN model with improved sentence representation.

Table 2. Statistics of the available and constructed DDAE dataset.

Details	Available 521 labelled DDAE dataset [17]	Constructed DDAE dataset
Abstracts	521	39,510
Unique Ds	1103	14,598
Unique DD pairs	3600	28,980
Unique Positive DD pairs	1626	34,481
Unique Negative DD pairs	124	5488
unique Null DD pairs	2649	36102
Unique Positive-Negative DD pairs	53	3254
Unique Positive-Negative-Null DD pairs	33	2589

5.1.2. Evaluation of different types of classified DDAs with enhanced literature-based DDA representation

DD pairs classified by ESEC-CNN model are of 3 types, namely, both positively and negatively associated, only negatively associated and null associated and their association scores are validated as discussed earlier in this section and the evaluation of the 3 types is shown in Tables 3–5 respectively. The association measures are calculated using DOSim package [5]. Further, the concordance of the classified DD pairs scores with each of the association metrics is evaluated on both 521 DDA labelled abstracts [17] and the constructed DDAE dataset.

For both positively and negatively associated DD pairs, as shown in Table 3, [47] and [18] derived DDA strength which are less correlated with all metrics when evaluated on both datasets with a count of 54 and 3254 DD pairs. The lower correlation is because, Sicknessminer considered the number of co-mentions ignoring the context and treated all co-mentioned pairs as equally contributing to DD association, while Gextext considered a direct positive association if the DD pair had an average occurrence in the whole corpus, thus missing out the negative context of the pairs. Hence, considering negative context for association quantification will balance the real context by which disease pairs are associated. Further, such consideration could lead to significant correlation achieved by association scores computed using literature-based positive, negative association matrix.

Table 3. Spearman's rank correlation between enhanced literature-based positive, negative DD association matrix and DO-based similarity metrics (Wang, Resnik, Relevance, Lin) for both positively and negatively associated DD pairs from different sets of labelled DDA dataset.

Association Type	Method	Wang et al. [24]	Resnik [25]	Schlicker et al. [88]	Lin [26]
#positively and negatively associated DD pairs = 54 [available 521 labelled dataset [17]]	GloVe-50 [93]	0.001	0.007	0.015	0.015
	SicknessMiner [47]	0.277131	0.203487	0.340086	0.340086
	GexText [18]	0.3982	0.3884	0.394	0.4013
	Enhanced literature-based positive,negative DDA representation	0.402411	0.417671	0.531724	0.531724
#positively and negatively associated DD pairs= 3254 [Enhanced DDA dataset]	GloVe-50 [93]	0.005	0.009	0.024	0.026
	SicknessMiner [47]	0.323	0.321	0.352	0.349
	GexText [18]	0.468	0.463	0.476	0.476
	Enhanced literature-based positive,negative DDA representation	0.515	0.521	0.575	0.573

In case of only negatively associated DD pairs as shown in Table 4, a total of 70 and 2234 DD pairs were found from the available and enhanced DDAE dataset, respectively, where their derived scores from the literature-based DDA matrix are positively correlated while other literature-based scores are negatively correlated indicating that considering the context of DD pairs occurrence plays a crucial role rather than taking only their occurrence frequency as in other methods.

Similarly, the associations discovered for 2649 (521 abstracts) and 36102 (enhanced dataset) null pairs from literature-based null similarity matrix, have also correlated better when compared to other methods shown in Table 5, as only few pairs co-occur and therefore Sicknessminer [47] which used the co-mention analysis for association, is less correlated. While GexText [18], resulted in strong association for DD pairs with higher occurrence in the corpus which may not be strongly associated and hence less correlated compared to our null similarity scores, as the null scores obtained, considered the surrounding context influencing the disease in the given pair. On the other hand, GloVe [93] generated less informative embeddings for association calculation and therefore less correlated in all the above cases.

Table 4. Spearman's rank correlation between enhanced literature-based positive, negative DD association matrix with DDA representation and DO-based similarity metrics (Wang, Resnik, Relevance, Lin) for only negatively associated DD pairs from different sets of labelled DDA dataset.

Association Type	Method	Wang et al. [24]	Resnik [25]	Schlicker et al. [88]	Lin [26]
#Only negatively associated DD pairs = 70 [available 521 labelled dataset [17]]	GloVe-50 [93]	−0.356	−0.245	−0.319	−0.319
	SicknessMiner [47]	−0.22114	−0.18152	−0.22987	−0.22987
	GexText [18]	−0.008	−0.125	−0.012	−0.011
	Enhanced literature-based positive, negative DDA	0.229156	0.208198	0.138969	0.138969
#Only negatively associated DD pairs= 2234 [Enhanced DDA dataset]	GloVe-50 [93]	−0.234	−0.301	−0.286	−0.284
	SicknessMiner [47]	−0.229	−0.298	−0.251	−0.253
	GexText [18]	−0.191	−0.226	−0.218	−0.218
	Enhanced literature-based positive, negative DDA	0.306	0.299	0.274	0.269

Table 5. Spearman's rank correlation between literature-based null similarity DD matrix and DO-based similarity metrics (Wang, Resnik, Relevance, Lin) for null associated DD pairs from different sets of labelled DDA datasets.

Association Type	Method	Wang et al. [24]	Resnik [25]	Schlicker et al. [88]	Lin [26]
#Null associated DD pairs = 2649 [available 521 labelled dataset [17]]	GloVe-50 [93]	0.07	0.018	0.0195	0.03
	SicknessMiner [47]	−0.0436	−0.009	0.003586	0.002185
	GexText [18]	0.231	0.236	0.196	0.193
	Literature-based Null DD similarity	0.333	0.307	0.281	0.281
#Null associated DD pairs= 36102 [Enhanced DDA dataset]	GloVe-50 [93]	0.006	0.005	0.002	0.003
	SicknessMiner [47]	0.024	0.017	0.0052	0.00549
	GexText [18]	0.168	0.186	0.210	0.208
	Literature-based Null DD similarity	0.423	0.419	0.454	0.456

5.2. Concept-based DDA evaluation

To characterize the concept-based DDA, the derived association embedding consisting of several components such as discovered new ancestors sets, mutli-source ancestorial embedding with root and leaf node, novel ancestorial level-based DDA quantification and finally, the proposed ontology-based joint multi-source association representation with the ontological relationship connections is evaluated with the association scores from DisGeNET and the human assessed combined dataset as discussed in Section 3.4.

For evaluating the concept embeddings represented using newly defined ancestors sets, the ontological sources such as the clinical classifications software for ICD-9-CM (diagnosis) from https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (accessed 2015) and anatomical therapeutic chemical classification system (ATC) (medications) https://www.whocc.no/atc/ (accessed 2 February 2018) are used.

5.2.1. Evaluation of discovered ancestors sets for DDA quantification

Table 6 shows the correlation effect of varying combinations of ancestors sets for $D_x$andD_y in D_x-D_y association quantification. With diseases in ontology, 7936 DD pairs found in common with DisGeNET, the embeddings derived with discovered new ancestors sets are better correlated compared to embeddings with all ancestors [86]. While considering only common ancestors without ancestors on longest path to D_x and D_y independently, shows good correlation than all ancestors but still less correlated when compared with the new ancestors sets. Since, association is not only influenced by commonality but also by ancestors on the longest path to each of the disease.

Table 6. Comparison of the effect of the new discovered ancestors sets to other ancestors sets of D_x and D_y for D_x-D_y quantification using spearman's rank correlation between association scores of DDA pairs obtained using different ancestors sets and DisGeNET DDA scores.

Source of multiple embeddings of ancestor	Ancestors' information	Spearman's rank correlation N=7936 DD pairs (DisGeNET)
Ontology Sources- CCS, ATC: Clinical Classifications Software for ICD-9-CM (CCS) (diagnosis) Anatomical Therapeutic Chemical classification system (ATC) (medications) [86]	All Ancestors of D_x and D_y : $\forall A_x$ and $\forall A_y$ [86]	0.759
	Common Ancestors of D_x and D_y : $\forall A_x \cap \forall A_y$	0.772
	New Ancestors Sets of D_x and D_y : New_Anc_Set(D_x) and New_Anc_Set(D_y)	0.779

5.2.2. Evaluation of multi-source embeddings for DDA quantification

With the best correlated newly defined ancestors sets and with all ancestors, the concept embeddings are further evaluated to show the effect of multi-source embeddings of those ancestors with and without including multi-source information of root and leaf nodes. In this regard, the concept embeddings are evaluated as shown in Table 7 for 2658 DD pairs. We observed that the concept embeddings using ancestorial embeddings from multiple conceptual sources including root node and leaf node multiple embeddings in addition to the new ancestors sets gives significantly higher correlation compared to the baseline that considers only semantic sources [86].

Table 7. Comparison of the effect of multi-source embeddings of ancestors with/without multi-source embeddings of root node and leaf node D_x or D_y for D_x-D_y quantification using spearman's rank correlation between association scores of DDA pairs obtained using multi-source ancestorial embeddings and DisGeNET DDA scores.

Sources of multiple embeddings of ancestor	**Different combination of ancestors sets for $\boldsymbol{D_x}$-$\boldsymbol{D_y}$ quantification**	**Without Root and leaf node multi-source embeddings**	**With Root and leaf node multi-source embeddings**
		Spearman's rank correlation N = 2658 DD pairs (DisGeNET)	
Ontology Sources: • Clinical Classifications Software for ICD- 9-CM (CCS) and Anatomical Therapeutic Chemical classification system (ATC) [86]	$\forall A_x$ *and* $\forall A_y$ [86]	0.612	0.618
	$\boldsymbol{New\_Anc\_Set(D_x)}$ *and* $\boldsymbol{New\_Anc\_Set(D_y)}$	**0.643**	**0.695**
Ontology Sources • **Disease Ontology (DO)** • **UMLS** **Biomedical Text** • **Clinical Notes** • **Insurance Claims Database** • **Journal Articles**	$\forall A_x$ *and* $\forall A_y$ [86]	0.726	0.730
	$\boldsymbol{New\_Anc\_Set(D_x)}$ *and* $\boldsymbol{New\_Anc\_Set(D_y)}$	**0.745**	**0.788**

5.2.3. Evaluation of novel ancestorial level-weight for DDA quantification

In order to evaluate the effect of level-weight or semantic value of LCS (D_x,D_y) in D_x-D_y association, we compared the level-weight of LCS computed by longest path of lower DAG with respect to D_x and D_y separately using the proposed novel level-weight and the upper DAG using Baseline_LCA in GOntoSim [65] as shown in Table 8. The results show that the DDA quantification by level-weight of LCS using lower part of DAG connecting D_x and D_y is better correlated with DisGeNET DDA scores than level-weight of LCS using upper part of DAG.

To demonstrate the effectiveness of adding novel level-weight to the multi-source ancestorial embeddings, we first introduce the effect of varying level-weight calculations of ancestors including LCS based on selection of children and then evaluated the effect of various combinations of level-weight with and without multi-source ancestorial embeddings. As shown in Table 9, with 1,75,939 DD pairs, the novel level weight, where the level-weight is contributed by the children that belongs only to the newly defined ancestors sets, even without multi-source ancestorial embeddings outperformed the baseline level-weight calculation [63]. In addition, the correlation is even better when the novel level-weight is applied on multi-source ancestorial embeddings.

Table 8. Comparison of the effect of upper and lower DAG-based level-weight or semantic value computation of LCS(D_x,D_y) in D_x-D_y association quantification using spearman's rank correlation between obtained association scores of DDA pairs by varying level-weight of LCS and and DDA scores from DisGeNET.

Calculation of level-weight of LCS($\mathrm{D_x}$, $\mathrm{D_y}$) for $\mathrm{D_x}$ - $\mathrm{D_y}$ quantification	Spearman's rank correlation N = 1,75,939 DD pairs (DisGeNET)
Baseline_LCA of GOntoSim: using upper DAG Level-weight of LCS($\mathrm{D_x, D_y}$) by ancestors on longest path to LCS($\mathrm{D_x, D_y}$) [65]	0.773
Novel ancestorial-level weight: using lower DAG Level-weight of LCS($\mathrm{D_x, D_y}$) by children on longest path to $\mathrm{D_x}$ and $\mathrm{D_y}$	0.782

Table 9. Comparison of the effect of novel ancestorial level-based to that of existing ancestorial level-based DDA quantification using spearman's rank correlation between association scores of DDA pairs obtained using level-weights of ancestors with and without ancestorial embeddings and DDA scores from DisGeNET.

Method	Without multi-source ancestorial embedding	With multi-source ancestorial embedding
	Spearman's rank correlation N = 1,75,939 DD pairs (DisGeNET)	
Semantic similarity of diseases [63]	0.610	0.720
Baseline_LCA of GOntoSim [65]	0.612	0.723
Novel ancestorial level-weight	0.619	0.756

5.2.4. Evaluation of ORVF for DDA quantification

In order to showcase the effect of the proposed ORVF, different combinations of the relationship connections as discussed earlier in Section 4.2.3. are considered. The performance of the effect of various ontological relationships is then evaluated through DDA quantification on DDA dataset as shown in Table 10.

Further, the proposed ontology-based joint multi-source association representation is evaluated against the state-of-art concept representation methods to project the effect of varying the ontological relationship connection of the given disease pair applied on to the association embedding derived by combining all the better performed components inferred from the sub-experiments as discussed earlier and is shown in Table 10 for 1756 DD pairs. The proposed model considering all subcomponents such as discovered new ancestors sets, multi-source ancestorial embedding with root and leaf node, novel ancestorial level-based DDA quantification and the ontological relationship connections is strongly correlated than other existing methods, because [86] considered only semantic ancestorial embeddings without level weight on all ancestors, is less correlated compared to other methods that considered other contextual and semantic type relations.

Table 10. Comparison of ontology-based joint multi-source association representation and the existing concept-based representation methods for DDA quantification using spearman's rank correlation between association scores of DDA pairs obtained using different concept-based representation methods and DDA scores from DisGeNET.

Different concept-based representation methods for concept-based DDA quantification	**Spearman's rank correlation N = 1756 DD pairs (DisGeNET)**
MMORE (CCS (diagnosis), ATC (medications)) [86]	0.703
Cui2vec (Clinical Notes, Claims Insurance, Journal articles) [20]	0.772
Retrofitted concept vector representation (PubMed, UMLS) [33]	0.781
Proposed **Ontology-based joint multi-source association representation**	
Ancestorial level-based + ontological relationship connection based-Parent, Grandparents only*	**0.787**
Ancestorial level-based + ontological relationship connection based-Parent, Grandparents & sibling only**	**0.790**
Ancestorial level-based + ontological relationship connection based-Parent, Grandparents, sibling, uncle & cousin relationships***	**0.802**

5.3. Evaluation of literature-based, concept-based and integrated approaches of disease representation for DDA quantification

The analysis presented so far shows the effectiveness of literature-based DDA and concept-based DDA. However, we need to evaluate integrated literature and concept based DDA representation. This requires representing each disease as a single disease vector representation, integrating literature-based and concept-based methods. This enhanced single vector representation of two diseases is then used to compute the DD association using cosine similarity. In order to show the effect of integrated disease representation, the association scores computed is compared with the other state-of-art methods using only literature-based, only concept-based and those with integrated literature-based and concept-based perspectives.

The disease representations produced by the models is evaluated across different perspectives of datasets. On the basis of type of DDA criteria, various angles of the datasets are used to evaluate the scores obtained by the generated disease representations. In this regard, we relied on disease-related biological domain database DisGeNet, where two association criteria were used to derive DDA scores. One is the disease-associated genes and other is disease-associated variants. Further, the Jaccard index similarity is used to compute association scores. In addition, we created a standard dataset covering the functional aspects of DDA using GO function. The disease-related GOs are obtained from Comparative Toxicogenomics Database (CTD). In order to calculate the DDA score in GO perspective, we employed the Jaccard index. Finally, we also evaluated against the human rated DD pairs obtained from a benchmark dataset. Details of the datasets used is discussed earlier in Section 3.4.

Table 11. Comparison of different aspects of disease vector representations using spearman's rank correlation between association scores of DDA pairs obtained across various angles of association criteria using DisGeNet (Gene and Variants), Standard dataset (GO) and human assessed scores.

Disease Vector Representation	**Spearman's Rank Correlation**			
Literature-based only	***DisGeNet Gene-based***	***DisGeNet Variant-based***	***Standard dataset GO-based***	***Human-rated***
	N = 2938 DD pairs (DisGeNET)			**N = 199 DD pairs**
Cui2vec (Clinical Notes, Claims Insurance, Journal articles) [20]	0.797	**0.254**	0.422	0.679
Enhanced Literature-Based Disease Vector Representation *(Vector-Similarity Fusion Without Integration)*	**0.799**	0.252	**0.427**	**0.682**
Concept-based only	***DisGeNet Gene-based***	***DisGeNet Variant-based***	***Standard dataset GO-based***	***Human-rated***
	N = 2638 DD pairs			**N = 50 DD pairs**
MMORE (CCS (diagnosis), ATC (medications)) [86]	0.716	0.144	0.541	0.790
Ontology-Based Joint Multi-Source Association Embedding **(Ancestral Level-Based + Ontological Relationship Connection Based)**	**0.808**	**0.146**	**0.551**	**0.809**
Integration of literature-based and concept-based	***DisGeNet Gene-based***	***DisGeNet Variant-based***	***Standard dataset GO-based***	***Human-rated***
	N = 1638 DD pairs			**N = 187 DD pairs**
Retrofitted Concept Vector Representation (PubMed, UMLS) [33]	0.801	0.213	0.592	0.810
Integration of literature-based and concept-based	***DisGeNet Gene-based***	***DisGeNet Variant-based***	***Standard dataset GO-based***	***Human-rated***
	N = 1638 DD pairs			**N = 187 DD pairs**
MORE [32]	0.811	0.220	0.609	0.813
Integrated Disease Vector Representation (*Literature-Based Positive, Negative & Concept-Based DDA*)	**0.816**	**0.225**	**0.624**	**0.818**
Enhanced Literature-Based & Concept-Based Joint Embedding Model for Disease Vector Representation (*Vector-Similarity Fusion with Integration*)	**0.822**	**0.227**	**0.626**	**0.821**

Table 11 summarizes the results of correlation of DDA scores obtained by different methods across various aspects of datasets. The DDA scores derived using only literature-based disease representation, shows better correlation than other literature-based method for DDA quantification in case of Gene-based, GO-based and human-rated scores. The reason may be that considering different context types in which DD pairs occur has a major influence on DDA scores as the additional features during the sentence representation learning can lead to better classified contexts. While, the correlation result on Variant-based dataset, is found to be less as the PubMed abstracts taken may not contain sentences that reveal much about variant related information or only limited contexts since we consider only disease mentioned sentences.

The DDA scores derived using only concept-based representation, found to have better correlation on all aspects of the datasets with only a slightly higher on variant-based. The proposed ontology-based method tries to embed a narrow information of concepts in ontology rather than generic concepts. This is achieved by controlling the contribution of ancestors on DDA in addition to varying the effect of different taxonomic relationships in ontology. Moreover, we select ancestors with respect to DDA rather than independently with respect to each of the diseases. All these has a major positive effect on DDA scores in different aspects.

On evaluation with the integrated approaches, the proposed method outperforms well compared to other baseline methods on all aspects of datasets. Integrating the enhanced literature-based contextual relations with enriched semantic relationships gives a broader coverage of relationships that might cover various influential factors affecting DDA. This basically includes indirect relationship information that can jointly eliminate false positives. Hence, the proposed work has shown promising results even for different aspects of DDA.

5.4. Implementation

The configurations of the machine include Intel(R) Xenon(R) 3.60 GHz (GPU), 64-bit OS (system) and 64 GB RAM (memory). Our system uses Python to implement the models. For literature-based DDA classification as discussed in Section 5.1.1, Table 12 shows the time taken by the baseline models and the proposed model for training and prediction tasks. On observation, we found that CNN models take less training time compared with other models since it involves less parameters calculation. However, LC-CNN and the proposed ESEC-CNN models, take almost equal time since only additional features have been added in the input sentence representation in ESEC-CNN model.

For concept-based DDA representation as discussed in Section 5.2.4, the proposed ontology-based joint multi-source embedding representation takes on an average of 22 seconds to derive DDA representation which is higher compared to other models. This arises from calculating different ancestors' information as discussed in earlier sections such as level weight, attention weights as well as the various ontological relationships to generate final representation of DDA. Other concept-based base-line models such as Cui2vec [20], Retrofitted concept vector representation [33] takes less time than MMORE [86] and the proposed model, as the former does not consider the deeper ancestors' information and ontological relationships. Compared with MMORE, the proposed model takes much more time since additional computation of ancestorial level weights and ontological relationships effect are involved. Though the proposed model, takes some time to obtain DDA representation, it is still able to produce quality embedding whose effectiveness is proved by the correlated results in Table 10.

Table 12. Comparison of computation time with base-line models.

Literature-based DDA sentence classification models	LSTM [21]	BiLSTM [89]	CNN [90]	BERT [91]	BioBERT [92]	LC-CNN [17]	**ESEC-CNN****
Training time per epoch (in seconds)	240s	237s	184s	262s	270s	196s	**200s**
Concept-based DDA representation	Retrofitted concept vector representation [33]	Cui2vec [20]	MMORE [86]	**Ontology-based joint multi-source association representation**	N/A	N/A	N/A
Average seconds to generate concept-based DDA representation	15s	16s	19s	**22s**	N/A	N/A	N/A

N/A-represents not applicable

5.5. Biological analysis

The significance of DDA scores obtained by the proposed framework is analysed in biological aspects: listing top 20 associated disease-disease pairs with normalized scores in Table 13, disease-wise most associated diseases in Table 13, top 5 category-wise associations and also the top 10 associated diseases with corresponding categories for a given disease.

For a given disease, Table 14 shows the most associated disease pairs comparatively to others.

The performance of disease representation in DDA quantification is further validated by disease categories, where the diseases are classified according to top 14-level DO categories such as "disease of cellular proliferation", "nervous system disease", "cardiovascular system disease", "musculoskeletal system disease", "endocrine system disease" and so on [72]. The strength of association between disease categories is measured by averaging the normalized DDA scores between disease categories. The disease category pairs are ranked based on the normalized score.

We find that disease associated within same category have high average association score than with diseases of other categories as shown in Table 15. On observation, diseases in "nervous system disease" category have relatively higher association scores across all other disease categories. On the other hand, we find that average association scores of diseases in "disease by infectious agent", "endocrine system disease", "urinary system disease" have lower association scores with all other categories compared to diseases within itself. In case of "nervous system disease" category, is comparatively higher within and with "cardiovascular system disease" and "musculoskeletal system disease". While the average association score of diseases in "disease of cellular proliferation", are far lower with diseases in "endocrine system disease" and "cardiovascular system disease" than for other categories.

Table 13. Top 20 associated disease pairs ranked by normalized DDA scores.

Disease 1	Disease 2	Association score
amyotrophic lateral sclerosis	motor neuron disease	0.999998
Hypertensive retinopathy	Vascular disease	0.096729
cardiovascular disease	intrinsic asthma	0.043286
autosomal dominant polycystic kidney disease	autosomal dominant polycystic kidney disease	0.033808
myopathy	Sjogren's syndrome	0.022195
congenital muscular dystrophy	muscular dystrophy	0.010661
cerebral folate receptor alpha deficiency	Down syndrome	0.007642
muscular dystrophy	myotonic dystrophy type2	0.006518
Alzheimer's disease	Moyamoya disease	0.004409
migraine without aura	Fibromyalgia	0.003855
diabetes mellitus	diabetic neuropathy	0.003613
diabetes mellitus	Hypoglycemia	0.003495
acute myeloid leukemia	acute monocytic leukemia	0.003449
lepromatous leprosy	Leprosy	0.002793
marasmus	anorexia nervosa	0.002623
lymphoblastic leukemia	lung disease	0.002265
cystic fibrosis	acute pancreatitis	0.002053
acute monocytic leukemia	acute leukemia	0.001959
Azoospermia	oligospermia	0.001876
acute myeloid leukemia	acute leukemia	0.001875

Table 14. Disease-wise most associated diseases.

Disease	**Most associated diseases**
amyotrophic lateral sclerosis	Motor neuron disease, lateral sclerosis
motor neuron disease	Motor neuron disease, cardiovascular disease
Hypertensive retinopathy	Vascular disease
Vascular disease	Hypertensive retinopathy, cardiovascular disease
Intrinsic asthma	Cardiovascular disease, lung disease
autosomal dominant polycystic kidney disease	autosomal recessive polycystic kidney disease
Myopathy	Sjogren's syndrome
Sjogren's syndrome	Myopathy, systemic scleroderma, Behcet's disease, systemic lupus erythematosus

Table 15. Top 5 associated category pairs ranked by average of normalized DDA scores between intra and inter disease categories.

DO Disease Category	Top 5 associated DO Disease Categories	Normalized average score of disease pairs category-wise
Nervous system disease	Nervous system disease	0.0223
	Cardiovascular system disease	0.0221
	Musculoskeletal system disease	0.0200
	Physical disorder	0.0085
	Disease of metabolism	0.0025
Physical disorder	Physical disorder	0.014
	Nervous system disease	0.0085
	Disease of metabolism	0.0052
	Disease of cellular proliferation	0.0038
	Endocrine system disease	0.00056
Disease of cellular proliferation	Disease of cellular proliferation	0.00400
	Respiratory system disease	0.00382
	Physical disorder	0.00380
	Endocrine system disease	0.000620
	Cardiovascular system disease	0.000621
Urinary system disease	Urinary system disease	0.00109
	Disease of cellular proliferation	0.00062
	Endocrine system disease	0.00059
	Gastrointestinal system disease	0.00054
	Disease by infectious agent	0.00048
Endocrine system disease	Endocrine system disease	0.000920
	Disease of cellular proliferation	0.000626
	Urinary system disease	0.000596
	Musculoskeletal system disease	0.000580
	Cardiovascular system disease	0.000564
Disease by infectious agent	Disease by infectious agent	0.00072
	Endocrine system disease	0.000486
	Nervous system disease	0.000485
	Respiratory system disease	0.000482
	Gastrointestinal system disease	0.000475

In addition, we have also shown the category-wise top 10 associated disease pairs for "Diabetes mellitus" of "endocrine system disease" and "cardiovascular disease" of "Cardiovascular system disease" in table 16.

Table 16. Top 10 associated diseases category-wise ranked by normalized DDA scores.

Disease with category	Top 10 associated disease pairs	Category of associated disease pair	Association score
Diabetes mellitus *Endocrine system disease*	diabetic neuropathy	endocrine system disease	0.003613
	hypoglycaemia	endocrine system disease	0.003495
	acute myocardial infarction	cardiovascular system disease	0.000944
	diabetic retinopathy	nervous system disease	0.000938
	stomach cancer	disease of cellular proliferations	0.000918
	kidney failure	urinary system disease	0.000728
	Hypothyroidism	endocrine system disease	0.000649
	disease of metabolism	disease of metabolism	0.000627
	autoimmune disease	musculoskeletal system disease	0.000603
	brain cancer	disease of cellular proliferations	0.000567
Cardiovascular disease *Cardiovascular system disease*	intrinsic asthma	respiratory system disease	0.043286
	nephrotic syndrome	urinary system disease	0.001028
	vascular disease	cardiovascular system disease	0.000399
	disease by infectious agent	parasetic infectious disease	0.000376
	vein disease	cardiovascular system disease	0.000359
	generalized atherosclerosis	cardiovascular system disease	0.000353
	Moyamoya disease	cardiovascular system disease	0.000351
	peripheral artery disease	cardiovascular sys. disease	0.000347
	Epilepsy	nervous system disease	0.000331
	intermediate coronary syndrome	cardiovascular system disease	0.000320

6. Conclusions

Representing a richer quality of disease vectors for a qualitative and quantitative measurement of DDA strength provides valuable information to the clinicians for better healthcare planning. The existing methods of integrated vector representation failed to consider various sentence contexts from literature and semantic embedding of concepts along with different ontological relationship connections from ontology for better quantification of biomedical associations. To address this issue, in this paper, we presented an enhanced and integrated DDA framework incorporating various types of sentence contexts such as positive, negative and null from literature with semantically embedded concepts and various ontological relationship connections affecting associations from ontology for a richer quality of disease vector representation. The enriched disease vectors achieved well correlated DDA scores especially on gene-based when evaluated in different aspects of datasets compared to other baseline literature-based, concept-based and integrated representations. Moreover, we also shown the top associated disease pairs and category-pairs. Any biomedical association quantification using biomedical entities representations could greatly be benefited from a richer vector representation using the enhanced and integrated framework. In future, the integrated representation can also be carried out for determining the strength of other biomedical associations such as disease-gene, gene-gene, disease-symptoms etc.

Acknowledgments

We would like to thank all biomedical data sources for data supporting.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. M. Žitnik, V. Janjić, C. Larminie, B. Zupan, N. Pržulj, Discovering disease- disease associations by fusing systems-level molecular data, *Sci. Rep.*, **3** (2013), 1–9. https://doi.org/10.1038/srep03202
2. S. Bang, J. H. Kim, H. Shin, Causality modeling for directed disease network, *Bioinformatics*, **32** (2016), 437–444. https://doi.org/10.1093/bioinformatics/btw439
3. A. Suratanee, K. Plaimas, DDA: A novel network-based scoring method to identify disease-disease associations, *Bioinform. Biol. Insights*, **9** (2015), 175–186. https://doi.org/10.4137/bbi.s35237
4. J. Yang, S. J. Wu, S. Y. Yang, J. W. Peng, S. N. Wang, F. Y. Wang, et al., DNetDB: The human disease network database based on dysfunctional regulation mechanism, *BMC Syst. Biol.*, **10** (2016), 1–8. https://doi.org/10.1186/s12918-016-0280-5
5. J. Li, B. Gong, X. Chen, T. Liu, C. Wu, F. Zhang, et al., DOSim: An R package for similarity between diseases based on disease ontology, *BMC Bioinformatics.*, **12** (2011), 1–10. https://doi.org/10.1186/1471-2105-12-266
6. D. A. Davis, N. V. Chawla, Exploring and exploiting disease interactions from multi-relational gene and phenotype networks, *PLoS One*, **6** (2011), 22670. https://doi.org/10.1371/journal.pone.0022670
7. Y. Li, P. Agarwal, A pathway-based view of human diseases and disease relationships, *PLoS One*, **4** (2009), 4346. https://doi.org/10.1371/journal.pone.0004346
8. S. S. Deepika, T. V. Geetha, Pattern-based bootstrapping framework for biomedical relation extraction, *Eng. Appl. Artif. Intell.*, **99** (2021), 104130. https://doi.org/10.1016/j.engappai.2020.104130
9. C. A. Hidalgo, N. Blumm, A. L. Barabási, N. A. Christakis, A dynamic network approach for the study of human phenotypes, *PLoS Comput. Biol.*, **5** (2009), 1000353. https://doi.org/10.1371/journal.pcbi.1000353

10. X. Zhang, R. Zhang, Y. Jiang, P. Sun, G. Tang, X. Wang, et al., The expanded human disease network combining protein-protein interaction information, *Eur. J. Hum. Genet.*, **19** (2011), 783–788. https://doi.org/10.1038/ejhg.2011.30

11. Y. Hu, M. Zhou, H. Shi, H. Ju, Q. Jiang, L. Cheng, Measuring disease similarity and predicting disease-related ncRNAs by a novel method, *BMC Med. Genomics*, **10** (2017), 67–74. https://doi.org/10.1186/s12920-017-0315-9

12. L. M. Schriml, J. B. Munro, M. Schor, D, Olley, C. McCracken, V. Felix, et al., The human disease ontology 2022 update, *Nucleic Acids Res.*, **50** (2022), 1255–1261. https://doi.org/10.1093/nar/gkab1063

13. S. Carbon, E. Douglass, N. Dunn, B. M. Good, N. L. Harris, S. E. Lewis, et al., The gene ontology resource: 20 years and still GOing strong, *Nucleic Acids Res.*, **47** (2019), 330–338. https://doi.org/10.1093/nar/gky1055

14. S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A Vasilevsky, et al., The human phenotype ontology in 2021, *Nucleic Acids Res.*, **49** (2021), 1207–1217. https://doi.org/10.1093/nar/gkaa1043

15. O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Res.*, **32** (2004), 267–270. https://doi.org/10.1093/nar/gkh061

16. D. L. Ngo, N. Yamamoto, V. A. Tran, N. G. Nguyen, D. Phan, F. R. Lumbanraja, et al., Application of word embedding to drug repositioning, *J. Biomed. Sci. Eng.*, **09** (2016), 7–16, http://doi.org/10.4236/jbise.2016.91002

17. P. T. Lai, W. L. Lu, T. R. Kuo, C. R. Chung, J. C. Han, R. T. H. Tsai, et al., Using a large margin context-aware convolutional neural network to automatically extract disease-disease association from literature: Comparative analytic study, *JMIR Med. Inform.*, **7** (2019), 14502. https://doi.org/10.2196/14502

18. R. O'Shea, Gextext: disease network extraction from biomedical literature, preprint, arXiv1911.02562.

19. M. V. Korff, B. Deffarges, T. Sander, Data mining in MEDLINE for disease-disease associations via second order co-occurrence, in *2015 IEEE Symposium Series on Computational Intelligence,* (2015), 314–321, https://doi.org/10.1109/SSCI.2015.54

20. A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, et al., Clinical concept embeddings learned from massive sources of medical data, *Pac. Symp. Biocomput.*, **25** (2018), 295–306.

21. A. B. Holmes, A. Hawson, F. Liu, C. Friedman, H. Khiabanian, R. Rabadan, Discovering disease associations by integrating electronic clinical data and medical literature, *PLoS One*, **6** (2011), 21132. https://doi.org/10.1371/journal.pone.0021132

22. S. Ghosh, P. Chakraborty, E. Cohn, J. S. Brownstein, N. Ramakrishnan, Characterizing diseases from unstructured text: A vocabulary driven Word2vec approach, preprint, arXiv.1603.00106.

23. J. Park, K. Kim, W. Hwang, D. Lee, Concept embedding to measure semantic relatedness for biomedical information ontologies, *J. Biomed. Inform.*, **94** (2019), 103182. https://doi.org/10.1016/j.jbi.2019.103182

24. J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, C. F. Chen, A new method to measure the semantic similarity of GO terms, *Bioinformatics*, **23** (2007), 1274–1281. https://doi.org/10.1093/bioinformatics/btm087

25. P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, preprint, arXiv:cmp-lg/9511007.
26. D. Lin, An information-theoretic definition of similarity, in *International Conference on Machine Learning,* **98** (1998), 296–304.
27. S. Mathur, D. Dinakarpandian, Automated ontological gene annotation for computing disease similarity, *Summit Transl Bioinform.*, (2010), 12–16.
28. S. Mathur, D. Dinakarpandian, Finding disease similarity based on implicit semantic similarity, *J. Biomed. Inform.*, **45** (2012), 363–371, https://doi.org/10.1016/j.jbi.2011.11.017
29. F. Z. Smaili, X. Gao, R. Hoehndorf, Onto2Vec: Joint vector-based representation of biological entities and their ontology-based annotations, *Bioinformatics*, **34** (2018), 52–60, doi: https://doi.org/10.1093/bioinformatics/bty259
30. L. Cheng, J. Li, P. Ju, J. Peng, Y. Wang, SemFunSim: A new method for measuring disease similarity by integrating semantic and gene functional association, *PLoS One*, **9** (2014), 99415. https://doi.org/10.1371/journal.pone.0099415
31. S. Su, L. Zhang, J. Liu, An effective method to measure disease similarity using gene and phenotype associations, *Front. Genet.*, **10** (2019), 1–8. https://doi.org/10.3389/fgene.2019.00466
32. S. Jiang, W. Wu, N. Tomita, C. Ganoe, S. Hassanpour, Multi-Ontology Refined Embeddings (MORE): A hybrid multi-ontology and corpus-based semantic representation model for biomedical concepts, *J. Biomed. Inform.*, **111** (2020), 103581. https://doi.org/10.1016/j.jbi.2020.103581
33. Z. Yu, B. C. Wallace, T. Johnson, T. Cohen, Retrofitting concept vector representations of medical concepts to improve estimates of semantic similarity and relatedness, preprint, arXiv.1709.07357.
34. E. Nourani, V. Reshadat, Association extraction from biomedical literature based on representation and transfer learning, *J. Theor. Biol.*, **488** (2020), 110112. https://doi.org/10.1016/j.jtbi.2019.110112
35. Y. Peng, Z. Lu, Deep learning for extracting protein-protein interactions from biomedical literature, in *Proceedings of the BioNLP 2017 workshop*, (2017), 29–38. http://doi.org/10.18653/v1/W17-2304
36. C. Quan, L. Hua, X. Sun, W. Bai, Multichannel convolutional neural network for biological relation extraction, *Biomed Res. Int.*, (2016), 1850404. https://doi.org/10.1155/2016/1850404
37. N. K. Rakhi, R. Tuwani, J. Mukherjee, G. Bagler, Data-driven analysis of biomedical literature suggests broad-spectrum benefits of culinary herbs and spices, *PLoS One*, **13** (2018), 1–20, doi: https://doi.org/10.1371/journal.pone.0198030
38. J. Li, X. Zhu, J. Y. Chen, Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts, *PLoS Comput. Biol.*, **5 (**2009), 1000450. https://doi.org/10.1371/journal.pcbi.1000450
39. H. W. Chun, Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, T. Hishiki, et al., Extraction of gene-disease relations from medline using domain dictionaries and machine learning, *Pac. Symp. Biocomput.,* **15** (2006), 4–15. https://doi.org/10.1142/9789812701626_0002
40. C. Perez-Iratxeta, P. Bork, M. A. Andrade, Association of genes to genetically inherited diseases using data mining, *Nat. Genet.*, **31** (2002), 316–319. https://doi.org/10.1038/ng895
41. S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, L. J. Jensen, DISEASES: Text mining and data integration of disease-gene associations, *Methods*, **74** (2015), 83–89. https://doi.org/10.1016/j.ymeth.2014.11.020

42. J. Lee, S. Kim, S. Lee, K. Lee, J. Kang, On the efficacy of per-relation basis performance evaluation for PPI extraction and a high-precision rule-based approach, *BMC Med. Inform. Decis. Mak.*, **13** (2013), 1–12. https://doi.org/10.1186/1472-6947-13-S1-S7
43. M. Song, W. C. Kim, D. Lee, G. E. Heo, K. Y. Kang, PKDE4J: Entity and relation extraction for public knowledge discovery, *J. Biomed. Inform.*, **57** (2015), 320–332. https://doi.org/10.1016/j.jbi.2015.08.008
44. L. Tari, J. Hakenberg, G. Gonzalez, C. Baral, Querying parse tree database of medline text to synthesize user-specific biomolecular networks, *Pac. Symp. Biocomput.*, **98** (2009), 87–98. https://doi.org/10.1142/9789812836939_0009
45. B. Bhasuran, J. Natarajan, Automatic extraction of gene-disease associations from literature using joint ensemble learning, *PLoS One*, **13** (2018), 1–22. https://doi.org/10.1371/journal.pone.0200699
46. Y. Zhang, Z. Lu, Exploring semi-supervised variational autoencoders for biomedical relation extraction, *Methods*, **166** (2019), 112–119. https://doi.org/10.1016/j.ymeth.2019.02.021
47. N. Rosário-Ferreira, V. Guimarães, V. S. Costa, I. S. Moreira, SicknessMiner: a deep-learning-driven text-mining tool to abridge disease-disease associations, *BMC Bioinformatics*, **22** (2021), 1–12. https://doi.org/10.1186/s12859-021-04397-w
48. M. Asada, M. Miwa, Y. Sasaki, Extracting drug-drug interactions with attention CNNs, in *Proceedings of the BioNLP 2017 workshop*, (2017), 9–18. http://doi.org/10.18653/v1/W17-2302
49. Y. Hsieh, Y. Chang, N. Chang, W. Hsu, Identifying rotein-protein interactions in biomedical literature using recurrent neural networks with long short-term memory, in *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, (2017), 240–245.
50. L. Hua, C. Quan, A shortest dependency path based convolutional neural network for protein-protein relation extraction, *Biomed Res. Int.*, (2016), 8479587. https://doi.org/10.1155/2016/8479587
51. X. Wang, L. Zhu, Z. Zheng, M. Xu, Y. Yang, Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision, *IEEE Trans. Multimedia*, (2022), 1–11. https://doi.org/10.1109/TMM.2022.3204444
52. R. Xu, L. Li, Q. Q. Wang, DRiskKB: A large-scale disease-disease risk relationship knowledge base constructed from biomedical text, *BMC Bioinformatics*, **15 (**2014) 105. https://doi.org/10.1186/1471-2105-15-105
53. D. Wei, T. Kang, H. A. Pincus, C. Weng, Construction of disease similarity networks using concept embedding and ontology, *Stud. Health Technol. Inform.*, **264** (2019), 442–446. https://doi.org/10.3233/shti190260
54. S. V. S. Pakhomov, G. Finley, R. McEwan, Y. Wang, G. B. Melton, Corpus domain effects on distributional semantic modeling of medical terms, *Bioinformatics*, **32** (2016), 3635–3644. https://doi.org/10.1093/bioinformatics/btw529
55. G. K. Mazandu, N. J. Mulder, Information content-based gene ontology semantic similarity approaches: Toward a unified framework theory, *Biomed Res. Int.*, (2013). https://doi.org/10.1155/2013/292063
56. X. Song, L. Li, P. K. Srimani, P. S. Yu, J. Z. Wang, Measure the semantic similarity of go terms using aggregate information content, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **11** (2014), 468–476. https://doi.org/10.1109/tcbb.2013.176

57. Z. Teng, M. Guo, X. Liu, Q. Dai, C. Wang, P. Xuan, Measuring gene functional similarity based on group-wise comparison of GO terms, *Bioinformatics*, **29** (2013), 1424–1432. https://doi.org/10.1093/bioinformatics/btt160
58. C. Zhao, Z. Wang, GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms, *Sci. Rep.*, **8** (2018), 1–10. https://doi.org/10.1038/s41598-018-33219-y
59. Z. Wu, M. Palmer, Verb semantics and lexical selection, in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, (1994), 133–138. https://doi.org/10.3115/981732.981751
60. R. Richardson, A. Smeaton, J. Murphy, Using WordNet as a knowledge base for measuring semantic similarity between words, *Tech. Rep. Work. Pap.* **9** (1994).
61. J. Cheng, M. S. Cline, J. Martin, D. Finkelstein, T. Awad, D. Kulp, et al., A knowledge-based clustering algorithm driven by Gene Ontology, *J. Biopharm. Stat.*, **14** (2004), 687–700. https://doi.org/10.1081/bip-200025659
62. H. Wu, Z. Su, F. Mao, V. Olman, Y. Xu, Prediction of functional modules based on comparative genome analysis and Gene Ontology application, *Nucleic Acids Res.*, **33** (2005), 2822–2837. https://doi.org/10.1093/nar/gki573
63. D. Wang, J. Wang, M. Lu, F. Song, Q. Cui, Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases, *Bioinformatics*, **26** (2010),1644–1650. https://doi.org/10.1093/bioinformatics/btq241
64. G. K. Mazandu, N. J. Mulder, A topology-based metric for measuring term similarity in the gene ontology, *Adv. Bioinformatics*, (2012), 975783. https://doi.org/10.1155/2012/975783
65. A. B. Kamran, H. Naveed, GOntoSim: A semantic similarity measure based on LCA and common descendants, *Sci. Rep.*, **12** (2022), 3818. https://doi.org/10.1038/s41598-022-07624-3
66. J. Camacho-Collados, M. T. Pilehvar, R. Navigli, NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities, *Artif. Intell.*, **240** (2016), 36–64. https://doi.org/10.1016/j.artint.2016.07.005
67. Z. H. Guo, Z. H. You, D. S. Huang, H. C. Yi, K. Zheng, Z. H. Chen, et al., MeSHHeading2vec: A new method for representing MeSH headings as vectors based on graph embedding algorithm, *Brief. Bioinform.*, **22** (2021), 2085–2095. https://doi.org/10.1093/bib/bbaa037
68. X. Zhong, R. Kaalia, J. C. Rajapakse, GO2Vec: Transforming GO terms and proteins to vector representations via graph embeddings, *BMC Genomics*, **20** (2019), 918. https://doi.org/10.1186/s12864-019-6272-2
69. F. Z. Smaili, X. Gao, R. Hoehndorf, OPA2Vec: Combining formal and informal content of biomedical ontologies to improve similarity-based prediction, *Bioinformatics*, **35** (2019), 2133–2140. https://doi.org/10.1093/bioinformatics/bty933
70. J. Lee, D. Lee, K. H. Lee, Literature mining for context-specific molecular relations using multimodal representations (COMMODAR), *BMC Bioinformatics*, **21** (2020), 250. https://doi.org/10.1186/s12859-020-3396-y
71. L. Deng, D. Ye, J. Zhao, J. Zhang, MultiSourcDSim: An integrated approach for exploring disease similarity, *BMC Med. Inform. Decis. Mak.*, **19** (2019), 269. https://doi.org/10.1186/s12911-019-0968-8
72. P. Li, Y. Nie, J. Yu, Fusing literature and full network data improves disease similarity computation, *BMC Bioinformatics*, **17** (2016), 326. https://doi.org/10.1186/s12859-016-1205-4

73. C. H. Wei, A. Allot, R. Leaman, Z. Lu, PubTator central: automated concept annotation for biomedical full text articles, *Nucleic Acids Res.*, **47** (2019), 587–593. https://doi.org/10.1093/nar/gkz389
74. J. Piñero, J. M. R. Anguita, J. S. Pitarch, F. Ronzano, E. Centeno, F. Sanz, et al., The DisGeNET knowledge platform for disease genomics: 2019 update, *Nucleic Acids Res.*, **48** (2020), 845–855. https://doi.org/10.1093/nar/gkz1021
75. L. E. Salnikova, E. V. Chernyshova, L. A. Anastasevich, S. S. Larin, Gene-and disease-based expansion of the knowledge on inborn errors of immunity, *Front. Immunol.*, **10** (2019), 2475. https://doi.org/10.3389/fimmu.2019.02475
76. T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, C. G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, *J. Biomed. Inform.*, **40** (2007), 288–299. https://doi.org/10.1016/j.jbi.2006.06.004
77. A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, J. Wiegers, T. C. Wiegers, et al., Comparative Toxicogenomics Database (CTD): Update 2021, *Nucleic Acids Res.*, **49** (2021), 1138–1143. https://doi.org/10.1093/nar/gkaa891
78. S. Manchanda, A. Anand, Representation learning of drug and disease terms for drug repositioning, in *2017 3rd IEEE International Conference on Cybernetics (CYBCON)*, (2017), 1–6. https://doi.org/10.1109/CYBConf.2017.7985802
79. J. J. Lastra-Díaz, J. Goikoetxea, M. A. Hadj Taieb, A. García-Serrano, M. Ben Aouicha, E. Agirre, A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art, *Eng. Appl. Artif. Intell.*, **85** (2019), 645–665. https://doi.org/10.1016/j.engappai.2019.07.010
80. M. de Marneffe, C. D. Manning, Stanford typed dependencies manual, 2008. Available from: https://downloads.cs.stanford.edu/nlp/software/dependencies_manual.pdf
81. Y. Tsuruoka, J. Tsujii, Bidirectional inference with the easiest-first strategy for tagging sequence data, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, (2005), 467–474. http://doi.org/10.3115/1220575.1220634
82. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, **2** (2013), 3111–3119.
83. S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, S. Ananiadou, Distributional semantics resources for biomedical text processing, *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, **5** (2013), 39–44.
84. Y. Tang, Deep Learning using Linear Support Vector Machines, preprint, arXiv:1306.0239.
85. E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, J. Sun, GRAM: Graph-based attention model for healthcare representation learning, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2017), 787–795. https://doi.org/10.1145%2F3097983.3098126
86. L. Song, C. W. Cheong, K. Yin, W. K. Cheung, B. C. M. Fung, J. Poon, Medical concept embedding with multiple ontological representations, in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, (2019), 4613–4619. https://doi.org/10.24963/ijcai.2019/641

87. Q. Le, T. Mikolov, Distributed representations of sentences and documents, preprint, arXiv:1405.4053
88. A. Schlicker, F. S. Domingues, J. Rahnenführer, T. Lengauer, A new measure for functional similarity of gene products based on gene ontology, *BMC Bioinformatics*, **7** (2006), 302. https://doi.org/10.1186/1471-2105-7-302
89. M. Miwa, M. Bansal, End-to-end relation extraction using LSTMs on sequences and tree structures, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, **1** (2016), 1105–1116. http://doi.org/10.18653/v1/P16-1105
90. D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, (2014), 2335–2344.
91. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1** (2019), 4171–4186. http://doi.org/10.18653/v1/N19-1423
92. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, et al., BioBERT: A pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, **36** (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682
93. J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* (2014), 1532–1543. http://doi.org/10.3115/v1/D14-1162