KGETCDA: an efficient representation learning framework based on knowledge graph encoder from transformer for predicting circRNA-disease associations

Jinyang Wu 1, Zhiwei Ning 1, Yidong Ding 1, Ying Wang 1, Qinke Peng 1, and Laiyi Fu 1,2,3,*

1School of Automation Science and Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi 710049, China, 2Research Institute of Xi’an Jiaotong University, Hangzhou, Zhejiang 311200, China and 3Sichuan Digital Economy Industry Development Research Institute, Chengdu, Sichuan 610036, China

ABSTRACT

Recent studies have demonstrated the significant role that circRNA plays in the progression of human diseases. Identifying circRNA-disease associations (CDA) in an efficient manner can offer crucial insights into disease diagnosis. While traditional biological experiments can be time-consuming and labor-intensive, computational methods have emerged as a viable alternative in recent years. However, these methods are often limited by data sparsity and their inability to explore high-order information. In this paper, we introduce a novel method named Knowledge Graph Encoder from Transformer for predicting CDA (KGETCDA). Specifically, KGETCDA first integrates more than 10 databases to construct a large heterogeneous non-coding RNA dataset, which contains multiple relationships between circRNA, miRNA, IncRNA and disease. Then, a biological knowledge graph is created based on this dataset and Transformer-based knowledge representation learning and attentive propagation layers are applied to obtain high-quality embeddings with accurately captured high-order interaction information. Finally, multilayer perceptron is utilized to predict the matching scores of CDA based on their embeddings. Our empirical results demonstrate that KGETCDA significantly outperforms other state-of-the-art models. To enhance user experience, we have developed an interactive web-based platform named HNRBase that allows users to visualize, download data and make predictions using KGETCDA with ease.

INTRODUCTION

Circular RNA (circRNA) is a unique type of endogenous non-coding RNA (ncRNA) molecules typically characterized by a covalently closed-loop structure without 3’ and 5’ polyadenylated tails (1). With the rapid development of high-throughput sequencing technologies, increasing studies have demonstrated that circRNA is extensively involved in numerous physiological and pathological processes in humans, including regulating gene expression (2), acting as miRNA sponges (3) and mediating cancer metastasis (4). Thus, exploring the associations between circRNAs and diseases can offer crucial insights into disease pathology at molecular resolution and help identify promising biomarkers for clinical diagnosis and treatment of complex diseases (5). However, despite the extensive cataloging of circRNAs (6, 7), the number of experimentally confirmed disease-related circRNAs (8, 9, 10) remains relatively small due to the time-consuming and expensive laboratory experiments.

To solve this problem, a series of computational methods have been proposed to identify potential circRNA-disease associations (CDA) recently. These approaches could be roughly divided into three types, including network-based, traditional machine learning-based and deep learning-based methods. Firstly, network-based algorithms construct heterogeneous networks based on known CDA and biological similarity calculation to predict CDA. For example, KATZHCDA (11), KATZCPDA (12), and IBNPKATZ (13) employ the KATZ algorithm to infer potential CDA. Random walk and label propagation are other frequently used methods. Huseyin et al. (14) applied random walk with restart (RWR) to predict CDA based on similarity matrices. Zhang et al. (15) adopted a linear neighborhood label propagation method (CD-LNLP) to identify potential CDA after constructing similarity networks. Secondly, traditional machine learning-based algorithms identify CDA by training classifiers after handcrafted feature construction. Lei et al. (16) proposed GBDTCDA to predict CDA, where gradient boosting decision tree was employed on manually constructed fusion features. Zhang et al. (17) used the weighted nuclear norm minimization to identify associations between circRNAs and diseases. Moreover, Matrix factorization (MF) have also been proposed for CDA identification recently.
due to MF’s exceptional performance in exploring latent information from sparse and noisy data. Wei et al. (18) designed a new computational method (iCircDA-MF) to calculate affinity scores of CDA using MF. Peng et al. (19) combined robust nonnegative matrix factorization with label propagation (RNMFLP) to infer potential CDA. Thirdly, deep-learning-based algorithms have demonstrated the superiority in automatically extracting latent features in many fields (20, 21, 22), which is also a significant branch in the domain of CDA prediction (23, 24, 25, 26). Specially, Lu et al. (24) put forward a computational model (CDASOR) based on convolutional and recurrent neural networks and Deepthi et al. (25) applied deep autoencoder to identify CDA (AE-RF). Considering the graph structure consisting of CDA, Niu et al. (27) used the graph markov neural network (GMNN) to capture deep representations of CDA and make predictions. Lan et al. (28) designed the knowledge graph with attention mechanism (KGANCDa) for inferring unknown CDA.

Although existing methods have achieved significant progress in the field of CDA prediction, they still face limitations due to data sparsity and their inability to efficiently explore high-order information that combines biological data characteristics. For example, most existing methods rely heavily on similar circRNAs to predict disease-related circRNAs, which can be challenging when working with sparse data and unable to effectively explore high-order information. Therefore, it is necessary to incorporate other circRNA-related relations and construct a large heterogeneous dataset to alleviate data sparsity and improve performance of current models to some extent. Additionally, it’s also necessary and significant to propose a method capable of efficiently processing multi-source data and extracting high-order information by integrating the intrinsic properties of biological data.

In this paper, we introduce a new framework for predicting CDA called Knowledge Graph Encoder from Transformer for CDA (KGETCDA). Our framework is based on representation learning and can effectively capture high-order interaction information between multi-source data and the intrinsic properties of biological molecules. We start by integrating over 10 databases to construct a heterogeneous ncRNA dataset, including associations between circRNA, miRNA, lncRNA and disease. After that, a biological knowledge graph is created based on the dataset. Additionally, knowledge graph representation learning and attentive embedding propagation layers are designed to generate optimal embeddings with captured high-order connectivity. Finally, multilayer perceptron is applied to obtain reliable matching scores of CDA based on their embeddings. The empirical experiments demonstrate that KGETCDA outperforms other state-of-the-art (SOTA) models. Moreover, we develop a user-friendly web-based platform for users to visualize, download data and make predictions using KGETCDA conveniently.

The remaining part of this paper proceeds as follows. The section ‘MATERIALS AND METHODS’ introduces the constructions of datasets, similarity measures and the overall architecture of KGETCDA. Then, evaluation criteria, performance comparison with SOTA methods, ablation experiments, parameter setting and case studies are discussed in the section ‘RESULTS’. After that, we present an interactive web-based platform for users in ‘WEB-BASED VISUALIZATION’. Finally, some discussions on KGETCDA and future works are introduced in ‘DISCUSSION AND CONCLUSION’.

MATERIALS AND METHODS

In this section, we present the proposed method KGETCDA, which efficiently captures high-order interaction information. Figure 1 illustrates the model framework, consisting of three parts: data collection, similarity measures and model prediction. Firstly, we collect multi-source data and construct a large, heterogeneous ncRNA dataset. Next, we prepare similarity measures for negative sampling. Subsequently, a biological knowledge graph is created, and optimal embeddings are obtained using Transformer-based knowledge representation learning and attentive embedding propagation layers. Finally, MLP-based prediction layers are applied to output the affinity scores of candidate CDA.

Datasets

In this study, we first obtain 330 circRNAs, 265 miRNAs, 404 lncRNAs, 79 diseases, 346 circRNA-disease associations, 106 miRNA-disease associations, 527 lncRNA-disease associations, 146 circRNA-miRNA associations and 202 miRNA-lncRNA associations from previous work (28) to get a commonly used dataset in previous works for dataset1. Then, for dataset2, we integrate multi-source data from 13 databases to construct a larger heterogeneous dataset. Specifically, for dataset2, we gather associations of circRNA-disease and circRNA-miRNA from Circad (10), MNDR (29), Lnc2Cancer (30), LncRNA2Disease (31), CircRNA2Disease (32), Circ2Disease (8) and CircR2Cancer (9). And the
Figure 1. The flowchart of KGETCDA. (A) Extract disease information by creating a directed acyclic graph. (B) Construct a heterogeneous network via integrating multi-source ncRNA-disease databases. (C) Similarity-based negative sampling. It first constructs the integrated similarity descriptor, and then obtains top-k similar circRNAs of candidate circRNA. Finally, negative samples (diseases) are selected from unknown associated diseases of these diseases. (D) Produce high-quality embeddings and identify potential circRNA-disease associations. (E) Web-based Visualization.

associations of miRNA-disease, miRNA-lncRNA, lncRNA-disease are obtained from MNDR (29), HMDD (33), Circ2Disease (8), StarBase (34), LncRNADisease (31) and Lnc2Cancer (30). We ensure the consistency by matching circRNA names with circBase (6) and circBank (7) as much as possible, addressing the inconsistent standard issue of current circRNA-disease datasets. Additionally, most of the disease names in dataset1 have been replaced with those in disease ontology (35). After removing redundant associations, 561 circRNAs, 658 miRNAs, 1043 lncRNAs, 190 diseases, 1399 circRNA-disease associations, 10154 miRNA-disease associations, 3280 IncRNA-disease associations, 1129 circRNA-miRNA associations and 9506 miRNA-lncRNA associations are obtained for dataset1. Table 1 summarizes the information of the two datasets. To visualize datasets, we show the degree distribution of diseases and the relationship between circRNA-disease and circRNA-position of dataset1 in Figure 2A, 2B and dataset2 in Supplementary Figure S1, and the process of constructing the dataset2 is shown in Figure 2C. We observe that some diseases associate with
more circRNAs than others, highlighting the significance of designing an effective architecture capable of extracting features from the above complex multi-relational scene.

**Similarity measures**

To obtain intrinsic properties of multi-source biological data, an integrated feature descriptor is constructed based on disease semantic similarity, circRNA functional similarity and gaussian interaction profile kernel similarity.

**Disease semantic similarity** Following previous studies (25, 36, 37), the disease information is obtained by calculating disease semantic similarity based on the medical subject headings (MeSH), which is provided by the National Library of Medicine (https://www.ncbi.nlm.nih.gov/). In the MeSH database, diseases are described as a directed acyclic graph (DAG), where nodes represent diseases and edges represent relationships between diseases. For example, given a disease \(d\), it can be represented as \(DAG_d = (d, A_d, R_d)\), where \(A_d\) represents the ancestral node set of \(d\) including itself and \(R_d\) represents the set of \(d\)'s corresponding relationships. Additionally, there are two approaches to calculate disease semantic similarity from different perspectives.

The first method is based on the assumption that different ancestral nodes of disease \(d\) in the same layer have the same contribution to \(d\). Specially, assuming that another disease \(k\) exits in \(DAG_d\), its contribution to disease \(d\) is described by:

\[
\begin{align*}
D_d(k) &= 1 & \text{if } k = d \\
D_d(k) &= \max \{ \lambda \cdot D_d(k') | k' \in \text{children of } k \} & \text{if } k \neq d
\end{align*}
\]

where \(\lambda\) controls the semantic contribution between disease \(k\) and its child diseases. Similar to previous work (36), \(\lambda\) is set to 0.5. Then, we can obtain the semantic value of disease \(d\) as follows:

\[
DV(d) = \sum_{k \in A_d} D_d(k).
\]

Finally, for disease \(m\) and \(n\), we can calculate their semantic similarity based on the assumption that two diseases sharing larger parts in DAG are more similar. The formula is

\[
DSS_1(d(m), d(n)) = \frac{\sum_{k \in A_d(m) \cap A_d(n)} D_d(m)(k) + D_d(n)(k)}{DV(d(m)) + DV(d(n))}
\]
The second method considers that diseases appearing less in DAG may be more important. Therefore, the contribution of disease $k$ can be calculated as follows:

$$D'_s(k) = -\log\left(\frac{\text{num}(DAGs(k))}{\text{num}(\text{diseases})}\right).$$

(4)

where $\text{num}(DAGs(k))$ denotes the number of $DAGs$ containing disease $k$, and $\text{num}(\text{diseases})$ denotes the number of all diseases. Then, similar to the first method, the second semantic similarity can be calculated by:

$$DSS_2(d(m), d(n)) = \sum_{k \in A_d(m) \cap A_d(n)} D'_s(k) + D'_s(k) \frac{D'V(d(m)) + D'V(d(n))}{2}.$$ (5)

$$DV'(d) = \sum_{k \in A_d} D'_s(k).$$ (6)

CircRNA functional similarity Under the hypothesis that circRNAs sharing similar disease groups tend to be functionally similar (36), we can obtain the functional similarity matrix for circRNA. Specifically, we define $D_i$ and $D_j$ to represent the disease set related to circRNA $c_i$ and $c_j$, respectively. Then, the functional similarity between circRNA and can be described as follows:

$$CFS(c_i, c_j) = \sum_{1 \leq m \leq |D_i|} DSS(d_m, D_j) + \sum_{1 \leq n \leq |D_j|} DSS(d_n, D_i) \frac{|D_i| + |D_j|}{|D_i| + |D_j|}.$$ (7)

$$DSS(d_m, D_j) = \max_{1 \leq k \leq |D_j|} (DSS(d_m, d_k)).$$ (8)

where $DSS(d_m, D_j)$ is the semantic similarity value between disease $d_m$ and disease group $D_j$.

Gaussian interaction profile kernel similarity Assuming that similar circRNAs associate with similar diseases and vice versa, the gaussian interaction profile kernel similarity (GIPS) for circRNAs can be calculated (38). Here, we define the association matrix $A \in \mathbb{R}^{n_c \times n_d}$, in which $n_c$ and $n_d$ denote the total number of circRNAs and diseases, respectively. Then the GIPKS for circRNA $c_i$ and $c_j$ can be obtained by:

$$CGS(c_i, c_j) = \exp(-\delta ||A(c_i) - A(c_j)||^2),$$

(9)

$$\delta = \frac{1}{\sum_{n \in I_c} ||A(c_i)||^2},$$

(10)

where $n_c$ is the number of all circRNAs, $\delta$ controls the kernel bandwidth, and $A(c_i)$ and $A(c_j)$ denote the $i$th and $j$th rows of $A$, respectively.

Similarity fusion To obtain a reasonable representation of disease semantic similarity, element-wise averaging of the above two disease similarity methods is applied to construct the disease semantic similarity descriptor:

$$DSS(d(m), d(n)) = \frac{\text{DSS}_1(d(m), d(n)) + \text{DSS}_2(d(m), d(n))}{2}. $$

(11)

Then, considering the information of different similarity methods in different feature spaces, similarities above are integrated to produce the fusion feature descriptor for circRNAs:

$$FV(c_i, c_j) = \begin{cases} 
CFS(c_i, c_j) & \text{if } i \text{ and } j \text{ have functional similarity} \\
CGS(c_i, c_j) & \text{otherwise}
\end{cases}.$$ (12)

**KGETCDA**

We now introduce the method KGETCDA and its detailed implementation here, which captures high-order interaction information in an end-to-end manner. An overview of the KGETCDA architecture is presented in Figure 1D, which consists of four parts: knowledge graph encoder from Transformer, attentive information propagation, MLP-based association prediction and negative sampling based on similarity measuring.

**Knowledge graph encoder from Transformer** Recently, knowledge graph (KG) as a form of structured human knowledge has been widely applied to various applications such as question answering (39), recommendation (40) and time series prediction (41). Due to KG’s outstanding feature representation capability, they are also gradually applied to computational biology like drug-drug prediction (42) and drug-target protein prediction (43). For our research, we can create a heterogeneous KG based on datasets construct before (illustrated in Figure 3). Specially, the KG of ncRNA-disease can be described as $G = (E, R, F)$, where $E$ denotes the entity set (circRNA, miRNA, lncRNA, disease), $R$ denotes the relation set (associations between entities), and $F$ denotes the fact set (a fact is a triplet $(h, r, t) \in E \times R \times E$, $h \in E$, $r \in R$, $t \in E$). Knowledge graph representation learning (KRL, also known as knowledge graph embedding) is an effective tool to model complex relational graphs like biological systems and produce reliable representations of entities and relations for downstream tasks. Furthermore, due to Transformer’s (44) remarkable performance in computer vision (45) and natural language processing (21, 22), considerable efforts have been devoted to integrating Transformers into KRL (46, 47). Inspired by these excellent works, we propose a Transformer encoder-based architecture. As illustrated in Figure 1D, multi-head self-attention mechanism of Transformer encoders is applied to effectively encode KG and produce high-quality embeddings. In particular, for a given triple $(h, r, t) \in F$, each $e \in E \cup R \cup \{\text{head}, \text{tail}\}$ is defined as the embedding vector for head entity, tail entity and relation, respectively. For simplicity, we assume that $d_{\text{head}} = d_{\text{tail}} = d$. Then, the concatenated embeddings $C \in R^{d \times |E|}$ of the pair $(h, r)$ are considered to be the input sequence, which also serves as the query $Q$, the key $K$ and the value $V$ for self-attention mechanism. Therefore, in order to obtain highly expressive information from different representation subspaces, multi-head attention mechanism is used as follows:

$$\text{Multthead}(Q, K, V) = [H_1, ..., H_{n_k}]W_o.$$ (13)
where $\| \|$ represents the concatenation operation, $n_h$ represents the number of heads, and $W_Q^h \in \mathbb{R}^{d_h \times d_e}$, $W_K^h \in \mathbb{R}^{d_h \times d_e}$, $W_V^h \in \mathbb{R}^{d_h \times d_e}$ are linear projection matrices. Additionally, more attention heads are applied than the standard Transformer (44), since increasing the number of attention heads would offer better representations and facilitate latent feature extraction in different subspaces (46). Afterwards, a token-wise feed-forward network following is utilized to further enhance model’s capability to produce powerful representations and perform various complex tasks.

\[
F^{\text{FFN}}(x) = \max(xW_1 + b_1, 0)W_2 + b_2,
\]

where $x$ is the output of multi-head attention mechanism, and $W_1 \in \mathbb{R}^{d_h \times d_f}$, $W_2 \in \mathbb{R}^{d_f \times d_b}$, $b_1 \in \mathbb{R}^{d_f}$, $b_2 \in \mathbb{R}^{d_b}$ are parameter matrices. Note that Yun et al. (48) demonstrated Transformer’s universal approximation power in various sequence-based tasks with the architecture of multi-head attention mechanism combined with a fully connected feed-forward network.

After obtaining learned embeddings of entities and relations based on the above architecture, a score function is designed for measuring the plausibility of a given triple $(h, r, t)$ similar to previous work (49, 51). In this paper, we use a simple method named TwoMult (46, 50), which is formulated as follows:

\[
f(h, r, t) = e_r^T e_t,
\]

where $e_r$ is the updated embedding of the relation passing the Transformer-encoder based architecture above and $e_t$ is the initial embedding of tail entity. Despite its simplicity, TwoMult is an effective method to decode embeddings and produce reliable matching scores of $(h, r, t)$ (46). Moreover, it’s our belief that the Transformer-encoder-based architecture has perfectly embedded the information of head entity into the relation embedding and thus we only use the updated relation embedding in TwoMult. Noteworthy, we use the simple inner product as scoring function here for simplicity, and leave further exploration as future works. Finally, a margin-based ranking criterion (49) following is applied to train KRL part, which aims to encourage the discrimination between positive and negative triples.

\[
L_1 = \frac{1}{|P_{\text{train}}|} \sum_{(h, r, t), (h', r, t') \in T} -\log \sigma(f(h, r, t) - f(h', r, t')),
\]

where $\sigma$ is the sigmoid function, $|P_{\text{train}}|$ is the number of all facts in training set, and $T$ is the set of all positive triples and equal proportion of negative triples, which are constructed by replacing head or tail entity in a positive triple once. A positive triple will be assigned a higher score while a negative triple gets a lower score. In summary, the architecture of knowledge graph encoder from Transformer constructs a hierarchical representation of complex biological KG by effectively merging high-order neighborhood information, which paves the way for downstream tasks such as CDA prediction.

**Attentive information propagation** Considering that our downstream task is CDA prediction and the properties of each entity in the above KG structure have no difference by default, we build an attentive information propagation architecture following (40) to further distinguish different neighbors and strengthen the representation of two specific entities, i.e. circRNA and disease. In the following, we will start with a single layer, consisting of two parts: information propagation and information aggregation, and then extend to multiple layers.

**Information propagation** To obtain high-quality expression, it’s crucial to fully exploit the information of neighbors of an entity. Therefore, we will describe how to generate an efficient combination of neighborhood information in this part. Specifically, given an entity $h$ (circRNA or disease), we can define its ego-network $N_h = \{(h, r, t)|h, t \in E, r \in R, (h, r, t) \in F\}$ (52), which represents triples consisting of the entity $h$, all neighbors of $h$ and their relations. Then, the integrated information of $h$‘s ego-network can be defined as follows:

\[
e_{N_h} = \sum_{(h, r, t) \in N_h} \alpha(h, r, t) e_t,
\]

where $e_t$ is the embedding of tail entity defined before, and $\alpha(h, r, t)$ is the propagation score deciding how much information is propagated from neighbor $t$ to $h$. Since different neighbors have different importance (40, 54), knowledge-aware attention is applied to calculate the propagation score, which is formulated as follows:

\[
\alpha(h, r, t) = \frac{\exp(\alpha(h, r, t))}{\sum_{(h', r', t') \in N_{h'}} \exp(\alpha(h', r', t'))}.
\]

**Information aggregation** After the information propagation, we can aggregate the representation of the entity $h$ and its ego-network. Graph convolutional network (GCN) aggregator (55) is employed, which is defined as follows:

\[
f_{\text{GCN}}(e_h, e_{N_h}) = \text{LeakyReLU}(W(e_h + e_{N_h})),
\]

where $W \in \mathbb{R}^{d_h \times d}$ is the trainable matrix and $d_h$ is the transformation size. Moreover, GraphSage aggregator (53) and Bi-Interaction aggregator (40) following are applied to compare with GCN aggregator to demonstrate the effectiveness of the GCN aggregator in our research.

\[
f_{\text{GraphSage}}(e_h, e_{N_h}) = \text{LeakyReLU}(W(e_h || e_{N_h})),
\]

\[
f_{\text{Bi-Interaction}}(e_h, e_{N_h}) = \text{LeakyReLU}(W_1(e_h + e_{N_h}) + \text{LeakyReLU}(W_2(e_h \odot e_{N_h}))),
\]

\[
f_{\text{GCN}}(e_h, e_{N_h}) = \text{LeakyReLU}(W(e_h + e_{N_h})),
\]

\[
f_{\text{GraphSage}}(e_h, e_{N_h}) = \text{LeakyReLU}(W(e_h || e_{N_h})),
\]

\[
f_{\text{Bi-Interaction}}(e_h, e_{N_h}) = \text{LeakyReLU}(W_1(e_h + e_{N_h}) + \text{LeakyReLU}(W_2(e_h \odot e_{N_h}))),
\]
where $\| \|$ denotes the concatenation operation and $\odot$ denotes the element-wise product.

After obtaining the first-order interaction information, we can stack more layers to further strengthen high-order representations of circRNAs and diseases, which contain abundant information from their multi-hop neighbors. In particular, for the entity $h$, the representation of $l$-th layer can be described as follows:

$$e_h^{(l-1)} = \text{GCN}(e_h^{(l-2)}, e_h^{(l-1)}),$$

$$e_h^{(l-1)} = \sum_{(h, r, t) \in \mathcal{N}_h} \alpha(h, r, t)c_{ij}^{(l-1)},$$

where $e_h^{(l-1)}$, $e_h^{(l-2)}$, $\mathcal{N}_h$ is the embedding of entity $h$, $t$, and ego-network of $h$ in the $(l-1)$-th layer, respectively. Clearly, after performing above attentive propagation layer architecture, the high-order interaction information is further explored, and the encoding representation of circRNAs and diseases are more expressive and reliable. Following (28, 40), $l$ is set to 4 because the fourth-order information is sufficient for accurately exploiting latent features in KG.

**MLP-based association prediction** With above multiple attentive propagation layers, high-quality embeddings of each entity (circRNA and disease) with accurately captured high-order information are obtained. Note that the embeddings of different layers represent the interaction information of corresponding orders. The concatenation operation of these embeddings is applied to further enrich the final representation and adjust the strength of information propagation as follows:

$$e_{c_{ij}} = e_{c_1}^{(0)} \| \cdots \| e_{c_1}^{(l)} \| \cdots \| e_{c_d}^{(0)} \| \cdots \| e_{c_d}^{(l)},$$

where $\|$ denotes the concatenation operation. Similar to (23, 25, 28), we also concatenate the embedding $e_{c_{ij}}$ and $e_{d_{ij}}$ of the two entity to get fusion descriptor of a candidate CDA. Then, the multilayer perceptron is applied on the concatenated features to produce reliable affinity scores, which is formulated as follows:

$$\hat{y}(i, j) = W_1\sigma_1(W_{l-2}\sigma_{l-2}(\cdots (W_1(e_{c_{ij}}) || e_{d_{ij}}) + b_1) \cdots ) + b_{l-1}) + b_l,$$

where $W_l$, $b_l$ and $\sigma_l$ are the trainable parameter matrix, the bias, and the activation function in the $l$-th linear layer. Moreover, we select ReLU (56) as the activation function in hidden layers and sigmoid in the output layer to control the output to be a probability value from 0 to 1. To optimize this part, we get the BPR loss (57) as:

$$L_{BPR} = \frac{1}{|P|} \sum_{(i, j, j') \in S} -\ln \text{sigmoid}(\hat{y}(i, j) - \hat{y}(i, j')),$$

where $|P|$ is the number of all pairs consisting of head entities and tail entities in training set, and $S$ is the set of all positive pairs and negative triples of equal proportion, which represents unobserved triples here.

Finally, to achieve the joint model optimization in an end-to-end manner, the whole loss is defined as follows:

$$L_{total} = L_1 + L_2 + \lambda |\alpha|^2,$$

where $\lambda$ is a regularization parameter to prevent overfitting, $\alpha$ is all parameters of the model, and $L_1$ and $L_2$ are the margin-based ranking loss and BPR loss, respectively.

**Negative Sampling based on similarity measuring** Recent studies have demonstrated the significant role that negative sampling plays in efficient model learning in various fields, including knowledge graph embedding (49), recommendation (59) and computational biology (26, 58). High-quality negative sampling can effectively facilitate the model learning towards correct direction. However, existing methods commonly rely on a random negative sampling strategy, which selects unobserved CDA as negative samples randomly and thus often result in unidentified false negative signals. Therefore, following (26), a similarity-based negative sampling strategy is applied in this paper. Specifically, given that similar circRNAs tend to associate with similar diseases and vice versa, we can obtain integrated similarity descriptor in the part ‘Similarity measures’. Afterwards, considering a circRNA $i$, we use $SC = \{c_1, c_2, \cdots, c_{k-1}, c_k\}$ and $SD = \{d_1, d_2, \cdots, d_m\}$ to denote the set of the top-$k$ most similar circRNAs selected based on the similarity descriptor, and the set of all diseases, respectively. Then, the related diseases $SD_i = \{d_1, d_2, \cdots, d_m\}$ of all circRNAs $SC$ in the training set can be obtained and they will not be chosen to consist of the negative CDA. We generate a negative pair ($c_i, d_j$) only by selecting the disease $j$ from $SD - SD_i = \{d_{m+1}, d_{m+2}, \cdots, d_n\}$. The specific architecture of negative sampling is illustrated in Figure 1C. It’s our belief that this similarity-based strategy could produce high-quality negative samples for better model learning and thus effectively integrate the intrinsic properties of biological data into our KG-based model.

**RESULTS**
To systematically verify the effectiveness of our method, a series of experiments are performed on the two above datasets. We first introduce the evaluation criteria and metrics. Then, the proposed method KGETCDA is compared with other SOTA methods. Moreover, ablation experiments are carried out to further evaluate the effectiveness of different modules. After that, we provide the best parameter settings of KGETCDA. Finally, we conducted case studies to evaluate the capability of our method for identifying potential associations between circRNAs and diseases. Through these experiments, we demonstrated the effectiveness of KGETCDA in accurately predicting potential CDA and capturing high-order interaction information with biological properties in an end-to-end manner.

**Evaluation criteria** Following previous studies (19, 28, 38), 5-fold cross-validation (CV) method is adopted to evaluate KGETCDA and compare with other SOTA models in this paper. Specifically, in the 5-fold CV, all samples of the whole dataset are randomly divided into five equal parts and each part is selected for testing in turn while the remaining parts is utilized to train the model. Then, the prediction scores of test samples and unknown associations are obtained and sorted in descending order. The final results are the average values of five folds. Note that such a strategy helps reduce the over-fitting by fully using all samples in training. Finally, some classical evaluation metrics are applied to quantify the obtained results, including accuracy (Acc), precision (Pre), recall
With the same i.e. metrics comparison and statistical hypothesis test. we will compare them systematically from two perspectives, DMFCDA (23), and deep learning-based (GMNN2CD (27), based (KATZHCDA (11), RWR (14) and CD-LNLP. We compare KGETCDA with 8 SOTAs, including network-based negative sampling (NS) methods. Then, the effect of the other eight methods maintain consistency with their experimental settings, the final results of 5-fold CV on two datasets are obtained for nine models. The parameters information is listed in Supplementary Table S1, where KGETCDA obtains the best results of 0.4956 on Acc, 0.9228 on Rec, 0.0093 on Pre, and 0.0173 F1. Moreover, in order to further verify the prediction ability of models, the average number of correctly predicted CDA of five folds from top 10 to 40 is obtained, as shown in Figure 4C. Note that KGETCDA significantly yields the best performance (25.8, 35.8, 46, 54 for top 10 to top 40), which again demonstrates its strong prediction capability. More details are shown in Supplementary Table S2.

For dataset2, as shown in Figure 5A and 5B, KGETCDA still obtains remarkable performance in AUC and AUPR, which again demonstrates our model’s strong prediction capability and excellent generalization. Additionally, we also take the average number of correctly predicted CDA of five folds from top 10 to 40 to further validate the prediction ability of our model on this large dataset. Noteworthy, as shown in 5C, our method significantly outperforms other methods (33.6, 51.2, 65.6, 78.4 for top 10 to top 40). On further analysis, due to the complex multi-relational scene, the performance of other methods is often limited. However, our method still maintains an outstanding prediction ability and produces reliable prediction results. More details are shown in the Supplementary File.

In summary, our Transformer-based representation learning framework is more capable of identifying potential CDA than the baselines via accurately capturing high-order interaction information such as circRNA-lncRNA-disease, circRNA-LncRNA-miRNA-lncRNA-disease.

**Statistical hypothesis test** To further evaluate the significant differences between KGETCDA and the baselines from statistical perspective, a nonparametric Wilcoxon signed-rank test-method (64) is applied, which is a powerful tool for making use of the magnitudes of the differences and thus effectively detect statistical differences between two models. The null hypothesis \( H_0 \) is considered that there is no obvious difference between KGETCDA and the baselines \( H_0: \text{difference}=0 \) while the alternative hypothesis \( H_1 \) is \( H_1: \text{difference} \neq 0 \). Similar to (62, 63), we set the default P-value threshold to 0.05. As shown in Supplementary Table S3, all P-values between KGETCDA and other models are less than 0.05, and thus we reject the null hypothesis. Clearly, KGETCDA is superior to SOTA methods from a statistical perspective.

To summarize, the proposed method KGETCDA significantly outperforms all baselines from both biological and statistical perspectives. It’s our belief that this representation learning framework has a strong potential for effectively applying in other areas and lead to further improvement.

**Ablation Study**

In this part, some ablation experiments are conducted to get deeper insights into the overall framework. We first study the performance of KGETCDA with different knowledge graph embedding (KGE) models and similarity-based negative sampling (NS) methods. Then, the effect of different tower structure of linear layers, dimensionality of embeddings, and the number of attention heads and layers...
in KRL are discussed. Finally, we explore the influence of different aggregators and scoring methods (illustrated in the Supplementary File). Here we present the ablation results on dataset1, and results on dataset2 are available in Supplementary Figure S2.

Different KGE and similarity-based NS methods
To better verify the effectiveness of different modules, we conduct an ablation study on KGE and similarity-based NS modules. Specifically, we replace the Transformer-based KGE with commonly used baselines TransD (65), TransH (66) or ConvKB (67), named as TransD+INS, TransH+INS and ConvKB+INS, respectively (INS means integrated similarity-based NS). Moreover, different similarity-based NS methods are compared, where we replace NS part with GIPKS (named as Transformer+GNS), functional similarity (Transformer+FNS), respectively. We obtain another variant by removing the similarity-based NS module and just use random NS strategy, named as Transformer+No NS. Note that only one module is replaced at once.

Five fold cross-validation AUC results are shown in Figure 6A. We observe that our design achieves slightly higher AUC than TransD+INS, TransH+INS and ConvKB+INS, which proves the powerful information extracting capability from multi-source data of our Transformer-based KRL. It’s promising that further exploring on this KRL framework in future work may lead to better improvements. Additionally, for the NS part, our design achieves remarkable performance, which indicates that integrating similarities from different representation spaces can effectively incorporate the intrinsic properties of biological data into KG and thus improve the representation learning. Note that all variants here outperform the current SOTA method GMNN2CD, which further demonstrate our method’s strong prediction capability.

Effect of the embedding dimensionality
To explore the effect of the embedding dimensionality, we vary the size of dimensionality in 64, 128, 256, 512, 1024, 2048, 4096, which are termed with corresponding size (for example, size 64 is KGETCDA/64), and the results are illustrated in Figure 6B by scaling up the metric values for better visualization. Note that the performance gradually improves while increasing the size of embedding dimensionality until 4096. Specially, KGETCDA/2048 achieves the highest AUC and AUPR, which indicates that high embedding dimension enhance KGETCDA’s model learning capability and thus improve the performance. However, too high dimensions may introduce unusual features, lead to over-fitting problem and reduce generalization capabilities of the model.

Effect of the number of attention heads and layers in KRL
We further conduct a study to investigate the effect of the number of attention heads and layers in KRL. From Figure 6C we notice that KGETCDA achieves the best...
Figure 6. The results of ablation experiments on dataset1. (A) The effect of different knowledge graph embedding models and similarity-based negative sampling strategies. Our design achieves the best performance in both mean and median, and all variants here outperform the current SOTA method GMNN2CD. (B) The effect of different embedding dimensionality. We select 2048 here after considering both AUC and AUPR performance, and computational consumption. (C) The effect of the number of attention heads and layers in knowledge representation learning. (D) The difference of different tower structure of linear layers.

performance while the number of attention heads and layers are set to 16 and 1, respectively. While increasing the number attention heads from 4 to 16, the model performs better overall. It demonstrates that more attention heads facilitate more efficient representation learning and help encode more expressive information in different representation subspaces into embeddings than using multiple layers. Additionally, the performance of KGETCDA no longer increases and even degrades when attention heads increase from 16 to 64. This may be because large features lead to sparse feature representation and thus easily introducing useless information and even noise.

Tower structure of linear layers As illustrated in (20, 61), reducing the number of hidden units after each hidden layer can facilitate more abstractive features learning and the tower structure has been proven to be particularly efficient and suitable, which reduces $1/2^n$ to for each layer. Specially, we consider that the reduction for each layer is searching in 1/2, 1/4, 1/8, 1/16, 1/32. From Figure 6D, we observe that KGETCDA with halving the layer size is superior to others and we use this structure in our research.

Parameter setting

We implement KGETCDA in PyTorch with an NVIDIA Titan Xp GPU and optimize the whole model with Adam (60) optimizer. The learning rate and weight decay on all parameters are set to 1e-4 and 1e-5 with 100 training epochs. To accurately exploit high-order features and obtain high-quality embeddings, the dimension of each propagation layer is set to 512, 256, 128, 64, respectively. To reduce the over-fitting on the train set, we apply dropout after each propagation layer with the same rate of 0.1. We also apply dropout to the input of KRL, the output of scaled dot product, multi-head attention and feed-forward network with a rate of 0.2, 0.3, 0.2, 0.3 by default, respectively. Additionally, after extensive experiments, our method achieves the best performance while the ratio of negative to positive samples is set to 8:1 and the number of similar-based sampling is 10. The detailed information of parameter setting are listed in Supplementary Table S6.

Case studies

To further evaluate the effectiveness of KGETCDA, case studies are conducted on both two datasets. Following (23,
Table 2. The top-10 predicted lung cancer or leukemia related candidate circRNAs

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Disease</th>
<th>Candidate circRNA</th>
<th>Rank</th>
<th>Evidence (PMID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset1</td>
<td>Leukemia</td>
<td>hsa_circ_0075001</td>
<td>1</td>
<td>28971903</td>
</tr>
<tr>
<td></td>
<td></td>
<td>circpan3</td>
<td>2</td>
<td>31401408, 30395908</td>
</tr>
<tr>
<td></td>
<td></td>
<td>circ-anapc7</td>
<td>3</td>
<td>29969755</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_100290</td>
<td>4</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0035381</td>
<td>5</td>
<td>28282919</td>
</tr>
<tr>
<td></td>
<td></td>
<td>circznf669</td>
<td>6</td>
<td>33259601</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0017446</td>
<td>7</td>
<td>28282919</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_000488</td>
<td>8</td>
<td>30037980</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0049657</td>
<td>9</td>
<td>28282919</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0008078</td>
<td>10</td>
<td>28282919</td>
</tr>
<tr>
<td>Dataset2</td>
<td>Lung Cancer</td>
<td>CDR1as</td>
<td>1</td>
<td>30022841</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0007534</td>
<td>2</td>
<td>30017736</td>
</tr>
<tr>
<td></td>
<td></td>
<td>circSOX4</td>
<td>3</td>
<td>32112500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>circWHSC1</td>
<td>4</td>
<td>34551195, 29722168</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0000735</td>
<td>5</td>
<td>30737027</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0015449</td>
<td>6</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0075829</td>
<td>7</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0067934</td>
<td>8</td>
<td>29863250, 30344708</td>
</tr>
<tr>
<td></td>
<td></td>
<td>circSMARCA5</td>
<td>9</td>
<td>33085761</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hsa_circ_0025033</td>
<td>10</td>
<td>unknown</td>
</tr>
</tbody>
</table>

Figure 7. The overview of web-based platform.

38, 62), we train the model KGETCDA on all known CDA, and then predict potential probabilities for all unknown CDA and sort them in descending order. Moreover, we collect experimental evidence to verify them in public databases and newly published literatures, and the results are listed in Table 2.

For dataset1, we choose leukemia, which has affected human health severely, and existing studies have demonstrated that the studies of circRNA benefit the treatment of leukemia. From Table 2, we can observe that top-10 leukemia-related leukemia, and 9 of 10 circRNAs (hsa_circ_0075001, circpan3, circ-anapc7, hsa_circ_0035381, circznf669, hsa_circ_0017446, hsa_circ_0000488, hsa_circ_0049657, hsa_circ_0008078) are confirmed. For example, circpan3 mediates drug resistance in leukemia via the miR-183-5p-XIAP/miR-153-5p axis (68). Hsa_circ_0000488 accelerates human leukemia progression via promoting PRKACB expression and suppressing microRNA 496 (69).

For dataset2, we choose lung cancer (LC), which is a common cancer worldwide, and accurately identifying breast cancer-related circRNAs can offer new insights into disease diagnosis and treatment. As shown in Table 2, top-10 lung cancer-related circRNAs are selected based on their predicted scores, and 7 of 10 circRNAs (CDR1as, hsa_circ_0007534, circSOX4, circWHSC1, hsa_circ_0000735, hsa_circ_0067934, circSMARCA5) are confirmed by existing literatures. For example, as an oncogene, overexpressed CDR1as ranked at top 1 can promote the tumor progression via miR-7 in lung cancer (70). The expression level of hsa_circ_0067934 ranked at top 8 affects the proliferation and metastasis of LC (71).
In summary, case studies conducted on two datasets demonstrate our proposed method’s significant role in identifying unknown CDA, and it deserves experimental tests to further verify these CDA.

WEB-BASED VISUALIZATION

Considering that existing computational methods often present results and offer collected data resources only in the form of static figures or tables, which may not be user-friendly relatively for researchers to use, we develop an online interaction web-based platform named HNRBase, which provides intuitive visualization and user-friendly web interaction interfaces for intelligent search, convenient data download, and reliable model prediction. In particular, as illustrated in Figure 7, our web interaction interface mainly consists of 4 core functions: intelligent search and browse, model prediction, information visualization, and advanced interaction. The detailed architecture design and description are illustrated in the Supplementary File. Note that our web-based platform enables novel visualization, accessible resources and user-friendly interaction. This design enhances user experience well and benefits further research in the field of CDA prediction. Moreover, to further utilize our method, we expand this framework to miRNA-disease and lncRNA-disease prediction due to the high similarity between them and CDA prediction.

DISCUSSION

Increasing studies have demonstrated that circRNA plays a significant role in the initiation and progression of complex human diseases, exhibiting rich biological functions. Effectively predicting potential CDA can offer novel insights into disease diagnosis and treatment. However, existing computational methods face challenges in effectively exploring high-order interaction information that incorporates biological intrinsic properties due to data sparsity problem. In this paper, we propose KGETCDA, a new CDA prediction method based on a Transformer-based knowledge graph representation learning framework, which produces high-quality feature representations. Empirical experiments conducted on two heterogeneous datasets reveal that KGETCDA significantly outperforms state-of-the-art (SOTA) methods, and accurately captures high-order information with biological properties in an end-to-end manner. Moreover, we develop an online interaction web-based platform named HNRBase, which provides user-friendly interfaces for intuitive visualization, intelligent search, convenient data download, and reliable model prediction. This presentation manner of results enhances user experience well and benefits further research in the field of CDA prediction.

Furthermore, as there are still some promising avenues for further research, we suggest several intuitive improvements worth exploring. Firstly, more biological data can be integrated to further verify KGETCDA’s effectiveness, and we hope that this representation learning framework can open up future research probabilities for other areas such as drug-protein prediction and gene-RNA prediction. Secondly, more efficient scoring methods in Transformer-based knowledge representation learning can be introduced to further improve performance. Thirdly, incorporating additional biological auxiliary knowledge, such as RNA sequence and gene expression information, into KG would be a promising approach. Finally, due to Transformer’s intensive computational and memory consumption, we anticipate investing more efforts to tackle this problem in the future.

DATA AVAILABLE

The code and datasets are publicly available at https://github.com/jinyangwu/KGETCDA.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/nar.

AUTHOR CONTRIBUTIONS STATEMENT

J.W. and L.F. conceived the project, conducted the experiments, analysed the results, and wrote the manuscript. Z.N., Y.D., Y.W. and Q.P. reviewed the manuscript.

FUNDING

This work was supported by Zhejiang Provincial Natural Science Foundation of China [LQ23F020018], Natural Science Basic Research Program of Shaanxi [2023-JC-QN-0737], Sichuan Science and Technology Program, China [2023NSFSC1416], and National Natural Science Foundation of China [61872288].

ACKNOWLEDGEMENTS

We acknowledge helpful discussions with members of the Peng lab.

Conflict of interest statement. None declared.

REFERENCES


14 Nucleic Acids Research, YYYY, Vol. xx, No. xx

2019, preprint: not peer reviewed.


