GRLDAMAN: A predictive framework for lncRNA-disease associations based on GraRep network embedding

Ping Zhang[1], Yongbin Zeng[2, *]
[1]School of Computer, Baoji University of Arts and Sciences, Baoji 721016, China
[2]School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China
*Corresponding email: zengyb2025@lzu.edu.cn

Abstract

Long non-coding RNAs (lncRNAs) play crucial roles in a variety of human diseases. As wet-lab experiments for identifying lncRNA-disease associations (LDAs) are often costly and time-consuming, computational prediction methods offer a valuable alternative. Such approaches can help elucidate the molecular mechanisms of diseases and contribute to the discovery of diagnostic biomarkers. Traditional LDA prediction methods primarily rely on capturing local features of lncRNAs or diseases. In contrast, we propose a novel computational framework, GRLDAMAN, for LDA prediction. To achieve robust predictive performance, GRLDAMAN utilizes the GraRep network embedding algorithm to learn informative and representative feature vectors. Compared to existing methods, GRLDAMAN demonstrates superior prediction performance, attaining an area under the curve (AUC) of 0.9201 under 5-fold cross-validation. This result indicates that the combination of GraRep and XGBoost yields stable and reliable predictions. Furthermore, case studies confirm that GRLDAMAN can holistically enhance the performance of LDA prediction.

Keywords

Features, Association, Prediction, GraRep, lncRNA, Disease

Introduction

Long non-coding RNAs (lncRNAs), defined as non-coding RNA molecules exceeding 200 nucleotides in length, were initially regarded as transcriptional noise [1]. However, accumulating evidence has now established that lncRNAs constitute a crucial category of regulatory ncRNAs, playing significant roles in fundamental cellular processes such as proliferation and differentiation [2,3]. Their dysregulation is frequently associated with complex human diseases, offering insights into pathogenic mechanisms and revealing potential avenues for novel therapies. For instance, the lncRNA X inactive specific transcript (XIST) has been implicated in the maintenance of human glioblastoma stem cells, while Wilms tumor antisense RNA (WT1-AS) promotes cell proliferation and invasion in gastric cancer [4,5]. Notably, the lncRNA HOX transcript antisense RNA (HOTAIR) is overexpressed by nearly 100-2000-fold in breast cancer metastases, and its expression levels correlate with metastasis and poor prognosis in various cancers, including lung, liver, gastric, and colorectal cancer [6,7]. Therefore, identifying associations between lncRNAs and diseases is essential not only for understanding molecular disease mechanisms but also for advancing diagnosis, treatment, and prevention. Nevertheless, the number of experimentally validated lncRNA-disease associations (LDAs) remains limited, as traditional wet-lab methods for their identification are often costly, time-consuming, and labor-intensive. Consequently, the well-known associations are far from comprehensive. Given these constraints, the development of effective and accurate computational models to predict potential LDA has become increasingly imperative.

Given the biological significance of LDAs, considerable effort in recent years has been directed toward developing computational models that are both stable and accurate. Consequently, numerous machine learning-based algorithms have been proposed. Generally, computational methods for predicting LDAs can be categorized into three main types. The first type comprises algorithms that leverage biological information about lncRNAs or diseases, operating under

the assumption that similar diseases are likely associated with functionally similar lncRNAs. For instance, Ha et al. predicted associations by analyzing the genomic neighborhood relationships between lncRNAs and genes, combined with known gene-disease associations [8]. In another study, Zeng et al. integrated multiple data sources including known associations, lncRNA functional similarity, expression profiles, disease semantic similarity, and Gaussian interaction profile kernels and applied the katz centrality (KATZ) measure for prediction [9]. Yu et al. developed a computational model named Naïve Bayesian Classifier for LncRNA–Disease Associations (NBCLDA), which employs a naïve Bayesian classifier to infer potential LDAs [10]. Furthermore, Ouyang et al. proposed a two-side sparse self-representation (TSSR) algorithm that uses estimated representations of lncRNAs and diseases for association prediction [11]. A common limitation of these methods, however, is their heavy dependence on known LDAs. Consequently, they are often not applicable to new diseases with no known associated lncRNAs, or to new lncRNAs without any known disease links, a scenario commonly referred to as the "cold start" problem.

A second major category of methods is based on data fusion through matrix factorization. Numerous approaches utilizing non-negative matrix factorization (NMF) have been proposed, analyzed, and implemented for various biological prediction tasks. For instance, Pei et al. fused diverse genomic and proteomic data sources and employed manifold-regularized NMF to predict protein-protein interactions [12]. In the context of predicting gene functions and pharmacological actions, Wodecki et al. developed DFMF, a model that factorizes multiple interrelated data matrices via penalized matrix tri-factorization [13]. For long non-coding RNA–disease association (LDA) prediction specifically, Li et al. proposed the similarity-based inductive matrix completion for lncRNA – disease associations (SIMCLDA) model, which applies inductive matrix completion to integrate feature vectors derived from the Gaussian interaction profile kernel of lncRNAs and the functional similarity of diseases [14]. The third category encompasses network-based methods, which contrast with the aforementioned approaches. These methods typically construct a heterogeneous network by integrating multiple relationship networks, such as known LDAs, disease similarity networks, and lncRNA similarity networks, and then implement propagation algorithms on this integrated structure. For example, Liu et al. directly utilized a random walk algorithm on a random walk protein complex network (RWPCN) to build a bipartite network for predicting and prioritizing disease-related genes [15]. Subsequently, Ding et al. proposed the tripartite graph for potential lncRNA-ddisease association identification (TPGLDA) model, inspired by tripartite graph theory. This model systematically predicts LDAs by integrating gene-disease and lncRNA-disease data within an lncRNA-disease-gene tripartite graph [16]. To improve predictive performance, Mori et al. incorporated biological sequence information into a disease-target-ncRNA tripartite network for association prediction [17]. Regarding bipartite networks, Ping et al. introduced a model that constructs a bipartite network following a power-law distribution to infer potential LDAs [18]. Furthermore, to identify these associations, Sumathipala et al. developed a multi-level complex network (a tripartite network) named LION. This model integrates protein-disease associations, protein-protein interactions, and lncRNA-protein interactions, applying a random walk with restart algorithm for network diffusion [19].

Constructing an effective predictive model fundamentally relies on the acquisition of reasonable feature representations. To enhance the prediction performance for LDAs and thereby facilitate the development of potential diagnostic biomarkers, numerous computational methods can be considered. However, it is important to note the following three limitations of existing approaches: (1) While methods based on biological information can be effective in certain scenarios, many conventional techniques for predicting LDAs primarily focus on the intrinsic attributes of lncRNAs or diseases, extracting only local features for representation. These methods often neglect to leverage global feature representations derived from network embedding algorithms. The features obtained in this manner are then directly used for classification tasks, which can result in suboptimal performance, such as lower AUC values or accuracy, due to a discrepancy between the captured patterns and the true predictive

relationships. (2) Most network-based models rely heavily on local topological information, which may limit their efficiency and generalizability when applied to lncRNAs or diseases with few or no known associations. (3) Existing methods often fail to adequately address the fact that interaction prediction within biological systems involves complex networks of multiple interacting biomolecules, a reality not fully captured by simpler, isolated models. Consequently, there is a clear need for a novel and more effective predictive framework.

In recent years, network embedding algorithms have been increasingly applied to association prediction tasks in domains such as social networks, computer networks, and biomedical networks, owing to their strong capability for feature engineering and their ability to deliver superior predictive performance [20,21]. In 2019, Guo et al. introduced a novel Molecular Associations Network (MAN) model, which integrates multiple molecular interactions, including those among miRNAs, lncRNAs, proteins, drugs, and diseases, enabling the prediction of various potential associations within a unified framework [22]. The MAN model is recognized as an effective approach for molecular association prediction due to its comprehensive network perspective and reliable performance. Building upon this advantage, we propose, analyze, and investigate a new global-view method for LDA prediction, following the design principles established by Guo et al. For clarity and ease of comparison, we refer to our model as GRLDAMAN (LDA prediction based on MAN via GraRep network embedding). In contrast to conventional feature extraction methods, GRLDAMAN leverages global molecular associations and preserves the intrinsic structures of node features and potential inter-node relationships during the learning process.

To evaluate the predictive performance of our newly proposed GRLDAMAN model based on the MAN framework, we conducted a series of experiments. These included a comparative analysis of different network representation learning algorithms paired with various classifiers, as well as an assessment of different node feature combinations. Using known LDA data obtained from the lncRNASNP2 and lncRNADisease databases, we compared two network embedding algorithms (GraRep and LINE) and three classifiers (XGBoost, SVM, and Naïve Bayes). The results of these cross-comparison experiments substantiate the efficacy of GRLDAMAN for LDA prediction.

Specifically, GRLDAMAN achieved reliable mean AUC values of 0.9201 and 0.8776 under 5-fold cross-validation (5-CV) when using the XGBoost classifier with GraRep and LINE embeddings, respectively, demonstrating its excellent predictivecapability.

Furthermore, experiments on different node feature combinations indicated that the "Behavior" feature set yields strong performance in predicting potential LDAs. Finally, in a case study application, GRLDAMAN was successfully employed to predict lncRNAs associated with breast cancer and colon cancer. Collectively, the construction of GRLDAMAN provides a new systematic perspective for identifying disease-related lncRNAs, and the numerical results confirm its effectiveness and superiority.

Materials

GRLDAMAN

To holistically and systematically construct the GRLDAMAN model, associations among various biological molecules, including transcripts, diseases, and drugs, were integrated from multiple databases. As the prediction of unknown LDAs is the primary objective, nine types of associations, including known lncRNA-disease associations, were incorporated into the GRLDAMAN model. Based on the MAN framework, we constructed a network comprising five types of biological entities: lncRNAs, diseases, miRNAs, proteins, and drugs. Nine relational associations among these entities were collected from publicly available databases. Following identifier unification, redundancy removal, and the elimination of irrelevant entries, the processed data were integrated to construct the GRLDAMAN network. The details of the final GRLDAMAN datasets are summarized in Table 1 and illustrated in Figures 1 and 2.

Table 1. The database of nine kinds of associations in GRLDAMAN.

Relationship Type	Database	URL
lncRNA-disease	LncRNADisease [23],	http://cmbi.bjmu.edu.cn/lncrnadisease/ http://bioinfo.life.hust.edu.cn/lncRNASNP2
miRNA-lncRNA	LncRNASNP2 [24]	http://bioinfo.life.hust.edu.cn/lncRNASNP2
lncRNA-protein	LncRNASNP2 [24]	http://123.59.132.21/lncrna2target/
miRNA-disease protein-disease drug-disease	LncRNA2Target [25]	http://www.cuilab.cn/hmdd
	HMDD [26]	http://www.disgenet.org/
	DisGeNET [27]	http://ctdbase.org/
miRNA-protein	CTD [28]	http://miRTarBase.mbc.nctu.edu.tw/
drug-protein protein-protein	miRTarBase [29]	http://www.drugbank.ca/
	DrugBank [30]	http://string-db.org/
miRNA-drug	STRING [31]	http://www.pharmaco-mir.org/
	PharmacomiR	http://bioinfo.hrbmu.edu.cn/SM2miR/
circRNA-disease	SM2miR	http://cgga.org.cn:9091/circRNADisease/
piRNA-disease	circRNADisease	http://www.regulatoryrna.org/database/piRNA/
	piRBase	http://www.piwirna2disease.org/index.php

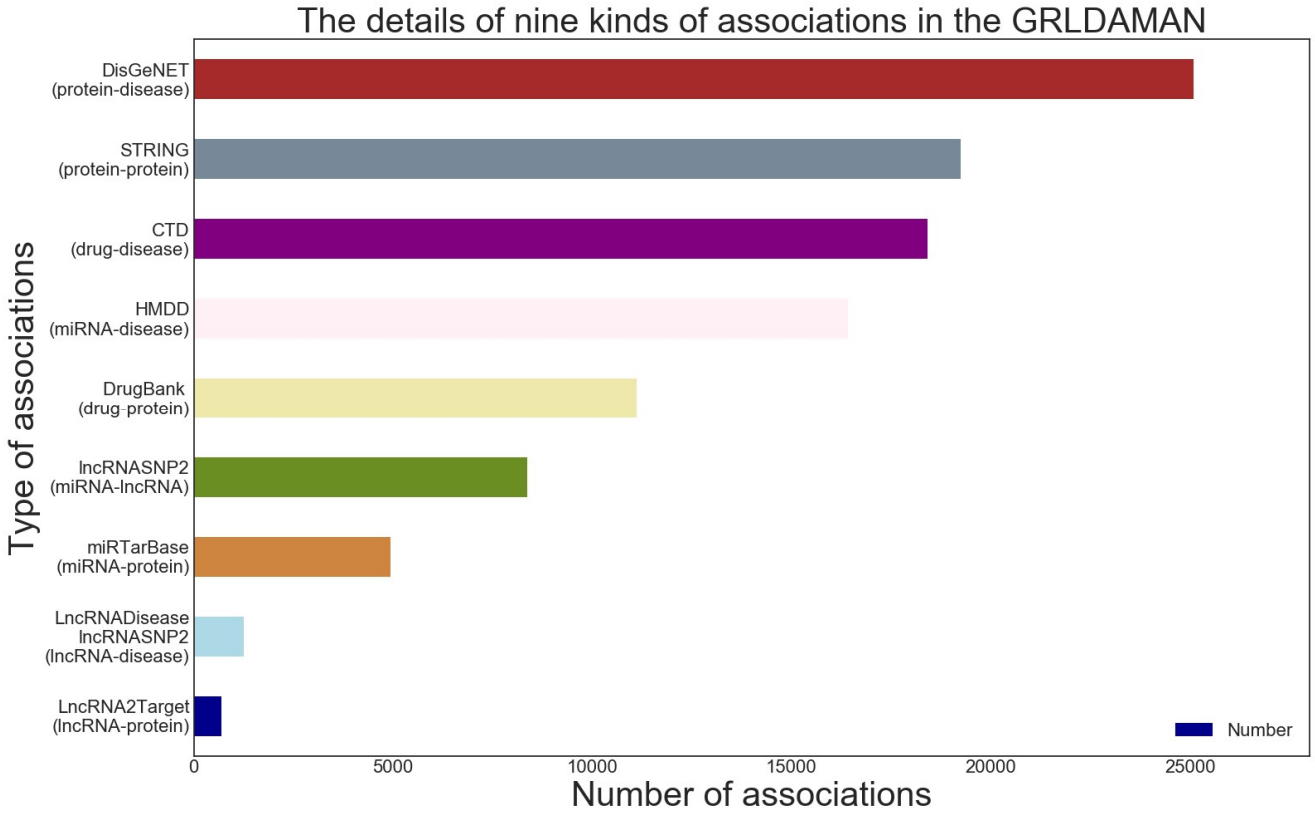

Figure 1. The details of nine kinds of associations in GRLDAMAN.

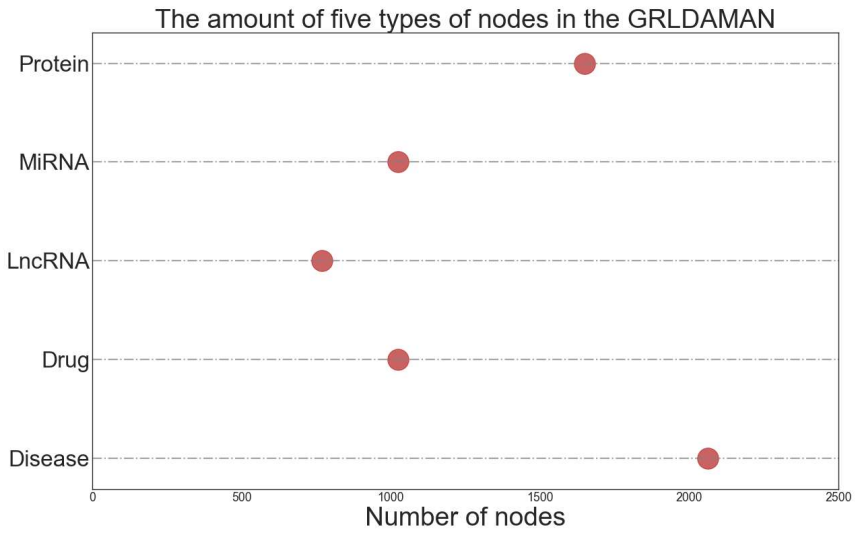

Figure 2. The amount of five types of nodes in GRLDAMAN.

Known lncRNA-disease associations

The lncRNASNP2 database provides comprehensive ratiresources of single nucleotide polymorphisms in human or mouse lncRNAs (e.g., SNP effects on lncRNA structure, SNPs in lncRNAs). As a resource dataset, LncRNADisease database includes experimentally supported lncRNA-disease association data, which provides a confidence score for each ncRNA-disease association and curates lncRNA interactions in various levels. Hence, known lncRNA-disease associations data were gathered from the lncRNASNP2 and LncRNADisease databases to estimate the prediction performance of our newly proposed GRLDAMAN. By processing data that describes the same lncRNA-disease relationships based on evidence from different experiments, we obtained 835 independent lncRNA-disease association pairs which include 252 different lncRNAs and 143 different diseases. As a train dataset, known lncRNA-disease associations data are used to fit together with other known molecular associations.

LncRNA, protein sequence and drug molecular fingerprint

In GRLDAMANMAN, the sequences of lncRNA, miRNA, and protein are encoded vectors in which a 64 (4×4×4) dimensional vector is used to encode ncRNA sequences (following the method of Xu et al. [32]). Moreover, each attribute in nodes represents the normalized frequency of the corresponding 3-mer appearing in the RNA sequence and protein sequence. For Drug Molecular Fingerprint, the SMILES of drugs a re transformed into the corresponding Morgan Fingerpri nt using the application programming interface (API) R DKit (OpenSource Cheminformatics Software), a onal discovery toolkit.

Directed acyclic graph (DAG)

In GRLDAMAN, each disease including all related annotation terms which can be obtained from MeSH can be represented by a DAG (Directed Acyclic Graph) (http://www.ncbi.nlm.nih.gov/). Generally speaking, DAG can be expressed as DAG = (D, N(D), E(D)), for a given disease D, N(D) denotes D itself together with all its ancestor nodes, while E(D) denotes all relationships connecting between nodes in the DAG(D). The Dd(t) of a disease t in a DAG to the semantics of disease D is defined as follows:

$$D_d(D) = 1$$
$$D_d(t) = max\{0.5 * D_d(t') | t' \epsilon \ children \ of \ t\} if t \neq d \quad (1)$$

The semantic similarity score between two diseases i and j can then be calculated by:

$$S(i,j) = \frac{\sum_{t \in T(i) \cap T(j)} \left(D_i(t) + D_j(t)\right)}{\sum_{t \in T(i)} D_i(t) + \sum_{t \in T(j)} D_j(t)} \quad (2)$$

Methods

GRLDAMAN framework

According to available datasets, the complex association network of biomolecules for Molecular Association Network (MAN) (as shown in Figure 3) consists of two parts: network embedding and classification.

By network embedding, the edges between any two nodes within the MAN can learn structural semantic

representation for LDAs. In order to improve the prediction performance of LDAs, the XGBoost classifier with ensemble learning is utilized to construct GRLDAMAN framework.

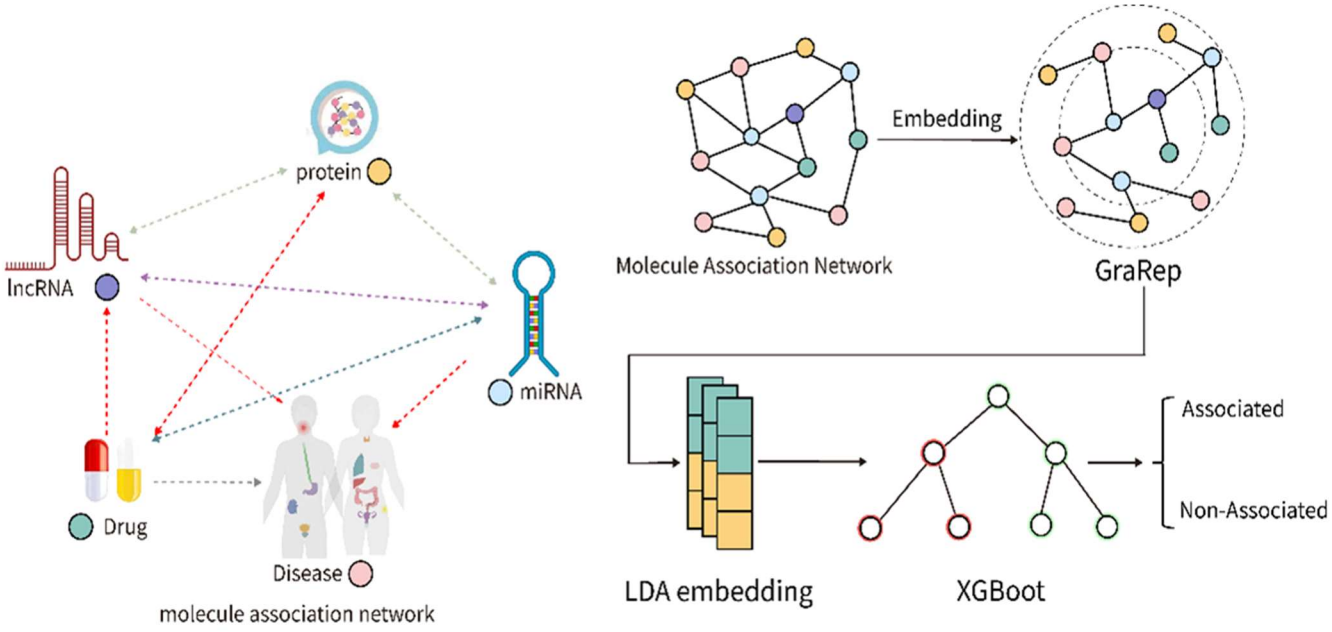

Figure 3. GRLDAMAN framework.

GraRep

GraRep proposes a framework for large-scale graph representations, which is based on SVD to obtain the dense low-dimensional real value representation of vertices. GraRep is a model to learn latent representations of vertices on graphs, which can capture global structural information associated with the graph. It also defines a more accurate loss function that allows nonlinear combinations of different local relational information to be integrated. Therefore, it is very pressing to find a method that is able to effectively represent the vertices of global structural information and local adjacent relations. Based on the above problem, to obtain the optimal network representation structure and optimal adjacent relations, we adopt GraRep to perform network embedding in GRLDAMAN. By following Following the GraRep method proposed by Ouyang et al., to implement and obtain the vertex representations, we first define the following k-step probability transition matrix.

$$F^k = \underbrace{F \cdots F}_{k} \tag{3}$$

For given $F_{i,j}^k$ (i.e., the transition probability from vertex i to vertex j based on k step), we can obtain:

$$p_k(c|v) = F_{v,c}^k \tag{4}$$

Wherein $F_{v,c}^k$ denotes the element from $v-th$ row and $c-th$ column of the matrix F^K. For a given k, all of k-step paths can be sampled from the given G, starting with v and ending with c. Note that, v denotes current vertex, while c denotes context vertex. we thus have k-step paths marked as (v,c). $pk(V)$ denotes the probability distribution of $V \in G$. The goals are divided into two parts, namely, to maximize the probability of $(v,c) \in G$ and minimize the probability of $(v,c) \notin G$ $(i.e.,(v,c') \in G, c' \in V)$.

Secondly, we use noise contrastive estimation (i.e., NCE) to define k-step object function:

$$O_k = \sum_{v \in V} O_k(v) \tag{5}$$

Wherein, for any $v \in V$. to compute the object function value and add them, then we can have:

$$O_k(v) = (\sum_{c \in V} p_k(c|v) \log\sigma(\vec{v} \cdot \vec{c})) + \lambda \mathbb{E}_{c' \sim p_k(V)}[\log\sigma(-\vec{v} \cdot \vec{c}')] \tag{6}$$

Wherein, we can regard $\sigma(\vec{v} \cdot \vec{c})$ as the probability of $(v,c) \in G$, while regard $\sigma(-\vec{v} \cdot \vec{c}')$ as the probability of $(v,c') \in G$. Note that, λ dnotes the amounts of negative samples $(i.e.,c')$, thus we have:

$\mathbb{E}_{c\prime\sim p_k(V)}[\log\sigma(-\vec{v}\cdot\vec{c}')] = p_k(c)\cdot\log\sigma(-\vec{v}\cdot\vec{c}') + \sum_{c\prime\in V\setminus\{c\}} p_k(c')\cdot\log\sigma(-\vec{v}\cdot\vec{c}')$ (7)

Where, $\mathbb{E}_{c\prime\sim p_k(V)}[\cdot]$ denotes the expectation withc′followed the distribution $pk\,(V)$ namely,

$O_k\,(v,c) = p_k(c|v)\cdot\log\sigma(\vec{v}\cdot\vec{c}))+\lambda\cdot p_k(c)\cdot\log\sigma(-\vec{v}\cdot\vec{c})$ (8)

Since the distribution $pk\,(c)$ can be computed as follows:

$p_k(c) = \sum_{v\prime} q(v')p_k(c|v') = \frac{1}{N}\sum_{v\prime} F^k_{v\prime,c}$ (9)

Therefore, we obtain the following loss function:

$Y^k_{i,j} = W^k_i\cdot C^k_j = \log\left(\frac{F^k_{i,j}}{\sum_t F^k_{t,j}}\right) - \log(\beta)$ (10)

To obtain a suitable decomposition for Y^k and reduce noise, thus replace all negative entries in Y^k with 0, we can obtain:

$X^k_{i,j}$= max $(Y^k_{i,j}, 0)$ (11)

then by SVD method, we have:

$X^k = U^k\Sigma^k(V^K)^T$ (12)

Owing to a d-dimensional matrix for network representation to obtain, thus further factorizes it and finally we can obtain:

$X^k \approx X^k_d = U^k_d\Sigma^k_d(V^k_d)^T$ (13)

$X^k \approx X^k_d = W^k C^k$ (14)

$W^k = U^k_d(\Sigma^k_d)^{\frac{1}{2}}$ (15)

$C^k = (\Sigma^k_d)^{\frac{1}{2}}V^{k^T}_d$ (16)

Where, W^k denotes network representations of current vertices as its column vectors, C^k while denotes network representations of context vertices as its column vectors.

Hence, the W^k is returned from the algorithm as the low d-dimensional representations of the vertices which capture k-step global structural information in the graph.

Results and discussion

Comparison with LINE based on different classifier

In this subsection, to assess the prediction performance of GRLDAMAN between network embedding algorithms and classifiers, we compare GraRep with LINE under 5-CV via different classifiers. As we know, choosing an appropriate classifier is also important in model practical applications. Therefore, XGBoost, SVM and Naive Bayes have been chosen to implement experiments for the prediction performance of our model. By validating different network embedding algorithms, AUCs based on 3 classifiers are obtained, and simulation results are shown in Table 2 and Figure 4.

Specifically, as seen from Table 2, the three classifiers are all effective on classification with high AUC values. By looking into the details of these results, we can observe that, as compared with LINE, the GraRep achieves higher AUC values via XGBoost. This substantiates that the GraRep performs better than the LINE under 5- CV. Moreover, we can also find that all classifier parameters are default values and only behavior features for nodes are appended to training. Obviously, each classification effect of three classifiers based on GraRep is better than LINE. This, to some extent, verifies the superiority of the GRLDAMAN on LDAs prediction. Besides, for Naive Bayes classifier (AUC<85), the lower classification effect may be caused by the classifier default parameters. In addition, since the assumption of sample attributes independence is used in Naive Bayes, it does not work well if the sample features are correlated. Hence, when the properties of the sample are independent of each other, then Naive Bayes can get better results. In this experiment, there are samples where the features are not independent, and thus cross-joining together affects the final Naive Bayes classification effect.

Table 2. AUCs of GraRep and LINE are based on three classifiers with default parameters.

Network Embedding	Classifier (AUCs) (%)		
	XGBoost	SVM	Naive Bayes
GraRep	92.01	85.15	83.37
LINE	87.76	85.09	81.54

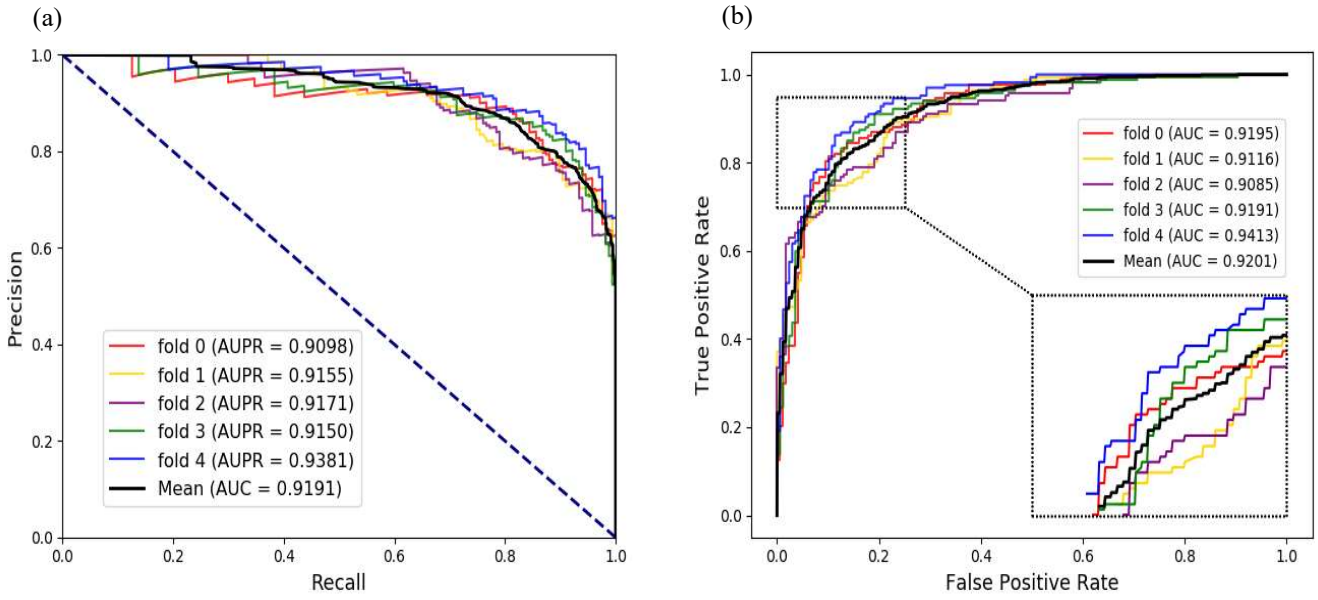

Figure 4. AUPRs and AUCs use GraRep network representation via XGBoost classifier with default parameters. (a) 5-Fold Cross-Validation PR Curves (b)5-Fold Cross-Validation ROC Curves.

Comparison of different node features

Among nodes in a GRLDAMAN network, behavior information is a critical association relationship for LDAs prediction. In other words, the main goal that we construct GRLDAMAN is trying to obtain the relation features, namely, behavior features. In the GRLDAMAN, each node can be represented by its intrinsic attributes and its relationship with other nodes (i.e., behavior features). For behavior information (64-dimensional vectors), the relationship of each node with others could be abstracted by GraRep. For attribute information (64-dimensional vectors), the attributes of the node itself can be the sequence of ncRNA, protein, the semantics of the disease, and the molecular fingerprint of the drug. Here, in comparison with the predictive performance of GRLDAMAN with different nodes feature combination. Particularly, we divid into three groups to respectively validate the different performances with "Attribute", "Behavior" and "Attribute + Behavior".

Specifically, as shown below, Figure 5 plots the ROC curves of the three combinations results and reports their AUROC values of five-fold cross validation. Figure 5 (a) shows results of the ''Behavior'' that 5-CV with pure behavior information as the characteristics of the node. Figure 5 (b) shows results of ''Attribute'' that 5-CV with pure attribute information as the characteristics of the node. Figure 5 (c) shows the results of ''Attribute + Behavior'' that is based on the feature combined attribute information with behavior. As shown in results, we find that different feature combinations lead to different performances of GRLDAMAN. There exists clear difference between feature combinations. We also see that the best performance is achieved by ''Behavior'' (AUC=0.9201). This implies that attributing information has a small impact on predictive performance. In contrast, the ''Attribute + Behavior'' has a lower performance, which only performs AUC of 0.8792 and the AUC of 0.8244 is obtained by the ''Attribute''.

Besides, it is worth pointing out that we chose GraRep to globally represent the behavior feature of nodes in the entire network and the flow of information directly or latently between other nodes, thus improving the performance. In addition, from these results, we can confirm that the ''Behavior'' has its own superiority in GRLDAMAN.

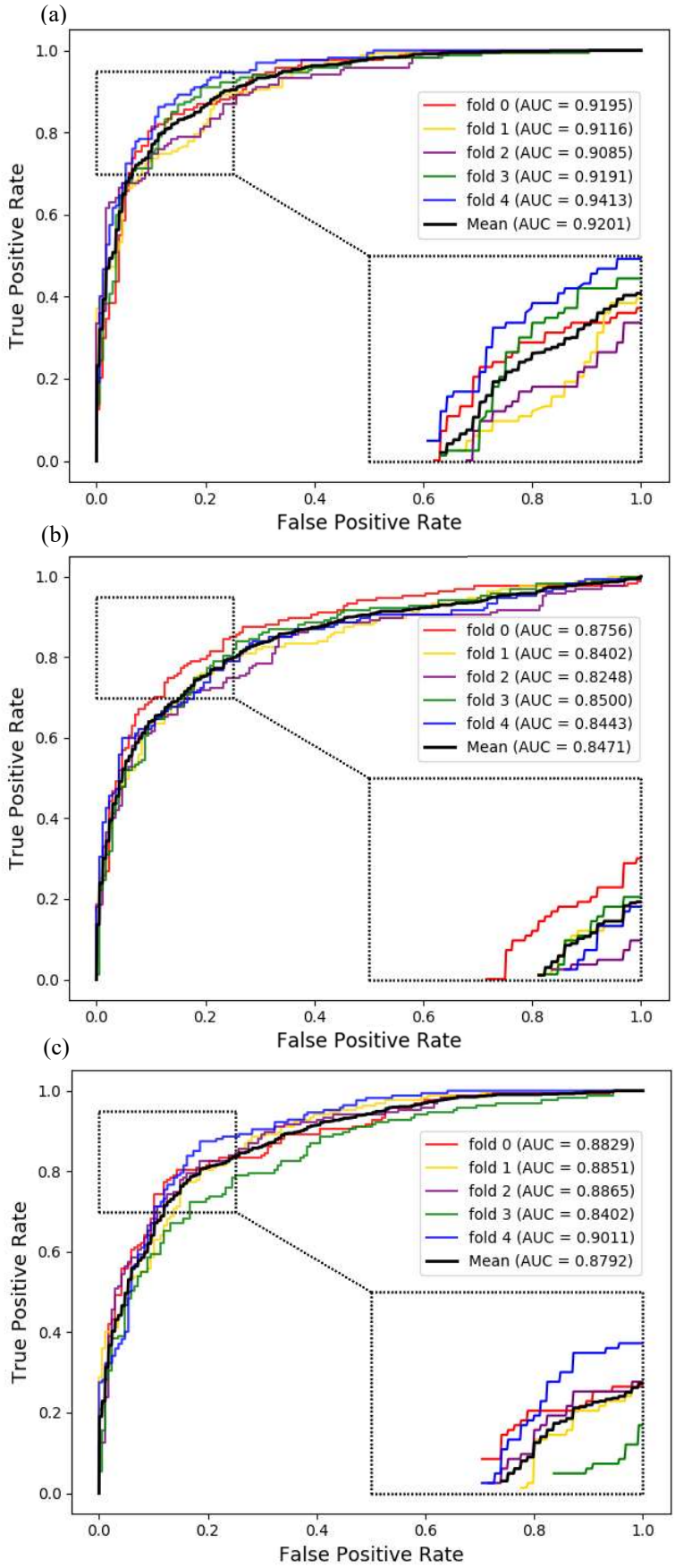

Figure 5. AUC for different feature combinations. (a) Behavior-only 5-fold CV ROC Curves (b) Attribute-only 5-fold CV ROC Curves (c) Attribute+Behavior 5-fold CV ROC Curves.

Case study

In addition to comparison experiments, the case study of GRLDAMAN based on three diseases is implemented to evaluate the prediction performance for a watched disease without any known association. Especially by removing the original lncRNA associations information of the investigated disease (i.e. Leukemia, colorectal neoplasm and prostatic neoplasm), we validated the top 5 predicted disease-related lncRNAs based on long non-coding RNA to Cancer database (Lnc2cancer), MNDR and LncRNADisease2.0 databases. Here, we select the top 5 associated lncRNAs which get the highest predicted ranks for each watched disease. Predictive results are supported by relevant databases, and the detaile of three diseases can be interpreted as follows.

Leukemia is one of the malignant diseases of hematopoietic stem cells worldwide in humans by inhibiting the normal hematopoietic function. It is reported that the five-year survival rate of Leukemia is lower for women and men has the high recurrence rate [33-35].

Therefore, it's obvious that the early detection of Leukemia is vital to cancer treatment. Among the top 5 Leukemia-related candidates ranked by GRLDAMAN (see table 3 and table4), all five potential lncRNAs are verified to be associated with Leukemia by related literature and databases. For instance, maternally expressed 3 (MEG3), H19 and growth arrest-specific 5 (GAS5) have been found to be involved in the formation of Leukemia successively.

A recent experimental result shows that MEG3 may play a role and inhibit cell proliferation and metastasis in Chronic Myeloid Leukemia [36]. Furthermore, recent research reports that GAS5 polymorphism predicts a poor prognosis of acute myeloid leukemia in Chinese patients by affecting hematopoietic reconstitution [37].

Table 3. predicted lncRNAs associated Leukemia and the corresponding evidence based on Lnc2Cancer database.

Disease	LncRNAs	Evidence	PMID
Leukemia	NONHSAT 017462.2	Lnc2Cancer	15645136, 29703210
Leukemia	NONHSAT 017460.2	Lnc2Cancer	24685695, 28765931
Leukemia	NONHSAT 007662.2	Lnc2Cancer	27951730
Leukemia	NONHSAT 137541.2	Lnc2Cancer	7981627
Leukemia	NONHSAT 022132.2	Lnc2Cancer	28713913

Colorectal neoplasm is one of several tumors with a high morbidity rate worldwide especially in men and women. According to the statistics, in the United States, approximately 160,000 new cases of colorectal neoplasms are diagnosed each year [38,39].

The top 5 colorectal neoplasms-related candidates ranked by GRLDAMAN (see Table 5) and four potential lncRNAs are verified to be associated with colorectal neoplasm by Lnc2Cancer database. Particularly, the GAS5 and XIST are confirmed by related literatures. For example, the endogenous suppressive effect of XIST promoted the proliferation and invasion of colon cancer cells by Wnt/β-catenin signaling activation [40]. The GAS5 also inhibits angiogenesis and metastasis of colorectal cancer through the Wnt/β-catenin signaling pathway [41].

Table 4. Predicted lncRNAs associated colorectal neoplasm and the corresponding evidence based on Lnc2Cancer database.

Disease	LncRNAs	Evidence	PMID
Colorectal neoplasm	NONHSAT130416.2	Lnc2Cancer	25058480
Colorectal neoplasm	NONHSAT041865.2	Unconfirmed	N/A
Colorectal neoplasm	NONHSAT137542.2	Lnc2Cancer	26820130
Colorectal neoplasm	NONHSAT137558.2	Lnc2Cancer	24809982
Colorectal neoplasm	NONHSAT137559.2	Lnc2Cancer	27314206

Prostate neoplasm is one of several tumors with a high morbidity rate worldwide, especially in white and African American men.

According to the statistics, there are about 20% African-American men and 17% white men being diagnosed with prostate cancer in their lifetime [42]. As shown in Table 5 and 6, the top 5 prostate neoplasms-related candidates are ranked by GRLDAMAN. Here, four potential lncRNAs are verified to be associated with prostate neoplasm by Lnc2Cancer and MNDR database.

Table 5. Predicted lncRNAs associated prostate neoplasm and the corresponding evidence based on Lnc2Cancer database.

Disease	LncRNAs	Evidence	PMID
prostate neoplasm	NONHSAT017462.2	Lnc2Cancer	25513185
prostate neoplasm	NONHSAT039742.2	Lnc2Cancer	20541999
prostate neoplasm	NONHSAT028514.2	Lnc2Cancer	27176634
prostate neoplasm	NONHSAT079510.2	Unconfirmed	N/A
prostate neoplasm	NONHSAT022125.2	Lnc2Cancer	23728290

Table 6. Predicted lncRNAs associated prostate neoplasm and the corresponding evidence based on MNDR database.

Disease	LncRNAs	Evidence
prostate neoplasm	NONHSAT017462.2	Lnc2Cancer/LncRNADisease MNDR
prostate neoplasm	NONHSAT039742.2	Lnc2Cancer/LncRNADisease MNDR
prostate neoplasm	NONHSAT028514.2	Lnc2Cancer MNDR
prostate neoplasm	NONHSAT079510.2	Unconfirmed
prostate neoplasm	NONHSAT022125.2	Lnc2Cancer/LncRNADisease MNDR

Conclusion

In this paper, owing to the important role of lncRNA in a variety of biological processes and human diseases, a novel model called GRLDAMAN is constructed to identify potential novel associations between lncRNAs and diseases. In GRLDAMAN, we proposed a new framework based on MAN to detect lncRNA-diseases associations between arbitrary nine molecules.

To evaluate the predictive performance of our approach, the comparative experiment of different classifiers based on different network representation as well as the comparative experiment of different nodes feature are implemented. In addition, the Case Study based on Leukemia, colorectal neoplasm and prostatic neoplasm is

also implemented to evaluate the prediction performance of GRLDAMAN. The superior prediction performance obtained by GRLDAMAN could be due to two main reasons: On one hand, the construction of molecular associations network in human cells, offer a new systematic view on understanding complex life activities and diseases. On the other hand, our approach integrates associations information of lncRNA, miRNA, diseases, drug, protein and their associated biomolecules with lncRNA and diseases by using a GraRep network embedding algorithm. Hence, GRLDAMAN could fully make use of the integrated associated data, and the "Behavior" can thus further improve prediction performance and has its own superiority, compared to the other two kinds feature combination.

Summarizing the above analysis, in this paper, a GRLDAMAN model based on MAN is presented, developed and investigated for the association prediction of lncRNA-disease. Validation results demonstrate that such a GRLDAMAN model can globally improve the performance. Besides, the case study also demonstrates the superior performance of GRLDAMAN model on potential lncRNAs related to three watched diseases. However, it is worth mentioning that only 835 known LDAs have been adopted by GRLDAMAN. The prediction accuracy of GRLDAMAN will improve higher if more known LDAs are added.

Funding

This work was not supported by any funds.

Author Contributions

Ping Zhang: conceptualization, methodology, software, investigation, resources, data curation, validation, visualization and formal analysis, funding acquisition and supervision. Yongbin Zeng: writing-review and editing, project administration.

Acknowledgments

The authors would like to show sincere thanks to those techniques who have contributed to this research.

Conflicts of Interests

The authors declare no conflict of interest.

References

[1] Rosandić, M., Paar, V. (2022) Standard genetic code vs. supersymmetry genetic code–alphabetical table vs. physicochemical table. *BioSystems*, 218, 104695.

[2] Mahat, D. B., Tippens, N. D., Martin-Rufino, J. D., Waterton, S. K., Fu, J., Blatt, S. E., Sharp, P. A. (2024) Single-cell nascent RNA sequencing unveils coordinated global transcription. Nature, 631(8019), 216-223.

[3] Zhang, S., Duan, J., Du, Y., Xie, J., Zhang, H., Li, C., Zhang, W. (2021) Long non-coding RNA signatures associated with liver aging in senescence-accelerated mouse prone 8 model. *Frontiers in Cell and Developmental Biology*, 9, 698442.

[4] Moftakhar, A., Khoshnam, S. E., Farzaneh, M., Abouali Gale Dari, M. (2024) Functional Roles of Long Non-coding RNAs on Stem Cell-related Pathways in Glioblastoma. *Current Signal Transduction Therapy*, 19(3), 40-52.

[5] Zhang, X., Jin, M., Yao, X., Liu, J., Yang, Y., Huang, J., Zhang, B. (2024) Upregulation of LncRNA WT1-AS inhibits tumor growth and promotes autophagy in gastric cancer via suppression of PI3K/Akt/mTOR pathway. *Current Molecular Pharmacology*, 17(1), e18761429318398.

[6] Liu, S. J., Dang, H. X., Lim, D. A., Feng, F. Y., Maher, C. A. (2021) Long noncoding RNAs in cancer metastasis. *Nature Reviews Cancer*, 21(7), 446-460.

[7] Bure, I. V., Nemtsova, M. V., Kuznetsova, E. B. (2022) Histone modifications and non-coding RNAs: mutual epigenetic regulation and role in pathogenesis. *International Journal of Molecular Sciences*, 23(10), 5801.

[8] Ha, J. (2024) LncRNA expression profile-based matrix factorization for predicting lncRNA-disease association. *IEEE Access*, 12, 70297-70304.

[9] Zeng, M., Lu, C., Fei, Z., Wu, F. X., Li, Y., Wang,

J., Li, M. (2020) DMFLDA: a deep learning framework for predicting lncRNA–disease associations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(6), 2353-2363.

[10] Yu, J., Ping, P., Wang, L., Kuang, L., Li, X., Wu, Z. (2018) A novel probability model for lncRNA–disease association prediction based on the naïve bayesian classifier. *Genes*, 9(7), 345.

[11] Ouyang, L., Huang, J., Zhang, X. F., Li, Y. R., Sun, Y., He, S., Zhu, Z. (2019) LncRNA-disease association prediction using two-side sparse self-representation. *Frontiers in Genetics*, 10, 476.

[12] Pei, F., Shi, Q., Zhang, H., Bahar, I. (2021) Predicting protein–protein interactions using symmetric logistic matrix factorization. *Journal of Chemical Information and Modeling*, 61(4), 1670-1682.

[13] Wodecki, J., Michalak, A., Zimroz, R., Wyłomańska, A. (2020) Separation of multiple local-damage-related components from vibration data using Nonnegative Matrix Factorization and multichannel data fusion. *Mechanical Systems and Signal Processing*, 145, 106954.

[14] Li, M., Liu, M., Bin, Y., Xia, J. (2020) Prediction of circRNA-disease associations based on inductive matrix completion. *BMC Medical Genomics*, 13(Suppl 5), 42.

[15] Liu, L., Zhu, S. (2021) Computational methods for prediction of human protein-phenotype associations: a review. *Phenomics*, 1(4), 171-185.

[16] Ding, L., Wang, M., Sun, D., Li, A. (2018) TPGLDA: Novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph. *Scientific Reports*, 8(1), 1065.

[17] Mori, T., Ngouv, H., Hayashida, M., Akutsu, T., Nacher, J. C. (2018) ncRNA-disease association prediction based on sequence information and tripartite network. *BMC Systems Biology*, 12(Suppl 1), 37.

[18] Ping, P., Wang, L., Kuang, L., Ye, S., Iqbal, M. F. B., Pei, T. (2018) A novel method for lncRNA-disease association prediction based on an lncRNA-disease association network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(2), 688-693.

[19] Sumathipala, M., Maiorino, E., Weiss, S. T., Sharma, A. (2019) Network diffusion approach to predict lncRNA disease associations using multi-type biological networks: LION. *Frontiers in Physiology*, 10, 888.

[20] Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., Sun, H. (2020) Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, 36(4), 1241-1251.

[21] Ouyang, M., Zhang, Y., Xia, X., Xu, X. (2023) Grarep++: flexible learning graph representations with weighted global structural information. *IEEE Access*, 11, 98217-98229.

[22] Guo, Z. H., Yi, H. C., You, Z. H. (2019) Construction and comprehensive analysis of a molecular association network via lncRNA–miRNA–disease–drug–protein graph. *Cells*, 8(8), 866.

[23] Lin, X., Lu, Y., Zhang, C., Cui, Q., Tang, Y. D., Ji, X., Cui, C. (2024) LncRNADisease v3. 0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Research*, 52(D1), D1365-D1369.

[24] Miao, Y. R., Liu, W., Zhang, Q., Guo, A. Y. (2018) lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Research*, 46(D1), D276-D280.

[25] Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Jiang, Q. (2019) LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Research*, 47(D1), D140-D144.

[26] Cui, C., Zhong, B., Fan, R., Cui, Q. (2024) HMDD v4. 0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Research*, 52(D1), D1327-D1332.

[27] Hu, Y., Guo, X., Yun, Y., Lu, L., Huang, X., Jia, S. (2025) DisGeNet: a disease-centric interaction database among diseases and various associated genes. *Database*, 2025, baae122.

[28] Wiegers, T. C., Davis, A. P., Wiegers, J., Sciaky, D.,

Barkalow, F., Wyatt, B., Mattingly, C. J. (2025) Integrating AI-powered text mining from PubTator into the manual curation workflow at the Comparative Toxicogenomics Database. Database, 2025, baaf013.

[29] Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., Huang, H. D. (2018) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1), D296-D302.

[30] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Wilson, M. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1), D1074-D1082.

[31] Fernando, P. C., Mabee, P. M., Zeng, E. (2020) Integration of anatomy ontology data with protein–protein interaction networks improves the candidate gene prediction accuracy for anatomical entities. *BMC Bioinformatics*, 21(1), 442.

[32] Xu, D., Xu, H., Zhang, Y., Chen, W., Gao, R. (2020) Protein-protein interactions prediction based on graph energy and protein sequence information. *Molecules*, 25(8), 1841.

[33] Kantarjian, H., Kadia, T., DiNardo, C., Daver, N., Borthakur, G., Jabbour, E., Ravandi, F. (2021) Acute myeloid leukemia: current progress and future directions. *Blood Cancer Journal*, 11(2), 41.

[34] Marrapodi, M. M., Mascolo, A., Di Mauro, G., Mondillo, G., Pota, E., Rossi, F. (2022) The safety of blinatumomab in pediatric patients with acute lymphoblastic leukemia: a systematic review and meta-analysis. *Frontiers in Pediatrics*, 10, 929122.

[35] Chen, E. C., Garcia, J. S. (2024) Immunotherapy for acute myeloid leukemia: current trends, challenges, and strategies. *Acta Haematologica*, 147(2), 198-218.

[36] Li, J., Zi, Y., Wang, W., Li, Y. (2018) Long noncoding RNA MEG3 inhibits cell proliferation and metastasis in chronic myeloid leukemia via targeting miR-184. *Oncology Research*, 26(2), 297.

[37] Ketab, F. N. G., Gharesouran, J., Ghafouri-Fard, S., Dastar, S., Mazraeh, S. A., Hosseinzadeh, H., Rezazadeh, M. (2020) Dual biomarkers long non-coding RNA GAS5 and its target, NR3C1, contribute to acute myeloid leukemia. *Experimental and Molecular pathology*, 114, 104399.

[38] Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., Bray, F. (2019) Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. *International Journal of Cancer*, 144(8), 1941-1953.

[39] Zavala, V. A., Bracci, P. M., Carethers, J. M., Carvajal-Carmona, L., Coggins, N. B., Cruz-Correa, M. R., Fejerman, L. (2021) Cancer health disparities in racial/ethnic minorities in the United States. British journal of cancer, 124(2), 315-332.

[40] Sun, N., Zhang, G., Liu, Y. (2018) Long non-coding RNA XIST sponges miR-34a to promote colon cancer progression via Wnt/β-catenin signaling pathway. *Gene*, 665, 141-148.

[41] Song, J., Shu, H., Zhang, L., Xiong, J. (2019) Long noncoding RNA GAS5 inhibits angiogenesis and metastasis of colorectal cancer through the Wnt/β-catenin signaling pathway. *Journal of Cellular Biochemistry*, 120(5), 6937-6951.

[42] Hsieh, T. C., Wu, J. M. (2020) Resveratrol suppresses prostate cancer epithelial cell scatter/invasion by targeting inhibition of hepatocyte growth factor (HGF) secretion by prostate stromal cells and upregulation of E-cadherin by prostate cancer epithelial cells. *International Journal of Molecular Sciences*, 21(5), 1760.