Network-based drug repurposing using network Proximity

Abhayram A
2019102017

Aswin Jose
2019113016

Abstract

We propose an improved version of Similarity Network Fusion Neural Network (SNF-NN) called Similarity Network Fusion Network Proximity Neural Network (SNF-NPNN) for the task of drug repurposing. Our approach includes network-based information on drug-target and disease-target genes in the SNF-NN pipeline, resulting in significant performance enhancement compared to the original model. We also utilized less sparse labels obtained from knowledge graphs during training, improving the performance of our model. Our method shows promising results for the drug-disease link prediction task.

1 Introduction

Drug repurposing has emerged as a promising approach to identify new therapeutic uses for existing drugs. By leveraging the known safety and efficacy profiles of approved drugs, this strategy can significantly reduce the time and cost associated with traditional drug development. However, identifying drug-disease interactions remains a challenge. In this work, we propose a model that can predict the usability of a particular drug for a given disease based on the relationship between the drug target genes and the disease target genes on the human protein-protein interaction (PPI) network.

Our goal is to reduce the clinical testing time required to evaluate multiple drugs for a specific disease. To achieve this, we build upon previous work in the field. One such work is SNF-NN[1], a computational method that predicts drug-disease interactions using similarity network fusion and neural networks. This method leverages drug-related similarity information, disease-related similarity information, and known drug-disease interactions to predict novel drug-disease interactions. A similar method is used in SNF-CVAE[2] where they use a variational auto-encoder instead. In another relevant work[3], we see a network-based approach for drug repurposing. This approach quantifies the network-based relationship between drug targets and disease proteins in the human protein-protein interactome to predict the drug's efficacy to treat the disease.

In this paper, we propose to combine the two methods used in [1] and [3] to create a model that takes into account both gene-level interactions between drugs and diseases as well as other properties of the drugs. Our core architecture remains similar to SNF-NN, but we modify the input information to include the network-based relationship between drug targets and disease proteins in the human protein-protein interactome. We also modify the SNF-NN pipeline to include drug-gene and disease-gene interaction information using a network proximity approach. We also obtain less sparse labels from the BIOSNAP knowledge-graph database[4]. We demonstrate that incorporating network information improves the link prediction ability of the algorithm.

1.1 Our work

In this work, we improved upon the SNF-NN architecture by implementing a similarity selection module and network proximity calculation module. We then used labels extracted from the drug-disease knowledge graph present in BIOSNAP database to do our predictions. We have also extracted

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

the drug-target interaction data and disease-gene interaction data from the drug-gene and disease-gene knowledge graphs, which were available in BIOSNAP accordingly.

2 Baseline Model

We use the SNF-NN as the baseline model for our analysis. The SND dataset with and without similarity network fusion were used as the input to this model and the results were compared against what was reported in the paper [1]. The SNF-NN is a simple MLP-based model consisting of 4 layers of fully connected neurons. We did 10-fold cross-validation with the same parameters as specified by the original work.

Data	Accuracy	F1	Recall	MCC	AUC-ROC	AUC-PR
Reported Values	0.796	0.800	0.816	0.593	0.867	0.876
Without Similarity Selection	0.82	0.72	0.64	0.60	0.85	0.79
With Similarity Selection	**0.83**	0.75	0.72	0.62	0.85	0.79

Table 1: Comparison of the performance reported in the paper SNF-NN and our model to establish the baseline.

3 Methodology

3.1 Dataset and Pre-Processing

The SND data set is used for obtaining the baseline. It is the gold-standard data consisting of 867 FDA-approved drugs, 803 diseases, and 8684 clinically reported and/or experimentally validated drug-disease interactions with 98.75 per cent sparsity(The drug-related data was obtained from DrugBank and RepoDB) was created for benchmarking drug-repurposing pipelines[2]. The similarity measures used for this dataset are given in the supplementary section.

A similarity network selection pipeline based on entropy was calculated for each similarity matrix and the similarity network fusion pipeline was run on it to combine all the features. The diagram given shows the rough representation of the above pipeline.

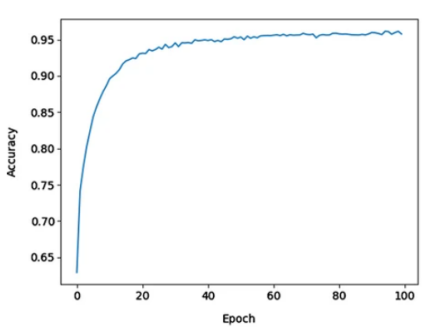
The data from the knowledge graphs were obtained from the BioSNAP dataset[4]. The following information was obtained from it:

- drug-gene (5017 drugs, 2324 genes, 15139 relations)
- disease-gene (5604 diseases, 17821 genes, 15509619 relations)
- drug-disease (1662 drugs, 5536 diseases, 466657 relations) interactions.

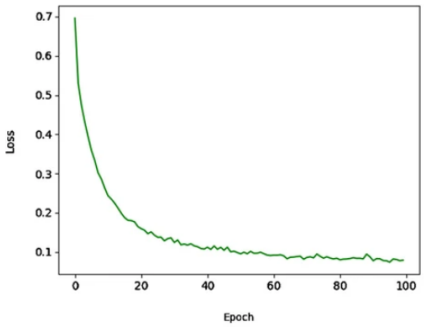
The drug-disease knowledge graph was used as labels to train the SNF-NPNN. The disease target gene-related data and the drug target gene-related data were extracted from the corresponding knowledge graphs mentioned above for the network proximity calculations. The protein-protein interaction network on which the target data is projected to is obtained from HumanNet-PI [5] which has 15,352 genes and 81.6% known human genome coverage with 158,499 links.

For the network proximity calculation task the drugs and the disease with target gene information are selected specifically. The target gene information was only available for 705 drugs and 436 diseases specifically. For network proximity based calculations, data associated with only these drugs and diseases were considered. We will be calling the filtered dataset SNDf in the following sections. To test the performance of SNF-NPNN we also used disease-disease and drug-drug similarity matrices from Cdataset [6] and LRSSL dataset [7].

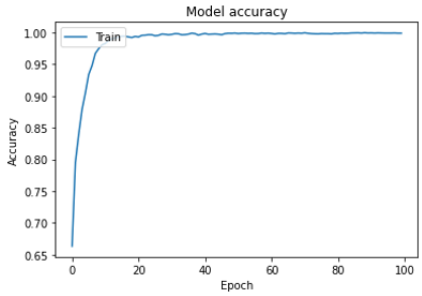
The drug identifiers used for the whole analysis is the Drugbank id and the disease id used is the MESH and OIMM ids. The disease ids for the datasets in hand was converted to the standard ids mentioned above by comparing them against the UMLS Metathesaurus database [8].



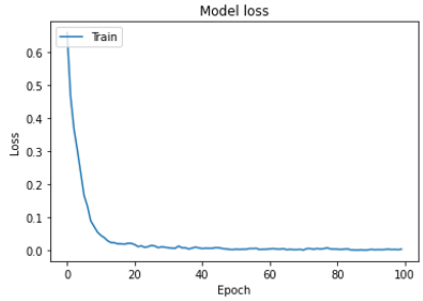
(a) Accuracy reported by paper



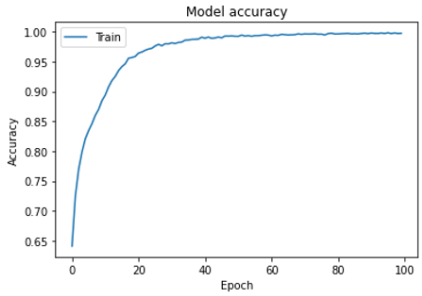
(b) Loss Reported by the paper



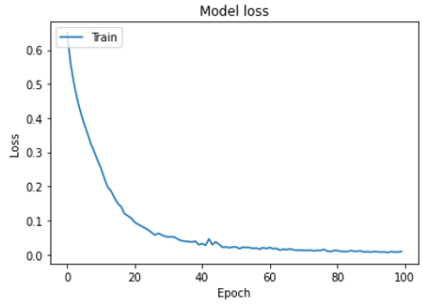
(c) Without feature selection



(d) Without feature selection



(e) With feature selection



(f) With feature selection

Figure 1: The accuracy and loss plots for the SND dataset with original labels on the baseline model

3.2 Similarity Network Selection

The entropy was calculated for each of the disease-disease similarity matrices and the drug-drug similarity matrices and it was thresholded (19 for drug-drug similarity matrices and 14 for disease-disease similarity matrices). The equation for entropy calculation is given below.

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

3.3 Similarity Network Fusion (SNF)

Similarity Network Fusion (SNF) is a method for integrating multiple heterogeneous data sources into a single network that can be used for various applications, including drug repurposing. SNF combines the individual similarity matrices from each data source and iteratively fuses them to generate a final integrated similarity matrix. The final matrix can then be used as input to various machine learning algorithms. The package SNFpy was used to perform SNF for our work. The similarity matrices

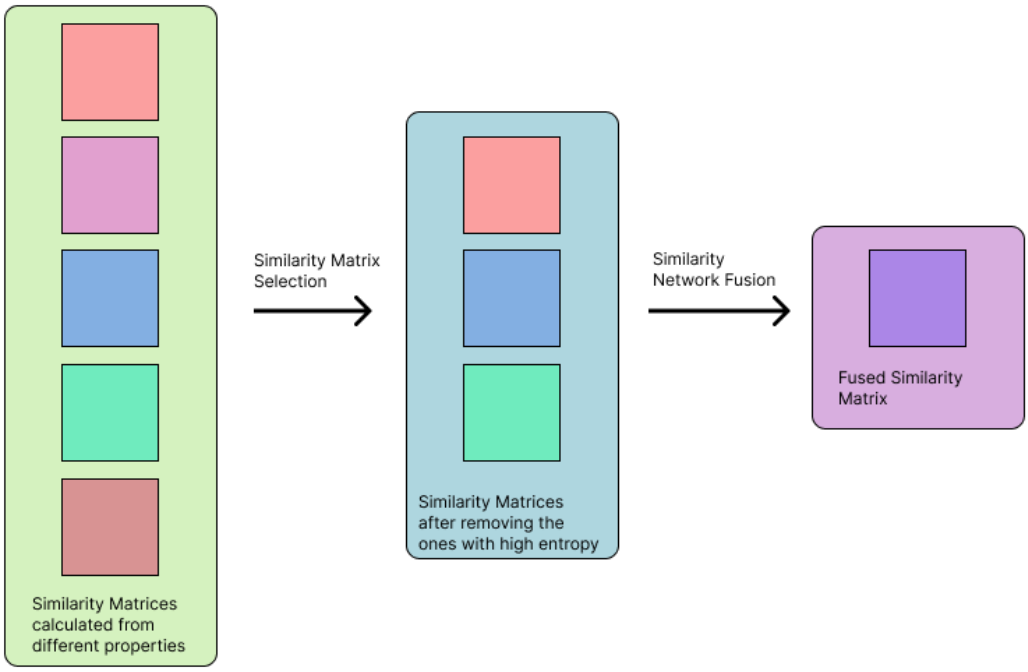


Figure 2: Similirarity network fusion pipeline

after thresholding were fed into the SNF module to obtain a combined drug-drug similarity and disease-disease similarity matrix.[9]

3.4 Network Proximity Calculations

The network proximity for a given two genes are calculated using the following distance measures mentioned below.

- **Shortest Path:** This method is based on Dijkstra's algorithm, which works by iteratively exploring the nodes in the network and updating the shortest path to each node as it is discovered.
- **Adamic Adar:** This is a similarity measure between nodes in a network/graph based on the common neighbors between them. The Adamic-Adar distance for two nodes is calculated by summing the inverse logarithm of the degree of each common neighbor. The idea behind this is that nodes with a high degree should have less weight in determining similarity because they are likely to be connected to many other nodes.

$$AA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{\log(deg(z))}$$

  where x and y are the nodes being compared, N(x) and N(y) are the neighbour sets of x and y, deg(z) is the degree of node z, and the summation is over all common neighbours of x and y.

- **Jaccard Index:** is a similarity measure between nodes in a network/graph based on the size of their shared neighborhood. The Jaccard index measures the ratio of the number of common neighbors to the total number of neighbors of both nodes.

$$J(x,y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}$$

where J(x, y) is the Jaccard index between nodes x and y, N(x) and N(y) are the neighbors of nodes x and y, |S| represents the size of set S, and intersection and union denote set intersection and set union, respectively.

- **Cosine Similarity Index:** This method is a similarity measure between nodes in a network/graph based on the cosine of the angle between their feature vectors. The cosine similarity index measures the cosine of the angle between two vectors in a high-dimensional space.

$$\text{cosine}(x, y) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}$$

where cosine(x, y) is the cosine similarity between nodes x and y and norm(x) and norm(y) are the Euclidean norms of the feature vectors of x and y, respectively.

- **Preferential Attachment:** This method calculates the product of the degree of the nodes in consideration.

The above metrics are used to compute the distances between each drug-target and disease-target genes in consideration and the mean, median, min, max of the computed distances are taken for the further downstream tasks.

3.5 Overall Pipeline

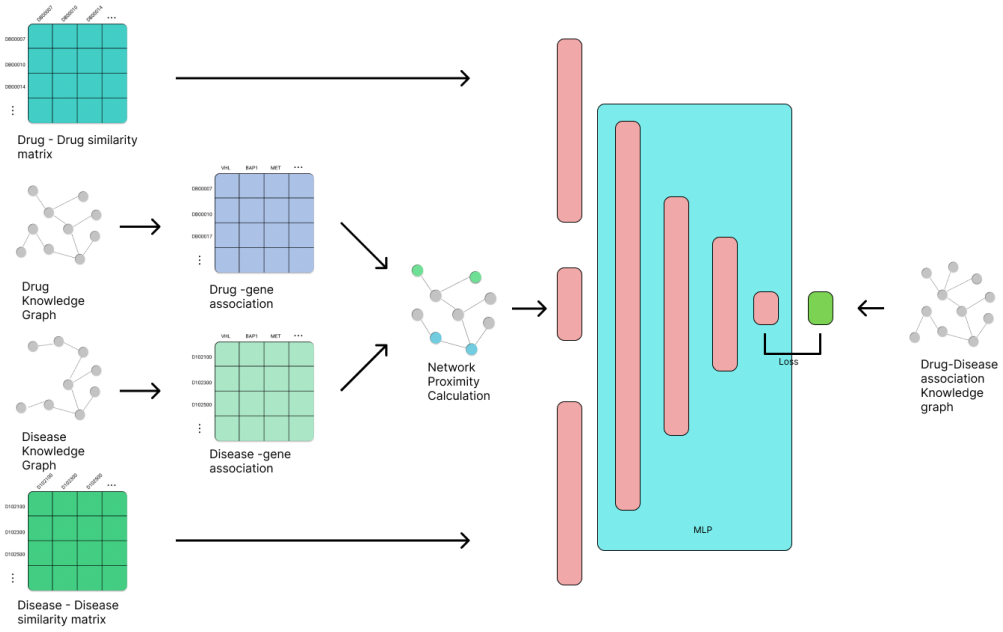


Figure 3: SNF-NPNN pipeline

- First, we construct drug-drug and disease-disease similarity matrices using available data. We then calculate the entropy value of these matrices to identify and remove matrices that contain less information.
- Next, we perform similarity network fusion on the remaining matrices to integrate the drug-drug and disease-disease similarity information. This step allows us to combine multiple sources of information to improve the accuracy of our predictions.
- In addition to the similarity network fusion, we also incorporate a network proximity approach to calculate the proximity between drugs and diseases in the human protein-protein interaction (PPI) network. This step allows us to take into account the gene-level interactions between drugs and diseases. Different distance measures mentioned above were used to

Distance Metric	Accuracy	F1	Recall	MCC	AUC-ROC	AUC-PR
Shortest Path	0.76	0.61	0.56	0.45	0.76	0.66
Jaccard Index	0.76	0.61	0.55	0.44	0.76	0.66
Cosine Similarity	0.76	0.60	0.56	0.43	0.76	0.65
Adamic Adar	0.76	0.62	0.57	0.45	0.76	0.66
Preferential Attachment	0.76	0.61	0.57	0.44	0.76	0.66
No distance	0.75	0.60	0.57	0.42	0.75	0.63

Table 2: Comparison of the performance of the model using different distance metrics on the SNDf dataset with the original labels provided in the paper SNF-NN and 3-fold cross validation was performed.

test the improvement in the performance of the model. There is no big difference in the performance of the final model using different distance metrics. The shortest path algorithm was used as the distance metric for further analysis.

- Finally, we pass the resulting information to a neural network based on the SNF-NN architecture. The neural network takes as input the fused similarity matrices and the network proximity information. It outputs a confidence score for the usability of a particular drug for a given disease. Stratified K-fold cross-validation is used to train and evaluate the model.

4 Experiments

4.1 Establishing the baseline model

The SNF was directly performed on all the drug-similarity and disease-similarity matrices in the SND dataset, and it was fed into the SNF-NN model. The second experiment was conducted by performing similarity network selection using the above-mentioned thresholds. The results are displayed in table 1. This model was run for 100 epochs using 10-fold cross-validation with the learning rate of 10^{-3}. The true negatives used for all the following studies are randomly generated from the set of drug-disease interactions that are not true positive.

4.2 Network Proximity Experiments

The SND-F dataset with the original labels provided by the paper SNF-NN were used to test the different network proximity measures. The results are provided in table 2 and it is compared against a model run without using any network distance measures. The differences are seen to be small since the dataset used for the analysis had only around 7000 labels where half of them were true positives. This dataset is used due to the time-complexity of each run. After further consideration the shortest path algorithm was used for further analysis.

4.3 Benchmarking SNF-NPNN

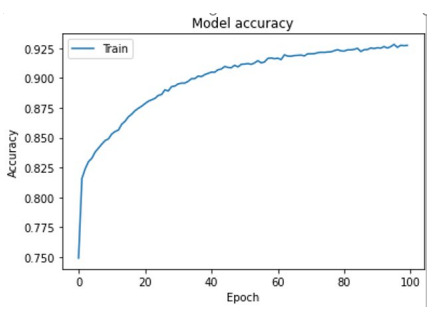
To benchmark the new pipeline we used the SNDf dataset with original labels and also with the labels obtained from the knowledge graphs. For both cases, the results are given in the tables (TABLES 3 and 4) respectively and it is compared against the case where the network distance metrics are not used.

Metric	Accuracy	F1	Recall	MCC	AUC-ROC	AUC-PR
Without Network Proximity	0.71	0.72	0.75	0.41	0.76	0.70
With Network Proximity	**0.73**	**0.75**	**0.78**	**0.47**	**0.80**	**0.78**

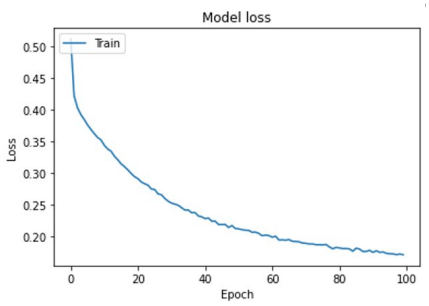
Table 3: SNF-NPNN performance on Cdataset-LRSSL similarity matrices. It had a total of 10,000 labels with half of them being true positives

Metric	Accuracy	F1	Recall	MCC	AUC-ROC	AUC-PR
Without Network Proximity	0.80	0.82	0.91	0.61	0.83	0.75
With Network Proximity	**0.91**	**0.91**	**0.93**	**0.82**	**0.98**	**0.98**

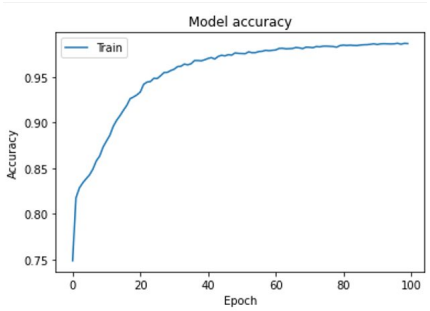
Table 4: SNF-NPNN performance on SNDf with labels from knowledge graph. The dataset had around 95,000 datapoints with 50% of them being true positive



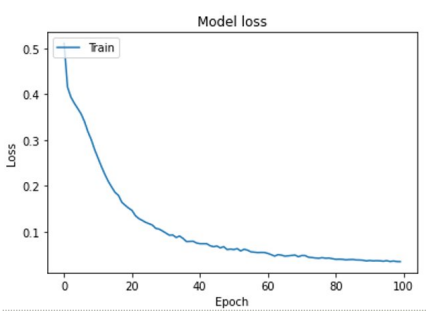
(a) Without proximity scores



(b) Without proximity scores



(c) With proximity scores



(d) With proximity scores

Figure 4: The accuracy and loss plots for the SNDf dataset with labels from knowledge graph on the SNF-NPNN architecture

4.4 Performance Difference on introducing class imbalance

The SNF-NPNN model is trained by introducing an imbalance in the class distribution. (All the earlier experiments made sure of equal class distribution). The below table shows the difference in the performance of the model for different class imbalances:

Imbalance	Accuracy	F1	Recall	MCC	AUC-ROC	AUC-PR	False Positives
1	0.91	0.91	0.93	0.82	0.98	0.98	1540
2	0.94	0.91	0.88	0.87	0.98	0.97	1432
3	0.96	0.90	0.86	0.88	0.98	0.96	1083
4	0.966	0.90	0.86	0.88	0.98	0.95	1019

Table 5: The performance of the model trained with class imbalance, here imbalance refers to the ratio of true negatives(which are randomly generated as mentioned above) with true positives

It is observed that the class imbalance plateaus at an imbalance of 4 and the performance of the model does not increase further.

5 Discussions and Conclusion

From the above results, it is evident that including network proximity-related information in the pipeline has improved the overall performance of the model. Throughout the experiments, the

network-based information, such as network proximity, has been shown to give a significant boost to the model performance as it might have captured the low-level interactions at the genomic level present in the human body between the drug targets and the genes affected by the diseases. This information is not captured by the similarity matrices shown above as it is a similarity computation within the drug dataset and the disease dataset separately. They are not taking into account any direct drug-disease associations, which can be useful in predicting the links. Hence our model works better compared to SNF-NN in the link prediction task.

Different similarity measures were explored and the shortest path and Adamic Adar distance were found to have the best performance compared to other metrics. Adding the knowledge graph-based labels have significantly improved the model's performance compared to the original labels provided in the dataset. It can be attributed to the lesser sparsity of the knowledge graph (87% sparse).

6 Future Work

The similarity matrix selection pipeline can be further optimized to take in lesser number of similarity matrices for the same analysis. The ML model can be modified to directly incorporate graph information rather than pre-calculated network proximity information. For example, using Graph Convolutional Networks (GCN's) could boost performance. More prior information can also be used by using attention mechanisms to target specific sets of important genes or pathways.

Code availability - https://github.com/ML4Sciences/final-project-codebase-meagoodboy.git

References

[1] T. N. Jarada, J. G. Rokne, and R. Alhajj, "Snf-nn: computational method to predict drug-disease interactions using similarity network fusion and neural networks," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–20, 2021.

[2] T. N. Jarada, J. G. Rokne, and R. Alhajj, "Snf–cvae: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder," *Knowledge-Based Systems*, vol. 212, p. 106585, 2021.

[3] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, and F. Cheng, "Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2," *Cell discovery*, vol. 6, no. 1, p. 14, 2020.

[4] S. M. Marinka Zitnik, Rok Sosič and J. Leskovec, "BioSNAP Datasets: Stanford biomedical network dataset collection." `http://snap.stanford.edu/biodata`, Aug. 2018.

[5] S. Hwang, C. Y. Kim, S. Yang, E. Kim, T. Hart, E. M. Marcotte, and I. Lee, "Humannet v2: human gene networks for disease research," *Nucleic acids research*, vol. 47, no. D1, pp. D573–D580, 2019.

[6] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F.-X. Wu, and Y. Pan, "Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm," *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, 2016.

[7] X. Liang, P. Zhang, L. Yan, Y. Fu, F. Peng, L. Qu, M. Shao, Y. Chen, and Z. Chen, "Lrssl: predict and interpret drug–disease associations based on data integration using sparse subspace learning," *Bioinformatics*, vol. 33, no. 8, pp. 1187–1196, 2017.

[8] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[9] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods*, vol. 11, no. 3, pp. 333–337, 2014.

Supplementary Information

SND dataset

The similarity matrices present in the dataset are given below:

Index	Disease Matrix	Entropy
0	GPGenes	16.5184
1	BefreeGenes	13.1649
2	LncRNA	11.3152
3	Resnik	16.1275
4	CureatedGenes	7.9051
5	MirRNA	12.9200
6	BefreeVariants	8.1269
7	GPHpo	12.4377
8	Wang	16.1275
9	Lin	16.3714
10	PSB	16.4659
11	DisGeNET	8.5215
12	Fun	13.3500
13	CuratedVariants	7.9051

Table 6: The disease similarity matrix entropy

Index	Disease Matrix	Entropy
0	Drug	18.9302
1	Therapeuticnet	18.6956
2	Metanet	18.7264
3	BPnet	19.0769
4	Chemicalnet	18.9843
5	Protein	15.0228
6	CCnet	19.0866
7	Wmnet	19.0759
8	MFnet	19.0282
9	SideEffect	18.5496

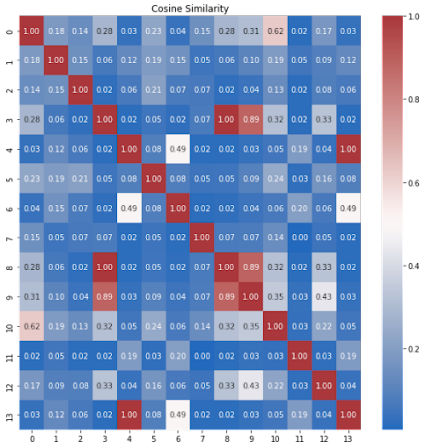
Table 7: The drug similarity matrix entropy

Further information about these matrices is provided in the paper SNF-NN[1].

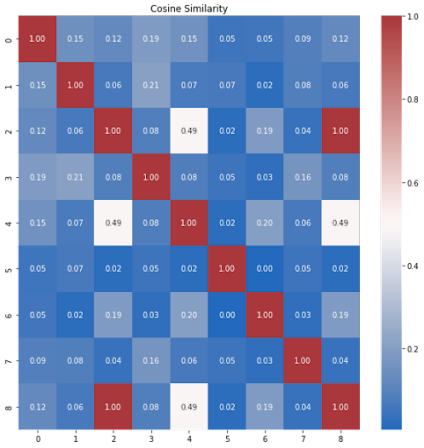
The computed pairwise similarity between the similarity matrices is also given below:

SNDf with Knowledge graph labels

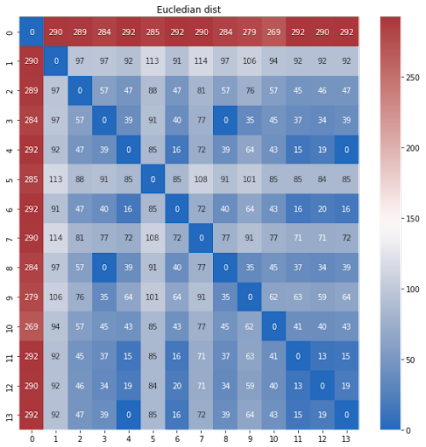
To obtain the drug labels and gene labels which matched the knowledge graph, an api was designed to extract the drug ids using drug concept ids(used in SND dataset) and the gene name where matched using the supplementary table provided by the BioSNAP database. Only 436 diseases could be matched using the API out of 803 diseases available and the gene target information was only available for 705 drugs out of 867 drugs. For this set of drug-disease pairs , 47981 true positive labels were available(84.4% sparsity).



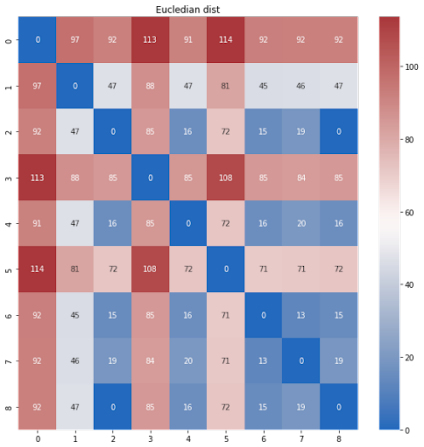
(a) Cosine similarity before feature selection



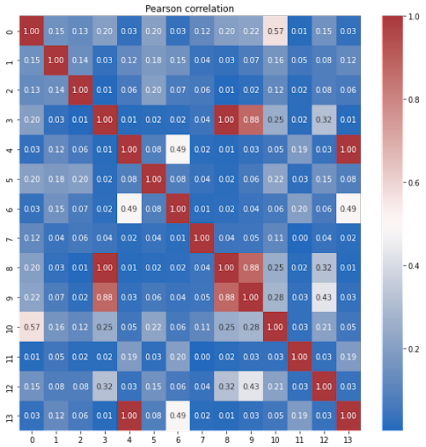
(b) Cosine similarity after feature selection



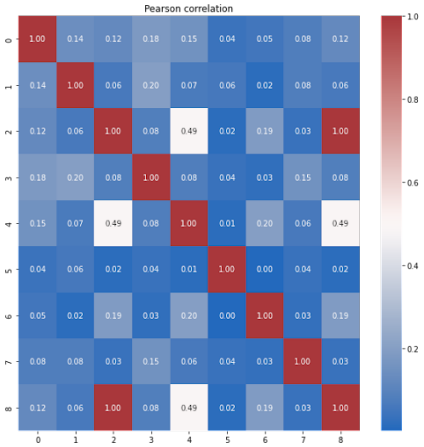
(c) Euclidean Distance before feature selection



(d) Euclidean Distance after feature selection

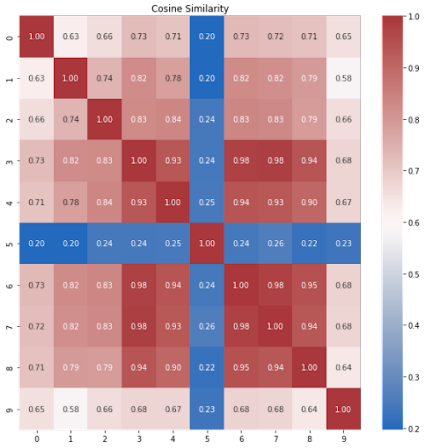


(e) Pearson Co-relation before feature selection

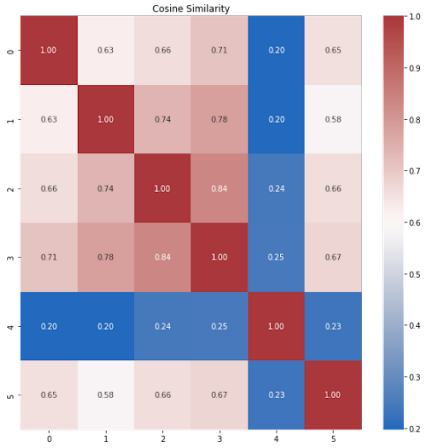


(f) Pearson Co-relation after feature selection

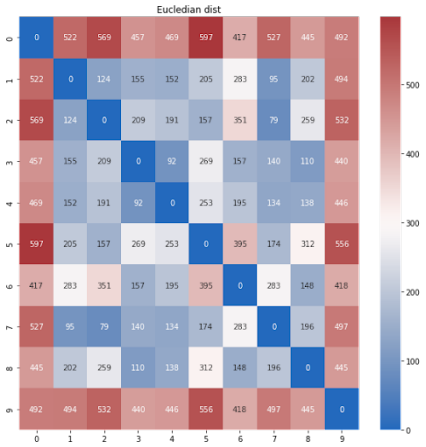
Figure 5: The pairwise correlation of disease similarity matrices before and after similarity selection, the numbers map to the indexes given in the Entropy table above



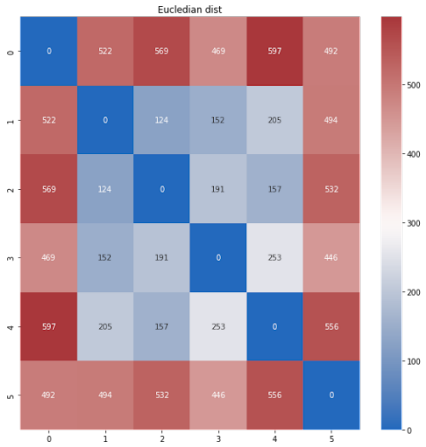
(a) Cosine similarity before feature selection



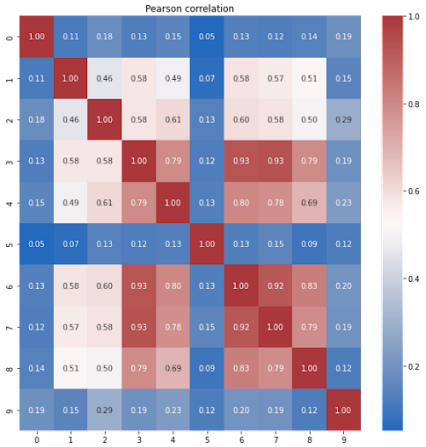
(b) Cosine similarity after feature selection



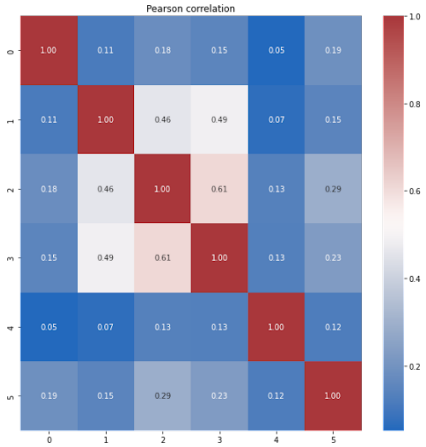
(c) Euclidean Distance before feature selection



(d) Euclidean Distance after feature selection



(e) Pearson Co-relation before feature selection



(f) Pearson Co-relation after feature selection

Figure 6: The pairwise correlation of drug similarity matrices before and after similarity selection, the numbers map to the order of indexes given in the Entropy table above