A Network-Based Embedding Method for Drug-Target Interaction Prediction

Poorya Parvizi[1], Francisco Azuaje[2,3], Evropi Theodoratou[1] and Saturnino Luz[1]

Abstract— **Integration of multi-omics and pharmacological data can help researchers understand the impact of drugs on dynamic biological systems. Network-based approaches to such integration explore the interaction of different cellular components and drugs. However, with ever-increasing amounts of data, processing these high-dimensional biological networks requires powerful tools. We investigate whether network embeddings can address this problem by providing an effective method for dimensionality reduction in drug-related networks. A neural network-based embedding method is employed to encode protein-protein, protein-disease, drug-drug and drug-disease networks for the prediction of novel drug-target interactions. We found that drug-target interaction prediction using embeddings of heterogeneous networks as input features performs comparably to state-of-the-art methods, exhibiting an area under the ROC curve of 84%, outperforming methods such as BLM–NII and NetLapRLS, and coming very close to the best performing network methods such as HNM, CMF and DTINet. These encouraging results suggest that further investigation of this approach is warranted.**

I. INTRODUCTION

Current research on drug discovery and drug repurposing is not restricted to traditional *in vivo* experiments. Novel, state-of-the-art, computational methods promote drug discovery by selecting the most suitable candidate compounds whilst also reducing time and cost of experiments [1], [2]. Moreover, the ever-increasing availability of omics and pharmacological data have created new opportunities for more effective computational drug design by aiding the exploration of chemical associations with pre-designed drugs as well as predicting mechanisms of drug action in targeting specific biomolecules [3]. Network (graph) based approaches have been proposed for the integration of the different data sources to enhance drug-target interaction (DTI) prediction [4].

Although integrating different data types in a single heterogeneous network could provide new valuable insights into the emerging field of drug discovery, interpreting such high-dimensional data has proved a challenge to conventional statistical methods. To achieve accurate and biologically meaningful predictions, the development of methods which can manage large high-throughput data is essential. Embedding methods might provide an effective way to overcome the complexity of network-based approaches in high-throughput data analysis. Neural network based embedding methods transform high-dimensional data into low dimensional vector spaces (features) whilst preserving topological properties of their higher dimensional counterparts [5].

In this study, features of drugs and proteins are extracted from their related networks using a neural network embedding method, and used in DTI prediction. We propose two pipelines to learn features of drugs/proteins using the node2vec embedding method [6], and compare these approaches with state-of-the-art methods for DTI prediction. The first pipeline, heterogeneous network embedding (HNE), entails constructing two heterogeneous networks; integration of drug-drug and drug-disease networks and integration of protein-protein (including homologous proteins) and protein-disease networks. The node2vec algorithm is then applied to each of these heterogeneous networks to retrieve low dimensional features of drugs and proteins interactions. In contrast, the second pipeline, individual network embedding (INE), applies the node2vec algorithm to each network separately without integrating them. Then features achieved from each network are combined to create drugs and proteins feature matrices. Following the embedding step in each pipeline, Inductive Matrix Completion (IMC) [7] is used to aid the prediction of DTI using the aforementioned drug and protein feature matrices.

II. RELATED WORK

Computational methods for DTI prediction encompass three main approaches: ligand-based DTI, docking simulation and chemogenomic approaches [8]. Ligand-based methods predict the affinity of the drugs for a given target by comparing new ligands with the known protein ligands. However, due to the existence of few known ligands this prediction method is inadequate [9]. Docking simulation predicts the physical complementarity of drugs with proteins using three-dimensional structures and calculates the binding energy between them [10]. However this prediction is challenging due to the low number of discovered protein structures. Chemogenomics is the novel approach in the prediction of DTIs by combining disciplines of chemistry, genomics and proteomics [11]. This method systematically screens libraries of small molecules against each drug target families in order to develop novel drugs. In this approach, machine learning methods have demonstrated promising performance at predicting interactions between drugs and targets.

One of the earliest chemogenomics methods is the Bipartite Local Models (BLM) which predicts the drug targets given a known drug, and the drugs given a known protein. In this supervised method, these two independent

[1]P. Parvizi, E. Theodoratou and S. Luz are with The Usher Institute, Edinburgh Medical School, The University of Edinburgh, United Kingdom {poorya.parvizi,e.theodoratou,s.luz}@ed.ac.uk.

[2]F. Azuaje was with Quantitative Biology Unit, Luxembourg Institute of Health (LIH), Luxembourg, Luxembourg.

[3]F. Azuaje's current affiliation is Data and Translational Sciences, UCB Celltech, Slough, United Kingdom Francisco.Azuaje@ucb.com.

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

predictions can later be aggregated to predict the DTI [12]. This method can be extended by integrating neighbour based interaction profile inference to BLM (BLM-NII). This further increases the power of the method toward DTI prediction of proteins that have no known interactions [13]. Manifold Laplacian regularized least square (LapRLS) is a regression method also based on the BLM concept [14]. Its extension, NetLapRLS, integrates a kernel from known DTI networks. Another method, Collaborative Matrix Factorization (CMF) uses known interactions as well as similarity amongst drugs and proteins [15] to predict DTI. This method is extended to Multiple Similarities CMF (MSCMF) which employs more than one similarity matrix of drugs and proteins by projecting them into common low-rank feature spaces.

In another chemogenomics approach to DTI prediction, Wang et al. constructed a heterogeneous network model (HNM) to explore the association between drug and target using node diffusion states [16]. Finally, a study by Luo et al [17], similar to our study, focuses on low-dimensional vector representations of each node in a heterogeneous network. However, their approach uses diffusion component analysis (DCA) [18] and singular-value decomposition (SVD) [19] to reduce the network's dimensionality. Then vector space projection is employed to predict the DTIs based on the aforementioned vector representations.

III. METHODS

A. Datasets

All data used in analyses are open access and publicly available, thus, ethical approval was not necessary for this study. Drug-drug and DTIs were extracted from the DrugBank (Version 3.0) [20]. This online free database provides information on drug structures, drug sequences, drug actions and their targets. Protein-protein interactions (PPI) were obtained from the Human Protein Reference Database (HPRD, release 9) [21]. Disease related interactions, drug-disease and protein-disease networks were downloaded from the Comparative Toxicogenomics Database (CTD) [22]. These networks converted to binary data, where 0 and 1 represent the absence or presence of an association respectively, between 708 drugs, 5603 diseases and 1512 proteins. Exploration of the node degree distributions of these networks show that all the networks, in particular the homogeneous networks, are in line with the scale-free network topology.

B. Network Embedding

1) Heterogeneous Network Embedding (HNE): As mentioned above, for this pipeline two heterogeneous networks were constructed: 1) a drug-related heterogeneous network consisting of drug-drug and drug-disease networks, and 2) a protein-related heterogeneous network constructed by the integration of protein-protein and protein-disease networks. Then node2vec was applied to these heterogeneous networks to construct vector representations of drugs and proteins.

Creating node embeddings with node2vec involves two algorithms: random walk with restart (RWR), and skip-gram model training. Skip-gram models have been widely used in natural language processing (NLP) research to predict the probability of surrounding words in sentences given a target word [23]. This model consists of a neural network with three layers: an input layer, a hidden layer, and an output layer. The softmax function is then used to constrain the results into a probability distribution that can be taken to correspond to the probabilities that certain words occur in the same context as a target word. Therefore, in order to measure the probability of two words being in the same context, the model preserves the features of each word in the hidden layer. In other words, this layer can be extracted and used as a low-dimensional vector representation. We calculated the relation probability of drugs and proteins using the same approach. The sequences to be estimated were generated by converting the heterogeneous networks into directed acyclic subgraphs. These subgraphs were generated by multiple random walks from each node of the network with the help of the RWR algorithm [24]. A schematic representation of this method is demonstrated in Figure 1.

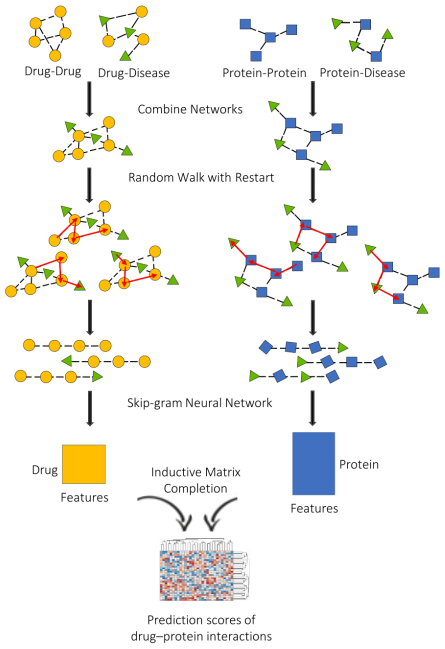

Fig. 1. Heterogeneous Network Embedding pipeline. The left side of the figure represents drug-related networks and the right side represents protein-related networks. Following the heterogeneous network construction, RWR and skip-gram neural network were applied and then drug and protein vector representations were fed into IMC to predict drug-target associations.

2) Individual Network Embedding (INE): In this pipeline, RWR and skip-gram were applied the same way as in the HNE approach. However, instead of integrating the networks, in this approach network embeddings were extracted from each network separately. Therefore, vector representations of drugs were obtained from both embeddings of a drug-drug network and embeddings of a drug-disease network, separately. Similarly, vector representations of proteins were obtained from embeddings of protein-protein and protein-disease networks.

C. Drug-Target Interaction Prediction

The IMC method [7] was employed to predict the DTI scores using known DTIs and vector representations of drugs and proteins as training data. The known DTI matrix contains 0 and 1 entries representing interaction and non-interaction respectively. We excluded some of the associations in the matrix and used the resulting data as a training set for the IMC algorithm. Excluded associations were used as the test set to assess the performance of the predictions. The aim of the matrix completion algorithm is to recover the missed entries of the low-rank matrix using a fewer number of inputs [25]. The DTINet algorithm [17] also uses the IMC method to predict DTIs and tests the performance of the method using 5-times ten-fold cross-validation procedure. It is possible that a different predictive model, such as integrating the node embedding layer into a deep neural network including a convolution layer might yield higher DTI prediction accuracy. However, in order to be able to compare our feature extraction method to DTINet we applied the same predictive model and performance procedures as Luo et Al [17].

IV. Results and Discussion

As explained above, we evaluated the performance of HNE and INE pipelines with IMC using 5-times ten-fold cross-validation withholding out 10% of the drug-target interactions and the matching number of non-interactions as test sets, using the rest of the data as training sets. Employing the non-interactions as test set allows the calculation of false positive and false negative rates. Due to this, in each cross-validation set the performance of the prediction was reported in the form of receiver operating characteristic (ROC) and precision-recall (PR) curves. The ROC curve plots the true positive rate versus false positive rate [26] and the larger the area under ROC (AUROC) curve the more successful predictions are. Similarly, the larger area under the PR curve (AUPR) shows high prediction performance with high precision and sensitivity. Here, for each pipeline, we report the average of AUROC and AUPR curves, as each fold in cross-validation has its own ROC and PR curves. The HNE pipeline exhibits an AUROC of 84%, while the INE pipeline obtains only 70% (Figure 2). The HNE method was compared to the state-of-the-art DTI prediction algorithms. The AUROC curve results show that the HNE pipeline outperforms methods such as BLM–NII (67%) and NetLapRLS (83%), and exhibits performance close to the best performing network methods, including HNM (85%), CMF (86%) and DTINet (91%) (Figure 3). The approximate predictions of mentioned DTI methods are obtained from Luo et al [17].

Similarly to the HNE pipeline, HNM and CMF are designed to predict DTIs by using heterogeneous data. These methods outperform methods which are designed to predict DTI by employing single networks, such as BLM–NII and NetLapRLS. This demonstrates the advantage of heterogeneous networks in the prediction of drug-target pairs. Heterogeneous networks harbour different types of biological

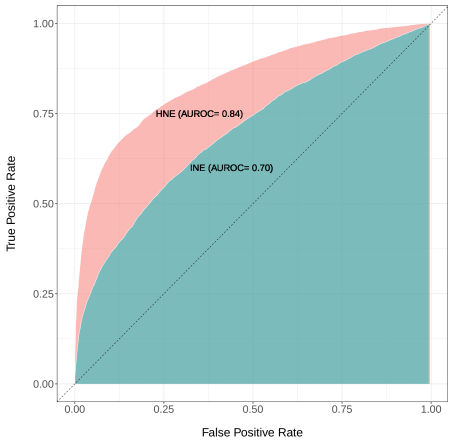

Fig. 2. This figure shows the ROC curves of the drug-target prediction performance in HNE and INE pipelines. Average AUROC curve values are 0.84 and 0.70 for HNE and INE respectively.

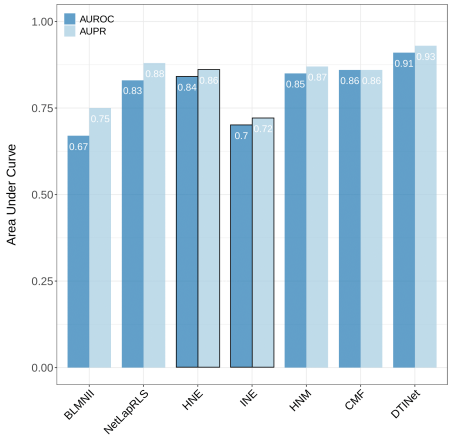

Fig. 3. This figure shows the average AUROC and AUPR curve values of different drug-target interaction methods including HNE and INE (highlighted with black borders). AUROC is the area under the curve of false positive against true positive rate. However, AUPR is the area under the curve of sensitivity versus precision.

data, and this interconnected structure uncovers invaluable new indirect connections in the data. In the case of the HNE pipeline, for instance, two networks describe the association of proteins, the PPI network and the protein-disease network. The integration of these networks creates new indirect protein-disease associations [27].

To the best of our knowledge this study is the first approach in using a neural network based embedding method in DTI prediction. As mentioned above, it is possible that prediction performance may be improved for HNE representations if they were used in conjunction with a deep neural network architecture, and hyperparameters were systematically tuned. For instance, the subgraphs which served as input to the the skip-gram algorithm were generated by RWR with the random movement from each node in the network. However, the number of walks and walk lengths are hyperparameters in RWR. The skip-gram model itself

has two hyperparameters, namely, windows size and number of nodes in the hidden layer. In this study, the most common hyperparameter settings were used, and no attempt was made to search for the best parameters for the DTI task. The reason for this is that our aim was to compare the dimensionality reduction methods used in the top performing DTI prediction model (DCA for DTINet) to the network embedding method. However, in future we aim to investigate the effect that hyperparameter setting might have on DTI prediction performance.

The results reported by BLM–NII, NetLapRLS, HNM, CMF, DTINet and our study's pipelines are based on original collected data sets of proteins, without removing homologous proteins. Homologous proteins have similar structures and function, and therefore may artificially boost the accuracy of DTI prediction [17]. Future work will also investigate the effect of homologous proteing removal on the embedding based methods presented in this paper.

V. CONCLUSION

This study demonstrates that neural network based node embedding method can reduce the high-dimensional biological networks to low-dimensional feature sets which can be successfully used in DTI prediction. The HNE pipeline exhibited the AUROC curve of 84% and outperformed the BLM–NII and NetLapRLS algorithms on this task. This pipeline also showed close performance to the HNM, CMF and DTINet methods. Further investigations of this method such as exploring the most suitable hyperparameters in network embedding and using of the features in a full-fledged deep neural network are therefore warranted, and may result in further prediction performance improvements. In addition, this method will be applied on different datasets such as cancer and neurodegenrative disease to evaluate the method and compare with other state-of-the-art methods.

ACKNOWLEDGMENTS

PP is funded by the University of Edinburgh's Global Research Scholarship and the Chancellor's fellowship awarded to SL. ET is supported by a CRUK Career Development Fellowship (C31250/A22804).

REFERENCES

[1] S. C. Basak, "Chemobioinformatics: The Advancing Frontier of Computer-Aided Drug Design in the Post-Genomic Era," *Current Computer Aided-Drug Design*, vol. 8, no. 1, pp. 1–2, mar 2012.

[2] S. Dibyajyoti, E. Talha Bin, and P. Swati, "Bioinformatics:The effects on the cost of drug discovery," *Galle Medical Journal*, vol. 18, no. 1, p. 44, may 2013.

[3] H. Matthews, J. Hanison, and N. Nirmalan, ""Omics"-informed drug and biomarker discovery: Opportunities, challenges and future perspectives," sep 2016.

[4] D. K. Arrell and A. Terzic, "Network Systems Biology for Drug Discovery," *Clinical Pharmacology & Therapeutics*, vol. 88, no. 1, pp. 120–125, jul 2010.

[5] W. Nelson, M. Zitnik, B. Wang, J. Leskovec, A. Goldenberg, and R. Sharan, "To embed or not: Network embedding as a paradigm in computational biology," 2019.

[6] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu. Association for Computing Machinery, aug 2016, pp. 855–864.

[7] H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon, "Large-scale multi-label learning with missing labels," in *International Conference on Machine Learning (ICML)*, vol. 32, 2014, pp. 593–601.

[8] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, "Machine Learning for Drug-Target Interaction Prediction," *Molecules*, vol. 23, no. 9, p. 2208, aug 2018.

[9] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature Biotechnology*, vol. 25, no. 2, pp. 197–206, feb 2007.

[10] V. Salmaso and S. Moro, "Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview," aug 2018.

[11] H. Kubinyi and G. Muller, *Chemogenomics in drug discovery : a medicinal chemistry perspective*. Wiley-VCH, 2006.

[12] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug-target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, sep 2009.

[13] J. P. Mei, C. K. Kwoh, P. Yang, X. L. Li, and J. Zheng, "Drug-target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, jan 2013.

[14] Z. Xia, L.-Y. Wu, X. Zhou, and S. T. C. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces." *BMC systems biology*, vol. 4 Suppl 2, p. S6, sep 2010.

[15] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-Target interactions," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. Part F1288. Association for Computing Machinery, aug 2013, pp. 1025–1033.

[16] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, oct 2014.

[17] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Communications*, vol. 8, no. 1, dec 2017.

[18] H. Cho, B. Berger, and J. Peng, "Diffusion component analysis: Unraveling functional topology in biological networks," 2015.

[19] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[20] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, "DrugBank 3.0: A comprehensive resource for 'Omics' research on drugs," *Nucleic Acids Research*, vol. 39, no. SUPPL. 1, jan 2011.

[21] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. I. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human Protein Reference Database - 2009 update," *Nucleic Acids Research*, vol. 37, no. SUPPL. 1, 2009.

[22] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The Comparative Toxicogenomics Database: Update 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D972–D978, jan 2017.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[24] H. Tong, C. Faloutsos, and J. Y. Pan, "Fast random walk with restart and its applications," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2006, pp. 613–622.

[25] P. Jain and I. S. Dhillon, "Provable Inductive Matrix Completion," *arXiv preprint*, pp. 1–22, jun 2013.

[26] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, apr 1993.

[27] K. Tsuyuzaki and I. Nikaido, "Biological Systems as Heterogeneous Information Networks: A Mini-review and Perspectives," *arXiv preprint*, dec 2017.