**Supplementary Materials for:**

**Prediction of circRNA-drug sensitivity using random auto-encoders and multi-layer heterogeneous graph Transformers**

Yinbo Liu 1,#, Xinxin Ren 1,#, Jun Li 1,#, Xiao Chen 2 and Xiaolei Zhu 1,\*

1 School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, Anhui, 230036, China

2 School of Mathemeatics and Statistics, Northwestern Polytechnical University, Xi’an 710072, China.

\* Corresponding authors: [xlzhu\_mdl@hotmail.com](mailto:xlzhu_mdl@hotmail.com)；

# Joint first authors

**Principles of Heterogeneous Graph Transformer**

The single-layer Heterogeneous Graph Transformer architecture for predicting circRNA-drug sensitivity associations is shown in Figure S1.



**Figure S1. The single-layer Heterogeneous Graph Transformer architecture for predicting circRNA-drug sensitivity associations**. (Given a sampled heterogeneous sub-space with target nodeand source node, edgesand their corresponding meta relationsare fed into HGT to acquire a contextualized representation for each node. HGT consists of three components, including *Heterogeneous Mutual Attention, Heterogeneous Message Passing and Target-Specific Aggregation*.)

**Definition 1(Heterogeneous graph) [1]**: A heterogeneous graph is a type of graph data structure where nodes and edges have different types. In this study, we represent a heterogeneous graph as . denotes nodes and denotes edges. represents the node type union, and represents the edge type union.

**Definition 2(Mapping functions for Node types and Edge types) [1]:** We define , and as functions that map node types and edge types, respectively.

**Definition 3(Target node and source node) [1]:** In HGT, a node in is considered as a target node, denoted as . Conversely, a node is identified as a source node, denoted as , if an edge between and in exists, represented as for convenience. When conducting HGT for information aggregation, the embedding of will be updated.

**Definition 4(Node meta relation) [1]:** For a pair of nodes and linked by an edge , the meta relation is denoted as .

Based on the input matrices, we build a bipartite heterogeneous graph, consisting of two types of nodes (circRNA and drug) and one type of edge (circRNA-drug edge). , where denotes all circRNAs, and represents all drugs. represents the edge between and . In the following sections, a node of a circRNA, , is considered as the target node and a node of a drug, , is considered as the source node, for introducing the principle of HGT.

***Multi-head attention mechanism and vector linear mapping***

Let denotes the embedding of the -th HGT layer . The embeddings of the and in the -th layer are denoted as and , respectively. A multi-head attention mechanism is utilized to evenly distribute both and into heads, which facilitates simultaneous consideration of information from diverse embeddings. Different heads can undergo the attention mechanism in parallel, leading to a reduction in computational time. For the -th head in the -th HGT layer, the and is updated from and . The and are the initial embeddings of and , respectively. Three linear projection functions are utilized to transform node embeddings into the -th vector representation. Specifically, the function maps the target node into the -th query vector, with dimension , where represents the dimension of and represents the vector dimension of each head. Similarly, the source node is mapped into the -th key vector and the -th value vector by function and , respectively.

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
|  | (3) |

Distinct linear projections [1] are used for different types of node which allows for the optimal modeling of distribution differences.

***Heterogeneous mutual attention***

The Attention operator [1] is used to assess the significance of each to .

|  |  |
| --- | --- |
|  | (4) |

where, is a concatenation function. The overall attention of and is obtained by concatenating the attention weights among all heads, which are subsequently processed through a softmax function. The attention can be calculated by:

|  |  |
| --- | --- |
|  | (5) |

where is a transformation matrix to model different types of relationships between nodes.is the transposal function and , as a prior tensor, represents the overall significance of each meta relation . This prior tensor serves as an adaptive scaling factor for attention computation, enabling more effective learning. The weights of attention heads are concatenated to form a single attention vector for each node pair. Then, the attention vectors of its neighboring nodes are gathered and an edge-softmax function is applied to normalize their values, ensuring that they sum up to one: .

***Heterogeneous message passing***

In HGT, the information of the source node need to be processed by a Message operator before passing to the target node . The multi-head Message [1] is defined as follows:

|  |  |
| --- | --- |
|  | (6) |

The -th head message for edge is defined as:

|  |  |
| --- | --- |
|  | (7) |

where, the source node was mapped into the -th message vector using a linear projection function : (in Eq.5), then a transformation matrix is used to incorporate the edge dependency.

***Target specific aggregation***

Based on the mutual attention and the message calculated in the above two sections, the information of source nodes can be aggregated to update the embedding of . The aggregation process [1] is defined as follow:

|  |  |
| --- | --- |
|  | (8) |

Then, the updated embedding of the target node needs to be mapped back into its distribution specific to its type. To achieve this, the updated vector undergoes a linear projection , followed by a residual connection.

|  |  |
| --- | --- |
|  | (9) |

After performing the above steps, the -th HGT layer’s output for the target node can be obtained.

**Evaluation metrics**

The performance assessment of the proposed MHGTCDA model includes conducting 5-fold cross-validation (5-CV) and 10-fold cross-validation (10-CV) experiments on the benchmark dataset. In the 5-CV trials, an equal number of negative circRNA-drug sensitivity pairs were randomly chosen to match the quantity of positive samples. These pairs were divided into five equally sized subsets. Each subset alternated as a test set, while the other four subsets were employed for training. This process was repeated five times to ensure reliable results. Note that the labels of examples in the test set were masked during the cross-validation processes. Specifically, only the associations in the training set were input into network to train the model. A similar procedure was followed for the 10-CV experiment. Additionally, a comprehensive set of evaluation metrics, encompassing various aspects such as area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, precision, recall, F1-score, and specificity, was utilized to evaluate the model's performance. Detailed definitions of these metrics can be found in equations (17-21).

|  |  |
| --- | --- |
|  | (10) |
|  | (11) |
|  | (12) |
|  | (13) |
|  | (14) |

In the above equation, TP, TN, FP, and FN represent different elements of a binary classification model's performance:

TP (True Positives): The number of correctly predicted positive instances (i.e., instances that are actually positive and are predicted as positive).

TN (True Negatives): The number of correctly predicted negative instances (i.e., instances that are actually negative and are predicted as negative).

FP (False Positives): The number of instances that are predicted as positive but are actually negative (i.e., instances that are negative but are predicted as positive).

FN (False Negatives): The number of instances that are predicted as negative but are actually positive (i.e., instances that are positive but are predicted as negative).

[1] Hu Z, Dong Y, Wang K, Sun Y. Heterogeneous Graph Transformer. Proceedings of The Web Conference 2020. Taipei, Taiwan: Association for Computing Machinery; 2020. p. 2704–10.