**DRSE: A Novel Method for Drug Repurposing Based on Drugs Side-Effects Similarities**

Authors: Sayed Mohammad Hassan Mir-ashrafia, Amir Lakizadeha, \*,

a Computer Engineering Department, University of Qom, Qom, Iran

\*Corresponding author

**Abstract**

The most significant drawback of the experimental methods in the field of drug development and discovery is that they are costly and time-consuming. Researches have indicated that designing a new drug from the first stage to its delivery to the consumer market lasts between 10 to 15 years. Moreover, this process costs between 0.8 to 1.5 billion dollars. Drug repurposing is seeking new indications as to the approved drugs plus, their main indications. Recently, some methods have been utilized regarding drug repurposing, and they were based on incorporating computational approaches. In this research, DRSE, a method for drug repurposing geared towards the integration of diverse and heterogeneous data sources is illustrated. DRSE is able to predict the drug-disease associations by a matrix factorization algorithm using the side-effect features of drugs. Having assessed 763 different kinds of drugs and 681 disease, confirmed that the proposed method for integration of multiple data sources considering drug side effects similarities and other similarities is able to raise the accuracy of drug repurposing task, the AUC and AUPR criteria have also improved this by 1.13% and 14.23% respectively, compared to the state-of-the-art methods.

**Keywords**: Drug repurposing, Link prediction, Matrix factorization, Drug side-effect

**1. Introduction**

There are a number of experimental approaches regarding drug development and discovery, however one of the most significant disadvantages of them is that not only do they cost a fortune, but also they are time-consuming. According to the current studies, the process of designing a new drug from the first stage to its delivery to the consumer market is taken approximately between 10 to 15 years. In addition, the process costs about 0.8 to 1.5 billion dollars. However, despite all the investments and advances in the last 20 years, the number of new drugs approved by the Food and Drug Administration (FDA) have not been significant. Every year, nearly 90% of evaluated medications by the FDA are rejected and prevented to be used for therapeutic purposes [1].

Drug repurposing (also called drug repositioning) refers to the process of identifying new indications of approved drugs plus their original indications. This strategy yields significant speed advantages over discovering a new drug, when treating a particular disease. In some cases, the spread rate of an acute disease such as Covid-19 is very high, and designing a new drug doesn’t seem to be prompt, as a result it is not reliable. On the other hand, spending billions of dollars to pharmaceutical companies in the development of novel medicines for a rare but dangerous disease is not economically reasonable, due to the fact that the scarcity of the disease reduces the number of affected patients. Therefore, researchers have to restrict themselves to evaluating the effectiveness of drugs already known as the treatment of these disease.

Predicting a new application for already known drugs with computational methods is in a way that a candidate drug-disease interaction will be evaluated, then other validation process fulfill through experiments. Two categories can be mentioned as these assessing computational experiments, the first one works according to a drug and a disease relation based on their common genes or protein complexes which is shared between drug and disease. The second one beside drugs, diseases, and targets interactions benefit from similarities between them. Few methods have predicted drug new application through a logistic regression model, a statistical model, sparse subspace learning, or similarity constraint matrix factorization.

Drug repositioning through integration of prior knowledge and projections of drugs and diseases. منبع

Among computational methods to investigate drug repurposing, supervised learning, because drugs with similar targets may have some similarities, can be considered a feasible method. Supervised learning have leaded to a suitable classification however, the rate of accuracy and correctness of classification depends on training data sets. Thus negative and positive samples are important. To identify drug-target interaction there are plenty of problems which should be considered first positive samples are rare, second true negative samples are mostly unavailable. Therefore, negative samples are constructed randomly in such methods [5].

L. Peng, W. Zhu, B. Liao, Y. Duan, M. Chen, Y. Chen, J. Yang, Screening drug-target interactions with positive-unlabeled learning, Scientific reports, 7 (2017) 1-17. منبع

این دو پاراگراف برای منطق intoduction هست

Sildenafil, with the "Viagra" commercial name, is a clear example of drug repurposing which brought a successful result. Sildenafil was first prescribed in the late 1980s as the treatment of chest pain. The drug was found to be ineffective during clinical trials; consequently, its development was suppressed. However, the patients reported unusual side effects, including prolonged erections. Researchers then re-purposed the drug to treat erectile dysfunction. One way of recovering from this problem of finding new therapeutic applications for drugs could be modeled using the network-based models and recommender systems.

One of the essential state-of-the-art methods concerning the recommender systems is the matrix factorization. In this method, a latent feature vector is extracted for each user and each item. The length of latent feature vectors is equal to the number of latent features of the drug or disease as well as the obtained values from the score matrix and the latent disease features depend on the content of the user and items.

In the initial stage of the process of Matrix factorization methods, the drugs and diseases are mapped into the latent feature spaces. The next step is that the missing values in the score matrix are estimated by the internal multiplication of the corresponding latent feature vectors inferred from the known values [2].

The researches regarding this field can be examined from the two perspectives: the type of data and the computational approach. Recently, with the rapid development of computational techniques in biology, various drug databases such as ChemBank, OMIM, Pubmed, DrugBank, and extensive genomic databases such as GenBank and MIPS were created. These databases contain the data related to one or more categories such as drugs, proteins, diseases, and genes [3].

Most of the bioactive molecules exert their activities by interacting with the proteins. The interactions can be analyzed using the computer modeling of the target protein and the drug 3D structures. This procedure is called molecular docking and is used in drug discovery strategies. In 2019, Arochi et al. have used molecular docking to re-purpose a known drug for the Pim-1 protein target, which is known to cause prostate cancer and acute bone marrow leukemia [4].

The biological data are incomplete and restricted in some cases whereas, they are capable of expressing complementary concepts. Therefore, using a combination of different data sources may lead to more accurate and precise predictions. The methods have discussed so far were based on specific biological concepts. In the following, a brief review of various techniques with different computational approaches is presented.

The support vector machine, minimum least squares, and logistic regression algorithms have been the leading machine learning methods used for drug repurposing. In 2017, Peng et al. have developed a method called NDTISE, which deeply examined unlabeled samples of drug-disease pairs (which have not been reported to be related) [5]. These samples were broken into two categories, and the uncertain samples were separated from the samples which can be defined as unrelated ones with certainty. In the next step, only negative samples were given to a backup classifier which performed the decision task.

In 2020, Kodia et al. have proposed a method based on non-negative matrix factorization that applies a non-negative constraint on the factorized matrices during multiplication and update operations [6]. The shortest paths between nodes at each layer of the network were calculated separately and used in the initialization of feature matrices and similarity matrices.

Recently, Yu et al. have designed a comprehensive package called BioNEV, which can execute two categories of methods: the link prediction and the multi-class classification in biological networks [7]. In the link prediction part, three networks, namely drug-disease, protein interactions, drug-drug interactions, have been implemented. In the classification part, the prediction of protein function and the classification of various medical terms methods have been implemented.

In an effort topological and structural features of nodes have been used to calculate similarity measures between the network nodes. The next step is that similarity matrices of nodes are given as an input to a deep learning method to extract secondary complex features which helps to predict the potential interactions. In the deep learning method, the characteristics of each drug, disease, gene, protein, etc. can be given as a binary vector in raw form without providing any additional information to the model. DeepDR is a drug repurposing model proposed by Zheng et al. in 2019 [8]. In DeepDR, a random walk is used to extract feature vectors representing the network properties. Then, they used an auto-encoder to integrate the data and dimension reduction. DeepDR finally used cVAE to predict new links and achieved AUC = 0.90 and AUPR = 0.92.

Xuan et al. منبع همونی که فرستادی have developed a method called DisDrugPred based on non-negative matrix factorization which captures both intra and inter-relationships among drugs and diseases. However, it is not benefit from a low dimensional feature vector which is also informative and geometric proximity of the mapped vectors.

Here, we proposed DRSE, a Drug Repurposing method based on drug Side Effects and some similarities information. It uses a random walk with a restart to merge drug and disease features including drug side-effects data. Also, it predicts drug-disease interactions by applying a matrix factorization method. The results confirmed that the proposed method for integration of multiple data sources considering multi similarities is able to improve the accuracy of drug repurposing task. The proposed method in contrast to the Xuan et al. همونی که فرستادی منبع can learn through informative low dimensional vectors which obtained by integration of diverse and noisy biological data. It also makes the prediction of new DTIs possible by benefiting from geometric proximity of the mapped vectors. قسمت بین سبزها همونی هست که خواستید بنویسم

The rest of the article is broken into three sections. Section 2 expounds the details of the proposed method. Section 3 represents the evaluation results of the proposed method and its comparison to the previously proposed methods. Moreover, a case study is performed on some of the predicted associations to investigate the model's performance further. Finally, Section 4 summarizes the main points of this research and the conclusions are drawn in this section, and some suggestions for future works are provided.

**2 Materials and Method**

**2.1 Data sources**

Some of the data used in this study have been downloaded from Liang et al. study [9]. They obtained and cleansed the required data from common databases such as PubChem, InterPro, UniProt, and DincRNA. This dataset consists of three types of feature matrices for drugs, an integrated similarity matrix, and a drug-disease association matrix, whose features are as follows:

* The chemical structures of drugs have been obtained from the PubChem database [10].
* Ontology characteristics of genes related to the target proteins were collected from the InterPro database [11].
* The 3D and spherical structures of the target proteins were downloaded from the UniProt database [12].
* The related genes for each disease were attained from the DincRNA database [13].
* The semantic features of diseases were extracted by analyzing the articles on the web related to each disease.
* Drug-disease association matrix downloaded from Liang et al. study [9]

The datasets which are used in this study and defined as above can be reviewed in table (1) which can illustrate the number of each data in the related group.

Table (1) the data and the number of each one, which have been used in this study

شماره جدولها نیاز به آپدیت داره این دوتا رو که رفرنس دادم و نوشتم هم داخل پرانتز هست هم قرمز هایلایت کردم

|  |  |  |  |
| --- | --- | --- | --- |
|  | Number Of Nodes #1 | Number Of Nodes #2 | Number Of Edges |
| Drug - Chemical structure | 763 | 623 | 96202 |
| Drug – Target protein domains | 763 | 1425 | 7593 |
| Drug – Gene Ontology terms of drug target | 763 | 4447 | 55918 |
| Drug – Side effects | 763 | 3046 | 9730 |
| Disease - Gene | 681 | 386 | - |
| Drug - Disease | 763 | 681 | 3051 |

In order to improve the performance of the method, another type of drug feature based on drug side-effects has also been calculated. Besides, the side effects of all drugs in the SIDER database were downloaded and converted into a usable format [14]. Besides disease related matrices which are mentioned were not available directly. Therefore, the aggregated matrix which is mentioned in Liang et al. study [9] have been used.

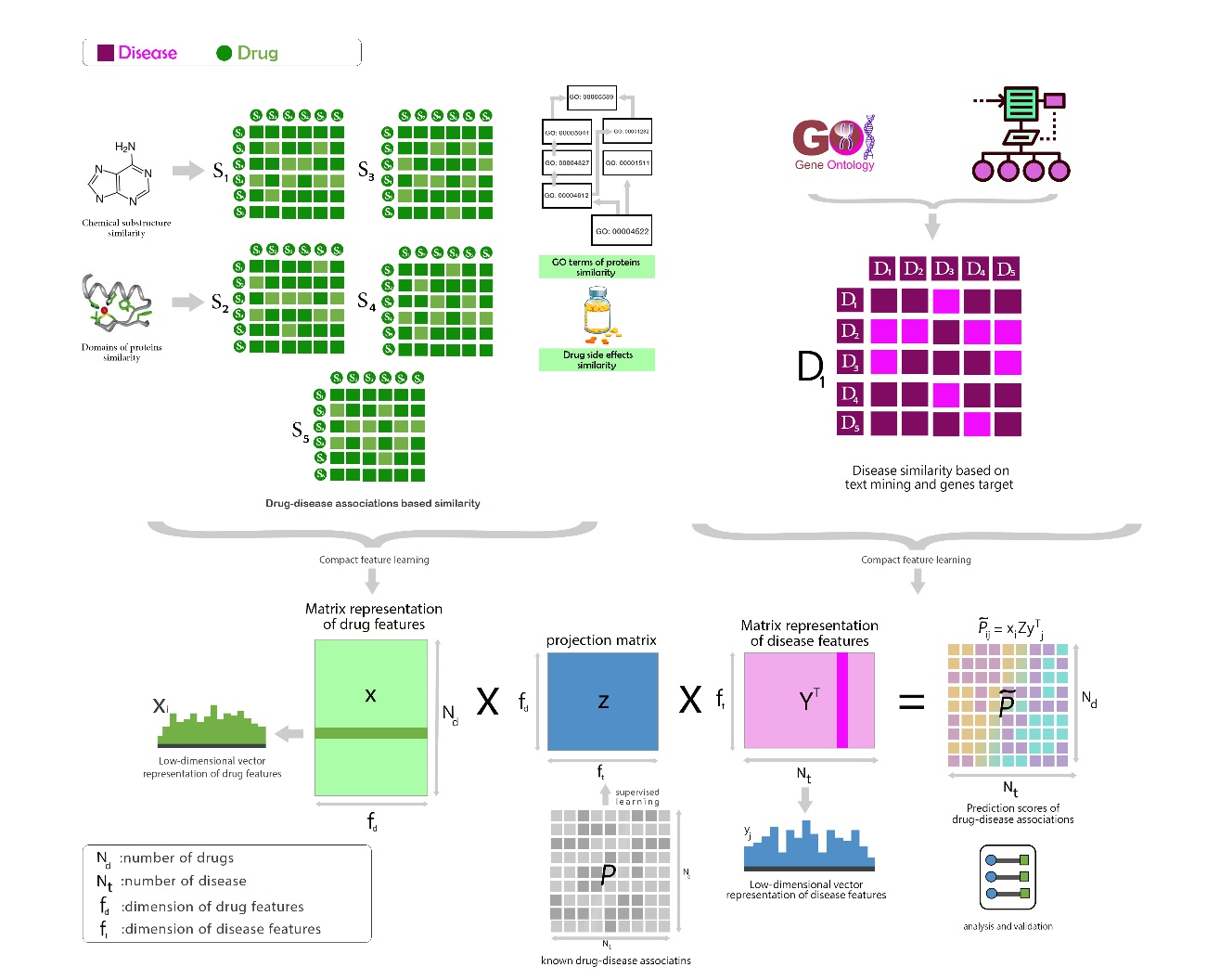
منبع چک شود لیانگ و همکاران

**2.2 Methods**

The method proposed (called DRSE, a Drug Repurposing method based on drug Side Effects information) includes the Random walk with restart technique to merge drug and disease features, diffusion component analysis (DCA) to extract more practical features, and the matrix factorization method to estimate final prediction. The steps of DRSE are summarized as follows:

* Compiling the different types of features for drugs and diseases
* Calculating the similarity matrix from the raw features using the Jaccard index function
* Applying a random walk with restart probability on each similarity matrix
* Integrating the resulted matrices and mapping them to a low-dimensional space, this action must be taken once either for the drug similarities or disease similarities
* Using the matrix factorization to predict a score for the possibility of drug-disease association

An overview of this method is illustrated in Figure 1.



**Figure 1.** The overview of the proposed method

**2.2.1 Computing similarity matrices**

The mentioned features were converted to the similarity matrices using the Jaccard index function with the following formula:

|  |  |
| --- | --- |
| (1) |  |

The Jaccard similarity coefficient for two objects with binary attributes is given in formula 1. It is important to note that there are two feature vectors *A* and *B,* each of which contains *n* cells with values 0 or 1. The symbols in the above equation are defined as follows:

* is the number of features that are 1 in both vectors *A* and *B*.
* is the number of features that are 0 in *A* and 1 in *B*.
* is the number of features that are equal to 1 in *A* and 0 in *B*.

<https://link.springer.com/article/10.1007/s11634-011-0090-y> منبع شاخص جاکارد

**2.2.2 Feature engineering**

Feature engineering denotes the process of computing features with lower dimensions than the original dimensions. In this research, the diffusion component analysis (DCA) has been taken into account with the aim of learning and creating a low-dimensional vector representations for drugs and disease similarity matrices [15]. This recently developed algorithm is a combination of network propagation method (random walk with restart) and dimensional reduction method, which is able to extract the topological properties of a network [16].

Random walk with restart (RWR) is an algorithm for network diffusion which has widely been used to analyze biological network data [17-21]. Unlike the conventional random walk method, RWR is using a predefined probability for restarting the process and returning back to the initial node in each iteration. Therefore, RWR can take local and global topological connectivity patterns into account to utilize direct or indirect links in the network.   
RWR have been applied to all five drug similarity matrices. The results have given the topological information of the network. Moreover, the dimensions of each matrix were similar to the dimension of the original matrix. Obtaining an informative but low-dimensional vector which represents nodes in the network derived from RWR (random walk with restart) and DCA (diffusion component analysis). This is called compact feature learning which produces a feature vector.

**2.2.3 Dimension reduction framework**

Diffusion probabilities obtained by the RWR process may not be completely accurate due to the relatively low quality and the high dimension of the biological data which is considered as a pitfall. As a matter of fact, the lack of information or misinformation in the network has a considerable impact on the results of RWR [22]. In addition, using high-dimensional data as the input features it is not appropriate. To overcome this problem, the proposed method used a dimensional reduction technique, called DCA, to reduce the dimensions of the feature space and extract the fundamental topological properties of the diffusion probabilities. For instance, the probability assigned to the node in the diffusion state of the node is computed as follows.

|  |  |
| --- | --- |
| (2) |  |

Where, , represents the context feature and denotes the node feature for theth node that both specify the topological properties of the network. The DCA considered a set of observed diffusion states as input. This procedure seeks to reduce the input dimension to a lower dimension using relative entropy (Kulback-Leibler divergence (KLD)); thus, the input and output distributions have the least difference. This operation was performed for all and all by minimizing the following cost function which computes the KLD function between the two distributions:

|  |  |
| --- | --- |
| (3) |  |

The DCA framework utilizes a standard quasi-Newton method called L-BFGS [23] in order to optimize this function. It is worth noting that the formulas have used to reduce the dimension, and both the integration selected and the final prediction made in this section were supplied by Luo et al [24].

**2.2.4 Integrating matrix information of different similarities**

**T**he above dimensional reduction framework has been extended to integrate several drug similarity matrices to consider all similarity matrices in the dimensional reduction process. If a drug similarity matrix were available, RWR would be performed on each network separately to incorporate different networks with various information (similarity matrices based on different features). Then the network diffusion state for each node in the network would be obtained through the following formula:

|  |  |
| --- | --- |
| (4) |  |

Finally, DCA optimizes the following objective function:

|  |  |
| --- | --- |
| (5) |  |

In this study, the restart probability for the RWR was set to 0.5. Moreover, according to the results of experiments in previous researches [31], the dimension of latent features for drug was specified to and for disease was set to, which were equal to 10% and 15% of the original dimensions. Besides, the maximum number of repetitions of the proposed method was set to 20.

**2.2.5 Drug-disease association prediction**

The low-dimensional vectors of drug and disease features derived from the compact feature engineering process have been used to predict new drug-disease associations, so that the compact feature vectors for each drug were geometrically close to its original associated diseases.  
Explaining mathematically, the expression of has been employed to represent the matrix of drug feature (the th of the matrix represents the features of the drug ) and used to represent the feature matrix of disease (the th of the matrix represents the features of the disease ). It should be noted that and indicate the number of drugs and the number of diseases, respectively. If denotes a matrix of drug-disease associations; So that would indicate the known association between drug and disease , while would illustrate the unknown link between them. With the aim of learning the prediction matrix, a bidirectional function was defined to predict the unknown links in (i.e., zero values in the association matrix). More precisely, the bidirectional function is defined as follows:

|  |  |
| --- | --- |
| (6) |  |

indicates drug-disease association matrix, are the results of compact feature engineering (RWR and DCA steps), and is the projection matrix that must be learned. Afterward, the following formula has been used to calculate the probability of an association between the pair of drug and disease :

|  |  |
| --- | --- |
| (7) |  |

The higher value of suggests the more likely association between drug and disease . The projection matrix has the dimension of as well. Typically there is a significant correlation between those drugs or diseases feature vectors that are geometrically close in space, which can consequently reduce the number of effective parameters needed for model drug-disease associations. For this reason, a low-rank constraint to has been applied to learn only a small number of hidden factors calculated through low-rank decomposition where . Not only does this low-rank constraint overcome the problem of overfitting but also improves computationally the optimization task [25].  
The optimization problem with low-rank constraint in the projection matrix is an NP-hard problem. A standard relaxation of the low-rank constraint is to minimize the trace norm (i.e., sum of singular values) of the matrix, which is equivalent to minimizing the Frobenius norm . Therefore, factorization of Z to G and H can be done by solving the following optimization problem:

|  |  |
| --- | --- |
| (8) |  |

Here, the parameter is the regularization parameter that controls the balance between minimizing squares of error in a known association of pair and the Frobenius norm. The optimization problem has been tackled by alternation minimization [26].

**3 Results and discussion**

**3.1 Evaluation measures**

In this study, the stratified five-fold cross-validation has been utilized and the standard classification criteria have been considered to evaluate the model. Stratified cross-validation was modified cross-validation and a more appropriate evaluation method has been employed for the unbalanced data. This technique prevents the over-fitting problem and helps impartially evaluating the model. In five-fold cross-validation, the data were divided into five equal sets. During the first phase, the proposed method was trained with four sets of data and tested with another set and after this step has been done, the evaluation criteria were calculated for the test set. This process had been repeated until all sections were considered as the test data once. When these steps have been completed, the average calculated criteria in all test sets were considered the overall performance of the model. Since classification is the main problem in this research that must be tackled, standard classification criteria should be used to be evaluated. A confusion matrix was used to calculate the classification criteria. This table introduced four measures TP, TN, FP, and FN.

It is also essential to note that when the interaction between a drug-disease pair is zero, there is no evidence of an association between them. Thus, they may be associated and it should be considered false negative. Therefore, in this type of problem, the number of true and false positives could not be accurately counted. The training process always required data from both labels zero and one data types, so some pairs which were considered as zero by the model in the training process may not be zero. Because the values of criteria such as accuracy, precision, and sensitivity depend on the threshold value, two more comprehensive criteria (AUC and AUPR) were considered in evaluating the models. AUC is the receiver operating characteristic curve created by plotting the true positive rate and the false-positive rate in different thresholds. In this study has been illustrated in figure 2.

|  |  |
| --- | --- |
| (9) |  |

|  |  |
| --- | --- |
| (10) |  |

AUPR is the area under the plot created based on precision and sensitivity which is shown in figure 3 for current study. AUC and AUPR evaluate the performance of the model independently of the threshold value. In cases where the samples are not balanced, and the association matrix has more zero, the AUPR criterion is a fairer metric for model evaluation.

**3.2 Results of the proposed method**

The proposed method was implemented in two ways once on each of similarity matrix's separately to investigate their importance which is shown in table 1. In the next step was also applied to the integrated similarity matrix. As shown in Table 1, the model applied to the matrix of drug-disease similarities and gene ontology has outperformed others. The combination of different features has resulted in the best results, which confirms the efficiency of DCA compaction and similarity integration by RWR. Using the integrated feature and similarities based on drug-disease associations has led to the best AUC and AUPR values.

**Table 1.** Performance of the proposed method on each similarity matrix

|  |  |  |
| --- | --- | --- |
| **Features** | **AUC** | **AUPR** |
| d1) Chemical structure-based similarity | 83.60 | 87.28 |
| d2) Protein target-based similarity | 84.42 | 88.45 |
| d3) Gene ontology-based similarity | 85.30 | 89.13 |
| d4) Side-effect based similarity | 80.58 | 85.42 |
| d5) Similarity based on drug-disease association matrix | 92.07 | 93.47 |
| d1+d2+d3+d5 | 92.87 | 94.53 |
| Integration of all above similarity matrices | 93.23 | 94.83 |

A comparison between the performance of the proposed method and the results of the previous methods have been made which is highlighted in Table 2. The results of اسم نویسنده یا نویسندگان مقاله که این اعداد رو آوردی ازش are reported in the three first rows in this table. Needless to say that the data set which are used are standard the same. A more detailed look at the results, reveals the fact, that the proposed method far outweigh the previous methods and can predict the existing drug-disease relationships more appropriately. From these comparisons, it is clear that the high values of AUC and AUPR in the proposed method indicate that the proposed method seems to be much successful in drug repurposing than previous methods. The proposed method has improved the AUC and AUPR criteria by 1.13% and 14.23%, respectively, compared to the previous methods. Since the AUPR criterion is a more appropriate criterion for evaluating the problem, the proposed method has the potential to outperform all the earlier methods. The AUC is sensitive to a large number of zeros in the association matrix, and the high number of zeros can lead to an insignificant increase in AUC. Overall, this method is the one that obtained the most robust results, for instance improving the AUPR criterion by the proposed method is a sign of the excellent performance of the model in returning known relations, which leads to the confidence of the model in predicting unknown links. This improvement over other methods may be due to the use of different features and their combinations. As outlined in Table 2, increasing the various features has led to improve the results. Moreover, using Random walk with restart and Matrix factorization with an efficient approach has led to combining different features conveniently. This combination method could consist of the topological features in the final prediction due to the use of the latent space. All these properties improved the AUPR by 14.23 percent which is an important finding in the understanding of this field.

**Table 2.** Comparison of the proposed method with previous studies

|  |  |  |
| --- | --- | --- |
| **Method** | **AUC** | **AUPR** |
| TL-HGBI 2014 [27] | 72.7 | 3.0 |
| MBiRW 2016 [28] | 85.1 | 4.4 |
| SCMFDD 2018 [29] | 63.8 | 6 |
| DisDrugPred 2019 [30] | 92.0 | 24.3 |
| BLM 2009 [31] | 74.75 | 76.89 |
| Graph Embedding Matrix Factorization-GraRep 2015 [32] | 75.7 | 69.3 |
| Graph Embedding Random walk based-struc2vec 2017 [33] | 84.4 | 80.6 |
| Graph Embedding Neural network based-SDNE 2016 [34] | 77.4 | 75.2 |
| **DRSE**(proposed) | **93.23** | **94.83** |

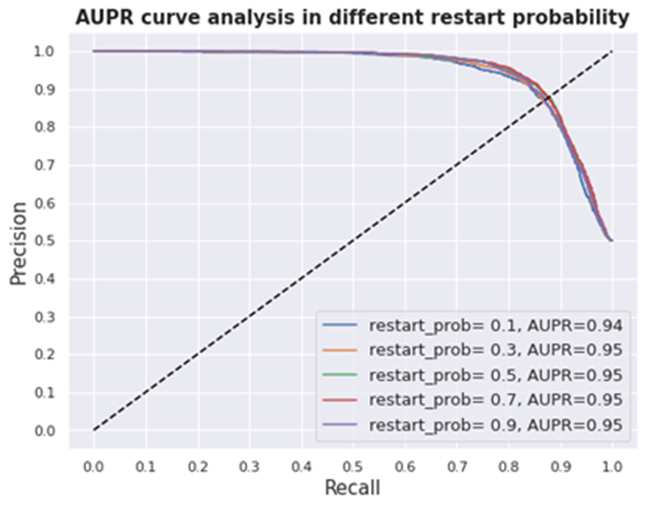
With the aim of investigating the impact of the distinct values ​​of the hyper parameters on the probability of returning to the initial node in the random walk and the latent space dimension in factorization, these two hyper parameters have been employed for the values of [0.9, 0.7, 0.5, 0.3, 0.1] and [90, 75, 50, 25] respectively, whose AUPR values are illustrated in Figures 2 and Figures 3.

In the performed scenario as it is illustrated in table 3, the similarities were first aggregated for the different values ​​of return to the initial node whose aim is integration, and the best result was obtained by setting the probability of return equal to 0.5. Afterward, the return probability value was kept constant at 0.5 and then the output of the work was measured per different values ​​of the hyper parameter dimension of the latent space. These settings were obtained through experimenting different scenarios which led us to set them as it is mentioned in table 3. یه روش می تونه این باشه که استاد خوشش نیومده.

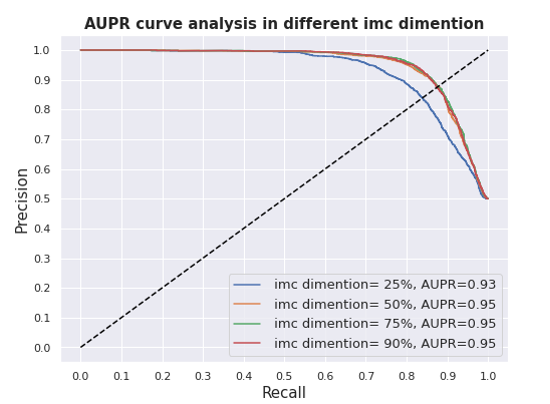
To set the mentioned hyper parameters different scenarios have been implemented and the best values are 0.5 and 50, for the probability of returning to the initial node and the latent space dimension in factorization, respectively. اگر این مدل رو بنویسی دیگه جدول می شه حذف بشه

**Table 3.** Hyperparameters Imc and restart probe setting

|  |  |  |
| --- | --- | --- |
| Imc = 50 | | |
| Restart prob | auc | aupr |
| 0.1 | 92.43 | 94.21 |
| 0.3 | 92.86 | 94.51 |
| 0.5 | 93.23 | 94.83 |
| 0.7 | 93.15 | 94.78 |
| 0.9 | 92.99 | 94.69 |
| restart prob=0.5 | | |
| Imc dimension | auc | aupr |
| 25 | 91.62 | 93.55 |
| 50 | 93.23 | 94.83 |
| 75 | 93.02 | 94.50 |
| 90 | 92.70 | 94.36 |



**Figure 2.** AUPR curve for different hyper parameter values the probability of returning to the initial node



**Figure 3.** AUPR curve for different hyper parameter values of latent space dimension in factorization

**3.3 Case Stud**y

It is notable that, since the zeros in the drug-disease association matrix denote the unknown associations (and not a definite absence of relation), these zeros may be converted to one in the future. Therefore, some false positives predicted by the proposed method have been examined. Table 4 highlights the evidence of the case studies.

**Table 4.** Evidence of case studies on false positive predictions

|  |  |  |
| --- | --- | --- |
| **Drug name** | **Disease name** | **Reference** |
| Cefpodoxime | Staphylococcus infection | [35] |
| Lovastatin | Hyperlipidemia Type II | [36] |
| Trifluoperazine | Schizophrenia and Paranoia | [37] |
| Cefdinir | Skin infection | [38] |
| Nabilone | Postoperative nausea and vomiting | [39] |
| Piperacillin | Klebsiella infection | [40] |
| Procainamide | Ventricular fibrillation | [41] |
| Flecainide | Ventricular fibrillation | [42] |
| Ertapenem | Staphylococcus | [43] |

These cases were the pairs that were zero in the original association matrix, but DRSE has predicted them as one. For instance, the association of Trifluoperazine drug with Schizophrenia disease has been predicted despite this link was zero in the original association matrix. Trifluoperazine is a member of the Phenothiazines group that its main function is to inhibit the dopamine receptors in brain cells. It controls the effects of dopamine stimulation by obstructing dopamine receptors in the central nervous system. It also blocks alpha receptors and neutralizes the activity of histamine and serotonin [37]. Another case is Nabilone drug for the treatment of Postoperative nausea and vomiting. In recent studies, Nabilone has been identified as a supplement to prevent Postoperative nausea and vomiting. It has been revealed previously to be clinically effective in treating chemotherapy-related nausea and vomiting [39]. Therefore, these examples showed how the proposed method can accurately predict the usage of a drug with an unknown target in another disease.

**4 summary**

Predicting the drug-disease associations is considered as an integral part of medical science because the drugs are playing a prominent role in the treatment of other diseases and their original indications. Drug repurposing has this ability to reduce the time and cost of drug discovery and development phases. Due to the importance of these problems, many studies have used bioinformatics and computational biology to predict drug-disease associations by analyzing various types of data, indeed the aim in this research is to present more sophisticated methods to overcome these problems. Based on the hypothesis that states similar drugs are able to have similar functions, and also similar diseases are able to have similar treatments, the similarity of drugs and diseases could be used for drug repurposing. In this study, DRSE, a method for drug-disease associations prediction using various drug features such as chemical structure, side-effects, target protein, and an integrated disease similarity matrix has been proposed. The proposed method in this research uses a random walk with restart and feature compacting technique along with matrix factorization for drug-disease association prediction.

The method performance was evaluated by stratified five-fold cross-validation with AUC and AUPR criteria. The method obtained AUC and AUPR of 93.23% and 94.83%, respectively. The efficiency of each similarity matrix in the improvement of model performance was also examined. Among the similarities, structural similarity and ontological similarity obtained better results than other features. In summary, this paper argued that the proposed method has a number of advantages over the previous ones. Studies have also revealed that using a combination of different features can boost the accuracy and prove the efficiency of the method in data compacting and integrating.

For further analysis, case studies have been performed on false-positive predictions. Evaluations demonstrated that the proposed method is able to predict unknown interactions by considering side effects and other features.

In the end, a number of recommendations for future research are given below:

* Using other types of features
* Combining drug features by other integration methods
* Checking the entropy of similarity matrices before considering them and then excluding matrices that may have less information
* The severity of the side-effects of drugs has not been recorded anywhere yet. If this data is available, it can be considered as a factor to reduce the risk of drug repurposing

**References**

[1] H. Xue, J. Li, H. Xie, Y. Wang, Review of drug repositioning approaches and resources, International journal of biological sciences, 14 (2018) 1232.

[2] M.H. Aghdam, M. Analoui, P. Kabiri, A novel non-negative matrix factorization method for recommender systems, Applied Mathematics & Information Sciences, 9 (2015) 2721.

[3] G. Jin, S.T. Wong, Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines, Drug discovery today, 19 (2014) 637-644.

[4] H. Arrouchi, W. Lakhlili, A. Ibrahimi, Re-positioning of known drugs for Pim-1 kinase target using molecular docking analysis, Bioinformation, 15 (2019) 116.

[5] L. Peng, W. Zhu, B. Liao, Y. Duan, M. Chen, Y. Chen, J. Yang, Screening drug-target interactions with positive-unlabeled learning, Scientific reports, 7 (2017) 1-17.

[6] G. Ceddia, P. Pinoli, S. Ceri, M. Masseroli, Matrix Factorization-based Technique for Drug Repurposing Predictions, IEEE Journal of Biomedical and Health Informatics, (2020).

[7] X. Yue, Z. Wang, J. Huang, S. Parthasarathy, S. Moosavinasab, Y. Huang, S.M. Lin, W. Zhang, P. Zhang, H. Sun, Graph embedding on biomedical networks: methods, applications and evaluations, Bioinformatics, 36 (2020) 1241-1251.

[8] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, F. Cheng, deepDR: a network-based deep learning approach to in silico drug repositioning, Bioinformatics, 35 (2019) 5191-5198.

[9] X. Liang, P. Zhang, L. Yan, Y. Fu, F. Peng, L. Qu, M. Shao, Y. Chen, Z. Chen, LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning, Bioinformatics, 33 (2017) 1187-1196.

[10] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, S.H. Bryant, PubChem: a public information system for analyzing bioactivities of small molecules, Nucleic acids research, 37 (2009) W623-W633.

[11] A. Mitchell, H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, The InterPro protein families database: the classification resource after 15 years, Nucleic acids research, 43 (2015) D213-D221.

[12] U. Consortium, The universal protein resource (UniProt) in 2010, Nucleic acids research, 38 (2010) D142-D148.

[13] L. Cheng, Y. Hu, J. Sun, M. Zhou, Q. Jiang, DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function, Bioinformatics, 34 (2018) 1953-1956.

[14] M. Kuhn, I. Letunic, L.J. Jensen, P. Bork, The SIDER database of drugs and side effects, Nucleic acids research, 44 (2016) D1075-D1079.

[15] H. Cho, B. Berger, J. Peng, Diffusion component analysis: unraveling functional topology in biological networks, International Conference on Research in Computational Molecular Biology, Springer, 2015, pp. 62-64.

[16] H. Cho, B. Berger, J. Peng, Compact integration of multi-network topology for functional analysis of genes, Cell systems, 3 (2016) 540-548. e545.

[17] M. Cao, C.M. Pietras, X. Feng, K.J. Doroschak, T. Schaffner, J. Park, H. Zhang, L.J. Cowen, B.J. Hescott, New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence, Bioinformatics, 30 (2014) i219-i227.

[18] S. Köhler, S. Bauer, D. Horn, P.N. Robinson, Walking the interactome for prioritization of candidate disease genes, The American Journal of Human Genetics, 82 (2008) 949-958.

[19] S. Navlakha, C. Kingsford, The power of protein interaction networks for associating genes with diseases, Bioinformatics, 26 (2010) 1057-1063.

[20] C.-S. Liao, K. Lu, M. Baym, R. Singh, B. Berger, IsoRankN: spectral methods for global alignment of multiple protein networks, Bioinformatics, 25 (2009) i253-i258.

[21] X. Chen, M.-X. Liu, G.-Y. Yan, Drug–target interaction prediction by random walk on the heterogeneous network, Molecular BioSystems, 8 (2012) 1970-1978.

[22] M. Kim, J. Leskovec, The network completion problem: Inferring missing nodes and edges in networks, Proceedings of the 2011 SIAM International Conference on Data Mining, SIAM, 2011, pp. 47-58.

[23] C. Zhu, R.H. Byrd, P. Lu, J. Nocedal, Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization, ACM Transactions on Mathematical Software (TOMS), 23 (1997) 550-560.

[24] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, J. Zeng, A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, Nature communications, 8 (2017) 1-13.

[25] H.-F. Yu, P. Jain, P. Kar, I. Dhillon, Large-scale multi-label learning with missing labels, International conference on machine learning, 2014, pp. 593-601.

[26] N. Natarajan, I.S. Dhillon, Inductive matrix completion for predicting gene–disease associations, Bioinformatics, 30 (2014) i60-i68.

[27] W. Wang, S. Yang, X. Zhang, J. Li, Drug repositioning by integrating target information through a heterogeneous network model, Bioinformatics, 30 (2014) 2923-2930.

[28] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F.-X. Wu, Y. Pan, Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm, Bioinformatics, 32 (2016) 2664-2671.

[29] W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang, F. Liu, Predicting drug-disease associations by using similarity constrained matrix factorization, BMC bioinformatics, 19 (2018) 1-12.

[30] P. Xuan, Y. Cao, T. Zhang, X. Wang, S. Pan, T. Shen, Drug repositioning through integration of prior knowledge and projections of drugs and diseases, Bioinformatics, 35 (2019) 4108-4119.

[31] K. Bleakley, Y. Yamanishi, Supervised prediction of drug–target interactions using bipartite local models, Bioinformatics, 25 (2009) 2397-2403.

[32] S. Cao, W. Lu, Q. Xu, Grarep: Learning graph representations with global structural information, Proceedings of the 24th ACM international on conference on information and knowledge management, 2015, pp. 891-900.

[33] L.F. Ribeiro, P.H. Saverese, D.R. Figueiredo, struc2vec: Learning node representations from structural identity, Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 385-394.

[34] D. Wang, P. Cui, W. Zhu, Structural deep network embedding, Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 1225-1234.

[35] Y.-C. Liu, W.-K. Huang, D.-L. Cheng, Antibacterial Activity of Cef podoxime in vitro, Chemotherapy, 43 (1997) 21-26.

[36] R. Carmena, G. Roederer, H. Mailloux, S. Lussier-Cacan, J. Davignon, The response to lovastatin treatment in patients with heterozygous familial hypercholesterolemia is modulated by apolipoprotein E polymorphism, Metabolism-Clinical and Experimental, 42 (1993) 895-901.

[37] R. Macdonald, T.S. Watts, Trifluoperazine Dihydrochloride (“Stelazine”) in Paranoid Schizophrenia, British medical journal, 1 (1959) 549.

[38] H.S. Sader, J.M. Streit, T.R. Fritsche, R.N. Jones, Potency and spectrum reevaluation of cefdinir tested against pathogens causing skin and soft tissue infections: a sample of North American isolates, Diagnostic microbiology and infectious disease, 49 (2004) 283-287.

[39] D.N. Levin, Z. Dulberg, A.-W. Chan, G.M. Hare, C.D. Mazer, A. Hong, A randomized-controlled trial of nabilone for the prevention of acute postoperative nausea and vomiting in elective surgery, Canadian Journal of Anesthesia/Journal canadien d'anesthésie, 64 (2017) 385-395.

[40] T. Pillay, D.G. Pillay, M. Adhikari, A.W. Sturm, Piperacillin/tazobactam in the treatment of Klebsiella pneumoniae infections in neonates, American journal of perinatology, 15 (1998) 47-51.

[41] D.T. Markel, L.S. Gold, J. Allen, C.E. Fahrenbruch, T.D. Rea, M.S. Eisenberg, P.J. Kudenchuk, Procainamide and Survival in Ventricular Fibrillation Out‐of‐hospital Cardiac Arrest, Academic emergency medicine, 17 (2010) 617-623.

[42] R.H. Falk, Flecainide-induced ventricular tachycardia and fibrillation in patients treated for atrial fibrillation, Annals of internal medicine, 111 (1989) 107-111.

[43] R.M. Gesser, K.A. McCarroll, G.L. Woods, Efficacy of ertapenem against methicillin-susceptible Staphylococcus aureus in complicated skin/skin structure infections: results of a double-blind clinical trial versus piperacillin-tazobactam, International journal of antimicrobial agents, 23 (2004) 235-239.