An Integrative Disease Information Network Approach to Similar Disease Detection

Wuli Xu, Lei Duan✉, Huiru Zheng, Jesse Li-Ling, Weipeng Jiang, Yidan Zhang, Tingting Wang, Ruiqi Qin

Abstract—Disease similarity analysis impacts significantly in pathogenesis revealing, treatment recommending, and disease-causing genes predicting. Previous works study the disease similarity based on the semantics obtaining from biomedical ontologies (e.g., disease ontology) or the function of disease-causing molecules. However, such methods almost focus on a single perspective for obtaining disease features, which may lead to biased results for similar disease detection. To address this issue, we propose a disease information network-based integrate approach named *MISSION* for detecting similar diseases. By leveraging the associations between diseases and other biomedical entities, the disease information network is established firstly. And then, the disease similarity features extracted from the aspects of disease taxonomy, attributes, literature, and annotations are integrated into the disease information network. Finally, the top-*k* similar disease query is performed based on the integrative disease information. The experiments conducted on real-world datasets demonstrate that *MISSION* is effective and useful in similar disease detection.

Index Terms—similar disease detection, disease information network, multimodal-information

1 Introduction

Similar disease detection has a wide range of biomedical applications, such as disease classification [1], pathogenesis understanding [2], disease-causing molecules inference [3], therapeutic drug prediction [4], and clinical decision-making systems improvement [5]. For a pair of diseases, multiple types of relationships can be employed to measure the similarity. For example, the "disease-gene-disease" relationship is used to find diseases sharing identical disease-causing genes. And the "disease-phenotype-disease" relationship is used to find diseases having similar symptoms.

In general, similar disease detection methods can be divided into the following two categories:

- *Homogeneous relationship based*: some studies focus on using a specific relationship to discover similar diseases. For example, Wang *et al.* [6] measured the disease similarity on the aspect of term semantics through the "disease-term-disease" relationship obtained from Disease Ontology (DO) [7]. Mathur *et al.* [8] measured the similarity between diseases on the aspect of the "disease-gene-disease" relationship retrieved from Gene Ontology (GO) [9]. Clearly, the results of similar disease detection depend on the selection of similar relationships. However, it is not easy to select a suitable similar relationship unless the available domain knowledge is enough.
- *Heterogeneous relationships based*: some studies consider the disease similarity by combining multiple similar relationships. For example, Qin *et al.* [10] provided a disease information network (DIN) constructed by multiple disease relationships to perform similar disease detection. Cheng *et al.* [11] combined the "disease-gene-disease" relationship (i.e., disease-causing gene functions) with the "disease-term-disease" relationship (i.e., disease term semantics) to measure the disease similarity.

Clearly, it's more flexible to consider heterogeneous relationships, compared with homogeneous ones, for similar disease detection. DIN [10] includes multi-type disease-related entities (nodes) and relationships (edges). As a result, it can provide a comprehensive perspective on disease information.

It's worth noting that the disease-related information in the DIN is structured. In other words, the semantic of any path in the DIN is explicitly defined. Besides the structured information represented in the DIN, we find that the unstructured information is useful for similar disease detection. For example, the literature contains rich and latest research progresses about diseases. The disease taxonomy (provided by DO) describes the hierarchical relationships among diseases. The disease annotations (provided by GO) introduce the functions of related genes. Thus, performing similarity analysis on the unstructured information can provide an extra contribution to similar disease detection.

As stated above, we can see that the unstructured information is multimodal. Multimodal-information has multiple types of information, providing different perspectives to describe the relevance between diseases. Thus, it is unreasonable to adopt a unified similarity measure. Instead, we propose an integrative approach to detect similar dis-

- *W. Xu is with the School of Computer Science, Sichuan University, Chengdu 610065, China.*
- *L. Duan is with the School of Computer Science, Sichuan University, Chengdu 610065, China. E-mail: leiduan@scu.edu.cn.*
- *H. Zheng is with the School of Computing, Ulster University, Northern Ireland, United Kingdom.*
- *J. Li-Ling is with the State Laboratory of Biotherapy, Sichuan University, Chengdu 610041, China.*
- *W. Jiang and Y. Zhang are with the School of Computer Science, Sichuan University, Chengdu 610065, China.*
- *T. Wang is with the School of Computer Science & IT, RMIT University.*
- *R. Qin is with SAP (China) Co., Ltd. Chengdu Branch.*

eases. Specifically, we first construct the DIN using structured information and compute the disease similarities. Secondly, we evaluate the disease similarities on each type of multimodal-information. Finally, similar diseases are detected by integrating similarities.

Technically, there are two challenges.

- How to design the similar measure for each type of multimodal-information?
- How to integrate disease similarities under different information scales?

To address these challenges, we propose a DIN-based integrative approach, named *MISSION* (short for multimodal-information-aided similar disease detection), to perform top-*k* similar disease query for a given disease. Briefly, *MISSION* starts with constructing the DIN which contains four entities (i.e., disease, phenotype, gene, and chemical) and the relationships among them. The DIN-based similarity is evaluated based on the meta structure connecting the query diseases. Next, for the multimodal-information, *MISSION* computes the similarity with each type of multimodal-information independently, and combines the results together. Finally, *MISSION* adopts the multimodal compact bilinear pooling to integrate the DIN-based similarity and multimodal-information-based similarity.

In summary, the main contributions of this work are as follows:

- We propose a DIN-based integrative approach *MISSION* to detect similar diseases. *MISSION* not only takes DIN-based similarity into consideration, but also can evaluate disease similarity from multimodal-information.
- We design similarity measures for disease taxonomy, disease attributes, disease literature, and disease annotations information, respectively, as well as an approach for similarity integration.
- We conduct extensive experiments using real-world datasets to demonstrate the effectiveness of *MISSION* for similar disease detection from multimodal-information.

A preliminary version of this work appeared in the proceedings of the IEEE BIBM 2020 conference [12]. Compared to that work, we made several major improvements. First, we incorporate another type of multimodal-information, i.e., disease annotations, to measure the relationships between diseases more comprehensively. Second, we optimize the interaction strategy between the disease information network and auxiliary multimodal-information. Third, we add an analysis of the works related to disease information retrieval. Finally, we conduct more extensive empirical evaluations: (1) adding more evaluation indicators in performance comparison; (2) performing a more detailed analysis of *MISSION* variants; (3) testing parameter sensitivity; and (4) providing an extra case study about top-*k* similar disease query.

The remainder of the paper is organized as follows. We briefly discuss the related works in Section 2. Section 3 introduces the architecture of *MISSION*, followed by experiments and results discussed in Section 4. In Section 5, we conclude our work and highlight the future directions.

2 Related Work

2.1 Disease Similarity Analysis

With available data increasing, many studies about disease similarity analysis were carried out, which advance the development of biology.

Some methods depended on DO to measure the semantic similarity of diseases. DO unifies the representations of diseases among varied vocabularies into a relational ontology, in which the term is the professional noun for referring to each disease. The Information Content (IC), indicating how specific and informative a term is, has been widely used for measuring the similarities among terms. For example, Resnik *et al.* [13] measured disease similarity based on the IC of the most informative common ancestor (MICA) between two terms. Lin *et al.* [14] incorporated the IC of both two terms and their MICA. Wang *et al.* [6] computed the similarities among terms by considering the contributions of all common ancestors in the ontology.

Moreover, some studies utilized the relationships between diseases and genes to detect similar diseases. Mathur *et al.* [15] first proposed a method called *BOG* that calculates similarity by comparing gene overlaps of related diseases. However, it does not consider the functional relations between disease-related genes. Furthermore, Mathur *et al.* [8] proposed a process similarity-based (*PSB*) method involving the GO biological process terms associated with genes.

However, only relying on disease-related genes greatly limits the utility of those methods mentioned above. Recently, Qin *et al.* [10] proposed *RADAR*, which derives a multi-layer similarity network from multiple disease associations, to learn the latent representation of diseases. Besides, Zhang *et al.* [16] considered the associations between diseases and non-coding RNAs, as well as functional associations and semantic associations between diseases to discover similar diseases.

Facilitated by disease similarity analysis, other researches were conducted. For example, Lei *et al.* [17] utilized the integration of disease semantic similarity and functional similarity, and adopted the bipartite network projection method to predict latent metabolite-disease associations. Moreover, Jarada *et al.* [18] leveraged disease-related similarity information, and combined the drug-related similarity information and the known drug-disease interaction information to predict drug-disease relationships.

It can be seen that various information is employed to analyze similar diseases. However, existing disease similarity analysis methods almost focus on one type of information, without a comprehensive description of the relationships among diseases.

2.2 Disease Information Retrieval

Information retrieval is an essential part of similar disease detection. Different types of disease information characterize disease features in different ways. Disease information retrieval can be divided into three main categories: the taxonomy information, the association information, and the literature information.

Some studies retrieved taxonomy information from disease terminologies, such as DO and the Human Phenotype

Ontology (HPO) [19]. For example, Cheng *et al.* [20] developed an online system, called *DisSim*, to explore significant similar diseases by measuring the similarities among DO terms. Moreover, in *SemFunSim* [11], DO was also used as a part to detect similar diseases. Besides, the information of HPO was employed for disease classification in Medical Subject Headings (MeSH) [21].

Some studies retrieved association information by building information networks. Specifically, Suthram *et al.* [22] proposed a quantitative framework to compare and contrast diseases by analyzing disease-related mRNA expression data and the human protein interaction network. Besides, Deng *et al.* [23] proposed a method *MultiSourcDSim* that integrates multiple disease similarity networks established based on disease associations to calculate disease similarity.

In addition, the clinical medical literature can provide detailed information about diseases, especially for new diseases. Kim *et al.* [24] proposed a literature-based method, called *LDDSim*, to estimate disease similarity. *LDDSim* uses all possible gene symbols and drug names in literature to characterize diseases and calculates feature values of diseases with the frequencies of co-occurrence of the two entities. Besides, Kafkas *et al.* [25] formed a database of pathogen and its phenotypes by extracting pathogen-disease relations from literature. What's more, Li *et al.* [26] presented a method to measure disease similarity, called *MedNetSim*, in which biomedical literature mining plays an important role.

Although many types of information describe diseases from different aspects, unfortunately, not all types of information are integrated to measure similar diseases.

3 Our Solution – MISSION

In this section, we present the detailed techniques used in *MISSION*. Figure 1 shows the framework of *MISSION*.

3.1 Disease Information Network based Similarity

In order to effectively calculate the similarities among diseases, we employ the disease information network, which is a typical heterogeneous information network, to model disease associations.

Definition 1 (Disease Information Network). A disease information network (DIN) is defined as an undirected graph $G = (V, E)$, in which V and E are the sets of nodes and edges between nodes, respectively. G is associated with a node type mapping $\phi : V \rightarrow A$ and an edge type mapping $\psi : E \rightarrow R$, where A refers to the type set of disease-related entities and R denotes the type set of all relations.

For the construction of DIN, we adopt the method of [27]. Please refer to [27] for the details of the process of building a disease information network, considering that this is not the focus of our study.

After constructing DIN, a critical point is to capture complex semantic relationships among diseases on DIN for calculating the similarities among diseases. Note that there are many approaches for disease similarity measurement (e.g., [28]), without depending on a specific similarity metric. Here, we adopt the meta structure, which has a strong ability to express complex relationships between two diseases [29].

Definition 2 (Disease Meta Structure). Given a DIN $G = (V, E)$, a disease meta structure $M = (\mathcal{V}, \mathcal{E}, d_i, d_j)$ is a sub-graph where $\mathcal{V} \subseteq V$ is a set of nodes and $\mathcal{E} \subseteq E$ is a set of edges. The meta structure M is a directed acyclic graph with a source disease node d_i and a target disease node d_j.

Then we denote $\mathcal{M} = \{M_1, M_2, ..., M_n\}$ as the set of meta structures between diseases d_i and d_j. We also define *disease meta structure instance set* between diseases d_i and d_j as $Ins_{d_i \rightarrow d_j}(M)$, which is a set of sub-graphs going from d_i to d_j induced by M. Please note that we design four meta structures based on DIN to express different relevance between two diseases, as illustrated in Figure 3.

Example 1. An example of disease information network is illustrated in Figure 2. There are a total of four node types (i.e., disease, gene, phenotype, and chemical). Figure 3 shows four designed meta structures. The corresponding multiple disease meta structure instances can be found in this disease information network. For example, the meta structure M_2 indicates that two diseases are associated with the same chemical. In addition, they have the same symptom at the same time. And, the corresponding two disease meta structure instances, i.e., "d_5-{c_3, p_2}-d_2" and "d_5-{c_5, p_2}-d_3" are shown in the shaded part of Figure 2.

Observation 1. In a disease information network, the more disease meta structure instances shared by two diseases, the more similar they are.

Example 2. In Figure 2, two diseases d_1 and d_4 share two disease meta structure instances, i.e., "d_1-{g_1, p_1}-d_4" and "d_1-{g_2, p_1}-d_4", while two diseases d_1 and d_2 have only one disease meta structure instance, i.e., "d_1-{g_2, p_3}-d_2". It can be seen that the relevance between d_1 and d_2 is weaker than the relevance between d_1 and d_4. In other words, d_4 is more similar to d_1 compared with d_2 to d_1.

Based on Observation 1, given a meta structure, we consider the similarity between any two diseases through the shared disease meta structure instances. Formally, given a DIN, the similarity between two diseases d_i and d_j based on the meta structure M can be defined as:

$$MetaSim_M(d_i, d_j) = |Ins_{d_i \rightarrow d_j}(M)|, \quad (1)$$

where the value of $|Ins_{d_i \rightarrow d_j}(M)|$ is the number of instances induced by meta structure M.

Typically, there may exist multiple different meta structures between two diseases. However, for many meta structures, the disease meta structure instances between two diseases appear less frequently. With the loss of generality, we choose the meta structure with the most disease meta structure instances to represent the relationships between two diseases. Thus, we calculate the similarity between two diseases based on DIN as:

$$NetSim(d_i, d_j) = \max_{M \in \mathcal{M}} (MetaSim_M(d_i, d_j)). \quad (2)$$

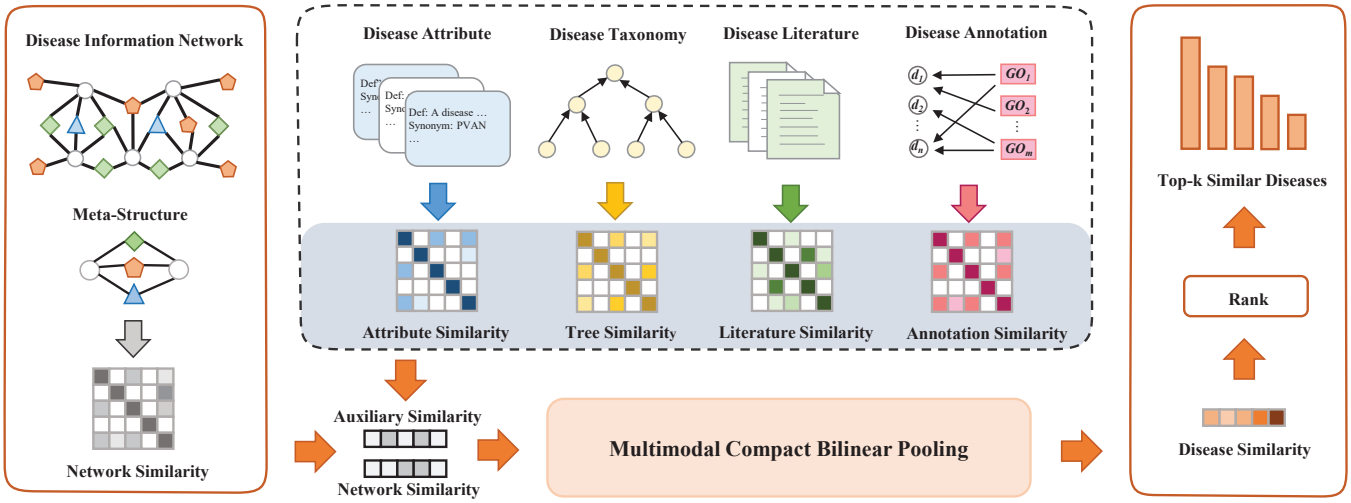


Fig. 1. The framework of *MISSION*.

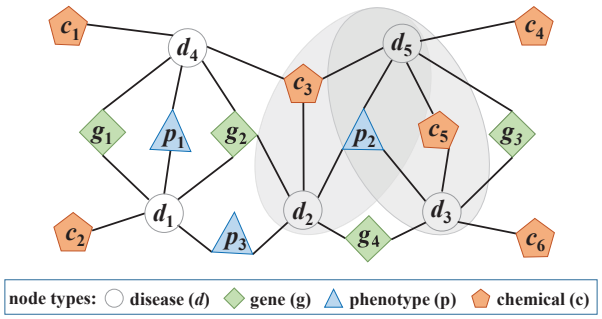


Fig. 2. An example of disease information network.

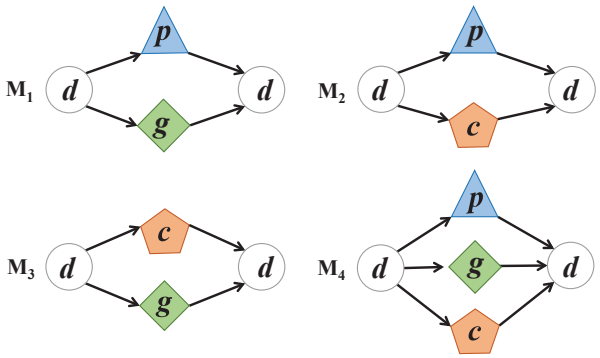


Fig. 3. The illustration of four meta structures.

3.2 Multimodal-Information based Similarity

In this section, we mainly focus on four types of multimodal-information, including the disease taxonomy, disease attributes, disease literature, and disease annotations. Note that, the disease taxonomy and disease attributes both come from DO database [7]. Disease literature is extracted from PubMed[1]. Disease annotations are obtained from GO database [9].

3.2.1 Disease Taxonomy Similarity

In DO database, diseases are linked into a hierarchical structure by a type of semantic relation, called 'IS_A' relation. We call such a hierarchical structure *disease taxonomy tree*, defined as:

Definition 3 (Disease Taxonomy Tree). A disease taxonomy tree (DT) is a polytree, which is a directed acyclic graph defined as $T = (D, \mathcal{E})$. D is the set of disease nodes, and $\mathcal{E}$ is the set of edges that represent the 'IS_A' semantic relation between diseases.

Then, we define the similarity between two diseases based on the disease taxonomy tree. As stated in [6], the similarity between two disease nodes d_i and d_j can be calculated by making full use of the information of their ancestors in the disease taxonomy tree. Let the set of ancestors of disease d be N_d. The semantic contribution of an ancestor $t \in N_d$ to disease d is $S_d(t)$, defined as:

$$S_d(t) = \begin{cases} 1 & t = d \\ \max\{w \cdot S_d(t') | t' \in \text{children of } t\} & t \neq d \end{cases}, \tag{3}$$

where node t' is a child of node t, and w represents the semantic contribution factor of ancestor t to t' and is set to 0.5 according to [6].

Thus, given two diseases d_i and d_j, their similarity based on the disease taxonomy tree is defined as:

$$TreeSim(d_i, d_j) = \frac{\sum\limits_{t \in N_{d_i} \cap N_{d_j}} (S_{d_i}(t) + S_{d_j}(t))}{\sum\limits_{t \in N_{d_i}} S_{d_i}(t) + \sum\limits_{t \in N_{d_j}} S_{d_j}(t)}, \tag{4}$$

where $\sum\limits_{t \in N_{d_i}} S_{d_i}(t)$ is the summation of all the contributions of N_{d_i} to disease d_i.

3.2.2 Disease Attribute Similarity

In addition to the hierarchical structure, DO also includes meta-data in the form of attributes. The meta-data can

1. https://pubmed.ncbi.nlm.nih.gov

provide abundant valuable information about diseases, and thus can be applied to analyze the similarities among diseases. Please note that some attributes among the meta-data are trivial due to many missing data. For better analysis, we select five attributes with less missing data, including *name*, *def*, *synonym*, *created_by*, and *creation_date*.

Next, it is required to comprehensively utilize multiple kinds of attribute information to characterize diseases. Here, we use Word2Vec [30], which is widely adopted to generate vector representation of words through fully considering the contextual information, to obtain feature vectors of diseases. Specifically, we first build an attribute corpora by generating sentences from disease attributes. For example, if the attribute of a disease d_i has a synonym s, the sentence "d_i synonym s" will be generated. Consequently, we use the pre-trained Word2Vec model [31] to assign a semantic to the attribute words (e.g., synonym). Next, the Skip-Gram model of Word2Vec is applied to retrain the pre-trained Word2Vec model based on the attribute corpus. Finally, for each disease d, we get a vector representation $\mathcal{F}_d^{att}$.

Accordingly, the similarity of disease pairs based on disease attributes can be calculated by utilizing the representation vectors of diseases d_i and d_j, denoted by:

$$AttSim(d_i, d_j) = \frac{\mathcal{F}_{d_i}^{att} \cdot \mathcal{F}_{d_j}^{att}}{|| \mathcal{F}_{d_i}^{att} || || \mathcal{F}_{d_j}^{att} ||}. \tag{5}$$

3.2.3 Disease Literature Similarity

Disease literature provides the latest rich and diverse information about diseases. Therefore, we can construct literature corpora to measure the similarities among diseases.

For a given disease, we crawl literature from PubMed website by using the disease name as the keyword to obtain query results sorted by relevance. Then we select the abstracts of the top 100 most relevant literature to construct the literature corpus of this disease. To reduce the noise in the raw literature, we perform the following preprocessing: (1) performing word segmentation processing; (2) removing the meaningless characters and words, such as punctuation marks and pronouns; (3) leveraging a Python package, WordNetLemmatizer, to transform several forms (e.g., tense) of a word into the dictionary form of the word.

After constructing literature corpora of all diseases, we use Latent Dirichlet Allocation (LDA) [32], which is a classical topic model, to extract hidden topics of literature corpora in the form of a probability distribution. Given the literature corpus of the disease d, the disease topic vector $\mathcal{F}_d^{lit}$ can be generated through the LDA model, where each dimension in this topic vector represents the probability distribution of the corresponding topic.

Correspondingly, for two diseases d_i and d_j, their literature-based similarity calculation can be converted to the cosine similarity between vector representations of these two diseases, as follows:

$$LitSim(d_i, d_j) = \frac{\mathcal{F}_{d_i}^{lit} \cdot F_{d_j}^{lit}}{|| \mathcal{F}_{d_i}^{lit} || || \mathcal{F}_{d_j}^{lit} ||}. \tag{6}$$

3.2.4 Disease Annotation Similarity

Since GO terms annotate the functions of genes involved in the occurrence of diseases, we can utilize GO terms to understand diseases from the characteristics of Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). With the genes as an intermediate conversion, each disease is annotated with a set of GO terms. So, we can use such disease annotations to measure the disease similarity.

Specifically, we need to calculate the IC value of each GO term first. Similar to [33], we define the count of the given GO term as the number of term hyponyms on the ontology structure. Based on the observation about the topology structure of GO, the count of GO term only depends on its child terms. Therefore, when the count of leaf terms is set to 1, the count of non-leaf terms can be calculated by recursively adding the count of children from bottom to top in the hierarchical structure, and given by:

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is a leaf term,} \\ \sum\limits_{z \in C_h(x)} f(z) & \text{otherwise,} \end{cases} \tag{7}$$

where $C_h(x)$ is the set of children of GO term x in the hierarchical structure of GO.

Consequently, the IC value of a term can be quantified as a function of the count of its hyponyms, denoted by:

$$IC(x) = -\lg(\frac{f(x)}{f(r)}), \tag{8}$$

where $f(r)$ is the count of root term in the ontology under consideration.

Based on the above calculation, each GO term is associated with its corresponding IC value. As each disease is annotated by a set of GO terms O_d, we can represent the feature vector of each disease by IC values of GO terms. Specifically, given a set of diseases D, we denote the feature vector of a disease $d_i \in D$ with respect to another disease $d_j \in D \backslash d_i$ as $\mathcal{F}_{d_i}^{ann} = [v_1, v_2, ..., v_m]$, where $m = |O_{d_i} \cup O_{d_j}|$, and each v_i corresponding to a GO term $o_i \in O_{d_i} \cup O_{d_j}$ is calculated by Equation 9.

$$v_i = \begin{cases} IC(o) & \text{if } o \in O_{d_i}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Subsequently, the similarity between two diseases d_i and d_j based on disease annotations can be calculated by $\mathcal{F}_{d_i}^{ann}$ and $\mathcal{F}_{d_j}^{ann}$, denoted by:

$$AnnSim(d_i, d_j) = \frac{\mathcal{F}_{d_i}^{ann} \cdot \mathcal{F}_{d_j}^{ann}}{|| \mathcal{F}_{d_i}^{ann} || || \mathcal{F}_{d_j}^{ann} ||}. \tag{10}$$

By performing the above similarity metrics for any two diseases, we can obtain four disease similarity matrices based on different types of multimodal-information. Since the distributions of similarity scores under different modalities are uneven, it is unreasonable to fuse the raw similarity scores. Thus, we perform normalization on each similarity matrix to adjust the distribution of values, enabling a balanced fusion of disease similarity features across multimodal-information. Without the loss of generalization, we employ the Hadamard product to obtain the similarity feature of multimodal-information, denoted by:

$$InfoSim(d_i, d_j) = TreeSim(d_i, d_j) * AttSim(d_i, d_j) * LitSim(d_i, d_j) * AnnSim(d_i, d_j). \quad (11)$$

Note that, we set the weight of each type of multimodal-information as 1.0 by default.

3.3 Similar Disease Query

In this section, we introduce how to interact the similarity feature based on DIN with the similarity feature based on multimodal-information for a similar disease query. Here, we adopt the idea of multimodal compact bilinear pooling [34].

According to Equations 2 and 11, for each disease $d \in D$, we calculate the similarity based on DIN and the similarity based on multimodal-information with respect to another diseases. In this case, we get two n-dimensional feature vectors of this disease, the similarity feature vector based on DIN $\mathbf{v}_{din}$ and the similarity feature based on multimodal-information $\mathbf{v}_{info}$, where n is the number of diseases. Relying on Count Sketch [35] projection function Ψ, each feature vector $\mathbf{v} \in \mathbb{R}^n$ can be projected to $\mathbf{y} \in \mathbb{R}^m$, where m is the output dimension. Correspondingly, for feature vectors $\mathbf{v}_{din}$ and $\mathbf{v}_{info}$, we can get the output vectors $\mathbf{y}_{din} = \Psi(\mathbf{v}_{din})$ and $\mathbf{y}_{info} = \Psi(\mathbf{v}_{info})$.

As stated in [34], the count sketch of the outer product of two vectors can be expressed as the convolution of each individual's count sketch. Meanwhile, the convolution is equivalent to the dot product in the frequency domain. Formally, the output vector of disease d with dimension m can be obtained by the following formula:

$$\mathcal{F}_d = FFT^{-1}(FFT(\mathbf{y}_{din}) \odot FFT(\mathbf{y}_{info})). \quad (12)$$

By this process, all elements of the two vectors based on DIN and multimodal-information interact with each other. In the end, we perform an element-wise signed squared root normalization.

Ultimately, given two diseases d_i and d_j, we calculate the similarity between them as:

$$MSim(d_i, d_j) = \frac{\mathcal{F}_{d_i} \cdot \mathcal{F}_{d_j}}{|| \mathcal{F}_{d_i} || || \mathcal{F}_{d_j} ||}. \quad (13)$$

So far, given a set of diseases D and a query disease $d \in D$, the top-k similar disease query can be quickly performed by calculating the similarity between d and another disease $d' \in D \setminus (d)$ based on Equation 13. Algorithm 1 shows the pseudo-code of top-k similar disease query.

Algorithm 1 Top-k Similar Disease Query

Input: d: the given query disease, k: the number of similar diseases, D: the set of diseases
Output: $\mathcal{R}$: the results of similar disease query

```
R ← ∅
S ← ∅
v_din ← Obtain the similarity feature of d based on DIN according to Section 3.1
v_info ← Obtain the similarity feature of d based on multimodal-information according to Section 3.2
Generate F_d according to Equation 12
for d' ∈ D \ d do
    v'_din ← obtain the similarity feature of d' based on DIN according to Section 3.1
    v'_info ← obtain the similarity feature of d' based on multimodal-information according to Section 3.2
    Generate F'_d according to Equation 12
    Calculate similarity sim according to Equation 13
    S ← S ∪ {(d', sim)}
end for
Sort S by sim in descending order
R ← Find the top-k similar diseases based on S
return R
```

4 Experiments and Discussion

We conduct experiments on real-world datasets to answer the following questions:

- **Q1**: How well does *MISSION* perform compared with the baselines in detecting similar diseases?
- **Q2**: Can multimodal-information further improve the performance in the detection of similar diseases?
- **Q3**: Does *MISSION* have the ability to detect similar diseases even with different levels of information richness?

4.1 Experimental Settings

4.1.1 Datasets

Different types of information were collected from various sources to validate the effectiveness of our proposed *MISSION*. The followings describe the details of each type of information.

First, a total of three types of disease associations were employed to construct the disease information network, i.e., disease-gene associations, disease-chemical associations, and disease-phenotype associations, which were derived from the DisGeNet database [36], Comparative Toxicogenomics Database (CTD) [37], and HSDN [38], respectively. Besides, the data of terms was downloaded from DO database [7], containing 162639 disease terms. After the following two steps: (1) screening the co-occurring diseases in all the disease associations, and (2) mapping the various IDs of all diseases to the corresponding ID in the DO database uniformly through utilizing the mapping relationships obtained from DisGeNet and DO database, 1754 diseases were finally extracted. The main statistics of the disease information network are summarized in Table 1.

Then, the abstracts of a total number of 253960 literature related to 1754 diseases were crawled from PubMed, which is a free resource supporting the retrieval of biomedical and life science literature. In addition, GO database [9] provided GO terms, and the annotation relationships of GO terms to genes were obtained from Gene Ontology Annotation (GOA) database [39]. Then, by taking the disease-gene associations obtained from the DisGeNet [36] database as intermediate conversions, the annotation relationships between 1754 diseases and 17920 GO terms were obtained.

4.1.2 Evaluation Metric

To evaluate *MISSION* quantitatively, the widely used metric AUC, which is defined as the area under the receiver oper-

TABLE 1
Statistics of the Disease Information Network

Dataset	Type of Nodes	#Nodes	#Edges	Source
Disease-Gene	disease gene	1754 17382	397702	DisGeNet [36]
Disease-Chemical	disease chemical	1754 3953	304443	CTD [37]
Disease-Phenotype	disease phenotype	1754 2029	43687	HSDN [38]

ating characteristic (ROC) curve enclosed by the coordinate axis, was adopted. Compared with the ROC curve, AUC can more clearly indicate which method performs better. And, the higher the AUC score, the better the performance. Besides, the benchmark set we used was integrated from two manually checked datasets of disease pairs, one obtained from the study of Suthram *et al*. [22] and the other derived from the work of Pakhomov *et al*. [40]. In total, the benchmark set contains 47 diseases and 70 disease pairs of high similarity. Then, the negative sample set, including 200 disease pairs, was randomly generated based on the disease set by excluding the disease pairs that already exist in the benchmark set. To weaken the impact of the bias caused by occasionality, the experiments were repeated 100 times on the newly generated negative sample set for each time, and then the average AUC score was calculated to represent the performance of *MISSION*.

4.1.3 Comparison Methods

To demonstrate the effectiveness of *MISSION*, six methods were considered as baselines, including homogeneous relationship based methods, i.e., Resnik's [13], Wang's [6], Lin's [14] and *PSB* [8], as well as heterogeneous relationships based methods, i.e., *RADAR* [10] and *SemFunSim* [11], which are all introduced in Section 2. Moreover, *MISSION-S*, a variant of *MISSION*, was also regarded as a baseline, which only bases on DIN (i.e., only using disease associations) without any multimodal-information. This setting ensures a comprehensive evaluation for the performance of our proposed method.

4.1.4 Parameter Settings

Generally, the parameters were set according to the default and optimal experiment effect. According to [31], the parameters used to retrain the Skip-Gram model were set as follows: the dimension of the word embedding is 200, the number of iterations is 100, the minimum count is 1, and the window is 5. Besides, the number of LDA topics was set to 85 based on the optimal experimental result. For each type of variant, to make the optimal output dimension m applicable for most variants, we set the output dimension m of the unimodal-information-aided variants, bimodal-information-aided variants, trimodal-information-aided variants and *MISSION* to 8000, 2000, 1000 and 1000, respectively.

All experiments were conducted on a PC with an Intel Xeon E5-2678 v3 2.50 GHz CPU and 64 GB main memory, running the Ubuntu 19.04. All algorithms were implemented in Python and compiled by Python 3.7. The source codes and dataset of *MISSION* are available on https://github.com/MangoXu98/MISSION.

TABLE 2
The Distribution of AUC Scores for Each Method

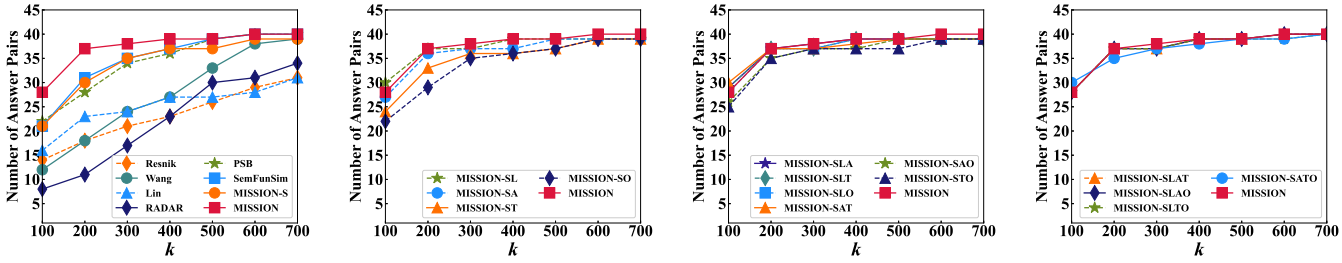
Method	AUC		
	Max	Min	Average
Resnik's	0.6933	0.624	0.6528
Wang's	0.7323	0.6444	0.6908
Lin's	0.7323	0.6831	0.7057
RADAR	0.9084	0.844	0.8741
PSB	0.9191	0.8634	0.8935
SemFunSim	0.9633	0.8987	0.9354
MISSION-S	0.9407	0.9002	0.9212
MISSION	**0.9766**	**0.9436**	**0.9627**

4.2 Performance Comparison with Baselines (Q1)

The disease similarity scores among 3525 diseases calculated by these methods (i.e., Lin's, Wang's, Resnik's, *PSB*, and *SemFunSim*) are prepared in the system *DincRNA* [41], which can be downloaded directly. Additionally, the disease similarity scores calculated by *RADAR* were obtained by utilizing the disease associations collected in our work. Considering the difference between the diseases in *DincRNA* and the diseases collected in our work, we picked the shared diseases from the two disease sets, and 834 common diseases were finally obtained. Correspondingly, 38 diseases and 40 disease pairs were extracted from the benchmark set and new negative sample sets were randomly generated. Based on this, we verify the effectiveness of *MISSION* from two aspects as follows.

First, we compared the performance of *MISSION* with baselines based on the AUC score. The comparison results are reported in Table 2, from which we can observe that our proposed method *MISSION* achieves the best performance, indicating the effectiveness and accuracy of *MISSION* in detecting similar diseases. Furthermore, the remaining findings are summarized as follows: (1) Although *MISSION-S* outperforms most baselines, it fails to exceed *SemFunSim* that simultaneously leverages disease associations and taxonomy. This finding indicates that considering multiple types of information is necessary, which also validates our motivation of incorporating multimodal-information. (2) Based on DIN, both *RADAR* and *MISSION-S* produce good results. But in contrast, *RADAR* is slightly inferior to *MISSION-S*, showing the superiority of meta structure in mining the network structure. (3) Most homogeneous relationship based methods, i.e., Resnik's, Wang's, and Lin's, perform the worst. The reason may lie in that only one type of information fails to adequately describe the relationships between diseases, making these methods unable to identify similar diseases well. (4) The removal of multimodal-information of *MISSION* (i.e., *MISSION-S*) results in significant performance degradation, showing the importance of multimodal-information for performance enhancement.

Next, for further comparison, another experiment was also carried out. Specifically, given the value of k in the



(a) Baselines (b) Unimodal-information (c) Bimodal-information (d) Trimodal-information

Fig. 4. Performance analysis of *MISSION* compared with (a) baselines, (b) unimodal-information-aided variants, (c) bimodal-information-aided variants, and (d) trimodal-information-aided variants respectively.

TABLE 3
The Distribution of AUC Scores for Each Type of Multimodal-Information and Each Variant of *MISSION*

Types	Multimodal-Information	AUC			Variant	AUC		
		Max	Min	Average		Max	Min	Average
-	-	-	-	-	MISSION-S	0.9407	0.9002	0.9212
Unimodal	L	0.9034	0.8099	0.8504	MISSION-SL	0.969	0.9386	0.9546
	A	0.8871	0.8239	0.8553	MISSION-SA	0.9626	0.9309	0.948
	T	0.8234	0.74	0.7822	MISSION-ST	0.9506	0.9191	0.9375
	O	0.7311	0.6603	0.6947	MISSION-SO	0.9401	0.9018	0.9244
Bimodal	LA	0.9279	0.8604	0.8903	MISSION-SLA	0.9728	0.9431	0.9587
	LT	0.9243	0.8473	0.88	MISSION-SLT	0.9739	0.9434	0.9586
	LO	0.912	0.8276	0.8624	MISSION-SLO	0.9733	0.9444	0.9585
	AT	0.9128	0.8552	0.8833	MISSION-SAT	0.9719	0.94	0.9571
	AO	0.8873	0.8268	0.8594	MISSION-SAO	0.9671	0.9356	0.9538
	TO	0.8278	0.7429	0.7868	MISSION-STO	0.9534	0.9178	0.939
Trimodal	LAT	**0.9431**	**0.8798**	**0.9063**	MISSION-SLAT	**0.9769**	0.943	**0.9629**
	LAO	0.9313	0.8675	0.8942	MISSION-SLAO	0.9774	0.9435	0.9626
	LTO	0.928	0.8531	0.8833	MISSION-SLTO	0.9763	0.9423	0.9622
	ATO	0.9059	0.843	0.8759	MISSION-SATO	0.9739	0.9395	0.9601
-	LATO	0.9418	0.8788	0.9061	MISSION	0.9766	**0.9436**	0.9627

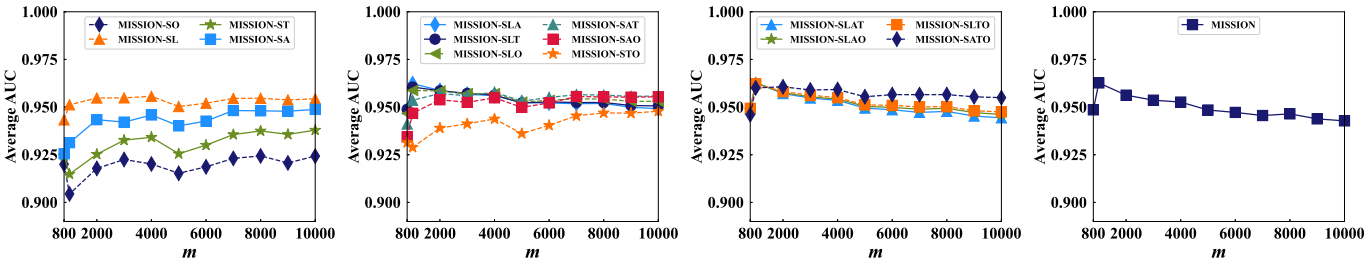
similar disease query, we checked how many disease pairs in the benchmark set (i.e., answer pairs) can satisfy this query. For example, for the disease pair *(Asthma, Chronic Obstructive Pulmonary Disease)* in the benchmark set, the ranking of disease *Chronic Obstructive Pulmonary Disease* relative to disease *Asthma* is 1. If the given value of k is 100, then the number of answer pairs for this top-100 query is increased by 1, since the relative ranking of *(Asthma, Chronic Obstructive Pulmonary Disease)* is within 100.

The experiment results are shown in Figure 4(a), and we observe that: (1) *MISSION* consistently outperforms all baselines (i.e., always finding the more answer pairs than all baselines), which demonstrates its effectiveness in performing top-k similar disease query; (2) when k is 200, *MISSION* finds nearly 92.5% disease pairs in the benchmark set, while all baselines cannot; and (3) the performance of *MISSION-S* is basically the same as that of *SemFunSim*, and the homogeneous relationship based methods perform worse. Overall, the performance of all methods is roughly consistent with the results based on the AUC score.

4.3 Performance Comparison with Variants (Q2)

In order to further explore the enhancement effect of different combinations of each type of multimodal-information on DIN for similar disease detection, a detailed comparative analysis was carried out. To be specific, by utilizing each type of multimodal-information as supplementary information, three types of variants of *MISSION* were obtained accordingly, including (1) unimodal-information-aided variants: only one type of multimodal-information was adopted, and 4 variants were produced, (2) bimodal-information-aided variants: 6 variants were generated by combining the two types of multimodal-information in pairs, and (3) trimodal-information-aided variants: three types of multimodal-information were combined simultaneously, and the corresponding 4 variants were obtained. Similar to Section 4.2, the results were analyzed from two aspects.

Each type of multimodal-information is represented by a capital letter, namely S stands for disease associations, L for disease literature, A for disease annotations, T for disease attributes, and O for disease taxonomy. Besides, the connected letters indicate that the corresponding type of



(a) Unimodal-information (b) Bimodal-information (c) Trimodal-information (d) Multimodal-information

Fig. 5. Performance analysis of different variants with respect to parameter m, namely (a) unimodal-information-aided variants, (b) bimodal-information-aided variants, (c) trimodal-information-aided variants, and (d) multimodal-information-aided variants (i.e., *MISSION*).

multimodal-information is adopted simultaneously. For example, *LA* means that both disease literature and annotation are utilized to measure disease similarity, and *MISSION-SLA* represents that disease literature and annotations serve as supplementary information for the disease information network at the same time. The experimental results of each variant divided by type are presented in Table 3, in which the best and second-best results are highlighted in bold and underline.

First, it can be seen that the average AUC scores based on each type of multimodal-information (i.e., *L* (only using disease literature), *A* (only using disease annotations), *T* (only using disease attributes) and *O* (only using disease taxonomy)) are 0.8504, 0.8553, 0.7822 and 0.6947, respectively, while the average AUC score of *MISSION-S* (only using disease associations) is 0.9212. This indicates that the disease associations describe the disease relationships in more detail than other types of information, that is, only relying on disease associations can also achieve a good result, which supports our idea of employing disease associations as a basis and others as auxiliary multimodal-information. Then, we hold an independent analysis of each type of variant below.

Unimodal-Information-Aided Variants. By adding each type of multimodal-information individually as supplementary information, the performance of unimodal-information-aided variants is improved to varying degrees in comparison with *MISSION-S*. Among the multimodal-information, it can be seen that the variant aided by disease literature performs best, while the enhancement effect of disease taxonomy is the worst. The reason is that disease taxonomy can only provide a few supplementary information about disease classification. Notably, although the AUC score of *L* is slightly lower than that of *A*, the variant *MISSION-SL* achieves a better performance than *MISSION-SA*. In other words, for DIN, disease literature contributes more supplementary information than disease annotations. This is also in line with our expectations because literature provides more supplementary information in the form of text (e.g., up-to-date disease information).

Bimodal-Information-Aided Variants. By adding another type of multimodal-information to each of the unimodal-information-aided variants, each variant achieves better performance. The variants aided with the combination of disease literature and another single type of multimodal-information have almost the same performance. Perhaps, this is because another type of multimodal-information contributes little useful information to the variant *MISSION-SL*. Moreover, when incorporating the disease taxonomy, only a slight performance improvement is observed in either auxiliary bimodal-information or variants. This is probably because little valuable information was provided by the disease taxonomy.

Trimodal-Information-Aided Variants. When adopting three types of multimodal-information at the same time, there is a general performance improvement for all trimodal-information-aided variants and the performance gap is narrowed. This also demonstrates that it is effective to adopt various types of multimodal-information to enhance DIN in similar disease detection. A further novel finding is that compared with *MISSION-SLAT*, the performance of *MISSION* decreases slightly with the addition of disease taxonomy. The reason might be that the information contained in *MISSION-SLAT* is already abundant, and considering disease taxonomy with a low confidence level introduce external noise instead.

Next, the experiment results of checking how many disease pairs satisfy the similar disease query for the given value of k are presented in Figure 4(b)(c)(d). Obviously, *MISSION* performs better than most of its variants, and variants with more types of multimodal-information generally have better performance than those with fewer types of multimodal-information. Moreover, when the number of types are more than one, the performance of variants is roughly close to that of *MISSION*. However, *MISSION* performs slightly worse than a few variants when k is 100. This is probably because *MISSION* raises the ranking of some disease pairs (e.g., disease subtypes) that are not recorded in the benchmark set but of highly similar, therefore lowing the ranking of corresponding disease pairs in the benchmark set.

4.4 Parameter Sensitivity

Here, we investigated how the output dimension m affects the performance of the model. And, the performances of each type of variant with changing the output dimension m are presented in Figure 5.

As can be seen from Figure 5(a), the performances of the unimodal-information-aid variants do not fluctuate much and the best results are generally achieved when m reaches

TABLE 4
Top-*k* Similar Diseases for the Given Queries

Query	Top-3 Results
Familial Combined Hyperlipidemia	Abdominal Obesity-metabolic Syndrome 1 Endometrial Cancer Familial Hypercholesterolemia
Scleroderma	Systemic Scleroderma Autoimmune Hypersensitivity Disease Cerebral Infarction
Alzheimer's Disease	Alzheimer's Disease 14 Alzheimer's Disease 13 Alzheimer's Disease 15
Asthma	Chronic Obstructive Pulmonary Disease Immune System Disease Endometriosis
Bronchitis	Pneumonia Myocardial Infarction Arthritis

8000. As shown in Figure 5(b), the AUC scores of the bimodal-information-aid variants have very little difference and generally perform best when m is 2000. Besides, for the trimodal-information-aid variants and *MISSION*, when the dimension m exceeds 1000, the performance shows a downward trend, and the best performance is achieved when m is 1000 as presented in Figure 5(c) and Figure 5(d).

Overall, the performance trends of the same type of variants along the dimension are similar. Moreover, our method *MISSION* is not critically sensitive to the output dimension m with a small performance fluctuation.

4.5 Case Study (Q3)

Two case studies were carried out: (1) top-*k* similar disease query; and (2) similar disease detection with different information richness.

4.5.1 Similar Disease Query

Five diseases, i.e., *Familial Combined Hyperlipidemia*, *Scleroderm*, *Alzheimer's Disease*, *Asthm* and *Bronchiti* were randomly selected from the benchmark disease set as the target diseases to perform top-*k* similar disease query. We retrieved the corresponding top-3 similar diseases based on *MISSION*. The query results are reported in Table 4. In general, all the top-3 diseases are strongly related to the given query disease, especially for *Alzheimer's Disease* that the most relevant diseases are all its subtypes. For diseases *Asthm* and *Bronchiti*, the most similar diseases are *Chronic Obstructive Pulmonary Disease* and *Pneumoni* respectively, which are recorded in benchmark set. Besides, for disease *Scleroderm*, the most relevant disease *Systemic Scleroderma* is one of its forms. Moreover, the most similar disease for *Familial Combined Hyperlipidemia* is *Abdominal Obesity-metabolic Syndrome 1*. Although this disease pair does not exist in the benchmark set, some studies have indicated that *Familial Combined Hyperlipidemia* develops against a background of *Abdominal Obesity* [42].

TABLE 5
The Ranking of Disease Pairs in Each Stage

Disease Pair	Ranking		
	Stage 1	Stage 2	Stage 3
(Leukopenia, Pneumonia)	102	17	5
(Cataract, Pancreatitis)	820	21	14
(Asthma, Chronic Obstructive Pulmonary Disease)	159	1	1
(Chronic Progressive External Ophthalmoplegia, Dilated Cardiomyopathy)	281	62	14
(Myocardial Infarction, Chronic Obstructive Pulmonary Disease)	187	37	19

4.5.2 Similar Disease Detection with Different Information Richness

Based on the process of disease research, we divided information into the three stages to detect similar diseases according to the different degree of information richness, as follows: (1) Stage 1: only disease-phenotype associations were obtained through clinical when appearing a new disease; (2) Stage 2: with in-depth research, extensive literature recording the latest research was published. Besides, some disease annotations and associations with other entities were discovered; (3) Stage 3: the DO and more disease associations were established and enriched.

Five disease pairs were selected from the benchmark set for research, i.e., *(Leukopenia, Pneumonia)*, *(Cataract, Pancreatitis)*, *(Asthma, Chronic Obstructive Pulmonary Disease)*, *(Chronic Progressive External Ophthalmoplegia, Dilated Cardiomyopathy)*, *(Myocardial Infarction, Chronic Obstructive Pulmonary Disease)*. The information corresponding to the above three stages was specified as follows: disease-phenotype associations (Stage 1), the disease associations with phenotypes and genes, disease annotations and literature (Stage 2), and all information (Stage 3).

The ranking results of disease pairs in each stage are presented in Table 5, in which this ranking refers to the relevance ranking of the second disease to the first one. Overall, *MISSION* can perform the similar disease query in all the three stages with different information richness, although the result is different. Furthermore, we observe that: (1) all the five target disease pairs are generally low-ranking in the first stage; (2) in the second and third stage, when more information is incorporated, the ranking of disease pairs increases accordingly, which is also in line with our expectations; (3) for some diseases, the most similar diseases can be found only by providing partial information. Take the disease pair *(Asthma, Chronic Obstructive Pulmonary Disease)* as an example, the similarity ranking is already 1 in the second stage. Nonetheless, for most diseases, more information is needed to be considered, and *MISSION* can meet the needs for mining similar diseases from a large amount of information.

To sum up, *MISSION* gains performance improvement

accordingly as more types of information are available. And, *MISSION* is able to detect similar diseases in various stages with different information richness of disease research.

4.6 Efficiency Analysis

We further performed an efficiency analysis of similarity measurement based on each type of information by randomly choosing 20, 40, 60, 80, 100 percent diseases in the dataset respectively.

As shown in Figure 6, the runtime increases with the number of diseases, although the growth rate differs from one another. Besides, the runtime of similarity measurement based on disease annotations (i.e., A) grows fast relatively, and the runtime of some types of information reaches 10^3 s when all diseases are adopted. Fortunately, they can work in parallel due to mutual independence, reducing the total time-cost of *MISSION*. Furthermore, the calculation results of each type of information can be saved offline, so that the subsequent similar disease query can be quickly performed.

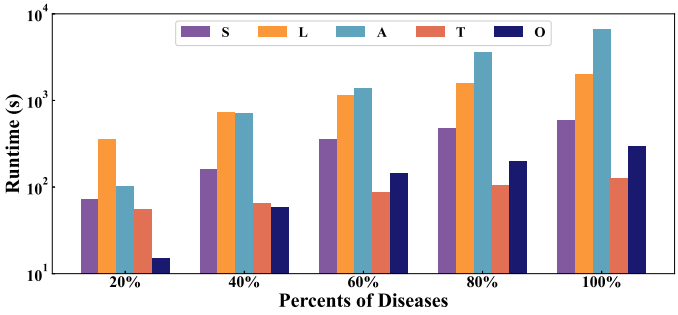


Fig. 6. Runtime w.r.t. the number of diseases.

5 Conclusion

Similar disease discovery can deepen our understanding of the field of bioinformatics. In order to flexibly integrate multiple types of information to comprehensively describe the characteristics of diseases, we propose a novel approach, *MISSION*, to perform the top-k similar disease query based on the integrative information. *MISSION* considers multiple types of disease-related information, i.e., disease taxonomy, attributes, literature and annotations, to enhance the disease information network, thereby providing a more reliable description of the relevance between diseases. Extensive experimental results on the real-world datasets suggest that *MISSION* outperforms all baselines, demonstrating its superiority. Besides, the performance comparison with variants confirms our idea of incorporating multimodal-information is helpful. Meanwhile, we further validate the effectiveness of *MISSION* in the two relevant case studies.

For future work, we intend to focus on the following tasks. First, we will look for a suitable method to automatically discover meaningful meta structures from the disease information network, without requiring domain expert knowledge to design manually. Besides, more types of information (e.g., electronic health records data) can be considered to further improve the accuracy of similar disease detection. Moreover, it would be interesting to explore the contribution rate of each type of multimodal-information to *MISSION*.

References

[1] M. G. Dozmorov, "Disease classification: from phenotypic similarity to integrative genomics and beyond," *Briefings in Bioinformatics*, vol. 20, no. 5, pp. 1769–1780, 2019.

[2] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, and L. I. Furlong, "Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases," *PLoS One*, vol. 6, no. 6, p. e20284, 2011.

[3] J. Zhu, Y. Qin, T. Liu, J. Wang, and X. Zheng, "Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles," *BMC Bioinformatics*, vol. 14, no. S-5, p. S5, 2013.

[4] M. Iida, M. Iwata, and Y. Yamanishi, "Network-based characterization of disease-disease relationships in terms of drugs and therapeutic targets," *Bioinformatics*, vol. 36, no. Supplement-1, pp. i516–i524, 2020.

[5] A. P. Quimbaya, R. A. Gonzalez, W. Bohórquez, O. M. Muñoz, O. M. García, and D. Londoño, "Improving decision-making for clinical research and health administration," in *Engineering and management of IT-based service systems*, 2014, vol. 55, pp. 179–200.

[6] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[7] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein *et al.*, "Human Disease Ontology 2018 update: classification, content and workflow expansion," *Nucleic Acids Research*, vol. 47, no. D1, pp. D955–D962, 2019.

[8] S. Mathur and D. Dinakarpandian, "Finding disease similarity based on implicit semantic similarity," *Journal of biomedical informatics*, vol. 45, no. 2, pp. 363–371, 2012.

[9] S. Carbon, E. Douglass, N. Dunn, B. M. Good, N. L. Harris, S. E. Lewis, C. J. Mungall, S. N. Basu, R. L. Chisholm, R. J. Dodson *et al.*, "The gene ontology resource: 20 years and still going strong," *Nucleic Acids Research*, vol. 47, no. D1, pp. D330–D338, 2019.

[10] R. Qin, L. Duan, H. Zheng, J. Li-Ling, K. Song, and Y. Zhang, "An ontology-independent representation learning for similar disease detection based on multi-layer similarity network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 183–193, 2021.

[11] L. Cheng, J. Li, P. Ju, J. Peng, and Y. Wang, "SemFunSim: A new method for measuring disease similarity by integrating semantic and gene functional association," *PLoS One*, vol. 9, no. 6, pp. 1–11, 2014.

[12] W. Xu, L. Duan, H. Zheng, J. Li-Ling, W. Jiang, M. Huang, and Y. Zhang, "MISSION: Multimodal-information-aided similar disease detection based on disease information network," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2020, pp. 369–374.

[13] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1995, pp. 448–453.

[14] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the International Conference on Machine Learning*, 1998, pp. 296–304.

[15] S. Mathur and D. Dinakarpandian, "Automated ontological gene annotation for computing disease similarity," *Summit on translational bioinformatics*, vol. 2010, pp. 12–16, 2010.

[16] N. Zhang, L. Juan, and T. Zang, "NCRR: A novel method for measuring disease similarity based on non-coding RNA regulation," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2020, pp. 1700–1707.

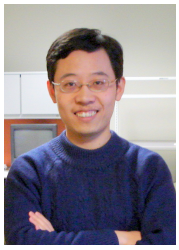
[17] X. Lei and C. Zhang, "Predicting metabolite-disease associations based on linear neighborhood similarity with improved bipartite network projection algorithm," *Complexity*, vol. 2020, pp. 9 342 640:1–9 342 640:11, 2020.

[18] T. N. Jarada, J. G. Rokne, and R. Alhajj, "SNF-NN: Computational method to predict drug-disease interactions using similarity network fusion and neural networks," *BMC Bioinformatics*, vol. 22, no. 1, p. 28, 2021.

[19] S. Köhler, L. Carmody, N. Vasilevsky, J. O. B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry *et al.*, "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1018–D1027, 2019.

[20] L. Cheng, Y. Jiang, Z. Wang, H. Shi, J. Sun, H. Yang, S. Zhang, Y. Hu, and M. Zhou, "DisSim: An online system for exploring significant similar diseases and exhibiting potential therapeutic drugs," *Scientific Reports*, vol. 6, no. 1, p. 30024, 2016.
[21] H. J. Lowe and G. O. Barnett, "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches," *The Journal of the American Medical Association*, vol. 271, no. 14, pp. 1103–1108, 1994.
[22] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, and A. J. Butte, "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets," *PLoS Computational Biology*, vol. 6, no. 2, pp. 1–10, 2010.
[23] L. Deng, D. Ye, J. Zhao, and J. Zhang, "MultiSourcDSim: An integrated approach for exploring disease similarity," *BMC Medical Informatics and Decision Making*, vol. 19, no. Suppl 6, pp. 269–269, 2019.
[24] H. Kim, Y. Yoon, J. Ahn, and S. Park, "A literature-driven method to calculate similarities among diseases," *Computer Methods Programs Biomedical*, vol. 122, no. 2, pp. 108–122, 2015.
[25] S. Kafkas and R. Hoehndorf, "Ontology based mining of pathogen-disease associations from literature," *Biomedical Semantics*, vol. 10, no. 1, pp. 15:1–15:5, 2019.
[26] P. Li, Y. Nie, and J. Yu, "Fusing literature and full network data improves disease similarity computation," *BMC Bioinformatics*, vol. 17, p. 326, 2016.
[27] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of The VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.
[28] Y. Bai, H. Ding, S. Bian, T. Chen, Y. Sun, and W. Wang, "SimGNN: A neural network approach to fast graph similarity computation," in *International Conference on Web Search and Data Mining*, 2019, pp. 384–392.
[29] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li, "Meta structure: computing relevance in large heterogeneous information networks," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1595–1604.
[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Conference on Neural Information Processing Systems*, 2013, pp. 3111–3119.
[31] S. F. Zohra, G. Xin, and H. Robert, "OPA2Vec: Combining formal and informal content of biomedical ontologies to improve similarity-based prediction," *Bioinformatics*, vol. 35, no. 12, pp. 2133–2140, 2019.
[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[33] P. Zhang, J. Zhang, H. Sheng, J. J. Russo, B. Osborne, and K. H. Buetow, "Gene functional similarity search tool (GFSST)," *BMC Bioinformatics*, vol. 7, p. 135, 2006.
[34] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 457–468.
[35] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 239–247.
[36] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, vol. 48, no. D1, pp. D845–D855, 2020.
[37] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database: update 2019," *Nucleic Acids Research*, vol. 47, no. D1, pp. D948–D954, 2019.
[38] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," *Nature Communications*, vol. 5, no. 1, pp. 1–10, 2014.
[39] C. Evelyn, M. Michele, B. Daniel, L. Vivian, D. Emily, M. John, B. David, H. Nicola, L. Rodrigo, and A. Rolf, "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Research*, vol. 32, no. Database-Issue, pp. 262–266, 2004.
[40] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. B. Melton, "Semantic similarity and relatedness between clinical terms: an experimental study," in *AMIA Annual Symposium Proceedings Archive*, vol. 2010, 2010, pp. 572–576.
[41] C. Liang, H. Yang, S. Jie, Z. Meng, and J. Qinghua, "DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 2018.
[42] C. J. Van Der Kallen, C. Voors-Pette, and T. W. De Bruin, "Abdominal obesity and expression of familial combined hyperlipidemia," *Obesity research*, vol. 12, no. 12, pp. 2054–2061, 2004.

Wuli Xu received her bachelor's degree in Software Engineering from Hunan University, China, in 2019. She is currently working towards her master degree in the School of Computer Science, Sichuan University, China. Her research interests include bioinformatics and data mining.



Lei Duan received his B.Sc. and PhD degrees both in Computer Science from Sichuan University, China, in 2003 and 2008, respectively. He was a visiting PhD student in the Department of Computer Science and Engineering, Wright State University, Dayton, Ohio from 2007 to 2008, and was a visiting scholar in the School of Computing Science, Simon Fraser University, Canada, from 2012 to 2013. He is currently a Professor in the School of Computer Science, Sichuan University, China. His research interests include data mining, knowledge management, evolutionary computation, bioinformatics and health-informatics.

Huiru Zheng IEEE Senior Member, is a Professor of Computer Science with School of Computing at Ulster University, UK; and a Fellow of the UK Higher Education Academy. She was awarded a PhD in Bioinformatics in 2003 and a Postgraduate Certificate in Teaching in Higher Education in 2005 from Ulster University. Prof. Zheng is an active researcher in bioinformatics and healthcare informatics. Within her broad interests in data mining, data integration, machine learning and healthcare decision support, Prof. Zheng has a particular research interest and expertise in integrative data analytics in the field of systems biology, and intelligent data analysis and assistive technology to support healthcare and independent living. She has published over 250 peer reviewed scientific research papers.

Jesse Li-Ling received the MD degree from West China University of Medical Sciences, China, in 1993 and the PhD degree from the University of Newcastle upon Tyne, U.K. in 2000. He conducted his postdoctoral research at Tsinghua University, China from 2001 to 2003. In 2007, he was promoted to full professor. He is currently working at Sichuan University (State Key Laboratory of Biotherapy), China, and his main research interests include medical genetics, bioinformatics and Traditional Chinese Medicine.

Weipeng Jiang received the bachelor's degree from in Computer Science from Sichuan University, China, in 2020. He is currently work in towards the master's degree in the School of Computer Science, Sichuan University, China. His research interests include data mining and Knowledge graph.

Yidan Zhang received the bachelor's degree in Biological Sciences and Software Engineering from Sichuan University, China, in 2018. She is currently working towards the PhD degree in the School of Computer Science, Sichuan University, China. Her research interests include bioinformatics and biomedicine.

Tingting Wang received the master's degree in Computer Science from Sichuan University, China, in 2020. She is currently working towards the PhD degree in the School of Computer Science & IT, RMIT University, Australia. Her research interests include data mining and database.

Ruiqi Qin received the master's degree from the School of Computer Science, Sichuan University, China, in 2020. She is currently working at SAP (China) Co., Ltd. Chengdu Branch. Her research interests include bioinformatics and data mining.