Journal of Cheminformatics

RESEARCH Open Access

Deep learning of multimodal networks with topological regularization for drug repositioning

Yuto Ohnuki[1], Manato Akiyama[1] and Yasubumi Sakakibara[1*]

Abstract

Motivation Computational techniques for drug-disease prediction are essential in enhancing drug discovery and repositioning. While many methods utilize multimodal networks from various biological databases, few integrate comprehensive multi-omics data, including transcriptomes, proteomes, and metabolomes. We introduce STRGNN, a novel graph deep learning approach that predicts drug-disease relationships using extensive multimodal networks comprising proteins, RNAs, metabolites, and compounds. We have constructed a detailed dataset incorporating multi-omics data and developed a learning algorithm with topological regularization. This algorithm selectively leverages informative modalities while filtering out redundancies.

Results STRGNN demonstrates superior accuracy compared to existing methods and has identified several novel drug effects, corroborating existing literature. STRGNN emerges as a powerful tool for drug prediction and discovery. The source code for STRGNN, along with the dataset for performance evaluation, is available at https://github.com/yuto-ohnuki/STRGNN.git.

Keywords Multimodal network, Multi-omics data, Drug-disease association, Topological regularization

Introduction

De-novo drug discovery, involving multiple rigorous stages to verify drug efficacy and safety, generally takes about 15 years and costs over one billion dollars. The high cost and lengthy development process pose significant challenges [1]. However, with the availability of diverse, high-throughput data in biological databases, in-silico drug discovery methods, which exhaustively search for drug candidate compounds using high-performance computing, have been actively developed. Drug repositioning, which seeks new applications for existing drugs, is gaining increasing attention.

In-silico methods for drug repositioning can be broadly classified into network propagation-based, recommendation-based, classical machine learning-based, and deep learning-based methods [2]. Network propagation methods represent biological entity interactions as networks, exploring new interactions through approaches like random walks. Node2Vec [3], a popular biased random walk method, is frequently applied in drug repositioning prediction [4]. Luo et al. [5] proposed RWHNDR, a random walk-based algorithm for heterogeneous networks in computational drug repositioning. Cheng et al. [6] quantified the network proximity between disease genes and drug targets in the human protein–protein interactome.

Recommendation-based methods interpret the interaction prediction problem as a recommendation task, primarily utilizing matrix decomposition and completion techniques. Luo et al. [7] proposed a drug-disease prediction algorithm DRRS using matrix interpolation

*Correspondence:
Yasubumi Sakakibara
yasu@bio.keio.ac.jp
[1] Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

for heterogeneous networks for drug repositioning. Xuan et al. [8] introduced DisDrugPred, a non-negative matrix factorization (NMF) method for predicting drug-disease associations using multiple similarity measures. Zhang et al. [9] developed the DRIMC method, a Bayesian-derived matrix imputation approach. These methods are advantageous as they do not require negative sampling, allowing for the flexible integration of extensive prior information. However, their application to large-scale data is limited due to the complexity of matrix operations [10].

Machine learning-based methods, assuming that similar drugs target similar diseases, treat interaction prediction as a binary classification problem [11]. Notable contributions include Nagamine and Sakakibara's statistical prediction method [12] using support vector machines for protein-chemical interactions, Gottlieb et al.'s PREDICT [13], a similarity-based method employing logistic regression, and Zhou et al.'s NEDD [14], a metapath-based learning approach using a random forest classifier. To expand on the machine learning-based methods for drug repurposing, it's crucial to consider how each method leverages specific data types and analytical frameworks to improve predictive accuracy. While statistical prediction and similarity-based methods primarily focus on existing relationships and known properties, metapath-based approaches integrate multiple data sources, enabling a more holistic view of potential drug-disease interactions. These methods each have their strengths; for example, support vector machines are robust to overfitting, logistic regression offers interpretability, and random forest classifiers excel in handling heterogeneous data. However, deep learning (DL) approaches often surpass these methods in capturing complex nonlinear relationships through layered data representations.

Deep learning methods have gained prominence in bioinformatics for their ability to acquire data representations at multiple abstraction levels. For instance, Watanabe et al. [15] proposed a method integrating molecular and interactome data for protein-compound interaction prediction. Notably, graph neural network (GNN) methods [16] have been actively developed, allowing for node vector representations that reflect network topology. This includes the Graph Attention Network [17], which integrates an attention mechanism, and the Relational GCN [18], applicable to heterogeneous networks. Zeng et al.'s deepDR [19] and Wan et al.'s NeoDTI [20] are examples of network-based deep learning methods integrating various drug-related information. Yu et al. [10] proposed LAGCN, a GNN method with a layer attention mechanism for accurate drug-disease association prediction. Wan et al. [21] constructed a heterogeneous network containing protein–protein interactions and proposed a bipartite graph convolutional model called BiFusion. Cai et al. [22] introduced DRHGCN, which extracts features from domain information in heterogeneous networks for drug repositioning.

Systems biology research, viewing human diseases as perturbations in biological networks, emphasizes the importance of multi-omics data analysis for understanding complex biological systems [23] and investigating disease mechanisms [24–26]. Computational methods like Iwata et al.'s approach [27] using gene expression profile data from LINCS [28] and Wang et al.'s DeepDRK [29], which integrates multi-omics data for predicting cancer cell drug responses, demonstrate the value of extracting and integrating modality-specific information [30]. However, existing golden datasets for drug-disease association prediction, such as Fdataset [13], Cdataset [31], and Ldataset [10], focus solely on drugs and diseases. Moreover, few studies have constructed datasets encompassing comprehensive omics data, including RNA and metabolites. Biological networks are inherently noisy and complex [2], necessitating a regularization mechanism to mitigate noise effects and obtain robust feature representations that efficiently integrate heterogeneous data sources [32]. Peng et al.'s RNMFLP [33] and methods like BNNR [34], which incorporates bounded nuclear norm regularization, represent attempts to address these challenges. However, they do not fully address the nuances of heterogeneous networks, and methods like EnMUGR [32], which perform denoising as preprocessing, may not effectively select modalities during training.

To address these challenges, our study introduces STRGNN (Sequentially Topological Regularization Graph Neural Network), a deep learning model that effectively predicts drug-disease associations using large-scale multimodal networks rich in omics data. It is important to note the distinction between multimodal and heterogeneous networks. While both integrate various data types, multimodal networks focus on different modalities or sources, and heterogeneous networks on different node and edge types. Our approach specifically utilizes multimodal networks, acknowledging the broader context of "heterogeneous networks" in the literature. Initially, we methodically gather and integrate interaction information from biological databases, forming a multimodal network comprising 6 biomolecular entity types (drugs, diseases, proteins, mRNAs, miRNAs, metabolites) and 20 interaction or association types, including multifunctional interactions. Subsequently, we introduce a novel mechanism in GNN learning, called topological regularization which effectively selects informative modalities and eliminates redundant data from the multimodal network. This approach has led to

Table 1 Comparison of drug-disease association prediction methods

Method	Category	Used modality	Main feature
RWHNDR [5]	Network propagation	Drugs, diseases, target proteins	Random walk
DisDrugPred [8]	Recommendation-based	Drugs, diseases	Non-negative matrix factorization
PREDICT [13]	Machine learning	Drugs, diseases	Logistic regression
NEDD [14]	Machine learning	Drugs, diseases	Random forest, metapath-based learning
deepDR [19]	Deep learning	Drugs, diseases, target proteins, drug side effects	Variational autoencoder
NeoDTI [20]	Deep learning	Drugs, diseases, target proteins, drug side effects	Neural Network, Graph Embedding
LAGCN [10]	Deep learning	Drugs, diseases	Graph attention mechanism
DRHGCN [22]	Deep learning	Drugs, diseases	Heterogeneous GCN
STRGNN	Deep learning	Drugs, diseases, target proteins, mRNAs, miRNAs, metabolites	Topological regularization GNN

superior prediction accuracy compared to other leading interaction prediction methods in drug repositioning, and has enabled the discovery of multiple new drug effects that align with empirical findings in the literature.

To provide a comprehensive review and discussion of machine learning-based methods for drug repurposing, we have compiled a comparative table of key approaches. Table 1 summarizes the characteristics of representative drug-disease association prediction methods, including their categories, used modalities, main features. This comparison highlights the unique aspects of our proposed STRGNN method, which integrates a diverse range of modalities and employs topological regularization GNN for effective processing of large-scale multimodal data.

Materials and methods

Dataset

In this study, in order to construct multimodal networks comprehensively including multi-omics information, a total of 6 types of biomolecular entities and 20 types of interactions or associations were collected from 9 databases (DrugBank [35], STRING [36], HuRI [37], CTD [38], HMDB [39], CREED [40], Harmonizome [41], RNAInter [42], HMDD [43]). As a result, compared to the golden data set Cdatasets, we succeeded in including approximately 4.8 times more drugs, 4.6 times more diseases, and 4.9 times more drug-disease associations. The number of entities of each biomolecular type and the number of interactions of each type are listed in Supplemental Table S1.

"Disease" entities were collected from KEGG [44]. "Drug-disease" associations were obtained from CTD, specifically focusing on pairs that have direct evidence of being used as a therapeutic drug in clinical practice. "Disease-protein" associations were also obtained from CTD. In CTD, manually curated association is available, and proteins with direct evidence such as markers or therapeutic targets were collected. "Disease-mRNA" associations were obtained from CREEDS and Harmonizome. In CREEDS, gene expression signatures based on Gene Expression Omnibus (GEO) [45] are available. Harmonizome standardized data formats and unified identifiers in gene expression profile information obtained from numerous databases. From these databases, information on fold changes in gene expression during disease was collected, and a network was constructed based on that information. "Disease-miRNA" associations were obtained from HMDD. Dysregulation of miRNAs is associated with numerous diseases such as cancer and neurodegenerative diseases [46]. "Disease-metabolite" interactions were obtained from HMDB. HMDB is the most comprehensive metabolomics database, annotating all known human metabolites. A weighted "disease-disease" network was constructed by calculating a similarity score based on the ICD-11 disease classification. Diseases are classified hierarchically in the form of a directed acyclic graph (DAG) structure using descriptors such as MeSH and ICD-11, and a method for evaluating semantic similarity between diseases based on this classification has been proposed [47]. Since diseases with similar molecular characteristics are expected to have similar drug targets, it is important to consider disease similarity. The semantic value $DV(A)$ of disease A is defined by the following formula.

$$DV(A) = \sum_{t \in T_A} D_A(t)$$

where T_A represents the set of ancestor nodes of disease A, and the contribution $D_A(t)$ of disease t to disease A is defined by

$$D_A(t) = \begin{cases} 1, & t = A \\ max\{\Delta * D_A(t') | t' \in children\, of\, t\}, & t \neq A \end{cases}$$

where Δ is the semantic contribution damping coefficient and is generally chosen to be $\Delta = 0.5$. The semantic similarity score $S(A, B)$ between disease A and disease B is defined by the following formula

$$S(A,B) = \frac{\sum_{t \in T_A \cap T_B}(D_A(t) + D_B(t))}{DV(A) + DV(B)}.$$

"Drug" entities were collected from DrugBank. DrugBank is a database that provides comprehensive information on FDA-approved drugs and experimental drugs in the approval process. "Drug-protein" interactions, "drug-mRNA" interactions, "drug-drug" interactions were retrieved from DrugBank. "Drug-metabolite" interactions were obtained from HMDB. Drugs are structurally modified into metabolites by various metabolic enzymes [48]. Therefore, metabolomics dealing with metabolites is regarded as an important tool for drug discovery [49]. In this study, in order to avoid duplication with compounds used as drug nodes, the metabolites were incorporated into the network independently. It is important to note that while our focus is on predicting drug-disease interactions, these interactions are not included in the input networks to avoid any bias. Our methodology aims to deduce these interactions based on the intricate patterns observed among other entities and their interactions.

"Proteins" play a vital role in a plethora of biological processes. Therefore, exploring the functional mechanisms of proteins lead to a comprehensive understanding of disease processes and drug mechanisms [50]. In particular, protein–protein interactions are not only known to play important roles in biological systems and disease states, but have also been shown to be important modalities in predicting drug repositioning [21]. "Protein–protein" interactions were obtained from STRING and Human Reference Protein Interactome (HuRI). STRING are accompanied with confidence scores based on experimental data and pathway databases. A confidence score of 900 or higher is considered the cutoff for the highest level of confidence [51], and protein–protein interaction networks were constructed accordingly. In addition, HuRI obtained direct interaction information experimentally verified by the Yeast Two-Hybrid System, and has achieved the most systematic and comprehensive screening of human protein interactions.

Transcriptome data are useful for systematic understanding of the functional mechanisms of drugs and diseases because they can capture the dynamic characteristics of cells [52]. Various types of RNA are involved in gene regulation, so we selected two of them, "mRNA" (coding-RNA) and "miRNA" (non-coding RNA), as entities to be included in the multimodal network. miRNA has the function of regulating gene expression by binding complementary to mRNA. Thus, interactions between RNAs are known to play important roles in various cellular processes [53]. While we included miRNA as a representative non-coding RNA entity due to its well-established role in drug interactions and its extensive documentation in databases, we acknowledge the importance of other non-coding RNAs such as lncRNA, piRNA, siRNA, and shRNA. The decision to focus on miRNA was based on data availability and its established relevance in drug-target interactions. RNA-related interactions were obtained from RNAInter. RNAInter integrates literature and multiple data sources to provide empirically-based reliability scores. "Drug-miRNA", "protein-mRNA" and "protein-miRNA" interactions were introduced from RNAInter as a weighted network using this reliability score. Interactions between RNAs were also obtained as weighted networks based on reliability scores obtained from RNAInter.

Model workflow

This section describes in details our method, STRGNN, for predicting drug-disease associations. The STRGNN consists of three main steps: (i) Attribute Encoder, which encodes structural information of biomolecules; (ii) Network Encoder, which encodes network information; (iii) Decoder, which predicts drug-disease associations. The framework of STRGNN is illustrated in Fig. 1.

The feature matrix obtained by the Attribute Encoder is used as the initial node feature in the Network Encoder. This low-dimensional distributed representation serves as the input to the Graph Neural Network layers, which then learn the node embeddings based on the multimodal network structure.

It is important to note that STRGNN does not require Attribute Decoders or Network Decoders, as the purpose of our method is to learn informative node representations for drug-disease association prediction. The final node embeddings obtained from the Network Encoder are directly utilized by the Decoder to compute the association scores.

In this study, we represent multimodal networks as graphs. A graph is defined as $G = (V, E)$ (V, E are nodes and edges respectively). Each modal network is represented by a graph of its own type. The node v of biomolecule type o belongs to the set V_o and the edge of interaction type r belongs to the set E_r, where the set of biomolecular types $O = \{$*drug, disease, protein, mrna, mirna, metabolite*$\}$ and the set of interaction types $R = \{$*drug-drug, drug-disease, drug-protein, drug-mrna, drug-mirna, drug-metabolite, disease-disease, disease-protein, disease-mrna, disease-mirna, disease-metabolite, protein–protein, protein-mrna, protein-mirna, mrna-mrna, mrna-mirna, mirna-mirna*$\}$. The multimodal network combines all those networks composed of biomolecular types and interaction types. Then the drug-disease association prediction is defined as the task of predicting the link probability of a pair (v_i, u_j) of a drug node

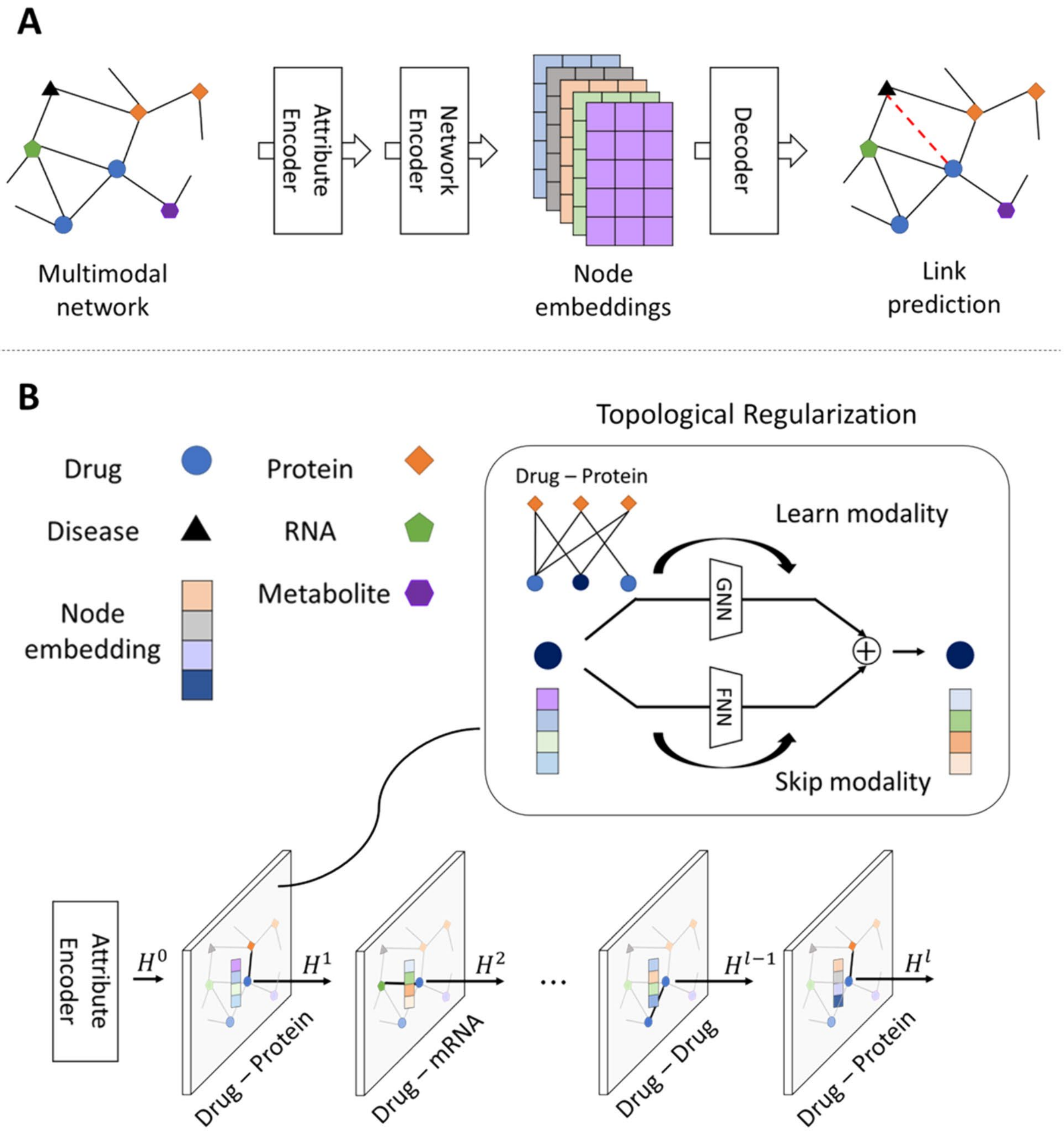

Fig. 1 Schematic diagram of STRGNN workflow. **A** STRGNN performs drug-disease link prediction using two encoders, Attribute Encoder and Network Encoder, and a Decoder. **B** Network Encoder sequentially applies Topological Regularization to each interacting network

$v_i \in V_{drug}$ and a disease node $u_j \in V_{disease}$ on the multimodal network.

Attribute Encoder encodes drugs (compounds) and proteins (amino acid sequences). The feature matrix H^0 obtained by the Attribute Encoder is used as the initial value of the node feature in the Network Encoder. Molecular structures of chemical compounds and amino acid sequences of proteins are important information representing their structural characteristics and functions, and are useful for interaction prediction related to proteins and drugs [54].

Drugs were converted from SMILES representations obtained from DrugBank to 2048-dimensional Extended Connectivity Fingerprint (ECFP) [55] using RDKit. Then, by applying a fully-connected feedforward neural network (FNN), we obtained a low-dimensional distributed representation of the drug node $H^0_{drug} \in \mathbb{R}^{N_{drug} \times d}$, where N_{drug} denotes the number of drug nodes, and d denotes the number of dimensions of the low-dimensional distributed representation.

Amino acid sequences of proteins were obtained from UniProt [56]. After converting the amino acid sequences into one-hot vectors, a one-dimensional convolutional neural network (CNN) was applied according to the method of ref. [15]. Then, by applying an FNN, we obtained a low-dimensional distributed representation $H^0_{protein} \in \mathbb{R}^{N_{protein} \times d}$ for the protein node, where $N_{protein}$ denotes the number of protein nodes, and d denotes the number of dimensions of the low-dimensional distributed representation.

These methods have been shown to be effective in many previous studies. ECFP-based drug representation is widely used in various tasks such as drug discovery and interaction prediction due to its high expressiveness and computational efficiency [57]. The application of CNN to protein amino acid sequences has demonstrated high performance in protein function prediction by effectively capturing local features of the sequences [58].

For diseases, RNA, and metabolites, initial values were sampled from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = \frac{1}{\sqrt{d}}$, and their low-dimensional distributed representations ($H^0_{disease}, H^0_{rna}, H^0_{metabolite}$ for disease, RNA and metabolite, respectively) were obtained.

Network Encoder acquires the distributed representation of each node (entity) based on the adjacency relations between nodes included in the multimodal network. Graph Neural Network (GNN) acquires node distribution representation based on adjacency relation for graph structure data according to the following update formula.

$$H^{l+1} = f(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^l W^l)$$

where A is the adjacency matrix of the input graph, H^l is the node feature (node distribution representation) matrix in the l-th layer, $D = diag(\sum_j A_{ij})$ is the degree matrix, W^l represents the weight parameters of GNN, and $f(\cdot)$ is the activation function.

In STRGNN that learns each modality network sequentially, it is inevitable to deal with the problem caused by learning with deep layers, that is, over-smoothing. In this study, we employed DropEdge [59], which is a regularization method that probabilistically removes edges in the input graph, and NodeNorm [60], which conducts nodewise scaling for the feature of each node (see Supplemental in detail).

The large-scale networks may contain redundant and unnecessary information and noise, which causes overfitting of learned models. The "topological regularization" proposed in this study introduces a path that skips a modality learning by adding an FNN in parallel with the GNN that learns the modality network. We call this path "skip connection" and defined with the following formula.

$$H^{l+1} = f\left(D^{-\frac{1}{2}}A^l D^{-\frac{1}{2}}H^l W^l_{gnn} + H^l W^l_{fnn}\right)$$

where H^l is the node feature matrix in the l-th layer, and W^l_{gnn}, W^l_{fnn} represent the weight parameters of GNN and FNN, respectively. By successively applying this model learning consisting of two paths to the network of each modality, the node distributed representation is learned in a sequential manner. In addition, in order to enhance the effect of skipping the learning of unnecessary modalities, the involvement of the redundant weight parameter W is removed by introducing L_1 regularization. Combined the skip connection with L_1 regularization, the mechanism of topological regularization was implemented.

The Inner Product Decoder [61] was used as a decoder to predict the association between drugs and diseases.

$$\widehat{A} = sigmoid(H_{drug}H^T_{disease})$$

where $\widehat{A} \in \mathbb{R}^{N_{drug} \times N_{disease}}$ is the prediction probability matrix, H_{drug} and $H_{disease}$ are the node feature matrices in the final layer for drugs and diseases, respectively, and the prediction score for the pair (v_i, u_j) of drug v_i and disease u_j is obtained as a $\widehat{a}_{i,j} \in \widehat{A}$.

The cross-entropy loss and the regularization term were adopted as the loss function. Regularization is important not only for suppressing overfitting, but also for suppressing bias between modalities in multimodal learning that integrates different modalities [62]. As stated above, in order to assist the effect of topological regularization, which skips the learning of unnecessary modalities, the contribution of the redundant weight parameter W is eliminated by introducing L_1 regularization.

$$loss = \sum_{i=0}^{N}\left\{-y_i\log(p_i) - \left(1 - y_i\right)\log(1 - p_i)\right\} + \lambda\|w\|_1$$

where $p_i \in \{0,1\}$ is the i-th prediction, y_i is the ground truth, w represents the learning weight, and λ is the coefficient of the regularization strength.

It is important to note that L1 regularization is applied to the weights of all layers in STRGNN. The motivation behind using L1 regularization is to induce sparsity in the weights, which allows for the selective inclusion or exclusion of interactions at each layer. By encouraging sparsity, L1 regularization helps the model to identify and focus on the most informative interactions while discarding the less relevant ones.

For performance evaluation, we used the following metrics: Area Under the Receiver Operating Characteristic curve (AUROC), Accuracy (ACC), and Area Under the Precision-Recall Curve (AUPRC). In particular, AUPRC is frequently used as an evaluation metric for imbalanced datasets. The model was evaluated using a nested cross-validation approach, where the drug-disease network was initially split into a test set and a training-validation set, following the approach described in the "Setting for drug repositioning prediction" section.

The training-validation set was then used for hyperparameter tuning using fivefold cross-validation. In each fold, the data was further split into training and validation sets, and the model was trained on the training set while the hyperparameters were tuned based on the performance on the validation set. The best-performing hyperparameters from this inner loop were then used to train the final model on the entire training-validation set.

Finally, the model's performance was assessed on the held-out test set, which was not used during the hyperparameter tuning process. This approach ensures an unbiased evaluation of the model's performance on unseen data.

There are several hyperparameters in STRGNN, including kernel size, channel size, dropout probability, drop edge probability, embedding dimensionality, regularization strength, and learning rate. These hyperparameters were extensively searched using the validation sets in the inner loop of the nested cross-validation procedure. The best-performing configuration is provided in the supplementary material.

Setting for drug repositioning prediction

In this study, we evaluated a model for drug repositioning. Specifically, we constructed a data set to test the potential reuse of drugs that have been confirmed to be effective for specific diseases to other diseases. Figure 2 illustrates the method for constructing the test data. First, we selected drugs whose degree deg(v) of the drug node $v \in V_{drug}$ in the drug-disease network satisfies deg(v) ≥ 2, that is, the number of edges connected to the node v is more than or equal to 2. This enabled us to select drugs with multiple target diseases for test dataset. Second, we compared these target diseases with disease classifications at the shallowest hierarchies of the ICD-11. During this process, we ensured that there was no overlap between the ICD-11 classification of diseases in the training dataset and the ICD-11 classification of diseases in the test dataset. As a result, drug-disease pairs suitable for the purpose of drug repositioning were extracted as test data, and prediction performance was evaluated based on these datasets.

Negative examples cannot be obtained with the database CDT that collects positive examples for drug-disease associations. Therefore, negative sampling using unknown or unlabeled drug-disease pairs as negative examples is necessary. Specifically, for a drug-disease edge for (v_i, u_j) that is a positive example, we select random drug $v_{k1} \in V_{drug}$ and disease $u_{k2} \in V_{disease}$ exclusively for positive examples. Then, the pairs (v_i, u_{k2}) and (v_{k1}, u_j) obtained by replacing the nodes are taken as negative examples.

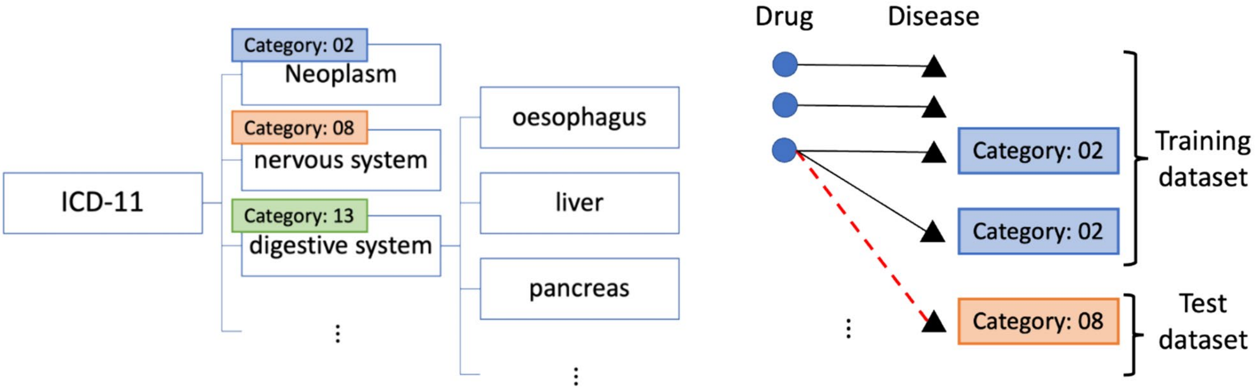

Fig. 2 Test data construction considering drug repositioning. (Left) Hierarchical classification based on disease characteristics and etiology in ICD-11. (Right) A pair of drug nodes with multiple target diseases and disease nodes with different ICD-11 is used as test data. For instance, the figure illustrates a drug targeting two diseases in category:02 and one in category:08. In this scenario, the pair consisting of the drug and the disease from category:08 is selected as test data

Table 2 STRGNN performance on different modality datasets

Modality	AUPRC	AUROC	ACC
Small	0.7772	0.8679	0.8180
Middle	0.7890	0.8830	0.8258
Large	0.8044	0.8792	0.8250

Table 3 Effectiveness of topological regularization

	AUPRC	AUROC	ACC
STRGNN with topological regularization	0.8044	0.8792	0.8250
STRGNN without skip connection	0.7874	0.8785	0.8037
STRGNN without L_1 regularization	0.7941	0.8887	0.8217

Results

Performance evaluation of STRGNN on different modality datasets

In order to verify that STRGNN can effectively learn the abundant modality information of multimodal networks, we compared the prediction accuracy when given different modalities. In the small-modality dataset, only drugs and diseases were handled as entities, and their association networks were provided as input. Specifically, the dataset involved drug-drug and disease-disease networks. For the middle-modality dataset, proteins were added as entities alongside drugs and diseases, with multiple interaction networks between them as input. This dataset incorporated drug-drug, drug-protein, protein–protein, disease-protein, and disease-disease networks. In the large-modality dataset, all modalities including RNA and metabolites were considered, utilizing their multimodal networks for model training. Table 2 shows the prediction performance for each modality dataset. Prediction performance improved with the introduction of more networks from different modalities. This result implies that it is important to utilize interaction networks based on multiple omics data in predicting drug-disease associations.

The performance improvement with the large modality dataset was not as significant as expected. This suggests that the additional modalities may not contribute substantially to prediction accuracy, possibly due to redundancy or irrelevance of the added information.

Effect of topological regularization

In order to verify the effect of the topological regularization proposed in this study, we compared the models with and without the FNN path that skips the learning of each modality, as well as with and without L_1 regularization. The result is shown in Table 3. The result shows that introducing both skip connection and L_1 regularization, that is topological regularization, achieves higher prediction performance. The novelty of STRGNN lies in the topological regularization, which integrates Graph Neural Networks (GNNs) and Feedforward Neural Networks (FNNs) through skip connections and combines them with L_1 regularization. This unique combination allows STRGNN to effectively exclude redundant modalities while learning from the most informative interactions in the multimodal network, as demonstrated by the improved performance when skip connections and L_1 regularization are used together. The low prediction performance without topological regularization implies the importance of selectively learning necessary modalities and discarding redundant ones from multimodal networks. In other words, a mechanism to eliminate potential noise is essential in interaction prediction based on biological data. In addition, by introducing L_1 regularization, the effect of modality selection is enhanced by competing for weight parameters that can be learned by GNN and skip connection in the STRGNN method.

Comparison with other state-of-the-art methods

Performance of STRGNN drug-disease association prediction was compared with existing state-of-the-art methods; NMF [63] as recommendation-based method, Node2Vec [3] as network propagation method, deepDR [19], Relational-GCN [18], NeoDTI [20], LAGCN [10], and DRHGCN [22] as most recent and representative deep learning methods.

To ensure a fair comparison, all methods were retrained using the Large-Dataset. For existing methods that do not support multimodal data, we utilized the maximum number of modalities available within each method. The retrained models were then tested using the same test set as the one used in our study, which was held out during the training process. This approach guarantees an unbiased evaluation of the models' performance on unseen data and ensures a direct comparison with STRGNN. The hyperparameters for the existing methods were kept the same as those proposed in their respective original papers to allow for a direct comparison without introducing additional tuning bias.

The result, shown in Table 4, indicates that STRGNN exhibits strong performance and the best on all indexes compared to other state-of-the-art methods. The bold values in the table highlight these best-performing results, emphasizing STRGNN's superiority. LAGCN and DRHGCN, which are state-of-the-art drug repositioning prediction methods, only utilize the network structure composed of drugs and diseases, and cannot be applied to multimodal networks with abundant interaction information.

Table 4 Performance comparison with state-of-the-art methods

Method	AUPRC	AUROC	ACC
STRGNN	**0.8044**	**0.8792**	**0.8250**
NMF	0.6802	0.8114	0.6676
Node2Vec	0.5553	0.6975	0.6956
deepDR	0.6800	0.8022	0.7067
Relational-GCN	0.7299	0.8269	0.7900
NeoDTI	0.7778	0.8788	0.8175
LAGCN	0.7502	0.8146	0.7202
DRHGCN	0.5485	0.7437	0.6761

Considering the importance of introducing multiple modalities, this drawback can be a serious bottleneck in drug-disease association prediction. On the other hand, the remarkably high performance of STRGNN and NeoDTI, which can incorporate multimodal interaction information, suggests the necessity of integrating multimodal networks.

In this experiment, STRGNN outperformed NeoDTI, potentially due to NeoDTI's inability to effectively extract features while considering the influence of noise and the importance of modalities contained in the network. Therefore, this result demonstrates the effectiveness of topological regularization to utilize diverse multimodal networks.

Performance evaluations were conducted with the modalities used by each method by default (shown in Supplementary Table S2).

Discussions

Effect of topological regularization

In order to verify the effect of topological regularization, which selectively learns important modalities, the heat map visualization of the strength of selected network modality is displayed in Fig. 3. The color of the heatmap represents the strength of the weights obtained by STRGNN for each interaction network, calculated separately for the Graph Neural Network (GNN) and the Feed-forward Neural Network (FNN) using the following formula:

$$\text{Weight Strength } = \sqrt{\sum_{i=1}^{N_{weight}} {w_i}^2}$$

where N_{weight} is the number of weight parameters and w_i is the value of the i-th weight parameter.

This result indicates that two learners, GNN for network learning and FNN for skip connection are competing to acquire weight parameters in learning each interaction network. For example, in the learning of disease-disease modality network and drug-protein modality network, GNN obtained larger weights, which is consistent with previous results that drug-target information and similarities in disease classification are important factors in drug discovery [64]. On the other hand, it is worth noting that the FNN for skip connection acquired larger weights in drug-drug networks, although drug-drug networks are frequently employed in other graph-based deep learning methods. Considering that the drug-drug network contains a larger amount of interaction information compared to other networks, as shown in Supplemental Table S1, it might contain much redundant interaction information that becomes noise in drug-disease association prediction. RNA-related networks, especially miRNA-miRNA interaction, also gained large weights. This result indicates the necessity of introducing RNA-related interactions that have not been addressed in previous studies in drug-disease association prediction.

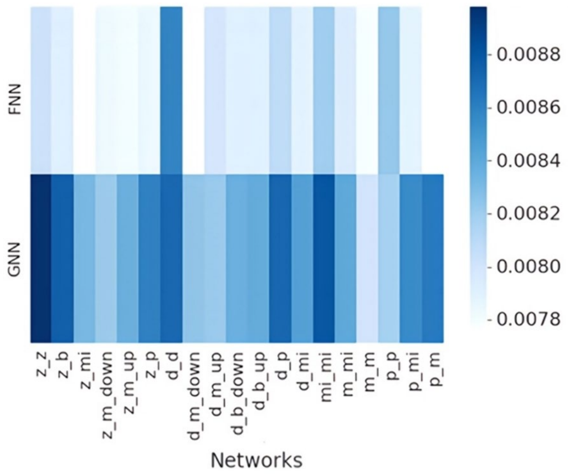

Fig. 3 Heatmap to display strength of weights obtained by STRGNN. The vertical axis repre-sents the GNN path and the FNN path. The horizontal axis represents each interaction network. Each modality is indicated by the following symbols; d: drug, z: disease, p: protein, m: mRNA, mi: miRNA, b: metabolite

Case studies for discovering new drug efficacy

A novel drug-disease association prediction was performed by applying the fully trained STRGNN to predict all drug-disease pairs for which no association was confirmed. Table 5 shows novel drug-disease associations predicted by STRGNN and the literature supporting the validity of the prediction. Among the predicted pairs, we focused on the drug Quercetin (DrugBank Accession Number: DB04216) and its target disease, Amyotrophic lateral sclerosis (ALS). Quercetin is a natural flavonoid and is known to have anti-inflammatory and anti-oxidant effects [65]. ALS pathogenesis involves misfolding

Table 5 Discovery of novel drug-disease association

Drug	Predicted disease	Refs.
Anthralin	Inflammatory bowel disease	[68]
Quercetin	Amyotrophic lateral sclerosis	[66, 67]
Amiodarone	Hypertension	[70]
Enalapril	Angina pectoris	[71]
Isotretinoin	Hyperbilirubinemia	[72]

and monomerization of the homodimeric protein SOD1. Molecular biological experiments have confirmed that Quercetin interacts with SOD1 and inhibits self-assembly [66, 67]. This promising prediction of interactions between Quercetin and ALS, which are expected to be used as therapeutic drugs, is a valid prediction result from the viewpoint of mechanism of action.

It is also worth noting that Anthralin (DrugBank Accession Number: DB11157) and its target disease, Inflammatory bowel disease had high predictive scores. Anthralin is classified as a drug for a skin disease by ICD-11 and is used as a treatment for psoriasis. Inflammatory bowel disease, on the other hand, is classified as disease of the digestive system in the ICD-11. This new efficacy prediction across different ICD-11 classifications is a useful case for drug repositioning. Anthralin is experimentally verified to inhibit cytokines such as TNF-α and IFN-γ [68]. Since TNF-α is related to the onset of Inflammatory bowel disease [69], it is expected that this new drug efficacy prediction considering the biological network could be correct.

Drug-disease association prediction for new drug

Existing methods for drug-disease association prediction based on graph deep learning have the problem that prediction cannot be applied to biological entities that are not included in multimodal networks. This implies that it is not possible to predict the association for novel compounds which are not registered in the database. On the other hand, in this study, we constructed an algorithm that can easily extend the association prediction of STRGNN to compounds outside the network. That is, the algorithm is to apply a loss function based on the similarity between the compound feature vector obtained from the Attribute Encoder and the Network Encoder. More specifically, by applying the cosine embedding loss, we projected a new compound onto the feature vector space based on the multimodal network using only the ECFP value of the new compound. In a preliminary experiment, we selected a limited number of drug nodes (614 nodes) as novel compounds and removed all edges connected to these drug nodes in the training dataset, and performed the association prediction (semi-inductive prediction) only between these drugs and diseases. It turned out that STRGNN exhibited high prediction performance (AUPRC: 0.833, AUROC:0.895, ACC:0.865) for novel drugs that did not exist in the multimodal network.

Conclusion

In this study, we developed STRGNN, a new framework for drug-disease association prediction from multimodal networks. STRGNN achieved higher prediction performance than other state-of-the-art methods by introducing Topological Regularization to selectively learn modalities from multimodal networks. In addition, STRGNN succeeded in predicting new efficacy of drugs with validity from the viewpoint of literature and mechanism of action. A future work is to perform association prediction considering more biological data sources. For example, by considering the RNA base sequence and the compound structure of metabolites, it is expected that a feature space reflecting more abundant information can be obtained.

Abbreviations

NMF	Non-negative matrix factorization
GNN	Graph neural network
STRGNN	Sequentially Topological Regularization Graph Neural Network
DAG	Directed acyclic graph
ECFP	Extended Connectivity Fingerprint
FNN	Feedforward neural network
CNN	Convolutional neural network
AUROC	Area under the Receiver Operating Characteristic Curve
AUPRC	Area under the Precision-Recall Curve
ALS	Amyotrophic lateral sclerosis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00897-y.

Supplementary Material 1.

Acknowledgements

Not applicable.

Author contributions

YO: implemented the software, analysed the data, and co-wrote the paper. MA: analysed the data and co-wrote the paper. YS: designed and supervised the research, analysed the data, and co-wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas "Frontier Research on Chemical Communications" [no. 17H06410] from the Ministry of Education, Culture, Sports, Science and Technology of Japan. This work was also supported by Grant-in-Aid for Transformative Research Area (A) "Latent Chemical Space" [no. 23H04885 (Y.S.)] from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Availability of data and materials

The source code for STRGNN, along with the dataset for performance evaluation, and instructions (README file) on how to use the program are available at our GitHub site: https://github.com/yuto-ohnuki/STRGNN.git.

Declarations

Competing interests

The authors declare no competing interests.

Received: 14 January 2024 Accepted: 12 August 2024
Published online: 23 August 2024

References

1. Hodos RA, Kidd BA, Shameer K et al (2016) In silico methods for drug repurposing and pharmacology. Wiley Interdiscip Rev Syst Biol Med 8:186–210
2. Luo H, Li M, Yang M et al (2021) Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinform 22:1604–1619
3. Grover A, Leskovec J (2016) node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, pp 855–864
4. Li J, Li J, Kong M et al (2021) SVDNVLDA: predicting lncRNA-disease associations by singular value decomposition and node2vec. BMC Bioinform 22:538
5. Luo H, Wang J, Li M et al (2019) Computational drug repositioning with random walk on a heterogeneous network. IEEE/ACM Trans Comput Biol Bioinform 16:1890–1900
6. Cheng F, Desai RJ, Handy DE et al (2018) Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun 9:2691
7. Luo H, Li M, Wang S et al (2018) Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics 34:1904–1912
8. Xuan P, Cao Y, Zhang T et al (2019) Drug repositioning through integration of prior knowledge and projections of drugs and diseases. Bioinformatics 35:4108–4119
9. Zhang W, Xu H, Li X et al (2020) DRIMC: an improved drug repositioning approach using Bayesian inductive matrix completion. Bioinformatics 36:2839–2847
10. Yu Z, Huang F, Zhao X et al (2020) Predicting drug–disease associations through layer attention graph convolutional network. Brief Bioinform 22:bbaa243
11. Chiang AP, Butte AJ (2009) Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther 86:507–510
12. Nagamine N, Sakakibara Y (2007) Statistical prediction of protein–chemical interactions based on chemical structure and mass spectrometry data. Bioinformatics 23:2004–2012
13. Gottlieb A, Stein GY, Ruppin E, Sharan R (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 7:496
14. Zhou R, Lu Z, Luo H et al (2020) NEDD: a network embedding based method for predicting drug-disease associations. BMC Bioinform 21:387
15. Watanabe N, Ohnuki Y, Sakakibara Y (2021) Deep learning integration of molecular and interactome data for protein-compound interaction prediction. J Cheminform 13:36
16. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv [cs.LG]
17. Velickovic P, Cucurull G, Casanova A et al (2017) Graph attention networks. Stat 1050:20
18. Schlichtkrull M, Kipf TN, Bloem P et al (2018) Modeling relational data with graph convolutional networks. The semantic web. Springer International Publishing, Berlin, pp 593–607
19. Zeng X, Zhu S, Liu X et al (2019) deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics 35:5191–5198
20. Wan F, Hong L, Xiao A et al (2019) NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. Bioinformatics 35:104–111
21. Wang Z, Zhou M, Arnold C (2020) Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. Bioinformatics 36:i525–i533
22. Cai L, Lu C, Xu J et al (2021) Drug repositioning based on the heterogeneous information fusion graph convolutional network. Brief Bioinform. https://doi.org/10.1093/bib/bbab319
23. Barabási A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56
24. Sun YV, Hu Y-J (2016) Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Adv Genet 93:147–190
25. Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18:83
26. Graw S, Chappell K, Washam CL et al (2021) Multi-omics data integration considerations and study design for biological systems and disease. Mol Omics 17:170–185
27. Iwata M, Sawada R, Iwata H et al (2017) Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. Sci Rep 7:40164
28. Duan Q, Flynn C, Niepel M et al (2014) LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res 42:W449–W460
29. Wang Y, Yang Y, Chen S, Wang J (2021) DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. Brief Bioinform. https://doi.org/10.1093/bib/bbab048
30. Lahat D, Adali T, Jutten C (2015) Multimodal data fusion: an overview of methods, challenges, and prospects. Proc IEEE 103:1449–1477
31. Luo H, Wang J, Li M et al (2016) Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics 32:2664–2671
32. Zhang X, Wang W, Ren C-X, Dai D-Q (2022) Learning representation for multiple biological networks via a robust graph regularized integration approach. Brief Bioinform. https://doi.org/10.1093/bib/bbab409
33. Peng L, Yang C, Huang L et al (2022) RNMFLP: predicting circRNA-disease associations based on robust nonnegative matrix factorization and label propagation. Brief Bioinform. https://doi.org/10.1093/bib/bbac155
34. Yang M, Luo H, Li Y, Wang J (2019) Drug repositioning based on bounded nuclear norm regularization. Bioinformatics 35:i455–i463
35. Wishart DS, Feunang YD, Guo AC et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074–D1082
36. Szklarczyk D, Gable AL, Nastou KC et al (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49:D605–D612
37. Luck K, Kim D-K, Lambourne L et al (2020) A reference map of the human binary protein interactome. Nature 580:402–408
38. Davis AP, Grondin CJ, Johnson RJ et al (2021) Comparative toxicogenomics database (CTD): update 2021. Nucleic Acids Res 49:D1138–D1143
39. Wishart DS, Guo A, Oler E et al (2022) HMDB 5.0: the human metabolome database for 2022. Nucleic Acids Res 50:D622–D631
40. Wang Z, Monteiro CD, Jagodnik KM et al (2016) Extraction and analysis of signatures from the gene expression omnibus by the crowd. Nat Commun 7:12846
41. Rouillard AD, Gundersen GW, Fernandez NF et al (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database. https://doi.org/10.1093/database/baw100
42. Kang J, Tang Q, He J et al (2022) RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. Nucleic Acids Res 50:D326–D332
43. Huang Z, Shi J, Gao Y et al (2019) HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res 47:D1013–D1017
44. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
45. Barrett T, Wilhite SE, Ledoux P et al (2012) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995
46. Esteller M (2011) Non-coding RNAs in human disease. Nat Rev Genet 12:861–874

47. Wang D, Wang J, Lu M et al (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics 26:1644–1650
48. Zhang Z, Tang W (2018) Drug metabolism in drug discovery and development. Acta Pharm Sin B 8:721–732
49. Wishart DS (2016) Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov 15:473–484
50. Jin S, Zeng X, Xia F et al (2021) Application of deep learning methods in biological networks. Brief Bioinform 22:1902–1917
51. Taboada B, Verde C, Merino E (2010) High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res 38:e130
52. Lotfi Shahreza M, Ghadiri N, Mousavi SR et al (2018) A review of network-based approaches to drug repositioning. Brief Bioinform 19:878–892
53. Tafer H, Hofacker IL (2008) RNAplex: a fast tool for RNA-RNA interaction search. Bioinformatics 24:2657–2663
54. Long Y, Wu M, Liu Y et al (2022) Pre-training graph neural networks for link prediction in biomedical networks. Bioinformatics. https://doi.org/10.1093/bioinformatics/btac100
55. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754
56. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480–D489
57. Deng J, Yang Z, Wang H et al (2023) A systematic study of key elements underlying molecular property prediction. Nat Commun 14:6395
58. Agrawal S, Sisodia DS, Nagwani NK (2023) Function characterization of unknown protein sequences using one hot encoding and convolutional neural network based model. In: Lecture Notes in Electrical Engineering. Springer Nature Singapore, Singapore, pp 267–277
59. Rong Y, Huang W, Xu T, Huang J (2019) DropEdge: towards deep graph convolutional networks on node classification. arXiv [cs.LG]
60. Zhou K, Dong Y, Wang K, et al (2021) Understanding and resolving performance degradation in deep graph convolutional networks. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Association for Computing Machinery, New York, pp 2728–2737
61. Kipf TN, Welling M (2016) Variational graph auto-encoders. arXiv [stat.ML]
62. Guo W, Wang J, Wang S (2019) Deep multimodal representation learning: a survey. IEEE Access 7:63373–63394
63. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
64. Wang W, Yang S, Zhang X, Li J (2014) Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics 30:2923–2930
65. Tang S-M, Deng X-T, Zhou J et al (2020) Pharmacological basis and new insights of quercetin action in respect to its anti-cancer effects. Biomed Pharmacother 121:109604
66. Bhatia NK, Modi P, Sharma S, Deep S (2020) Quercetin and baicalein act as potent antiamyloidogenic and fibril destabilizing agents for SOD1 fibrils. ACS Chem Neurosci 11:1129–1138
67. Ip P, Sharda PR, Cunningham A et al (2017) Quercitrin and quercetin 3-β-d-glucoside as chemical chaperones for the A4V SOD1 ALS-causing mutant. Protein Eng Des Sel 30:431–440
68. Tang L, Cao L, Pelech S et al (2003) Cytokines and signal transduction pathways mediated by anthralin in alopecia areata-affected Dundee experimental balding rats. J Investig Dermatol Symp Proc 8:87–90
69. Palladino MA, Bahjat FR, Theodorakis EA, Moldawer LL (2003) Anti-TNF-alpha therapies: the next generation. Nat Rev Drug Discov 2:736–746
70. Somberg JC, Timar S, Bailin SJ et al (2004) Lack of a hypotensive effect with rapid administration of a new aqueous formulation of intravenous amiodarone. Am J Cardiol 93:576–581
71. Trevelyan J, Brull DJ, Needham EWA et al (2004) Effect of enalapril and losartan on cytokines in patients with stable angina pectoris awaiting coronary artery bypass grafting and their interaction with polymorphisms in the interleukin-6 gene. Am J Cardiol 94:564–569
72. Lowenstein EB, Lowenstein EJ (2011) Isotretinoin systemic therapy and the shadow cast upon dermatology's downtrodden hero. Clin Dermatol 29:652–661

Publisher's Note