Evaluation of Knowledge Graph Embedding Approaches for Drug-Drug Interaction Prediction using Linked Open Data

Remzi Çelebi[1,2], Erkan Yasar[1], Huseyin Uyar[1], Ozgur Gumus[1], Oguz Dikenelli[1] and Michel Dumontier[2]

[1] Ege University Computer Engineering Department, Izmir, Turkey,
{erkan.yasar,ozgur.gumus,oguz.dikenelli}@ege.edu.tr,
uyarhuseyin1@gmail.com
[2] Institute of Data Science, Maastricht University, Maastricht, Netherlands,
{remzi.celebi,michel.dumontier}@maastrichtuniversity.nl

Abstract. Current approaches to identifying drug-drug interactions (DDIs), which involve clinical evaluation of drugs and post-marketing surveillance, are unable to provide complete, accurate information, nor do they alert the public to potentially dangerous DDIs before the drugs reach the market. Predicting potential drug-drug interaction helps reduce unanticipated drug interactions and drug development costs and optimizes the drug design process. Many bioinformatics databases have begun to present their data as Linked Open Data (LOD), a graph data model, using Semantic Web technologies. The knowledge graphs provide a powerful model for defining the data, in addition to making it possible to use underlying graph structure for extraction of meaningful information. In this work, we have applied Knowledge Graph (KG) Embedding approaches to extract feature vector representation of drugs using LOD to predict potential drug-drug interactions. We have investigated the effect of different embedding methods on the DDI prediction and showed that the knowledge embeddings are powerful predictors and comparable to current state-of-the-art methods for inferring new DDIs. We have applied Logistic Regression, Naive Bayes and Random Forest on Drugbank KG with the 10-fold traditional cross validation (CV) using RDF2Vec, TransE and TransD. RDF2Vec with uniform weighting surpass other embedding methods.

Keywords: linked open data, knowledge graph embedding, drug-drug interaction prediction, machine learning

1 Introduction

Adverse Drug Events (ADEs) are a significant threat to public health. A study by [14] estimates 6.7% of hospitalized patients experience serious adverse drug effects with fatality rate 0.32% in the USA. In 2014, 807,270 cases of serious

ADEs were reported in the United States, resulting in 123,927 lost lives [3] . ADEs present a financial burden to the healthcare system due to the costs of further hospitalization, morbidity, mortality, and health-care utilization. The majority of adverse drug effects are caused by unintended drug-drug interactions (DDIs), which occasionally arise through co-prescription of a drug with other drug(s) [2]. Patient groups such as elderly patients and cancer patients are more likely to take multiple drugs simultaneously, which increases their risk of DDIs [20, 9]. Current approaches to identifying DDIs, which involve clinical evaluation of drugs and post-marketing surveillance, are unable to provide complete, accurate information, nor do they alert the public to potentially dangerous DDIs before the drugs reach the market [19]. Predicting potential drug-drug interaction helps reduce unanticipated drug interactions and drug development costs and optimizes the drug design process. Thus, there is clear need for automated methods for detecting drug-drug interactions.

In recent years, biological data and knowledge bases have been increasingly built on Semantic Web technologies and knowledge graphs are used for information retrieval, data integration, and federation. Many bioinformatics databases have begun to present their data as Linked Open Data (LOD), a graph data model, using Semantic Web technologies [22, 13]. The knowledge graphs provide a powerful model for defining the data, in addition to making it possible to use underlying graph structure for extraction of meaningful information. In this work, we have applied Knowledge Graph Embedding approaches to extract feature vector representation of drugs using DrugBank LOD from Bio2RDF to predict potential drug-drug interactions. This study also aims to investigate the effect of different embedding methods and linked data sources on the DDI prediction task.

Researchers have used various approaches and data sources to predict novel drug interactions [19]. These approaches include extracting DDI statements from medical texts and drug event reports [27], inferring DDI mechanism [17] by integration knowledge from several sources and using network proximities [5]. The machine learning based approaches have commonly used pharmacological similarities of drugs as features [6]. Gottlieb et al. [10], by using different drug similarity metrics, developed a new prediction framework called INDI. INDI trained a logistic classifier using 7 similarities, also using them to calculate their maximum likelihood by using known drug-drug interactions. Cheng et al. [7] presented the HNAI framework for predicting drug interactions using phenotypic, therapeutic, structural, and genomic similarities of drugs. Cami et al. [5] have trained a logistic classifier by extracting the pharmacological and graph/network qualities between drugs. Zhang et al. [25] used a label propagation method on drug chemical infrastructure, drug side effect and drug off-side effects. Li et al. [15] have developed a Bayesian network that combines drug molecular similarity and drug phenotypic (side effect) similarity to predict the combination effect of drugs. Zhang et al. [26] collects a variety of drug data and thus predicts drug-drug

[3] https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070461.htm

interactions by integrating chemical, biological, phenotypic and network data. Abdelaziz et al. [1] presented Tiresias, a similarity-based framework for predicting DDIs. They used 1,014 features derived from pharmacological similarities and from drug text and similarity based on the Knowledge Graph embeddings (TransE and HolE). Each feature represents the similarity value of the known interacting drug pair to the most similar drug pair.

This work differs from other previous machine learning based approaches in the following aspects: i) Many existing methods have used similarities of drugs based on the properties such as targets, side-effects, fingerprint and indications [10, 26, 23, 7]. Each similarity is used as a feature for a binary classifier, but there is a limited number of these features. In the proposed approach a drug is characterized with a feature vector large enough to increase the classifier's predictive power. It is possible to use these feature vectors in other drug-related machine learning tasks (e.g. drug-target, drug-adverse effect). ii) We are able to make predictions for the drugs that have missing or inadequate information. Owing to linked open data, the presence of an entity (drug) is sufficient to enable embedding vectors for machine learning to be extracted. Most drugs and hence DDIs could be included in the training set with this intention, enabling deep learning models to be used. Similarity-based approaches, in contrast, do not allow for the calculation of various similarities for many drugs due to lack of drug information.

We used Knowledge Graph Embedding-based drug vectors to train various classifiers for DDI prediction. Our results show that performance of drug vector representation is comparable to the existing pharmacological similarity-based DDI prediction methods. The AUC score of 0.93 and F-Score of 0.86 were achieved based on ten cross-validations with the vector representations of drugs for the Drugbank dataset. Finally, we make our work open and freely available so that others can use or extend this methodology [4].

2 Materials and Method

2.1 Materials

Drugbank v5.0 [24] contains 288,856 distinct pairwise DDIs spanning 2,551 drugs. We were able to extract features for 2,124 drugs of these 2,551, filtering out the drugs that have no calculated feature vector. Thus, the number of DDIs was reduced to 253,449 .

2.2 Method

The steps of our RDF Graph Embedding based DDI prediction methodology are shown in Figure 1. The first step is to construct knowledge graph data in RDF format. And then as second step, the feature vector of drugs is extracted using the knowledge graph by applying different Graph Embedding approaches namely

[4] https://github.com/rcelebi/GraphEmbedding4DDI/

RDF2VEC, TransE and TranD. The last step is to predict drug interactions using extracted feature vectors by applying three different classifiers: Logistic Regression (LR), Naive Bayes (NB) and Random Forest (RF).

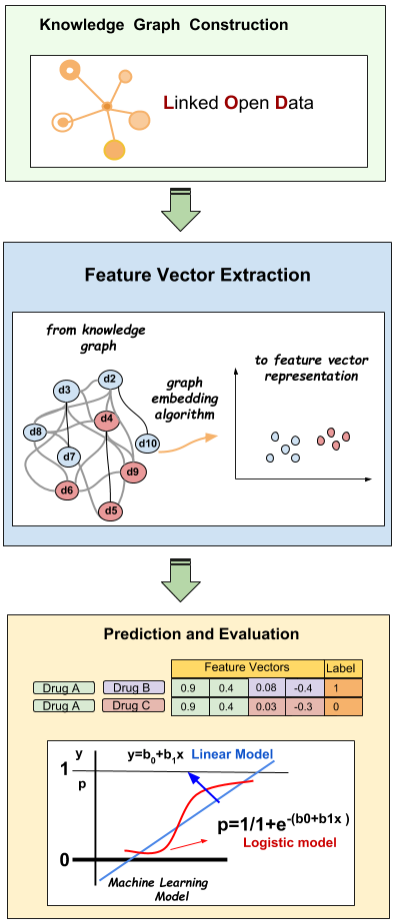

Fig. 1. Overview of our methodology.

Knowledge Graph Construction We used an already linked open biological dataset, Bio2RDF [4], as background knowledge to extract drug features. Bio2RDF is an open-source project that integrates numerous Life Sciences databases available on different websites, providing a data integration service for scientific researchers. Bio2RDF created a large RDF graph that interlinks data from major biological databases related to biological entities such as drug, protein, pathway and disease. In this study, Drugbank dataset within Bio2RDF project release 4.0 was used as the background knowledge graph after removing the drug-drug interaction information ('drugbank_vocabulary:ddi-interactor-in' relations). The number of triples, entities and relation types in the Drugbank dataset are 2,588,933, 574,152 and 76 respectively.

Feature Vector Extraction We have tested multiple successful approaches for knowledge graph embeddings to generate vector representation of drugs from graphs such as RDF2Vec [21], TransE [3] and TransD [12]. To represent feature vector of a drug pair, we concatenated embedding vectors of each drug in the pair. These approaches are explained in detail in the following subsections.

– **RDF2Vec**
RDF2Vec is a recently published methodology that adapts the language modeling approach of word2Vec [16] to RDF Graph Embeddings. Word2Vec

trains a neural network model to learn vector representation of words, called word embeddings. It maps each word to a vector of latent numerical values in which semantically and syntactically closer words will appear closer in the vector space. The hypothesis which underlies this approach is that closer words in word sequence are statistically more dependent. RDF2Vec applies a similar approach to RDF Graph, considering the entities and relations between entities by converting the graph into set of sequences (walks or paths) and training the neural network model to learn vector representation of entities from the RDF graph.

Graph Walks
G (V, E) is a graph with V nodes and E edges. The random walk algorithm was used to generate P_v paths at depth d starting at each vertex v in V. At first iteration, the algorithm traverses the direct outgoing edges of a root vertex (vr), then randomly exploring the connected edges through visited vertices until d iterations is reached . The union of all the P_{vr} walks, starting from all entities (vr) in the knowledge network were used as a set of sequences to train artificial neural network models.
By biasing the walks, we could capture more meaningful information and therefore better representation of entities. To do this, each edge is assigned a weight and the walks will follow an edge with a probability based on its weight in selection method similar to roulette wheel selection. In their study, Cochez et al. [8] discussed three successful weighting strategies for RDF2Vec; Uniform, PageRank[18], PageRank split. Uniform weight is the standard approach taken by RDF2Vec where each edge has equal probability to be followed. PageRank weighting assigns or splits (divides) PageRank score of a node to its incoming edges.

Neural Network Training
Each word (entity) is trained to maximize its log probability according to the context words within the fixed-size window. Each word in the vocabulary is represented by two vectors; input and output vectors. While learning the input vectors is cheap, learning the output vectors is very expensive. Approximation techniques such as hierarchical softmax and negative sampling have been developed for efficient training. Word2vec introduces two architectures to obtain vector embedding representation of words: Continuous Bag-of-Words (CBOW) and Skip-Gram.

Continuous Bag-of-Words Model
The CBOW model is a two-layer artificial neural network model that predicts a target word using context words in near proximity. Given word sequence $w_1, w_2, w_3, .., w_T$, CBOW tries to maximize the average log probability of the target word as follows:

$$\frac{1}{T}\sum_{t=1}^{T} logp(w_t|w_{t-c} + \cdots + w_{t+c}) \tag{1}$$

where c is the context window and p defined as :

$$p(w_t|w_{t-c}+\cdots+w_{t+c}) = \frac{exp(\overline{v}^T v'_{w_t})}{\sum_{w=1}^{V} exp(\overline{v}^T v'_w)} \tag{2}$$

where v'_w is output vector of word w, V is the complete vocabulary of words and $\overline{v}$ is the averaged input vector of all the context words.

Skip-Gram Model
While CBOW predicts the word given the context, the Skip-gram predicts the context of the given word. It tries to find useful word representations to predict the words around the target word in a training document or sentences. Given word sequence $w_1, w_2, w_3, .., w_T$ and context window size c, Skip-gram maximizes the average log probability as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j=0} logp(w_{t+j}|w_t) \tag{3}$$

where p is defined using softmax function as follows:

$$p(w_{t+j}|w_i) = \frac{exp(v'^T_{w_{t+j}} v_{w_t})}{\sum_{w=1}^{V} exp(v'^T_{w_k} v_{w_t})} \tag{4}$$

where v_w and v_w are the input and the output vector of the word w, and V is the complete vocabulary of words.

– **TransE**
TransE embeds every entity and relation in the knowledge graph (KG) into low-dimensional vectors where the relations are represented as translation from head entity to tail entity. For a triple (h, r, t) in KB, the embedding head h is close to the embedding tail t by adding the embedding relation r, that is $h + r \approx t$. A vector representation of every entity and relation in the KG could be computed by learning a neural network model, which minimizes difference between its head entity and its tail entity in embedding space. TransE is convenient for modeling one-to-one relations, but is insufficient for one-to-many, many-to-one and many-to-many relations.

– **TransD**
In TransD, each entity or relation is defined by two vectors; one being the embedding vector of an entity or a relation, the other the projection vector. The projection vector represents the way to project an entity vector into a relation vector space to be used to construct mapping matrices. Every entity-relation pair has a unique mapping matrix. Thus, it can handle one-to-many, many-to-one and many-to-many relations. In addition, TransD has no matrix-by-vector operations which can be replaced by vectors operations.

Prediction and Evaluation

- **Data balance:** For DDI prediction using supervised machine learning, a binary classifier needs negative and positive example sets. In previous studies the negative set typically was chosen randomly from unknown interactions. Alternatively, the set of all unknown interactions could be designated as the negative set, but designating all unknown interactions as the negative set creates a data balance issue, influencing performance metrics (such as AUPR and F1-score). Other studies accounted for this issue through a random undersampling from these unknown interactions at a ratio corresponding to the positive set [7], or inferring negatives by clustering [11]. In this study, the negative samples were taken from unknown drug pairs in sample size equivalent to the positive samples.
- **Evaluation Metrics:** While many studies use the AUC score in computational prediction for drug-drug interactions, some studies [1, 11] have emphasized that this score is insufficiently accurate, therefore metrics such as AUPR and F1 score are viable alternatives. We used the evaluation metrics including AUC, F1 score and AUPR to accurately measure the performance of our classifiers.
- **Parameters:** We combined the generated walks to be used as input to RDF2Vec where the graph walk parameters are depth = 1,2,3,4 and walks per entity = 250. And we trained the word2vec model using CBOW and SG neural network architectures with the following parameters; window size = 5, number of iterations = 5, negative samples = 25 and dimension = 100. The size of each drug vector is 100. Thus, the classifiers used 200 features for prediction of DDIs. The default parameters given by OpenKE (openke.thunlp.org) were used for TransE and TransD models. Logistic Regression (LR), Naive Bayes (NB) and Random Forest (RF) were trained using Scikit-learn machine learning package. The parameters used for building the classifiers are as follows; C=0.01 for LR, Gaussian version for NB and number of estimators = 200 for RF.

3 Results

We first performed the experiments applying Logistic Regression, Naive Bayes and Random Forest on Drugbank KG with ten repetitions of 10-fold traditional cross validation (CV) using three well known knowledge graph embedding methods, namely RDF2Vec, TransE and TransD. The results of the experiments are shown in Table 1. RDF2Vec using uniform weight strategy has performed better than the other graph walks generation methods and embedding methods. RDFVec uniform-weight embedding vectors using Skip-Gram Neural Network achieved the best performance values. The best AUC value obtained is 0.932 and the best F-Score value is 0.860 using Random Forest learning algorithm.

Comparison with the state-of-art methods: In spite of the high number of methods which have been proposed for DDI prediction, their results have had insufficient basis for comparison because of the differing terms of their datasets

Table 1. AUC and F-score values for Drugbank Knowledge Graph using different embedding methods

Embedding Method			AUC			F-score		
			NB	LR	RF	NB	LR	RF
RDF2Vec	Uniform	CBOW	0.703	0.787	0.927	0.640	0.715	0.853
		SG	0.739	0.799	**0.932**	0.683	0.730	**0.860**
	PageRank	CBOW	0.625	0.682	0.891	0.495	0.622	0.815
		SG	0.642	0.695	0.893	0.621	0.634	0.816
	PageRank Split	CBOW	0.618	0.676	0.893	0.548	0.619	0.817
		SG	0.636	0.706	0.894	0.635	0.644	0.817
TransE			0.734	0.763	0.912	0.686	0.700	0.835
TransD			0.714	0.738	0.914	0.676	0.679	0.836

(known DDIs) and evaluation methodologies of the studies. The most noteworthy of these studies is the Tiresias framework [1], which uses both pharmacological similarities and similarities from embedding features. Tiresias has reported an F-score of 0.851 and AUPR of 0.919, all features included, as their best results and an F-score of 0.813 and AUPR of 0.887 with only pharmacological similarity features (equivalent performance with INDI [10]) using Drugbank version 4.0. Our embedding using the same DDI dataset with similar settings achieved a high F-score of 0.867 and AUPR of 0.918. It shows that the proposed embedding based method is comparable to current state-of-the-art methods.

4 Conclusion

To the best of our knowledge, this study used the largest DDI dataset (the number of known DDIs is 253,449) available to be used as input for our machine learning models. Our methodology enabled us to extract features for a large number of samples which is essential for deep learning methods to be applied. Previous studies used much lesser known DDI samples ($\approx 40 - 50K$).

We have applied Logistic Regression, Naive Bayes and Random Forest on Drugbank KG with the 10-fold traditional cross validation (CV) using RDF2Vec, TransE and TransD. RDF2Vec with uniform weighting surpass other embedding methods.

In this study, knowledge graph embedding feature vectors were used to predict the potential DDIs. We have investigated the effect of different embedding methods on the DDI prediction. We showed that the knowledge embeddings are powerful predictors and comparable to current state-of-the-art methods for inferring new DDIs. One limitation of our method is that it does not provide the mechanistic explanations for predicted potential DDIs since the embedding features were constructed using a black-box model (neural network). Through the integration with Electronic Health Record (EHR) system, the predictions made by our approach could be valuable to the system, as it would prevent co-

prescription of potentially hazardous interacting drugs. Consideration of these predictions would also be helpful to the design large-scale clinical trials.

Acknowledgements

The first named author (R.C.) is grateful to TUBITAK for providing financial support under 2214-A programme. This work was supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3454-01-01, FCC/1/1976-08-01, and FCS/1/3657-02-01. This work was also supported by Ege University Research Fund through the 16-MUH-095 BAP project.

References

1. Abdelaziz, I., Fokoue, A., Hassanzadeh, O., Zhang, P., Sadoghi, M.: Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. Web Semantics: Science, Services and Agents on the World Wide Web 44, 104–117 (2017)
2. Bates, D.W., Cullen, D.J., Laird, N.: Incidence of adverse drug events and potential adverse drug events. Int. J. Trauma Nurs. 1(4), 117 (1995)
3. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems. pp. 2787–2795 (2013)
4. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) The Semantic Web: Semantics and Big Data. pp. 200–212. Lecture Notes in Computer Science, Springer Berlin Heidelberg (May 2013)
5. Cami, A., Manzi, S., Arnold, A., Reis, B.Y.: Pharmacointeraction network models predict unknown drug-drug interactions. PLoS One 8(4), e61468 (Apr 2013)
6. Celebi, R., Mostafapour, V., Yasar, E., Gumus, O., Dikenelli, O.: Prediction of Drug-Drug interactions using pharmacological similarities of drugs. In: 2015 26th International Workshop on Database and Expert Systems Applications (DEXA) (2015)
7. Cheng, F., Zhao, Z.: Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J. Am. Med. Inform. Assoc. 21(e2), e278–86 (Oct 2014)
8. Cochez, M., Ristoski, P., Ponzetto, S.P., Paulheim, H.: Biased graph walks for RDF graph embeddings. In: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics - WIMS '17 (2017)
9. Doubova Dubova, S.V., Reyes-Morales, H., Torres-Arreola, L.d.P., Suárez-Ortega, M.: Potential drug-drug and drug-disease interactions in prescriptions for ambulatory patients over 50 years of age in family medicine clinics in mexico city. BMC Health Serv. Res. 7, 147 (Sep 2007)
10. Gottlieb, A., Stein, G.Y., Oron, Y., Ruppin, E., Sharan, R.: INDI: a computational framework for inferring drug interactions and their associated recommendations. Mol. Syst. Biol. 8 (2012)

11. Hameed, P.N., Verspoor, K., Kusljic, S., Halgamuge, S.: Positive-Unlabeled learning for inferring drug interactions based on heterogeneous attributes. BMC Bioinformatics 18(1), 140 (Mar 2017)
12. Ji, G., He, S., Xu, L., Liu, K., Zhao, J.: Knowledge graph embedding via dynamic mapping matrix. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). vol. 1, pp. 687–696 (2015)
13. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., Jenkinson, A.M.: The EBI RDF platform: linked open data for the life sciences. Bioinformatics 30(9), 1338–1339 (May 2014)
14. Lazarou, J., Pomeranz, B.H., Corey, P.N.: Incidence of adverse drug reactions in hospitalized patients. JAMA 279(15), 1200 (1998)
15. Li, P., Huang, C., Fu, Y., Wang, J., Wu, Z., Ru, J., Zheng, C., Guo, Z., Chen, X., Zhou, W., Zhang, W., Li, Y., Chen, J., Lu, A., Wang, Y.: Large-scale exploration and analysis of drug combinations. Bioinformatics 31(12), 2007–2016 (Jun 2015)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
17. Noor, A., Assiri, A., Ayvaz, S., Clark, C., Dumontier, M.: Drug-drug interaction discovery and demystification using semantic web technologies. J. Am. Med. Inform. Assoc. 24(3), 556–564 (May 2017)
18. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web (Nov 1999)
19. Percha, B., Altman, R.B.: Informatics confronts drug–drug interactions. Trends Pharmacol. Sci. 34(3), 178–184 (2013)
20. Riechelmann, R.P., Tannock, I.F., Wang, L., Saad, E.D., Taback, N.A., Krzyzanowska, M.K.: Potential drug interactions and duplicate prescriptions among cancer patients. J. Natl. Cancer Inst. 99(8), 592–600 (Apr 2007)
21. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: Lecture Notes in Computer Science, pp. 498–514 (2016)
22. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome Biol. 6(5), R46 (Apr 2005)
23. Vilar, S., Uriarte, E., Santana, L., Lorberbaum, T., Hripcsak, G., Friedman, C., Tatonetti, N.P.: Similarity-based modeling in large-scale prediction of drug-drug interactions. Nat. Protoc. 9(9), 2147–2163 (Sep 2014)
24. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 36(suppl_1), D901–D906 (2007)
25. Zhang, P., Wang, F., Hu, J., Sorrentino, R.: Label propagation prediction of Drug-Drug interactions based on clinical side effects. Sci. Rep. 5, 12339 (Jul 2015)
26. Zhang, W., Chen, Y., Liu, F., Luo, F., Tian, G., Li, X.: Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC Bioinformatics 18(1), 18 (Jan 2017)
27. Zhang, Y., Zheng, W., Lin, H., Wang, J., Yang, Z., Dumontier, M.: Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics 34(5), 828–835 (Mar 2018)