Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph

Konstantinos Bougiatiotis[1,2], Fotis Aisopos[1], Anastasios Nentidis[1,3], Anastasia Krithara[1], and Georgios Paliouras[1]

[1] Institute of Informatics and Telecommunications, National Center Scientific Research Demokritos, Athens, Greece
[2] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
[3] School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
kbogas@di.uoa.gr, {fotis.aisopos, tasosnent, akrithara, paliourg}@iit.demokritos.gr

Abstract. Knowledge Graphs provide insights from data extracted in various domains. In this paper, we present a new approach to discover probable drug-to-drug interactions, through the generation of a Knowledge Graph from disease-specific literature. The Graph is generated using natural language processing and semantic indexing of open biomedical publications and manually annotated resources. Then, the semantic paths connecting different drugs in the Graph are extracted and aggregated into feature vectors representing drug pairs. Finally, a classifier is trained on known interactions and is used to discover other possible interacting drug pairs. We evaluate this approach on two use cases, Alzheimer's Disease and Lung Cancer. A manually curated drug database is utilized as a golden standard and our system is shown to outperform competing graph embedding approaches, while also recommending new drug-drug interactions that are validated retrospectively.

Keywords: Literature Mining · Knowledge Graph · Path Analysis · Knowledge Discovery · Drug Interactions.

1 Introduction

Drug-Drug Interactions (DDIs) often occur in cases of simultaneous administration of multiple drugs. This may pose a serious problem for patient safety, as it seriously affects the biological action of the implicated drugs and may result in various adverse drug effects. The extent of the problem becomes more evident given that in the United States alone, DDIs are responsible for up to 195,000 hospital admissions [11].

However, this may pose a serious challenge due to the absence of sufficient clinical data and knowledge. Thus, automated software solutions that discover potential drug interactions can be valuable tools to improve health care and help pharmacovigilance. Many approaches try to address this by assessing structural or other kinds of drug similarities, based mainly on targets, pathways and

transporters [14]. However, most of these approaches fail to capture and combine information from heterogeneous sources of data which are important to address the complexity of the task.

The current paper proposes a holistic framework towards DDI prediction, based on a Biomedical Literature Knowledge Graph (DDI-BLKG)[4]. In the proposed framework, we extract knowledge items from biomedical publications and manually curated databases, using automated Natural Language Processing (NLP) tools. The results are integrated in a disease-specific Knowledge Graph (KG). Then, a human-curated drug database is used to train a classifier that identifies patterns of interactions between drug pairs. As features for the patterns, the classifier uses the semantic relations in the paths connecting interacting drugs. We showcase the usefulness of our approach by testing it on drug interactions for two prevalent diseases: Alzheimer's Disease (AD) and Lung Cancer (LC). Our experiments show that the proposed approach achieves better results than other graph embedding techniques on the same task. Moreover, through a small-scale qualitative evaluation we showcase the predictive potential of the method and its usefulness in providing novel DDIs.

Overall, the main contributions of this work correspond to the following:

- We present an automated DDI prediction approach, utilizing a disease-specific biomedical literature Knowledge Graph.
- We propose the use of the semantic relations connecting different drugs in the literature, as features for the DDIs.
- We make available for further experimentation two real-world disease-specific KGs, related to Alzheimer's Disease (AD) and Lung Cancer (LC) respectively, alongside the probable DDIs predicted by our model.

2 Related Work

Various existing approaches aim to extract associations and identify relations between biomedical entities directly from text [1,12]. However, in order to extend over the narrow scope of a sentence that rarely contains all the information needed, one needs to combine multiple sources of information. Such an example is the method proposed in [5], which builds a heterogeneous network and performs link prediction to construct an integrative model of drug efficacy.

Most relevant to our approach is the work in [2,17], presenting drug discovery methods, based on biomedical knowledge graphs. The former method focuses on treatment and causative relations exploiting connections of biomedical entities as found in literature. The latter publication presents SemaTyP, a method for discovering drug-disease relations based on a literature Knowledge Graph. Its successor, GrEDeL [16] extends the previous model by employing graph embedding techniques and deep learning approaches.

During the last years, there have been many other approaches to utilize graph embeddings for DDIs. Authors in [18] present KMR, a procedure for similarity computation based on chemical structure and side effects. On the other

[4] https://github.com/kbogas/DDI_BLKG

hand, Shtar et al. [19] employ adjacency matrix factorization to embed the drugs based on their interactivity as derived by DrugBank. Finally, authors in [6] also construct a biomedical Knowledge Graph from structured data (i.e. DrugBank, PharmGKB and KEGG databases), but with a limited set of drug-related nodes and relations.

In summary, various methods aim at completing a Graph and using it for DDI prediction. However, most of these approaches focus on either specific databases or literature without combining the two. In what follows, we illustrate that a multi-type and disease-specific approach could provide significant benefits.

3 Approach

3.1 General Workflow

Our approach to the creation of the biomedical literature Knowledge Graph and the development of a link prediction model consists of a sequence of distinct steps as shown in Figure 1. In the following sections, we will discuss each one in detail, leading to the predictive model to be validated.

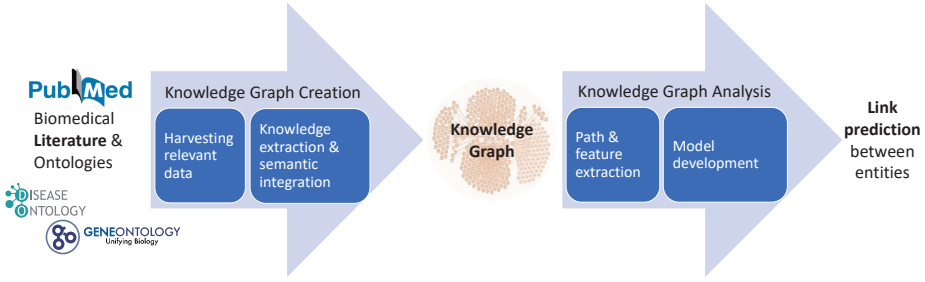

Fig. 1. Overview of the steps for the knowledge graph creation and link prediction task.

3.2 Knowledge Graph Creation

First, a disease-specific KG is created based on disease-related biomedical literature and structured resources. To this end, we utilize the project iASiS Open Data Graph [8] generation pipeline. According to this, biomedical articles related to the disease of interest are fetched from PubMed. These are analyzed using SemRep [13], a UMLS-based [3] tool that extracts biomedical predications i.e. semantic triples in the form of subject-predicate-object. Moreover, edges denoting co-occurrence of terms and topic annotations of the articles are also added to the KG. Finally, biomedical ontologies (e.g. Gene Ontology, Disease Ontology)

are also indexed under the UMLS schema and integrated into the resulting KG. The KG is built using the capabilities of the graph database Neo4j[5].

3.3 Knowledge Graph Analysis

The result of the previous step is a multi-relational KG, as seen in the upper-left part of Figure 2. The next step is to identify probable interactions between biomedical entities. Given a pair of drugs, the drug-drug interaction problem can be formulated as a link prediction problem on the KG. In our setting, we model this task as a supervised learning task where data samples are generated from different drug pairs found in the KG and the goal is to find which drug pairs interact.

In order to generate the data samples, we use DrugBank as a source of known interactions (ground truth) and map the drugs to the corresponding UMLS entities. For example, *Bivalirudin* (DB00006), an antithrombin drug, is mapped to two entities under the UMLS schema, *Hirulog* (C0210057) and *Bivalirudin* (C0168273). This process is depicted in the bottom-left part of Figure 2.

Then, we aggregate all possible paths in the KG between the examined pair of drug nodes. Let $E = \{e_1, e_2, ..., e_M\}$ denote all the relations (i.e. the edges) found in the graph. Also, let d_1, d_2 be two drug nodes and let π^l be a path of length l connecting d_1 and d_2. This path π consists of a series of relations starting from node d_1 and ending in node d_2 in the form of: $d_1 e_0 e_1 e_2 ... e_{l-1} d_2$. In this work, we limit $l \leq 3$ after observing that longer paths were of lower quality due to the high interconnectedness of the graph (i.e. with $l \geq 4$ almost all nodes were within reach from any other node). Thus, the representation of each path between a pair of drugs becomes: $\pi = e_0 e_1 e_2$.

Given two drugs d_1 and d_2 let $\mathit{\Pi} = \{\pi^l_1, \pi^l_2, ..., \pi^l_{N_{d_1,d_2}}\}$ be the set of all possible paths between them. An illustrated example of two such paths, as found between two drugs in the KG, can be seen in the upper-right part of Figure 2. Once the desired paths are retrieved, each one is processed in order to extract a feature representation of it.

We use 35 unique relation types from the UMLS Semantic Network, after merging the semantically similar ones and downsizing them from 55. We use one-hot encoding of these 35 relations as features on every possible hop. Therefore with $l = 3$ hops, the feature vector will have: $3 \times 35 = 105$ features. Each feature value in this vector will be either zero or one, denoting whether the specific relation was found in the specific hop.

Then, for a specific path $\pi^l_{d1,d2}$ between the drugs d_1,d_2 the corresponding feature vector will be: $x_i = [c_{r_1}, c_{r_2}, ..., c_{r_{105}}]$ where c_{r_j} is either 0 or 1 as mentioned before. For paths $\pi^m_{d1,d2}$ with $m < l = 3$ the last $(l-m)\times 35$ elements of the vector are set to zero. The result of this one-hot encoding process is illustrated in the middle-right part of Figure 2.

Eventually, we want to aggregate the information of all the paths into a single feature vector for each drug pair. Thus, we combine the feature vectors of the

[5] https://neo4j.com/

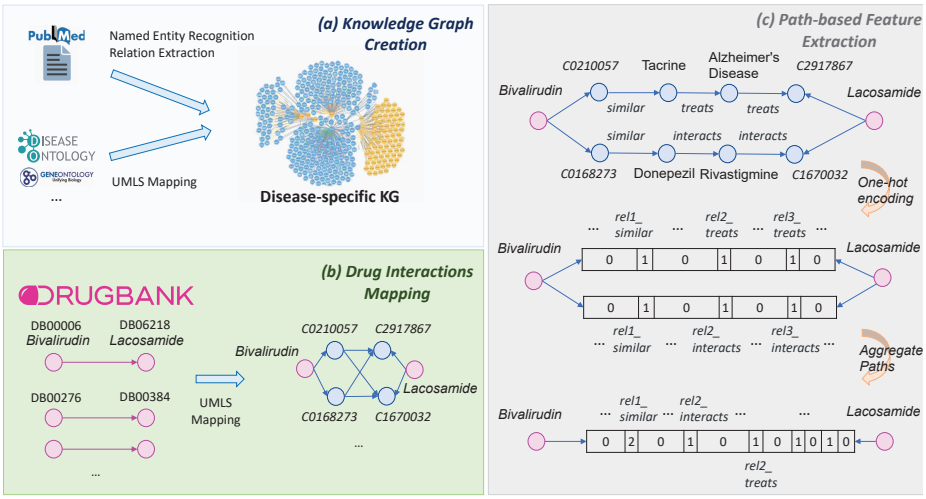

Fig. 2. Overview of the process leading to the final feature representations of the drug pairs.

individual paths leading from d_1 to d_2 by summing their corresponding feature vectors. Hence, the feature vector for a drug pair with N_{d_1,d_2} paths will be :

$$x_{d_1,d_2} = \sum_{i=1}^{N_{d_1,d_2}} x_i = \sum_{i=1}^{N_{d_1,d_2}} [c_{r_1}, c_{r_2}, ..., c_{r_{105}}]_i = [\sum_i c^i_{r_1}, \sum_i c^i_{r_2}, ..., \sum_i c^i_{r_{105}}] \quad (1)$$

where $c^i_{r_j}$ denotes the feature c_{r_j} of path x_i. The final outcome is shown in the bottom-right part of Figure 2. These features will be used downstream to train a classifier and predict new probable DDIs.

4 Evaluation and Results

In this section, we first introduce the details of the generated Knowledge Graph and the corresponding dataset for the experiments. Then, the competing methods are outlined and the evaluation task and procedure are described. Finally, the results are presented and discussed.

4.1 Knowledge Graph Creation

The pipeline described above was configured with the MeSH descriptors "Dementia" (D003704) and "Lung Neoplasms" (D008175) to create two disease-specific KGs related to two prevalent diseases with high societal impact. As a result, more than 250,000 documents were analysed yielding more than 27 million semantic triples for both diseases. Moreover, hierarchical relations were integrated

from the Disease Ontology, the Gene Ontology and the MeSH hierarchy adding more than 250,000 semantic triples in each KG. More details about the composition of the data and the type of nodes and edges as ingested by different data sources can be found in the repository of the proposed method.

4.2 Drug-Drug Interaction

Focusing on the task of predicting interactions between drugs, we used DrugBank 5.0.3[6] taking into account only AD and LC related drugs. There were 94 and 68 drugs respectively, according to their textual description. Using these drugs and their corresponding interactions with all other drugs in DrugBank we found 1326 and 4494 interacting drug pairs existing in the two KGs.

Additionally, we generated the (implicitly) negative pairs. We opted for a *corrupted sampling* procedure adapted to our setting. Specifically, let D^+ be the set of positive pairs. Then, for each positive pair $(d_1, d_2 \in D^+)$ a negative pair $(d_1, d' \notin D^+)$ was formed. This way of creating the dataset allows for the most "interacting" drugs to be represented equally in the two classes, marginalizing the factor of the different "interactivity" levels among the drugs. Moreover, to mimic the real world where actual DDIs are rare, in comparison to all possible drug pairs, we retrieved as many negative pairs could be found in the KGs, essentially oversampling the negative class. Details on the final datasets can be seen on Table 1.

Table 1. Drug pairs for each use case to be used for training and validation.

	Positive	**Negative**
AD	1,326	6,554
LC	4,494	28,752

4.3 Competing Methods

In order to evaluate the semantic path approach, as a way for creating expressive features for the drug pairs, we compared our method against several graph embedding approaches that are popular for link prediction tasks over KGs. To this end, we generated entity and relation embeddings, by training the respective models on the triples that make up each KG. We experimented with methods from different families of embeddings (i.e. translational, semantic matching, etc.), focusing heavily on tensor decomposition methods, due to their robustness in link prediction tasks [15].

The details and references on the competing methods can be seen in Table 2. The symbol ; symbolizes the concatenation of the embeddings, D_1 and D_2 the embedding of the first and second drug of the pair respectively and $R_{INTERACTS}$ the embedding of the relation *INTERACTS*.

[6] https://www.drugbank.ca/releases/5-0-3

Table 2. Feature representations of the drug pairs for each competing method.

Method	Feature Vector From Embeddings	Size
TransE [4]	$[D_1; R_{INTERACTS}; D_2]$	300
RESCAL [10]	$[(D_1 \times R_{INTERACTS}); D_2]$	200
HolE [9]	$[D_1; R_{INTERACTS}; D_2]$	300
DistMult [20]	$[D_1; R_{INTERACTS}; D_2]$	300

It is worth noting that we also tried using directly the score for each drug pair as generated by the corresponding graph embedding model. However, the results were much worse compared to using the embeddings generated from each model as feature vectors for a classifer.

4.4 Experimental Setup

Having generated the various feature presentations, the effectiveness of each one is measured using an extensive cross-validation (cv) procedure. Specifically, a nested cv scheme with an outer 10-fold cv to estimate the performance of the model and an inner 5-fold cv to tune the hyperparameters of the classifier was used. For all the different feature representations a *Random Forest* was used as the final predictor. The hyperparameters of each forest (e.g. number of trees, maximum depth, etc.) were optimized independently for each model. Performance is calculated using the *area under the receiver-operating characteristic* (AUROC), while also the F_1-score and the *area under the precision-recall curve* (AUPRC) for the positive class are calculated. Higher values always indicate better performance for all measures.

Finally, regarding the graph embedding procedures, the *TorchKGE*[7] library was used to train the models and generate the embeddings. All methods were allowed to train for a maximum of 100 epochs while early stopping was used, utilizing 10% of the data for validation purposes. For each model, 100-sized embeddings were used, as they seemed to converge faster and increasing the embedding size did not provide better results.

4.5 Results

The results of the evaluation can be seen in Figure 3. We can clearly see that the *DDI-BLKG* method outperforms the competing approaches on both disease-specific KGs. Although all of the models seem to be doing well when focusing on their ROC-AUC scores, our model surpasses the competition by far when focusing on the positive class. This indicates that the proposed methodology, captures important information regarding the DDIs and the predicted interactions are precise despite the class imbalance.

In order to evaluate the predictive capabilities of our method, we also performed a small-scale qualitative analysis of its top predictions. Taking the top-10

[7] https://torchkge.readthedocs.io

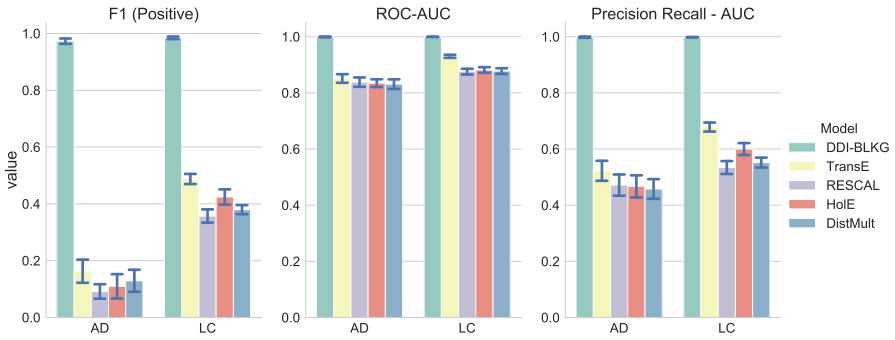

Fig. 3. Results of the 10-fold validation for the both use-cases.

scoring false positives (i.e. DDIs that were not present in DrugBank 5.0.3), we evaluated them using a newer version (5.3.1). The results were promising as 5/10 drug pairs for the AD, and 7/10 for the LC graph, were actual DDIs in the more recent version. The predicted drug-pairs that were validated on the new database can be seen in Table 3. As an example of the AD drugs, *Estradiol* is a form of estrogen used to treat menopause symptoms and its relation with *Memantine* (medication for mild-to-severe AD) has been thoroughly studied [7]. Taking into account the interaction with *Rivastigmine* (another medication for moderate AD), our model has captured the possible DDIs that emerge with hormone therapy trials after menopause and the adverse effects they can cause in Alzheimer's patients.

Table 3. Most probable DDIs that have been retrospectively validated in DrugBank 5.3.1.

Disease	Drug	Interacting Drugs
AD	Estradiol	Memantine, Rivastigmine, Trientine
AD	Trientine	Leuprolide, Rivastigmine
LC	Cisplatin	Flunitrazepam, Magnesium hydroxide Sparfloxacin
LC	Gemcitabine	Abacavir, Anagrelide, Isosorbide Mononitrate
LC	Docetaxel	Pranlukast

Regarding the LC drugs, *Cisplatin* (a chemotherapy drug) is commonly provided in conjunction with the predicted interacting drugs to combat the heavy side-effects of chemotherapy (e.g. *Flunitrazepam* as an anti-emetic treatment and *Mangesium hydroxide* to combat cisplatin-induced hypomagnesemia). However, adverse effects may arise, as in these DDIs, with unwanted changes in the metabolism and the serum concentration of the drugs. It is also worth noting that the predicted DDIs are related to a small set of drugs, namely *Estradiol* and

Trientine for the AD use case and *Cisplatin* and *Gemcitabine* for the LC use case. It is interesting to look further into the predicted DDIs, taking into account the interactivity of the drugs, their popularity (e.g. in how many publications they were mentioned) and their type. Overall, 60% of the predicted DDIs are present in the new version of DrugBank. This is an indication of the usefulness of our method and its capability to propose probable DDIs for further validation.

5 Conclusion

In this work, we proposed a new approach to predicting DDIs as a downstream task in a semantically-rich Knowledge Graph. First, we utilized a KG creation workflow to integrate knowledge extracted from biomedical literature and structured databases into a common semantic graph representation. Then, we extracted expressive features for the drug pairs based on the semantic paths that connect them. The experimental results validated the basic premise of our research regarding the latent knowledge of the extracted paths and that our method can be used to effectively discover interacting drug pairs. Source code for the methods and pre-processed datasets to replicate the results have also been made available.

As next steps, further experiments on the importance of each feature and the common patterns indicating a DDI should be conducted. Moreover, an extended retrospective analysis using the latest versions of DrugBank should be done, to assess the predictive performance of the method. As a final note, it is important to stress that the presented methodology can be used in many different tasks involving different pairs of nodes such as drug-disease, drug-side effect, gene-disease, etc. This stems from the fact that the feature extraction methodology was use case agnostic and dependent only on the topological and relational aspects of the paths between different pairs of nodes.

Acknowledgments

This work is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 727658, project iASiS[8] (Integration and analysis of heterogeneous big data for precision medicine and suggested treatments for different types of patients). We would also like to acknowledge the helpful comments and discussions with all iASiS partners that lead to the completion of this work.

References

1. Arnold, P., Rahm, E.: Semrep: A repository for semantic mapping. Datenbanksysteme für Business, Technologie und Web (BTW 2015) (2015)

[8] http://project-iasis.eu/

2. Bakal, G., Talari, P., Kakani, E.V., Kavuluru, R.: Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. Journal of biomedical informatics **82**, 189–199 (2018)
3. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
4. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 2787–2795. Curran Associates, Inc. (2013)
5. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S.E.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife **6**, e26726 (2017)
6. Karim, M.R., Cochez, M., Jares, J.B., Uddin, M., Beyan, O., Decker, S.: Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-lstm network. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 113–123 (2019)
7. Lamprecht, M.R., Morrison III, B.: A combination therapy of 17β-estradiol and memantine is more neuroprotective than monotherapies in an organotypic brain slice culture model of traumatic brain injury. Journal of neurotrauma **32**(17), 1361–1368 (2015)
8. Nentidis, A., Bougiatiotis, K., Krithara, A., Paliouras, G.: Semantic integration of disease-specific knowledge. In: IEEE 33rd International Symposium on Computer Based Medical Systems (CBMS) (to appear) (2020)
9. Nickel, M., Rosasco, L., Poggio, T.: Holographic embeddings of knowledge graphs. In: Thirtieth Aaai conference on artificial intelligence (2016)
10. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: Icml. vol. 11, pp. 809–816 (2011)
11. Percha, B., Altman, R.B.: Informatics confronts drug–drug interactions. Trends in pharmacological sciences **34**(3), 178–184 (2013)
12. Percha, B., Altman, R.B.: A global network of biomedical relationships derived from text. Bioinformatics **34**(15), 2614–2624 (2018)
13. Rindflesch, T.C., Fiszman, M.: The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of biomedical informatics **36**(6), 462–477 (2003)
14. Rohani, N., Eslahchi, C.: Drug-drug interaction predicting by neural network using integrated similarity. Scientific reports **9**(1), 1–11 (2019)
15. Rossi, A., Firmani, D., Matinata, A., Merialdo, P., Barbosa, D.: Knowledge graph embedding for link prediction: A comparative analysis (2020)
16. Sang, S., Yang, Z., Liu, X., Wang, L., Lin, H., Wang, J., Dumontier, M.: Gredel: A knowledge graph embedding based method for drug discovery from biomedical literatures. IEEE Access **7**, 8404–8415 (2018)
17. Sang, S., Yang, Z., Wang, L., Liu, X., Lin, H., Wang, J.: Sematyp: a knowledge graph based literature mining method for drug discovery. BMC bioinformatics **19**(1), 193 (2018)
18. Shen, Y., Yuan, K., Yang, M., Tang, B., Li, Y., Du, N., Lei, K.: Kmr: knowledge-oriented medicine representation learning for drug–drug interaction and similarity computation. Journal of cheminformatics **11**(1), 22 (2019)
19. Shtar, G., Rokach, L., Shapira, B.: Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures. PloS one **14**(8) (2019)
20. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)