International Journal of
Molecular Sciences

MDPI

Article

GATNNCDA: A Method Based on Graph Attention Network and Multi-Layer Neural Network for Predicting circRNA-Disease Associations

Cunmei Ji [1,*], Zhihao Liu [1], Yutian Wang [1], Jiancheng Ni [1] and Chunhou Zheng [2,*]

[1] School of Cyber Science and Engineering, Qufu Normal University, Qufu 273165, China; liuzhihao19971002@gmail.com (Z.L.); wytfuture@163.com (Y.W.); nijch@163.com (J.N.)
[2] School of Artificial Intelligence, Anhui University, Hefei 230601, China
* Correspondence: cunmeiji@126.com (C.J.); zhengch99@126.com (C.Z.)

Abstract: Circular RNAs (circRNAs) are a new class of endogenous non-coding RNAs with covalent closed loop structure. Researchers have revealed that circRNAs play an important role in human diseases. As experimental identification of interactions between circRNA and disease is time-consuming and expensive, effective computational methods are an urgent need for predicting potential circRNA–disease associations. In this study, we proposed a novel computational method named GATNNCDA, which combines Graph Attention Network (GAT) and multi-layer neural network (NN) to infer disease-related circRNAs. Specially, GATNNCDA first integrates disease semantic similarity, circRNA functional similarity and the respective Gaussian Interaction Profile (GIP) kernel similarities. The integrated similarities are used as initial node features, and then GAT is applied for further feature extraction in the heterogeneous circRNA–disease graph. Finally, the NN-based classifier is introduced for prediction. The results of fivefold cross validation demonstrated that GATNNCDA achieved an average AUC of 0.9613 and AUPR of 0.9433 on the CircR2Disease dataset, and outperformed other state-of-the-art methods. In addition, case studies on breast cancer and hepatocellular carcinoma showed that 20 and 18 of the top 20 candidates were respectively confirmed in the validation datasets or published literature. Therefore, GATNNCDA is an effective and reliable tool for discovering circRNA–disease associations.

Keywords: circRNA–disease associations; graph attention network; multi-layer neural network

Citation: Ji, C.; Liu, Z.; Wang, Y.; Ni, J.; Zheng, C. GATNNCDA: A Method Based on Graph Attention Network and Multi-Layer Neural Network for Predicting circRNA-Disease Associations. *Int. J. Mol. Sci.* **2021**, *22*, 8505. https://doi.org/10.3390/ijms22168505

Academic Editor: Alessandro Desideri

Received: 21 July 2021
Accepted: 3 August 2021
Published: 7 August 2021

1. Introduction

Circular RNAs (circRNAs) are a new class of endogenous non-coding RNA lacking a 5′ cap and a 3′ polyadenylated tail [1,2]. Since circRNAs were first discovered, in the 1970s, they have been considered as splicing errors [3,4]. In the past decade, with the development of high-throughput sequencing technology, a large number of circRNAs have been identified in mammalian cells [5,6]. Researchers have found that circRNAs are widely expressed in human tissues, and have stable structure and tissue-specificity. The mechanism of circRNA expression remains unknown, and how the biogenesis of circRNA affects its unique regulatory pattern remains limited [7]. Studies have revealed that many circRNAs perform their biological functions by acting as sponges of microRNA or RNA-binding proteins, by regulating protein function or by being translated themselves [8–10].

Cumulative evidence has indicated that many circRNAs are involved in human diseases, especially cancers [11]. For example, circHIPK3 has been found significantly up-regulated in colorectal cancer (CRC) tissues by sponging miR-7 to inhibit miR-7 activity [12]. Hsa_circ_0000190 was down-regulated in gastric cancer (GC) tissues and plasma from patients with GC. Compared with common biomarkers such as CEA and CA19-9, it has better sensitivity and specificity, and can be used as a novel biomarker for diagnosis of gastric cancer [13]. Researchers have identified that the expression of hsa_circ_0005075 is

significantly different between hepatocellular carcinoma (HCC) and normal tissues [14]. The expression of Hsa_circ_0001649 was significantly different between HCC and normal liver tissues [15]. Moreover, circRNAs have also been related to other human diseases. CircANRIL is related to atherosclerotic disease by binding to pescadillo homolog 1 (PES1), which then impairs pre-rRAN processing and ribosomal biogenesis, results in the activation of p53, and thereby induces apoptosis and inhibits proliferation [14]. Recent studies have shown that the circRNA level in the brain is associated with Alzheimer's disease (AD) [16]. Compared with the control group, Li et al. have found that 112 circRNAs were up-regulated and 51 circRNAs were down-regulated in AD patients [17], which also were enriched in AD-related pathways, and the clinical guidance of circ-AXL, circ-GPHN and circ-PCCA in disease management of AD patients was identified.

As researchers have realized that circRNAs are abundant in mammalian cells, evolutionarily conserved and stable, and could serve as better biomarkers [18], databases of rich circRNA information, such as circBase [19], circ2traits [20], CircFunBase [21] have been built for study. Furthermore, researchers have also manually curated evidence from published literature, established databases such as circRNADisease [19], CircR2Disease [22], Circ2Disease [23], and circAtlas [24]. While experimental verification is expensive and time-consuming, computational methods have gradually introduced inferring potential circRNA–disease associations. Lei et al. first proposed a path weighted method to predict disease-related circRNAs. They calculated disease semantic similarity, disease functional similarity and integrated with the Gaussian Interaction Profile (GIP) kernel similarities. Then, they constructed a heterogeneous network and adopted the depth-first search (DFS) to traverse nodes in the network and calculate the predictive score [25]. Yan et al. developed the DWNN-RLS method based on Regularized Least Squares of Kronecker product kernel for predicting circRNA–disease associations, and obtained AUC values of 0.8854, 0.9205 and 0.9701 in fivefold, 10-fold and leave-one-out cross validation, respectively [26]. Another graph-based method KATZHCDA achieved the best AUC values of 0.7936 and 0.8469 in fivefold CV and LOOCV, respectively [27]. Xiao et al. developed a weighted low-rank approximation optimization method with dual-manifold regulations to infer potential circRNA–disease associations [28].

Deep learning algorithms have also been introduced in this field. Deepthi et al. proposed an ensemble method named AE-RF, which extracted features via deep autoencoder, and then used random forest for prediction. As a result, this method achieved 0.9486 and 0.9552 in fivefold and 10 fold CV, respectively [29]. Li et al. used DeepWalk to extract node features in the circRNA–disease network, and used a network consistency projection algorithm for circRNA–disease interactions prediction [30]. Wang et al. designed GCNCDA using FastGCN to extract high-level features, and by applying Forest PA classifier for prediction [31]. As a result, it achieved an AUC value of 0.909 in fivefold CV based on circR2Disease dataset. Bian et al. developed GATCDA method based on graph attention network to obtain representation of circRNAs and diseases, calculated the probability score by dot production [32], and yielded an AUC value of 0.9011.

In this study, we proposed a novel computational method named GATNNCDA to predict potential circRNA–disease associations, based on graph attention network and multi-layer neural network. To be specific, GATNNCDA first integrates circRNA functional similarity, disease semantic similarity and the GIP similarities. Secondly, GATNNCDA utilizes linear transformation to project the integrated similarity matrices into the same space, and applies a graph attention network to extract dense representations of nodes in the heterogeneous circRNA–disease graph. Furthermore, a multi-layer neural network is constructed to infer the associations between circRNAs and diseases. The framework of GATNNCDA is shown in Figure 1. In summary, our contributions are listed as follows:

- We proposed an end-to-end framework for inferring disease-related circRNAs, which can effectively and accurately infer the potential associations between circRNAs and diseases.

- We made use of GAT to extract low-dimensional dense representations of circRNAs and diseases, and these presentations had rich structural and semantic information of the heterogeneous circRNA–disease graph.
- We proposed a NN-based classifier, and applied a sampling strategy to construct balanced samples. In addition, we designed cross-entropy loss with L2 regularization to make the training process fast and robust.
- We demonstrated the predictive performance of our method by extensive experiments via fivefold cross validation and case studies, and achieved competitive results on CircR2Disease and circRNADisease datasets.

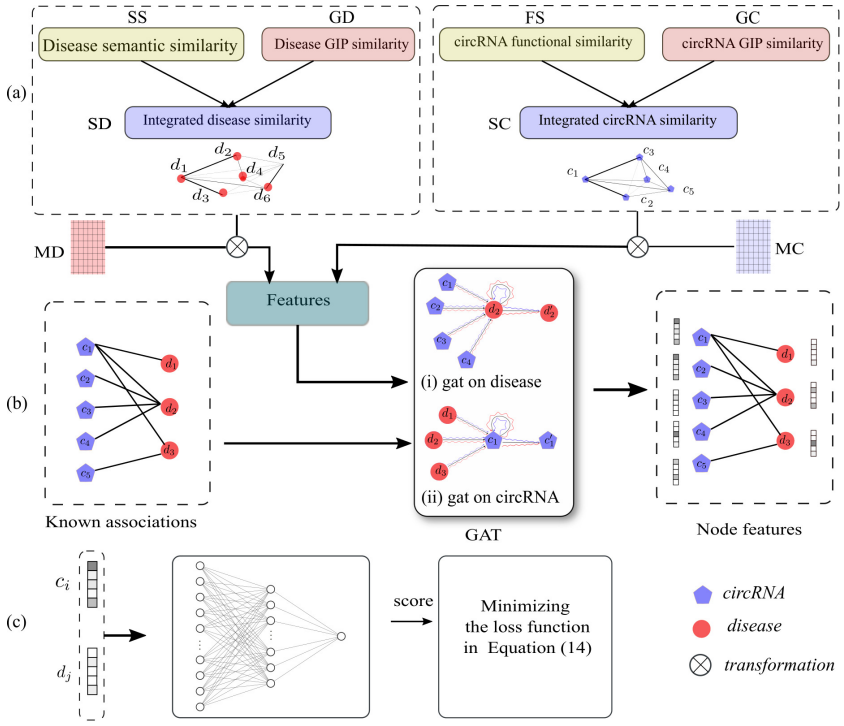

Figure 1. The framework of GATNNCDA. It consists of three steps: (**a**) similarity integration for circRNA and disease, (**b**) GAT-based feature extraction, and (**c**) NN-based classification.

2. Results and Discussion

2.1. Experiments Settings

In our experiments, we conducted fivefold cross-validation (fivefold CV) to evaluate the prediction performance of GATNNCDA. In particular, we randomly split all samples into five groups, of which four of them were used for training and the other group for validation. Furthermore, we carried out several commonly used criteria in this field [33–35] to quantitatively analyze the performance of our method, such as accuracy, precision, recall and F1-score. Moreover, we also plotted the receiver operating characteristic curve (ROC) and precision-recall (PR) curve, and calculated the area under the ROC curve (AUC) and the area under the PR curve (AUPR).

The implementation of our method was based on Python machine learning library PyTorch v1.6.0 [36]. Graph attention network was developed by using PyTorch Geometric deep learning library [37]. We carried out our experiments on the Ubuntu 16.04, with two Tesla V100 GPUs. The default settings for GAT are 2 GAT layers and 4 heads. While the dimension size is set to 32, the classifier is 2-layer fully-connected layers. In addition, We used Adam optimizer [38] to update parameters of GATNNCDA iteratively.

2.2. Performance Analysis

To evaluate the performance of our method, we conducted the fivefold CV on CircR2Disease [22]. Here, $N_c = 585$ and $N_d = 88$ denote the number of circRNAs and diseases. We performed fivefold CV 50 times on CircR2Disease, and the best performance is shown in Table 1, with average accuracy of 0.9315, precision of 0.9714, recall of 0.9615, F1-score of 0.9336, AUC of 0.9742 and AUPR of 0.9707. We also plotted the ROC and PR curves as shown in Figure 2. The average AUC and AUPR values of 50 times are 0.9619 and 0.9452, respectively.

We also performed the fivefold CV on another commonly used circRNA–disease association dataset, cicRNADisease [19]. In circRNADisease, the number of circRNAs $N_c = 313$, and the number of disease $N_d = 44$. We can construct a circRNA–disease graph, calculate the similarities and train and validate GATNNCDA by similar criteria. The results are shown in Table 2. It can be seen that GATNNCDA obtained an average accuracy of 0.9638, precision of 0.9852, recall of 0.9910, F1-score of 0.9649, AUC of 0.9882 and AUPR of 0.9848. Therefore, the results on CircR2Disease and cicRNADisease showed that GATNNCDA performed well and can promote the prediction performance of potential disease-related circRNAs.

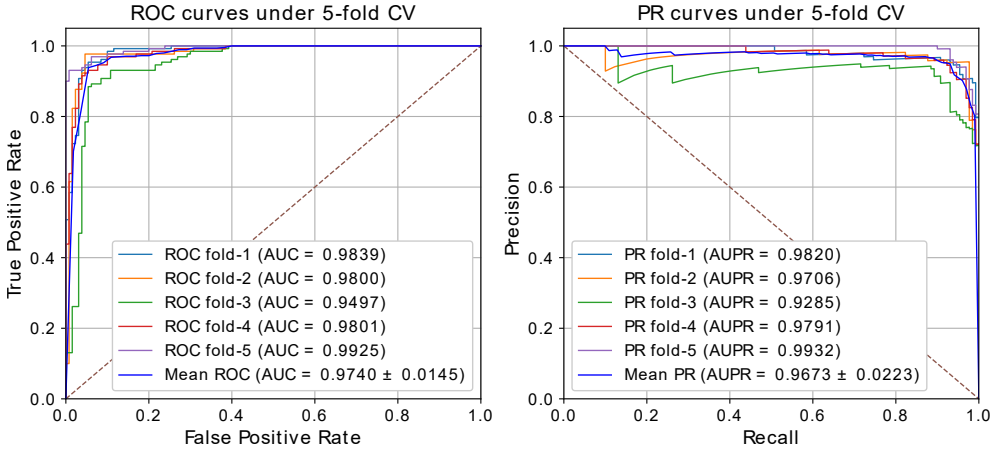

Figure 2. The framework of GATNNCDA.

Table 1. Results of fivefold CV based on CircR2Disease dataset of best performance.

Test Fold	Accuracy	Precision	Recall	F1-Score	AUC	AUPR
1	0.9346	0.9821	0.9692	0.9368	0.9839	0.9820
2	0.9346	0.9720	0.9769	0.9373	0.9800	0.9706
3	0.9077	0.9305	0.9308	0.9098	0.9497	0.9285
4	0.9308	0.9793	0.9692	0.9333	0.9801	0.9791
5	0.9500	0.9933	0.9615	0.9506	0.9925	0.9932
Average	0.9315	0.9714	0.9615	0.9336	0.9742	0.9707

Table 2. Results of fivefold CV based on cicRNADisease dataset of best performance.

Test Fold	Accuracy	Precision	Recall	F1-Score	AUC	AUPR
1	0.9478	0.9799	1.0000	0.9504	0.9826	0.9794
2	0.9627	0.9944	0.9701	0.9630	0.9938	0.9943
3	0.9776	0.9719	1.0000	0.9781	0.9831	0.9703
4	0.9776	0.9879	0.9851	0.9778	0.9895	0.9877
5	0.9531	0.9921	1.0000	0.9552	0.9922	0.9920
Average	0.9638	0.9852	0.9910	0.9649	0.9882	0.9848

2.3. Comparison with Other Methods

As some methods have been proposed for inferring circRNA–disease association, we compared the performance of GATNNCDA with other state-of-the-art methods by fivefold CV. Some methods used different evaluation criteria or datasets. To compare fairly, we chose nine methods and mainly used CircR2Disease dataset and AUC as the criteria, including DWNN-RLS [26], PWCDA [25], KATZHCDA [27], NCPCDA [39], AE-RF [29], Wang's method, [40], iCircDA-MF [41], GCNCDA [33] and GATCDA [29]. We performed the experiment 50 times, and selected the best performance and the average performance for comparison, denoted as GATNNCDA-best and GATNNCDA-average. The results are shown in Table 3. It can be seen that GATNNCDA is superior to the other nine methods. It is worth noting that the latter two methods are graph neural network based. We found that the data used in GCNCDA and GATCDA are not exactly the same as for us. GCNCDA uses all known circRNA–disease associations in the CircR2Disease dataset, while GATCDA integrates the data with other datasets. However, the AUC value of our method outperforms these methods by a large margin, which demonstrates that GCTNNCDA can effectively and accurately predict underlying disease-related circRNAs.

Table 3. The fivefold CV AUC comparison with the other nine methods based on CircR2Disease dataset.

Models	AUC
DWNN-RLS [26]	0.8854
PWCDA [25]	0.8900
KATZHCDA [27]	0.7936
NCPCDA [39]	0.9201
AE-RF [29]	0.9486
Wang's method [40]	0.8667
iCircDA-MF [41]	0.9178
GCNCDA [31]	0.9090
GATNNCDA [42]	0.9011
GATNNCDA-best	0.9742
GATNNCDA-average	0.9613

2.4. Ablation Study

In this section, we quantitatively evaluated the effect of different components, such as similarity integration, GAT-based feature extraction, and multi-layer NN-based classification, we performed the ablation study by using fivefold CV based on the CircR2Disease dataset. Specially, we defined the variants of GATNNCDA as follows:

- GATNNCDA w/o features: It uses randomly initialized SD and SC as initial node features, instead of integrated similarities.
- GATNNCDA w/o GAT: It removes the GAT from GATNNCDA, and uses the integrated similarities as features and a two-layer NN as a predictor.
- GATNNCDA w/o NN: It uses dot production to calculate the prediction score, instead of a two-layer NN as a predictor.

The results are shown in Figure 3. GATNNCDA w/o features has the lowest values of AUC and AUPR, indicating that the integration similarities as initial node features can greatly improve the performance. GATNNCDA w/o GAT and GATNNCDA w/o NN have about 10% performance degradation. Therefore, our proposed method, GATNNCDA, combines the advantages of these components to obtain the best performance.

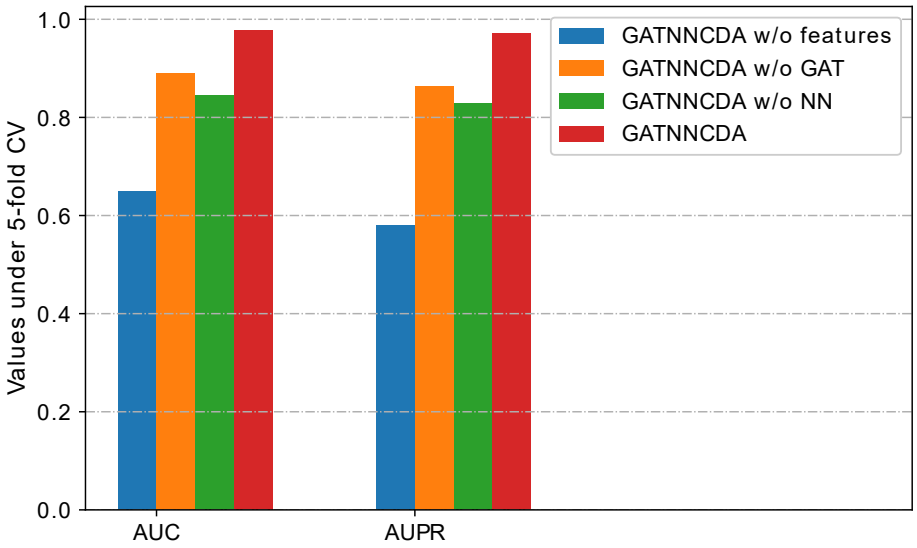

Figure 3. The comparison results of GATNNCDA and its variants based on CircR2Disease dataset.

2.5. Effect of the Parameters

GATNNCDA has several hyper-parameters that also affect the predictive performance. In this section, we performed the experiments to evaluate the effect of the parameters, such as the dimension size of nodes, number of heads in GAT, and the regularization factor, based on CircR2Disease dataset. Figure 4 shows the results of AUC and AUPR under different parameter values.

Recall dimension size of nodes not only affect the similarity parameter matrices *MC*, *MD*, but also impact the input features in the GAT and the NN-based classifier. In our experiment, we chose the values of {8, 16, 32, 64, 128, 256} to test the influence of dimension size. As shown in Figure 4a, we can see that GATNNCDA achieves the lowest AUC and AUPR when the dimension is set to 8, and obtains the best performance at 32. As the dimension increases beyond 32, the performance degrades slightly. The result demonstrate that too small dimensions could lead to under expression of diseases and circRNAs, while too large dimensions may lead to high noise. Therefore, we set 32 as our default dimension.

As reported in a previous study, the deeper GNN can degrade the performance [43]. We set 2 as the default number of the GAT layer. Then, we conducted an experiment on the different number of heads of GAT. Figure 4b shows that GATNNCDA achieves the best AUC at four GAT heads, and the best AUPR at one GAT head. Considering most methods use AUC as a criteria in performance comparison, we finally choose four as the default number of heads of GAT. In addition, we also designed the experiment to evaluate the regularization factor λ. As shown in Figure 4c, GATNNCDA acquires the best AUC and AUPR at $\lambda = 1 \times 10^{-2}$.

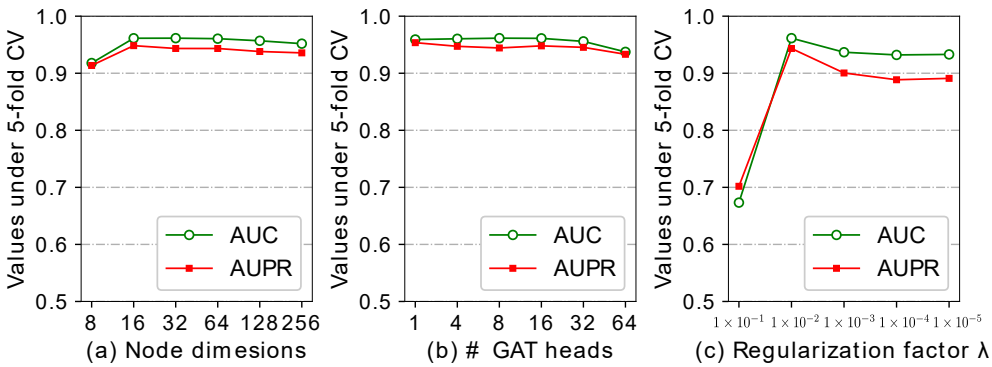

Figure 4. The comparison results of different parameter values based on CircR2Disease dataset.

2.6. Case Studies

To further evaluate the prediction ability of our proposed method, we performed two case studies in this section. We trained GATNNCDA on CircR2Disease dataset [22], and then verified the candidates on circRNADisease [19] and circAtlas v2.0 [24] datasets. The first case study was conducted on breast cancer, which is one of the most common cancers in women. In particular, we constructed the positive samples with all known associations between circRNAs and diseases in the CircR2Disease. Meanwhile, we randomly chose the same number of negative samples from the unknown associations. Based on these training samples, we built the GATNNCDA and calculated the scores between breast cancer and each circRNA. Finally, we selected the top 20 related circRNAs for analysis. As shown in Table 4, 18 of the top 20 are confirmed by the validation datasets. The other two candidates have been verified in the recently published literature.

The second case study is performed on hepatocellular carcinoma. It is the most common form of liver cancer, with a higher incidence in patients with long-term liver diseases [44]. We utilized GATNNCDA to calculate the correlation score with circRNAs and then sorted by descending order. The top 20 hepatocellular carcinoma related circRNAs are listed in Table 5. We can see that 10 of the top 20 are verified by the validation datasets, and the other eight candidates have been conformed in relevant literature, e.g., hsa_circ_0000520 is one of the three circRNAs that showed significantly different expression levels in HCC tissues [14]. Therefore, the unknown associations with high scores are likely to be correlated.

Table 4. Top 20 predicted circRNAs related to Breast cancer based on circR2Disease dataset.

Rank	circRNA	Evidence	Rank	circRNA	Evidence
1	hsa_circ_0007534	*II*	11	hsa_circ_0068033	*I*; *II*
2	hsa_circ_0011946	*II*	12	circamotl1hsa_circ_0004214	*I*; *II*
3	hsa_circ_0093859	*II*	13	hsa_circ_0006528	*I*; *II*
4	circrna-000911	*II*	14	hsa_circ_0002874	*I*; *II*
5	circrna-001283	PMID:29431182	15	hsa_circ_0001667	*I*; *II*
6	circrna-001175	*II*	16	hsa_circ_0085495	*I*; *II*
7	circrna-100438	PMID:29431182	17	hsa_circ_0086241	*I*; *II*
8	hsa_circ_0001982	*I*; *II*	18	hsa_circ_0092276	*I*; *II*
9	hsa_circ_0001785	*I*	19	hsa_circ_0003838	*I*; *II*
10	hsa_circ_0108942	*I*; *II*	20	circvrk1	*I*; *II*

I, *II* denote circRNADisease, circAtlas v2.0.

Table 5. Top 20 predicted circRNAs related to hepatocellular carcinoma based on circR2Disease dataset.

Rank	circRNA	Evidence
1	circc3p1	*II*
2	hsa_circ_0067531	*II*
3	circarsp91hsa_circ_0085154	*II*
4	circmto1hsa_circrna_0007874hsa_circrna_104135	*II*
5	hsa_circ_0005986	*I*;*II*
6	hsa_circrna_100338circsnx27	PMID:28710406
7	hsa_circrna_104075	*I*;*II*
8	hsa_circrna_102049	PMID:28710406
9	circrna_000839	*II*
10	circzkscan1hsa_circ_0001727	*I*;*II*
11	hsa_circ_0004018	*I*;*II*
12	hsa_circ_0005075	*II*
13	hsa_circrna_100571	PMID: 29609527
14	hsa_circrna_400031	PMID:29609527
15	hsa_circrna_102032	PMID: 29609527
16	hsa_circrna_103096	PMID:29609527
17	hsa_circrna_102347	PMID:29609527
18	hsa_circrna_000167hsa_circ_0000518	unknown
19	hsa_circ_0000520	PMID:27258521
20	hsa_circ_0000172	unknown

I, *II* denote circRNADisease, circAtlas v2.0.

3. Materials and Methods

3.1. Known circRNA-Disease Associations

The experimentally verified circRNA–disease association dataset used in this paper is CircR2Disease [22]. We directly downloaded the dataset from the website (http://bioinfo.snnu.edu.cn/CircR2Disease, retrieved 7 June 2021). It contains 739 experimentally validated associations collected from some published studies, and includes 661 circRNAs and 100 diseases. After preprocessing, we obtained 585 circRNAs and 88 diseases. We defined the adjacent matrix $Y \in \mathbb{R}^{N_c \times N_d}$ to denote the known circRNA–disease associations. The element $Y(c_i, d_j)$ is 1 if the association between circRNA c_i and disease d_j has been verified in CircR2Disease. Otherwise, $Y(c_i, d_j)$ is 0. $N_c = 585$ and $N_d = 88$ are the number of circRNAs and diseases.

3.2. Disease Semantic Similarity

We used the Disease Ontology dataset (DO) to calculate the similarity score between disease–disease pairs, which can be download from https://disease-ontology.org (retrieved 7 June 2021). Every disease has a term structure, including a unique ID, name, and the is-a relation with its parents. Given a disease d, we can build a Directed Acyclic Graph (DAG) represented as $DAG_d = (T_d, E_d)$. T_d and E_d denote the nodes and edges in the DAG_d. Based on the assumption that the more shared the nodes in the DAGs between two diseases are, the more similar they are, we can calculate the semantic similarity between disease d_i and d_j using DOSE package, and denote matrix $SS \in \mathbb{R}^{N_d \times N_d}$ as the semantic similarities between diseases.

3.3. circRNA Functional Similarity

As proposed in the previous work for computing functional similarity between miRNAs [45], we assumed that the more similar the diseases connected to two circRNAs, the more similar their functions will be [45]. In particular, we denoted circRNA functional similarity between circRNA c_i and c_j as $CS(c_i, c_j)$. Let D_i and D_j represent the related disease groups that were calculated from the known circRNA–disease associations. Then, we defined the functional similarity between circRNA c_i and c_j as following:

$$FS(c_i, c_j) = \frac{\sum_{d_k \in D_j} S(d_k, D_i) + \sum_{d_l \in D_i} SS(d_l, D_j)}{|D_i| + |D_j|} \quad (1)$$

where $S(d, D) = \max_{d_i \in D}(SS(d, d_i))$ is the disease similarity between disease d and group D. $|D_i|$ and $|D_j|$ are the number of diseases in the group D_i and D_j.

3.4. Gaussian Interaction Profile Kernel Similarity for Disease

Based on the assumption that similar circRNAs are more likely connected to similar diseases [46], we denoted i-row of Y and j-column of Y as the representations of circRNA c_i and disease d_j, and then calculated the we Gaussian interaction profile (GIP) kernel similarities between two circRNAs or diseases as follows:

$$GC(c_i, c_j) = \exp(-\gamma_c \|Y_{i\cdot} - Y_{j\cdot}\|^2) \quad (2)$$

$$GD(d_i, d_j) = \exp(-\gamma_d \|Y_{\cdot i} - Y_{\cdot j}\|^2) \quad (3)$$

where γ_d and γ_c are the kernel bandwidth control parameters, and are defined by the following equations:

$$\gamma_c = \frac{1}{\frac{1}{N_c}\sum_{i=1}^{N_c} \|Y_{i\cdot}\|^2} \quad (4)$$

$$\gamma_d = \frac{1}{\frac{1}{N_d}\sum_{j=1}^{N_d} \|Y_{\cdot j}\|^2} \quad (5)$$

3.5. Integrated Similarities for circRNA and Disease

We observed that the similarity matrices SS and FS are very sparse. Therefore, we integrated GIP similarities to improve the expression of disease similarity and circRNA similarity. The formulas are as follows:

$$SC(c_i, c_j) = \begin{cases} FS(c_i, c_j) & \text{if } FS(c_i, c_j) \neq 0 \\ GC(c_i, c_j) & \text{otherwise} \end{cases} \tag{6}$$

$$SD(d_i, d_j) = \begin{cases} SS(d_i, d_j) & \text{if } SS(d_i, d_j) \neq 0 \\ GD(d_i, d_j) & \text{otherwise} \end{cases} \tag{7}$$

where $SC \in \mathbb{R}^{N_c \times N_c}$ and $SD \in \mathbb{R}^{N_d \times N_d}$ are integrated similarities.

3.6. Feature Extraction Based on Graph Attention Network

Graph attention network (GAT) is a powerful graph-based method whose node can aggregate its neighbor's information by an attention mechanism [47]. In this section, we used GAT in the circRNA–disease graph to learn the rich representations of circRNAs and diseases. We first constructed the circRNA–disease graph based on adjacency matrix Y, and defined it as $G = (V, E)$. $V = \{v_1, v_2, \ldots, v_{N_c+N_d}\}$ are vertices, E represents the edges between circRNA and disease. In particular, edges in the circRNA–disease graph are un-directional, so G can be considered as a bidirectional graph.

As the integrated similarities SC and SD are in different dimension size, we introduced two parameter matrices $MC \in \mathbb{R}^{N_c \times F}$ and $MD \in \mathbb{R}^{N_d \times F}$ to transform SC and SD to the same size, and defined the initial node features in graph G as follows:

$$X = concat(SC \times MC, SD \times MD) \tag{8}$$

where F is the dimension size, and *concat* denotes matrix concatenation. We denoted the input of l-layer of GAT as $H^{(l)} = \{\boldsymbol{h}_1^{(l)}, \boldsymbol{h}_2^{(l)}, \ldots, \boldsymbol{h}_N^{(l)}\}, \boldsymbol{h}_i^{(l)} \in \mathbb{R}^{F^{(l)}}$, and we set $H^{(0)} = X$ as the initial input to GAT. In the circRNA–disease graph, some vertices have no connections with others. To keep the dimensions of GAT output the same as the dimensions of input node features, we set $F^{(l)} = F$. Then, we defined the coefficient between node v_i and the neighborhood v_j by the following formula:

$$e_{ij}^{(l)} = a(\mathbf{W}^{(l)}\boldsymbol{h}_i^{(l)}, \mathbf{W}^{(l)}\boldsymbol{h}_j^{(l)}) \tag{9}$$

where $\mathbf{W}^{(l)}$ is the l-layer shared parameter, and a represents a single-layer neural network with LeakyReLU as the activation function. Similarly, we calculated the coefficients over the neighbor $\mathcal{N}_i$, and normalized the score of node v_j as follows:

$$\alpha_{ij}^{(l)} = softmax_j^{(l)}(e_{ij}^{(l)}) = \frac{\exp(e_{i,j}^{(l)})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik}^{(l)})} \tag{10}$$

For node v_i, the output of l-layer over multi-head attention mechanisms can be defined as follows:

$$\boldsymbol{h}_i^{(l+1)} = \sigma\left(\Big\Vert_{k=1}^{K} \sum_{j \in \mathcal{N}_i} \alpha_{ij}^{(l,k)} \mathbf{W}^{(l,k)} \boldsymbol{h}_j^{(l)}\right) \tag{11}$$

where σ is a nonlinear activation function. K is the number of independent attention heads. $\Vert$ denotes concatenation of K heads except averaging in the last GAT layer. As the L-layer GAT calculation, we obtained the final node features, and defined as $H^{(L+1)} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{N_c}, \boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{N_d}\}$.

3.7. circRNA-Disease Association Prediction

In this section, we constructed a NN classifier to predict the associations between circRNAs and diseases. The k-layer output of the NN classifier can be defined as follows:

$$h^{(k+1)} = \sigma(W^{(k)} \times h^{(k)} + b^{(k)}) \tag{12}$$

where $h^{(0)} = concat(\boldsymbol{c}, \boldsymbol{d})$ is the input to NN classifier, concatenated by the vectors of circRNA c and disease d. σ denotes LeakyReLU activation function. $W^{(k)}$ and $b^{(k)}$ are the parameters of weight and bias in the k-layer of NN classifier. In the last layer (K-layer) of the NN classifier, we can calculate the correlation score as follows:

$$f(c, d) = h^{(K+1)} = \sigma(W^{(K)} \times h^{(K)} + b^{(K)}) \tag{13}$$

where σ is a $sigmoid(\cdot)$ activation function which ensure the score is between 0 and 1. In GATNNCDA, known pairs of circRNA and disease are taken as positive samples, and labeled as 1. However, there are no negative samples in the CircR2Disease; we randomly selected the same numbers of negative samples from the unknown associations, and marked them as 0. The training samples can be denoted as $\mathcal{G}$. Finally, we can define our loss function by the following equation:

$$\mathcal{L} = -\frac{1}{N} \sum_{(c,d) \in \mathcal{G}} (y \log f(c, d) + (1 - y) \log(1 - f(c, d))) + \lambda \|\Theta\|^2 \tag{14}$$

where N is the number of training samples. λ denotes the control factor to the regularization, and Θ is the parameters of our model.

4. Conclusions

Cumulative evidence has shown that circRNAs play an important role in progression of human diseases, and are suitable as promising disease biomarkers for prevention, diagnosis and treatment. As traditional biological identification is very costly and time-consuming, more and more computational methods have been introduced in this field. In this study, we proposed a novel computational method called GATNNCDA for predicting potential circRNA–disease associations. GATNNCDA achieved a better performance than other state-of-the-art methods by combining similarity integration, graph attention network and multi-layer neural network. In particular, we performed fivefold CV for evaluation, and obtained the best performance of AUC of 0.9742, AUPR of 0.9707. The average values of AUC and AUPR for under 50 experiments were 0.9613 and 0.9452. Furthermore, case studies on breast cancer and hepatocellular carcinoma have also demonstrated that GATNNCDA can be a useful tool for predicting potential disease-related circRNAs.

However, GATNNCDA still has some limitations. The initial node features may not be perfect. Recall that similarity integration as initial node representations would affect the final performance. Nonetheless, known interactions between circRNA–disease associations are insufficient. In addition, circRNA functional similarity and GIP similarity may be inaccurate. Therefore, more biological information such as circRNA–miRNA association or circRNA sequence will be used for further study to construct more accurate node features, especially for some unseen circRNAs. Furthermore, the NN-based classifier of GATNNCDA requires negative samples for training, which are rarely reported in the literature. Randomly sampling from the unknown associations in a CircR2Disease dataset would introduce bias. In the future, we will seek a better negative sampling strategy to promote the performance of GATNNCDA.

Author Contributions: C.J. and Z.L. proposed the prediction method; Y.W., C.Z. and J.N. participated in the methodology design; C.J. draft the manuscript; C.Z. and J.N. reviewed and edited the manuscript; C.J. and Z.L. developed the software; Z.L. and Y.W. analyzed the results and revised the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant numbers U19A2064, 61873001, 61872220, and 61802227), and the Natural Science Foundation of Shandong Province (grant number ZR2020KC022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are present within the manuscript or available by request to corresponding author, Cunmei Ji (cunmeiji@126.com).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Memczak, S.; Jens, M.; Elefsinioti, A.; Torti, F.; Krueger, J.; Rybak, A.; Maier, L.; Mackowiak, S.D.; Gregersen, L.H.; Munschauer, M.; et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **2013**, *495*, 333–338. [CrossRef]
2. Meng, S.; Zhou, H.; Feng, Z.; Xu, Z.; Tang, Y.; Li, P.; Wu, M. CircRNA: Functions and properties of a novel potential biomarker for cancer. *Mol. Cancer* **2017**, *16*, 1–8. [CrossRef]
3. Sanger, H.L.; Klotz, G.; Riesner, D.; Gross, H.J.; Kleinschmidt, A.K. Viroids are single stranded covalently closed circular RNA molecules existing as highly base paired rod like structures. *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 3852–3856. [CrossRef]
4. Coca-Prados, M.; Hsu, M.T. Electron microscopic evidence for circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* **1979**, *280*, 339–340.
5. Salzman, J.; Gawad, C.; Wang, P.L.; Lacayo, N.; Brown, P.O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* **2012**, *7*, e30733. [CrossRef]
6. Jeck, W.R.; Sharpless, N.E. Detecting and characterizing circular RNAs. *Nat. Biotechnol.* **2014**, *32*, 453–461. [CrossRef] [PubMed]
7. Chen, L.L. The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nat. Rev. Mol. Cell Biol.* **2020**, *21*, 475–490. [CrossRef]
8. Zheng, Q.; Bao, C.; Guo, W.; Li, S.; Chen, J.; Chen, B.; Luo, Y.; Lyu, D.; Li, Y.; Shi, G.; et al. Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nat. Commun.* **2016**, *7*, 1–13. [CrossRef]
9. Abdelmohsen, K.; Panda, A.C.; Munk, R.; Grammatikakis, I.; Dudekula, D.B.; De, S.; Kim, J.; Noh, J.H.; Kim, K.M.; Martindale, J.L.; et al. Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by CircPABPN1. *RNA Biol.* **2017**, *14*, 361–369. [CrossRef]
10. Kristensen, L.S.; Andersen, M.S.; Stagsted, L.V.; Ebbesen, K.K.; Hansen, T.B.; Kjems, J. The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.* **2019**, *20*, 675–691. [CrossRef]
11. Vo, J.N.; Cieslik, M.; Zhang, Y.; Shukla, S.; Xiao, L.; Wu, Y.M.; Dhanasekaran, S.M.; Engelke, C.G.; Cao, X.; Dan, R.; et al. The landscape of circular RNA in cancer. *Cell* **2020**. *176*, 869–881.e13. [CrossRef]
12. Zeng, K.; Chen, X.; Xu, M.; Liu, X.; Hu, X.; Xu, T.; Sun, H.; Pan, Y.; He, B.; Wang, S. CircHIPK3 promotes colorectal cancer growth and metastasis by sponging miR-7 article. *Cell Death Dis.* **2018**, *9*. [CrossRef]
13. Chen, S.; Li, T.; Zhao, Q.; Xiao, B.; Guo, J. Using circular RNA hsa_circ_0000190 as a new biomarker in the diagnosis of gastric cancer. *Clin. Chim. Acta* **2017**, *466*, 167–171. [CrossRef]
14. Shang, X.; Li, G.; Liu, H.; Li, T.; Liu, J.; Zhao, Q.; Wang, C. Comprehensive circular RNA profiling reveals that hsa-circ-0005075, a new circular RNA biomarker, is involved in hepatocellular crcinoma development. *Medicine* **2016**, *95*, e3811. [CrossRef]
15. Qin, M.; Liu, G.; Huo, X.; Tao, X.; Sun, X.; Ge, Z.; Yang, J.; Fan, J.; Liu, L.; Qin, W. Hsa-circ-0001649: A circular RNA and potential novel biomarker for hepatocellular carcinoma. *Cancer Biomark.* **2016**, *16*, 161–169. [CrossRef]
16. Lukiw, W.J. Circular RNA (circRNA) in Alzheimer's disease (AD). *Front. Genet.* **2013**, *4*, 1–2. [CrossRef]
17. Li, Y.; Fan, H.; Sun, J.; Ni, M.; Zhang, L.; Chen, C.; Hong, X.; Fang, F.; Zhang, W.; Ma, P. Circular RNA expression profile of Alzheimer's disease and its clinical significance as biomarkers for the disease risk and progression. *Int. J. Biochem. Cell Biol.* **2020**, *123*, 105747. [CrossRef]
18. Chen, L.L. The biogenesis and emerging roles of circular RNAs. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 205–211. nrm.2015.32. [CrossRef]
19. Glažar, P.; Papavasileiou, P.; Rajewsky, N. CircBase: A database for circular RNAs. *RNA* **2014**, *20*, 1666–1670. rna.043687.113. [CrossRef]
20. Ghosal, S.; Das, S.; Sen, R.; Basak, P.; Chakrabarti, J. Circ2Traits: A comprehensive database for circular RNA potentially associated with disease and traits. *Front. Genet.* **2013**, *4*, 1–9. [CrossRef]
21. Meng, X.; Hu, D.; Zhang, P.; Chen, Q.; Chen, M. CircFunBase: A database for functional circular RNAs. *Database* **2019**, *2019*, baz003. [CrossRef]
22. Fan, C.; Lei, X.; Fang, Z.; Jiang, Q.; Wu, F.X. CircR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases. *Database* **2018**, *2018*, bay044. [CrossRef]
23. Yao, D.; Zhang, L.; Zheng, M.; Sun, X.; Lu, Y.; Liu, P. Circ2Disease: A manually curated database of experimentally validated circRNAs in human disease. *Sci. Rep.* **2018**, *8*, 1–6. [CrossRef]

24. Wu, W.; Ji, P.; Zhao, F. CircAtlas: An integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol.* **2020**, *21*, 1–14. [CrossRef]
25. Lei, X.; Fang, Z.; Chen, L.; Wu, F.X. Pwcda: Path weighted method for predicting circrna-disease associations. *Int. J. Mol. Sci.* **2018**, *19*, 3410. [CrossRef] [PubMed]
26. Yan, C.; Wang, J.; Wu, F.X. DWNN-RLS: Regularized least squares method for predicting circRNA–disease associations. *BMC Bioinform.* **2018**, *19*, 520. [CrossRef]
27. Fan, C.; Lei, X.; Wu, F.X. Prediction of circRNA–disease associations using KATZ model based on heterogeneous networks. *Int. J. Biol. Sci.* **2018**, *14*, 1950–1959. [CrossRef]
28. Xiao, Q.; Luo, J.; Dai, J. Computational Prediction of Human Disease- Associated circRNAs Based on Manifold Regularization Learning Framework. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 2661–2669. [CrossRef] [PubMed]
29. Deepthi, K.; Jereesh, A.S. Inferring Potential circRNA–disease Associations via Deep Autoencoder-Based Classification. *Mol. Diagn. Ther.* **2021**, *25*, 87–97. [CrossRef]
30. Li, G.; Luo, J.; Wang, D.; Liang, C.; Xiao, Q.; Ding, P.; Chen, H. Potential circRNA–disease association prediction using DeepWalk and network consistency projection. *J. Biomed. Inform.* **2020**, *112*, 103624. [CrossRef]
31. Wang, L.; You, Z.H.; Li, Y.M.; Zheng, K.; Huang, Y.A. GCNCDA: A new method for predicting circRNA–disease associations based on Graph Convolutional Network Algorithm. *PLoS Comput. Biol.* **2020**, *16*, e7568. [CrossRef]
32. Bian, C.; Lei, X.J.; Wu, F.X. GATCDA: Predicting circRNA–disease associations based on graph attention network. *Cancers* **2021**, *13*, 2595. [CrossRef]
33. Lei, X.; Fang, Z.; Guo, L. Predicting circRNA–disease associations based on improved collaboration filtering recommendation system with multiple data. *Front. Genet.* **2019**, *10*, 897. [CrossRef]
34. Le, N.Q.; Do, D.T.; Hung, T.N.; Lam, L.H.; Huynh, T.T.; Nguyen, N.T. A Computational Framework Based on Ensemble Deep Neural Networks for Essential Genes Identification. *Int. J. Mol. Sci.* **2020**, *21*, 9070. [CrossRef]
35. Ho Thanh Lam, L.; Le, N.H.; Van Tuan, L.; Tran Ban, H.; Nguyen Khanh Hung, T.; Nguyen, N.T.; Huu Dang, L.; Le, N.Q. Machine Learning Model for Identifying Antioxidant Proteins Using Features Calculated from Primary Sequences. *Biology* **2020**, *9*, 325. [CrossRef] [PubMed]
36. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703.
37. Fey, M.; Lenssen, J.E. Fast graph representation learning with pytorch geometric. *arXiv* **2019**, arXiv:1903.02428.
38. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, Diego, CA, USA, 7–9 May 2015.
39. Li, G.; Yue, Y.; Liang, C.; Xiao, Q.; Ding, P.; Luo, J. NCPCDA: Network consistency projection for circRNA–disease association prediction. *RSC Adv.* **2019**, *9*, 33222–33228. [CrossRef]
40. Wang, L.; You, Z.H.; Huang, Y.A.; Huang, D.S.; Chan, K.C. An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network. *Bioinformatics* **2020**, *36*, 4038–4046. [CrossRef] [PubMed]
41. Wei, H.; Liu, B. iCircDA-MF: Identification of circRNA–disease associations based on matrix factorization. *Briefings Bioinform.* **2019**, *21*, 1356–1367. [CrossRef]
42. Chen, X.; Xie, D.; Wang, L.; Zhao, Q.; You, Z.H.; Liu, H. BNPMDA: Bipartite network projection for MiRNA–Disease association prediction. *Bioinformatics* **2018**, *34*, 3178–3186. [CrossRef] [PubMed]
43. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017—Conference Track Proceedings, Toulon, France, 24–26 April 2017; pp. 1–14.
44. Llovet, J.M.; Kelley, R.K.; Villanueva, A.; Singal, A.G.; Pikarsky, E.; Roayaie, S.; Lencioni, R.; Koike, K.; Zucman-Rossi, J.; Finn, R.S. Hepatocellular carcinoma. *Nat. Rev. Dis. Prim.* **2021**, *7*. [CrossRef]
45. Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **2010**, *26*, 1644–1650. [CrossRef] [PubMed]
46. van Laarhoven, T.; Nabuurs, S.B.; Marchiori, E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* **2011**, *27*, 3036–3043. [CrossRef]
47. Velicković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.