Current Research in Structural Biology 7 (2024) 100122

Contents lists available at ScienceDirect

Current Research in Structural Biology

journal homepage: www.journals.elsevier.com/current-research-in-structural-biology

Research Article

SAGESDA: Multi-GraphSAGE networks for predicting SnoRNA-disease associations

Biffon Manyura Momanyi [a], Yu-Wei Zhou [b], Bakanina Kissanga Grace-Mercure [c], Sebu Aboma Temesgen [c], Ahmad Basharat [c], Lin Ning [b,c], Lixia Tang [c], Hui Gao [a,**], Hao Lin [c,*], Hua Tang [d,e,f,***]

[a] School of Computer Science and Engineering, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China
[b] School of Health Care Technology, Chengdu Neusoft University, Chengdu, China
[c] School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China
[d] School of Basic Medical Sciences, Southwest Medical University, Luzhou, 646000, China
[e] Basic Medicine Research Innovation Center for Cardiometabolic Diseases, Ministry of Education, Luzhou, 646000, China
[f] Central Nervous System Drug Key Laboratory of Sichuan Province, Luzhou, 646000, China

ARTICLE INFO

Handling Editor: Prof N Chandra

Keywords:
Small nucleolar RNAs
Diseases
snoRNA-disease associations
GraphSAGE
Heterogeneous network

ABSTRACT

Over the years, extensive research has highlighted the functional roles of small nucleolar RNAs in various biological processes associated with the development of complex human diseases. Therefore, understanding the existing relationships between different snoRNAs and diseases is crucial for advancing disease diagnosis and treatment. However, classical biological experiments for identifying snoRNA-disease associations are expensive and time-consuming. Therefore, there is an urgent need for cost-effective computational techniques that can enhance the efficiency and accuracy of prediction. While several computational models have already been proposed, many suffer from limitations and suboptimal performance. In this study, we introduced a novel Graph Neural Network-based (GNN) classification model, called SAGESDA, which is implemented through the GraphSAGE architecture with attention for the prediction of snoRNA-disease associations. The classifier leverages local neighbouring nodes in a heterogeneous network to generate new node embeddings through message passing. The mini-batch gradient descent technique was applied to divide the graph into smaller sub-graphs, which enhances the model's accuracy, speed and scalability. With these advancements, SAGESDA attained an area under the receiver operating characteristic (ROC) curve (AUC) of 0.92 using the standard dot product classifier, surpassing previous related studies. This notable performance demonstrates that SAGESDA is a promising model for predicting unknown snoRNA-disease associations with high accuracy. The SAGESDA implementation details can be obtained from https://github.com/momanyibiffon/SAGESDA.git.

1. Introduction

With the advent of sequencing technologies and technical advancements, substantial studies have been conducted to explore the associations between small nucleolar RNAs (snoRNAs) and human diseases. It has been increasingly evident that snoRNAs play a critical role in various biological processes involved in the development and progression of diseases (Zhang and Liu, 2022) such as cancer, hereditary disorders, hematopoiesis etc. (Liao et al., 2010)

These snoRNAs, which measure 60–300 nucleotides in length, are unique structural components found throughout the nucleoli of eukaryotic cells that regulate the maturation and post-transcriptional alteration of ribosomal RNAs(Esteller, 2011). They are classified into two groups: box C/D snoRNAs and box H/ACA snoRNAs(Liao et al., 2010), distinguished by their motifs, architectures, and chemical alterations (Reichow et al., 2007). SnoRNAs are integral components of the extensively studied ribonucleoprotein complexes known as snoRNPs. They belong to a diverse family of short non-coding RNAs that are

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: huigao@uestc.edu.cn (H. Gao), hlin@uestc.edu.cn (H. Lin), huatang@swmu.edu.cn (H. Tang).

https://doi.org/10.1016/j.crstbi.2023.100122
Received 4 October 2023; Received in revised form 30 November 2023; Accepted 24 December 2023
Available online 29 December 2023

crucial for ribosome synthesis in all eukaryotes (Dieci et al., 2009).

Increasing evidence has revealed the transcriptional and post-transcriptional regulatory functions of snoRNAs. For example, SNHG2 has been identified as a tumour suppressor and its downregulation is associated with lung cancer (Lin et al., 2020). Additionally, snoRNAs such as SNORD115-32 and SNORD114-22 have been implicated in cerebral cavernous deformities in the brainstem (Kar et al., 2018), while the micro-deletion of SNORD116 is the major cause of Prader Willi syndrome (Qi et al., 2016). As a result, understanding snoRNA-disease associations (SDA) and other complex disease mechanisms is of great importance for identifying biomarkers and advancing disease diagnosis and prognosis.

Traditional clinical studies are time-consuming, labour-intensive, and costly, making computational methods highly valuable for exploring non-coding RNA-disease interactions (Momanyi et al., 2023). Several computational models have demonstrated cost-effectiveness and impressive performance in this field. For example, (Chen et al., 2021) developed a DBNMDA model based on Boltzmann machine to predict the associations between microRNA (miRNA) and diseases, achieving a high area under the receiver operating characteristic curve (AUC) values through global and local leave-one-out cross-validation (LOOCV). Similarly, (Yan et al., 2022) proposed PDMDA based on GNN, which not only predicted miRNA-disease interactions but also identified the different types of miRNA-disease associations. In the realm of long non-coding RNA (lncRNA) and diseases, (Lu et al., 2018) presented the SIMCLDA technique for locating probable lncRNA-disease interactions based on inductive matrix completion, while (Ping et al., 2018) introduced a bipartite network-based predictive model for lncRNA-disease interactions based on the data acquired from LncRNADisease, Lnc2Cancer, and MNDR databases. In another study, (Lan et al., 2020) proposed the LDICDL model for predicting lncRNA-disease associations through collaborative deep learning. In this model, matrix decomposition was performed before applying an automatic encoder. To address the limitations of matrix decomposition, a hybrid model was developed to denoise multiple lncRNA feature information and multiple disease feature information. All these computational methods mentioned above achieved notable AUC values for predicting known lncRNA-disease associations.

Compared to these types of RNA-based link prediction models, such as drug-disease (Yu et al., 2021), lncRNA-disease (Yu et al., 2021) (Ping et al., 2018) miRNA-disease (Yan et al., 2022) and circular RNA (circRNA)-disease (Yang and Lei, 2021) associations, there is still lack of sufficient computational methods for predicting snoRNA-disease associations. Nowadays, there are only a few research tools on the relationships between snoRNAs and diseases. (Sun et al., 2022) introduced the PSnoD methodology based on bounded nuclear norm regularization (BNNR). Using the 5-fold stratified shuffling, this matrix completion approach achieved an AUC of 0.90 and an area under the precision-recall curve (AUPRC) of 0.55. Another network-based model, called iSnoDi-LSGT, was proposed by Zhang and Liu (2022). This model incorporated snoRNA sequences and disease similarity as local similarity constraints, utilizing network embedding technology to extract snoRNA and disease characteristics. Global topological constraints were then calculated to identify snoRNA-disease associations. Similarly, Liu et al. (2021) presented GCNSDA, a graph neural network (GNN)-based model for snoRNA-disease associations identification. By leveraging the bipartite graph of snoRNAs and diseases, GCNSDA achieved an average AUC of 0.8865 using the advanced GNN algorithm and 5-fold cross-validation. However, a notable constraint of this model was its limited applicability to the prediction of novel snoRNAs or diseases, requiring graph reconstruction and model retraining when new snoRNAs or diseases were introduced.

In summary, while substantial progress has been made in understanding snoRNA-disease associations, computational models for predicting these interactions are still relatively scarce. However, the existing computational models have demonstrated cost-effectiveness and impressive performance, paving the way for further advancements in this area of research.

To address the limitations identified in previous studies, this research introduces the SAGESDA model for predicting potential snoRNA-disease associations. The approach involves constructing a heterogeneous snoRNA-disease network by integrating snoRNA similarity, disease similarity, and snoRNA-disease association networks. These networks are then divided into mini-batches to extract snoRNA and disease embeddings from multiple perspectives using three heterogeneous networks. The GraphSAGE GNN framework is employed to generate new node embeddings, resulting in a comprehensive representation that combines snoRNA and disease characteristics as edge features. The potential associations are predicted using the dot product classifier. This novel model aims to overcome the challenges faced by previous approaches and provide improved predictions for snoRNA-disease associations.

2. Materials and methods

2.1. Datasets

In this study, the dataset used was obtained from Sun et al. (2022) which was deemed reliable for efficient prediction (Zulfiqar et al., 2022). The dataset consists of 27 diseases, 220 snoRNAs and 459 known snoRNA-disease associations. The original source of this dataset is the Mammalian ncRNA-Disease Repository (MNDR) v3.1 (http://www.rna-society.org/mndr/), which curates experimentally verified and predicted ncRNA-disease associations from the literature and other relevant resources (Ning et al., 2021).The snoRNA-disease associations were represented in an adjacency matrix $A \in \mathbb{R}^{(sn \times dn)}$ where variables *sn* and *dn* represent the total number of snoRNAs and diseases, respectively.

2.2. SnoRNA and disease pairwise similarity

To obtain snoRNA pairwise similarity information (SS), we utilized the similarity information provided by Sun et al. (2022). Following the implementation of k-mer (4-mer) for feature extraction, they obtained a fixed length of numeric vectors from the different length sequences which led to an $s \times 4^k$ feature matrix denoted as $F_{s \times 4^k}$ using a sample set of snoRNAs denoted as S. Subsequently, the Tanimoto coefficient which evaluates the angle difference and length difference between the two vectors using dot products and square lengths in the denominator was applied to calculate the snoRNA pairwise similarity (Lipkus, 1999; Kryszkiewicz, 2013). Additionally, recording both the angle and length variances can improve the performance of specific domains (Anastasiu and Karypis, 2017), thus the vector similarity for real-valued vectors can be determined as given in Eq. (1) using the Tanimoto coefficient, which extends the Jaccard similarity coefficient (Ayub et al., 2020).

$$\boldsymbol{d}(s_i s_j) = \frac{\langle s_i s_j \rangle}{\| s_i \|^2 + \| s_j \|^2 - \langle s_i s_j \rangle} \tag{1}$$

where $s_i s_j$ refers to the vector dot product of snoRNAs s_i and s_j calculated as shown in Eq. (2), while $\| s_i \| = \sqrt{\langle s_i s_j \rangle}$ refers to the Euclidean distance between the two snoRNAs.

$$\langle s_i s_j \rangle = \sum_{k=1}^{n} s_{i,k} \times s_{j,k} \tag{2}$$

To determine the disease pairwise similarity (DS), we adopted the semantic similarity approach based on directed acyclic graphs (DAG) proposed by Wang et al. (Tao, 2019). The DAG principle states that the semantic similarity between two diseases is influenced by the number of shared ancestors they have. The contribution of similarity is also weighted based on the proximity of the ancestors to the target node. Thus, common ancestors may have different impacts on the semantic

similarity score for diseases, with closer common ancestors indicating higher similarity between the diseases. This strategy was employed because the MeSH terms were organized in a hierarchical tree-like structure, resembling a standard DAG graph. The relevant disease Mesh ID and the gene ID of each snoRNA are listed in the MNDR data. MeSH, developed by the National Library of Medicine (NLM), is a controlled vocabulary used for indexing, cataloguing, and searching articles related to health and bio-medicine in repositories like PubMed/MEDLINE(Lipscomb, 2000). Fig. 1 illustrates the feature distributions of the disease and snoRNA similarity matrices used in our proposed model.

2.3. Heterogeneous network construction

In contrast to homogeneous networks, heterogeneous networks consist of nodes and edges of various types, each associated with unstructured contents (Zhang et al., 2019). For instance, a user-based network may depict associations between users and items, while an academic graph may represent relationships between authors and papers. After obtaining datasets containing snoRNA and disease pairwise similarities and their association network, we identified a total of 459 known associations between 220 snoRNAs and 27 diseases. The association information was then presented in an adjacency matrix $A \in \mathbb{R}^{(sn \times dn)}$, where A_{ij} is equal to 1 if there exists a known association between snoRNA *i* and disease *j*, and 0 otherwise. With snoRNA and disease nodes as the two types of nodes, we constructed a heterogeneous graph $G = (V, E)$ that incorporates snoRNA and disease nodes along with their corresponding edges based on *A*, the graph structure is shown in Fig. 2, illustrating the snoRNA-disease associations and two-way message passing for feature aggregation. The graph data object was generated comprising 247 nodes and 459 edges, which were further split into training, validation and test sets using the RandomLinkSplit function from PyTorch Geometric to facilitate model training and evaluation. The function randomly divides the edges in a way that the training split does not include edges from other splits, while the validation split does not include the edges found in the test split set. Therefore, 80% of the graph data was used for training, 10% for evaluation and 10% for testing. In the training edges, 70% were used for message passing and 30% for supervision, this was determined by the disjoint_train_ratio parameter of the RandomLinkSlit function which was set to 0.3. Also, fixed negative edges were generated at a rate of 0.1 for model evaluation, while during training, the negative edges were generated on the fly with mini batches generated at a negative sampling rate of 0.2.

According to the definitions provided by Chen et al. (2022), a heterogeneous network is denoted as $G = (V, E, D, R)$, where *V* represents the collection of nodes i.e., snoRNA and disease nodes, *E* denotes the collection of edges in the network, *D* signifies the set of node types denoted as $D = \{D_1, D_2, ..., D_n, ..., D_N\} (N \leq |V|)$, *R* represents the set of edge types. Each node $v_i \in V$ belongs to a given type of node and can be represented as $\varphi(v_i) = D_n \in D (1 \leq n \leq N)$. The total number of node types is represented by $N = |D|$. For a given edge $e_j = (v_i v_j) \in E$, it belongs to a given relation type indicated as $\varphi(e_j) \in R$, where the number of edge types can be defined as $M = |R|$. Corresponding to another definition by Chen et al. (2022), in the proposed network embedding, a mapping function $f : V \rightarrow X \in R^{|V| \times b}, b \ll |V|$ is trained to generate new vector representations of the nodes, which capture both the structural and semantic links between different nodes. The heterogeneous network developed for the proposed model is described based on the known interactions, as indicated in Eq. (3).

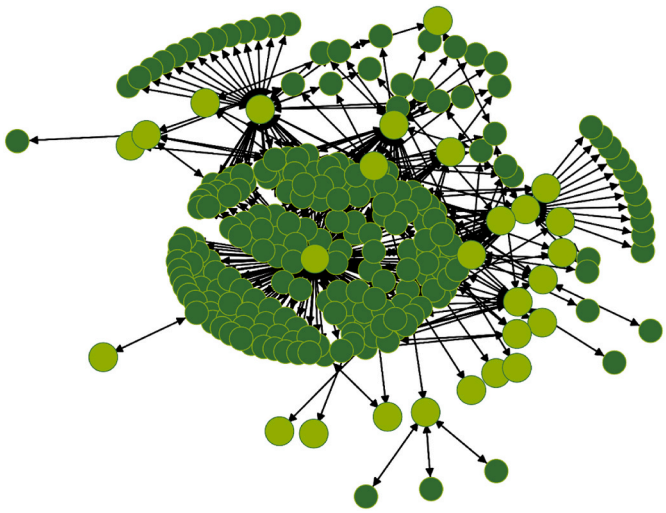

Fig. 2. An illustration of the heterogeneous graph G = (V, E) comprising the two types of nodes (snoRNA and disease nodes) and their edges indicating the known associations between two adjacent nodes. The dark and light green nodes represent the disease and snoRNA nodes, respectively, while the undirected links represented by two direction arrows were set to facilitate two-way message-passing during node feature aggregation. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

$$G = (V, E, D, R), = \begin{bmatrix} SS & A^T \\ A & DS \end{bmatrix} \quad (3)$$

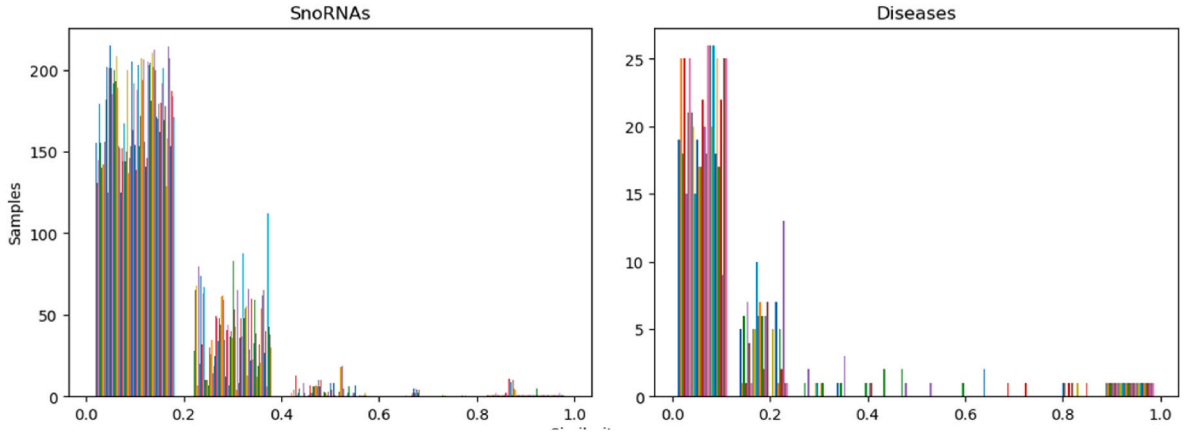

Fig. 1. SnoRNA and disease matrices' pairwise similarity distribution demonstrating the degree of correlation for specific snoRNA and disease pairs. A pair correlation that is weaker and stronger, respectively, is indicated by the values closer to 0 and 1.

2.4. Case study

After training and testing the SAGESDA model, a case study was conducted to further validate and assess the effectiveness of the model in predicting snoRNA-disease associations. In this case study, a specific disease and its associated snoRNAs were selected to generate a sub-graph, which served as the testing data for the model. Known associations between the selected disease and associated snoRNAs were intentionally removed from the sub-graph to simulate previously unseen data. To evaluate the predictions, the top ten scores obtained from the model were ranked in descending order, with higher scores indicating a high likelihood of association and vice-versa. For this case study, two diseases namely, lung neoplasms and stomach neoplasms, were selected due to their high prevalence worldwide. Lung cancer, also known as lung neoplasms, is the leading cause of cancer-related deaths globally (Tao, 2019), while gastric cancer, also known as stomach neoplasms, ranks fifth and fourth globally in incidence and fatality, respectively (Sung et al., 2021). The top ten predicted snoRNAs associated with lung neoplasms and stomach neoplasms were then verified by the RNADisease v4.0 database (Chen et al., 2023). Eight out of the top 10 snoRNAs predicted to be associated with both lung and stomach neoplasms were successfully confirmed as shown in Table 1, a clear indication of the SAGESDA's effectiveness in accurate prediction of snoRNA-disease associations.

3. Theory / calculation

The proposed method, known as SAGESDA, utilizes multiple graph neural networks (GNNs) and is developed using the GraphSAGE framework along with a dot product classifier for the prediction of probable snoRNA-disease interactions. To achieve this, the snoRNA and disease pairwise similarity data, along with their known associations, are mapped into three heterogeneous networks. These networks are then utilized in a three-layer GraphSAGE network for training, resulting in the generation of snoRNA and disease embeddings. These embeddings serve as the final feature representations for predicting snoRNA-disease associations. The link-level prediction is performed using a dot product classifier, which identifies likely associations. The complete SAGESDA process is illustrated in Fig. 3.

3.1. The GraphSAGE with attention

GraphSAGE (Graph Sampling and Aggregation) is a framework for GNNs that enables inductive representation learning. It generates node embeddings, which are low-dimensional representations capturing both structural and semantic information of each node. Unlike other GNN frameworks that rely on a single fixed graph, GraphSAGE is capable of generalizing to new nodes efficiently (Afoudi et al., 2023). It achieves this by using the immediate neighbourhood of each node to train its embedding (Zhang et al., 2020). This approach allows for the anticipation of embeddings for unseen nodes without requiring model retraining (Hamilton et al., 2017).

GraphSAGE collects node information through aggregator functions, considering a specific number of hops away from the target node. Each iteration involves two essential operations: sampling and aggregation. For each node u, a fixed number of neighbours are chosen using a random walk-based sampling strategy during the sampling phase i.e., $N(u) = \{v_1, v_2, ..., v_n\}(v \in G)$, and in the aggregation phase, the embeddings of the sampled neighbours are combined to create a new representation for the central node, denoted as u (Zhang et al., 2020). This process is illustrated in Eq. (4), where the mean aggregation function is applied.

$$\boldsymbol{h}_u^{(k)} = \boldsymbol{\sigma}\left(\boldsymbol{W}.\boldsymbol{MEAN}\left(\left\{\boldsymbol{h}_u^{(k-1)}\right\} \cup \left\{\boldsymbol{h}_u^{(k-1)}, \forall \boldsymbol{v} \in \boldsymbol{N}(\boldsymbol{u})\right\}\right)\right) \tag{4}$$

where σ (.) refers to the non-linear activation function, W is the weight matrix, v is the neighbours of the target node u, h_u^{k-1} refers to the previous node representation, while the aggregated neighbourhood node information in the layer k can be represented as $h^k_{N(u)}$. In addition to considering feature significance (Zulfiqar et al., 2022), the Graph Attention Network (GAT) introduced multi-head attention, enabling the model to selectively attend to different neighbours based on their importance and various input feature elements across different heads. GAT learns multiple representations of each graph, capturing complex interactions to improve performance, as described in Eq. (5), where $\alpha_{uv} = \frac{1}{|N(u)|}$ refers to the weighing factor determining the importance of the message of node v to node u.

$$\boldsymbol{h}_u^k = \boldsymbol{\sigma}\left(\sum_{v \in N(u)} \boldsymbol{\alpha}_{vu} \boldsymbol{W}^k \boldsymbol{h}_v^{k-1}\right) \tag{5}$$

For two-way message passing, Eq. (4) was adjusted as shown in Eq. (6), where $m^k_{j,i}$ represents the incoming message from node j to node i in layer k, and $N^{'}(i)$ represents the set of nodes sending messages to node i. The neighbouring node representations and incoming messages from the nodes that send messages to node i are used as the inputs for the aggregation function. This function applies a non-linear activation function σ and estimates the mean of the representations and incoming messages. The resulting vector becomes the updated representation of node i in the next layer of the network.

$$\boldsymbol{AGGRERATE}\left(\left\{\boldsymbol{h}_j^k : \boldsymbol{j} \in \boldsymbol{N}(\boldsymbol{i})\right\} \cup \left\{\boldsymbol{m}_{j,i}^k : \boldsymbol{j} \in \boldsymbol{N}^{'}(\boldsymbol{i})\right\}\right) = \left(\frac{\mathbf{1}}{|\boldsymbol{N}(\boldsymbol{i})| + |\boldsymbol{N}^{'}(\boldsymbol{i})|}\left(\sum_{j \in N(i)} \boldsymbol{h}_j^k + \sum_{j \in N^{'}(i)} \boldsymbol{m}_{j,i}^k\right)\right) \tag{6}$$

To deploy the SAGESDA network, we combine multiple heterogeneous networks trained using similarity and direct interaction information for snoRNA and disease nodes, as shown in Eq. (7). Here, G represents the combined heterogeneous graph, n is the total number of combined networks, V_i and E_i denote the nodes and edges in the i^{th}

Table 1
The top 10 snoRNA predictions associated with Lung and Stomach neoplasms as obtained by the SAGESDA model, confirmed via the RNADisease database.

Lung neoplasms		Evidence	Stomach cancer		Evidence
snoRNA ID	RNA Symbol		snoRNA id	RNA Symbol	
692235	SNORD103B	Confirmed	692198	SNORD78	Confirmed
767569	SNORD113-9	Confirmed	8944	SNORD73A	Confirmed
767577	SNORD114-1	Confirmed	26770	SNORD79	Confirmed
692084	SNORD13	Unconfirmed	692215	SNORD112	Unconfirmed
595097	SNORD16	Confirmed	94161	SNORD46	Confirmed
9301	SNORD27	Confirmed	9301	SNORD27	Confirmed
26798	SNORD51	Confirmed	26818	SNORD33	Confirmed
26788	SNORD60	Confirmed	619564	SNORD72	Confirmed
767566	SNORD113-6	Confirmed	100113382	SNORD105B	Confirmed
100033454	SNORD115-16	Unconfirmed	692234	SNORD103A	Unconfirmed

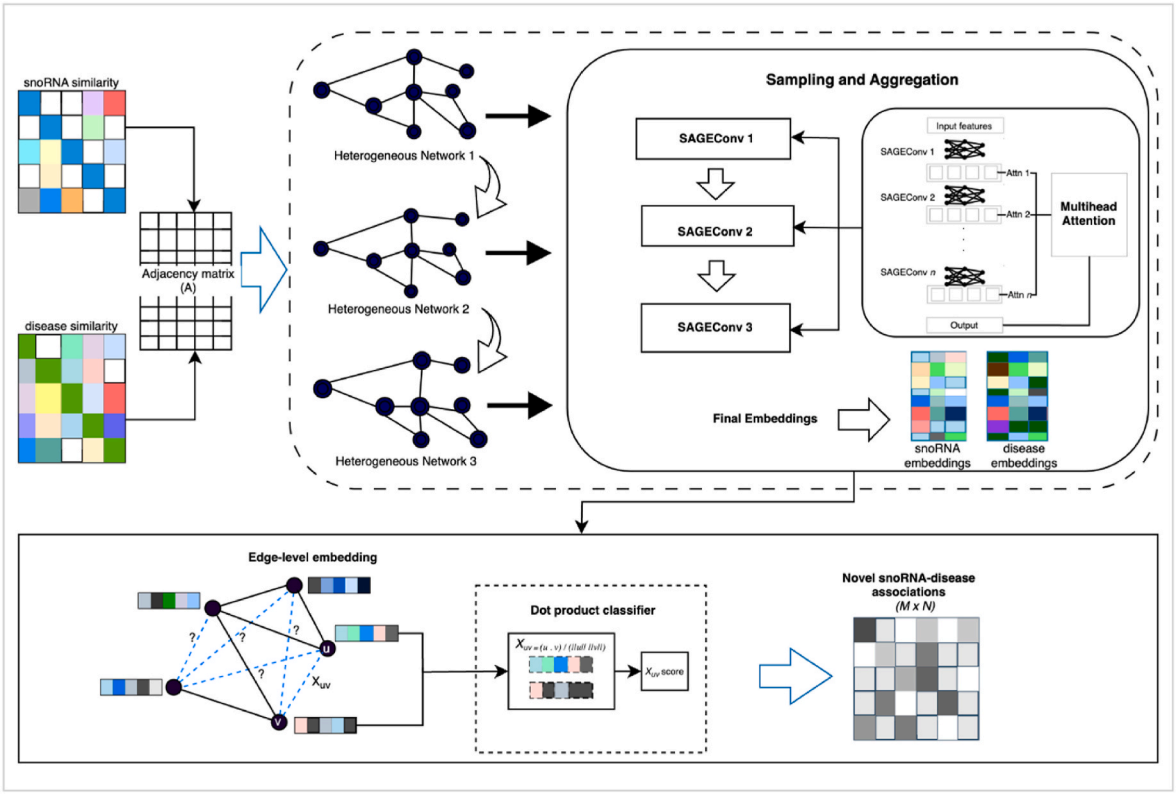

Fig. 3. An illustration of the proposed SAGESDA model for the prediction of snoRNA-disease associations. The dot product classifier is primarily used to forecast potential interactions, while sampling and aggregation via the GraphSAGE framework helps to obtain low-dimensional data.

GraphSAGE network, and D_i and R_i are the node and edge types of V_i and E_i, respectively.

$$G = \bigcup_{i=1}^{n} (V_i E_i D_i R_i) \tag{7}$$

Combining multiple graphs captures relationships from different perspectives, and shared weights across the graphs facilitate learning of common representations and information sharing between the graphs. Also, D_i and R_i ensure accurate identification of nodes from each distinct network in the final graph G. To obtain edge-level embeddings for link prediction, we perform feature aggregation between the two adjacent nodes v_i and v_j using a dot product as shown in Eqs. (8) and (9). This efficient calculation allows us to derive edge embeddings that reflect the similarity between nodes in the embedding space.

$$v_{i,j} = \sigma\left(W.CONCAT\left(v_i v_j\right)\right) \tag{8}$$

$$v_i.v_j = \sum_{x=1}^{n} v_i v_j \tag{9}$$

where $v_{i,j}$ in Eq. (8) refers to the edge-level embedding, σ refers to the activation function, and W and $CONCAT$ refer to the weight matrix and concatenation operation, respectively.

To control node contribution and prevent data overfitting, we introduce a penalty factor λ through a linear layer applied to the output embedding M, as shown in Eq. (10) which represents the first input of the SAGESDA. The result is then normalized to obtain the final node embeddings Ms and Md for snoRNAs and diseases, respectively.

$$M = \begin{bmatrix} \lambda \sim SS & A^T \\ A & \lambda \sim DS \end{bmatrix} \tag{10}$$

where λ is a value ranging from 0 to 1, with 0 or 1 denoting that M is only contributed to by either DS or SS, respectively. As a result, a λ value in-between enables a trade-off between DS and SS contributions.

4. Results and discussion

4.1. Implementation findings

The proposed SAGESDA model aims to predict snoRNA-disease associations by training three heterogeneous networks using a three-layer GraphSAGE network. This approach generates a rich set of final node embeddings that facilitate the prediction of potential associations. The model utilizes shared weights across different graphs to learn common node representations, enabling it to learn from multiple perspectives and capture detailed embeddings. To evaluate the performance of SAGESDA, we used the testing data and measured the AUC as the primary evaluation metric, where a higher AUC value indicates better model performance and vice versa. Besides, SAGESDA performance was evaluated based on the area under the precision-recall curve (AUPRC) and the F1-score for enhanced evaluation. AUC refers to a single scaler value popularly adopted in machine learning models to provide an aggregate measure of the model's ability of discriminating between different classes in classification tasks by evaluating the trade-off between true and false positive rates (Zhu et al., 2023; Zhang et al., 2023a; Zou et al., 2023). AUPRC on the other hand considers the trade-off between the model's precision and recall (Yang et al., 2023; Zhang et al., 2022).

By implementing a dot-product classifier, we successfully obtained the final embeddings and predicted potential associations based on the similarity scores between adjacent snoRNA-disease nodes. The dot product methodology calculates the cosine similarity between two node vectors u and v, represented as $score(u, v) = (u.v) / (\| u \| \| v \|)$, where $(u.v)$ denotes the dot product and $||u||$ and $||v||$ refer to the Euclidean norms of the vectors u and v, respectively. The similarity score ranges between −1 and 1, with a higher score indicating a stronger association between the two vectors, and vice versa.

To maintain the continuous nature of the similarity scores and rank the anticipated links based on their strength, we avoided using a threshold value. Instead, we evaluated the model performance by computing the AUC score over the predictions and corresponding ground-truth edges from the evaluation data, which includes both positive and negative edges. SAGESDA achieved an AUC of 0.92 and an AUPRC of 0.90 as depicted in Fig. 4 **(A)** and 4 **(B)**, respectively, alongside the training loss curve in Fig. 4 **(C).** The model also attained an F1-score of 0.80 indicating a balance between precision and recall**.** This performance was achieved using a learning rate of 0.00001 and 1500 epochs, which were determined as the optimal and most significant parameters for SAGESDA. The model was implemented using the PyTorch geometric deep learning library version 2.1.0.

The performance of the SAGESDA, as a state-of-the-art GNN-based approach, exhibited a significant improvement compared to previously proposed models for snoRNA-disease associations, such as PSnoD (Sun et al., 2022) and GCNSDA (Liu et al., 2021), which obtained AUCs of 0.90 and 0.8865, respectively.

In addition to comparing SAGESDA with other specifically designed models for snoRNA-disease associations, we evaluated its performance against a few other non-coding RNA-based computational models discussed in the literature. For instance, DBNMDA (Chen et al., 2021) attained an AUC of 0.9184 in global LOOCV, while PDMDA (Yan et al., 2022) obtained an average AUC of 0.7917 across three study tasks related to miRNA-disease associations. IMCLDA (Lu et al., 2018) achieved an AUC of 0.8237, and Ping et al. (2018) reported 0.9292 using the MNDR database for the prediction of lncRNA-disease interactions. Table 2 contains a summarized comparison through which SAGESDA proves to be a reliable model for snoRNA-disease association prediction, outperforming previous models and demonstrating its effectiveness in this field.

Table 2
The SAGESDA model performance in comparison with other baseline models previously proposed.

Model	AUC	Association Data
SAGESDA	0.92	SnoRNA - disease data (originally from MNDR)
PSnoD	0.90	
GCNSDA	0.8865	

4.2. Interpretation and analysis

SnoRNAs, a class of non-coding RNA molecules, play a crucial role in RNA processing and modification, and emerging evidence has highlighted their involvement in various diseases, including cancer, neurodegenerative disorders, and metabolic diseases etc. (Mannoor et al., 2012; Liu et al., 2020; Zhang et al., 2021; Ning et al., 2022; Ren et al., 2022). Understanding the associations between altered snoRNA expression levels and these diseases can offer valuable insights into their underlying molecular mechanisms (Liu et al., 2020; Zhang et al., 2021, 2023b; Ricciuti et al., 2016; Ren et al., 2023). To predict potential associations between snoRNAs and diseases, one promising approach is the utilization of Graph Neural Networks (GNN), a cutting-edge technique capable of capturing the intricate relationships between snoRNAs and diseases through heterogeneous networks. GNNs provide an effective means of modelling the complex interplay between snoRNAs and diseases, opening avenues for improved prediction and understanding of their associations.

In this study, we introduced SAGESDA, a novel model based on the GraphSAGE GNN architecture, for the prediction of potential snoRNA-disease associations. SAGESDA leverages three heterogeneous networks, along with a dot-product classifier for link prediction. The model was trained on a three-layer GraphSAGE network, and mini-batch loaders were utilized to generate sub-graphs for the model input. By incorporating these networks and employing GraphSAGE, SAGESDA offers an innovative approach to effectively capture the intricate relationships between snoRNAs and diseases, enhancing the accuracy and reliability of association predictions. This enriches the dataset given the initial low snoRNA-snoRNA and disease-disease similarity scores with the majority ranging between 0 and 0.2 as seen in Fig. 1. While the model's performance was not directly affected by the low similarity scores, it is evident that most of the items are not strongly correlated.

The utilization of mini-batch loaders played a crucial role in enhancing the efficiency of the model. By updating model parameters more frequently using smaller sub-datasets, the computational efficiency was improved. Additionally, processing a reduced amount of data at each iteration facilitated the model's ability to handle larger datasets, while simultaneously improving its generalization performance. This capability is particularly advantageous for biological networks, which tend to be extensive and intricate (Blundell et al., 2015). Furthermore, mini-batch loaders mitigated the risk of over-smoothing, a common challenge in GNN. Over-smoothing occurs when models generate similar embeddings for different nodes, resulting in the loss of valuable information and limiting the model's capacity to capture diverse graph information (Hamilton et al., 2017). Hence, the utilization of mini-batch loaders in our approach effectively addressed these issues, contributing to the overall effectiveness and robustness of the model.

Moreover, to regulate the contribution of each node during feature aggregation and prevent data overfitting, a penalty factor λ was introduced through a linear layer (Li et al., 2017). This addition significantly improved the performance of the model. However, the incorporation of

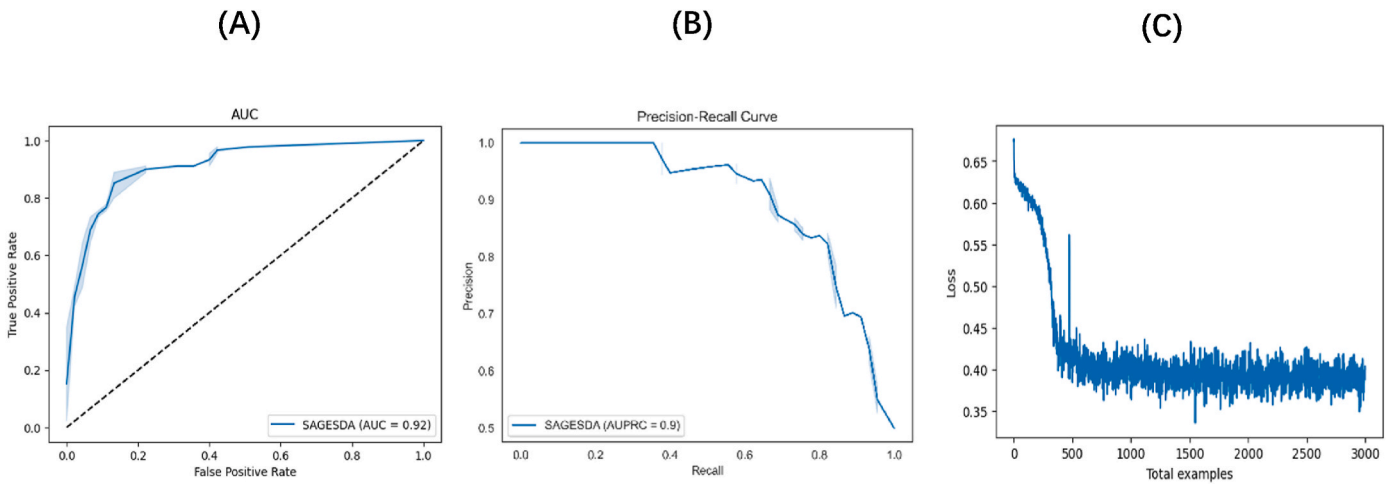

Fig. 4. An illustration of the testing AUC (A) and AUPRC (B) obtained by the SAGESDA using the dot product classifier, and the model training loss curve (C).

multi-head attention using the Graph Attention Network (GAT) framework increased the model's time complexity. The attention mechanism required additional computations, and the GraphSAGE-based model involved multiple rounds of neighbourhood aggregation. Despite the increased time complexity, the superiority of SAGESDA in learning edge-specific weights and emphasizing the most relevant neighbouring nodes outweighed this drawback. Similar to attention mechanisms, SAGESDA excelled at identifying and selecting the most important neighbours while disregarding less relevant ones. To evaluate the performance of SAGESDA, popular metrics such as the AUC were employed and the performance was compared to that of other previously proposed models. The results obtained using the dot-product classifier demonstrated that SAGESDA outperformed other baseline models and achieved state-of-the-art performance in predicting potential snoRNA-disease associations, achieving an impressive AUC of 0.92.

Note that we specifically employed a dot-product classifier for link prediction due to its simplicity and strong performance, particularly in scenarios with a large number of potential associations, as seen in the prediction of snoRNA-disease associations. The dot-product classifier is also notably powerful, especially on embedding-based models where the dot product between two adjacent nodes can be an effective measure of association based on the dot products. Additionally, during model training, we observed that reducing the learning rate improved the model performance. The best results were obtained with a learning rate of 0.00001 and the Adam optimizer, trained over 1500 epochs. The loss curve is illustrated in Fig. 4 **(C)**, with a steady downward trend as the model learns. However, some notable peaks in the loss curve occurred as a result of a challenging batch of data, leading to a temporary spike in loss.

While the GraphSAGE GNN architecture relied heavily on local neighbourhood information in this study, SAGESDA introduced several key innovations. Firstly, it leveraged local neighbourhood information to generate more accurate and reliable node embeddings, enhancing prediction efficiency. This is particularly crucial for large graphs where high time complexity poses a significant challenge. Secondly, SAGESDA employed multiple heterogeneous graphs, enabling detailed node representations by capturing essential information from different perspectives. Thirdly, the model incorporated a penalty factor that effectively regulated each node's contribution, preventing data overfitting. Another significant contribution was the ability to compute node embeddings for new snoRNA or disease nodes without requiring model retraining. This overcomes a major limitation of previous studies such as GCNSDA (Liu et al., 2021), which necessitated graph reconstruction and model retraining when introducing new snoRNAs or diseases.

With the SAGESDA, it becomes possible to generate node embeddings for newly introduced snoRNA or disease nodes and predict their associations with existing nodes. This is achieved with the help of the GraphSAGE architecture, which learns a function to create embeddings by aggregating and combining features from the local neighbourhood rather than the entire fixed graph. Despite the remarkable success of SAGESDA, it is important to acknowledge the model's major limitations of high time complexity, which increased notably after the introduction of multi-head attention and the model's high reliance on local neighbourhood. Besides, the model was implemented on a single dataset due to limited availability of compatible datasets. Therefore, future studies can be aimed at addressing these issues for improved model capacity. Other classification techniques can also be implemented for the classification of edges to determine the potential snoRNA-disease associations for further improvement of the SAGESDA's performance. Also, the SAGESDA model utilized the GraphSAGE architecture which was found to be the most suitable, however, future studies can implement other architectures for comparison with the performance of the SAGESDA.

5. Conclusion

As studies continuously highlight the significant role of snoRNAs in the emergence and progression of complex human diseases such as cancer, cardiovascular diseases, neurodegenerative diseases, etc., there is a pressing need for efficient computational models that can identify potential snoRNA-disease associations. The development of such models holds great promise in enhancing disease diagnosis, prognosis, and treatment by providing valuable insights into disease pathogenesis and potential therapeutic targets.

In this study, we proposed the SAGESDA model, based on the GraphSAGE GNN architecture, which has demonstrated promising performance compared to existing methods, achieving an AUC of 0.92. The high performance of SAGESDA positions it as a valuable tool in the healthcare sector. For example, it can assist researchers in identifying potential snoRNA-disease associations, leading to the development of new diagnostic and therapeutic strategies. Additionally, it can also contribute to the prediction of potential drug side effects by identifying snoRNAs associated with adverse drug reactions. Furthermore, the model supports the prioritization of drug targets by identifying snoRNAs linked to disease-specific molecular pathways. Hence, the model has significant implications for disease diagnosis, treatment, and drug development. Moreover, the SAGESDA model addresses several limitations identified in previous studies, making it a guiding principle for future experimental research. By overcoming these limitations, the model paves the way for further advancements in the field. Overall, the development of the SAGESDA model fills a crucial gap in computational models for snoRNA-disease associations and opens up new research avenues, ultimately contributing to improved healthcare outcomes.

Funding

This work was supported by the National Natural Science Foundation of China (62172343), the Sichuan Science and Technology Program (2022YFS0614) and Medico-Engineering Cooperation Funds from UESTC (ZYGX2021YGLH202).

CRediT authorship contribution statement

Biffon Manyura Momanyi: Conceptualization, Methodology, Software, Writing – original draft, preparation, Investigation. **Yu-Wei Zhou:** Formal analysis, Investigation. **Bakanina Kissanga Grace-Mercure:** Validation, Resources. **Sebu Aboma Temesgen:** Investigation, Formal analysis. **Ahmad Basharat:** Data curation. **Lin Ning:** Data curation, Writing – review & editing. **Lixia Tang:** Investigation, Formal analysis. **Hui Gao:** Supervision, Funding acquisition, Project administration. **Hao Lin:** Resources, Supervision, Funding acquisition. **Hua Tang:** Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

Afoudi, Y., Lazaar, M., Hmaidi, S., 2023. An enhanced recommender system based on heterogeneous graph link prediction. Eng. Appl. Artif. Intell. 124, 106553.

Anastasiu, D.C., Karypis, G., 2017. Efficient identification of Tanimoto nearest neighbors: all-pairs similarity search using the extended Jaccard coefficient. Int. J. Data Sci. Anal. 4, 153–172.

Ayub, M., Ghazanfar, M.A., Khan, T., et al., 2020. An effective model for Jaccard coefficient to increase the performance of collaborative filtering. Arabian J. Sci. Eng. 45, 9997–10017.

Blundell, C., Cornebise, J., Kavukcuoglu, K., et al., 2015. Weight uncertainty in neural network. In: International Conference on Machine Learning. PMLR, pp. 1613–1622.

Chen, X., Li, T.-H., Zhao, Y., et al., 2021. Deep-belief network for predicting potential miRNA-disease associations. Briefings Bioinf. 22, bbaa186.
Chen, X., Hao, T., Han, L., et al., 2022. Heterogeneous network embedding based on random walks of type and inner constraint. Mathematics 10, 2623.
Chen, J., Lin, J., Hu, Y., et al., 2023. RNADisease v4.0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. Nucleic Acids Res. 51, D1397–D1404.
Dieci, G., Preti, M., Montanini, B., 2009. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. Genomics 94, 83–88.
Esteller, M., 2011. Non-coding RNAs in human disease. Nat. Rev. Genet. 12, 861–874.
Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. 30.
Kar, S., Bali, K.K., Baisantry, A., et al., 2018. Genome-wide sequencing reveals small nucleolar RNAs downregulated in cerebral cavernous malformations. Cell. Mol. Neurobiol. 38, 1369–1382.
Kryszkiewicz, M., 2013. Using non-zero dimensions for the cosine and tanimoto similarity search among real valued vectors. Fundam. Inf. 127, 307–323.
Lan, W., Lai, D., Chen, Q., et al., 2020. LDICDL: LncRNA-disease association identification based on collaborative deep learning. IEEE ACM Trans. Comput. Biol. Bioinf 19, 1715–1723.
Li, J., Cheng, K., Wang, S., et al., 2017. Feature selection: a data perspective. ACM Comput. Surv. 50, 1–45.
Liao, J., Yu, L., Mei, Y., et al., 2010. Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. Mol. Cancer 9, 1–10.
Lin, Y., Holden, V., Dhilipkannah, P., et al., 2020. A non-coding RNA landscape of bronchial epitheliums of lung cancer patients. Biomedicines 8, 88.
Lipkus, A.H., 1999. A proof of the triangle inequality for the Tanimoto distance. J. Math. Chem. 26, 263–265.
Lipscomb, C.E., 2000. Medical Subject headings (MeSH). Bull. Med. Libr. Assoc. 88, 265–266.
Liu, T.Y., Zhang, Y.C., Lin, Y.Q., et al., 2020. Exploration of invasive mechanisms via global ncRNA-associated virus-host crosstalk. Genomics 112, 1643–1650.
Liu, D., Luo, Y., Zheng, J., et al., 2021. GCNSDA: predicting snoRNA-disease associations via graph convolutional network. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 183–188.
Lu, C., Yang, M., Luo, F., et al., 2018. Prediction of lncRNA–disease associations based on inductive matrix completion. Bioinformatics 34, 3357–3364.
Mannoor, K., Liao, J., Jiang, F., 2012. Small nucleolar RNAs in cancer. Biochim. Biophys. Acta Rev. Canc 1826, 121–128.
Momanyi, B.M., Zulfiqar, H., Grace-Mercure, B.K., et al., 2023. CFNCM: collaborative filtering neighborhood-based model for predicting miRNA-disease associations. Comput. Biol. Med., 107165
Ning, L., Cui, T., Zheng, B., et al., 2021. MNDR v3. 0: mammal ncRNA–disease repository with increased coverage and annotation. Nucleic Acids Res. 49, D160–D164.
Ning, L., Liu, M., Gou, Y., et al., 2022. Development and application of ribonucleic acid therapy strategies against COVID-19. Int. J. Biol. Sci. 18, 5070–5085.
Ping, P., Wang, L., Kuang, L., et al., 2018. A novel method for lncRNA-disease association prediction based on an lncRNA-disease association network. IEEE ACM Trans. Comput. Biol. Bioinf 16, 688–693.
Qi, Y., Purtell, L., Fu, M., et al., 2016. Snord116 is critical in the regulation of food intake and body weight. Sci. Rep. 6, 1–15.
Reichow, S.L., Hamma, T., Ferré-D'Amaré, A.R., et al., 2007. The structure and function of small nucleolar ribonucleoproteins. Nucleic Acids Res. 35, 1452–1464.
Ren, L., Xu, Y., Ning, L., et al., 2022. TCM2COVID: a resource of anti-COVID-19 traditional Chinese medicine with effects and mechanisms. Imeta e42.
Ren, L., Ning, L., Yang, Y., et al., 2023. MetaboliteCOVID: a manually curated database of metabolite markers for COVID-19. Comput. Biol. Med. 167, 107661.
Ricciuti, B., Mencaroni, C., Paglialunga, L., et al., 2016. Long noncoding RNAs: new insights into non-small cell lung cancer biology, diagnosis and therapy. Med. Oncol. 33, 1–12.
Sun, Z., Huang, Q., Yang, Y., et al., 2022. PSnoD: identifying potential snoRNA-disease associations based on bounded nuclear norm regularization. Briefings Bioinf. 23, bbac240.
Sung, H., Ferlay, J., Siegel, R.L., et al., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J. Clin. 71, 209–249.
Tao, M.-H., 2019. Epidemiology of Lung Cancer, Lung Cancer and Imaging.
Yan, C., Duan, G., Li, N., et al., 2022. PDMDA: predicting deep-level miRNA–disease associations with graph neural networks and sequence features. Bioinformatics 38, 2226–2234.
Yang, J., Lei, X., 2021. Predicting circRNA-disease associations based on autoencoder and graph embedding. Inf. Sci. 571, 323–336.
Yang, H., Luo, Y.M., Ma, C.Y., et al., 2023. A gender specific risk assessment of coronary heart disease based on physical examination data. NPJ Digit Med. 6, 136.
Yu, Z., Huang, F., Zhao, X., et al., 2021. Predicting drug–disease associations through layer attention graph convolutional network. Briefings Bioinf. 22, bbaa243.
Zhang, W., Liu, B., 2022. iSnoDi-LSGT: identifying snoRNA-disease associations based on local similarity constraints and global topological constraints. RNA 28, 1558–1567.
Zhang, C., Song, D., Huang, C., et al., 2019. Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage. Association for Computing Machinery, AK, USA, pp. 793–803.
Zhang, Z., Cui, P., Zhu, W., 2020. Deep learning on graphs: a survey. IEEE Trans. Knowl. Data Eng. 34, 249–270.
Zhang, Y., Liu, T., Wang, J., et al., 2021. Cellinker: a platform of ligand-receptor interactions for intercellular communication analysis. Bioinformatics 37 (14), 2025–2032.
Zhang, Z.Y., Ning, L., Ye, X., et al., 2022. iLoc-miRNA: extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism. Brief. Bioinform. 23 (5), bbac395.
Zhang, Y.-F., Wang, Y.-H., Gu, Z.-F., et al., 2023a. Bitter-RF: a random forest machine model for recognizing bitter peptides. Front. Med. 10, 1052923.
Zhang, Y., Pan, X., Shi, T., et al., 2023b. P450Rdb: a manually curated database of reactions catalyzed by cytochrome P450 enzymes. J. Adv. Res. https://doi.org/10.1016/j.jare.2023.10.012.
Zhu, W., Yuan, S.S., Li, J., et al., 2023. A first computational frame for recognizing heparin-binding protein. Diagnostics 13, 2465.
Zou, X., Ren, L., Cai, P., et al., 2023. Accurately identifying hemagglutinin using sequence information and machine learning methods. Front. Med. 10, 1281880.
Zulfiqar, H., Huang, Q.-L., Lv, H., et al., 2022. Deep-4mCGP: a deep learning approach to predict 4mC sites in Geobacter pickeringii by using correlation-based feature selection technique. Int. J. Mol. Sci. 23, 1251.