Proceedings of the 41st Chinese Control Conference
July 25-27, 2022, Hefei, China

An Approach for Anticancer Drug Response Prediction Based on Knowledge Graph Embedding

Xinping Xie[1] Guanfu Wang[1] Weiwei Zhu[2,3] Shasha Shi[1] Xiaodong Du[4*] Hongqiang Wang[2,3*]

[1]School of mathematics and physics, Anhui Jianzhu University, Hefei, China
E-mail：592745039@qq.com，1808040693@qq.com
[2] Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China
[3] Zhongqi AI Lab, Hefei, China
E-mail: juway@126.com，hqwang126@126.com
[4]Experimental Teaching Center, Hefei University, Hefei, China
E-mail: xdongdu@hfuu.edu.cn

Abstract： Predicting the response of tumor cells to antitumor drugs is a hot research topic in precision medicine. Currently, most existing methods use genomics data to build prediction models that often ignore regulatory relationships between genes. In this study, we propose a drug sensitivity prediction method based on knowledge graph embedding, which can effectively tap the drug response related gene regulatory features. First, tumor cells are treated as nodes and fused into gene regulatory networks, and the gene regulation-fused expression representation of tumor cell nodes is obtained using the knowledge graph embedding technique (KGE), and then, based on the new expression representation of tumor cells, a logistic regression drug response is established for each drug. Evaluation experiments on the GDSC dataset showed the method can extract drug response-related features and obtain a better response prediction performance with an average AUC of 0.703 for 189 drugs.

Keywords: Precision medicine, gene regulatory networks, knowledge graph embedding, drug sensitivity prediction, logistic regression

1 Introduction

The International Agency for Research on Cancer (IARC) recently released data showing that about 20 million people worldwide were diagnosed with cancer in 2020, and 9.96 million died of cancer [1]. One of the biggest challenges in cancer treatment is to predict the therapeutic effect of anti-cancer drugs on patients because of the individual variability of patients and the large differences in therapeutic effect of the same anti-cancer drug for different patients. Accurate selection of appropriate anti-cancer drugs for cancer patients can effectively increase the possibility of recovery for cancer patients in view of their individual differences. Clinically trying drugs to cancer patients one by one often delays correct drugs and leads to large side effects on cancer patients' bodies, so preclinical prediction of drug response is highly anticipated.

Currently, many drug response prediction methods have been proposed based on tumor cell genomic data[2]. Early methods were mainly based on traditional statistical models or machine learning models, including regression, elastic networks, random forests, support vector machines or shallow neural networks. For example, Dong [3] et al. proposed an SVM classification model in which the obtained AUC of drugs, SNX-2112, BIX02189 and Belinostat, was greater than 0.8.

With the recent development of information technology, deep learning models have received increasing attention in the field of drug response prediction research. For example, Chang [4] et al. proposed a one-dimensional deep convolutional neural network-based drug response prediction model. The method encoded 3072 bits of chemical information of drugs by PaDEL, and obtained the R^2 values as high as 0.851. Although drug response prediction has achieved good results, some studies found

that drug targets and protein interaction information also take non-trivial impacts the prediction of personalized drug response [5], and by building a network, protein interaction information can be effectively integrated with the genomic information of tumor cells.

In this study, we proposed a knowledge graph embedding-based drug sensitivity prediction method KGE-LR: firstly, tumor cells were treated as nodes by fusing into gene regulatory networks, and the gene regulation fused expression representations of tumor cell nodes were obtained using the knowledge graph embedding technique, and then a logistic regression (LR) drug response prediction model was developed for each drug based on the new expression representations of tumor cell nodes. The prediction performance of KGE-LR was validated in the GDSC dataset, with a mean AUC of 0.703 on 189 drugs surpassing previous methods. In particular, for some drugs SNX-2112, CAY10603, and AT-7519, the resulted AUC is greater than 0.9.

2 Method

In this section, we will introduce the detailed information about the KGE-LR. As illustrated in Fig. 1, KGE-LR involves four steps: (**A**) Step 1: Construction of cell-gene fusion regulatory network; (**B**) Step 2: Learning new tumor cell representations; (**C**) Step 3: Development of logistic regression drug response prediction model for each drug; (**D**) Step 4: Prediction of tumor cell response (sensitivity or resistance) with anti-cancer drugs.

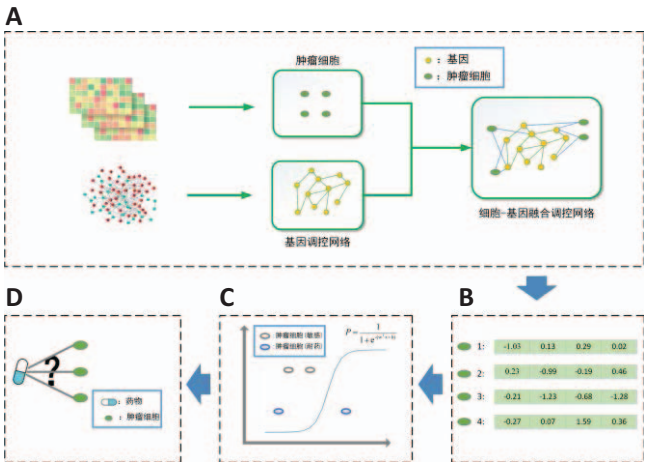

Fig. 1: The overall workflow of our KGE-LR method.(**A**) Step 1: Construction of cell-gene fusion regulatory network; (**B**) Step 2: Learning new tumor cell representations; (**C**)Step 3: Development of logistic regression drug response prediction model for each drug; (**D**) Step 4: Prediction of tumor cell response (sensitivity or resistance) with anti-cancer drugs.

2.1 Construction of cell-gene fusion regulatory network

Knowledge graph is a special data structure consisting of nodes and edges, where graph nodes represent entities and edges represent relationships between entities, using triples in the form of (head entity, relationship, tail entity). In the gene regulation knowledge graph, it is represented as (source gene, regulation relationship, destination gene). In this study, five gene regulatory repositories from PathMe [6], KEGG [7], Reactome [8],, BioGrid [9] and IntAct [10] were acquired and integrated into a gene regulatory network. The gene regulatory network contains a total of 8070 genes and ~21,000 regulatory relationships belong to five kinds of regulatory relationships: Associations, Increases, Decreases, Regulates, and hasComponent.

We connect the tumor cells as new nodes with the characteristic gene nodes, to get the cell-gene fusion regulatory network. There are various ways to select the feature genes, and the method of negative binomial distribution is chosen in this study. Specifically, for each gene, the expression value of the tumor cells is fitted to the negative binomial distribution, and the cells falling within the percentile threshold are selected and connected to the gene, and in particular, the cells at the end of the high expression value form an edge called up_expr, while the cells at the end of the low expression value form an edge called down_expr. The specific steps of the approach are as follows:

1）Set threshold α（0-50%）；

2）For a gene, fit the expression value of all cell samples to the negative binomial distribution；

3）Select the cells falling within the percentile threshold α and connect to the gene,；

4）Repeat steps 2) to 3) for each gene and finally obtain the cell-gene fusion regulatory network;

In this study, we use $G=(V,E)$， $V=(V_c \cap V_g)$ to denote the cell-gene fusion regulatory network， where $V_c=\{v_1,...,v_L\}$ denotes the L cell nodes of this network and $V_g=(v_{L+1},...,v_{L+N})$ denotes the N gene nodes of this network.

2.2 Learning new tumor cell representations

The main idea of KGE is to embed the entities and relationships in the knowledge graph into a continuous vector space. For the cell-gene fusion network constructed

above, we aim to embed the gene nodes, the tumor cell nodes, the regulatory relationships among genes and the expression relationships between tumor cells and genes into a continuous vector space, and then obtain the gene regulation fusion expression representation of tumor cell nodes. The new representation integrates the gene expression information and the information of multiple regulatory relationships of genes in the continuous vector space. In this study, we use the knowledge graph embedding method RotatE [11] to compute the representations of the tumor cell nodes as follows：

1）Extract the set of *T* positive triple in the cell-gene fusion regulatory network；

2）Initialize the tumor cells, gene nodes and edges with k=64-512 dimensional vectors with the following：

$$v_d^{Em0} = (w_1, w_2, ..., w_k), d = 1,2,...,Q$$
$$\mathrm{e}_a^{Em0} = (w_{k+1}, w_{k+2}, ..., w_{2k}), a = 1,2,...,H$$

where v_d^{Em0} denotes the initialized tumor cell node d vector, Q denotes the number of tumor cells , e_d^{Em0} denotes the initialized edge d vector, and H denotes the number of edges, and w $\in (-6/\sqrt{k}, 6/\sqrt{k})$;

3）Sample *M* negative triple by randomly replacing the trailing genes or trailing tumor cells in the triple set;

4）For each triple (h, r, o),define the distance function of RotatE as:

$$d_r(h, r, o) = \| h \circ r - o \| \tag{1}$$

where ∘ is the Hardman product, h denotes the vector representation of the head gene or head tumor cell , o denotes the vector representation of the tail gene or tail tumor cell , r denotes the vector representation of the positive regulatory relationship or the under- or over-expression of the gene on the tumor cell .

Calculate the total loss error Loss by scoring the *T* positive triple from step 1) and the *M* negative triple from step 3):

$$L = -\log \sigma(\gamma - \sum_{j=1}^{T} d_r(h_j, r_j, o_j)) - \sum_{i=1}^{M} \mathrm{g}(h_i', r_i', o_i') \log \sigma(d_r(h_i', r_i', o_i') - \gamma) \tag{2}$$

where h_j denotes the vector representation of the head gene or head tumor cell of the positive triad sample j, o_j denotes the vector representation of the tail gene or tail tumor cell of the positive triple sample j, r_j denotes the vector representation of the positive regulatory relationship or the under- or over-expression of the gene on the tumor cell of the positive triple sample *j*. h_i' denotes the vector representation of the head gene or head tumor cell of the negative triad sample i, o_i' denotes the vector representation of the tail gene or tail tumor cell of the negative triple sample i, r_i' denotes the vector representation of the negative regulatory relationship or the under- or over-expression of the gene on the tumor cell of the negative triple sample *i*. $\mathrm{g}(h_i', r_i', o_i')$ is the weight of negative triple samples i ,

$$g(h_i', r_i', o_i') = \frac{\exp \alpha d_r(h_i', r_i', o_i')}{\sum_{i=1}^{M} \exp \alpha d_r(h_i', r_i', o_i')} \tag{3}$$

where, α is a constant, represents the sampling rate;

5）Update the representation of the regulatory fusion features of all nodes and edges using the Adam optimization algorithm with the following expressions.；

$$v_d^{Em} = (c_1, c_2, ..., c_k), d = 1,2,...,Q$$
$$\mathrm{e}_a^{Em} = (c_{k+1}, c_{k+2}, ..., c_{2k}), \quad a = 1,2,...,H$$

where v_d^{Em} denotes the updated regulatory fusion feature representation of node v_d , e_j^{Em} denotes the updated regulatory fusion feature representation of edge e_j^{Em};

6）Repeat steps 3)-5) until the loss function converges, and finally obtain the optimal gene regulatory fusion expression characterization of the tumor cell *y*：

$$Embed_y = (\mathrm{z}_1, \mathrm{z}_2, ..., \mathrm{z}_k)$$

where z_i denotes the value on the *i*th dimension of the new representation.

2.3 Anticancer drug response classification prediction model

In order to predict the cellular response to a specific drug, in this study, we adopt logistic regression (LR) model based on the resulted regulation-fused expression profiles of each tumor cell.

We label the response of a tumor cell to a drug sensitive (1) or resistant (0) as a binary classification problem. Let *P* be the probability of a tumor cell being sensitive, then 1- *P* the probability of a tumor cell being resistant. By LR, *P* can be calculated as

$$P = \frac{1}{1 + \mathrm{e}^{-(W^T X + b)}} \tag{4}$$

where *X* denotes the gene regulation-fused expression representation of tumor cells, *W* denotes the coefficients of the logistic regression algorithm and b denotes bias.

We use the gradient descent algorithm to minimize the loss function as (5) to obtain W and b.

$$J(w,b) = -\frac{1}{m}\left(\sum_{i=1}^{m}\left(y_i \ln P(X_i) + (1-y_i)\ln(1-P(X_i))\right)\right) + \frac{1}{C}\sqrt{\sum_{j=1}^{n} w_j^2} \quad (5)$$

where m is the number of training sample, X_i is the gene fusion regulatory expression representation of training sample i, y_i is the label corresponding to training sample i, C={0.001,0.01,0.1,1,10} denotes the hyperparameter controlling the degree of $L2$ regularization, n denotes the dimension of the gene regulation-fused expression representation. For the drug response classification prediction model of each drug, we optimized the parameter of the model by five-fold cross-validation.

By the KGE-based LR model (KGE-LR), we predict a tumor sample to be sensitive if $P>\theta$=0.5 and to be resistant otherwise, i.e.

$$y = \begin{cases} 1 & P > \theta \\ 0 & P \le \theta \end{cases} \quad (6)$$

3 Experimental data

Preprocessed gene expression data [(RMA) normalized data] of 962 tumor cells were obtained from the GDSC [12] database. The response measurements of 189 drugs against 962 cells are expressed as log-normalized half-maximal inhibitory concentration (IC50). According to the work of Iorio et al [13]., we labeled drug-resistant or sensitive cells: less than an IC50 threshold as sensitive (1); greater than the threshold value as resistant (0). The final labeled data of drug-sensitive relationships between 189 drugs and 962 cells were obtained. For different drugs, the threshold was set according to https://www.cell.com/cell/fulltext. The transcriptome data contain 17419 genes. The gene regulatory network contains a total of 8070 genes. The two common gene sets have 6994 common genes.

By setting five different percentile thresholds (α), five cell-gene fusion networks were obtained, which in turn led to five sets of positive triples, denoted as G_1, G_2, G_3, G_4, G_5.

4 Results

4.1 Model evaluation strategy

To evaluate the predictive ability of the proposed method (KGE-LR), we used the five-fold cross-validation procedure, as shown in Fig. 1. In each validation, five-fold cross-validation was performed by further repartitioning the training data into training subsets (80%) and validation subsets (20%) 5 times to determine the optimal hyperparameter settings. Then the model was trained on the complete training data with the resulted hyperparameters, and was tested on the test set. The entire evaluation step was repeated five times for model evaluation.

The ROC curve was used to visualize the prediction performance of KGE-LR and the area under the ROC curve, AUC , was used to evaluate the prediction performance of the model. AUC is the area under the receiver operating characteristic curve, which is widely used to evaluate the performance of the model in machine learning binary classification tasks, and a larger AUC value indicates better performance.

Fig. 2: five-fold cross-validation flowchart

4.2 Performance evaluation

We obtained gene fusion regulatory expression representations at thresholds of 1%, 2%, 5%, 10%, and 20%, and calculated the AUCs on the drug TG101348, as shown in Fig. 3.

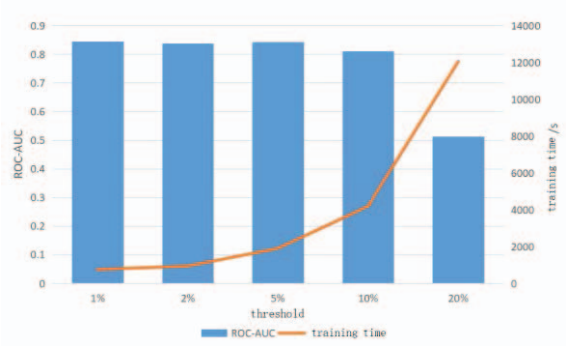

Fig. 3: The relationship between different threshold settings and training time and model performance

As the threshold value increases, the required training time is also longer and grows very fast. The reason is that the larger the threshold value is, the more edges and more triple are generated. The five thresholds led to five-fold cross-validation AUCs: 0.845, 0.837, 0.842, 0.811, and 0.512, respectively, suggesting that the feature vectors generated by the threshold 1%, gave the highest predictive ability. This should be because the cell nodes in the cell-gene fusion regulatory network cannot be effectively distinguished from each other when the number of cell lines and gene features are connected too densely.

4.3 Methods comparison

To gain a better understand of the performance of our approach, we compared it with two previous methods. The first is the method proposed by Stanfield et al.[14]. The method constructed a heterogeneous network to calculate the network profiles of cell lines and drugs, and then performed a random walk to predict the association between cell lines and drugs; the second is an SVM-based prediction method (SVMDRP) proposed by Dong et al [3].

We applied KGE-LR (1%), KGE-LR (5%), KGE-LR (20%), SVMDRP, and Stanfield's Method to the GDSC data for 189 drugs, respectively, and the results are shown in Fig. 4. From this figure, we can clearly see that the mean AUC values obtained by KGE-LR (1%) and KGE-LR (5%) were higher than those of SVMDRP and Stanfield's Method. Fig. 5 shows the corresponding ROC curves by KGE-LR (1%), KGE-LR (5%), KGE-LR (20%), SVMDRP, and Stanfield's on six of the 189 drugs.

In addition to AUC, we introduced five other accuracy evaluation metrics to further demonstrate the performance of KGE-LR compared to SVMDRP and Stanfield's Method: False positive rate(*FPR*), False negative rate(*FNR*), Accuracy(*ACC*),Positive predictive value (*PPV*),Matthews coefficient constant (*MCC*),and the results are shown in Table 1.The formula for calculating these five evaluation indicators is shown in (7)、(8)、(9)、(10) and (11).

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

$$FNR = \frac{FN}{FN + TP} \tag{8}$$

$$PPV = \frac{TP}{TP + FP} \tag{9}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

where *TP* , *FP*, *TN* and *FN* are the numbers of true positives, false positives, true negatives and false negatives, respectively.

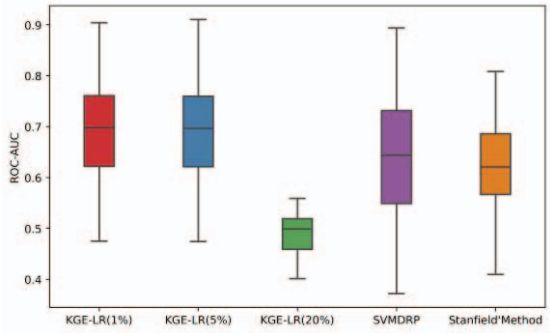

Fig. 4:AUC box plots on 189 drugs by different methods

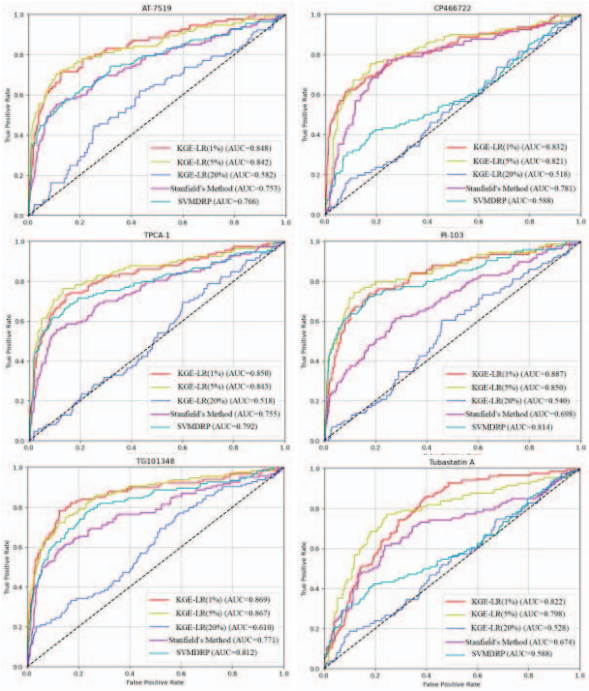

Fig. 5: ROC curves of KGE-LRs, Stanfield's and SVMDRP on six drugs

Table 1: Performance comparison(%,mean±SD) of KGE-LR and previous methods on 189 drugs.

	FPR	FNR	ACC	PPV	MCC	AUC
Stanfield's Method	10.7±8.3	70.3±14.7	87.1±4.2	65.7±4.4	35.5±12.9	62.3±8.1
SVMDRP	12.2±2.1	61.5±8.7	87.4±5.2	60.1±10.8	33.7±14.0	62.2±12.1
KGE-LR(1%)	10.5±4.7	72.5±4.3	88.5±6.1	72.9±5.3	39.4±10.1	70.3±9.3
KGE-LR(5%)	12.4±8.7	77.5±2.9	89.3±5.4	70.1±9.8	42.3±15.3	69.4±11.2
KGE-LR(20%)	53.5±4.1	51.9±6.3	79.5±3.1	49.6±5.7	1.2±9.7	46.8±5.5

5 Conclusion and discussion

We have proposed a new knowledge graph embedding-based drug response prediction method, KGE-LR, and evaluated it on the dataset GDSC. The method can address the complexity and high dimensionality of biological high-throughput data by constructing a fusion network and compute low-dimensional feature vectors for representing tumor cells. Experiments showed that the proposed method outperformed previous methods by a large margin.

Although the proposed KGE-LR in this study has achieved good results in the drug response prediction problem, there are still improvements to be done in future. For example, only gene regulation and expression information is used, and there are still other meaningful information, such as copy number variation information, drug-target interactions, etc. The information can be used to integrate into the network in future work to further improve the prediction performance of KGE-LR.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 61973295, 81872276), the Key Project of Scientific Research of Anhui Provincial Education Department (No. KJ2021A0633) and the Key Research and Development Program of Anhui Province (No. 201904a07020092).

Reference

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians 71(3) (2021) 209-249.

[2] A. George, R. Ladislav, S. Zhaleh, S. Petr, H.-K. Benjamin, G. Anna, Machine learning approaches to drug response prediction: challenges and recent progress, NPJ precision oncology 4(Suppl. 6) (2020).

[3] D. Zuoli, Z. Naiqian, L. Chun, W. Haiyun, F. Yun, W. Jun, Z. Xiaoqi, Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection, BMC cancer 15(1) (2015).

[4] Y. Chang, H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T.S. Kim, J. Jung, J.-M. Shin, Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature, Scientific Reports 8(1) (2018).

[5] S. Sotudian, I.C. Paschalidis, Machine Learning for Pharmacogenomics and Personalized Medicine: A Ranking Model for Drug Sensitivity Prediction, IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM PP(99) (2021) 1-1.

[6] D. Domingo-Fernández, S. Mubeen, J. Marín-Llaó, C.T. Hoyt, M. Hofmann-Apitius, PathMe: merging and exploring mechanistic pathway knowledge, BMC Bioinformatics 20(1) (2019).

[7] K. Minoru, F. Miho, T. Mao, S. Yoko, M. Kanae, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic acids research 45(D1) (2017).

[8] J. Bijay, M. Lisa, V. Guilherme, G. Chuqiao, L. Pascual, F. Antonio, S. Konstantinos, C. Justin, G. Marc, H. Robin, L. Fred, M. Bruce, M. Marija, R. Karen, S. Cristoffer, S. Veronica, S. Solomon, V. Thawfeek, W. Joel, W. Guanming, S. Lincoln, H. Henning, D.E. Peter, The reactome pathway knowledgebase, Nucleic acids research 48(D1) (2020).

[9] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, The BioGRID interaction database: 2019 update, Nucleic acids research 47(D1) (2019) D529-D541.

[10] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N.H. Campbell, G. Chavali, C. Chen, N. Del-Toro, The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases, Nucleic acids research 42(D1) (2014) D358-D363.

[11] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, arXiv preprint arXiv:1902.10197 (2019).

[12] W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J.A. Smith, I.R. Thompson, Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic acids research 41(D1) (2012) D955-D961.

[13] F. Iorio, T.A. Knijnenburg, D.J. Vis, G.R. Bignell, M.P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, A landscape of pharmacogenomic interactions in cancer, Cell 166(3) (2016) 740-754.

[14] Z. Stanfield, M. Coşkun, M. Koyutürk, Drug response prediction as a link prediction problem, Scientific reports 7(1) (2017) 1-13.