DDA-SKF: Predicting Drug–Disease Associations Using Similarity Kernel Fusion

Chu-Qiao Gao, Yuan-Ke Zhou, Xiao-Hong Xin, Hui Min and Pu-Feng Du*

College of Intelligence and Computing, Tianjin University, Tianjin, China

Drug repositioning provides a promising and efficient strategy to discover potential associations between drugs and diseases. Many systematic computational drug-repositioning methods have been introduced, which are based on various similarities of drugs and diseases. In this work, we proposed a new computational model, DDA-SKF (drug–disease associations prediction using similarity kernels fusion), which can predict novel drug indications by utilizing similarity kernel fusion (SKF) and Laplacian regularized least squares (LapRLS) algorithms. DDA-SKF integrated multiple similarities of drugs and diseases. The prediction performances of DDA-SKF are better, or at least comparable, to all state-of-the-art methods. The DDA-SKF can work without sufficient similarity information between drug indications. This allows us to predict new purpose for orphan drugs. The source code and benchmarking datasets are deposited in a GitHub repository (https://github.com/GCQ2119216031/DDA-SKF).

Keywords: drug repositioning, drug–disease association, similarity kernel fusion, Laplacian regularized least squares, orphan drugs

1 INTRODUCTION

In recent decades, although investment in pharmaceutical research and development has increased substantially, the discovery of a new drug is still a time-consuming, risky, and challenging process (Chong and Sullivan, 2007; Paul et al., 2010; Pammolli et al., 2011; Li et al., 2016). The pharmaceutical industry did not receive rational benefits from investments (Kola and Landis, 2004; Munos, 2009; Schuhmacher et al., 2016). A large number of new drug candidates failed in the FDA evaluations, thereby preventing their applications in therapies (Weng et al., 2013; Mullard, 2017; Mullard, 2018; Mullard, 2019; Mullard, 2020; Mullard, 2021). Today, drug repositioning has gained importance in identifying new therapeutic purposes for already-approved drugs (Parvathaneni et al., 2019). Repositioning of “old” drugs to treat both common and rare diseases is becoming an attractive proposition because it involves the use of safe compounds, with potentially lower overall development costs and shorter development cycles (Pushpakom et al., 2019). So far, drug repositioning has achieved many successes (Swanson, 1990; Soignet et al., 1998; Ashburn and Thor, 2004). Currently, rare diseases are a serious threat to human health (Wästfelt et al., 2006; Harari, 2016). Several orphan drugs are intended to treat rare diseases. Developing a new drug intended to treat a rare disease is costly and time-consuming (Attwood et al., 2018). Drug repositioning is considered as a more efficient strategy than traditional drug development (Scherman and Fetro, 2020). In fact, some successfully repurposed orphan drugs (Gordon et al., 2018; Kurolap et al., 2019; Scherman and Fetro, 2020) have attracted much attention. Therefore, systematic computational drug repositioning is an important research topic in both medical science and computational life science.
Many computational methods have been proposed to predict potential drug–disease associations for drug repositioning. Machine learning approaches play important roles in predicting associations between drugs and diseases. For example, Gottlieb et al. (2011) integrated several drug similarities and disease similarities to score the novel drug–disease associations by learning a logistic regression classifier. Recently, a method, which is called DRIMC (Zhang et al., 2020), introduced a drug-repositioning approach by using Bayesian inductive matrix completion. Xuan et al. (2019) presented DisDrugPred method based on nonnegative matrix factorization to predict the drug-related candidate indications. SCMFDD (Zhang et al., 2018b) proposed a similarity-constrained matrix factorization method for the drug–disease association prediction. PREDICT (Gottlieb et al., 2011) integrated several similarities to score the novel associations by learning a logistic regression classifier. LRSSL (Liang et al., 2017) proposed a Laplacian regularized sparse subspace learning method to identify novel indications.

Recently, the systems biology–based methods have been overwhelmingly studied in predicting drug–disease associations. Most methods of this kind constructed a heterogenous network model before calculating prediction scores. The MBiRW model (Luo et al., 2016) utilized bi-random walk (BiRW) algorithm to predict unknown drug–disease associations on a heterogenous network. Luo et al. (2018) designed a drug repositioning recommendation system (DRRS) based on the singular value decomposition to complete the potential association matrix of the heterogenous network. BNNR (Yang et al., 2019a) employs a bounded norm regularization method to complete the drug–disease matrix under the low-rank assumption. NTSIM (Zhang et al., 2018a) used the network topological similarity–based inference method. Yu et al. (2021) proposed a novel computational method, which is named as a layer attention graph convolutional network, to predict unobserved associations.

Previous studies have achieved many successes. It should be noted that most existing methods employ only one type of drug or disease similarity rather than integrating multiple similarities (Yang et al., 2019a; Yang et al., 2019b). Some models utilized linear combinations to integrate multiple similarities (Jiang et al., 2019; Yan et al., 2019), which loses the high-order interactions between different similarities.

In this work, we proposed the similarity kernel fusion (SKF) to integrate different similarity kernels with Laplacian regularized least squares (LapRLS) algorithms. We named our method as the DDA-SKF method. The overall workflow of DDA-SKF was illustrated in Figure 1. First, we constructed multiple similarity kernels for drugs and diseases. These similarity kernels were integrated by the SKF iterative process into two comprehensive similarity kernels. Finally, the Laplacian regularized least squares (LapRLS) algorithms were used to obtain the prediction association matrix. To demonstrate the effectiveness of our model, we compared it with several state-of-the-art methods.

2 MATERIALS AND METHODS

2.1 Drug–Disease Association Dataset

We adapted two benchmarking datasets in this article, which are the PREDICT dataset and the LRSSL dataset. The PREDICT dataset was obtained from the literature (Gottlieb et al., 2011), which was originally collected from the DrugBank database (Wishart et al., 2018) and the OMIM (Online Mendelian Inheritance in Man) database (Hamosh et al., 2005). It contains 1933 known drug–disease associations between 593 drugs and 313 diseases. The LRSSL dataset was extracted from the literature (Liang et al., 2017). It contains 3,051 drug–disease associations between 763 drugs and 681 diseases.

Without losing generality, we take the PREDICT dataset as an example to describe our method. Let $p_i$ (i = 1, 2, ..., 313) be the i-th disease and $d_j$ (j = 1, 2, ..., 593) the j-th drug. We defined the relationship between $p_i$ and $d_j$ as:

$$ a_{ij} = \begin{cases} 0 & p_i \text{ and } d_j \text{ have no association,} \\ 1 & \text{otherwise} \end{cases} \quad (1) $$

An adjacency matrix was then created to describe the drug–disease associations in the whole dataset which can be noted as $A \in \mathbb{R}^{313 \times 593}$.

2.2 Similarity Kernels for Drug and Disease

The basic assumption of this work is that drug-related diseases are more likely to be similar when the drugs are more similar. Three different similarity kernels of drugs and two different similarity kernels of diseases were applied. By fusing these similarity kernels, our method can identify potential associations between drugs and diseases.

2.2.1 Drug Chemical Similarity

We obtained the drug chemical similarity kernel from the literature (Zhang et al., 2020). PaDEL software (Yap, 2011) was used to compute PubChem fingerprint descriptors for each drug. The pairwise similarity between drugs was measured by the Jaccard coefficient between PubChem fingerprints. The drug chemical similarity kernel was noted as a $593 \times 593$ matrix $M_{d,1}$. $M_{d,1} (i, j)$, which is the element in the i-th row and the j-th column of $M_{d,1}$, represents the chemical similarity between the i-th drug and the j-th drug. Because of the Jaccard coefficient definition, 0 $\leq M_{d,1} (i, j) \leq 1$.

2.2.2 Drug Functional Similarity Kernel

The drug function was described by the Gene Ontology terms of the target genes. The Jaccard coefficient was used to measure the similarity between the Gene Ontology annotations of two different target genes. We obtain the similarity values from the literature (Zhang et al., 2020). The drug functional similarity was noted also as a $593 \times 593$ matrix $M_{d,2}$. The element $M_{d,2} (i, j)$ is the functional similarity between the i-th drug and the j-th drug.

2.2.3 Association Similarity Kernels for Drugs and Diseases

The association profile of the i-th drug is defined as the i-th column of the association matrix $A$, which can be noted as $A_i$. The drug association similarity kernel was represented by a $593 \times 593$ matrix $M_{d,3}$; the element $M_{d,3} (i, j)$ is the association similarity between $d_i$ and $d_j$, which can be defined as:
where

\[ M_{d,3}(i, j) = \exp\left(-\gamma_d \| A_{i,1} - A_{j,1} \|^2\right), \]  

and \( \| . \| \) is the 2-norm operator.

Similarly, the association profile of the \( i \)-th disease is defined as the \( i \)-th row of the association matrix \( A \), which can be noted as \( A_{i,1} \).

The disease association similaritykernel was represented by a \( 313 \times 313 \) matrix \( M_{p,1} \); the element \( M_{p,1}(i, j) \) is the association similarity between disease \( p_1 \) and disease \( p_2 \), which can be calculated as:

\[ M_{p,1}(i, j) = \exp\left(-\gamma_p \| A_{i} - A_{j} \|^2\right), \]  

where

\[ y_p = 313 / \sum_{i=1}^{313} \| A_{i} \|^2. \]  

### 2.2.4 Disease Semantic Similarity Kernel

A disease semantic similarity kernel is a \( 313 \times 313 \) matrix \( M_{p,2} \). The element \( M_{p,2}(i, j) \) is the similarity between disease \( p_1 \) and disease \( p_2 \). The disease semantic similarity is accessible by MimMiner (van Driel et al., 2006), which is a text mining approach to quantify phenotype relationships between human disease phenotypes from the OMIM database. We obtain the values of \( M_{p,2} \) from the literature (Zhang et al., 2020).

### 2.3 Similarity Kernel Fusion

The similarity kernel fusion (SKF) algorithm was applied to integrate three drug similarity kernels into a drug comprehensive similarity kernel and two disease similarity kernels into a disease comprehensive similarity kernel. Without losing generality, we take the drug similarity kernels as an example to explain this fusing process.

First, we normalize the aforementioned three drug similarity kernels \( M_{d,u} \), \( u = 1, 2, 3 \) by using the following formula:

\[ \theta_{d,u}(i, j) = \frac{M_{d,u}(i, j)}{\sum_{v=1}^{593} M_{d,u}(v, j)}, \]  

where \( \theta_{d,u}(i, j) \) represents a normalized similarity corresponding to \( M_{d,u}(i, j) \). The matrix composed by the normalized kernel is noted as:

\[ \Theta_{d,u} = \{ \theta_{d,u}(i, j) \}_{593 \times 593}. \]

Second, we established a neighbor-constrained normalization kernel for each drug similarity kernel. Given the drug \( d \) and \( M_{d,u} \), we collected the \( k \) most similar drugs as a set \( N_{d,u}(i, k) \). The neighborhood-constrained normalization of the \( M_{d,u} \) can be defined as follows:

\[ \phi_{d,u}(i, j) = \frac{M_{d,u}(i, j) I_{d,u,k}(i, j)}{\sum_{v=1}^{593} M_{d,u}(v, j) I_{d,u,k}(i, v)}, \]  

where

\[ I_{d,u,k}(i, j) = \begin{cases} 1 & d_j \in N_{d,u}(i, k) \backslash \{d_i\}, \\ 0 & d_j \notin N_{d,u}(i, k). \end{cases} \]

The matrix composed by the neighborhood-constrained normalization is noted as:

\[ \Phi_{d,u} = \{ \phi_{d,u}(i, j) \}_{593 \times 593}. \]

Finally, the normalized kernel \( \Theta_{d,u} \) and the neighborhood-constrained normalization kernel \( \Phi_{d,u} \) were fused by using the following iterative process:

\[ \Theta_{d,u}(t + 1) = \frac{1}{2} \alpha \left( \Theta_{d,u} \sum_{r \neq u} \Theta_{d,r}(t) \Phi_{d,u}^T(t) + \frac{1}{2} (1 - \alpha) \sum_{r \neq u} \Theta_{d,r}(0) \right), \]  

where \( t \) is the number of iterations, \( \alpha \), a weight parameter between 0 and 1, \( T \), the transpose operator in matrix algebra, and

\[ \Theta_{d,r}(0) = \Theta_{d,r}. \]
TABLE 2 | Performance comparison analysis with PREDICT, NTSIM, and LRSSL.

| Method      | Dataset | AUROC a | AUPRb | SENc | SPEd | PREe | ACCf | F1g | MCC h |
|-------------|---------|---------|-------|------|------|------|------|-----|-------|
| PREDICT     | PREDICT | 0.902   | 0.151 | 0.341 | 0.993 | 0.091 | 0.992 | 0.144 | 0.163 |
| NTSIM       | PREDICT | 0.921   | 0.338 | 0.368 | 0.999 | 0.462 | 0.998 | 0.402 | 0.407 |
| DDA-SKF     | PREDICT | 0.929 ± 0.0015j | 0.497 ± 0.0043 | 0.455 ± 0.0072 | 0.996 ± 0.0002 | 0.565 ± 0.0120 | 0.991 ± 0.0002 | 0.504 ± 0.0051 | 0.503 ± 0.0054 |
| LRSSL       | LRSSL   | 0.825   | 0.179 | 0.217 | 0.999 | 0.199 | 0.998 | 0.202 | 0.203 |
| NTSIM       | LRSSL   | 0.902   | 0.269 | 0.308 | 0.999 | 0.376 | 0.999 | 0.338 | 0.337 |
| DDA-SKF     | LRSSL   | 0.931 ± 0.0011 | 0.382 ± 0.0051 | 0.395 ± 0.0084 | 0.997 ± 0.0002 | 0.467 ± 0.0038 | 0.997 ± 0.0002 | 0.427 ± 0.0057 | 0.426 ± 0.0041 |

aAUROC: area under receiver operating characteristics.
bAUPR: area under precision-recall curve.
cSEN: sensitivity.
dSPE: specificity.
ePRE: precision.
fACC: accuracy.
gF1: F1-score.
hMCC: Matthew’s correlation coefficient.
iThe performance values are taken from Zhang et al. (2018a).
jThe value is represented as “average ± standard deviation” form for 5 times of repeated cross-validations.

After $z$ rounds of iterative computations, we obtained the integration kernel as follows:

$$
\Theta_d = \frac{1}{3} \left( \Theta_{d,1}(z) + \Theta_{d,2}(z) + \Theta_{d,3}(z) \right),
$$  \hspace{1cm} (13)

Although useful information was extracted in the fusion process, noise is inevitable simultaneously. The information of $k$ most similar drugs for each drug was further extracted to suppress the noise influence by defining a weight matrix; we defined an indicator function as follows:

$$
w_{d,k}(i,j) = \begin{cases} 
1 & I_{d,1,k}(i,j) = I_{d,2,k}(i,j) = I_{d,3,k}(i,j) = 1 \\
0 & \text{otherwise} 
\end{cases} \hspace{1cm} (14)
$$

Finally, the drug comprehensive similarity kernel was defined as the following formula:

$$
S_{d,k} = \{ \theta_d(i,j)w_{d,k}(i,j) \}_{593 \times 593^3},
$$  \hspace{1cm} (15)

where $\theta_d(i,j)$ is the element in the $i$-th row and the $j$-th column of the matrix $\Theta_d$.

By applying Eqs 6–15 and employing two disease similarities, we obtained the disease comprehensive similarity kernel $S_{p,k}$. The value of $k$ in the computing disease comprehensive similarity kernel is not necessarily the same as that of the drugs.

2.4 Laplacian Regularized Least Squares

In this work, Laplacian regularized least squares (LapRLS) were employed to build the prediction model to uncover the potential
drug–disease associations. Based on the characteristics of the model, we could predict the drug–disease associations from either the drug subspace or disease subspace.

In terms of the drug subspace, the Laplacian similarity matrix $\mathbf{L}_d$ was represented as follows:

$$\mathbf{L}_d = \mathbf{D}_d^{1/2} (\mathbf{D}_d - \mathbf{S}_{d,k}) \mathbf{D}_d^{1/2},$$  \hspace{1cm} (16)

where $\mathbf{D}_d$ is a diagonal matrix whose diagonal elements are the sum of the corresponding row elements of the matrix $\mathbf{S}_{d,k}$.

The drug–disease association prediction matrix $\mathbf{F}_d$ was calculated by minimizing the following loss function:

$$\min_{\mathbf{F}_d} \left( \| \mathbf{A} - \mathbf{F}_d \|_F^2 + \beta_d \| \mathbf{F}_d^T \mathbf{L}_d \mathbf{F}_d \|_F^2 \right).$$  \hspace{1cm} (17)

where $\mathbf{A}$ is the original drug–disease association matrix, $\mathbf{F}_d$ the predicted association matrix in the drug subspace, $\beta_d$ the weighting coefficient, and $\| \cdot \|_F$ the F-norm operator. The first term of the loss function aims to reduce the difference between the original matrix and prediction matrix. The second term is used to avoid the over-fitting problem.

The derivation of the optimization algorithm of LapRLS was introduced in Xia et al. (2010). We can calculate the predicted association matrix in the drug subspace by using the following formula:

$$\mathbf{F}_d = \mathbf{S}_{d,k} (\mathbf{S}_{d,k} + \beta_d \mathbf{L}_d \mathbf{S}_{d,k})^{-1} \mathbf{A}.$$  \hspace{1cm} (18)

Similarly, by using Eqs 16–18 on the disease subspace, the predicted association matrix $\mathbf{F}_p$ in the disease subspace was calculated as follows:

$$\mathbf{F}_p = \mathbf{S}_{p,k} (\mathbf{S}_{p,k} + \beta_p \mathbf{L}_p \mathbf{S}_{p,k})^{-1} \mathbf{A},$$  \hspace{1cm} (19)

where $\mathbf{L}_p$ is the normalized similarity matrix in the disease subspace, $\mathbf{L}_d$ was defined as follows:

**TABLE 4** | Top five candidate diseases for typical orphan drugs.

| Orphan drug | Disease | OMIM ID | Evidence |
|-------------|---------|---------|----------|
| Celecoxib   | Osteoarthritis susceptibility 2 | 140600   | CTD*     |
| Celecoxib   | Osteoarthritis susceptibility 1 | 168720   | CTD      |
| Celecoxib   | Progressive pseudorheumatoid dysplasia | 208230 | NA*      |
| Celecoxib   | Mitochondrial recessive ataxia syndrome | 607459 | NA       |
| Celecoxib   | Osteoarthritis susceptibility 3 | 607850   | CTD      |
| Methotrexate| Mismatch repair cancer syndrome 1 | 276300   | NA       |
| Methotrexate| Breast neoplasms | 114480   | CTD      |
| Methotrexate| Acute lymphoblastic leukemia | 247640 | NA       |
| Methotrexate| Autoimmune diseases | 109100   | CTD      |
| Methotrexate| Pyogenic arthritis–pyoderma gangrenosum–acne | 604416 | NA       |
| Doxorubicin | Mismatch repair cancer syndrome 1 | 276300   | NA       |
| Doxorubicin | Breast neoplasms | 114480   | CTD      |
| Doxorubicin | Dohle bodies and acute leukemia | 223350  | NA       |
| Doxorubicin | Breast neoplasms | 114480   | CTD      |
| Doxorubicin | Acute myeloid leukemia | 601620  | CTD      |

*CTD*: available on the Comparative Toxicogenomics Database.

**TABLE 5** | Top five candidate drugs for complex diseases.

| Disease | Drug | Evidence* |
|---------|------|-----------|
| Alzheimer’s disease | Pyridostigmine | Agnoli et al. (1983) |
| Alzheimer’s disease | Benzatropine | Chen and Yeong, (2020) |
| Alzheimer’s disease | Scopolamine | El-Moursy et al. (2009) |
| Alzheimer’s disease | Carbipoda | Bennett et al. (2016) |
| Parkinson’s disease | Biperiden | Kostelnik et al. (2017) |
| Parkinson’s disease | Levodopa | LeWitt, (2015) |
| Parkinson’s disease | Bromocriptine | Lieberman and Goldstein, (1985) |
| Parkinson’s disease | Trihexyphenidyl | Jilani et al. (2021) |
| Parkinson’s disease | Rivastigmine | Espay et al. (2021) |
| Amyotrophic lateral sclerosis | Baclofen | Bousicault et al. (2020) |
| Amyotrophic lateral sclerosis | Meleltine | Adiao et al. (2020) |
| Amyotrophic lateral sclerosis | Colchicine | Mandrioli et al. (2019) |
| Amyotrophic lateral sclerosis | Ranolazine | Djamegoz and Orkai, (2013) |
| Amyotrophic lateral sclerosis | Prilocaine | NA |

*Evidence: Evidences from the literature.

*NA: Evidences are not available.
\[ L_p = D_p^{-1/2} (D_p - S_{p,k}) D_p^{-1/2}, \]  

where \( D_p \) is the diagonal matrix of the matrix \( S_{p,k} \).

With the two predicted association matrices \( F_d \) and \( F_p \), we defined the final predicted association matrix as follows:

\[ F = \lambda F_p + (1 - \lambda) F_d, \]

where \( \lambda \in (0, 1) \) is a weighting parameter.

### 2.5 Performance Evaluation

We used novel association prediction and novel drug prediction to estimate the prediction performance of our method. In the novel association prediction, we applied both 5-fold and 10-fold cross-validation schemes. For the 10-fold cross validation, all drug–disease associations in the benchmarking dataset were divided into 10 nonoverlapping subsets randomly with almost the same size. While for the 5-fold cross validation, the number of the nonoverlapping subsets is five. The novel drug prediction was used to evaluate the prediction performance on new drugs. All drugs, not the drug–disease associations, were randomly divided into 10 subsets of approximately equal size. In each trial, one set was used in turn to act as the testing set, and other sets were used as the training set. In all aforementioned cross-validations, when one subset was used as the testing set, all prior knowledge of the testing set was removed before computing the association similarity kernels for drugs and diseases. The known associations corresponding to the testing set were reset to unknown. This guarantees that there is no information leak in the testing process.

All pairs of drug–disease associations were scored. Given a threshold, the drug–disease pair with a score larger than the threshold is predicted to be associated, otherwise nonassociated.

Due to the requirement of performance comparison and the performance values that are reported by the existing methods, we took a different set of performance measures in different contexts.
We used the AUROC (area under receiver operating characteristics) curve and AUPR (area under precision-recall) curve to evaluate the prediction performance of our method in the context of 10-fold cross-validations. AUROC and AUPR are threshold-free metrics that are capable of measuring the overall performance of prediction models (Jiao and Du, 2016). While AUROC is powerful and popular as a main performance evaluation index, the value of AUROC may be misleading when the dataset is highly imbalanced (Saito and Rehmsmeier, 2015). It should be noticed that the value of AUPR can be used as an alternative metric to evaluate the prediction performance when the data is imbalanced (Davis and Goadrich, 2006). The value of AUPR tends to be smaller relative to the value of AUROC because of the highly imbalanced dataset (Saito and Rehmsmeier, 2015). In the context of 5-fold cross-validations, we applied several additional performance measures other than the AUROC and the AUPR. Six statistics, including sensitivity (SEN), specificity (SPE), precision (PRE), accuracy (ACC), F1-score (F1), and Matthew’s correlation coefficient (MCC), were applied in measuring the prediction performance of DDA-SKF in the context of 5-fold cross-validations. The threshold was optimized to maximize the F1-score. These performance measures were defined as follows:

\[ SEN = \frac{TP}{TP + FN} \]  

(22)
In this section, to evaluate the prediction performance of our model on the benchmarking datasets using both novel association prediction and the novel drug prediction. We compared our model against other state-of-the-art methods, including BNNR.
(Yang et al., 2019a), DisDrugPred (Xuan et al., 2019), DRRS (Luo et al., 2018), SCMFDD (Zhang et al., 2018b), MBiRW (Luo et al., 2016), and DRIMC (Zhang et al., 2020). We also compared our method against the PREDICT (Gottlieb et al., 2011), LRSSL (Liang et al., 2017), and NTSIM methods (Zhang et al., 2018a).

To eliminate the randomness in the cross-validation, we repeated each cross-validation five times with different random data partition schemes. The average value and the standard deviations of every performance measure were reported. However, it should be noted that the standard deviations were not reported by other methods in comparison.

For the novel association prediction, the evaluation results of all the methods on the benchmarking dataset are collected in Table 1. As in Table 1, DDA-SKF obtained good values in AUROC and AUPR. DDA-SKF achieved an AUROC of 0.937, which is 0.536, 5.281, 0.861, 31.601, and 2.854%, respectively, higher than that of BNNR’s 0.932, DisDrugPred’s 0.890, DRRS’s 0.929, SCMFDD’s 0.712, and MBiRW’s 0.911. DDA-SKF also has an AUPR of 0.533, which is only slightly lower than that of BNNR. Therefore, we believe that DDA-SKF has a better, or at least comparable, performance to all state-of-the-art methods in novel association prediction. The integration of disease and drug similarity kernels is effective.

For the novel drug prediction, we evaluated the performance of all models in the benchmarking dataset. The results are recorded in Table 1. DDA-SKF achieved an AUROC of 0.845, which is 8.892, 1.198, 10.458, 15.280, and 5.890%, respectively, higher than that of BNNR’s 0.776, DisDrugPred’s 0.835, DRRS’s 0.765, SCMFDD’s 0.733, and MBiRW’s 0.798. DDA-SKF also obtained an AUPR of 0.270, which is only slightly lower than that of DRIMC. Therefore, we conclude that DDA-SKF outperformed most state-of-the-art methods in novel drug predictions.

In the comparison with PREDICT, LRSSL, and NTSIM methods, 5-fold cross-validation was applied rather than 10-fold cross-validations, as all three existing methods reported their performances in 5-fold cross validations. According to the results in Table 2, DDA-SKF achieved better performances in almost every comparison. DDA-SKF achieved an AUROC of 0.929 and 0.931 for the PREDICT dataset and LRSSL dataset, which are 0.869 and 3.215% higher than that of the second model NTSIM, respectively. As for AUPR, DDA-SKF achieves an AUPR of 0.497 and 0.382 on the PREDICT dataset and LRSSL dataset, which are the best values among the four. For the F1-score, DDA-SKF achieves 0.504 and 0.427 for the PREDICT dataset and LRSSL dataset, which are 25.373 and 26.331% higher than that of NTSIM, respectively. In summary, DDA-SKF achieved better performances in this comparison.

### 3.2 Case Study

Although the prediction performances of DDA-SKF in terms of AUROC and AUPR are slightly lower than those of the best method for the novel association prediction and novel drug prediction, our method can work without enough disease similarities. This is useful when the orphan drugs, whose indication is still very limited, are considered.

In this section, the capability of our model in predicting novel drug–disease associations is tested. We performed three tests based on the PREDICT dataset in this part: one is drug-repositioning prediction, the second is orphan drug indications prediction, and the last is complex disease prediction.

To predict novel indications for all drugs, all known drug–disease associations in the benchmarking dataset were used as the training set, and the unknown drug–disease associations were regarded as the candidate set. DDA-SKF was applied to obtain scores for all candidate drug–disease associations. All candidate associations were ranked according to the prediction scores. Top-ranked candidate associations were identified as novel drug–disease associations.

We verified newly predicted associations by comparing them against the Comparative Toxicogenomics Database (CTD) (Davis et al., 2019). CTD contains curated and inferred chemical–disease relationships, which are divided into marker/mechanism and therapeutic. We only compared the therapeutic relationships for verification. For each of the 593 drugs, we collected the top-5 and the top-20 prediction results. The predictions for all drugs were listed in Supplementary Table S1 and Supplementary Table S2 as OMIM IDs in supplementary materials. In the DDA-SKF prediction results, 156 of top-5 and 377 of top-20 predictions have been confirmed by CTD.

We listed several top-5 prediction results in Table 3. Although some top-ranked predictions had not been verified, we believe that these may provide new indications for approved drugs. For example, vincristine for B chronic lymphocytic leukemia (CLL), cisplatin for acute lymphoblastic leukemia (ALL), and methotrexate for neuroblastoma. These articles provide some possible hints that consist of our predictions (Lj et al., 1997; Vilpo et al., 2000; Lau et al., 2015).

In particular, DDA-SKF can perform drug–disease association prediction without disease similarities. This may be important in studying some orphan drugs. Since the indications of orphan drugs are usually very limited, the similarities between diseases may be misleading. Therefore, by excluding the disease similarities from the model, DDA-SKF should have more potential in finding new indications for orphan drugs, which may eventually decrease the cost of orphan drug utilization.

We selected several orphan drugs from Orphanet (http://www.orpha.net), including celecoxib, methotrexate, and doxorubicin. For each orphan drug, all known drug–disease associations were removed. The orphan drug would have no association information during the process of prediction. We fed the drug similarity kernels with orphan drug information and the aforementioned association matrix to our model. The results of top-20 predictions for each orphan drug were recorded in Supplementary Table S3 in supplementary materials. The top-5 predicted results are summarized in Table 4. These successful prediction instances further confirm that DDA-SKF has the potential to predict novel indications for orphan drugs without disease similarity information.

Last, several commonly studied complex diseases were chosen to evaluate DDA-SKF. For instance, Alzheimer’s disease (AD), Parkinson’s disease (PD), and amyotrophic lateral sclerosis (ALS) were considered. We listed the top-5 predicted results in Table 5 along with the literature evidences. For each disease, the uncovered associations will be ranked.
based on the prediction scores, and the top-ranked unknown associations were identified as the prediction results. The AD and PD results were obtained on the PREdict dataset, while the ALS results were obtained using the LRSSL dataset. These results indicate that DDA-SKF has the potential to uncover new drugs for these complex diseases.

### 3.3 Comparison With Single Similarity

In this work, different drug similarity kernels were integrated as a drug comprehensive similarity kernel in the drug subspace. Also, different disease similarity kernels were integrated as a disease comprehensive similarity kernel. To demonstrate the effectiveness in integrating multiple similarity kernels, we calculated the values of AUROC and AUPR with every single similarity kernel on the PREdict dataset. The results are shown in Figure 2 and Figure 3. We also compared the effectiveness of the multiple similarity kernels fusion and all single similarity kernels. The results are recorded in Table 6. We can see that the comprehensive kernel produced higher performances than every single kernel.

### 4 CONCLUSION

In this study, we proposed DDA-SKF (drug–disease associations prediction based on the similarity kernels fusion) for predicting drug–disease associations. Several similarity kernels of drugs and diseases were integrated. The SKF method has a better, or at least comparable, performance than the existing methods, in terms of AUROC and AUPR. The novel drug prediction test indicated that DDA-SKF can identify potential indications of approved drugs, which may be useful in drug repositioning. The DDA-SKF can also make predictions without disease similarity. This allows it to be applied on orphan drugs, which is useful in exploring potential indications of such drugs. The evaluation on several complex diseases illustrates that our method can provide valuable information and potential indications for clinical studies. However, it should be noted that there is still a room for further improvement. One is to integrate similarities from more biological knowledge, and the other is to integrate information of the drug-target information. Due to the limited time and resources of this study, these works will be conducted in future.

### DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/GCQ2119216031/DDA-SKF.

### AUTHOR CONTRIBUTIONS

C-QG collected the data, implemented the algorithm, performed the experiments, analyzed the results, and partially wrote the manuscript; Y-KZ designed the algorithm and analyzed the results; X-HX analyzed the results and partially wrote the manuscript; HM partially collected the data and partially analyzed the results; P-FD directed the whole study, conceptualized the algorithm, supervised the experiments, analyzed the results, and wrote the manuscript.

### FUNDING

This work was supported by the National Natural Science Foundation of China (No. NSFC 61872268) and National Key R&D Program of China (No. 2018YFC0910405).

### SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2021.784171/full#supplementary-material

### REFERENCES

Agnoli, A., Martucci, N., Manna, V., Conti, L., and Fioravanti, M. (1983). Effect of Cholinergic and Anticholinergic Drugs on Short-Term Memory in Alzheimer’s Dementia: a Neuropsychological and Computerized Electroencephalographic Study. *Clin. Neuropharmacol.*, 6, 311–323. doi:10.1097/00002826-198312000-00005

Ashburn, T. T., and Thor, K. B. (2004). Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. *Nat. Rev. Drug Discov.* 3, 673–683. doi:10.1038/nrd1468

Attwood, M. M., Rask-Andersen, M., and Schiöth, H. B. (2018). Orphan Drugs and Their Impact on Pharmaceutical Development. *Trends Pharmacol. Sci.* 39, 1077–1535. doi:10.1016/j.tips.2018.03.003

B Adiao, K. J., I Espiritu, A., and C Bagnas, M. A. (2020). Efficacy and Safety of Mexiletine in Amyotrophic Lateral Sclerosis: a Systematic Review of Randomized Controlled Trials. *Neurodegener Dis. Manag.* 10, 397–407. doi:10.2217/nmt-2020-0026

Bennett, J., Burns, J., Welch, P., and Bothwell, R. (2016). Safety and Tolerability of R(-) Pramipexole in Mild-To-Moderate Alzheimer’s Disease. *J. Alzheimers Dis.* 49, 1179–1187. doi:10.3233/JAD-150788

Boussicault, L., Laffaire, J., Schmitt, P., Rinaudo, P., Callizot, N., Nabirotchkin, S., et al. (2020). Combination of Acamprosate and Baclofen (PXT864) as a Potential New Therapy for Amyotrophic Lateral Sclerosis. *J. Neurosci. Res.* 98, 2435–2450. doi:10.1002/jnr.24714

Chen, W. N., and Yong, K. Y. (2020). Scopolamine, a Toxin-Induced Experimental Model, Used for Research in Alzheimer’s Disease. *CNS Neurol. Disord. Drug Targets* 19, 85–93. doi:10.2174/187152731966620014104331

Chong, C. R., and Sullivan, D. J. (2007). New Uses for Old Drugs. *Nature* 448, 645–646. doi:10.1038/448645a

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegens, J., et al. (2019). The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Res.* 47, D948–D954. doi:10.1093/nar/gky868

Davis, J., and Goadrich, M. (2006). “The Relationship between Precision-Recall and ROC Curves,” in Proceedings Of the 23rd International Conference on Machine Learning ICML ’06 (New York, NY, USA: Association for Computing Machinery). 233–240. doi:10.1145/1143844.1143874

Djangoz, M. B., and Onkal, R. (2013). Persistent Current Blockers of Voltage-Gated Sodium Channels: a Clinical Opportunity for Controlling Metastatic Disease. *Recent Pat Anticancer Drug Discov.* 8, 66–84. doi:10.2174/15748928130107
Yang, M., Luo, H., Li, Y., Wu, F. X., and Wang, J. (2019b). Overlap Matrix Completion for Predicting Drug-Associated Indications. *Plos Comput. Biol.* 15, e1007541. doi:10.1371/journal.pcbi.1007541

Yap, C. W. (2011). PaDEL-descriptor: an Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* 32, 1466–1474. doi:10.1002/jcc.21707

Yu, Z., Huang, F., Zhao, X., Xiao, W., and Zhang, W. (2021). Predicting Drug-Disease Associations through Layer Attention Graph Convolutional Network. *Brief Bioinform.* 22, bbaa243. doi:10.1093/bib/bbaa243

Zhang, W., Xu, H., Li, X., Gao, Q., and Wang, L. (2020). DRIMC: an Improved Drug Repositioning Approach Using Bayesian Inductive Matrix Completion. *Bioinformatics* 36, 2839–2847. doi:10.1093/bioinformatics/btaa062

Zhang, W., Yue, X., Huang, F., Liu, R., Chen, Y., and Ruan, C. (2018a). Predicting Drug-Disease Associations and Their Therapeutic Function Based on the Drug-Disease Association Bipartite Network. *Methods* 145, 51–59. doi:10.1016/j.ymeth.2018.06.001

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018b). Predicting Drug-Disease Associations by Using Similarity Constrained Matrix Factorization. *BMC Bioinformatics* 19, 233. doi:10.1186/s12859-018-2220-4

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gao, Zhou, Xin, Min and Du. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.