HNet-NeuMF: Predicting Drug–Disease Associations with Neural Matrix Factorization Based on Heterogeneous Network Features

Anuj
2018220
Indraprastha Institute of Information Technology, Delhi
Delhi, India
anuj18220@iiitd.ac.in

Bhavey Wadhwa
2018135
Indraprastha Institute of Information Technology, Delhi
Delhi, India
bhavey18135@iiitd.ac.in

Abstract: Despite all the huge investments in drug research and development, it is a time consuming task. Hence, some situations of crisis presses an urgent demand of discovering novel indications of an approved drug, which is also known as drug repositioning. It is an efficient and economical way of drug discovery. Given the high volumes of chemical, genomic, and pharmacological data sets generated by the high-throughput technique, the task of drug repositioning can be assisted by various computational approaches to find the best indication for a drug. In past few years, deep learning has proven to be successful in many research fields such as natural language processing, machine learning, computer vision and as well as recommender systems. Many researches have incorporated meta data using deep learning techniques to achieve better accuracy in recommendations. The use of metadata information is crucial in bioinformatics as every slight information is crucial. In this paper, we introduce HNet-NeuMF, which utilizes generalised matrix factorization(GMF), multi layer perceptron(MLP) and neural matrix factorization(NeuMF), to predict new drug disease associations, based on the features extracted from the drug disease heterogeneous network. Instead of concatenating the chemical and phenotypic features as, we used these raw features of drugs and diseases as input to our model to construct a disease-disease similarity network and a drug-drug similarity network, and then built a drug disease heterogeneous network by combining known drug disease associations. Then, we extracted topological features for drugdisease associations from the same heterogeneous network which we used to train our model. Then, GMF is applied to the known drug-disease association, whereas MLP is applied to metadata. Using NeuMF, the outputs for GMF and MLP are then concatenated and input to a neural network for predicting drug-disease association. The results of the proposed algorithm have been validated on two standard drug-disease association databases (Cdataset and PREDICT) by evaluating matrices such as AUC, ROC, etc. across a 10-fold cross validation split. The results of the experiments demonstrate that the proposed approach effectively exploits the features of the heterogeneous network to enhance the performance of drug-disease association over the baseline and the state of the art methods.

I. Introduction

Drug research and development is well-known to be an expensive and time consuming task. There is a very limited number of drugs approved in the last few decades [1]. This can be attributed to the long time it takes to discover and approve a drug as market-ready. This calls for an alternative, such as drug repositioning which has emerged as a research hotspot. It is an efficient and economical alternative to drug discovery as less time and money would be spent on approving and launching the drug. There have been some successful repositioned drugs done by investigating manually but this approach is inefficient and non-scalable given the huge volumes of chemical, genomic, and pharmacological data sets generated by the high-throughput technique. Therefore, computational approaches have been used over the past years to systematically prune down the massive search space for researchers, saving huge amounts of efforts, time and cost; predicting novel indications. This explains the utter need of using advance computational techniques to find new associations between drug and diseases. In this work, we present a heterogeneous network-based Neural matrix factorisation model, referred to as HNet-NeuMF, to predict new drugdisease associations. We collected the chemical structure of drugs, phenotypic attributes of diseases, and known drugdisease associations to use them to build similarity networks as input to our model. Then, we trained our model with extracted features for drugdisease associations from the heterogeneous network. Then, GMF is applied to the known drug-disease associations, whereas MLP is applied to metadata. Using NeuMF, the outputs for GMF and MLP are then concatenated and input to a neural network for predicting drug-disease association.

II. Literatue Review

Initial attempts in this field were based on profiles of gene expression [2], [3]. [2] proposed a database which ranked drug response gene expression with a gene signature specific

to a disease. Drug response profiles which correlated and anti-correlated were identified. However, this approach lacked validation from a large scale dataset. It might not be precise enough owing to the conditions under which response profiles were created for different drugs. Later, network based models were suggested like [4] proposed a method that predicted drug-disease association using various drug disease similarities to find feature vectors which were further used to train a logistic regression model. Another work, uses molecular activity, structure and phenotype data to construct a kernel function and correlate the drugs and diseases and then train on SVM classifier to make predictions [5]. [6] used a 3 layer heterogeneous network based model where nodes corresponded to drug, diseases and targets and edges between two nodes of same category were weighed with their similarity whereas those between two nodes of different category were based on relationship/association between them like drug-disease, drug-target, etc. Guilt-by-association principle was then used to find out the missing edges. [7] proposed a network based prioritization method using drug, disease and target information for predicting novel indications. [8] proposed an approach that uses heterogeneous networks embedding to characterize the drug-disease association and then trains an SVM for making predictions. [9] proposed a model which trains a dual convolutional neural network on two associations layers at the same time, one for encoding the drug-disease characteristics while the other for associating neighbourhood information. Another method [10], extracts feature descriptors from Gaussian interaction profile based and other similarities in drug and disease using auto-encoder and trained a RF classifier to predict indications. [11] uses aggregated neighbourhood information with drug disease associations and similarity matrices to train a neural network and minimise the loss while training it on the heterogeneous data between initial and recovered matrices. [12] used probabilistic matrix factorisation to classify drug-disease associations. [13] proposed integrating genomic space into matrix factorisation framework to exploit the molecular biological information to make predictions.

In recent years, deep learning has captured a lot of attention from researchers in bioinformatics and computational biology. [14] found out that Deep Neural Networks(DNN) can more effective in learning high level representations of proteins and drug structures to make better predictions than traditional approaches. Present studies have a common thing that they take high-dimensional ontology features directly as input to Deep learning models. Like, [15] encodes all the compounds by their chemical fingerprints and all the proteins by their structural descriptors and then concatenated both the features to represent drug-protein interactions. [16] also took the conjunction of drugs chemical fingerprints and protein domain attributes as input.

III. Data Sets

A. Description

We've used two drug–disease association benchmark data sets, called the PREDICT [4] and CDataset [17]. PREDICT includes 1933 high confidence associations between 593 drugs from DrugBank [18] and 313 diseases. The CDataset which includes 2352 known drug-disease associations between 663 unique drugs and 409 unique diseases is an expansion of PREDICT.

B. Processing

For every drug, simplified molecular-input line-entry system(SMILES) was obtained from the PubChem [19] Dataset to calculate chemical fingerprints. On the other hand, medical subject headings(MeSH) terms were extracted from Online Mendelian Inheritance in Man(OMIM) database [20].

IV. Methodology

In this section, a modification of Meta Embedding Deep Collaborative Filtering(MEDCF) [21] framework is described. Under this framework, we've described Generalized Matrix Factorization(GMF) and explored neural networks for Collaborative Filtering using Multilayer Perceptron(MLP). In the end, the neutral Matrix Factorization(NeuMF) method is used to integrate GMF and MLP models under a single framework. This is done to take into the account the linear interactions in GMF and non-linear interactions in MLP to model non-linear interactions using NeuMF.

A. Generalised Matrix Factorization(GMF)

Matrix Factorization is considered as a popular recommendation technique and is used in many factorization models. The input layer of the GMF model comprises of v_u^U and v_i^I which represents binarized one-hot encoded feature vector for drug u and disease i. Next, we apply embedding layer on vectors v_u^U and v_i^I to obtain the latent vector p_u and q_i for drug u and disease i respectively. The mapping function of GMF is described as the element-wise inner product between p_u and q_i vectors.

$$\phi(p_u, q_i) = p_u^T.q_i \tag{1}$$

The final result after applying multiplying with the weights of the output layer h and applying the activation function a_{out} is described as below.

$$\hat{y}_{ui}^G = a_{out}(h^T(p_u^T.q_i)) \tag{2}$$

We use the non-linear sigmoid activation function as a_{out} and Mean Squared Error(MSE) loss objective function. This is done to make our model more expressive than the linear MF models.

B. Multi Layer Perceptron(MLP)

We utilize the features extracted from the heterogeneous network of Drug and Disease Associations and concatenate the drug and disease latent vectors. The latent vectors, p_r' and q_s' are obtained by passing the drug and disease feature vectors, w_r^R and w_s^S, through a layer with linear activation function However, to provide a Collaborative Filtering effect, we add several hidden layers on top of the concatenated layer to learn

from the drug and disease latent vectors collaboratively. We summarize the structure of MLP model as follows.

$$z_1 = \phi_1(p'_r, q'_s) = \begin{bmatrix} p'_r \\ q'_s \end{bmatrix} \tag{3}$$

$$z_2 = \phi_2(z_1) = a_2(W_2^T z_1 + b_2) \tag{4}$$

$$z_L = \phi_L(z_{L-1}) = a_L(W_L^T z_{L-1} + b_L) \tag{5}$$

$$\hat{y}_{rs}^M = \sigma(h^T z_L) \tag{6}$$

where b_x, a_x, W_x denote the bias vector, activation function, weight matrix of the x-th layer of MLP, respectively. We use ReLU activation function because it is known to be a non-saturated function. It is used when the data is sparse and it prevents overfitting of data. The structure of MLP model is a hierarchical one where every successive layer has a fewer number of hidden neuron units than the immediate previous layer. The network is made deep to learn more complex interaction between the drug and disease features, and hence able to predict the associations better.

C. Neural Matrix Factorization(NeuMF) - The fusion of GMF and MLP

We discussed above that GMF allows us to model linear interactions between the drug and disease latent vectors, while MLP allows to model the non-linear interaction between the drug and disease latent vectors obtained from features from the drug-disease heterogeneous network. Here, we fuse the GMP and MLP models such that they compliment each other and form an enhanced model to predict the drug-disease association better. We presume the embedding layers of GMF and MLP and combine the output of their interaction functions. We use separate embedding for both MLP and GMF so that they learn from their respective features more effectively. Finally the outputs of the final hidden layers of MLP and GMF are concatenated as follows.

$$\phi^G = p_u^T . q_i, \tag{7}$$

$$\phi_M = a_L(W_L^T(a_{L-1}(\dots a_2(W_2^T \begin{bmatrix} p'_r \\ q'_s \end{bmatrix} + b_2)\dots)) + b_L), \tag{8}$$

$$\hat{y}_{ui} = ReLU(h^T \begin{bmatrix} \phi^G \\ \phi^M \end{bmatrix}). \tag{9}$$

The fused model with ReLU activation functions is called Neural Matrix Factorization(NeuMF), which jointly binds the linearity of GMF model and non-linearity of MLP to learn from known drug-disease associations and drug and disease features.

D. Drug-Disease Heterogeneous Network

The procedure followed is quite similar to the one followed in [22]

Algorithm 1 Predicting association using modified MEDCF

procedure GMF(v_u^U, v_i^I)
for each v_u^U, v_i^I **do**
Apply embedding layer on v_u^U and v_i^I to give p_u and q_i, respectively
Apply GMF layer to compute $\hat{y}_{ui}^G \leftarrow p_u^T . q_i$
end for
end procedure
procedure MLP(w_r^R, w_s^S)
for each w_r^R, w_s^S **do**
Apply linear layer on w_r^R and w_s^S to give p'_u and q'_i, respectively
Apply final MLP layer to compute $\hat{y}_{rs}^M$
$\sigma(h^{T(}a_L(W_L^T(a_{L-1}(\dots a_2(W_2^T \begin{bmatrix} p'_r \\ q'_s \end{bmatrix} + b_2)\dots)) + b_L)))$
end for
end procedure
procedure NEUMF(w_r^R, w_s^S)
$\phi^G = p_u^T . q_i$,
$\phi_M = a_L(W_L^T(a_{L-1}(\dots a_2(W_2^T \begin{bmatrix} p'_r \\ q'_s \end{bmatrix} + b_2)\dots)) + b_L)$,
Apply NeuMF layer to merge GMF and MLP output to compute final association as:
$\hat{y}_{ui} = ReLU(h^T \begin{bmatrix} \phi^G \\ \phi^M \end{bmatrix})$.
end procedure
return $\hat{y}_{ui}$ =0

1) Drug Disease Association Network: We construct the drug-disease association network using verified associations between the set of drugs and diseases, which are manually curated by the biocurators. Let C = $c_1, c_2, \dots, c_A$ denote the set of A drugs, and D = $d_1, d_2, \dots, d_B$ denote the set of B diseases. Let A be the adjacency matrix of the drug disease association network where element A_{ij} is 1 if drug i is associated with disease j, and 0 otherwise.

2) Drug-Drug Similarity Network: As mentioned above, SMILES was obtained from the PubChem Dataset to calculate the chemical fingerprint for every drug. Technically, SMILES is a specification describing structure of chemical species using short ASCII strings. Using SMILES as input, an 881-dimensional binary vector is obtained, representing 881 types of chemical fingerprints. An element is labelled 1 if the corresponding fingerprint is contained in the drug, and 0 otherwise. This is similar to tf-IDF approach. To calculate the drug-drug similarity, we calculate the jaccard similarity between chemical fingerprint vector of two drugs. The jaccard similarity is defined as the ratio of the number of common footprints of both the drug and their total number of footprints.

$$S_{ij}^{(c)} = \frac{|c_i \cap c_j|}{|c_i \cup c_j|} \tag{10}$$

where $S_{ij}^{(c)}$ denotes the jaccard similarity and c_i and c_j denotes the chemical footprint vector of drug i and drug j respectively.

3) Disease-Disease Similarity Network: The disease-disease similarity is calculated by considering the phenotype characteristics of diseases. We use the MimMiner [23] algorithm which characterizes each disease related phenotype by a vector of MeSH concepts extracted from the OMIM database. Further, the cosine similarity between the MeSH concept vectors of two diseases is calculated. Let $d_i = d_{i1}, d_{i2}, \ldots, d_{iK}$ and $d_j = d_{j1}, d_{j2}, \ldots, d_{jK}$ be the MeSH concept vectors of disease i and disease j. The cosine similarity S^c_{ij} based of phenotype is calculated as

$$S^c_{ij} = \frac{\sum_{k=1}^{K} d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^{K} d^2_{ik}} \cdot \sqrt{\sum_{k=1}^{K} d^2_{jk}}} \tag{11}$$

where K is the size of unique MeSH terms.

4) Feature Extraction from Heterogeneous Network: After computing the drug-drug and disease-disease similarities, the next step is to consider the drug-disease association between drugs and diseases to extract features from the heterogeneous structure. The feature vector for drug c_i consists of following two components.

1) A_i corresponds to the vector of known associations between drug c_i and disease set D. It is i-th the row of the Graph Adjacency Matrix
2) $S^{(c)}_i$ corresponds to the drug-drug similarities between drug c_i and the drug set C.

The final feature vector F^{c_i} is defined as.

$$F^{c_i} = \begin{bmatrix} A_i \\ S^{(c)}_i \end{bmatrix} \tag{12}$$

Similarly, the feature vector for disease d_j constitutes the following two components.

1) A_j corresponds to the vector of known associations between disease d_j and drug set C. It is the j-th row of the Graph Adjacency Matrix
2) $S^{(d)}_j$ corresponds to the drug-drug similarities between disease d_i and the disease set D.

The final feature vector F^{d_i} is defined as.

$$F^{c_i} = \begin{bmatrix} A_i \\ S^{(c)}_i \end{bmatrix} \tag{13}$$

For clarity, the process of extracting features is describe in Figure 1 One thing to be noted here is that the drug-disease associations in the test set are not included in the drug-disease heterogeneous network so that the features extracted contain no information about the associations in the test set.

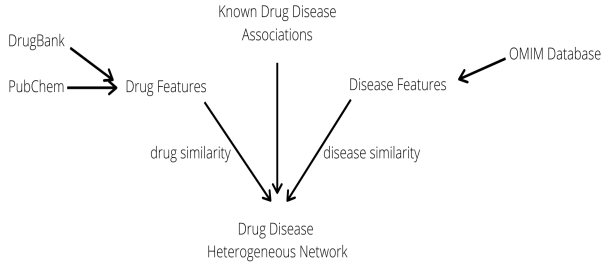

Fig. 1. Flowchart of drug-disease heterogeneous network

E. Training GMF, MLP and NeuMF models

Firstly, we train the GMF model on the train set and save the weights with latent vector size equal to 100. The second step consists of training the MLP model with latent vector size to be 32 and saving the weights. Finally, to train the NeuMF model, we load the saved pre-trained weights from GMF and MLP models, concatenate the output layer of both models, pass it through a layer with linear activation function. The size of the output layer is 1, which indicates the likelihood of the drug i being associated with disease j. The sigmoid(inverse-logit) activation function is used at the output layer.
Training the GMF, MLP and NeuMF models requires to minimize the difference between the predicted class and the actual class. We measure this difference in terms of Mean Squared Error(MSE) loss function and use optimizer Adam [24] to train the model(Backpropagation). We use the weight penalty coefficient under the L2 norm to regularize the weights. In order to prevent overfitting, we've used the dropout [25] technique with probability 0.1. The proposed framework is as shown in figure 2

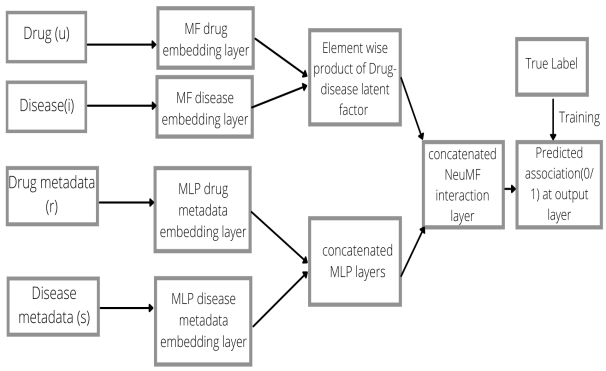

Fig. 2. Modified MEDCF framework

V. Experimental Results

We tested our models on the above-mentioned datasets, PREDICT and CDataset. Both the datasets were partitioned in 3 subsets, 80% of the data was used for training, 10% as validation and the remaining 10% was used as test set. We decompose the dataset such that each drug in the train set has at least one association with a certain disease.

1) Evaluation Metrics: We evaluated our model HNet-NeuMF using evaluation metrics of Precision, Recall, Accuracy, F1-score and Area under the Receiver Operating Curve(AUC).

Precision(Pre) = $\frac{TP}{TP+FP}$

TABLE I
HYPERPARAMETERS

Hyperparameter	Value
Weight Decay	1e-6
Dropout Rate	0.1
Epochs	100
# of Hidden Layers(MLP)	4
Act Func(Output Layer)	Sigmoid
Act Func(Hidden Layers)	ReLU

TABLE II
PERFORMANCE COMPARISON ON PREDICT BETWEEN HNET-NEUMF, HNET-DNN, SVM, RF METHODS

methods	Acc	Pre	Rec	F1	AUC
HNet-NeuMF	0.9425	0.9253	0.958	0.941	0.943
HNet-DNN	0.874	0.891	0.859	0.875	0.930
SVM	0.721	0.729	0.720	0.729	0.722
RF	0.825	0.820	0.844	0.832	0.889

Recall(Rec) = $\frac{TP}{TP+FN}$

Accuracy(Acc) = $\frac{TP+TF}{TP+FP+TN+FN}$

F1 Score = $\frac{2*Pre*Rec}{Pre+Rec}$

where TP stands for True Positives, TN stands for True Negatives, FP stands for False Positives and FN stands for True Negatives.

2) Hyperparameters: We've used 10 Fold Cross Validation to evaluate our models on both the datasets.The results are reported for the best performing fold. The value of the hyperparameters used in mentioned in table I.

The sizes of the hidden layers of MLP are kept as 64, 32, 16, and 8. The input size for MLP in case of PREDICT dataset is 1812 (906 features, each of drug and disease), while the same is 2144 (1072 features, each of drug and disease) for the CDataset.

3) Evaluation on PREDICT Dataset: We split the PREDICT dataset into 8:1:1 ratio of train, validation and test set. We evaluate our performance using 10 fold Cross Validation on combined test and validation set and compare the performance on the best performing fold with typical classifiers like Support Vector Machines(SVM) and Random Forests(RF) and our baseline method HNet-DNN [22].

We perform significantly better than the typical classifiers SVM and RF in terms of AUC, Accuracy, Precision, Recall and F1 Score. We also outperform our baseline method HNet-DNN in terms of AUC, Accuracy, Recall, and F1 Score. The performance comparison can be viewed in table II

4) Evaluation of CDataset: Similar to PREDICT, we split the CDataset into 8:1:1 ratio of train, validation and test set. On evaluating our performance on 10 Fold Cross Validation, we found that HNet-NeuMF outperforms the SVM and RF classifiers in terms of all the evaluation metrics for the best fold. It also performs better than the baseline HNet-DNN method, except in terms of Recall. The values of the evaluation metrics are compared in table III.

We also compared our performance with the existing DrugNet [7]and HGBI [5] methods in terms of AUC. DrugNet [7] and HGBI [5] reported AUC values of 0.804 and 0.856 respectively, while HNet reported AUC value of 0.9034, clearly outperforming the other two methods.

We evaluate our method on unseen test data whose drug-disease associations are excluded from the heterogeneous network. We outperform the baseline method HNet-DNN there as well. HNet-DNN reports the accuracy of 0.52, while HNet-NeuMF gives 0.799 as accuracy. The AUC of HNet-NeuMF and HNet-DNN is 0.8638 and 0.525 respectively. The ROC and Precision-Recall curves are presented in figure 3 and 4 respectively.

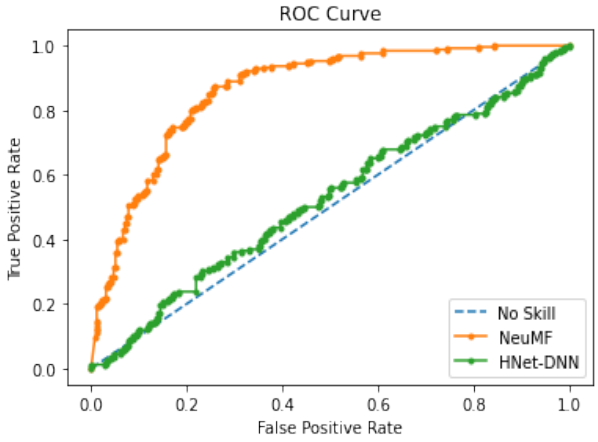

Fig. 3. Performance of HNet-NeuMF and HNet-DNN on Test Set: ROC curve

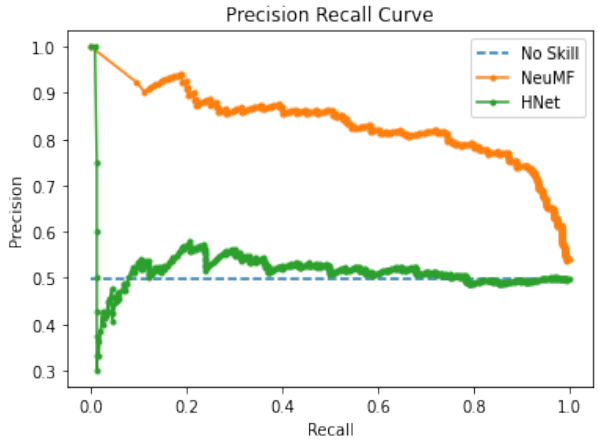

Fig. 4. Performance of HNet-NeuMF and HNet-DNN on Test Set: Precision Recall curve

TABLE III
PERFORMANCE COMPARISON ON CDATASET BETWEEN HNET-NEUMF, HNET-DNN, SVM, RF METHODS

methods	Acc	Pre	Rec	F1	AUC
HNet-NeuMF	0.9342	0.9122	0.974	0.937	0.935
HNet-DNN	0.899	0.888	0.913	0.900	0.946
SVM	0.775	0.769	0.782	0.776	0.775
RF	0.860	0.844	0.881	0.862	0.915

VI. CONCLUSION AND OUR CONTRIBUTION

In this paper, we have proposed to use the drug-disease association information and features extracted from the heterogeneous network constructed using chemical structure of drugs and phenotype information about disease to predict new associations between drugs and diseases. The experimental results have shown that our proposed method HNet-NeuMF not only performs better than the baseline method HNet-DNN, but also outperforms the typical classifiers SVM and RF as well as the pioneer methods like DrugNet and HGBI. DrugNet and HGBI predict new drug disease associations using bi-random walk algorithm.

Several studies have applied Neural Matrix Factorization under the Meta Embedding Deep Collaborative Filtering(MEDCF) framework to predict the user-item interactions to recommend items to users. This framework directly utilizes the metadata and information about the user-item interactions to produce better recommendation results. In this paper, we've proposed a modification of MEDCF framework to utilize the features from drug-disease heterogeneous network in order to predict drug-disease associations. We constructed the drugdisease heterogeneous network based on the chemical structure of drugs and phenotypic characteristics of diseases and then extracted relatively low-dimensional topological features from the heterogeneous network to train the MLP and consequently the NeuMF models. By using the side-information about drugs and diseases, we're able to effectively reduce the overfitting problem.

REFERENCES

[1] W. P. Walters, J. Green, J. R. Weiss, and M. A. Murcko, "What do medicinal chemists actually make? a 50-year retrospective," *Journal of medicinal chemistry*, vol. 54, no. 19, pp. 6405–6416, 2011.

[2] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[3] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," *PloS one*, vol. 4, no. 8, p. e6536, 2009.

[4] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "Predict: a method for inferring novel drug indications with application to personalized medicine," *Molecular systems biology*, vol. 7, no. 1, p. 496, 2011.

[5] Y. Wang, S. Chen, N. Deng, and Y. Wang, "Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data," *PloS one*, vol. 8, no. 11, p. e78518, 2013.

[6] 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, 2014.

[7] V. Martinez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "Drugnet: network-based drug–disease prioritization by integrating heterogeneous data," *Artificial intelligence in medicine*, vol. 63, no. 1, pp. 41–49, 2015.

[8] K. Yang, X. Zhao, D. Waxman, and X.-M. Zhao, "Predicting drug-disease associations with heterogeneous network embedding," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 12, p. 123109, 2019.

[9] P. Xuan, H. Cui, T. Shen, N. Sheng, and T. Zhang, "Heterodualnet: A dual convolutional neural network with heterogeneous layers for drug-disease association prediction via chou's five-step rule," *Frontiers in pharmacology*, vol. 10, p. 1301, 2019.

[10] H.-J. Jiang, Y.-A. Huang, and Z.-H. You, "Predicting drug-disease associations via using gaussian interaction profile and kernel-based autoencoder," *BioMed research international*, vol. 2019, 2019.

[11] Y. Wang, G. Deng, N. Zeng, X. Song, and Y. Zhuang, "Drug-disease association prediction based on neighborhood information aggregation in neural networks," *IEEE Access*, vol. 7, pp. 50 581–50 587, 2019.

[12] J. Yang, Z. Li, X. Fan, and Y. Cheng, "Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization," *Journal of chemical information and modeling*, vol. 54, no. 9, pp. 2562–2569, 2014.

[13] W. Dai, X. Liu, Y. Gao, L. Chen, J. Song, D. Chen, K. Gao, Y. Jiang, Y. Yang, J. Chen *et al.*, "Matrix factorization-based prediction of novel drug indications by integrating genomic space," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[14] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, "Deepppi: boosting prediction of protein–protein interactions with deep neural networks," *Journal of chemical information and modeling*, vol. 57, no. 6, pp. 1499–1510, 2017.

[15] M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, and H. Lu, "Deep-learning-based drug–target interaction prediction," *Journal of proteome research*, vol. 16, no. 4, pp. 1401–1409, 2017.

[16] K. Tian, M. Shao, Y. Wang, J. Guan, and S. Zhou, "Boosting compound-protein interaction prediction by deep learning," *Methods*, vol. 110, pp. 64–72, 2016.

[17] 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.

[18] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda *et al.*, "Drugbank 5.0: a major update to the drugbank database for 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[19] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu *et al.*, "Pubchem 2019 update: improved access to chemical data," *Nucleic acids research*, vol. 47, no. D1, pp. D1102–D1109, 2019.

[20] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 30, no. 1, pp. 52–55, 2002.

[21] R. Nahta, Y. K. Meena, D. Gopalani, and G. S. Chauhan, "Embedding metadata using deep collaborative filtering to address the cold start problem for the rating prediction task," *Multimedia Tools and Applications*, pp. 1–29, 2021.

[22] H. Liu, W. Zhang, Y. Song, L. Deng, and S. Zhou, "Hnet-dnn: Inferring new drug–disease associations with deep neural network based on heterogeneous network features," *Journal of chemical information and modeling*, vol. 60, no. 4, pp. 2367–2376, 2020.

[23] M. A. Van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen, "A text-mining analysis of the human phenome," *European journal of human genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.