A Machine Learning-Based Biological Drug–Target Interaction Prediction Method for a Tripartite Heterogeneous Network

Ying Zheng and Zheng Wu

ABSTRACT: Drug repositioning is the identification of interactions between drugs and target proteins in pharmaceutical sciences. Traditional large-scale validation through chemical experiments is time-consuming and expensive, while drug repositioning can drastically decrease the cost and duration taken by traditional drug development. With the rapid advancement of high-throughput technologies and the explosion of various biological and medical data, computational drug repositioning methods have been used to systematically identify potential drug–target interactions. Some of them are based on a particular class of machine learning algorithms called kernel methods. In this paper, we propose a new machine learning prediction method combining multiple kernels into a tripartite heterogeneous drug–target–disease interaction spaces in order to integrate multiple sources of biological information simultaneously. This novel network algorithm extends the traditional drug–target interaction bipartite graph to the third disease layer. Meanwhile, Gaussian kernel functions on heterogeneous networks and the regularized least square method of the Kronecker product are used to predict new drug–target interactions. The values of AUPR (area under the precision–recall curve) and AUC (the area under the receiver operating characteristic curve) of the proposed algorithm are significantly improved. Especially, the AUC values are improved to 0.99, 0.99, 0.97, and 0.96 on four benchmark data sets. These experimental results substantiate that the network topology can be used for predicting drug–target interactions.

1. INTRODUCTION

In the past few decades, financial investment in drug research and development has increased dramatically. However, the increasing demand for new drugs still cannot be met.1 Drug repositioning is a creative and resourceful approach to increase the number of therapies by exploiting available and approved drugs.2 Since the safety and effectiveness of specific signs have been tested and approved, the investment risk can also be greatly reduced.3,4 Drug repositioning aims to identify new therapeutic opportunities for existing drugs, which can reduce the time, costs, and risk of traditional drug development and shorten the period of drug approval and launch.5 Many success cases of drug repositioning greatly inspired the global pharmaceutical industries to explore the new uses of the existing drugs.6 Moreover, research on drug repositioning can provide biologists and pharmacists with drug–target interaction candidates for further research and clinical trials.7

Traditional experimental methods are mainly divided into target-based prediction methods and ligand-based prediction methods.8–14 Target-based prediction methods include docking or reverse docking, but they need complete 3D structures of the target, and the performance of the method is poor when facing some unknown target. Ligand-based prediction methods mainly rely on comprehensive information about the drug, but when the ligand information is inadequate, this method is ineffective.15–19

Recently, machine learning has been widely applied in many fields.20–24 Machine learning methods and network-based inference methods have been successfully introduced in drug repositioning. The machine learning methods can extract topological structural features, and the features can be used to calculate the similarities of the drugs and the targets.25 The chemical–chemical interactions and chemical–protein interactions are used to select the candidate drugs that have association with approved lung cancer drugs and related genes,26 and the permutation test and K-means clustering algorithm are introduced to exclude candidate drugs with low possibilities of treating lung cancer. Two distributed label propagation algorithms for heterogeneous networks named DHLP-1 and DHLP-2 are developed. Additionally, they measured the efficiency of DHLP-1 and DHLP-2 algorithms on a biological network consisting of drugs, diseases, and targets.27 A method named deepDR28 is proposed to learn high-level features of drugs from the heterogeneous networks by a multimodal deep autoencoder, which achieved a mean...
The rest of this paper is organized as follows. In Section 2, we introduce some related work, and then we present the general framework and relevant methods with detail in Section 3. In Section 4, the performance of our proposed THN_KRLS method is evaluated through extensive experiments. Some discussion is also provided in Section 5, and Section 6 concludes this paper.

2. RELATED WORK

2.1. Data Sets. In our experiments, we use the data sets with different focuses to build a tripartite heterogeneous network model and evaluate the performance of THN_KRLS on benchmark data sets. These benchmark data sets are named after four main targets: enzymes, ion channels, GPCR (the G protein-coupled receptors), and nuclear receptors. Table 1 shows the standard data sets we used during the experiment. Also, for the imported data for the disease layer, the main resource of data set is DisGeNET. In order to compare and evaluate the performance with other algorithms, we used the benchmark data sets proposed in ref.34 Table 2 gives the benchmark data sets, which are downloaded from ChEMBL.34

Another database that needs to be mentioned here is DrugBank.35 The version we used is 5.1.4, which was released on 2019/07/02. It contains 13,463 drug entries, which includes 2621 approved small molecule drugs and 1349 approved targets, biological preparations (proteins, peptides, vaccines, and allergens), 130 kinds of nutritional drugs, and more than 6350 kinds of experimental drugs (discovery stage).

2.2. Tripartite Heterogeneous Network. Based on the related ideas of pharmacology, the therapeutic effect of a single drug is relatively limited for diseases that are complex multiple pathological.36–38 Therefore, at the pharmacological perspective, it is necessary to set up a network for multiple drugs acting on multiple target proteins caused by complex diseases.39–41 Hence, the problem of drug repositioning can be formulated as the problem of predicting missing graph edges in graph theory, that is, predicting possible edges on a bipartite network. Figure 1a is the part of the visualization of the ion channel data set, in which the red node is the drug and the green node is the target; the line from the red node to the green node indicates the drug−target interaction. Figure 1b is the bipartite graph model of a part of Figure 1a, the red node in Figure 1b is the target, and the black node is the drug.
In this paper, we extend the previous drug–target bipartite structure model to a tripartite drug–target–disease heterogeneous network. It is worth mentioning that the disease layer plays a role in setting up the target–target interaction space. Additionally, drug repositioning aims to discover that the old drugs whether can act on new targets or not, the tripartite heterogeneous structure helps to provide a new direction for this, that is, we can predict new drug–disease interactions directly in the target layer without loss.

In a tripartite heterogeneous network, we mainly focus on the drug–target interaction. Disease layer and similarity of drug chemical structure are introduced as an auxiliary characteristic matrix. We hope to show the diversity of the reason why the chemical structure similarity and the disease-target interaction due to their similar chemical structure or the two targets are related to a certain disease before.

However, it is still difficult for computers to predict new drugs or targets from scratch. Since for an abstract network, a new drug or new target can only be introduced as an isolated node. It is difficult to explain that it can interact with other drugs or targets based on its network structure alone. That is the reason why the chemical structure similarity and the disease layer are introduced to construct a tripartite heterogeneous network structure.

Therefore, we construct a tripartite network that included three types of vertices: drugs, targets, and diseases. Correspondingly, two types of associations, drug–target interaction and disease–target interaction, are used as the edges to connect the vertices, as Figure 2 shows. The network is constructed based on the knowledge (i.e., associations) from two existing knowledge base ChEMBL (Table 3) and DisGeNET (Table 2).

As can be seen in Figure 2, $\text{drug} = \{d_1, d_2, ..., d_m\}$ represents the drug layer, $\text{target} = \{t_1, t_2, ..., t_n\}$ represents the target layer, and $\text{disease} = \{r_1, r_2, ..., r_k\}$ represents the disease layer. In this example, the solid line represents the similarity within the level, and the dotted line represents the interaction between the two levels.

### 3. RELEVANT MODEL AND METHODS

#### 3.1. Similarity of Medicinal Chemical Structures

In order to obtain the final prediction score, this chapter analyzes the drug–target heterogeneous network structure and related biological characteristics from multiple dimensions to form multiple similarity matrices. Both the drug–target interaction and disease-target interaction can be formulated to the bipartite graphs. On the heterogeneous network, we use the Gaussian kernel to make the similarity matrix space and use the Kronecker product based on the regularized least square classification to predict the highest score of the drug–target interaction. Finally, a 10-fold cross-validation was performed on the existing results.

In the THN_KRLS model, it is necessary to ensure that the chemical structure of all drugs is unique and easy to handle. Due to the existence of heterogeneous structures in the chemical formula, the simplified molecular-input line-entry system (SIMILES) is adopted here. That is, the specification of the molecular structure is clearly described by ASCII strings to ensure that each chemical structure has a unique corresponding string. Moreover, the corresponding string is converted into a string of 166 bits binary chemical fingerprint, in which each bit matches to a specific molecular feature. This also guarantees the uniqueness of the chemical structure. Finally, the Tanimoto coefficient is the result of the calculation of binary vectors. The specific calculation formula is

![Figure 1. Example of a bipartite graph model for drug–target interactions.](image1)

![Figure 2. Tripartite heterogeneous network model.](image2)

### Table 3. Result of FLapRLS, RLS-Kron, and THN_KRLS

| data sets | method  | AUC  | sensitivity | specificity | AUPR  |
|-----------|---------|------|-------------|-------------|-------|
| enzyme    | FLapRLS | 0.985| 0.913       | 0.999*      | 0.92  |
|           | RLS-Kron| 0.978| 0.905       | 0.997       | 0.915 |
|           | THN_KRLS| 0.99 | 0.979*      | 0.998*      | 0.99* |
| ion channel| FLapRLS | 0.991*| 0.688       | 0.986       | 0.89  |
|           | RLS-Kron| 0.984| 0.721       | 0.98        | 0.943 |
|           | THN_KRLS| 0.99 | 0.977*      | 0.998*      | 0.99* |
| GPCR      | FLapRLS | 0.944| 0.737       | 0.986       | 0.83  |
|           | RLS-Kron| 0.954| 0.753       | 0.975       | 0.79  |
|           | THN_KRLS| 0.97 | 0.934*      | 0.990*      | 0.97* |
| nuclear receptor| FLapRLS | 0.746| 0.52        | 0.915       | 0.608 |
|           | RLS-Kron| 0.92 | 0.713       | 0.937       | 0.684 |
|           | THN_KRLS| 0.96 | 0.930*      | 0.993*      | 0.94* |

*For each data set, * indicates the highest value.
\[ SIM_{\text{chem}} = \frac{|f(dx) \times f(dy)|}{|f(dx)| + |f(dy)| - |f(dx) \times f(dy)|} \]  
(1)

where \( f(dx) \) is the binary chemical fingerprint of drug \( x \). Hence, a matrix of chemical structure similarity is constructed for all drugs.

3.2. Drug/Target Gaussian Kernel Similarity. The similarity of the drug chemical structure cannot be the only measure of similarity matrix.\(^\text{42} \) Therefore, the interaction between the drug and target must be analyzed and calculated based on the bipartite graph structure. The Gaussian kernel is defined as the unimodal of the Euclidean distance between any two points in the space. The specific calculation formula is as follows:

\[ K_{\text{GIP},d}(D_i, D_j) = \exp(-\gamma d_i^2 - yd_j^2) \]  
(2)

\[ \gamma = \gamma_d' \left( \frac{1}{m} \sum_{i=1}^{m} |yd_i|^2 \right) \]  
(3)

\[ K_{\text{GIP},t}(T_i, T_j) = \exp(-\gamma t_i^2 - yt_j^2) \]  
(4)

\[ \gamma = \gamma_t' \left( \frac{1}{n} \sum_{i=1}^{n} |yt_i|^2 \right) \]  
(5)

where \( D_i \) is the \( i \)th drug in the drug set and \( m \) is the size of the drug set. \( T_i \) is the \( i \)th target in the target set, and \( n \) is the size of the target set. The adjacency matrix \( Y \in m \times n \) represents the known drug-target interactions. If the drug and the target have an existing interaction, then the value is 1, otherwise the value is 0. \( yd_i = \{y_{i1}, y_{i2}, ..., y_{in}\} \) is defined as the correlation vector between the drug \( d_i \) and all targets. \( yd_i \) and \( yt_i \) are the adjustment parameters that controls the width of the kernel, where \( \gamma_d' \) and \( \gamma_t' \) are set to 1 according to the experience of using Gaussian kernel.

3.3. Disease–Target Similarity. The same with the drug–target similarity, the human disease data set describes the interactions between the diseases and the targets. The disease–target network is constructed on this interaction, and the Gaussian kernel is also calculated. The specific calculation formula is as follows:

\[ K_{\text{GIP},d}(T_i, T_j) = \exp(-\gamma_d y_t^2 - yd_i^2) \]  
(6)

\[ \gamma_d = \gamma_d' \left( \frac{1}{k} \sum_{i=1}^{k} |yd_i|^2 \right) \]  
(7)

We set \( S = \{t_{i1}, t_{i2}, ..., t_{ik}\} \) as the set of targets derived from the disease–target interaction, and \( k \) is the number of targets; the adjacency matrix \( Y \in k \times n \) represents the known disease–target association. If there is a known correlation between the target \( t_i \) and the disease \( d_k \), then the value of \( \gamma_d \) is 1, otherwise the value is 0. \( yd_i = \{y_{i1}, y_{i2}, ..., y_{in}\} \) is defined as the correlation vector between the target \( t_i \) and all diseases. \( yd_i \) is an adjustment parameter that controls the width of the kernel, and \( \gamma_d \) set to 1 according to the experience of using Gaussian kernels.

3.4. Similarity Matrix Fusion. We construct the kernel containing the spatial information of the drug and the target from the above multiple similarity matrices. Since the similarity matrix is not a positive definite matrix, a two-class prediction is required in the end. We linearly fit the drug chemical structure similarity matrix and the drug Gaussian nuclear similarity matrix and the target Gaussian kernel similarity matrix and the disease Gaussian kernel similarity matrix separately, and we set the weighted factors empirically.

\[ SIM_{\text{drug}}(d_i, d_j) = (1 - \alpha) \times K_{\text{GIP},d}(d_i, d_j) + \alpha \times SIM_{\text{chem}}(d_i, d_j) \]  
(8)

\[ SIM_{\text{tan}}(t_i, t_j) = (1 - \alpha) \times K_{\text{GIP},t}(t_i, t_j) + \alpha \times K_{\text{GIP},d}(t_i, t_j) \]  
(9)

In the latter experiments, we adopt the standard of equal distribution for the parameter fitting constructed by all similarity spaces, that is, the ratio of 0.5:0.5 is used when fitting the similarity matrix.

3.5. Regularized Least Square Method of the Kronecker Product. Since the similarity matrix is a non-positive definite square matrix, a multiple similarity matrix should be merged into a large similarity matrix. Therefore, a more appropriate method named the regularized least square method is used to calculate the Kronecker product of the two matrices.

The final Kronecker product is expressed as \( W = W_d \otimes W_t \), where \( W \) is a matrix of size \((MN \times MN)\). Each position in the matrix represents the specific score of the drug pair \((d_i, d_j)\) multiplied by the target \((t_p, t_q)\) pair. The number of drugs \( M \) and targets \( N \) are on the size of \( 10^4-10^5 \) due to the current external data sets, so the final Kronecker product calculation is difficult to directly compute, store, and operate. Therefore, it is necessary to perform eigenvalue decomposition here. There are \( W_d = v_d \lambda_d v_d^T \) and \( W_t = v_t \lambda_t v_t^T \), where \( \lambda \) is an orthogonal matrix composed of the eigenvectors of the drug or target and \( \lambda \) is a diagonal matrix composed of the eigenvalues of the drug or target. Therefore, the final Kronecker product result is

\[ W = v \lambda v^T \]  
(10)

where \( \lambda = \lambda_d \otimes \lambda_t \). However, the size of the calculated matrix is still maintained \((MN \times MN)\), and the result is still difficult to save and operate. It is necessary to introduce the regularized least square method for calculation, and it is defined as follows:

\[ VEC(Y^T) = W(W + s\lambda)^{-1}VEC(Y^T) \]  
(11)

where \( \sigma \) is the regularization parameter and \( VEC(Y^T) \) is the column direction formed by stacking all the columns of matrix \( Y \), so eqs 10 and 11 are substituted into eq 12:

\[ VEC(Y^T) = (v_d \otimes v_t)(\lambda_d \otimes \lambda_t)(\lambda_d \otimes \lambda_t + s\lambda)^{-1} \]

\[ (v_d \otimes v_t)VEC(Y^T) \]  
(12)

According to the properties of the matrix equation of the Kronecker product and the Kronecker product, transposition operation transforms to the distribution law as follows:

\[ (B^T \otimes A)VEC(Y^T) = VEC(AY^TB) \]  
(13)

\[ (v_d \otimes v_t)^T = v_d^T \otimes v_t^T \]  
(14)

Therefore, eq 12 can be simplified as
VEC(Ŷ\trans) = (ν_d \otimes ν_t)(λ_d \otimes λ_t)(λ_d \otimes λ_t + σI)^{-1}V
EC(ν_d \trans Ŷ \trans ν_d)

Let X = (λ_d \otimes λ_t)(λ_d \otimes λ_t + σI)^{-1}VEC(ν_d \trans Ŷ \trans ν_d), matrix equation properties can still be used when X is a column vector, then VEC(Ŷ\trans) = (ν_d \otimes ν_t)X has

VEC(Ŷ\trans) = VEC(ν_d \times ν_d\trans)

If the drug−target pair has a higher prediction result score, then it indicates that they have a higher possibility of interaction. The prediction result can be achieved.

Based on the above similarity matrix, the complexity of our algorithm is O(n^2), where n is the sum of the size of drug, target, and disease. In the later experiment, the regularized least square method of the Kronecker product is used. It is easy to see that the algorithm can be applied to large-scale data sets.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Evaluation Criteria. In order to compare the performance of these methods, we conduct systematic experiments to simulate the analysis process of biological data on four heterogeneous interaction networks. In each run of the method, each drug−target pair (interacting or non-interacting) in the test sets is excluded by setting it to not exist in the adjacency matrix. Then, we try to use the remaining data to restore its true label.

Note that during the 10-fold cross-validation process, the matrix needs to be rebuilt based on the current training set, so the Gaussian kernel must be recalculated.

We assess the performance of the methods with two quality measures generally used in this paper: AUC and AUPR (the area under the precision−recall curve). We compute the ROC (receiver operating characteristic) curve and regard the AUC as the main quality measurement. The precision−recall curve is the chart of the true positive rate between all positive predictions of each given recall rate. The AUPR value provides a quantitative assessment. These two kinds of quality measurements have become the standard criteria for evaluating methods.

To systematically evaluate the performances of these methods, we use the 10-fold cross-validation theory. Before the experiment, a heterogeneous network model was constructed on 10 sets of drug−target interaction, and the top N rankings and thresholds were obtained. During each iteration, one set of the drug−target interaction and the top N rankings is obtained after complete training is combined to form the test sets. Additionally, the remaining nine set observation results were regarded as the training set. After we complete the algorithm based on the training set, according to the final association weight to the target, the tested drug is sorted in a descending order with all other drugs. For each specific ranking threshold, if the test rank is higher than the threshold, then it is regarded as a true positive. The number of true positives found in all possible drug−target interactions is regarded as the true positive rate matching to the specified threshold. The ROC curve and the precision−recall curve are constructed from this, and the quality measures AUC and AUPR of the area under the curve are obtained. At the same time, the relevant parameter is 1 empirically in the Gaussian kernel function and the regularized least square method.

In Figure 3, the precision−recall curve of THN_KRLS is the blue curve, the result of FLapRLS is green, and RLS-Kron is red. The areas under the precision−recall curves of these three algorithms are listed in Table 3. Under the same conditions and evaluation criteria, the AUPR results obtained by THN_KRLS have been significantly improved. Additionally, the values of sensitivity on the GPCR data set (Figure 3c) and nuclear receptor data set (Figure 3d) are increased by 0.14 and 0.256, respectively. It is shown that THN_KRLS maintains a high accuracy of prediction.

(c) GPCR (d) Nuclear Receptor

In Figure 4, we use the same standard and color to show the ROC curves of different methods. For AUC values, the THN_KRLS algorithm still maintains a high performance. For example, on the GPCR data set (Figure 4c) and nuclear receptor data set (Figure 4d), the value is increased by 0.016 and 0.04, respectively. On the ion channel data set (Figure 4b), our algorithm is only 0.001 less than the latest best method (FLapRLS). In general, we have better performance on most data sets.

Based on the tripartite heterogeneous network structure and the Gaussian kernel similarity matrix between the three layers, the ROC curves and the AUC values are obtained from the four benchmark data sets. The AUC results are improved to 0.99, 0.99, 0.97, and 0.96 on each data set. The value of sensitivity is increased by 0.181 on the GPCR data set. At the same time, the AUPR results are improved to 0.99, 0.99, 0.97, and 0.94 on each data set. FLapRLS has pointed out that AUPR needs to be separated from the real non-action score in the face of real existence score that has become a more important test performance because it punishes the existence of false positive examples found in the best ranking prediction score. This conclusion proves the high performance of the THN_KRLS method.
From Table 3 and the definitions of sensitivity and specificity, the ROC curve has the best cutoff point in the abscissa: 1-specificity and ordinate: sensitivity. The point means that the ROC curve is closest to the upper left corner. The distance between the best cutoff point and the (0, 1) point in the upper left corner of the coordinate axis is the radius, and the (0, 1) point is the center of the circle to form a quarter circle. The ROC curve is tangent to the quarter circle at the best cutoff point, so the AUC value must be less than 1 minus the area of the semicircle. Here, we can infer that some data may have problems.

As shown in the value of the FLapRLS method on the ion channel data set in Table 3, the ROC curve has the best cutoff point at the point (0.014, 0.688), which is shown in Figure 5. Then, the value of the AUC in any case at this time must be less than 1 minus the area of the quarter circle, which is approximately equal to 0.93. However, the AUC value of FLapRLS is 0.991, which does not match the deduction.

4.2. New Interaction Predictions. In order to analyze the practical relevance of the method for predicting novel drug–target interactions, we conduct an external data validation similar to FLapRLS. Table 4 shows the highest ranking new interaction in the enzyme data set, the bold pairs are the verified drug–target interactions in external data sets such as DrugBank. We verify the results on these data sets. At the same time, four interactions are proven to be existing, which proves that our experiments have a high confidence. In addition, since these data sets are not complete, if a predicted drug–target is not in the verified data sets, then it does not mean that the drug–target has no interaction in reality.

In the top 10 list of the enzymes data set, the verified target is cytochrome P450 2C19 (CYP2C19, UniProt ID: P33261). Also, (S)-mephenytoin hydroxylase is certified to be CYP2C19, which is involved in the metabolism of several clinically useful drugs. Additionally, as a terminal oxygenase, it participates in the synthesis of sterol hormones in the
Table 4. Top 10 New Drug–Target Interactions on the Enzyme Data Set

| enzyme rank | pair & name |
|-------------|-------------|
| DrugBank    | D00394: trimipramine (DrugBank ID DB00726), Hsa:1557: cytochrome P450 2C19 (UniProtKB P33261) |
|             | D00225: alprazolam (DrugBank ID DB00404), Hsa:1557: cytochrome P450 2C19 (UniProtKB P33261) |
| DrugBank    | D00380: tolbutamide (DrugBank ID DB01124), Hsa:1557: cytochrome P450 2C19 (UniProtKB P33261) |
|             | D00394: trimipramine (DrugBank ID DB00726), Hsa:28: histo-blood group ABO system transferase (UniProtKB P16442) |
| DrugBank    | D00225: alprazolam (DrugBank ID DB00404), Hsa:28: histo-blood group ABO system transferase (UniProtKB P16442) |
| DrugBank    | D01071: hexobarbital (DrugBank ID DB01355), Hsa:1557: cytochrome P450 2C19 (UniProtKB P33261) |
|             | D00380: tolbutamide (DrugBank ID DB01124), Hsa:28: histo-blood group ABO system transferase (UniProtKB P16442) |
| DrugBank    | D00574: aminoglutethimide (DrugBank ID DB00357), Hsa:1557: cytochrome P450 2C19 (UniProtKB P33261) |
|             | D01071: hexobarbital (DrugBank ID DB01355), Hsa:28: histo-blood group ABO system transferase (UniProtKB P16442) |
| DrugBank    | D00139: methoxsalen (DrugBank ID DB00553), Hsa:1557: cytochrome P450 2C19 (UniProtKB P33261) |

Table 5. Top 10 New Drug–Target Interactions on the GPCR Data Set

| GPCR rank | pair & name |
|-----------|-------------|
| DrugBank  | D01699: darifenacin (DrugBank ID DB00496), Hsa:1128: muscarinic acetylcholine receptor M1 (UniProtKB P11229) |
|           | D00465: oxybutynin (DrugBank ID DB01062), Hsa:57105: cysteinyl leukotriene receptor 2 (UniProtKB QN5S75) |
|           | D00465: oxybutynin (DrugBank ID DB01062), Hsa:10800: cysteinyl leukotriene receptor 1 (UniProtKB QY2271) |
|           | D00645: bretylium (DrugBank ID DB01158), Hsa:1128: muscarinic acetylcholine receptor M1 (UniProtKB P11229) |
| Kegg      | D00765: rocuronium (DrugBank ID DB00728), Hsa:1128: muscarinic acetylcholine receptor M1 (UniProtKB P11229) |
|           | D001699: darifenacin (DrugBank ID DB00496), Hsa:1129: muscarinic acetylcholine receptor M2 (UniProtKB P08172) |
|           | D00465: oxybutynin (DrugBank ID DB01062), Hsa:134: adenosine receptor A2a (UniProtKB P39274) |
|           | D00645: bretylium (DrugBank ID DB01158), Hsa:1129: muscarinic acetylcholine receptor M2 (UniProtKB P08172) |
|           | D00465: oxybutynin (DrugBank ID DB01062), Hsa:135: adenosine receptor A2a (UniProtKB P39274) |
| DrugBank  | D00765: rocuronium (DrugBank ID DB00728), Hsa:1129: muscarinic acetylcholine receptor M2 (UniProtKB P08172) |

shows that the results we verified are credible. The top-ranked drug–target interactions are muscarinic acetylcholine receptor M1 and darifenacin. Muscarinic acetylcholine receptor excitation can cause bladder smooth muscle contraction and saliva secretion, which has been proven effective.45 A significant fraction of the predictions (4 out of 10) is found in one or more of these data sets. It is worth mentioning that a large fraction of the interactions in these data sets are already included in the training data and hence are not counted as new interactions. Moreover, these data sets are incomplete, so if a predicted interaction is not present at one of the used data sets, then it does not necessarily mean it does not exist.

5. DISCUSSION

We present a new kernel method that leads to good predictive performance on the task of predicting interactions between drugs and target proteins.

Moreover, we also demonstrate that the THN KRLS method performs better than the other existing methods when known drug–target interactions are missing in the training data. This shows practical assessments of the predictive power of THN KRLS for real scenarios of drug–target interaction predictions.

However, we still have some problems to be solved in the future. First, how to apply the classical algorithms in graph theory to the heterogeneous network is a problem, such as the maximum matching problem on the bipartite graph, the classical binary classification algorithm, and so on. When a new drug or target enters the heterogeneous network structure, it must rely on some former information in the layers to complete the prediction, for example, the similarity of the drug–drug chemical structure. It is hard to detect the presence of drugs or targets with these graph algorithms. Second, the features and parameters in our experiments are only obtained by an empirical or ordinary weighted method, which makes it difficult to decide whether over-fitting or under-fitting is done in the experiment. Additionally, no definite standard can be used to decide the similarity in the biological research. Finally, since the Gaussian kernel function only plays a role in detecting common neighbors under this model, the Gaussian kernel that builds the similarity matrix space can be replaced by other machine learning methods.

6. CONCLUSIONS

We introduce a tripartite heterogeneous network model and the regularized least square method of the Kronecker product to fit multiple kernels and receive better prediction performance of drug repositioning. The method proposed in this paper achieves significantly more accurate results than the other network methods under different prediction task settings and on different data sets.

APPENDIX

In this paper, our main prediction results include the following drugs and targets as Table A1 shows. The drugs and targets come from the Kegg, DrugBank, and UniProt databases. We have given the final results in Tables 4 and 5. Therefore, the following content is not given as a supporting information file.

AUTHOR INFORMATION

Corresponding Author
Ying Zheng — School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha 410000, China; orcid.org/0000-0003-2023-259X; Phone: +86-186-0731-0657; Email: zhengying@csust.edu.cn

Author
Zheng Wu — School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha 410000, China; orcid.org/0000-0002-8096-0065
Table A1. One-to-One ID Information of the Drugs and Targets Involved, Including Kegg ID, DrugBank ID, or UniProt ID and the Drug Name or Target Name

| Kegg ID | DrugBank ID or UniProt ID | Drug name or target name |
|---------|---------------------------|--------------------------|
| D00394  | DB00726                   | trimipramine             |
| D00225  | DB00404                   | alprazolam               |
| D001071 | DB01355                   | hexobarbital             |
| D00380  | DB01124                   | tolbutamide              |
| D00574  | DB00357                   | aminoglutethimide        |
| D00139  | DB00533                   | methoxsalen              |
| D01699  | DB00496                   | darifenacin              |
| D00465  | DB01062                   | oxynutrin                |
| D00645  | DB01158                   | bretylium                |
| D00765  | DB00728                   | rocuronium               |
| Hsa:28  | P16442                    | histo-blood group ABO system transferase |
| Hsa:1557| P33261                    | cytochrome P450 2C19     |
| Hsa:57105| QN875                     | cysteinyl leukotriene receptor 2 |
| Hsa:10800| QY271                     | cysteinyl leukotriene receptor 1 |
| Hsa:1128| P11129                    | muscarinic acetylcholine receptor M1 |
| Hsa:1129| P08172                    | muscarinic acetylcholine receptor M2 |
| Hsa:134 | P30542                    | adenosine receptor A1    |
| Hsa:135 | P29274                    | adenosine receptor A2a    |

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c05377

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS
We thank Professor Jin Wang for his comments on the overall structure and feasibility of this article. We thank the DrugBank database team for their support in terms of experimental data.

REFERENCES
(1) Emig, D.; Ivlev, A.; Pustovolova, O.; Lancashire, L.; Bureeva, S.; Nikolsky, Y.; Bessarabova, M. Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. PLoS One 2013, 8, e60618.
(2) Mahmoud, D. B.; Afifi, S. A.; El Sayed, N. S. Crown Ether Nanovesicles (Crownosomes) Repositioned Phenytoin for Healing of Corneal Ulcers. Mol. Pharmacol. 2020, 17, 3952–3965.
(3) Liao, Z.; Peng, J.; Chen, Y.; Zhang, J.; Wang, J. A Fast Q Learning Based Data Storage Optimization for Low Latency in Data Center Networks. IEEE Access 2020, 8, 90630–90639.
(4) Trouilé, K.; López-Fernández, H.; García-Martín, S.; Reboiro-Jato, M.; Carretero-Puche, C.; Martorell-Marugán, J.; Martín-Serrano, G.; Carmona-Sáez, P.; González-Peña, D.; Al-Shahrour, F.; Gómez-López, G. DREIMT: a drug repositioning database and prioritization tool for immunomodulation. OA Bioinf. 2020, 1, 1–8.
(5) Li, L.; He, X.; Borgwardt, K. Multi-target drug repositioning by bipartite block-wise sparse multi-task learning. BMC Syst. Biol. 2018, 12, S5.
(6) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A comprehensive map of molecular drug targets. Nat. Rev. Drug Discovery 2017, 16, 19.
(7) Chen, Y.; Tao, J.; Zhang, Q.; Yang, K.; Chen, X.; Xiong, J.; Xia, R.; Xie, J. Salience Detection via the Improved Hierarchical Principal Component Analysis Method. Wireless Commun. Mobile Comput. 2020, 2020, 1–12.
(8) Chen, H.-R.; Sherr, D. H.; Hu, Z.; DeLisi, C. A network based approach to drug repositioning identifies plausible candidates for breast cancer and prostate cancer. BMC Med. Genomics 2016, 9, 51.
(9) Zheng, S.; Li, Y.; Chen, S.; Xu, J.; Yang, Y. Predicting drug–protein interaction using quasi-visual question answering system. Nat. Mach. Intell. 2020, 2, 134–140.
(10) Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. Nat. Commun. 2017, 8, 573.
(11) Wang, J.; Ji, C.; Gao, Y.; Sangaiah, A. K.; Kim, G.-J. A PSO based energy efficient coverage control algorithm for wireless sensor networks. Comput. Mater. Continua 2018, 56, 433–446.
(12) Olayan, R. S.; Ashoor, H.; Bajic, V. B. DDR: Efficient computational method to predict drug–Target interactions using graph mining and machine learning approaches. Bioinformatics 2018, 34, 1164.
(13) Cowen, L.; Ideker, T.; Raphael, B. J.; Sharan, R. Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. 2017, 18, 551–562.
(14) Li, W.; Xu, H.; Li, H.; Yang, Y.; Sharma, P. K.; Wang, J.; Singh, S. Complexity and Algorithms for Superposed Data Uploading Problem in Networks with Smart Devices. IEEE Internet Things J. 2019, 7, 8882–8891.
(15) Hao, M.; Bryant, S. H.; Wang, Y. Open-source chemogenomic data-driven algorithms for predicting drug-target interactions. Briefings Bioinf. 2019, 20, 1465–1474.
(16) Lee, A.; Lee, K.; Kim, D. Using reverse docking for target identification and its applications for drug discovery. Expert Opin. Drug Discovery 2016, 11, 707–715.
(17) Shahreza, M. L.; Ghadiri, N.; Mousavi, S. R.; Varshosaz, J.; Green, J. R. A review of network-based approaches to drug repositioning. Briefings Bioinf. 2018, 19, 878–892.
(18) Park, K.; Cho, A. E. Using Reverse Docking to Identify Potential Targets for Ginsenosides. J. Ginseng Res. 2017, 41, S34–S39.
(19) Madhukar, N. S.; Khade, P. K.; Huang, L.; Gayvert, K.; Galletti, G.; Stogniew, M.; Allen, J. E.; Giannakakou, P.; Elemento, O. A Bayesian machine learning approach for drug target identification using diverse data types. Nat. Commun. 2019, 10, 5221.
(20) Rifaioğlu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Briefings Bioinf. 2019, 20, 1878–1912.
(21) Wang, J.; Yang, Y.; Wang, T.; Sherratt, R.; Zhang, J. Big Data Service Architecture: A Survey. J. Internet Technol. 2020, 21, 393–405.
(22) Ezatt, A.; Zhao, P.; Wu, M.; Li, X.-L.; Kwoh, C.-K. Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization. IEEE/ACM Trans. Comput. Biol. Bioinf. 2016, 14, 646–656.
(23) Hao, M.; Bryant, S. H.; Wang, Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. Sci. Rep. 2017, 7, 40376.
(24) Wang, J.; Wu, W.; Liao, Z.; Sherratt, R. S.; Kim, G.-J.; Alijarra, O.; Alzubi, A.; Tolba, A. A Probability Preferred Priori Offloading Mechanism in Mobile Edge Computing. IEEE Access 2020, 8, 39758–39767.
(25) Luo, H.; Li, M.; Yang, M.; Wu, F.-X.; Li, Y.; Wang, J. Biomedical data and computational models for drug repositioning: a comprehensive review. Briefings Bioinf. 2020, 1, 26.
(26) Lu, J.; Chen, L.; Yin, J.; Huang, T.; Bi, Y.; Kong, X.; Zheng, M.; Cai, Y.-D. Identification of new candidate drugs for lung cancer using chemical–chemical interactions, chemical–protein interactions and a K-means clustering algorithm. J. Biomed. Struct. Dyn. 2016, 34, 906–917.
(27) Maleki, E. F.; Gohir, N.; Shahreza, M. L.; Maleki, Z. DHLBP1&2: Giraph based distributed label propagation algorithms on heterogeneous drug-related networks. Expert Syst. Appl. 2020, 159, 113640.
(28) Zeng, X.; Zhu, S.; Liu, X.; Zhou, Y.; Nussinov, R.; Cheng, F. DeepDR: A network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019, 35, 5191.

(29) Martínez, V.; Navarro, C.; Cano, C.; Fajardo, W.; Blanco, A. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.* 2015, 63, 41–49.

(30) Kunimoto, R.; Bajorath, J. Design of a tripartite network for the prediction of drug targets. *J. Comput.-Aided Mol. Des.* 2018, 32, 321–330.

(31) Lu, L.; Yu, H. DR2DI: a powerful computational tool for predicting novel drug-disease associations. *J. Comput.-Aided Mol. Des.* 2018, 32, 633–642.

(32) van Laarhoven, T.; Nabuurs, S. B.; Marchiori, E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 2011, 27, 3036–3043.

(33) Yu, W.; Cheng, X.; Li, Z.; Jiang, Z. Predicting drug-target interactions based on an improved semi-supervised learning approach. *Drug Dev. Res.* 2011, 72, 219–224.

(34) Yamanishi, Y.; Arai, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008, 24, i232–i240.

(35) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2018, 46, D1074–D1082.

(36) Goh, K.-I.; Cusick, M. E.; Valle, D.; Childs, B.; Vidal, M.; Barabasi, A.-L. The human disease network. *Proc. Natl. Acad. Sci.* 2007, 104, 8685–8690.

(37) Chen, H.; Cheng, F.; Li, J. IDrug: Integration of drug repositioning and drug-target prediction via cross-network embedding. *PLoS Comput. Biol.* 2020, 16, e1008040.

(38) BJ, B. N. Computational Approaches Towards Drug Repositioning by Combining Heterogeneous Network Model and Random Walk Algorithm for Medications in Neurological Disorders. *Int. J. Pure Appl. Math.* 2020, 4707–4711.

(39) Chen, H.Y.; Li, J. Modeling Relational Drug-Target-Disease Interactions via Tensor Factorization with Multiple Web Sources. In *The World Wide Web Conference*; 2019, 218–227, DOI: 10.1145/3308558.3313476.

(40) Parisi, D.; Adasme, M. F.; Sveshnikova, A.; Bolz, S. N.; Moreau, Y.; Schroeder, M. Drug repositioning or target repositioning: A structural perspective of drug-target-indication relationship for available repurposed drugs. *Comput. Struct. Biotechnol. J.* 2020, 18, 1043–1055.

(41) Chen, H.; Li, J. Learning Data-Driven Drug-Target-Disease Interaction via Neural Tensor Network. *Proc. Twenty-Ninth Int. Joint Conf. on Artif. Intell.* 2020, 3452–3458.

(42) Bleakley, K.; Yamanishi, Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 2009, 25, 2397–2403.

(43) Kirchheiner, J.; Müller, G.; Meineke, I.; Wernecke, K.-D.; Roots, I.; Brockmöller, J. Effects of polymorphisms in CYP2D6, CYP2C9, and CYP2C19 on trimipramine pharmacokinetics. *J. Clin. Psychopharmacol.* 2003, 23, 459–466.

(44) Desta, Z.; Zhao, X.; Shin, J.-G.; Flockhart, D. A. Clinical Significance of the Cytochrome P450 2C19 Genetic Polymorphism. *Clin. Pharmacokinet.* 2002, 41, 913–958.

(45) Jha, S.; Parsons, M. Treatment of overactive bladder in the aging population: focus on darifenacin. *Clin. Interv. Aging* 2006, 1, 309–316.