International Journal of
Molecular Sciences

MDPI

Article

Prediction of Potential Drug–Disease Associations through Deep Integration of Diversity and Projections of Various Drug Features

Ping Xuan [1], Yingying Song [1], Tiangang Zhang [2,*] and Lan Jia [1]

[1] School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China
[2] School of Mathematical Science, Heilongjiang University, Harbin 150080, China
* Correspondence: zhang@hlju.edu.cn; Tel.: +86-188-4503-0636

Received: 2 July 2019; Accepted: 20 August 2019; Published: 22 August 2019

Abstract: Identifying new indications for existing drugs may reduce costs and expedites drug development. Drug-related disease predictions typically combined heterogeneous drug-related and disease-related data to derive the associations between drugs and diseases, while recently developed approaches integrate multiple kinds of drug features, but fail to take the diversity implied by these features into account. We developed a method based on non-negative matrix factorization, DivePred, for predicting potential drug–disease associations. DivePred integrated disease similarity, drug–disease associations, and various drug features derived from drug chemical substructures, drug target protein domains, drug target annotations, and drug-related diseases. Diverse drug features reflect the characteristics of drugs from different perspectives, and utilizing the diversity of multiple kinds of features is critical for association prediction. The various drug features had higher dimensions and sparse characteristics, whereas DivePred projected high-dimensional drug features into the low-dimensional feature space to generate dense feature representations of drugs. Furthermore, DivePred's optimization term enhanced diversity and reduced redundancy of multiple kinds of drug features. The neighbor information was exploited to infer the likelihood of drug–disease associations. Experiments indicated that DivePred was superior to several state-of-the-art methods for prediction drug-disease association. During the validation process, DivePred identified more drug-disease associations in the top part of prediction result than other methods, benefitting further biological validation. Case studies of acetaminophen, ciprofloxacin, doxorubicin, hydrocortisone, and ampicillin demonstrated that DivePred has the ability to discover potential candidate disease indications for drugs.

Keywords: drug–disease association; non-negative matrix factorization; projections of drug features; diversity representation; specific features of different drug views

1. Introduction

Developing a new drug is a complex, time-consuming, and expensive process [1,2], which typically proceeds through preliminary compound testing, pre-clinical and animal experiments, clinical research, and Food and Drug Administration (FDA) review, before it finally yields a new drug that reaches the market after 10–15 years, costing approximately 0.8–1.5 billion dollars [3–6]. Even with a substantial time commitment and capital investment, the successful development of a new drug is still associated with considerable risks [1,7,8]. Because the number of new drugs approved by the FDA has been declining since the 1990s [9,10], there is an urgent need to find alternative approaches that will reduce the development costs. Drug repositioning refers to the identification of new indications for drugs that have been approved by regulatory agencies. Compared to the development of a new drug for a

certain indication, drug repositioning can shorten the drug development cycle to 6.5 years at the cost of approximately 0.3 billion dollars due to the known safety, tolerability, and efficacy profile of the drug candidate [11–13].

Computational prediction of new drug-related disease annotations can generate reliable drug–disease association candidates for further validation [14,15]. Previous prediction methods can be broadly divided into two categories. In first the category, the potential associations between drugs and diseases are usually related to shared target genes, and the more shared target genes there are, the higher the likelihood of a drug–disease association is. Therefore, several methods for predicting the association of drugs with diseases based on related target genes or gene expression profiles have been proposed [16,17]. Similarly, the possibility of a drug–disease association can be estimated based on the targeted protein complexes shared by the drugs and diseases [18] and the perturbed genes they have in common [19]. However, these methods are limited to drugs and diseases with shared genes or proteins.

The second category uses a variety of data types, including drug similarity, disease similarity, and target similarity, as well as interactions and association between drugs, targets, and diseases for drug repositioning. Wang et al. applied a kernel function to integrate similarity information drugs and diseases to predict potential drug–disease associations [20]. Several approaches integrate the information on drugs, targets, and diseases to create heterogeneous networks that infer drug candidates by information flow or random walks [21–24]. Some methods use the data of drugs and diseases to infer drug–disease association candidates using the logistic regression model [25], a statistical model [26], Laplacian regularized sparse subspace learning model [27], similar constraint matrix decomposition model [28] or non-negative matrix factorization model [29]. These methods include information from different sources and confirm that this information is important for predicting associations between drugs and diseases. However, multiple kinds feature of drugs, such as the chemical substructures and the target protein domains have diversity, and these methods did not take the diversity into account.

In this study, we present a new method, DivePred, for predicting potential drug–disease associations. DivePred deeply integrates not only the projection of multiple drug features in low-dimensional space but also the diversity of drug features. Projecting multiple high-dimensional drug features into the same dimension as the disease assists in measuring the distance between the drugs and the diseases, which is a critical parameter for the possibility of a drug–disease association. The chemical substructures of the drugs, the target protein domains, and the ontology annotation of the target gene, along with its associated disease annotations reflect the characteristics of the drugs from different perspectives. Therefore, retaining the diversity of multiple drug features can fully integrate information from different drug views. Thus, we created a unified model and developed an iterative optimization algorithm to derive drug–disease association scores. Experimental results based on cross-validation indicated that DivePred achieved better prediction performance than several state-of-the-art methods. Case studies of five drugs further demonstrated that DivePred could detect potential drug-related diseases.

2. Experimental Evaluation and Discussion

2.1. Evaluation Metrics

We used five-fold cross-validation to evaluate the performance of DivePred in predicting potential drug–disease associations. The known drug–disease associations were randomly divided into five equal subsets, four of which were used to train our model, while the remaining set was used to perform the test. In each cross-validation, $X^{(4)}$ contained only the drug–disease associations of the training set, and R_4 was calculated based on the known associations in matrix $X^{(4)}$. For a certain drug $r_i (1 \leq r_i \leq N_r)$, its associated diseases in the test set was called the positive sample, and the other unmarked diseases were called negative samples. In the test results, a high positive sample rate of drug r_i was correlated with an improved predictive performance for this drug.

A threshold θ was set, and when the score obtained by the sample estimate was higher than θ, it was identified as a positive example; otherwise, it was identified as a negative example. The *TPRs* (true-positive rates) and the *FPRs* (false-positive rates) under various θ can be calculated as follows,

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{TN + FP} \quad (1)$$

where TP is the number of positive cases that were correctly identified, and TN indicates the number of negative examples that were correctly identified. FN and FP are the numbers of positive and negative examples that were misidentified, respectively. After calculating *TPRs* and *FPRs* for different θ values, the receiver operating characteristic curve (ROC) was be plotted. The area under the curve (AUC) was used as a measure to predict the performance of potentially associated disease with drug r_i. The overall performance of the prediction method was the average of the AUC values of all drugs.

Due to the imbalance of the number of positive and negative samples in the sample data, the precision–recovery rate (P–R curve) can provide additional information; precision and recall were defined as follows,

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN} \quad (2)$$

The precision ratio refers to the proportion of correctly identified positive samples in the search samples, and the recall rate is the same as the TPR. The area under the P–R curve (AUPR) was also used to measure the performance for predicting potential drug–disease associations.

Biologists typically choose the top-ranked candidates for further experimentation. It was our goal to increase the number of positive samples in the top-ranked section. To create another evaluation index, we calculated the recall rate of the top-ranked samples, which is the proportion of positive samples correctly identified in the top k of the list among the total of positive samples.

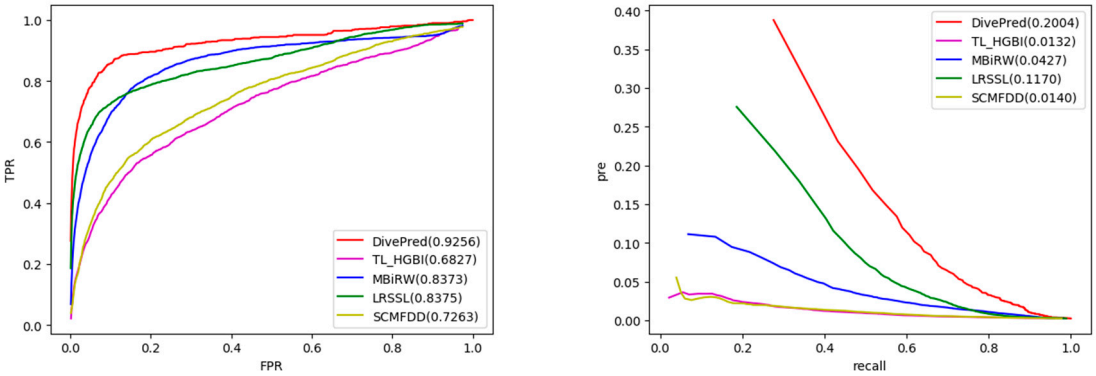
2.2. Comparison with Other Methods

To evaluate the performance of our prediction method, DivePred, we also compared it with several state-of-the-art methods for predicting potential drug–disease associations, including: TL_HGBI [21], MBiRW [22], LRSSL [27], and SCMFDD [28]. In our method of comparison, we need to fine-tune the hyperparameters. Based on five-fold cross-validation, we selected the hyperparameters values for α_1, α_2, α_3, α_4 and α_5 in DivePred from as $\{10^{-2}, 10^{-1}, 1, 10, 100\}$. DivePred achieved the best performance at $\alpha_1 = 1$, $\alpha_2 = 10$, $\alpha_3 = 0.1$, $\alpha_4 = 0.1$, and $\alpha_5 = 0.1$. To perform a fair comparison with the four other methods, we used the best value provided by the authors to set the hyperparameters (i.e., $\alpha = 0.4$ and $\beta = 0.3$ for TL_HGBI; $\alpha = 0.3$, $l = 2$ and $r = 2$ for MBiRW; $\mu = 0.01$, $\lambda = 0.01$, $\gamma = 2$, and $k = 10$ for LRSSL; $k = 45\%$, $\mu = 1$ and $\lambda = 4$ for SCMFDD).

As shown in Figure 1a, DivePred achieved the best average performance, on a set of 763 drugs (AUC = 0.9256). Specifically, the performance score of DivePred was 24.29% better than that of the TL_HGBI algorithm, 8.83% better than the MBiRW algorithm, 8.81% better than the LRSSL algorithm, and 19.93% better than the SCMFDD algorithm. In addition, we tested 15 drugs using DivePred and the other four methods. The AUC values of the 15 drugs are shown in Table 1, DivePred preforms the best on 12 of these drugs. Among these comparison methods, LRSSL achieved a good performance because similar to DivePred, it considers the information on multiple drug features, although it does not consider the diversity of multiple feature information of the drugs. The MBiRW algorithm only considers a feature of the drugs, limiting its performance. The SCMFDD algorithm and TL_HGBI algorithm were relatively poor. The weak performance of the former might be due to the excessive dependence on the accuracy of similarity calculations; the latter may have problems due to the introduction of noise when calculating drug–drug similarity. Compared with those methods, DivePred was superior to those methods because it captures the specific features of each aspect of the drugs.

As shown in Figure 1b, the average PR curve of 763 drugs was higher for DivePred than those for the other methods, indicating that DivePred has the best performance for drug–disease association

prediction (AUPR = 0.2004). Compared with the AUPR values of SCMFDD, TL_HGBI, MBiRW, and LRSSL, the DivePred values were 18.7%, 15.8%, 8.3%, and 18.6% higher, respectively. The AUPR values of the 15 drugs are shown in Table 2, and DivePred is the best performer on 10 of these drugs.



(**a**) ROC curves (**b**) P–R curves

Figure 1. Two types of curves for evaluating the predicting performance of DivePred and other methods. (**a**) receiver operating characteristic (ROC) curves; (**b**) precision–recall (P–R) curves.

Table 1. Area under ROC curve (AUC) values of 15 drugs using DivePred and other methods.

Drug Name	AUC DivePred	TL_HGBI	MBiRW	LRSSL	SCMFDD
ampicillin	**0.944**	0.751	0.932	0.962	0.895
cefepime	**0.976**	0.910	0.970	0.971	0.914
cefotaxime	**0.992**	0.917	0.929	0.950	0.953
cefotetan	**0.996**	0.808	0.918	0.948	0.848
cefoxitin	**0.979**	0.890	0.912	0.979	0.894
ceftazidime	**0.985**	0.845	0.931	0.936	0.922
ceftizoxime	0.797	0.960	0.961	0.923	**0.962**
ceftriaxone	0.907	0.945	0.898	**0.955**	0.811
ciprofloxacin	**0.957**	0.811	0.813	0.928	0.820
doxorubicin	**0.949**	0.487	0.921	0.727	0.460
erythromycin	**0.962**	0.827	0.887	0.918	0.764
itraconazole	**0.952**	0.445	0.877	0.845	0.730
levofloxacin	**0.975**	0.943	0.975	0.964	0.872
moxifloxacin	0.794	0.812	0.948	0.957	0.932
ofloxacin	**0.958**	0.902	0.943	0.904	0.774
Average AUC	**0.926**	0.683	0.837	0.838	0.726

The bold values indicate the higher AUCs.

Table 2. Area under precision–recall curve (AUPR) values of 15 drugs using DivePred and other methods.

Drug Name	AUPR DivePred	TL_HGBI	MBIRW	LRSSL	SCMFDD
ampicillin	**0.189**	0.032	0.023	0.285	0.068
cefepime	**0.744**	0.163	0.315	0.625	0.054
cefotaxime	**0.770**	0.071	0.292	0.283	0.105
cefotetan	**0.486**	0.054	0.197	0.512	0.059
cefoxitin	**0.580**	0.151	0.394	0.286	0.065
ceftazidime	0.675	0.032	0.201	0.488	**0.694**
ceftizoxime	**0.647**	0.212	0.244	0.455	0.096
ceftriaxone	0.409	0.056	0.223	**0.673**	0.077
ciprofloxacin	**0.425**	0.082	0.118	0.280	0.064
doxorubicin	0.164	0.005	0.051	**0.180**	0.004
erythromycin	**0.425**	0.023	0.038	0.144	0.022

Table 2. *Cont.*

Drug Name	AUPR DivePred	TL_HGBI	MBIRW	LRSSL	SCMFDD
itraconazole	0.188	0.006	**0.253**	0.042	0.008
levofloxacin	0.504	0.136	0.071	**0.539**	0.098
moxifloxacin	**0.565**	0.049	0.065	0.384	0.088
ofloxacin	**0.378**	0.091	0.130	0.201	0.078
Average AUC	**0.200**	0.013	0.043	0.117	0.014

The bold values indicate the higher AUPRs.

We evaluated the prediction results of 763 drugs by using a Wilcoxon test, and the results of the evaluation showed that DivePred was significantly better than other methods. These results were observed using a *p*-value threshold of 0.05, with DivePred showing better performance in terms of not only AUCs of ROC curves but AUCs of P–R curves as well (Table 3).

Table 3. Results of Wilcoxon test on DivePred and four other contrast methods for 763 drugs.

p-Value Between DivePred and Another Method	TL_HGBI	MBiRW	LRSSL	SCMFDD
p-value of ROC curve	5.631×10^{-42}	7.181×10^{-156}	3.735×10^{-78}	6.596×10^{-73}
p-value of PR curve	1.332×10^{-21}	2.635×10^{-32}	1.562×10^{-16}	8.452×10^{-29}

In addition, the recall rates for the top *k* candidate diseases were assessed. A high recall rate for the top *k* candidate diseases indicated that the predictive method performed well in identifying diseases that are truly associated with a drug. The average recall rates of all 763 drugs at different top *k* values are shown in Figure 2. DivePred was always superior to the other methods in the range for of the top 30 to the top 240 candidates. Among the top 30, 90, and 150 candidate diseases, the recall rates for which were 74.6%, 87.4%, and 90.0%, respectively; the second-best method was LRSSL, where the recall rate was 63.4% in the top 30, 75.2% in the top 90, and 79.6% in the top 150; followed by MBiRW, for which the recall rates among the top 30, 90, and 150 candidates were 52.9%, 74.2%, and 82.6%, respectively; the worst performers were TL_HGBI and SCMFDD. Their recall rates were relatively close. For the former method, the recall rates were 28.8%, 49.6%, and 58.5% among the top 30, 90, and 150 candidate diseases, respectively. The recall rates for the latter method, SCMFDD, were 30.6%, 52.5%, 62.1% in the top 30, 90, and 150 respectively.

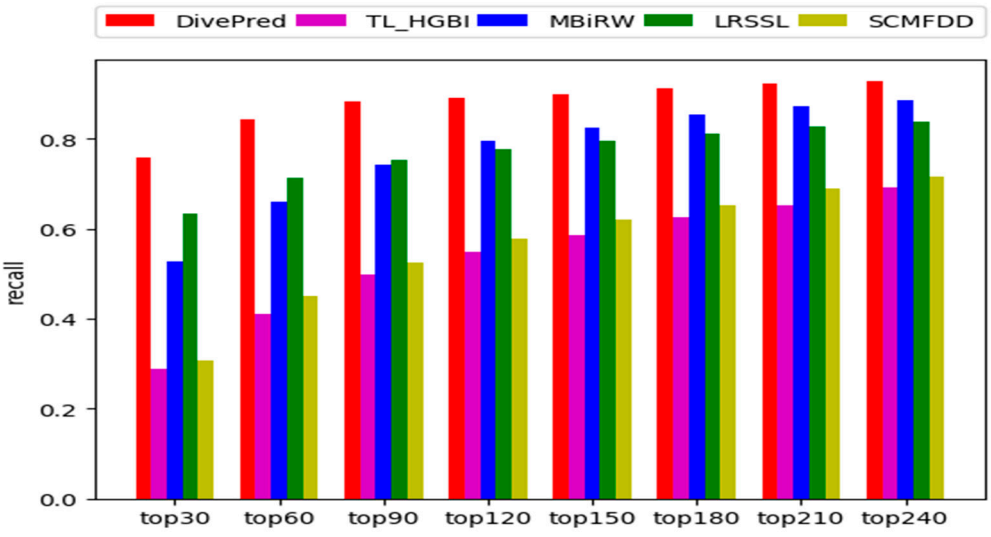


Figure 2. Average recall rates of all drugs at different top *k*.

2.3. Case Studies on Five Drugs

To further demonstrate the ability of DivePred to discover candidate diseases for drugs, we conducted case studies on five drugs, including acetaminophen, ciprofloxacin, doxorubicin, hydrocortisone, and ampicillin. For each of the five drugs, we scored the drug–disease association predictions and ranked them accordingly. The top 15 diseases with the highest association scores were considered candidate diseases for the drug. A total of 75 candidate diseases were predicted, as shown in Table 4.

Table 4. The top 15 related candidate diseases for acetaminophen, ciprofloxacin, doxorubicin, hydrocortisone, and ampicillin.

Drug Name	Rank	Disease Name	Description	Rank	Disease Name	Description
Acetaminophen	1	Osteoarthritis	CTD	9	Arthritis	DrugBank
	2	Arthritis, Rheumatoid	CTD	10	Pain, Postoperative	CTD
	3	Inflammation	CTD	11	Rheumatic Fever	PubChem
	4	Dysmenorrhea	inferred candidate by 1 literature	12	Arthritis, Gouty	CTD
	5	Arthritis, Juvenile Rheumatoid	DrugBank	13	Premenstrual Syndrome	DrugBank
	6	Gout	DrugBank	14	Menorrhagia	unconfirmed
	7	Spondylitis, Ankylosing	Clinicaltrials	15	Rheumatic Diseases	Clinicaltrials
	8	Bursitis	literature [30]			
Ciprofloxacin	1	Salmonella Infections	CTD	9	Pyelonephritis	CTD
	2	Streptococcal Infections	DrugBank	10	Bacterial Infections	CTD
	3	Bronchitis	CTD	11	Serratia Infections	DrugBank
	4	Pneumonia, Bacterial	CTD	12	Tuberculosis, Pulmonary	CTD
	5	Chlamydia Infections	CTD	13	Plague	CTD
	6	Gram-Negative Bacterial Infections	CTD	14	Brucellosis	PubChem
	7	Enterobacteriaceae Infections	CTD	15	Chlamydiaceae Infections	PubChem
	8	Soft Tissue Infections	CTD			
Doxorubicin	1	Leukemia, Myeloid, Acute	CTD	9	Rhabdomyosarcoma	CTD
	2	Precursor Cell Lymphoblastic Leukemia-Lymphoma	CTD	10	Histiocytosis	Clinicaltrials
	3	Carcinoma, Non-Small-Cell Lung	PubChem	11	Trophoblastic Neoplasms	DrugBank
	4	Mycosis Fungoides	PubChem	12	Stomach Neoplasms	CTD
	5	Leukemia, Lymphocytic, Chronic, B-Cell	inferred candidate by 14 literatures	13	Hodgkin Disease	CTD
	6	Head and Neck Neoplasms	CTD	14	Melanoma	CTD
	7	Sarcoma, Kaposi	CTD	15	Leukemia, Myelogenous, Chronic, BCR-ABL Positive	DrugBank
	8	Leukemia, Lymphoid	CTD			

Table 4. *Cont.*

Drug Name	Rank	Disease Name	Description	Rank	Disease Name	Description
Hydrocortisone	1	Asthma	CTD	9	Shock, Septic	CTD
	2	Rhinitis, Allergic, Perennial	DrugBank	10	Acne Vulgaris	unconfirmed
	3	Dermatitis	PubChem	11	Rosacea	CTD
	4	Skin Diseases	CTD	12	Addison Disease	CTD
	5	Pruritus	PubChem	13	Hyperhidrosis	literature [31]
	6	Keratosis	inferred candidate by 1 literature	14	Hematologic Diseases	inferred candidate by 1 literature
	7	Hypersensitivity	inferred candidate by 7 literatures	15	Pityriasis Rosea	unconfirmed
	8	Psoriasis	PubChem			
Ampicillin	1	Proteus Infections	CTD	9	Osteomyelitis	Clinicaltrials
	2	Streptococcal Infections	CTD	10	Impetigo	unconfirmed
	3	Septicemia	DrugBank	11	Serratia Infections	CTD
	4	Pneumonia, Bacterial	CTD	12	Peritonitis	CTD
	5	Bone Diseases, Infectious	PubChem	13	Bacterial Infections	CTD
	6	Staphylococcal Skin Infections	DrugBank	14	Enterobacteriaceae Infections	DrugBank
	7	Wound Infection	CTD	15	Cellulitis	CTD
	8	Pseudomonas Infections	PubChem			

Comparative Toxicogenomics Database (CTD) is a powerful public database that provides relevant drugs information and the effects of drugs on diseases; this information is compiled from published literatures. DrugBank database is supported by the Canadian Institutes of Health Research, the Alberta Innovats-Health Solutions and the Metabolomics Innovation Centre. It provides clinical trial information on the drugs, including the drugs and the diseases being tested. PubChem is an open chemical database supported by the National Institutes of Health (NIH), which contains from various data sources with many informational entries on drugs and diseases. As shown in Table 4, 38 drug–disease association information were included in the CTD, 12 association information were contained in the DrugBank, and 10 association information were recorded by PubChem, indicating that these candidate diseases are indeed associated with the corresponding drugs.

Secondly, ClinicalTrials.gov (https://clinicaltrials.gov/) is an online clinical trial database managed by the National Library of Medicine (NLM) and the Food and Drug Administration (FDA), which contains a large amount of clinical research information on various drugs and diseases. Four drug–disease association predictions matched entries in the ClinicalTrials database. In addition, two candidates were labelled with "literature", indicating that there is literature supporting that the candidate disease is being treated with the corresponding drug.

In addition, the CTD database also contains potential associations from literature data, which we included as "inferred candidate by *k* literatures", where *k* represents the number of documents reporting that a drug that could be associated with a disease according to the CTD. A total of five candidates were tagged, indicating that this drug is more likely to be associated with the corresponding disease candidates. Of the 75 candidates, four could not be confirmed by observational evidence; they were labelled as "unconfirmed".

2.4. Prediction of Novel Drug–Disease Associations

After evaluating its prediction performance by cross-validation, case studies, and the Wilcoxon test, we applied DivePred to predict novel drug–disease associations. All the known drug–disease

associations were utilized to train DivePred's prediction model. High-confidence candidate diseases of drugs were obtained using DivePred. Results are listed in supplementary Table ST1_candidates.

3. Materials and Methods

3.1. Datasets for Drug–Disease Association Prediction

We obtained drug feature data, disease similarity data, and drug–disease association data from previous studies by Wang et al., which included 763 drugs and 681 diseases, and 3051 drug–disease associations. The initial data were sourced from several databases: The chemical substructures of the drugs were represented by the chemical fingerprints defined in the PubChem database [32]; the domain composition of the proteins targeted by the drugs was obtained from the InterPro database; the protein ontology characteristics (molecular functions and biological processes) of the target proteins were extracted from the UniProt database.

3.2. Representation of Multi-Source Data

Our primary goal was to predict and rank diseases potentially associated with drugs that are of interest to us. A non-negative matrix factorization model was established by integrating multiple data about drug features, drug similarities, disease similarities, and drug–disease associations. Drug r_i and disease d_j association scores can be computed using our model. The higher the association score, the more likely is an association between r_i and d_j. Three characteristic information representations of drugs including chemical drug features form an 881-dimensional binary chemical substructure vector, represented by the feature matrix $X^{(1)} \in R^{881 \times N_r}$, where N_r is the number of drugs, $\left(\left(X^{(1)}\right)^T\right)_j$ is the jth row of the transposed of $X^{(1)}$ that indicates the case where the drug r_j contains various chemical substructures. The term $\left(X^{(1)}\right)_{ij}$ is 1 if r_j has a chemical substructure c_i, or it is 0 otherwise. The 1426-dimensional target protein domain features are represented by matrix; similarly, the jth column of $X^{(2)}$ indicates whether drug r_j is associated with each protein domain. Using the matrix $X^{(3)} \in R^{4447 \times N_r}$ to represent the 4447-dimensional target gene ontology feature $\left(X^{(3)}\right)_{ij}$ indicates whether the protein targeted by drug r_j has the ith gene ontology; if so, the term $\left(X^{(3)}\right)_{ij} = 1$ applies or it is 0 otherwise.

Calculation and representation of three types of drug similarities. In this study, the similarity between drugs was assessed based on drug features and on the assumption that drug-related diseases are more likely to be similar when the drugs are more similar. For these three types of drug features, the more chemical substructures (or protein domains, or gene ontology attributes) are shared between two drugs, then the more similar they are (Figure 3a). Cosine similarity was computed to determine the similarity between drug r_i and r_j based on the three drug feature criteria, which are denoted as $(R_v)_{ij}$, where $R_v \in R^{N_r \times N_r}$ represents the similarity matrix of the vth feature data, $v = [1, 2, 3]$. Then, the cosine similarity was used to construct the similarity matrix of the vth drug feature,

$$(R_v)_{ij} = \frac{(X_v)_i \cdot (X_v)_j}{\| (X_v) \|_i * \| (X_v)_j \|} \tag{3}$$

where $\| \cdot \|$ is the modulus of a vector.

Calculation and representation of the fourth drug similarity. From a previous publication, we used the drug–disease association data [17], and if two drugs are associated with more similar diseases, the more similar they are. We constructed the fourth drug feature matrix $\left(X^{(4)}\right)^{N_d \times N_r}$, where N_d represents the number of diseases, and $\left(X^{(4)}\right)_{ij}$ is 1 if drug r_j and disease d_i are related or it is 0 otherwise. To compute the similarity feature matrix of the fourth criterion, $R_4 \in R^{N_r \times N_r}$, we obtained

the disease sets associated with drug r_i and drug r_j [33] and recorded them as $D_i = \{d_1, d_3\}$ and $D_j = \{d_2, d_3, d_5\}$. The fourth similarity of r_i and r_j was calculated as follows,

$$(R_4)_{ij} = \frac{\sum_{a=1}^{m} \underset{1 \le b \le n}{max} (D(d_{1a}, d_{2b})) + \sum_{b=1}^{n} \underset{1 \le a \le m}{max} (D(d_{2b}, d_{1a}))}{m + n} \quad (4)$$

where $D(d_{1a}, d_{2b})$ is the semantic similarity between d_{1a} belonging to D_i and disease d_{2b} belonging to D_j; m and n represent the number of diseases in D_i and D_j, respectively. According to a previous study, Equation (4) calculates the semantic similarity between two diseases [33].

Representation of the drug–disease association. An association matrix $Y \in R^{N_r \times N_d}$ was established based on known drug–disease associations. Each row of Y corresponds to a drug, and each column corresponds to a disease. Y_{ij} is 1 if there is a known association between drug r_i and disease d_j or it is 0 otherwise.

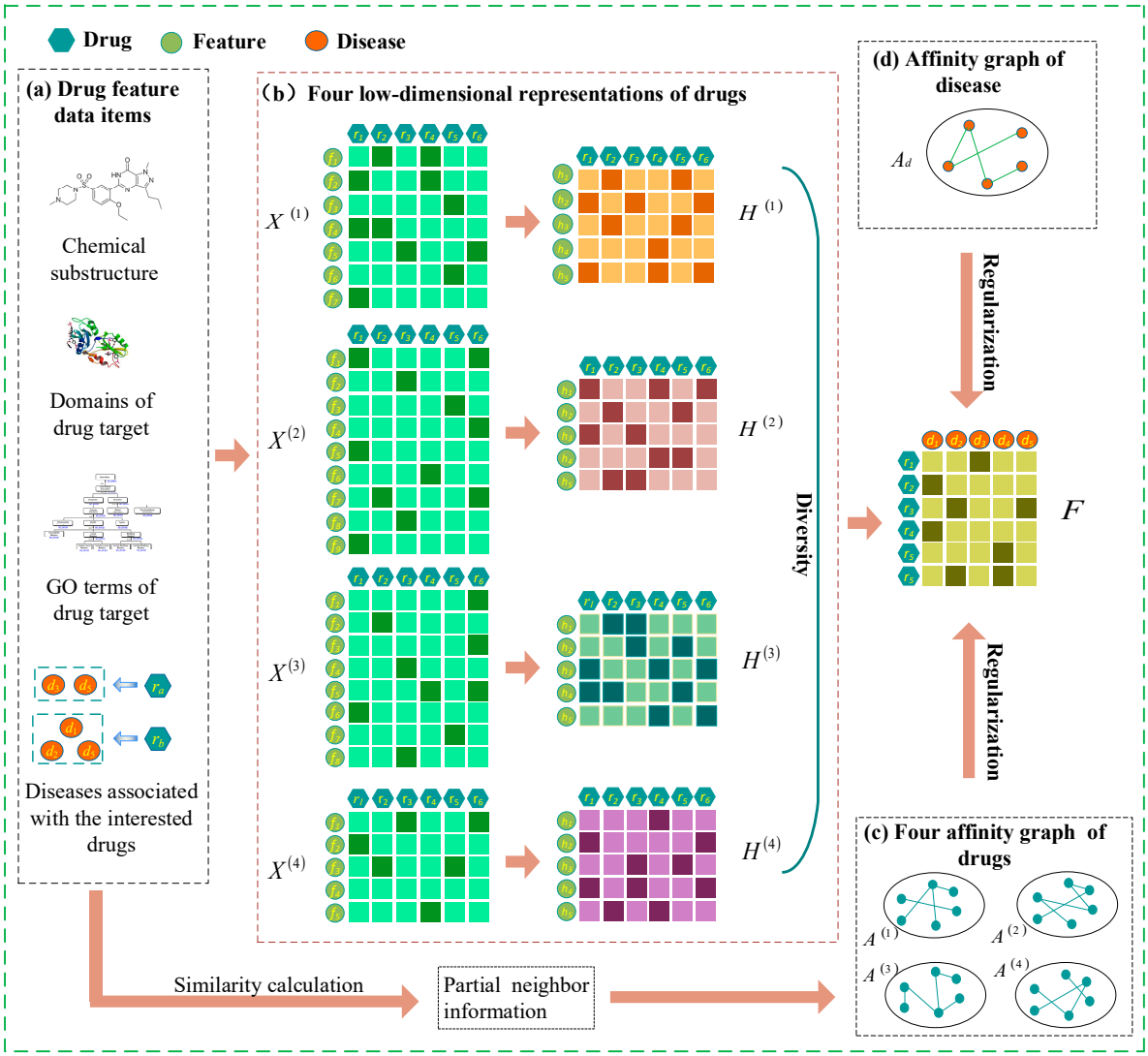


Figure 3. Representation of data from drugs and diseases from multiple sources and representation of drug–disease predictive association matrix F. (**a**) Drug feature data sets from multiple sources; (**b**) four low-dimensional representation of drugs; (**c**) four affinity maps of the drugs were obtained by similarity calculation; (**d**) extract the similarity of the diseases and obtain the affinity map of the disease.

3.3. Drug–Disease Association Prediction Model

Our new predictive model, DivePred, merges various drug features and can be used to predict new indications for drugs. We know that if two drugs share more of the same features, they are more likely to have a high similarity, indicating a potential association with similar diseases, which is at the core of our new model.

Modelling drug–disease association relationships. We introduced the matrix $F = \left(F_{ij}\right) \in R^{N_r \times N_d}$ to represent the association score matrix of N_r drugs and N_d diseases to better describe the model. In the model, F_i is the ith row of the association score matrix that represents the possibility of an association of drug r_i with all diseases. F_{ij} was the predicted association score between drug r_i and disease d_j, and a high F_{ij} indicates a stronger possibility of an association between r_i and d_j. Since the non-zero elements in Y are very sparse, previous studies using sparse cases usually built optimizations based on observed relationships only [34–36]. Here, we assume that the known set of observed drug–disease association information is Ω, and the construct matrix is $M = \left(M_{ij}\right) \in R^{N_r \times N_d}$, where M_{ij} was 1 if $\left(r_i, d_j\right) \in \Omega$, or it is 0 otherwise (in fact, $M = Y$). All known related drug–disease pairs should also be included in the predictions, i.e., there are known associations drug–disease should have a higher score in the prediction results. Therefore, the squared loss function was defined as,

$$min \parallel M \odot (F - Y) \parallel_F^2 \tag{5}$$

where $\| \cdot \|_F^2$ represents the Frobenius norm of a matrix, and $\odot$ is the Hadamard product.

Integrating multiple drug features into the model. We replaced the original feature matrix with a new matrix obtained by non-negative matrix factorization to fuse different types of drug features. $X^{(v)}$ indicates the vth feature matrix of drugs, and a new drug feature matrix $H^{(v)} \in R^{N_d \times N_r} (1 \le v \le 4)$ is obtained by matrix factorization of $X^{(v)}$ (Figure 3b); $\left(\left(H^{(v)}\right)^T\right)_i$ is the ith row of the transposed of $H^{(v)}$, representing the new feature vector of drug r_i in the vth view. While $W^{(v)} \in R^{d_v \times N_d} (1 \le v \le 4)$ denotes the basic matrix of the vth drug feature, the jth row of $W^{(v)}$, $\left(W^{(v)}\right)_j$, indicates the weight of each new feature to the original jth feature. $\left(W^{(v)}\right)_j \left(\left(H^{(v)}\right)^T\right)_i$ indicates the condition in which the drug r_i has the original features f_j. To ensure that the new drug feature matrix represents the original feature matrix as much as possible, $\left(W^{(v)}\right)_j \left(\left(H^{(v)}\right)^T\right)_i$ should match $\left(X^{(v)}\right)_{ji}$ as much as possible,

$$\min_{H^{(v)}, W^{(v)} \ge 0} \parallel M \odot (F - Y) \parallel_F^2 + \alpha_1 \sum_{v=1}^{4} \parallel X^{(v)} - W^{(v)} H^{(v)} \parallel_F^2 \tag{6}$$

where α_1 is a trade-off parameter that controls the weight of all drug feature information.

The multitude of drug similarities reflects the degree of similarity among the drugs from different aspects. There is consistency between the information from multiple aspects, but each view also has its own specific information. To ensure the diversity of each drug feature vector among the different views, we also require that each drug feature vector is as orthogonal as possible between the various views [37]. For example, $h_i^{(v)}$ and $h_i^{(w)}$ are the representation vectors of the drug r_i in the two drug feature views. To ensure that $h_i^{(v)}$ and $h_i^{(w)}$ are as different as possible, their dot product should approach zero.

$$\parallel h_i^{(v)} \circ h_i^{(w)} \parallel_1 = \sum_{j=1}^{K} h_{ji}^{(v)} \cdot h_{ji}^{(w)} \tag{7}$$

To derive a feature profile unique to every drug in each view, Formula (7) was introduced into the objective function.

$$\min_{H^{(v)},W^{(v)}\geq 0} \| M\odot(F-Y) \|_F^2 + a_1 \sum_{v=1}^{4} \| X^{(v)} - W^{(v)}H^{(v)} \|_F^2 + a_2 \sum_{w\neq v}^{4} tr\left(H^{(v)^T} H^{(w)}\right) \tag{8}$$

where $tr\left(H^{(v)^T} H^{(w)}\right) = \sum_{i=1}^{N_r}\sum_{j=1}^{N_d} h_{ji}^{(v)} \cdot h_{ji}^{(w)}$, and a_2 is used to control the contribution of the third term.

Modelling the drug–disease association score. In the drug–disease association score matrix F, the ith row of F, F_i, records the potential association score between drug r_i and various diseases. Furthermore, F_i is also the characteristic vector of r_i at the disease level. The ith column of $H^{(v)}$, $\left(\left(H^{(v)}\right)^T\right)_i$, is a new feature vector obtained after the original feature vector of the drug r_i is projected onto the disease dimension. $H^{(v)}$ plays a guiding role in the assessment of drug–disease association scores, $\left(\left(H^{(v)}\right)^T\right)_i$, and F_i should be as consistent as possible. The extended objective function was defined as:

$$\begin{aligned}\min_{H^{(v)},W^{(v)}\geq 0} \| M\odot(F-Y) \|_F^2 + a_1 \sum_{v=1}^{4} \| X^{(v)} - W^{(v)}H^{(v)} \|_F^2 + a_2 \sum_{w\neq v}^{4} DIVE\left(H^{(v)}, H^{(w)}\right) + \\ a_3 \sum_{v=1}^{4} \| F - H^{(v)^T} \|_F^2\end{aligned} \tag{9}$$

where α_3 is the super-parameter that regulates the contribution of drug characteristic information throughout the model.

Modelling the smoothness term. Drug r_i and its k neighbours are more likely to be associated with similar diseases. Hence, we established corresponding maps based on the drug neighbour information derived from the similarity of the four drugs. The corresponding adjacency matrix $A^{(v)}$ was obtained according to the vth figure (Figure 3c). $A^{(v)}$ was defined as,

$$\left(A^{(v)}\right)_{ij} = \begin{cases} 1, & \text{if the drug } r_j \text{ is one of the k most similar neighbours} \\ & \text{of the drug } r_i \text{ based on the vth drug similarity} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Since drug r_i and its neighbour r_j are more likely to be associated with a similar group of diseases, a drug-related smoothing term can be created,

$$\begin{aligned} &\tfrac{1}{2}\sum_{v=1}^{4}\sum_{i,j=1}^{N_r} \left(A^{(v)}\right)_{ij} \| F_i - F_j \|^2 \\ &= \sum_{v=1}^{4} \left(Tr\left(F^T U^{(v)} F\right) - Tr\left(F^T A^{(v)} F\right)\right) \\ &= \sum_{v=1}^{4} Tr\left(F^T L^{(v)} F\right) \end{aligned} \tag{11}$$

where F_i and F_j denote the ith and jth row vectors of F, respectively, and indicate the cases of a potential association of drug r_i and r_j with all diseases. $U^{(v)} \in R^{N_r\times N_r}$ is a diagonal matrix, where $\left(U^{(v)}\right)_{ii} = \sum_{j=1}^{N_r}\left(A^{(v)}\right)_{ij}$ and the Laplacian matrix of the vth feature graph is $L^{(v)} = U^{(v)} - A^{(v)}$.

Similarly, the disease d_i and its k neighbours are more likely to be associated with similar drugs. Therefore, we established a graph with disease as a node according to disease similarity and obtained the adjacency matrix A_d defined as (Figure 3d),

$$(A_d)_{ij} = \begin{cases} 1, & \text{if the disease } d_j \text{ is one of the } k \text{ most} \\ & \text{similar neighbours of the disease } d_i \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Therefore, disease–related regularization items were created as follows,

$$\begin{aligned} &\tfrac{1}{2}\sum_{i,j=1}^{N_r}(A_d)_{ij} \parallel \left(F^T\right)_i - \left(F^T\right)_j \parallel^2 \\ &= Tr\left(FU_dF^T\right) - Tr\left(FA_dF^T\right) \\ &= Tr\left(FL_dF^T\right) \end{aligned} \tag{13}$$

where $\left(F^T\right)_i$ and $\left(F^T\right)_j$ were the ith and jth row of F^T, respectively. They represent the potential association of disease d_i and d_j with all drugs. $U_d \in R^{N_d \times N_d}$ was a diagonal matrix, $(U_d)_{ii} = \sum_{j=1}^{N_d}(A_d)_{ij}$, and $L^{(v)} = U^{(v)} - A^{(v)}$ was the Laplace matrix of the characteristic graph of the disease. Then, we added a smoothness term to the objective function,

$$\begin{aligned} \min_{H^{(v)},W^{(v)}\geq 0} \parallel M \odot (F - Y) \parallel_F^2 + \alpha_1 \sum_{v=1}^{4} \parallel X^{(v)} - W^{(v)}H^{(v)} \parallel_F^2 + \alpha_2 \sum_{w\neq v}^{4} DIVE\left(H^{(v)}, H^{(w)}\right) + \\ \alpha_3 \sum_{v=1}^{4} \parallel F - H^{(v)^T} \parallel_F^2 + \alpha_4\left(\sum_{v=1}^{4} Tr\left(F^TL^{(v)}F\right) + Tr\left(FL_dF^T\right)\right) \end{aligned} \tag{14}$$

where α_4 adjusts the contribution of the smoothing term.

Considering the sparsity of drug–disease associations. The potential associations between drugs and diseases was limited. Thus, drug–disease associations have sparse properties. We used the l_1-norm to adjust the association matrix for sparse associations. We created the final objective function after adding the sparse item,

$$\begin{aligned} \min_{H^{(v)},W^{(v)}\geq 0} \parallel M \odot (F - Y) \parallel_F^2 + \alpha_1 \sum_{v=1}^{4} \parallel X^{(v)} - W^{(v)}H^{(v)} \parallel_F^2 + \alpha_2 \sum_{w\neq v}^{4} DIVE\left(H^{(v)}, H^{(w)}\right) \\ + \alpha_3 \sum_{v=1}^{4} \parallel F - H^{(v)^T} \parallel_F^2 + \alpha_4\left(Tr\left(\sum_{v=1}^{4} F^TL^{(v)}F\right) + Tr\left(FL_dF^T\right)\right) + \alpha_5 \parallel F \parallel_1 \end{aligned} \tag{15}$$

where α_5 is a regulation parameter.

3.4. Optimization

Since the objective Function (15) with the variables F, $H^{(v)}$ and $W^{(v)}$ is a non-convex function, it was impractical to derive a global optimal solution. Therefore, we divided the optimization problem into three subproblems and performed iterative optimization, converging each subproblem to a local minimum.

***F*-subproblem**. We updated F with fixed $W^{(v)}$ and $H^{(v)}$, and the resulting formula contains only the unknown variable F,

$$\begin{aligned} minL(F) = \parallel M \odot \quad & (F - Y) \parallel_F^2 + \alpha_3 \sum_{v=1}^{4} \parallel F - H^{(v)^T} \parallel_F^2 \\ & + \alpha_4\left(Tr\left(\sum_{v=1}^{4} F^TL^{(v)}F\right) + Tr\left(FL_dF^T\right)\right) + \alpha_5 \parallel F \parallel_1 \end{aligned} \tag{16}$$

The item containing the Frobenius norm in Equation (16) was changed to the form of the matrix trace, which can be rewritten as,

$$\begin{aligned} L(F) = Tr(M\odot \quad & \left(FF^T - FY^T - YF^T + YY^T\right)\right) \\ & +\alpha_3 \sum_{v=1}^{4} Tr\left(FF^T - FH^{(v)} - H^{(v)^T}F^T + H^{(v)^T}H^{(v)}\right) \\ & +\alpha_4\left(Tr\left(\sum_{v=1}^{4} F^T L^{(v)} F\right) + Tr\left(FL_dF^T\right)\right) + \alpha_5 \| F \|_1 \end{aligned} \tag{17}$$

By setting the derivative of $L(F)$ with respect to F to 0, we obtained,

$$2M\odot(F-Y) + 2\alpha_3 \sum_{v=1}^{4}\left(F - H^{(v)^T}\right) + 2\alpha_4\left(\sum_{v=1}^{4}\left(U^{(v)} - A^{(v)}\right)F + F(U_d - A_d)\right) + \alpha_5 = 0 \tag{18}$$

where all elements in matrix $B = \left[B_{ij}\right] \in \Re^{N_r \times N_d}$ are 1. By multiplying both sides of Equation (18) with F_{ij}, the following equation was obtained,

$$\left(2M\odot(F-Y) + 2\alpha_3 \sum_{v=1}^{4}\left(F - H^{(v)^T}\right) + 2\alpha_4\left(\sum_{v=1}^{4}\left(U^{(v)} - A^{(v)}\right)F + F(U_d - A_d)\right) + \alpha_5 B\right)_{ij} F_{ij} = 0. \tag{19}$$

We updated F according to the coordinate gradient descent Algorithm [38], and derived an updated formula,

$$F_{ij}^{new} \leftarrow F_{ij} \frac{\left(2M * Y + 2\alpha_3 \sum_{v=1}^{4} H^{(v)^T} + 2\alpha_4 \sum_{v=1}^{4} A^{(v)}F + 2a_4 FA_d\right)_{ij}}{\left(2M * F + 8F + 2\alpha_4 \sum_{v=1}^{4} U^{(v)}F + 2a_4 FU_d + \alpha_5 B\right)_{ij}} \tag{20}$$

***H*$^{(v)}$-subproblem**. We updated $H^{(v)}$ with fixed F and $W^{(v)}$. The function that only containing the variable $H^{(v)}$ was as follows,

$$\min_{H^{(v)}\geq 0} L\left(H^{(v)}\right) = \alpha_1 \| X^{(v)} - W^{(v)}H^{(v)} \|_F^2 + \alpha_2 \sum_{w\neq v}^{4} DIVE\left(H^{(v)}, H^{(w)}\right) + \alpha_3 \sum_{v=1}^{4} \| F - H^{(v)^T} \|_F^2 . \tag{21}$$

The term of the Frobenius norm in Equation (21) was changed to the form of the matrix trace. Assuming that $\eta_{ij}^{(v)}$ is the Lagrangian multiplier of constraint $H_{ij}^{(v)} \geq 0$, and $\eta^{(v)} = \left[\eta_{ij}{}^{(v)}\right]$, the resulting Lagrangian function of $H^{(v)}$ was as follows,

$$\begin{aligned} \min_{H^{(v)}\geq 0} L\left(H^{(v)}\right) = \quad & \alpha_1 Tr\left(X^{(v)}X^{(v)^T} - X^{(v)}H^{(v)^T}W^{(v)^T}\right. \\ & \left. -W^{(v)}H^{(v)}X^{(v)^T} + W^{(v)}H^{(v)}H^{(v)^T}W^{(v)^T}\right) + \alpha_2 \sum_{w\neq v}^{4} Tr\left(H^{(v)}H^{(w)^T}\right) \\ & +\alpha_3 Tr\left(FF^T - FH^{(v)} - H^{(v)^T}F^T + H^{(v)^T}H^{(v)}\right) + Tr\left(\eta^{(v)}H^{(v)}\right). \end{aligned} \tag{22}$$

By setting the derivative of $L\left(H^{(v)}\right)$ with respect to $H^{(v)}$ to 0, we obtained,

$$\alpha_1\left(2W^{(v)^T}W^{(v)}H^{(v)} - 2W^{(v)^T}X^{(v)}\right) + \alpha_2 \sum_{w\neq v}^{4} H^{(w)} + \alpha_3\left(2H^{(v)} - 2F^T\right) + \eta^{(v)} = 0 \tag{23}$$

According to the KTT condition $\eta_{ij}^{(v)} H_{ij}^{(v)} = 0$, we derived the following formula,

$$\left(2\alpha_1 W^{(v)^T} W^{(v)} H^{(v)} - 2\alpha_1 W^{(v)^T} X^{(v)} + \alpha_2 \sum_{w \neq v}^{4} H^{(w)} + 2\alpha_3 H^{(v)} - 2\alpha_3 F^T\right)_{ij} H_{ij}^{(v)} = 0 \tag{24}$$

Then we obtained the updated formula for $H^{(v)}$,

$$\left(H_{ij}^{(v)}\right)^{new} \leftarrow H_{ij}^{(v)} \frac{\left(2\alpha_1 W^{(v)^T} X^{(v)} + 2\alpha_3 F^T\right)_{ij}}{\left(2\alpha_1 W^{(v)^T} W^{(v)} H^{(v)} + \alpha_2 \sum_{w \neq v}^{4} H^{(w)} + 2\alpha_3 H^{(v)}\right)_{ij}} \tag{25}$$

$W^{(v)}$-subproblem. By using fixed F and $H^{(v)}$, we could update $W^{(v)}$. The subproblem with $W^{(v)}$ as the only variable was as follows,

$$\min_{W^{(v)} \geq 0} L\left(W^{(v)}\right) = \alpha_1 \parallel X^{(v)} - W^{(v)} H^{(v)} \parallel_F^2 \tag{26}$$

Then, we changed the term containing the Frobenius norm in Equation (26) to the form of the matrix trace, and let $\beta^{(v)} = \left[\beta_{ij}{}^{(v)}\right]$ be the Lagrangian multiplier with the constraint $W^{(v)} \geq 0$. The resulting Lagrangian function for $W^{(v)}$ was as follows,

$$\begin{aligned} \min_{W^{(v)} \geq 0} L\left(W^{(v)}\right) = \quad & \alpha_1 \left(X^{(v)} X^{(v)^T} - 2X^{(v)} H^{(v)^T} W^{(v)^T} - W^{(v)} H^{(v)} X^{(v)^T} + W^{(v)} H^{(v)} H^{(v)^T} W^{(v)}\right) \\ & + Tr\left(\beta^{(v)} W^{(v)}\right) \end{aligned} \tag{27}$$

By setting the derivative of $L\left(W^{(v)}\right)$ to $W^{(v)}$ to 0, we created the following formula,

$$2\alpha_1 W^{(v)} H^{(v)} H^{(v)^T} - 2\alpha_1 X^{(v)} H^{(v)^T} + \beta^{(v)} = 0 \tag{28}$$

Similarly, according to the KTT condition $\beta_{ij}^{(v)} W_{ij}^{(v)} = 0$, we derived,

$$\left(2\alpha_1 W^{(v)} H^{(v)} H^{(v)^T} - 2\alpha_1 X^{(v)} H^{(v)^T}\right)_{ij} W_{ij}^{(v)} = 0 \tag{29}$$

Therefore, the updated formula for $W^{(v)}$ was as follows,

$$\left(W_{ij}^{(v)}\right)^{new} \leftarrow W_{ij}^{(v)} \frac{\left(X^{(v)} H^{(v)^T}\right)_{ij}}{\left(W^{(v)} H^{(v)} H^{(v)^T}\right)_{ij}} \tag{30}$$

We solve F, $H^{(v)}$, and $W^{(v)}$ iteratively by using the above updating rules. Finally, F_{ij} is regarded as the estimated association score between drug r_i and disease d_j (Algorithm 1).

Algorithm 1 DivePred algorithm for predicting the potential drug-disease associations.

Input: A drug-disease association matrix $Y \in \Re^{N_r \times N_d}$ and the drugs character matrix $X_1 \in \Re^{881 \times N_r}$, $X_2 \in \Re^{1426 \times N_r}$, $X_3 \in \Re^{4447 \times N_r}$, $X_4 \in \Re^{N_d \times N_r}$.

Output: Drug-disease association score matrix F, where F_{ij} is the association score for drug r_i and disease d_j.

1. Randomly initialize the elements in $F, H^{(v)}, W^{(v)}$ $(1 \le v \le 4)$ with the values between 0 and 1.
2. While $L\left(F^{(v)}, H^{(v)}, W^{(v)}\right)$ not converged do
3. Fix $W^{(v)}$ and $H^{(v)}$, along with an update for F, using the rule:

$$F_{ij}^{new} \leftarrow F_{ij} \cdot \frac{\left(2M * Y + 2\alpha_3 \sum_{v=1}^{4} H^{(v)^T} + 2\alpha_4 \sum_{v=1}^{4} A^{(v)} F + 2a_4 F A_d\right)_{ij}}{\left(2M * F + 8F + 2\alpha_4 \sum_{v=1}^{4} U^{(v)} F + 2a_4 F U_d + \alpha_5 B\right)_{ij}}$$

4. For $v = 1$ to 4
5. Fix F and $W^{(v)}$, along with an update for $H^{(v)}$, using the rule:

$$\left(H_{ij}^{(v)}\right)^{new} \leftarrow H_{ij}^{(v)} \frac{\left(2\alpha_1 W^{(v)^T} X^{(v)} + 2\alpha_3 F^T\right)_{ij}}{\left(2\alpha_1 W^{(v)^T} W^{(v)} H^{(v)} + \alpha_2 \sum_{w \ne v}^{4} H^{(w)} + 2\alpha_3 H^{(v)}\right)_{ij}}$$

6. End for
7. For $v = 1$ to 4
8. Fix F and $H^{(v)}$, along with an update for $W^{(v)}$, using the rule:

$$\left(W_{ij}^{(v)}\right)^{new} \leftarrow W_{ij}^{(v)} \frac{\left(X^{(v)} H^{(v)^T}\right)_{ij}}{\left(W^{(v)} H^{(v)} H^{(v)^T}\right)_{ij}}$$

9. End for
10. End While

4. Conclusions

A method based on non-negative matrix factorization, DivePred, was developed to infer the potential associations between drugs and diseases. DivePred captures a variety of information on each drug, including four kinds of drug features and specific features associated with different aspects of the drugs. Meanwhile, it also captures disease–disease similarities and drug–disease associations. The projection of multiple kinds of drug features, along with the drugs and diseases neighbour information, was completely integrated to enhance the inference of drug–disease associations. An iterative algorithm was developed to estimate drug–disease association scores that can be used to prioritize disease candidates for each drug. DivePred outperforms other methods in AUCs and AUPRs. For biologists, DivePred is very useful because more real drug–disease associations were included in DivePred's top-ranking candidate list. Case studies on five drugs demonstrated that DivePred could detect potentially new indications for drugs. DivePred can serve as a prioritization tool to screen the potential candidates for subsequent discovery of real drug–disease associations through biological validation.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/17/4102/s1.

Author Contributions: P.X. and Y.S. conceived the prediction method, and Y.S. wrote the paper. Y.S. and L.J. developed the computer programs. P.X. and T.Z. analyzed the results and revised the paper.

Funding: The work was supported by the Natural Science Foundation of China (61972135), the Natural Science Foundation of Heilongjiang Province (LH2019F049, LH2019A029), the China Postdoctoral Science Foundation (2019M650069), the Heilongjiang Postdoctoral Scientific Research Staring Foundation (BHL-Q18104), the Fundamental Research Foundation of Universities in Heilongjiang Province for Technology Innovation (KJCX201805), and the Fundamental Research Foundation of Universities in Heilongjiang Province for Youth Innovation Team (RCYJTD201805).

Acknowledgments: We would like to thank Editage (www.editage.com) for English language editing.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Li, J.; Zheng, S.; Chen, B.; Butte, A.J.; Swamidass, S.J.; Lu, Z. A survey of current trends in computational drug repositioning. *Brief. Bioinform.* **2015**, *17*, 2–12. [CrossRef] [PubMed]
2. Neuberger, A.; Oraiopoulos, N.; Drakeman, D.L. Renovation as innovation: Is repurposing the future of drug discovery research? *Drug Discov. Today* **2019**, *24*, 1. [CrossRef] [PubMed]
3. Adams, C.P.; Brantner, V.V. Estimating the cost of new drug development: Is it really $802 million? *Health Aff.* **2006**, *25*, 420–428. [CrossRef] [PubMed]
4. Dickson, M.; Gagnon, J.P. Key factors in the rising cost of new drug discovery and development. *Nat. Rev. Drug Discov.* **2004**, *3*, 417. [CrossRef] [PubMed]
5. Tamimi, N.A.; Ellis, P. Drug development: From concept to marketing! *Nephron Clin. Pract.* **2009**, *113*, c125–c131. [CrossRef] [PubMed]
6. Pushpakom, S.; Iorio, F.; Eyers, P.A.; Escott, K.J.; Hopper, S.; Wells, A.; Doig, A.; Guilliams, T.; Latimer, J.; McNamee, C. Drug repurposing: Progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **2019**, *18*, 41. [CrossRef] [PubMed]
7. Paul, S.M.; Mytelka, D.S.; Dunwiddie, C.T.; Persinger, C.C.; Munos, B.H.; Lindborg, S.R.; Schacht, A.L. How to improve r&d productivity: The pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203.
8. Sultana, J.; Calabró, M.; Garcia-Serna, R.; Ferrajolo, C.; Crisafulli, C.; Mestres, J. Biological substantiation of antipsychotic-associated pneumonia: Systematic literature review and computational analyses. *PloS ONE* **2017**, *12*, e0187034. [CrossRef]
9. Grabowski, H. Are the economics of pharmaceutical research and development changing? *Pharmacoeconomics* **2004**, *22*, 15–24. [CrossRef]
10. Kinch, M.S.; Griesenauer, R.H. 2017 in review: Fda approvals of new molecular entities. *Drug Discov. Today* **2018**, *23*, 1469–1473. [CrossRef]
11. Ashburn, T.T.; Thor, K.B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **2004**, *3*, 673. [CrossRef] [PubMed]
12. Nosengo, N. Can you teach old drugs new tricks? *Nat. News* **2016**, *534*, 314. [CrossRef] [PubMed]
13. Pritchard, J.L.E.; O'Mara, T.A.; Glubb, D.M. Enhancing the promise of drug repositioning through genetics. *Front. Pharmacol.* **2017**, *8*, 896. [CrossRef] [PubMed]
14. Haupt, V.J.; Schroeder, M. Old friends in new guise: Repositioning of known drugs with structural bioinformatics. *Brief. Bioinform.* **2011**, *12*, 312–326. [CrossRef] [PubMed]
15. Lotfi Shahreza, M.; Ghadiri, N.; Mousavi, S.R.; Varshosaz, J.; Green, J.R. A review of network-based approaches to drug repositioning. *Brief. Bioinform.* **2017**, *19*, 878–892. [CrossRef]
16. Sirota, M.; Dudley, J.T.; Kim, J.; Chiang, A.P.; Morgan, A.A.; Sweet-Cordero, A.; Sage, J.; Butte, A.J. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Translat. Med.* **2011**, *3*, 96ra77. [CrossRef] [PubMed]
17. Wang, L.; Wang, Y.; Hu, Q.; Li, S. Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks. *CPT: Pharmacomet. Syst. Pharmacol.* **2014**, *3*, 1–9. [CrossRef]
18. Yu, L.; Huang, J.; Ma, Z.; Zhang, J.; Zou, Y.; Gao, L. Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomic.* **2015**, *8*, S2. [CrossRef]
19. Peyvandipour, A.; Saberian, N.; Shafi, A.; Donato, M.; Draghici, S. A novel computational approach for drug repurposing using systems biology. *Bioinformatics* **2018**, *34*, 2817–2825. [CrossRef]
20. Wang, Y.; Chen, S.; Deng, N.; Wang, Y. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PloS ONE* **2013**, *8*, e78518.

21. Wang, W.; Yang, S.; Zhang, X.; Li, J. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* **2014**, *30*, 2923–2930. [CrossRef] [PubMed]
22. Luo, H.; Wang, J.; Li, M.; Luo, J.; Peng, X.; Wu, F.X.; Pan, Y. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics* **2016**, *32*, 2664–2671. [CrossRef] [PubMed]
23. Liu, H.; Song, Y.; Guan, J.; Luo, L.; Zhuang, Z. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinformatics* **2016**, *17*, 539. [CrossRef] [PubMed]
24. Luo, H.; Li, M.; Wang, S.; Liu, Q.; Li, Y.; Wang, J. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* **2018**, *34*, 1904–1912. [CrossRef] [PubMed]
25. Gottlieb, A.; Stein, G.Y.; Ruppin, E.; Sharan, R. Predict: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **2011**, *7*, 496. [CrossRef] [PubMed]
26. Iwata, H.; Sawada, R.; Mizutani, S.; Yamanishi, Y. Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. *J. Chem. Inf. Modeling* **2015**, *55*, 446–459. [CrossRef] [PubMed]
27. Liang, X.; Zhang, P.; Yan, L.; Fu, Y.; Peng, F.; Qu, L.; Shao, M.; Chen, Y.; Chen, Z. Lrssl: Predict and interpret drug–disease associations based on data integration using sparse subspace learning. *Bioinformatics* **2017**, *33*, 1187–1196. [CrossRef]
28. Zhang, W.; Yue, X.; Lin, W.; Wu, W.; Liu, R.; Huang, F.; Liu, F. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics* **2018**, *19*, 233. [CrossRef]
29. Xuan, P.; Cao, Y.; Zhang, T.; Wang, X.; Pan, S.; Shen, T. Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics* **2019**. [CrossRef]
30. Bradley, J.D.; Brandt, K.D.; Katz, B.P.; Kalasinski, L.A.; Ryan, S.I. Comparison of an antiinflammatory dose of ibuprofen, an analgesic dose of ibuprofen, and acetaminophen in the treatment of patients with osteoarthritis of the knee. *N. Engl. J. Med.* **1991**, *325*, 87–91. [CrossRef]
31. Stolman, L.P. Hyperhidrosis: Medical and surgical treatment. *Eplasty* **2008**, *8*, e22. [PubMed]
32. Wang, Y.; Xiao, J.; Suzek, T.O.; Zhang, J.; Wang, J.; Bryant, S.H. Pubchem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633. [CrossRef] [PubMed]
33. Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics* **2010**, *26*, 1644–1650. [CrossRef] [PubMed]
34. Natarajan, N.; Dhillon, I.S. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics* **2014**, *30*, i60–i68. [CrossRef] [PubMed]
35. Chen, X.; Wang, L.; Qu, J.; Guan, N.N.; Li, J.Q. Predicting mirna–disease association based on inductive matrix completion. *Bioinformatics* **2018**, *34*, 4256–4265. [CrossRef]
36. Zhao, Y.; Chen, X.; Yin, J. A novel computational method for the identification of potential mirna-disease association based on symmetric non-negative matrix factorization and kronecker regularized least square. *Front. Genet.* **2018**, *9*. [CrossRef]
37. Wang, J.; Tian, F.; Yu, H.; Liu, C.H.; Zhan, K.; Wang, X. Diverse non-negative matrix factorization for multiview data representation. *IEEE Trans. Cybern.* **2017**, *48*, 2620–2632. [CrossRef]
38. Tan, V.Y.; Févotte, C. Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1592–1605. [CrossRef]