Bioinformatics, 37(8), 2021, 1156–1163
doi: 10.1093/bioinformatics/btaa913
Advance Access Publication Date: 23 October 2020
Original Paper

Data and text mining

A representation model for biological entities by fusing structured axioms with unstructured texts

Peiliang Lou [1,2], YuXin Dong[1], Antonio Jimeno Yepes[3] and Chen Li[1,4,*]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China, [2]Key Laboratory of Intelligent Networks and Network Security (Xi'an Jiaotong University), Ministry of Education, Xi'an, Shaanxi 710049, China, [3]IBM Research Australia, Southbank, VIC 3006, Australia and [4]National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 26, 2020; revised on September 4, 2020; editorial decision on October 9, 2020; accepted on October 13, 2020

Abstract

Motivation: Structured semantic resources, for example, biological knowledge bases and ontologies, formally define biological concepts, entities and their semantic relationships, manifested as structured axioms and unstructured texts (e.g. textual definitions). The resources contain accurate expressions of biological reality and have been used by machine-learning models to assist intelligent applications like knowledge discovery. The current methods use both the axioms and definitions as plain texts in representation learning (RL). However, since the axioms are machine-readable while the natural language is human-understandable, difference in meaning of token and structure impedes the representations to encode desirable biological knowledge.

Results: We propose ERBK, a RL model of bio-entities. Instead of using the axioms and definitions as a textual corpus, our method uses knowledge graph embedding method and deep convolutional neural models to encode the axioms and definitions respectively. The representations could not only encode more underlying biological knowledge but also be further applied to zero-shot circumstance where existing approaches fall short. Experimental evaluations show that ERBK outperforms the existing methods for predicting protein–protein interactions and gene–disease associations. Moreover, it shows that ERBK still maintains promising performance under the zero-shot circumstance. We believe the representations and the method have certain generality and could extend to other types of bio-relation.

Availability and implementation: The source code is available at the gitlab repository https://gitlab.com/BioAI/erbk.

Contact: cli@xjtu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

To address the growing need of automatically discovering knowledge and relationships of bio-entities, structured semantic resources, for example, biological knowledge bases (KBs) (Consortium *et al.*, 2018; Hastings *et al.*, 2016; Fabregat *et al.*, 2016) and ontologies (Consortium, 2018; Köhler *et al.*, 2019; Jupp *et al.*, 2016) are being actively studied using deep learning models as encyclopedic information about bio-entities contained in the resources could be useful biases to improve their vector representations (Alshahrani *et al.*, 2017; Alshahrani and Hoehndorf, 2018; Smaili *et al.*, 2018a,b).

Structured axioms and unstructured texts are two important sources of the information in the KBs and ontologies. The structured axioms formally describe properties of bio-entities and relations between them using the Semantic Web language OWL description logic (DL) (Grau *et al.*, 2008) which supports DL style reasoning and advanced querying; the unstructured texts describe bio-entities using the natural language which enables human experts to understand the precise meaning of bio-entities (Hoehndorf *et al.*, 2015b). Figure 1 represents some of the axioms and texts extracted from Gene Ontology (GO) (Consortium, 2018) and Uniprot (Consortium *et al.*, 2018) as a graph in which the circles represent GO classes, the diamond represents a protein and the labeled arrows represent types of the axioms that hold between these classes; `GO:0016774` and `GO:0008776` are the GO classes' identifiers while `B5Y7W0` is the protein's identifier. Take two axioms in Figure 1 as examples:

`GO:0008776` `is_a` `GO:0016774`

`B5Y7W0` `classified_with` `GO:0008776`

The first axiom states that the class 'acetate kinase activity' (`GO:0008776`) is a subclass of 'phosphotransferase activity'

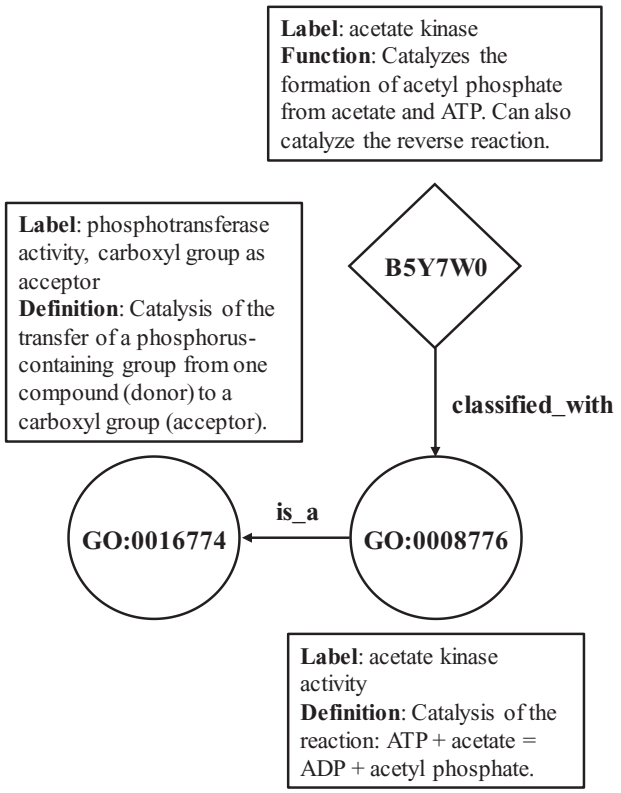

Fig. 1. An example of the structured axioms and unstructured texts from GO and Uniprot. The figure shows ontology classes as circles, a protein as a diamond, labels and definitions in boxes and types of axioms as edges

(`GO:0016774`) while the second axiom states that the protein 'acetate kinase' (`B5Y7W0`) is annotated by `GO:0008776`, which means that the protein has a molecular function of 'acetate kinase activity' (`GO:0008776`). The unstructured texts, as shown in the boxes of Figure 1, provide definitions for ontology classes or descriptions for bio-entities; the unstructured texts also contain other information including labels or synonyms associated with ontology classes and bio-entities. The structured axioms and unstructured texts contain rich and complementary information about bio-entities and thereby express biological reality accurately and comprehensively. Hence, considering both of them could lead to better representation learning (RL) of bio-entities and accordingly benefit various downstream task, for example, biological knowledge discovery.

There are two existing strategies for using the structured semantic resources to learn vector representations of bio-entities. The first strategy is applied by Alshahrani *et al.* (2017) and Alshahrani and Hoehndorf (2018) to construct knowledge graphs (KGs) using the structured axioms and apply a graph embedding technique to learn vector representations of bio-entities. The second strategy is applied by Smaili *et al.* (2018a) and Smaili *et al.* (2018b) to treat the axioms and the unstructured texts as a textual corpora; a word2vec model (Mikolov *et al.*, 2013a,b) is then used to learn vector representations of bio-entities.

We argue that the representations learned by these two strategies encode the knowledge from the KBs and ontologies only to a limited degree. The first strategy neglects to use the unstructured texts. However, the texts contain a great deal of information about bio-entities and ontology classes. Moreover, as ontology-based annotations are not available for all bio-entities, the unstructured texts are supplemental and essential resources for RL model to learn. As for the second strategy, although it makes use of both the structured axioms and unstructured texts, it does not fuse the information in a reasonable manner. The sentences in the corpora mix the Semantic Web language and natural language. For example, as shown in Figure 1, the definition of the ontology class whose id is `GO:0016774` is expressed as the sentence '<http://purl.obolibrary.org/obo/GO_0016774> <http://purl.obolibrary.o rg/obo/IAO_0000115> Catalysis of the transfer of a phosphorus-containing group from one compound (donor) to a carboxyl group (acceptor).', where '<http://purl.obolibrary.org/obo/GO_0016774>' is the Internationalized Resource Identifiers (IRI) of the ontology class, and '<http://purl.obolibrary.org/obo/IAO_0000115>' is the IRI of the annotation property `definition`. Since the two languages have different syntax rules regarding the structure and meanings of tokens, for example, the semantics of operators in OWL are different from words in the natural language, the representations learned from word co-occurrence might not be able to capture the underlying knowledge within the axioms and texts. It remains a big challenge for RL of biological entities how to seamlessly fuse the heterogeneous information in the unstructured texts and structured axioms. What's more, for an entity without any ontology-based annotation, the above-mentioned methods could only learn a representation by random guessing. Therefore, they could not be applied to the zero-shot circumstance, that is, predicting bio-relations for which at least one participating entity has no ontology-based annotation. How to deal with this challenging circumstance is another significant obstacle for widely applying these embedding methods.

To overcome the challenges mentioned above, we propose a novel RL method named **E**nhanced **R**epresentation with **B**iological **K**nowledge (ERBK). Instead of regarding the texts and axioms as a whole corpus, ERBK encodes the texts and axioms separately. Specifically, given an entity, the axioms encoding the relationships between the entity and the other entities or ontology classes are converted into triples and encoded using a knowledge graph embedding algorithm which is TransH (Wang *et al.*, 2014), while its textual definition is encoded using deep convolutional neural networks (CNN). To fuse the features learned from the texts and axioms, a training objective is used to maximize the likelihood of predicting the relationships and definitions simultaneously. In order to evaluate ERBK, we apply it on Gene Ontology (GO) and GOA (Gene Ontology Annotations) to generate vector representations of proteins, and on PhenomeNET (Hoehndorf *et al.*, 2011; Rodríguez-García *et al.*, 2017) ontology and its annotations to generate vector representations of genes and diseases. We then evaluate the representations on two bio-relation prediction tasks, which are PPI prediction and gene–disease association prediction. The results show that our method consistently outperforms existing methods on the two tasks. What's more, we evaluate our model under the zero-shot circumstance. The results reveal that, even in the absence of the structured axioms, the representations learned using only the unstructured texts still maintain promising performance on the two tasks. Further analyses illustrate that ERBK has the potential to discover novel bio-relations such as gene–disease associations. By taking advantage of the unstructured and structured information, more underlying knowledge is captured by the representations learned using ERBK. We believe the representations and the method have certain generality and could be applied to other types of entities and support several downstream applications.

2 Related works

In recent years, several approaches have been proposed to learn vector representations of biological entities. Some methods attempt to learn vector representations based on features derived from high-throughput techniques, such as sequential information of DNA or proteins (Chen *et al.*, 2019; You *et al.*, 2018b; Kulmanov *et al.*, 2018) or structural information of molecules (De Cao and Kipf, 2018; You *et al.*, 2018a; Jin *et al.*, 2018), others by incorporating prior knowledge from ontologies and biological KBs (Alshahrani *et al.*, 2017; Alshahrani and Hoehndorf, 2018; Smaili *et al.*, 2018a,b), which type of methods will be mainly discussed in this section.

Alshahrani *et al.* (2017) and Alshahrani and Hoehndorf (2018) construct KGs using ontologies and ontology-based annotations and apply graph embedding methods to generate vector representations of ontology classes and the annotated entities. Specifically, Alshahrani *et al.* (2017) firstly build a KG with ontology classes and

the annotated entities as nodes and types of axioms as edges. Then it uses DeepWalk (Perozzi *et al.*, 2014) to generate a textual corpus consisting of a set of edge-labeled random walks, that is, sequences composed of names of the nodes and edges. Then, a skip-gram model is used to generate vector representations of the nodes and edges. The method used by Alshahrani and Hoehndorf (2018) is similar, but the generated sequences include only the node names. The representations learned by Alshahrani *et al.* (2017) could be applied to predict functions of biological entities or drug target relations, while the representations learned by Alshahrani and Hoehndorf (2018) could be used to predict gene–disease associations based on phenotypic similarity.

In contrast to the above-mentioned methods, Smaili *et al.* (2018a) and Smaili *et al.* (2018b) regard an axiom expressed in OWL as a sentence, thus the axioms could be used to construct a textual corpus. Within each sentence, Smaili *et al.* (2018a) use IRI to denote ontology classes and OWL properties. Smaili *et al.* (2018b), the successor of Smaili *et al.* (2018a), further incorporates ontology meta-data into the corpus. A skip-gram model and pre-trained word vectors on biomedical literature are used to learn vector representations of bio-entities. The representations learned by these two methods could be used to predict PPIs and gene–disease associations.

Our work is inspired by DKRL model proposed by Xie *et al.* (2016). DKRL is a RL method for knowledge graphs taking advantage of unstructured descriptions of entities. DKRL applies TransE (Bordes *et al.*, 2013) to encode fact triples and two encoders to encode the descriptions which are continuous bag-of-words encoder (CBOW) and CNN. Considering that TransE does not do well in dealing with one-to-many/many-to-one/many-to-many relations which usually occur in the structured axioms, we replace TransE with TransH. Compare with the above-mentioned methods, our method further leverages the unstructured texts which are widely available and contain detailed information associated with bio-entities, and integrates the unstructured texts with the structured axioms in a reasonable manner making the representations encode more knowledge. What's more, our method is able to learn a representation for an entity with no ontology-based annotation in which circumstance the mentioned methods fall short.

3 Materials and methods

We propose ERBK in this paper to enhance vector representations of bio-entities by fusing the information of the structured axioms and the unstructured texts provided by biological KBs and ontologies. The representations could be used for automatically discovering bio-relations, such as PPIs and gene–disease associations. We firstly construct fact triples based on the structured axioms. Then, TransH is applied to encode the fact triples while CNN is applied to encode the unstructured texts. Finally, vector representations of bio-entities are generated by fusing semantic features learned from the triples and texts.

3.1 Construction of fact triples

The structured axioms represented in OWL may be complex and not easily be converted into fact triples (Rodríguez-García and Hoehndorf, 2018; Hoehndorf *et al.*, 2015b), for example, an axiom involving an object property which is associated with a complex class description instead of a single class. We construct fact triples by selecting the structured axioms which could be easily mapped to fact triples and contain crucial information of bio-entities. We build two sets of fact triples: one is built on GO and GOA of proteins in order to obtain vector representations of proteins, while the other is built on PhenomeNET ontology and its annotations of genes and diseases to obtain their representations.

We downloaded GO (http://www.geneontology.org/ontology/) in OWL format which was released on June 10, 2019. We downloaded GOA (http://www.ebi.ac.uk/GOA) of human proteins which was released on June 2, 2019 and GOA of yeast proteins released on July 1, 2019. We removed all the annotations with evidence code *IEA* as well as *ND*. In total, we obtain 194 569 GO to human protein annotations, 32 009 GO to yeast protein annotations. The number of unique GO classes associated with human and yeast proteins is 44 990, the number of unique human proteins is 15 718 and the number of unique yeast proteins is 3856. For a fact triple constructed from a GO axiom, a GO class is regarded as *head entity* or *tail entity*, while *relation* is constructed by the object and annotation properties of GO including *ends_during*, *regulates*, *has_part*, *negatively_regulates*, *positively_regulates*, *starts_during*, *part_of*, *occurs_in*, *happens_during* and *subClassOf*. For a fact triple constructed from a GOA axiom, an annotated protein is regarded as *head entity* while the GO class is *tail entity*. We define four types of *relation* which are *has_function*, *is_involved_in*, *is_located_within* and *is_annotated_with*; *is_annotated_with* links proteins with GO classes without any type information, while the other three types of *relation* connect proteins with GO classes delineating the particular molecular functions they have, the biological processes they are involved in and the certain cellular components they are located within, which allow the representations to encode in-depth knowledge regarding biological properties of proteins.

We downloaded PhenomeNET ontology in OWL format from the AberOWL repository (http://aber-owl.net) (Hoehndorf *et al.*, 2015a) on December 1, 2019. The mouse phenotype to mouse gene annotations were downloaded from Mouse Genome Informatics database (MGI) (http://www.informatics.jax.org/) (Smith and Eppig, 2015) on December 7, 2019. The human phenotype to human gene annotations and the human phenotype to human disease annotations were downloaded from Human Phenotype Ontology database (HPO) (https://hpo.jax.org/app/download/annotation) (Robinson *et al.*, 2008) on December 7, 2019. In total, we obtain 170 048 unique mouse phenotype to mouse gene annotations, 53 188 unique human phenotype to human gene annotations, and 44 557 unique human phenotype to human disease annotations. The ontology and the annotations contain 224 626 PhenomeNET ontology classes, 3131 unique diseases, 8534 unique mouse genes and 1379 unique human genes. Similarly, for a fact triple constructed from a PhenomeNET axiom, we regard PhenomeNET ontology as *head entity* or *tail entity*, while *relation* is constructed by the object and annotation properties of PhenomeNET including *equivalentClass*, *has_quality*, *has_modifier* and *subClassOf*; we further define *has_entity* as another type of *relation*. *equivalentClass* links two PhenomeNET ontologies which are semantically similar. As a PhenomeNET ontology refers to a phenotype, *has_entity* links the PhenomeNET ontology to an ontology referring to the affected entity of that phenotype; *has_quality* links the PhenomeNET ontology to an ontology referring to the specific quality of that entity being affected; *has_modifier* links the PhenomeNET ontology to an ontology referring to a modifier that specifies how the quality is affected. These links are based on EQ definitions (Mungall *et al.*, 2010) of PhenomeNET ontologies. To make it more clear, we take the PhenomeNET ontology 'increased plasma cell number' as an example; *has_entity* will link it to 'plasma cell', *has_quality* will link it to 'increased amount' while *has_modifier* will link it to 'abnormal'. For a fact triple constructed from a PhenomeNET annotation axiom, an annotated gene or disease is regarded as *head entity* while the PhenomeNET ontology is *tail entity*, and *relation* is defined as *has_phenotype*. These fact triples link genes or diseases not only to related phenotypes but also to related molecular and anatomical information as well as the mapping phenotypes of other species, which enable the representations to encode comprehensive relationships between genes and diseases in terms of phenotype.

3.2 Representation learning

As introduced in Section 3.1, the fact triples we construct have many more unique entities than unique relations in which case one-to-many/many-to-one/many-to-many relations are very common. Therefore, we apply TransH rather than TransE since TransH is more capable of dealing with these kinds of relations while maintaining computational efficiency. Inspired by DKRL, our model firstly learns vector representations of the structured axioms and unstructured texts independently, and then fuse the representations through optimizing an objective function.

We use $(h, r, t) \in F$ to denote a fact triple; $h, t \in E$ denote *head entity* and *tail entity* while $r \in R$ denotes *relation*; F, E, R denote the set of the fact triples, entities and relations respectively. $\mathbf{h_g}$ and $\mathbf{t_g}$ denote the representation of *head entity* and *tail entity* learned from a fact triple which take value in $\mathbb{R}^k$. Given a fact triple, TransH (Wang *et al.*, 2014) firstly projects $\mathbf{h_g}$ and $\mathbf{t_g}$ to a hyperplane $\mathbf{w}_r$ and the projections are denoted by $\mathbf{h}_\perp$ and $\mathbf{t}_\perp$ with restricting $||\mathbf{w}_r||_2 = 1$, and then minimizes a score function to learn $\mathbf{h_g}$ and $\mathbf{t_g}$ as follows:

$$\mathbf{h}_\perp = \mathbf{h_g} - \mathbf{w}_r^\top \mathbf{h_g} \mathbf{w}_r, \quad \mathbf{t}_\perp = \mathbf{t_g} - \mathbf{w}_r^\top \mathbf{t_g} \mathbf{w}_r \tag{1}$$

$$f_r(\mathbf{h_g}, \mathbf{t_g}) = ||\mathbf{h}_\perp + \mathbf{d}_r - \mathbf{t}_\perp|| \tag{2}$$

$\mathbf{d}_r$ denotes a relation-specific translation vector which is in the hyperplane. The score function indicates that the projection of *tail entity* $\mathbf{t}_\perp$ should be the nearest neighbor of $\mathbf{h}_\perp + \mathbf{d}_r$.

As for the unstructured texts, we use descriptions of biological entities provided by biological KBs and definitions of ontology classes provided by ontologies. Specifically, the protein descriptions are obtained from Uniprot; the gene descriptions are obtained from Alliance (The Alliance of Genome Resources Consortium, 2020); the disease definitions are provided by Human Disease Ontology (Schriml *et al.*, 2019); the definitions of the ontology classes are provided by GO or PhenomeNET. Based on DKRL Xie *et al.* (2016), CNN is applied to learn semantic information in the texts. The CNN consists of two layers. The input of the CNN is a word embedding sequence $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, where $\mathbf{x}$ represents a word embedding and $\mathbf{N}$ is the length. The output of the second convolution layer is calculated as follows:

$$\mathbf{x}_i^{(1)} = \mathbf{x}_{i:i+k-1} = [\mathbf{x}_i^T, \mathbf{x}_{i+1}^T, \ldots, \mathbf{x}_{i+k-1}^T]^T \tag{3}$$

$$\mathbf{z}_i^{(1)} = \sigma(\mathbf{W}^{(1)} \mathbf{x}_i^{(1)} + \mathbf{b}_i^{(1)}) \tag{4}$$

$$\mathbf{x}_i^{(2)} = \max(\mathbf{z}_{n \cdot i}^{(1)}, \ldots, \mathbf{z}_{n \cdot (i+1)-1}^{(1)}) \tag{5}$$

$$\mathbf{z}^{(2)} = \sum_{i=1,\ldots,m} \frac{\sigma\left(\mathbf{W}^{(2)} \mathbf{x}_i^{(2)} + \mathbf{b}_i^{(2)}\right)}{m} \tag{6}$$

$\mathbf{x}_i^{(1)}$ and $\mathbf{x}_i^{(2)}$ denote the ith inputs of the first and second convolutional layers. $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are the convolution kernels, $\mathbf{b}_i^{(1)}$ and $\mathbf{b}_i^{(2)}$ are the biases, and σ is the activation function. $\mathbf{z}_i^{(1)}$ denotes the ith output of the first convolution layer. m is the number of the input vectors of the second convolution layer; $\mathbf{z}^{(2)}$ denotes its output which is the representation learned from the unstructured texts (referred to as $\mathbf{h_e}$ and $\mathbf{t_e}$ in this paper) taking value in $\mathbb{R}^k$. Figure 2 shows an overview of ERBK model taking the protein and GO classes in Figure 1 as an example. $\mathbf{h_g}$, $\mathbf{t_g}$, $\mathbf{h_e}$ and $\mathbf{t_e}$ are trained by minimizing the same margin-based objective as TransH (Wang *et al.*, 2014):

$$\mathcal{L}(\mathbf{h}, \mathbf{d}_r, \mathbf{t}) = \sum_{(h,r,t) \in F} \sum_{(h',r',t') \in F'} \max(\gamma + f_r(\mathbf{h}, \mathbf{t}) - f_{r'}(\mathbf{h}', \mathbf{t}'), 0) \tag{7}$$

Critically, $\mathcal{L}(\mathbf{h}, \mathbf{d}_r, \mathbf{t})$ represents the sum of $\mathcal{L}(\mathbf{h_g}, \mathbf{d}_r, \mathbf{t_g})$, $\mathcal{L}(\mathbf{h_e}, \mathbf{d}_r, \mathbf{t_e})$, $\mathcal{L}(\mathbf{h_g}, \mathbf{d}_r, \mathbf{t_e})$ and $\mathcal{L}(\mathbf{h_e}, \mathbf{d}_r, \mathbf{t_g})$. By sharing $\mathbf{d}_r$ and $\mathbf{w}_r$, the two types of representations are mapped into a unified vector space in order to fuse the semantic information of the structured axioms and unstructured texts. $\gamma > 0$ is a margin hyper-parameter. F' denotes the negative set of the fact triples which is constructed the same way as TransH Wang *et al.* (2014).

3.3 Implementation details

ERBK takes plain texts and fact triples as input and outputs two types of entity representations. The vectors $\mathbf{d}_r$, $\mathbf{w}_r$, $\mathbf{h_g}$ and $\mathbf{t_g}$ are initialized randomly; while $\mathbf{h_e}$ and $\mathbf{t_e}$ are initialized by using the pre-trained word embeddings on PubMed Central articles provided by Smaili *et al.* (2018b). The dimension of the word embeddings and the two kinds of representations is 200. The Adam method (Kingma and Ba, 2014) is used for optimization.

4 Results

We conduct experiments on two entity-related bio-relation prediction tasks which are PPI prediction and gene–diseases association prediction. We further evaluate our method under the zero-shot circumstance where the bio-relations are predicted for which at least one participating entity has no ontology-based annotation.

4.1 Dataset

Different from the ontologies and annotations we downloaded to learn vector representations of bio-entities (see Section 3.1), the dataset introduced in this section is used to evaluate the representations for predicting PPI and gene–disease association.

The PPI dataset for human (*Homo sapiens*) and yeast are obtained from the STRING database (Szklarczyk *et al.*, 2019). Following the same settings of the experiments in (Smaili *et al.*, 2018b), we consider the interactions from the STRING positive. However, we generate the negative interaction set in a different way. Inspired by (Guo *et al.*, 2008), we construct a negative interaction set by pairing proteins in different cellular compartments. Specifically, we firstly obtain the sub-cellular localization information of proteins from Uniprot. We then categorize proteins into eight groups based on the eight types of localization: cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, golgi apparatus, peroxisome, vacuole and cytoplasm & nucleus. The negative cases are constructed by pairing proteins from one group with proteins from the

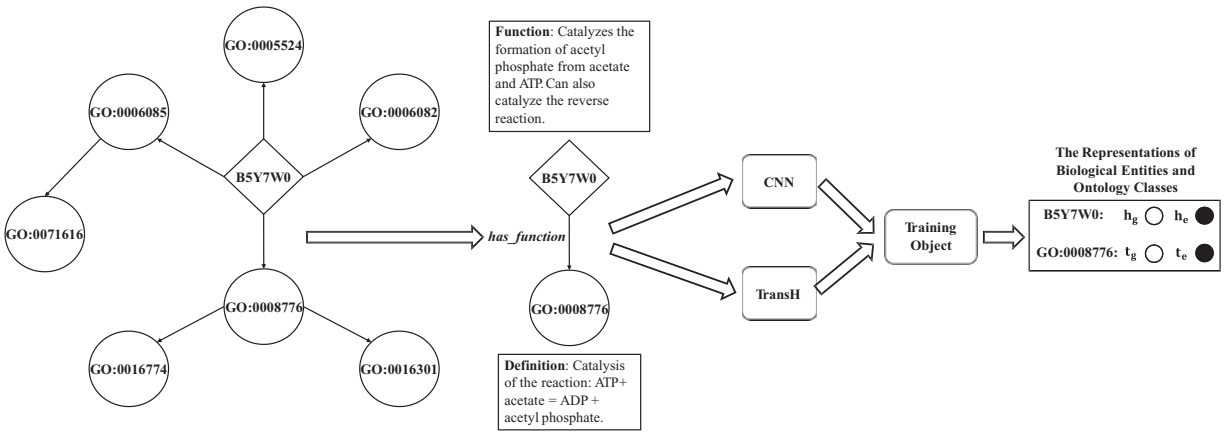

Fig. 2. Overview of ERBK model

Table 1. Statistics of the PPI dataset

Species	Dataset	Interactions	Proteins
Human	Train	200 000	8000
	(*e-e*)	2000	3065
	(*e-d*) and (*d-e*)	2000	399
	(*d-d*)	2000	396
Yeast	Train	100 000	2000
	(*e-e*)	1000	1179
	(*e-d*) and (*d-e*)	1000	199
	(*d-d*)	1000	200

Note: The third column indicates the number of PPI and the fourth column indicates the number of unique proteins contained in the corresponding dataset.

Table 2. Statistics of the gene–disease association dataset

Species	Dataset	Associations	Disease	Gene
Mouse	Train	4069	1299	1136
	(*e-e*)	540	390	382
	(*e-d*) and (*d-e*)	400	339	319
	(*d-d*)	494	276	278
Human	Train	4573	2092	945
	(*e-e*)	360	334	277
	(*e-d*) and (*d-e*)	400	363	301
	(*d-d*)	465	339	238

Note: The third column indicates the number of gene–disease associations, the fourth and fifth column indicates the number of unique diseases and gens contained in the corresponding dataset.

other while excluding the positive cases. Compared with randomly sub-sampling cases among all the pairs not occurring in the STRING as negative as (Smaili *et al.*, 2018b) does, the negative cases constructed in our way are more likely to be true. The number of the positive cases and the number of the negative cases are equal. We split the data into a training and testing set and the statistics are listed in Table 1. We construct three datasets for testing denoted as (*e-e*), (*e-d*) and (*d-d*) respectively, and the (*e-d*) and (*d-d*) datasets are used in the zero-shot circumstance. *e* denotes a protein with at least one GOA which in turn has a pre-trained representation, while *d* denotes a protein which does not have any ontology-based annotation and therefore has no pre-trained representation. For a protein with no pre-trained representation, ERBK takes its description obtained from Uniprot as input and the output of the CNN is used as the protein's representation. We ensure that all of the proteins in the training and testing data have descriptions provided by Uniprot.

The gene–disease association dataset are provided by MGI including human gene-human disease associations and mouse gene-human disease associations. We consider all of the associations not occurring in the data as negative. We also split the data into a training and testing set and the statistics are listed in Table 2. Similarly, we construct the (*e-e*), (*e-d*) and (*d-d*) datasets and the gene descriptions and disease definitions are used to generate representations for the genes or diseases without any pre-trained representation. We ensure that all of the genes or diseases in the dataset have the related descriptions or definitions.

We perform two-class classification and use F-measure and AUC value under ROC curve as our evaluation measures.

4.2 Methods for comparison

We mainly compare our method with Alshahrani *et al.* (2017); Alshahrani and Hoehndorf (2018); Smaili *et al.* (2018b). The method proposed by Alshahrani *et al.* (2017) is named 'Neuro-Symbolic method', the method proposed by Alshahrani and Hoehndorf (2018) is named 'SmuDGE' and the method proposed by Smaili *et al.* (2018b) is named 'OPA2Vec'. We do not compare our method with Smaili *et al.* (2018a) since Smaili *et al.* (2018b) achieves better performance than it. We also implement TransH (Wang *et al.*, 2014) and DKRL (Xie *et al.*, 2016) (using CNN as the sentence encoder) as the baselines.

Our model learns two vector representations for each entity and the experimental results using $\mathbf{h}_e$, $\mathbf{t}_e$ are numerically close to the results of $\mathbf{h}_g$ and $\mathbf{t}_g$. Therefore, we report only the experimental results using $\mathbf{h}_e$ and $\mathbf{t}_e$.

A neural network model is used to predict PPI and gene–disease association using the representations learned by ERBK and the baselines. The network model has three layers including an input layer, a hidden layer and a softmax layer. The network takes concatenation of two entities' representations as input; the dimension of the hidden layer is 256; the Adam method is used for optimization.

Table 3. Evaluation results of PPI on the non-splitted dataset

Species	Model	Accuracy	Recall	Precision	F1	AUC
Human	OPA2Vec	66.2	49.9	69.4	58.1	70.8
	SmuDGE	64.7	73.4	66.5	69.8	67.8
	Neuro-Symbolic method	65.3	77.0	65.9	71.0	69.2
	TransH	63.2	40.2	66.6	50.1	66.5
	DKRL	82.3	73.6	89.3	80.7	89.7
	ERBK	**82.8**	**77.7**	87.4	**82.3**	**90.3**
Yeast	OPA2Vec	64.3	75.7	65.7	70.3	69.3
	SmuDGE	63.8	75.2	63.3	68.7	68.9
	Neuro-Symbolic method	66.9	77.0	67.7	72.0	71.5
	TransH	62.8	40.8	64.5	50.0	68.3
	DKRL	82.3	75.0	87.8	80.9	88.7
	ERBK	**83.0**	75.3	**89.1**	**81.6**	**88.9**

Note: The results are given in percentage. The best results are bold.

4.3 PPI prediction

This task predicts if two proteins interact. Table 3 shows the evaluation results of PPI on the non-splitted dataset, that is, the dataset combining the (*e-e*), (*e-d*) and (*d-d*) datasets of PPI. The results reveal that ERBK outperforms the baseline methods. To further understand the characteristics of ERBK, we evaluate it by using the splitted datsets.

Table 4 shows the results of PPI on the (*e-e*) dataset. We have the following observations: (i) ERBK outperforms the baseline methods in F1 and AUC on both the human and yeast data. The results indicate that vector representations of proteins are well enhanced through our method. (ii) ERBK outperforms OPA2Vec while both methods use the structured axioms and unstructured texts; as for the other three baseline methods using only the structured axioms, SmuDGE and Neuro-Symbolic method outperform TransH. On the one hand, the results indicate that the ability of ERBK to encode the structured axioms might be limited by TransH. On the other hand, since ERBK still outperforms other methods, it demonstrates the effectiveness of our information fusing strategy. (iii) Compared with TransH, ERBK achieves better performance which shows that incorporating the unstructured texts could improve the representations. (iv) ERBK outperforms DKRL which indicates that TransH has more advantage over TransE. Although the improvement of ERBK may not be evident in the (*e-e*) dataset, the advancement of our method is substantial under the zero-shot circumstance.

Under the zero-shot circumstance, Neuro-Symbolic method (Alshahrani *et al.*, 2017), OPA2Vec (Smaili *et al.*, 2018b) and SmuDGE (Alshahrani and Hoehndorf, 2018) have to predict PPI randomly since there are no pre-trained vector representations for the proteins. However, ERBK could make predictions based on their

Table 4. Evaluation results of PPI on the (*e-e*) dataset

Species	Model	Accuracy	Recall	Precision	F1	AUC
Human	OPA2Vec	83.6	82.2	84.7	83.4	91.0
	SmuDGE	84.4	80.0	87.9	83.8	91.3
	Neuro-Symbolic method	83.8	78.3	88.1	82.9	90.6
	TransH	82.7	79.4	85.2	82.2	90.0
	DKRL	84.7	79.3	89.2	83.9	91.2
	ERBK	**85.3**	**83.4**	86.9	**85.1**	**91.5**
Yeast	OPA2Vec	83.6	80.4	86.2	83.2	91.1
	SmuDGE	84.3	84.1	84.8	84.4	91.4
	Neuro-Symbolic method	84.3	79.4	88.3	83.7	91.1
	TransH	82.8	82.7	83.2	82.9	90.8
	DKRL	84.1	79.4	87.9	83.5	91.2
	ERBK	**84.6**	81.1	**89.1**	**84.9**	**91.6**

Note: The results are given in percentage. The best results are bold.

Table 5. Evaluation results of PPI on the (*e-d*), (*d-e*) and (*d-d*) datasets

Test set	Species	Model	Accuracy	Recall	Precision	F1	AUC
e-d or *d-e*	Human	DKRL	82.7	74.8	88.7	81.2	90.7
		ERBK	**82.9**	**78.0**	86.5	**82.0**	**91.0**
	Yeast	DKRL	82.2	75.8	87.2	81.1	89.1
		ERBK	**83.0**	75.8	**88.9**	**81.8**	89.1
d-d	Human	DKRL	79.5	66.7	89.7	76.5	87.2
		ERBK	**80.3**	**69.8**	88.5	**78.0**	**87.9**
	Yeast	DKRL	80.7	71.6	87.9	78.9	85.0
		ERBK	**81.3**	**71.8**	**89.0**	**79.5**	**85.2**

Note: The results are given in percentage. The best results are bold.

descriptions obtained from Uniprot. For the sake of brevity, Table 5 compares ERBK with only DKRL; the results of other models are shown in Supplementary Tables S1 and S2.

Compared with the results on the (*e-e*) dataset, we have the following observations: (i) most of the measures in the (*e-d*) and (*d-d*) datasets experience a decrease which might be due to the lack of prior knowledge encoded in the structured axioms. (ii) Our method maintains stable performance on the (*e-d*) dataset and still achieves relatively good results on the (*d-d*) dataset. These results show that, even in the absence of the structured axioms, the representations learned by ERBK using only the unstructured texts could still be reliable for PPI prediction.

4.4 Gene–disease association prediction

This task predicts gene–disease association. Table 6 shows the evaluation results of gene–disease association on the non-splitted dataset. The results reveal that ERBK outperforms the baseline methods which further demonstrate that the representations of genes and diseases are well enhanced through our method. Similarly, we evaluate it by using the different splitted datasets.

Table 7 shows the results on the (*e-e*) dataset. We have the following observations: (i) ERBK outperforms the baselines on the human data; ERBK achieves the best F1 and a comparable AUC on the mouse data. To further investigate the potential of our method on the mouse data, we removed the fact triples involving *equivalentClass* and *has_modifier* and re-trained vector representations of the genes and the diseases using our method and TransH. The results on the (*e-e*) dataset are denoted as 'TransH-pruned' and 'ERBK-pruned'. As shown in Table 7, after simplifying the relationships of the fact triples, the performance of TransH and our method improve. The results indicate that TransH is not able to take full

Table 6. Evaluation results of gene–disease association on the non-splitted dataset

Species	Model	Accuracy	Recall	Precision	F1	AUC
Human	OPA2Vec	66.2	39.4	60.7	47.8	67.3
	SmuDGE	65.8	40.1	69.7	50.9	68.4
	Neuro-Symbolic method	65.6	54.4	66.5	59.8	67.2
	TransH	61.7	33.3	57.6	42.2	63.5
	DKRL	81.6	70.4	89.4	78.8	91.7
	ERBK	**85.3**	**76.5**	**91.9**	**83.5**	**93.0**
Mouse	OPA2Vec	61.9	38.4	58.4	46.3	63.7
	SmuDGE	62.1	41.8	61.7	49.7	64.0
	Neuro-Symbolic method	57.7	34.8	51.8	41.6	57.9
	TransH	57.3	38.4	53.7	44.8	56.1
	DKRL	72.6	67.8	71.3	69.5	79.9
	ERBK	**75.1**	**74.2**	**73.3**	**73.7**	78.2

Note: The results are given in percentage. The best results are bold.

Table 7. Evaluation results of gene–disease association on the (*e-e*) dataset

Species	Model	Accuracy	Recall	Precision	F1	AUC
Human	OPA2Vec	90.7	95.2	84.3	89.4	96.4
	SmuDGE	92.1	87.1	93.1	90.0	97.1
	Neuro-Symbolic method	86.1	82.3	83.6	82.9	93.4
	TransH	82.8	77.4	80.0	78.7	90.1
	TransH-pruned	84.8	88.6	81.1	84.7	92.5
	DKRL	89.1	86.2	90.3	88.2	95.9
	ERBK	93.2	93.4	92.4	92.9	97.9
	ERBK-pruned	**94.0**	93.9	**93.4**	**93.7**	**98.2**
Mouse	OPA2Vec	82.4	78.6	77.6	78.1	87.4
	SmuDGE	79.0	70.2	75.6	72.8	84.8
	Neuro-Symbolic method	72.4	60.7	67.1	63.8	75.2
	TransH	71.0	57.1	65.8	61.1	72.3
	TransH-pruned	73.0	65.7	68.9	67.3	77.9
	DKRL	76.9	70.9	71.8	71.3	82.9
	ERBK	81.1	81.7	76.8	79.2	84.5
	ERBK-pruned	**82.8**	**84.7**	75.3	**79.7**	86.8

Note: The results are given in percentage. The best results are bold.

advantage of the structured axioms and might therefore hamper the performance of our method in predicting associations between mouse genes and human diseases. Therefore, there is a great potential of the strategy fusing the structured axioms and unstructured texts as much value of the structured axioms remains to be exploited. (ii) ERBK achieves better performance than TransH and DKRL and ERBK-pruned outperforms TransH-pruned. The results demonstrate the advantage of incorporating the unstructured texts and the effectiveness of our information fusing strategy. The results of TransH-pruned and ERBK-pruned on the (*e-d*), (*d-d*) and non-splitted datasets are shown in Supplementary Tables S3, S4 and S5.

As for the zero-shot circumstance, for the sake of brevity, Table 8 compares the F-measure and AUC values of only DKRL and ERBK; the results of other models are shown in Supplementary Tables S3 and S4. Consistent with the results of PPI, our method achieves better results compared to other baselines, which further proves the applicability of our method under the zero-shot circumstance. What's more, since the unstructured texts play a main role in the performance of our method under the zero-shot circumstance, we investigate their effect by using different human gene descriptions. Specifically, we replaced the human gene descriptions

Table 8. Evaluation results of gene–disease association on the (*e-d*), (*d-e*) and (*d-d*) datasets

Test set	Species	Model	Accuracy	Recall	Precision	F1	AUC
e-d or *d-e*	Human	DKRL	81.9	68.4	91.5	78.3	92.5
		DKRL-RGD	82.8	76.5	86.1	81.0	90.0
		ERBK	84.8	75.2	91.5	82.5	93.5
		ERBK-RGD	**86.6**	**81.7**	89.4	**85.4**	**93.7**
	Mouse	DKRL	72.6	72.4	71.1	71.7	78.6
		ERBK	**74.3**	68.8	**75.5**	**72.0**	77.7
d-d	Human	DKRL	75.6	59.7	87.9	71.1	88.5
		DKRL-RGD	76.2	64.9	84.7	73.5	84.0
		ERBK	78.3	64.1	89.7	74.8	88.6
		ERBK-RGD	**81.2**	**73.4**	87.2	**79.7**	89.8
	Mouse	DKRL	68.2	60.3	71.6	65.5	75.8
		ERBK	**70.0**	**71.7**	69.4	**70.5**	74.3

Note: The results are given in percentage. The best results are bold.

downloaded from the Alliance database by the descriptions downloaded from RefSeq database (O'Leary *et al.*, 2016), which are human-curated and contain more comprehensive information regarding human gene functions. We then re-trained vector representations of human genes using DKRL and ERBK. The results are shown in Table 8 denoted as 'DKRL-RGD' and 'ERBK-RGD'. The results using the RefSeq descriptions outperform the results using the Alliance descriptions. This finding reveals that the better unstructured texts lead to more reliable predictions of our model which not only demonstrates the advantage of our model but also the potential of the strategy fusing the unstructured texts and the structured axioms.

5 Discussion

5.1 Potential for discovering novel gene–disease associations

We analyze the potential of our model in predicting novel bio-relations. We apply our model to predict 19 novel human gene-human disease pairs proposed by (Wang *et al.*, 2019) [The pairs are extracted from Table 7 and Table 8 in Wang *et al.* (2019)]; all of the pairs are validated by publications and are not included in our gene–disease association dataset. The pairs contain 9 unique diseases and 14 unique genes; 4 diseases have the pre-trained representations while other diseases and genes do not have any pre-trained representation and thereby 11 pairs are *d-e* and 8 pairs are *d-d*. Among all the pairs, 9 *d-e* pairs and 5 *d-d* pairs are predicted by ERBK correctly; the results are shown in Supplementary Table S5. It reflects the potential of our representations to be used for automatically discovering bio-relations, such as gene–disease associations. We believe that by learning the representations for more bio-entities, ERBK could be adopted to perform early bio-relations discovery that is beyond the reach of current experimental approaches.

5.2 Potential applications of ERBK

As ERBK is applicable to multiple circumstances and shows robustness, it could be used to support several downstream applications. In addition to PPI and gene–disease association, ERBK could further be applied to other bio-relations prediction tasks such as genotype–phenotype relationship prediction, RNA–disease association prediction, wherever there are ontologies, annotations, ontology definitions and entity descriptions. What's more, the representations of ontologies, relations and entities learned by ERBK could further be exploited by other machine learning or deep learning methods such as graph neural networks (GNNs) regarding ontology and its annotations as graph-structured data to support more biological tasks. For example, in order to combine genotype–phenotype data of different species from multiple databases, it is required to construct links between phenotype ontologies. This problem could be addressed by GNN as a task of network embedding and matching which network consists of a phenotype ontology, its affected entity(s), quality and the modifier based on its EQ definition. The representations learned by ERBK could be used as the input features. Moreover, ERBK could also be applied to the field of biomedical text mining to support named entity recognition (Habibi *et al.*, 2017) or even event-based mining (Yu et al., 2018; Lou *et al.*, 2020).

5.3 Limitations and future work

Our work has several limitations and we intend to address them as our future work. As Tables 6, 7 and 8 show, compared with the results on the human data, the results of our method on the mouse data experience a decrease. An association between a gene and a disease is predicted based on phenotypic similarity between the disease and the gene. For the human data, both the human genes and the human diseases are annotated by HPO ontology; for the mouse data, the human diseases are annotated by HPO ontology while the mouse genes are annotated by MP ontology (Smith *et al.*, 2004). Compared with computing phenotypic similarity between two HPO ontologies, computing the similarity between a MP ontology and a HPO ontology relies on more complex axioms, such as the axioms involving *has_quality*, *has_modifier* or *equivalent_Class* provided by PhenomeNET. As analyzed in Section 4.4, TransH might not be able to fully exploit the complex axioms regarding mouse genes and human diseases. Therefore, in order to exploit the potential of the strategy fusing the unstructured texts and the structured axioms, it will be necessary to further investigate how to encode more knowledge within the structured axioms and fuse it with the unstructured texts in a better way.

In contrast to the structured axioms and unstructured text, a larger scale of knowledge information is contained in biomedical literature. Learning representations by further taking advantage of these data holds the promise of modeling more entity-centric knowledge. It has been validated that large-scale pre-trained language models such as BERT (Devlin *et al.*, 2018) implicitly capture real-world knowledge from natural language texts (Petroni *et al.*, 2019; Logan *et al.*, 2019). Therefore, in the future, we expect to investigate how to make the pre-trained models to focus on capturing knowledge about bio-entities from biomedical literature.

6 Conclusion

The RL model, ERBK, proposed in this paper, improves the bio-entities' representations by fusing the heterogeneous information from the structured axioms and unstructured texts contained in biological ontologies and KBs. The representations could not only encode more biological knowledge so that support other computational methods achieve better performance on several tasks including bio-relation prediction, knowledge discovery etc., but also be further applied to the zero-shot circumstance where existing approaches fall short. The representations are evaluated on the task of PPI and gene–disease association prediction. The experimental results show that our method outperforms other baselines. Furthermore, our method maintains good performance under the zero-shot circumstance. We believe the representations and the method have certain generality and could be applied to other types of entities and support several downstream applications.

Funding

This work was supported by the National Key Research and Development Program of China [2018YFC0910404]; National Natural Science Foundation of China [61772409]; the consulting research project of the Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China); Project of China Knowledge Centre for Engineering Science and Technology; the Innovation Team from the Ministry of Education [IRT_17R86]; the Innovative Research Group of the National Natural Science Foundation of China [61721002]; and

Professor Chen Li' s Recruitment Program for Young Professionals of 'The Thousand Talents Plan'.

Conflict of Interest: none declared.

References

Alshahrani,M. and Hoehndorf,R. (2018) Semantic disease gene embeddings (SMUDGE): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics*, **34**, i901–i907.

Alshahrani,M. *et al.* (2017) Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics*, **33**, 2723–2730.

Bordes,A. *et al.* (2013) Translating embeddings for modeling multi-relational data. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, Volume 2, Curran Associates Inc, Red Hook, NY, USA. pp. 2787–2795.

Chen,M. *et al.* (2019) Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics*, **35**, i305–i314.

Consortium,G.O. (2018) The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.*, **47**, D330–D338.

Consortium,U. *et al.* (2018) Uniprot: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.

De Cao,N. and Kipf,T. (2018) Molgan: an implicit generative model for small molecular graphs. *arXiv preprint arXiv : 1805.11973*.

Devlin,J. *et al.* (2018) BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv : 1810.04805*.

Fabregat,A. *et al.* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.

Grau,B.C. *et al.* (2008) OWL 2: the next step for owl. *Web Semant. Sci. Serv. Agents World Wide Web*, **6**, 309–322.

Guo,Y. *et al.* (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res.*, **36**, 3025–3030.

Habibi,M. *et al.* (2017) Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, **33**, i37–i48.

Hastings,J. *et al.* (2016) Chebi in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.

Hoehndorf,R. *et al.* (2011) Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.*, **39**, e119–e119.

Hoehndorf,R. *et al.* (2015a) Aber-OWL: a framework for ontology-based data access in biology. *BMC Bioinformatics*, **16**, 26.

Hoehndorf,R. *et al.* (2015b) The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinf.*, **16**, 1069–1080.

Jin,W. *et al.* (2018) Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, 2323–2332.

Jupp,S. *et al.* (2016) The cellular microscopy phenotype ontology. *J. Biomed. Semant.*, **7**, 28.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations*, May 7-9, 2015, San Diego, CA, USA.

Köhler,S. *et al.* (2019) Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res.*, **47**, D1018–D1027.

Kulmanov,M. *et al.* (2018) DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, **34**, 660–668.

Logan,R. *et al.* (2019) Barack's wife Hillary: using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 5962–5971. doi: 10.18653/v1/P19-1598.

Lou,P. *et al.* (2020) BioNorm: deep learning-based event normalization for the curation of reaction databases. *Bioinformatics*, **36**, 611–620.

Mikolov,T. *et al.* (2013a) Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Mikolov,T. *et al.* (2013b) Efficient estimation of word representations in vector space. *arXiv preprint arXiv : 1301.3781*.

Mungall,C.J. *et al.* (2010) Integrating phenotype ontologies across multiple species. *Genome Biol.*, **11**, R2.

O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.

Perozzi,B. *et al.* (2014) DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 701–710.

Petroni,F. *et al.* (2019) Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2463—2473.

Robinson,P.N. *et al.* (2008) The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.

Rodríguez-García,M.Á. and Hoehndorf,R. (2018) Inferring ontology graph structures using owl reasoning. *BMC Bioinformatics*, **19**, 7.

Rodríguez-García,M.Á. *et al.* (2017) Integrating phenotype ontologies with phenomenet. *J. Biomed. Semant.*, **8**, 58.

Schriml,L.M. *et al.* (2019) Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.

Smaili,F.Z. *et al.* (2018a) Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, **34**, i52–i60.

Smaili,F.Z. *et al.* (2018b) Opa2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics*, **35**, 2133–2140.

Smith,C.L. and Eppig,J.T. (2015) Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. *J. Biomed. Semant.*, **6**, 11.

Smith,C.L. *et al.* (2004) The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.

Szklarczyk,D. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.

The Alliance of Genome Resources Consortium. (2020) Alliance of genome resources portal: unified model organism research platform. *Nucleic Acids Res.*, **48**, D650–D658.

Wang,X. *et al.* (2019) Predicting gene–disease associations from the heterogeneous network using graph embedding. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 504–511.

Wang,Z. *et al.* (2014) Knowledge graph embedding by translating on hyperplanes. In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*. pp. 1112—1119.

Xie,R. *et al.* (2016) Representation learning of knowledge graphs with entity descriptions. In: *Thirtieth AAAI Conference on Artificial Intelligence*.

You,J. *et al.* (2018a) Graph convolutional policy network for goal-directed molecular graph generation. In: *Advances in Neural Information Processing Systems*, pp. 6410–6421.

You,R. *et al.* (2018b) DeepText2GO: improving large-scale protein function prediction with deep semantic text representation. *Methods*, **145**, 82–90.

Yu,K. *et al.* (2018) Automatic extraction of protein–protein interactions using grammatical relationship graph. *BMC Med. Inf. Decision Mak.*, **18**, 42.