IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 June 2023, accepted 12 July 2023, date of publication 18 July 2023, date of current version 24 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296474

RESEARCH ARTICLE

Progressively Helical Multi-Omics Data Fusion GCN and Its Application in Lung Adenocarcinoma

JUNXUAN ZHU[1], JINHAN ZHANG[1], LIYAN WANG[1], HAO HUANG[1], ZHIBO ZHANG[2], KAI SONG[1,3], AND XIAOFEI ZHANG[2]

[1] School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China
[2] Tianjin Hospital, Tianjin 300211, China
[3] Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin 300072, China

Corresponding authors: Kai Song (ksong@tju.edu.cn) and Xiaofei Zhang (zhang_xiaofei_cool@126.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC3004502, and in part by the Ministry of Science and Technology of the People's Republic of China.

ABSTRACT Compared to single-omics data, utilizing multi-omics data helps to gain a more comprehensive understanding of the occurrence and development of cancer, which emphasizes the necessity of developing efficient multi-omics data fusion approaches. In this study, a novel framework based on graph convolution neural networks with a progressively helical multi-omics data fusion strategy, named phMFGCN, is proposed to effectively integrate multiple omics data. To demonstrate the effectiveness of our framework in addressing the challenges of multi omics data fusion, phMFGCN and other widely-used machine learning methods conducted comparative experiments on predicting gene-gene interactions in lung adenocarcinoma. The results illustrated that phMFGCN outperforms other models with an accuracy of 97.94%. Additionally, 506 new gene-gene interactions predicted by this framework have been validated in databases such as BioGrid. Finally, it was used to perform gene function prediction, and the results were inconsistent with other existing research, for examples: Sam68, DHX9, and HNRNPK were involved in regulating multiple lung adenocarcinoma related pathways simultaneously. All these results demonstrate the universality of phMFGCN for different clinical tasks and it can provide reference target genes or gene-gene interactions for cancer mechanism research and treatment research in clinical practice.

INDEX TERMS Multi-omics data fusion, graph convolution neural network, lung adenocarcinoma, gene-gene interaction.

I. INTRODUCTION

Gene-gene interactions are fundamentally important for understanding the structure and function of genetic pathways in cancer, Alzheimer and other complex diseases [1], [2], [3], [4]. Mapping interactions among genes is expected to make breakthroughs in revealing biological mechanisms of diseases, assessing risk factors in individual disease, and developing treatment strategies for precision medicine [5]. Although single-omics data are quite massive [6], [7], [8], numerous studies have shown that the information provided by single-omics data is relatively monotonous. On the contrary, reasonable exploitations of multi-omics data may provide complementary information and explore complex biological processes more holistically than single-omics data [9], [10], [11], [12], [13]. Therefore, the development of novel and effective models that can extract and aggregate meaningful information from heterogeneous datasets has getting increasing attention. Unfortunately, the integration and utilization of multi-omics data face the following challenges [14], [15]:

1) The relationships of multi-omics features in biological processes are usually nonlinear and complex, which requires the integration models to have strong ability to extract such features and to capture these relationships.
2) Different types of omics features have different contributions to the target results. It requires integration models to have the ability to pay different attention on these different contributions accordingly.

The associate editor coordinating the review of this manuscript and approving it for publication was Giacomo Fiumara.

3) Multi-omics data are often high-dimensional, complex, and heterogeneous, which makes it difficult for the model to extract valid knowledge.

Comparatively higher ratios of data noise and label loss also bring challenges. To overcome these challenges, efforts have been made to develop different kinds of algorithms for different biomedical downstream tasks.

- **Similarity-based methods**: including SNF [16], Multiplex Fusion Algorithm [17], NEMO [18] and so on. They firstly construct a patient similarity network as an integration basis for each type of omics data and then update the network by network iteration nonlinearly. As an improved method, Multiplex Fusion Algorithm does distinguish the contribution of different types of omics data. However, in the iterative steps, this type of methods only focuses on similarity matrices derived from different omics data and does not explicitly utilize node features matrices, which causes their inability to accurately capture the relationships among multiple omics data.
- **Network propagation methods**: These methods make outstanding contributions to the integration of multi-omics data. For example, Matthew et al. used protein-protein interaction network to integrate the mutation data and gene expression data on sample level by network propagation [19]. Aix et al. proposed a label propagation-based method to combine the known biological interaction network with gene expression signatures [20]. Although the network propagation approaches can integrate multiple data types, they also use multi-dimensional node features in an indirect way, which leads to their inadequate utilization of the information of the multi-omics data.
- **Manifold learning and convolutional neural networks (CNNs) methods**: This type of methods use t-distributed stochastic neighbor embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), etc. to create a low dimension representation (i.e. in a 2d dimensional map) for the omics and integrate them by CNNs [21], [22]. Li et al. created a gene similarity network map for each omic using t-SNE before being merged into the residual neural network classification model [23]. Bashier et al. exploited UMAP to embed multi-omics data to a lower dimension by creating 2d dimensional RGB images, then the images were used with CNNs to predict the Gleason score levels of prostate cancer patients and the tumor stage in breast cancer patients [24]. This type of methods focuses only on feature information rather than structure information of the data, it leads to their incomplete utilization of information.
- **Graph convolution neural network (GCN)-based methods**: They can explicitly utilize both graph structure and node features, which is why their performance is generally better than the above two types of methods. For instance, Schulte-Sasse et al. introduced EMOGI to carry out cancer gene prediction task by concatenating multi-omics pan-cancer data to a gene feature matrix as input of GCN [13]. Ma et al. firstly inputted each type of data into GCN, then concatenated the outputs to an integrated embedding matrix [14]. Although these are novel and effective data combining methods, the data integration strategies they adopt are simple, which may vulnerable to lower signal-to-noise ratio in any data type [16], [25]. To avoid this and better capture the relationships between different omics types, GCN-Cox composes the multiple omics data of each patient into a feature matrix, and executes dimensionality reduction on each patient feature matrix by GCN, then combine the reduced dimension results into a new embedding matrix [26]. Besides, MOGONET [27], MOGCN [28], and pDenseGCN [29] also discarded the simple usage of concatenation operation and proposed new integration strategies, but their strategies only focus on the fusion of multi-omics similarity matrix (topological structure) but ignore the fusion of node feature data, which makes themselves hardly overcome all above-mentioned challenges.

In this paper, inspired by the iterative update steps of SNF and advantages of GCN-based models, a new GCN-based framework, named as progressively helical multi-omics data fusion GCN (phMFGCN), was proposed to integrate multi-omics data by focusing on the fusion of both topological structure and node features, which can overcome all above mentioned challenges once for all. Additionally, our framework decouples and lightens GCN layers, which alleviates the over-smoothing issue and enables the model to run on a comparatively bigger model depth.

On the other hand, as a cancer type with high incidence and mortality, LUAD has been widely concerned. The multi-omics sequencing technologies and database related to lung adenocarcinoma are very mature, which provides commendable data source for our study [15]. However, only a part of the gene-gene interactions related to LUAD has been confirmed so far. Numerous gene interactions are unknown and difficult to verify. Predicting new possible gene interactions can be regarded as a binary edge classification task, we strive to complete this task by utilizing LUAD-related multi-omics data with our framework. Experiments about gene-gene interactions prediction and gene function prediction in LUAD proves that our framework is not only potential to address the previous challenges together but also outperforms other machine learning methods.

II. MATERIALS AND METHODS

A. LUAD DATASETS

In this paper, LUAD data mainly includes two parts: multi-omics datasets and label datasets.

1) MULTI-OMICS DATASETS

downloaded from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/repository). The Gene expression (GE) data, DNA methylation (ME) data, and somatic mutation (MU) data of TCGA-LUAD patients were used in our study, which were collected via the Illumina HiSeq 2000 RNA Sequencing Version 2 platform, Simple Nucleotide Variation Data of the whole exome sequencing measured with MuTect2 Variant Calling Pipeline, and Human Methylation 450 platform, respectively.

2) LABEL DATASETS

In gene-gene interaction prediction task, the gene interaction dataset related to LUAD was downloaded from GRNdb (http://www.grndb.com/) [30], which is a continuously updated database containing a wide range of gene-gene interactions. In functional gene prediction task, functional gene sets were download from Gene Set Enrichment Analysis (GSEA, http://www.gsea-msigdb.org/gsea/index.jsp) [31]. We used three LUAD-related signaling pathways 'KEGG MAPK SIGNALING PATHWAY', 'KEGG P53 SIGNALING PATHWAY', and 'KEGG CELL CYCLE' as functional gene sets [32], [33].

To ensure the validity of data and facilitate its subsequent use, we referred to the annotation file to align GE, ME and MU data according to gene symbols. After data preprocessing, 9449 common genes in 449 common samples among GE, MU, ME, and GRNdb data were obtained. The raw feature matrix of MU is obtained with the TTZ feature extracting algorithm proposed by us previously [34].

B. PRELIMINARIES

Gene-gene interactions can be denoted as a graph or network $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ is a set of nodes (i.e. genes) $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_\mathrm{n}\}$, and $\boldsymbol{E}$ is a set of edges or links (i.e. interactions between genes). The graph can be represented by an adjacency matrix $\mathbf{A}$. If there exists interaction between gene i and gene j, which is shown in the graph as an edge connecting node v_i and v_j then $A_{i,j} = 1$, and $A_{i,j} = 0$ otherwise.

GCN-based model is suitable for graph-based gene-gene interaction prediction tasks. The implementation of GCN relies on the structure neighborhood information which describes the topological structure of a graph. With the structure neighborhood information, GCN layer updates each node's information through message transmission [35]. Its inputs are the normalized feature matrix $\mathbf{X}_n$ and the adjacency matrix $\mathbf{A}_n$. The process of a regular GCN can be represented as $f(\mathbf{X}_n, \mathbf{A}_n)$:

$$f(\mathbf{X}_n, \mathbf{A}_n) = \sigma\left(\tilde{\mathbf{D}}_n^{-\frac{1}{2}} \tilde{\mathbf{A}}_n \tilde{\mathbf{D}}_n^{-\frac{1}{2}} \mathbf{X}_n \mathbf{W}_n\right) \tag{1}$$

where $\tilde{\mathbf{A}}_n = \mathbf{A}_n + \mathbf{I}_n$ is $\mathbf{A_n}$ with added self-loop and $\mathbf{I}_n$ is an identity matrix; $\tilde{D}_{i,i} = \sum_j \tilde{A}_{i,j}$ is the diagonal degree matrix of $\tilde{\mathbf{A}}_n$; $\mathbf{W}_n$ is a trainable parameter matrix; σ is a non-linear activation function. By selecting an appropriate number of GCN layers, an embedding aggregated feature information of the central node, its neighborhood nodes, and the local topology structure can be obtained.

C. THE FRAMEWORK OF PROGRESSIVELY HELICAL MULTI-OMICS DATA FUSION

Inspired by the iteratively and non-linearly updating process of SNF [16], we proposed a GCN-based framework with progressively helical multi-omics data fusion strategy. Considering about the fact that the fusion iteration under multi-omics makes the model bulky and causes a huge parameter amount, we decouple the framework by splitting the feature transformation and neighborhood aggregation to ensure the lightweight of the whole framework.

The whole framework consists of five steps: 1) construct an initial graph; 2) decouple; 3) integrate multi-omics data by progressively helical multi-omics data fusion strategy; 4) generate edge embedding vectors; 5) calculate the probability of edges. The framework and the overall algorithm are shown in Fig. 1 and Algorithm 1, respectively.

1) CONSTRUCT AN INITIAL GRAPH

An initial graph (adjacency matrix) is required to be the input of our framework. Because of the powerful learning ability of phMFGCN provided by the following steps, it doesn't require a very precise one. Therefore, methods even as simple as Pearson correlation or Mutual information can be used to determine this initial graph.

In this study, Pearson correlation matrix is applied to construct the initial graph. Due to the limitation of Pearson correlation, the initial graph is unavoidably noisy, correspondingly, a threshold is introduced as a simple filter and the adjacency matrix element of n-th omics data is defined as:

$$An, (i, j) = \begin{cases} 1 & if\ |Pn, (i, j)| \geq \theta n \\ 0 & otherwise \end{cases} \quad (n = 1, 2, \ldots N) \tag{2}$$

where $Pn, (i, j)$ is the element of Pearson correlation matrix obtained from the omics data n. $An, (i, j)$ is the element of the initial graph corresponding to $\mathbf{P}_n$. N is the total number of omics types. θn is the threshold.

2) DECOUPLE

Let $\mathbf{x}_i$ be the feature vector of node $v_i \in$ V. From the perspective of nodes, (1) can be decomposed into neighborhood aggregation and feature transformation:

$$\left(\frac{1}{\tilde{D}_{i,i}} \mathbf{x}_i^{(l)} + \sum\nolimits_{j \in N1(i)} \frac{1}{\sqrt{\tilde{D}_{i,i} \tilde{D}_{j,j}}} \mathbf{x}_j^{(l)}\right)$$

(neighborhood aggregation)

$$\sigma\left(\frac{1}{\tilde{D}_{i,i}} W_i^{(l)} \mathbf{x}_i^{(l)} + \sum\nolimits_{j \in N1(i)} \frac{1}{\sqrt{\tilde{D}_{i,i} \tilde{D}_{j,j}}} W_j^{(l)} \mathbf{x}_j^{(l)}\right)$$

(feature transformation) (3)

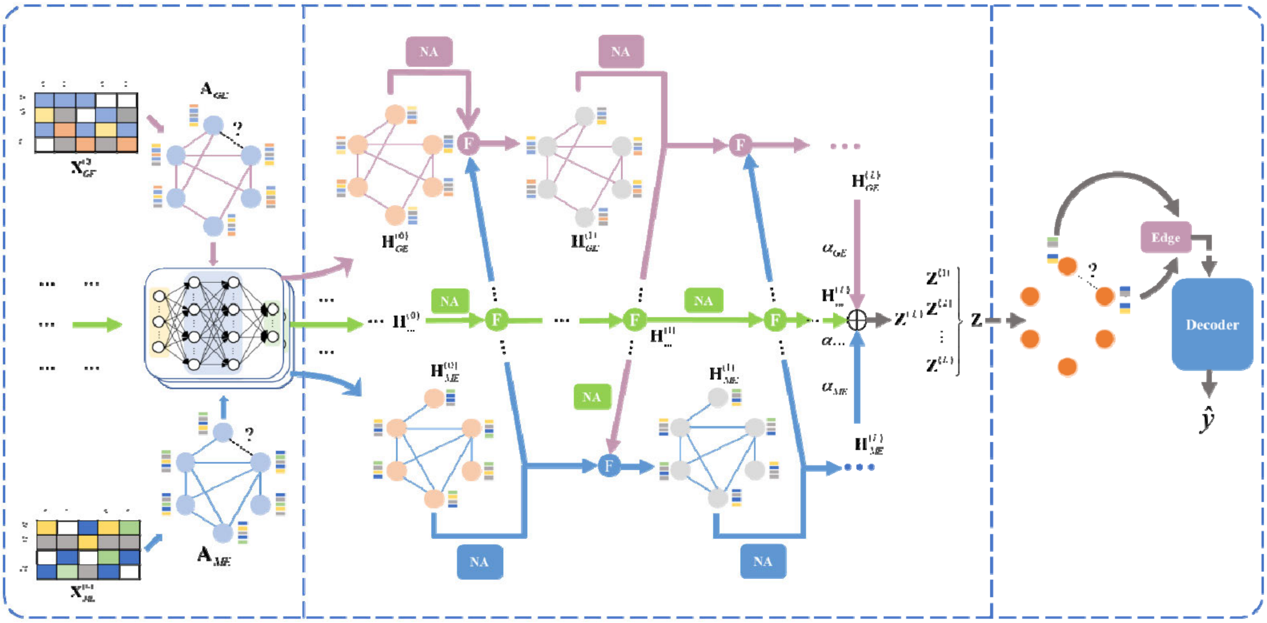

FIGURE 1. Framework of progressively helical multi-omics data fusion GCN. (Left) Constructing an initial graph and feature transformation with MLP. (Middle) Integrating multi-omics data by progressively helical multi-omics data fusion strategy. (Right) Generating edge-embedding vectors and calculating the probabilities of edges. The whole framework consists of five steps: 1) construct an initial graph; 2) decouple; 3) integrate multi-omics data by progressively helical multi-omics data fusion strategy; 4) generate edge embedding vectors; 5) calculate the probability of edges. 1) and 2) steps are shown in the left part, 3) and 4) are shown in the middle part, and 5) is shown in the right part of this figure.

Although GCN-based models have numerous advantages, the entanglement of neighborhood aggregation and feature transformation intensify with the increasing of the model depth, which is more common in the case of multi-omics integration and adversely affects the performance and robustness of models [36], [37]. Additionally, the number of parameter matrices increases dramatically with the depth of the GCN-based model, which is not conducive to the further expansion of the model and even causes the over-smoothing problem [38].

Inspired by the research of Liu et al. [38], in our framework, we decouple the original CGN layer into feature transformation and neighborhood aggregation and swap their order to avoid the entanglement.

In the feature transformation section, the original GCN requires multi-step nonlinear transformation of node information through multiple parameter matrices. An MLP is used to approximate this process. It reduces the amount of parameter matrices and facilitate the deep expansion of the framework. The equation is shown as follows:

$$\mathbf{H}_i^{(0)} = MLP(\mathbf{X}_i) \tag{4}$$

where $\mathbf{X}_i$ is preprocessed feature matrix of omics i, and $\mathbf{H}_i^{(0)}$ is the initial node embedding matrix of omics i.

In the neighborhood aggregation section, the setting of original GCN is followed. For omics i, the single-omics neighborhood aggregation equation is shown as follows:

$$\mathbf{NA}_i^{(l)} = \tilde{\mathbf{L}}_i \mathbf{H}_i^{(l)} \tag{5}$$

where $\tilde{\mathbf{L}}_i = \tilde{\mathbf{D}}_i^{-\frac{1}{2}} \tilde{\mathbf{A}}_i \tilde{\mathbf{D}}_i^{-\frac{1}{2}}$ is symmetric normalized Laplacian matrices.

Decoupling makes the model lighter, which enables it to easier to be extended to deeper layers, and avoid prematurely triggering the over-smoothing problem.

3) INTEGRATE MULTI-OMICS DATA WITH PROGRESSIVELY HELICAL MULTI-OMICS DATA FUSION STRATEGY

Our proposed fusion strategy consists of: ① multi-omics feature fusion; ② multi-omics contribution allocation; ③ node embedding matrix aggregation.

① Multi-omics feature fusion

The multi-omics feature fusion is operated within the network layers and advances in a progressive helical way to ensure the fully mutual integration of different omics data. Take GE, ME, and MU data as examples, the fusion strategy is shown as follows:

$$\begin{aligned}
\mathbf{H}_{GE}^{(l+1)} &= \sigma(Fusion(\tilde{\mathbf{L}}_{GE}\mathbf{H}_{GE}^{(l)}, \tilde{\mathbf{L}}_{ME}\mathbf{H}_{ME}^{(l)}, \tilde{\mathbf{L}}_{MU}\mathbf{H}_{MU}^{(l)}))\\
\mathbf{H}_{ME}^{(l+1)} &= \sigma(Fusion(\tilde{\mathbf{L}}_{GE}\mathbf{H}_{GE}^{(l+1)}, \tilde{\mathbf{L}}_{ME}\mathbf{H}_{ME}^{(l)}, \tilde{\mathbf{L}}_{MU}\mathbf{H}_{MU}^{(l)}))\\
\mathbf{H}_{MU}^{(l+1)} &= \sigma(Fusion(\tilde{\mathbf{L}}_{GE}\mathbf{H}_{GE}^{(l+1)}, \tilde{\mathbf{L}}_{ME}\mathbf{H}_{ME}^{(l+1)}, \tilde{\mathbf{L}}_{MU}\mathbf{H}_{MU}^{(l)}))
\end{aligned} \tag{6}$$

Fusion() is a function to mix different omics data after each single-omics neighborhood aggregation. According to different downstream tasks and performance requirements, the summation, concatenation, max-pooling, and other operators can be selected as *Fusion*(). In our study, considering the balance between computational complexity and model performance, summation is chosen as the fusion function.

After fusion, different omics data will update each other through progressive iteration.

This fusion strategy consists of linear or non-linear features fusion, nonlinear feature transformation, and progressively iterative update among multi-omics data, which simultaneously integrates multi-omics feature matrices and topological structure and is potential to capture complex relationships among different omics data. The equations for the feature fusion about more omics types are shown in Algorithm 1.

② **Multi-omics contribution allocation**

After each time of fusion, the omics node-embedding matrices will be updated. To make better use of each omics node-embedding matrix, weight is assigned to each type of omics. For the result of iteration L, we have:

$$\mathbf{Z}^{(L)} = \alpha_{GE}\mathbf{H}_{GE}^{(L)} + \alpha_{ME}\mathbf{H}_{ME}^{(L)} + \alpha_{MU}\mathbf{H}_{MU}^{(L)} \tag{7}$$

where α_i is the weight coefficient of node-embedding matrix of omics i. It is a hyper-parameter used to measure the contribution of the corresponding omics features. For different omics, the sum of these weight coefficients is 1. $\mathbf{Z}^{(L)}$ is the integrated node-embedding matrix obtained after the L-th iteration.

③ **Integrated node-embedding matrix aggregation**

Multiple integrated node-embedding matrices are obtained after L times iteration, which can be represented as $\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(L)}\}$. According to the previous description, the node-embedding matrix obtained from the l-th iteration mainly reflects the aggregation results of the l-hop neighborhood of each central node. To fully exploit the information of each hop neighborhood of the central node, we concatenate the results of each layer to obtain a single embedding matrix, which contains more useful information than only use the last layer result. The concatenation result is showed as follows:

$$\mathbf{Z} = [\mathbf{Z}^{(1)}||\mathbf{Z}^{(2)}||\ldots||\mathbf{Z}^{(\mathrm{L})}] \tag{8}$$

We hence homeopathically get the final embedding vector of each node from $\mathbf{Z}$.

4) GENERATE EDGE EMBEDDING VECTORS

After obtaining a pair of target nodes v_i and v_j, with their corresponding node embedding vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, an edge generation strategy is constructed to determine the edge existence of the target node pairs:

$$\mathbf{r}_{i,j} = Edge(\mathbf{z}_i, \mathbf{z}_j) \tag{9}$$

where $Edge(\cdot, \cdot)$ is the function to obtain the edge vector between target node pair v_i and v_j (e.g. hadamarD product, summation, concatenation and some other operators). $\mathbf{r}_{i,j}$ is the edge embedding vector of the edge between the target node pair.

5) CALCULATE THE PROBABILITY OF EDGES

After obtained the edge embedding vector between two target nodes, a decoder is defined to calculate the probability of the existence of the corresponding edge:

$$\hat{y}_{i,j} = \sigma(\mathbf{W}\mathbf{r}_{i,j} + \mathbf{b}) \tag{10}$$

where $\mathbf{W}$ is a parameter matrix, and $\mathbf{b}$ is a bias. Because the existence of the edge (i.e. gene-gene interaction prediction) is a binary classification task, sigmoid function is chosen as the activation function σ.

For model training, the binary cross-entropy loss function is used to optimize the model parameters:

$$\mathrm{L} = \sum_{(i,j)\in \mathrm{V}\times\mathrm{V}} -y_{i,j}\log\hat{y}_{i,j} - (1-y_{i,j})\log(1-\hat{y}_{i,j}) \tag{11}$$

where $y_{i,j}$ is the true label for the target nodes v_i and v_j that are sampled during training via mini-batching.

D. EVALUATION METRICS

Accuracy (ACC), precision (PRC), recall, AUROC and AUPRC were used to fully evaluate all different methods. They are described in **Supplementary Note 1**.

E. EXPERIMENT AND PARAMETER SETTINGS

In the experiments, the edges in GRNdb were used as positive samples. For negative edge set, we randomly selected node pairs to construct negative edges and ensured that the intersection of positive and negative edge set is empty. As a result, we obtained positive and negative edges of the same number to compose the label set with 99052 edges. For all label edges, we selected different randomized seeds 10 times to reduce the randomness of the results, and for each dataset partition, we randomly sampled 60% of them as the training data, 20% as the validation data and the rest 20% as the test data. The average value of each type of evaluation was used as the final result.

All algorithms were run on a sever with an AMD Ryzen Threadripper 3990X 64-CORE 2.90 GHz CPU, a 192GB RAM and a NVIDIA TITAN RTX GPU. The hyper-parameter search spaces are: the learning rate of Adam optimizer in [0.01, 0.005, 0.001], dropout rate in [0.1, 0.2, ..., 0.9], number of layers in [1, 2, ...,10], mini-batch size in [64, 128, 256, 512, 1024, 2048], multi-omics weighted coefficients in [0.1, 0.2, ..., 0.9]. Each model was trained with grid search to determine the optimal parameters. The optimal parameters of our framework are shown in **Supplementary Note 2**.

III. RESULTS AND DISCUSSION

There are three group experiments in this section: A) the verification of the performance of phMFGCN compared with other methods including results of gene-gene interaction prediction task and robustness to sample size; B) investigation of phMFGCN including effect of different omics data, effect of hyper-parameter α_i, ablation study, and effect of model depth; C) the generalized applications of phMFGCN to predict novel gene-gene interactions and gene functions.

Algorithm 1 The Algorithm of phMFGCN Framework

```
Input: adjacency matrices with added self-loop
    $\tilde{\mathbf{A}}_{GE}, \tilde{\mathbf{A}}_{ME}, \ldots, \tilde{\mathbf{A}}_{MU}$
normalized node feature matrices
    $\mathbf{X}_{GE}, \mathbf{X}_{ME}, \ldots, \mathbf{X}_{MU}$
Feature transformation with MLP network:
    $\mathbf{H}_{GE}^{(0)} \leftarrow$ Feature Transformation($\mathbf{X}_{GE}$) via (4)
    $\mathbf{H}_{ME}^{(0)} \leftarrow$ Feature Transformation($\mathbf{X}_{ME}$) via (4)
    $\ldots$
    $\mathbf{H}_{MU}^{(0)} \leftarrow$ Feature Transformation($\mathbf{X}_{MU}$) via (4)
Calculate the normalized graph Laplacian matrices:
    $\tilde{\mathbf{L}}_{GE} = \text{LaplacianNormalize}\left(\tilde{\mathbf{A}}_{GE}\right)$
    $\tilde{\mathbf{L}}_{ME} = \text{LaplacianNormalize}\left(\tilde{\mathbf{A}}_{ME}\right)$
    $\ldots$
    $\tilde{\mathbf{L}}_{MU} = \text{LaplacianNormalize}\left(\tilde{\mathbf{A}}_{MU}\right)$
Neighborhood aggregation and progressively helical multi-omics data fusion strategy:
For $l = 0, 1, 2, 3, \ldots, L$ do:
    $\mathbf{H}_{GE}^{(l+1)} \leftarrow$ FeatureFusion($\tilde{\mathbf{L}}_{GE}\mathbf{H}_{GE}^{(l)}, \tilde{\mathbf{L}}_{ME}\mathbf{H}_{ME}^{(l)}, \ldots, \tilde{\mathbf{L}}_{MU}\mathbf{H}_{MU}^{(l)}$) via (6)
    $\mathbf{H}_{ME}^{(l+1)} \leftarrow$ FeatureFusion($\tilde{\mathbf{L}}_{GE}\mathbf{H}_{GE}^{(l+1)}, \tilde{\mathbf{L}}_{ME}\mathbf{H}_{ME}^{(l)}, \ldots, \tilde{\mathbf{L}}_{MU}\mathbf{H}_{MU}^{(l)}$) via (6)
    $\ldots$
    $\mathbf{H}_{MU}^{(l+1)} \leftarrow$ FeatureFusion($\tilde{\mathbf{L}}_{GE}\mathbf{H}_{GE}^{(l+1)}, \tilde{\mathbf{L}}_{ME}\mathbf{H}_{ME}^{(l+1)}, \ldots, \tilde{\mathbf{L}}_{MU}\mathbf{H}_{MU}^{(l)}$) via (6)
    $\mathbf{Z}^{(l+1)} = \text{ContributionAllocation}\left(\mathbf{H}_{GE}^{(l+1)}, \ldots, \mathbf{H}_{MU}^{(l+1)}\right)$ via (7)
    $\mathbf{Z} = \text{concat}\left(\mathbf{Z}, \mathbf{Z}^{(l+1)}\right)$ via (8)
Construct the edge embedding vector between target nodes $v_i$ and $v_j$:
    $\mathbf{r}_{i,j} = \text{EdgeGeneration}(\mathbf{z}_i, \mathbf{z}_j)$via (9)
Calculate the probability of the existence of edges between $v_i$ and $v_j$:
    $\hat{y}_{i,j} = \text{Decoder}(\mathbf{r}_{i,j})$via (10)
Calculate the loss L with the probability $\hat{y}_{i,j}$ and the true label y, then update the model parameters via gradient descent algorithm.
    $\mathrm{L} = \text{LossFunction}\left(y_{i,j}, \hat{y}_{i,j}\right)$ via (11)
```

A. THE VERIFICATION OF THE PERFORMANCE OF phMFGCN

1) RESULTS OF GENE-GENE INTERACTION PREDICTION TASK

In this part, we aim to compare phMFGCN with several types of representative machine learning methods using three-omics data (GE, ME, and MU): XGBoost-AD [39] (eXtreme Gradient Boosting and autoencoder), Node2vec, MVGCN [14] and GCN-Cox [26]. These models are implemented with Pytorch [40]. ACC, PRC, Recall, AUROC and AUPRC are used to evaluate the performance of gene-gene interaction prediction of these methods. We tuned hyper-parameters for all models individually, and all performance data were recorded under the most optimal parameters of the models. The results of different methods are shown in Table 1.

According to the results listed in Table 1, because Node2vec only utilizes graph structure information and XGBoost-AD only exploits node features, they are two methods with the worst performance, which indicates that model using single data type (graph structure or node features) can only obtain limited performance. For the three GCN-based methods, however, because of combining graph topology and node features simultaneously, their performances are generally better than the first two methods.

Three GCN-based models differ in the way of fusion multi-omics data, which determines their ability to capture the complex relationship among different omics data. MVGCN uses a regular fusion strategy whose embedding matrices of each omics data are derived by basic GCN and then are concatenated to an integrated matrix; GCN-Cox generates a graph for each sample, and the node feature is formed by splicing each omics data of the sample. Then the omics information is fused together through GCN, which abandons concatenation but only focus on feature matrix fusion. For phMFGCN, the fusion of omics information is implemented through progressively helical fusion strategy within GCN layers, so that the fusion of topological structure and feature matrices can be considered simultaneously. The results show that phMFGCN performs best in all performance metrics, which proves that our fusion strategy is effective and potential to address the challenge of extracting features and catching the underlying multi-omics data association.

2) THE ROBUSTNESS TO SAMPLE SIZE

In this part, aiming to explore the performance of our framework in bad data quality (e.g. small sample size), we compared the performance of phMFGCN and other GCN-based models with different training sample sizes. Each model was trained with [30%, 40%, 50%, . . . , 90%] of the original training set. In this experiment, ACC and AUROC were used as the performance evaluations. The results are shown in Fig. 2.

It is obvious that the performance of all types of models decreases with the decreases of sample size. However, phMFGCN outperforms other methods and the test ACC and AUROC decrease more slowly than other models with any training percentage. When training percentage >= 50%, the ACC and AUROC values of phMFGCN almost do not decrease while the ACC and AUROC values of both GCN-Cox and MVGCN show a significant decline (decreased by 7.36% and 5.16% compared to the highest ACC, decreased by 7.47% and 6.69% compared to the highest AUROC). When training percentage <= 40%, the ACC and AUROC of phMFGCN are still above 90% ($ACC = 90.95\%$ and $AUROC = 92.29\%$ when training percentage = 30%), whereas the ACC and AUROC values of GCN-Cox and MVGCN are all lower than 87% ($ACC_{COX} = 84.20\%$ and $ACC_{MVGCN} = 79.41\%$,

TABLE 1. Comparison with other machine learning methods.

Method	ACC(%)	PRC(%)	Recall(%)	AUROC(%)	AUPRC(%)
XGBoost-AD	94.88±1.21	95.53±1.21	94.16±1.25	94.88±1.21	92.89±1.68
Node2vec	89.97±0.15	91.21±0.30	89.03±0.18	89.99±0.15	86.83±0.29
MVGCN	95.66±0.31	96.64±0.49	94.71±1.05	98.76±0.05	98.24±0.10
GCN-Cox	97.57±0.13	97.02±0.22	98.15±0.45	99.30±0.02	99.05±0.04
phMFGCN	**97.94±0.08**	**97.12±0.21**	**99.05±0.17**	**99.39±0.02**	**99.08±0.06**

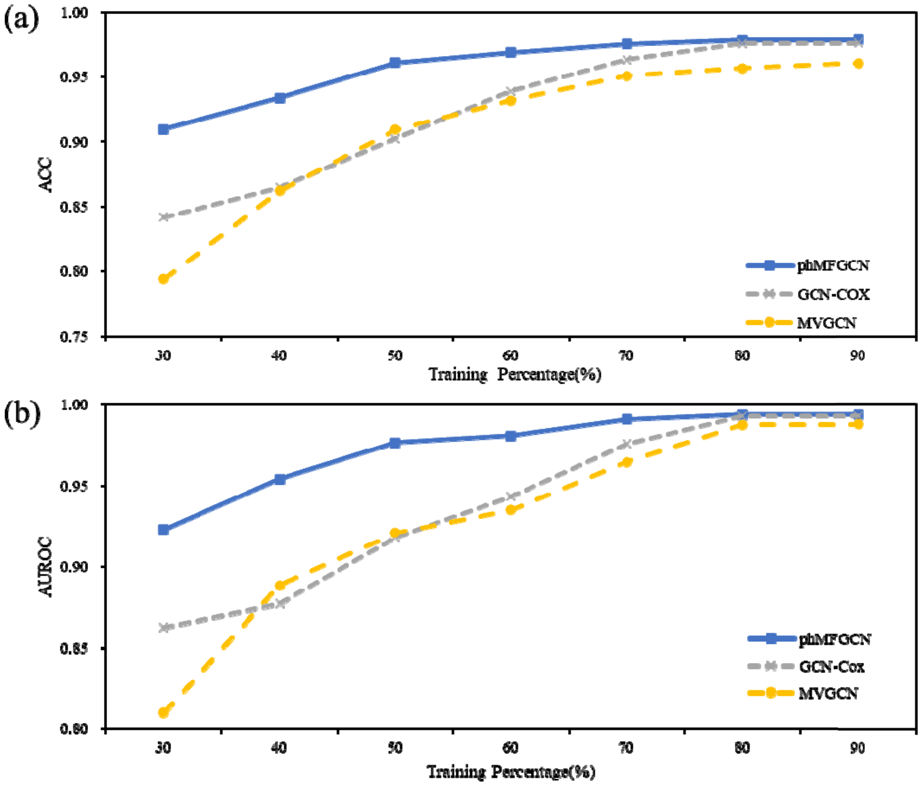

FIGURE 2. (a) The ACC comparison on phMFGCN and other GCN-based models in different training percentages (%). (b) The AUROC comparison on phMFGCN and other GCN-based models in different training percentages (%). phMFGCN outperforms other methods and the test ACC and AUROC decrease more slowly than other models with any training percentage.

$AUROC_{COX} = 86.29\%$ and $AUROC_{MVGCN} = 80.98\%$ when training percentage = 30%, respectively). The above results indicate that our framework is not sensitive to sample size and potential to overcome the problem of bad data quality.

B. THE INVESTIGATION ON phMFGCN

1) THE EFFECT OF DIFFERENT OMICS DATA

To explore the impact of different omics data or omics combinations on model performance, we carried out the following experiments:

①Basic GCN with only GE data.
②Basic GCN with only ME data.
③Basic GCN with only MU data.
④phMFGCN with GE and ME data.
⑤phMFGCN with GE and MU data.
⑥phMFGCN with ME and MU data.
⑦phMFGCN with GE, ME, and MU data.

It can be found that the performance of phMFGCN increases with the number of used omics data types. When using single-omics data, the best ACC of our framework is

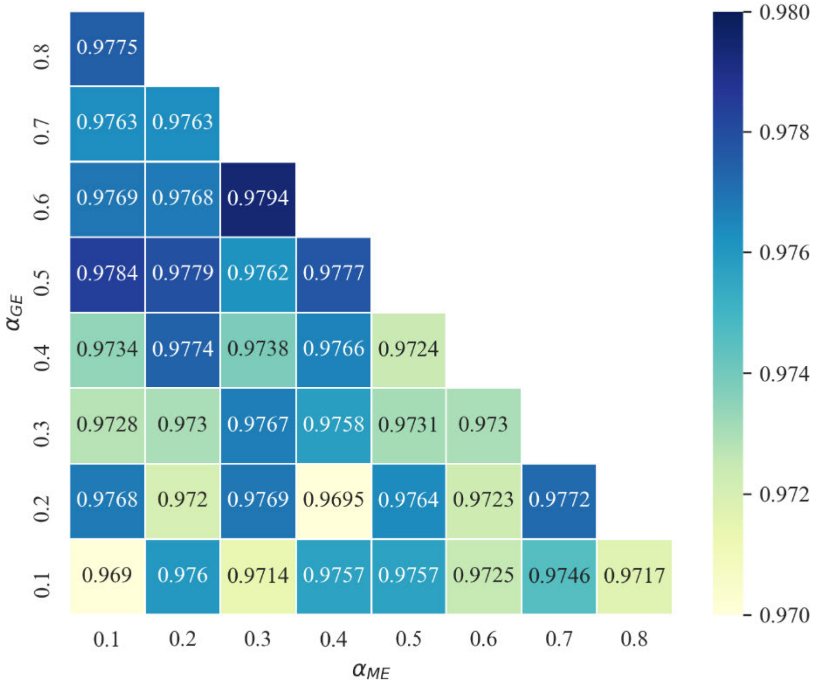

FIGURE 3. The effect of hyper-parameter α_i in phMFGCN. With the increase of α_{GE}, the test ACC of phMFGCN basically shows an upward trend. When $\alpha_{GE} = 0.6$, $\alpha_{ME} = 0.3$, and $\alpha_{MU} = 0.1$, phMFGCN obtains the best ACC 0.9794.

94.09%, however, when two omics data are used, the best ACC and worst ACC of phMFGCN are 97.84% and 97.27%, respectively, which are much bigger than the best ACC of phMFGCN with single-omics data. The optimal performance is obtained when the framework is trained with three-omics data, and the best ACC value is 97.94%.

The above results confirm that multiple omics data indeed outperform single-omics data, and the performance significantly improves as the number of omics data increases. It indicates that different omics data contain complementary information and phMFGCN has the ability to capture such complementary information.

2) THE EFFECT OF HYPER-PARAMETER α_I

The hyper-parameters in our study are purposed to measure the contribution of different omics features. In this part, experiments were carried out in phMFGCN with GE, ME and MU data, so the corresponding hyper-parameters are α_{GE}, α_{ME} and α_{MU}. Keeping $\alpha_{GE} + \alpha_{ME} + \alpha_{MU} = 1$ and using grid-search method to find the optimal parameters. The results of ACC in different hyper-parameters are shown in Fig. 3. It can be found that in phMFGCN the best ACC is 0.9794, which is obtained when $\alpha_{GE} = 0.6$, $\alpha_{ME} = 0.3$, and $\alpha_{MU} = 0.1$. Additionally, when $\alpha_{GE} >= 0.5$, the lowest ACC value is 0.9762, which is bigger than the ACC values in most cases when $\alpha_{GE} < 0.5$. The above results indicate that GE data contributed more than ME and MU data to the gene-gene interaction prediction task and our framework is able to pay right attention to the contributions of different omics data.

3) ABLATION STUDIES ON PHMFGCN

To investigate the contributions of different components of phMFGCN, we carry out ablation studies on phMFGCN with different components. The results are shown in Table 3.

(a)Without multi-omics feature fusion in progressively helical multi-omics data fusion strategy, using $\mathbf{H}_{GE}^{(l+1)} = \sigma(\tilde{\mathbf{L}}_{GE}\mathbf{H}_{GE}^{(l)}), \ldots, \mathbf{H}_{MU}^{(l+1)} = \sigma(\tilde{\mathbf{L}}_{MU}\mathbf{H}_{MU}^{(l)})$ to replace it.

(b)Without multi-omics contribution allocation.

(c)Without integrated node embedding matrices aggregation, only using the node embedding matrix derived from the last layer.

(d)Without decoupling. The equation of phMFGCN without decoupling (PWD) is shown in Supplementary Note 3.

It can be found from Table 3 that the complete phMFGCN gets the best performance, while study (a) gets the worst performance whose ACC is 2.16% lower than that of phMFGCN. This significant performance difference indicates the importance of our multi-omics feature fusion in the whole framework. As shown in study (b), compared to the complete phMFGCN, ACC decreases by 0.62% without omics allocation strategy. Since different omics has different contribution to the target results, it is necessary to allocate different omics data different weight to measure their contribution, which is also proven in the previous section. The ACC of Study (c) is 0.67% lower than that of phMFGCN, which illustrates that the information of integrated node-embedding matrix derived from the last layer is not as sufficient as that from phMFGCN. As shown in study (d), its performance is 0.33% lower than that of phMFGCN, which demonstrates that in two layers

TABLE 2. The performance of different models with different omics data or omics combinations.

Method	ACC(%)	PRC(%)	Recall(%)	AUROC(%)	AUPRC(%)
GCN-GE	94.09±0.26	94.08±1.35	94.20±1.25	98.09±0.25	97.19±0.38
GCN-ME	93.84±0.58	95.19±1.09	92.40±2.14	98.23±0.13	97.61±0.19
GCN-MU	93.00±0.21	92.31±0.23	93.90±0.36	97.73±0.15	97.19±0.37
GE+ME	97.84±0.09	97.06±0.23	98.72±0.20	99.37±0.02	99.06±0.06
GE+MU	97.27±0.07	96.54±0.17	98.07±0.34	99.16±0.04	98.74±0.11
ME+MU	97.34±0.14	96.52±0.38	98.26±0.63	99.14±0.05	98.75±0.10
GE+ME+MU	**97.94±0.08**	**97.12±0.21**	**99.05±0.17**	**99.39±0.02**	**99.08±0.06**

TABLE 3. The results of ablation studies on phMFGCN with different components.

Study	ACC(%)	PRC(%)	Recall(%)	AUROC(%)	AUPRC(%)
(a)	95.78±0.30	96.62±0.39	94.80±0.95	98.80±0.05	98.21±0.12
(b)	97.32±0.20	96.64±0.34	98.07±0.52	99.15±0.55	98.73±0.12
(c)	97.27±0.29	96.70±0.39	97.89±0.78	99.19±0.07	98.80±0.10
(d)	97.61±0.08	96.54±0.24	98.80±0.21	99.22±0.05	98.84±0.11
phMFGCN	**97.94±0.08**	**97.12±0.21**	**99.05±0.17**	**99.39±0.02**	**99.08±0.06**

(small model depth), the GCN-based model has begun to exhibit slight entanglement. It demonstrates the necessity of decoupling operation.

4) THE EFFECT OF MODEL DEPTH IN PHMFGCN

Theoretically, more steps neighborhood aggregation (i.e. large layer numbers or large model depth) from the central node can enable itself to aggregate information in bigger hop neighborhood.

However, in traditional GCN, large model depth does not always lead to better performance, it may cause over-smoothing problem and lead to poor performance. In practice, it is difficult to know the optimal model depth of the prediction task in advance, so the model should have the ability to run on a large model depth and determine the optimal depth, which is also the difference between phMFGCN and traditional GCN.

To explore the effect with different model depth, we mainly compared the ACC and AUROC in gene interaction prediction task with phMFGCN, phMFGCN without decoupling, GCN-cox, MVGCN under different layer numbers. The results are shown in Fig. 4. Fig. 4 indicates that all models obtain the best test ACC and AUROC in 2 layers. Starting from the third layer, the ACC and AUROC decrease to different levels. However, in the case of decoupling, phMFGCN can be applied to a larger model depth. It achieves the best ACC and AUROC in 2 layers and the worst ACC and AUROC in 10 layers ($\triangle ACC_{phMF}$ = 97.94% - 96.14% = 1.80%, $\triangle AUROC_{phMF}$ = 99.39% - 97.02% = 2.37%), whereas the performance of other models gets seriously degraded as the receptive field or the depth of

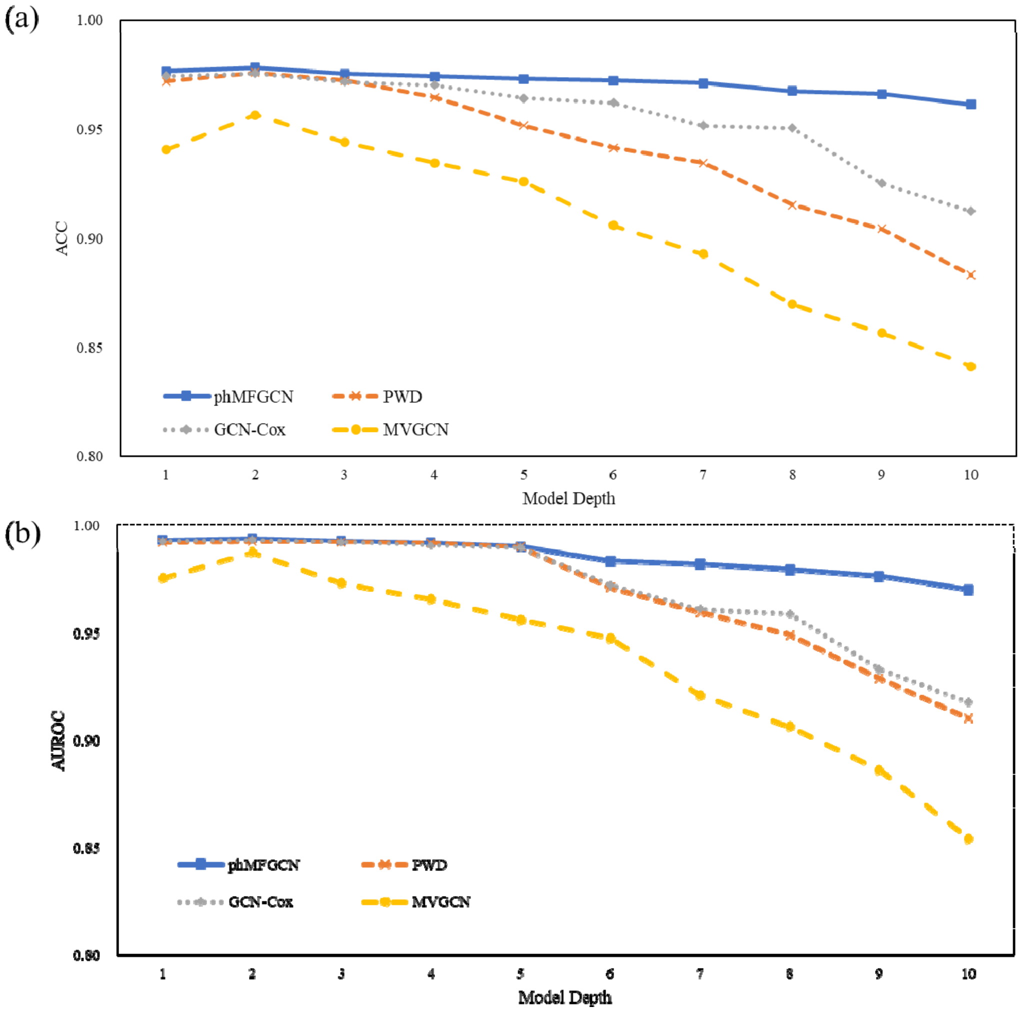

FIGURE 4. (a) The ACC comparison on phMFGCN and other GCN-based models with different model depth. (b) The AUROC comparison on phMFGCN and other GCN-based models with different model depth. phMFGCN has the best ACC and AUROC at any depth and can alleviate over-smoothing problem.

model increasing ($\triangle ACC_{PWD}$ = 97.61% - 88.32% = 9.29%, $\triangle ACC_{Cox}$ = 97.57% - 91.25% = 6.32%, $\triangle ACC_{MVGCN}$ = 95.66% - 84.15% = 11.51%; $\triangle AUROC_{PWD}$ = 99.22% - 91.03% = 8.19%, $\triangle AUROC_{Cox}$ = 99.30% - 91.77% = 7.53%, $\triangle AUROC_{MVGCN}$ = 98.76% - 85.43% = 13.33%). Obviously, the decoupling operation in phMFGCN significantly delays the decline of performance with the increasing of model depth, which indicates that phMFGCN takes effect in alleviating the over-smoothing problem.

C. THE GENERALIZED APPLICATIONS OF phMFGCN

1) NOVEL GENE-GENE INTERACTION PREDICTION

In fact, the purpose of gene-gene interaction prediction task is to predict the existence of unknown interactions in original graph. In this experiment, we used phMFGCN to calculate the probability of unknown interactions. We ran the phMFGCN 10 times and take the average value of ten probabilities of each gene pair as the final score of edge existence. Then the gene pairs with highest scores were picked and sent to the literature database to see whether there is evidence to support their existence. Among the top 8000 (about 0.1‰) gene-gene interactions we predict: 464 pairs of gene interactions are found in BioGrid [41], 22 pairs in Trrust V2 [42], and 20 pairs in RegNetwork [43].

Additionally, we chose the novel edges with high scores to form the gene interaction network, then the degree of each gene in the network was calculated. We selected the top 50 genes with the highest degree as hub genes. In order to explore the biological function of the hub genes derived by predicted edges, the KEGG pathway analysis was carried out on these hub genes. KEGG explains the hub gene's function from the perspective of functional pathway. $P < 0.01$ was considered statistically significant. The results of KEGG analysis are shown in Fig. 5 and Table 4.

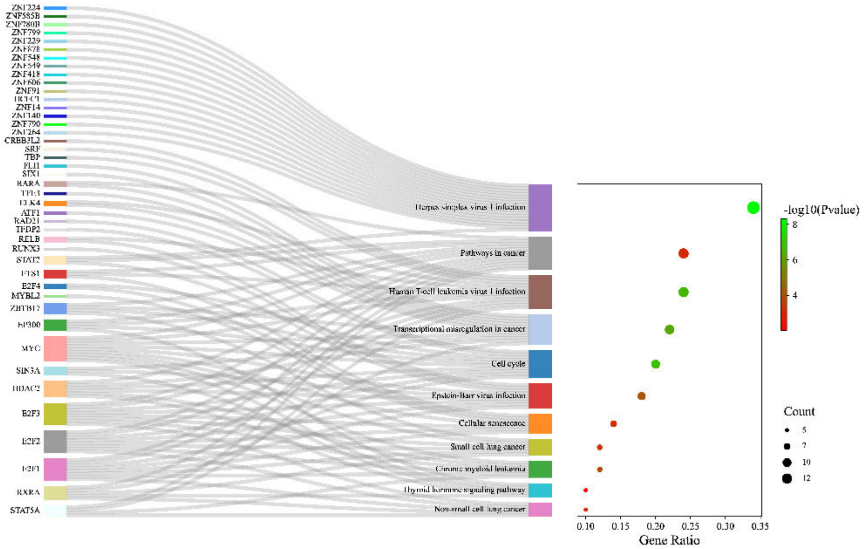

FIGURE 5. The Sankey diagram of KEGG pathway analysis of 50 hub genes (p<0.01). These 50 predicted hub genes are mainly enriched in the following pathways: pathways in cancer, transcriptional misregulation in cancer, cell cycle, non-small cell lung cancer, cellular senescence, and thyroid hormone signaling pathway, which are closely related to the occurrence and development of LUAD.

TABLE 4. The KEGG pathways of 50 hub genes.

KEGG Code	Pathways	Gene Count	P Value
hsa05168	Herpes simplex virus 1 infection	19	4.86E-09
hsa05166	Human T-cell leukemia virus 1 infection	12	2.74E-07
hsa05200	**Pathways in cancer**	12	9.08E-04
hsa05202	**Transcriptional misregulation in cancer**	11	6.79E-07
hsa04110	**Cell cycle**	10	1.84E-07
hsa05169	Epstein-Barr virus infection	9	6.89E-05
hsa04218	**Cellular senescence**	7	6.99E-04
hsa05220	Chronic myeloid leukemia	6	1.72E-04
hsa05222	Small cell lung cancer	6	4.20E-04
hsa05223	**Non-small cell lung cancer**	5	1.52E-03
hsa04919	**Thyroid hormone signaling pathway**	5	9.74E-03

These 50 predicted hub genes are mainly enriched in the following pathways: pathways in cancer, transcriptional misregulation in cancer, cell cycle, non-small cell lung cancer, cellular senescence, and thyroid hormone signaling pathway, which are closely related to the occurrence and development of LUAD. Especially, senescence cells have been proven to have immunogenicity and can promote anti-tumor immune responses [45], which means cellular senescence is a pathway closely related to the development of LUAD. As for thyroid hormone signaling pathway, Liu et al.'s research shows that thyroid hormone stimulates cancer cell proliferation through the imbalance of molecular and signal pathways [44], that is, this pathway is also very likely to affect the occurrence and development of LUAD by promoting the unlimited proliferation of lung adenocarcinoma cells. According to the above analysis, the 50 ''Hub'' genes involved in these regulatory pathways are important genes that directly or indirectly regulate the development of lung adenocarcinoma, and they may be key targets for gene diagnosis and treatment of lung adenocarcinoma. In addition, this result is an indirect proof of the reliability of our newly predicted gene-gene interaction, and also indicates the practical value of our framework in clinical practice.

2) GENE FUNCTION PREDICTION

We next explored the application of phMFGCN in the node classification tasks. In this part, we mainly tested the performance of phMFGCN in the task of gene function prediction.

TABLE 5. A part of top predicted genes for LUAD-related GSEA pathways.

Cell cycle	
Sam68	Sam68, involved in a variety of cellular processes, plays an significant role in cell cycle regulation[46].
DHX9	DHX9 is upregulated in many cancers and promotes cell cycle regulation[47, 48].
HNRNPK	HNRNPK is involved in multiple cellular processes and several of its target genes are involved in controlling the cell cycle, and participates in cancer development and progression via different pathways[49].
UBE2D3	UBE2D3 affects cancer progression, cell cycle and DNA damage repair[50].
CDCA5	CDCA5 regulates the activity of cell cycle-associated proteins and transcription factors, accelerates the proliferation of cancer cells [51].
MAPK signaling pathway	
DLC1	DLC1 inhibits LUAD cell proliferation, migration and invasion through regulating MAPK signaling pathway[52].
KABK3	KANK3 mediates MAPK pathway to regulate LUAD cells in proliferation and invasion[53].
STAMBP	STAMBP promotes LUAD metastasis by regulating MAPK signaling pathway[54].
IMP4	The silence of IMP4 inhibits the malignancy of LUAD via ERK Pathway[55].
TRAF3	TRAF3 upregulates the MAPK pathway and promotes LUAD proliferation[56].
P53 signaling pathway	
HNRNPK	HNRNPK is a regulator of p53 expression at the transcriptional and potentially translational levels[57].
DHX9	The downregulation of DHX9 causes p53-mediated apoptosis[47].
PTBP3	PTBP3 can stabilize UBE4A to regulate P53 expression and may serve as a prognostic biomarker[58].
UBF	The depletion of UBF induces p53 independent apoptosis and death in transformed cells[59]
Sam68	Sam68 may have tumor suppressor activities like p53[60].

This was accomplished by searching functional gene sets from GSEA, then these genes were considered as labeled functional genes to participate in model training. To balance the label sets, each gene in functional gene sets was set as positive gene with label 1 and the same number of genes were selected as negative genes with label 0. The negative gene set selected strategy was as followed: Firstly, remove positive sample genes from the whole gene set. Then, we calculated Pearson correlation coefficient one by one between remain genes and positive sample genes. Finally, we selected a threshold and chose the same number of genes with the correlation coefficient among each positive gene less than the threshold to form a negative set. In this way, the negative genes and positive genes have a low correlation, which is suitable for negative sample selection rules. The threshold value started from 0.1. If the number of negative sample genes selected is not enough, the threshold value will be changed to a little bit higher value.

We focused on studying three pathways selected from the GSEA database that are closely related to the occurrence and development of LUAD, namely Cell cycle, MAPK signaling pathway, and P53 signaling pathway. The genes in these pathways are considered as their corresponding functional gene sets, with GE, ME, and MU data as the original feature datasets. Perform node classification tasks on the remaining genes separately to explore their correlation with the corresponding functional gene set (pathway) and reveal their gene functions. Table 5 show a portion of high scoring predictive genes related to the LUAD KEGG pathway (the top 10 genes with scoring rankings were selected from the predicted results of each functional gene set, and due to space limitations, only 5 predicted results for each pathway were listed here). All these predictive results are supported by existing research or literature, indicating that these genes are involved in or regulate corresponding pathways. This proves the excellent performance of our phMFGCN framework in node classification tasks, which is also the result of its perfect integration of multiple omics data information.

Through the results of the above table, we can find that Sam68, DHX9, and HNRNPK are simultaneously involved in regulating the cell cycle pathway and P53 signaling pathway, which further reveals their close relationship with the occurrence and development of cancer. It is worth conducting medical experiments to further explore their properties and functions, which may provide important reference value for the corresponding cancer gene diagnosis and treatment and the development of targeted drugs.

IV. CONCLUSION

In this paper, we proposed a novel GCN-based framework phMFGCN to overcome the difficulties in multi-omics data fusion, which consists of: 1) construct an initial graph; 2) decouple; 3) integrate multi-omics data with progressively helical multi-omics data fusion strategy; 4) generate edge embedding vectors; 5) calculate the probability of edges. In gene-gene interaction prediction task, phMFGCN achieved the best test results in widely used evaluation metrics when compared with other machine learning models. phMFGCN also acquired superior performance in predicting novel gene-gene interactions and some prediction results were supported by existing research and literatures. Additionally, in functional gene prediction task, the top genes that phMFGCN predicted were authentically closely related to or participate in regulation of their corresponding GSEA pathway. Under various tasks and experimental conditions, phMFGCN performed well, which showed the effectiveness and progress of our framework in addressing the challenges of multiple omics data integration.

In the future, with the increasing development of multi omics sequencing technology, the types of multi omics data will become more diverse, and our framework has the potential to integrate these omics data and provide reliable guidance for clinical practice. Additionally, our framework can be further extended to more complex diseases and clinical tasks to make greater contributions in biomedical applications.

SUPPLEMENTARY MATERIALS

Supplementary Note 1: Evaluation metrics.

Supplementary Note 2: Optimal parameters.

Supplementary Note 3: The multi-omics feature fusion equation of phMFGCN without decoupling.

REFERENCES

[1] T. Cui, K. El Mekkaoui, J. Reinvall, A. S. Havulinna, P. Marttinen, and S. Kaski, ‘‘Gene–gene interaction detection with deep learning,’’ *Commun. Biol.*, vol. 5, no. 1, p. 1238, 2022.

[2] M. Costanzo, E. Kuzmin, J. van Leeuwen, B. Mair, J. Moffat, C. Boone, and B. Andrews, ‘‘Global genetic networks and the genotype-to-phenotype relationship,’’ *Cell*, vol. 177, no. 1, pp. 85–100, Mar. 2019.

[3] R. Bhara, S. Tiwari, S. Vijayraghavalu, and M. Kumar, ‘‘Genetic polymorphisms of xenobiotic metabolizing genes (GSTM1, GSTT1, GSTP1), gene-gene interaction with association to lung cancer risk in North India; a case control study,’’ *Asian Pacific J. Cancer Prevention*, vol. 20, no. 9, pp. 2707–2714, Sep. 2019.

[4] R. Zhang et al., ‘‘Independent validation of early-stage non-small cell lung cancer prognostic scores incorporating epigenetic and transcriptional biomarkers with gene-gene interactions and main effects,’’ *Chest*, vol. 158, no. 2, pp. 808–819, Aug. 2020.

[5] P. C. Phillips, ‘‘Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems,’’ *Nature Rev. Genet.*, vol. 9, no. 11, pp. 855–867, Nov. 2008.

[6] R. Upstill-Goddard, D. Eccles, J. Fliege, and A. Collins, ‘‘Machine learning approaches for the discovery of gene-gene interactions in disease data,’’ *Briefings Bioinf.*, vol. 14, no. 2, pp. 251–260, Mar. 2013, doi: 10.1093/bib/bbs024.

[7] P. V. Johnsen, S. Riemer-Sørensen, A. T. DeWan, M. E. Cahill, and M. Langaas, ‘‘A new method for exploring gene–gene and gene–environment interactions in GWAS with tree ensemble methods and SHAP values,’’ *BMC Bioinf.*, vol. 22, no. 1, p. 230, Dec. 2021, doi: 10.1186/s12859-021-04041-7.

[8] R. Luss, S. Rosset, and M. Shahar, ‘‘Efficient regularized isotonic regression with application to gene–gene interaction search,’’ *Ann. Appl. Statist.*, vol. 6, no. 1, pp. 253–283, Mar. 2012, doi: 10.1214/11-AOAS504.

[9] S. Chakraborty, M. I. Hosen, M. Ahmed, and H. U. Shekhar, ‘‘Onco-multi-OMICS approach: A new frontier in cancer research,’’ *Biomed. Res. Int.*, vol. 2018, Oct. 2018, Art. no. 9836256, doi: 10.1155/2018/9836256.

[10] S. Ren, Y. Shao, X. Zhao, C. S. Hong, F. Wang, X. Lu, J. Li, G. Ye, M. Yan, Z. Zhuang, C. Xu, G. Xu, and Y. Sun, ‘‘Integration of metabolomics and transcriptomics reveals major metabolic pathways and potential biomarker involved in prostate cancer,’’ *Mol. Cellular Proteomics*, vol. 15, no. 1, pp. 154–163, Jan. 2016.

[11] S. Huang, K. Chaudhary, and L. X. Garmire, ‘‘More is better: Recent progress in multi-omics data integration methods,’’ *Frontiers Genet.*, vol. 8, p. 84, Jun. 2017.

[12] J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin, ‘‘Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data,’’ *Briefings Bioinf.*, vol. 19, no. 6, pp. 1370–1381, Jun. 2017.

[13] R. Schulte-Sasse, S. Budach, D. Hnisz, and A. Marsico, ‘‘Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms,’’ *Nature Mach. Intell.*, vol. 3, no. 6, pp. 513–526, Apr. 2021, doi: 10.1038/s42256-021-00325-y.

[14] T. Ma, C. Xiao, J. Zhou, and F. Wang, ‘‘Drug similarity integration through attentive multi-view graph auto-encoders,’’ 2018, *arXiv:1804.10850*.

[15] Y. Tong, Q. He, J. Zhu, E. Ding, and K. Song, ‘‘Multi-omics differential gene regulatory network inference for lung adenocarcinoma tumor progression biomarker discovery,’’ *AIChE J.*, vol. 68, no. 4, Apr. 2022, Art. no. e17574, doi: 10.1002/aic.17574.

[16] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, ‘‘Similarity network fusion for aggregating data types on a genomic scale,’’ *Nature Methods*, vol. 11, no. 3, pp. 333–337, Mar. 2014, doi: 10.1038/nmeth.2810.

[17] C. Angione, M. Conway, and P. Lió, ‘‘Multiplex methods provide effective integration of multi-omic data in genome-scale models,’’ *BMC Bioinf.*, vol. 17, no. S4, pp. 257–269, Feb. 2016.

[18] N. Rappoport and R. Shamir, ‘‘NEMO: Cancer subtyping by integration of partial multi-omic data,’’ *Bioinformatics*, vol. 35, no. 18, pp. 3348–3356, Sep. 2019, doi: 10.1093/bioinformatics/btz058.

[19] M. Ruffalo, M. Koyutürk, and R. Sharan, ‘‘Network-based integration of disparate omic data to identify ‘silent players’ in cancer,’’ *PLOS Comput. Biol.*, vol. 11, no. 12, Dec. 2015, Art. no. e1004595, doi: 10.1371/journal.pcbi.1004595.

[20] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, P. Cau, E. Remy, and A. Baudot, ‘‘Random walk with restart on multiplex and heterogeneous biological networks,’’ *Bioinformatics*, vol. 35, no. 3, pp. 497–505, Feb. 2019, doi: 10.1093/bioinformatics/bty637.

[21] B. Yesilkaya, M. Perc, and Y. Isler, ‘‘Manifold learning methods for the diagnosis of ovarian cancer,’’ *J. Comput. Sci.*, vol. 63, Sep. 2022, Art. no. 101775.

[22] M. Surucu, Y. Isler, M. Perc, and R. Kara, ‘‘Convolutional neural networks predict the onset of paroxysmal atrial fibrillation: Theory and applications,’’ *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 31, no. 11, Nov. 2021, Art. no. 113119, doi: 10.1063/5.0069272.

[23] L. Zhou, M. Rueda, and A. Alkhateeb, ‘‘Classification of breast cancer Nottingham prognostic index using high-dimensional embedding and residual neural network,’’ *Cancers*, vol. 14, no. 4, p. 934, Feb. 2022.

[24] Z.-J. Cao and G. Gao, "Multi-omics single-cell data integration and regulatory inference with graph-linked embedding," *Nature Biotechnol.*, vol. 40, no. 10, pp. 1458–1466, Oct. 2022, doi: 10.1038/s41587-022-01284-4.

[25] R. G. Verhaak et al., "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, Jan. 2010, doi: 10.1016/j.ccr.2009.12.020.

[26] Y. Wang, Z. Zhang, H. Chai, and Y. Yang, "Multi-omics cancer prognosis analysis based on graph convolution network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1564–1568.

[27] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature Commun.*, vol. 12, no. 1, p. 3445, Jun. 2021.

[28] X. Li, J. Ma, L. Leng, M. Han, M. Li, F. He, and Y. Zhu, "MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis," *Frontiers Genet.*, vol. 13, p. 127, Feb. 2022.

[29] H. Xiao, B. Wang, H. Xiong, J. Guan, J. Wang, T. Tan, K. Lin, S. Zou, Z. Hu, and K. Wang, "A novel prognostic index of hepatocellular carcinoma based on immunogenomic landscape analysis," *J. Cellular Physiol.*, vol. 236, no. 4, pp. 2572–2591, Apr. 2021, doi: 10.1002/jcp.30015.

[30] L. Fang, Y. Li, L. Ma, Q. Xu, F. Tan, and G. Chen, "GRNdb: Decoding the gene regulatory networks in diverse human and mouse conditions," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D97–D103, Jan. 2021, doi: 10.1093/nar/gkaa995.

[31] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.

[32] S. Xu, R. Liu, and Y. Da, "Comparison of tumor related signaling pathways with known compounds to determine potential agents for lung adenocarcinoma," *Thoracic Cancer*, vol. 9, no. 8, pp. 974–988, Aug. 2018, doi: 10.1111/1759-7714.12773.

[33] D. Anusewicz, M. Orzechowska, and A. K. Bednarek, "Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of Notch, Hedgehog, Wnt, and ErbB signalling," *Sci. Rep.*, vol. 10, no. 1, p. 21128, Dec. 2020, doi: 10.1038/s41598-020-77284-8.

[34] Q. He, Z. Qiu, Y. Tong, and K. Song, "A new TTZ feature extracting algorithm to decipher tobacco related mutation signature genes for the personalized lung adenocarcinoma treatment," *IEEE Access*, vol. 8, pp. 89031–89040, 2020.

[35] H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, and B. Yang, "Geom-GCN: Geometric graph convolutional networks," 2020, *arXiv:2002.05287*.

[36] D. Bo, X. Wang, C. Shi, and H. Shen, "Beyond low-frequency information in graph convolutional networks," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 5, pp. 3950–3957.

[37] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6861–6871.

[38] M. Liu, H. Gao, and S. Ji, "Towards deeper graph neural networks," Presented at the 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2020.

[39] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103761, doi: 10.1016/j.compbiomed.2020.103761.

[40] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[41] R. Oughtred, J. Rust, C. Chang, B. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-Aryamontri, K. Dolinski, and M. Tyers, "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein Sci.*, vol. 30, no. 1, pp. 187–200, Jan. 2021.

[42] H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, and E. Kim, "TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic acids Res.*, vol. 46, no. D1, pp. D380–D386, 2018.

[43] Z.-P. Liu, C. Wu, H. Miao, and H. Wu, "RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse," *Database*, vol. 2015, Jan. 2015, Art. no. bav095.

[44] Y.-C. Liu, C.-T. Yeh, and K.-H. Lin, "Molecular functions of thyroid hormone signaling in regulation of cancer progression and anti-apoptosis," *Int. J. Mol. Sci.*, vol. 20, no. 20, p. 4986, Oct. 2019. [Online]. Available: https://www.mdpi.com/1422-0067/20/20/4986

[45] I. Marin, O. Boix, A. Garcia-Garijo, I. Sirois, A. Caballe, E. Zarzuela, I. Ruano, C. S.-O. Attolini, N. Prats, J. A. López-Domínguez, M. Kovatcheva, E. Garralda, J. Muñoz, E. Caron, M. Abad, A. Gros, F. Pietrocola, and M. Serrano, "Cellular senescence is immunogenic and promotes antitumor immunity," *Cancer Discovery*, vol. 13, no. 2, pp. 410–431, Feb. 2023, doi: 10.1158/2159-8290.Cd-22-0523.

[46] S. J. Taylor, R. J. Resnick, and D. Shalloway, "Sam68 exerts separable effects on cell cycle progression and apoptosis," *BMC Cell Biol.*, vol. 5, p. 5, Jan. 2004, doi: 10.1186/1471-2121-5-5.

[47] U. Thacker, T. Pauzaite, J. Tollitt, M. Twardowska, C. Harrison, A. Dowle, D. Coverley, and N. A. Copeland, "Identification of DHX9 as a cell cycle regulated nucleolar recruitment factor for CIZ1," *Sci. Rep.*, vol. 10, no. 1, p. 18103, Oct. 2020, doi: 10.1038/s41598-020-75160-z.

[48] O. Sergeeva and T. Zatsepin, "RNA helicases as shadow modulators of cell cycle progression," *Int. J. Mol. Sci.*, vol. 22, no. 6, p. 2984, Mar. 2021, doi: 10.3390/ijms22062984.

[49] Z. Wang, H. Qiu, J. He, L. Liu, W. Xue, A. Fox, J. Tickner, and J. Xu, "The emerging roles of hnRNPK," *J. Cellular Physiol.*, vol. 235, no. 3, pp. 1995–2008, Mar. 2020, doi: 10.1002/jcp.29186.

[50] H. Yang, L. Wu, S. Ke, W. Wang, L. Yang, X. Gao, H. Fang, H. Yu, Y. Zhong, C. Xie, F. Zhou, and Y. Zhou, "Downregulation of ubiquitin-conjugating enzyme UBE2D3 promotes telomere maintenance and radioresistance of ECA-109 human esophageal carcinoma cells," *J. Cancer*, vol. 7, no. 9, pp. 1152–1162, 2016, doi: 10.7150/jca.14745.

[51] J. Ji, T. Shen, Y. Li, Y. Liu, Z. Shang, and Y. Niu, "CDCA5 promotes the progression of prostate cancer by affecting the ERK signalling pathway," *Oncol. Rep.*, vol. 45, no. 3, pp. 921–932, Jan. 2021, doi: 10.3892/or.2021.7920.

[52] N. Niu, X. Ma, H. Liu, J. Zhao, C. Lu, F. Yang, and W. Qi, "DLC1 inhibits lung adenocarcinoma cell proliferation, migration and invasion via regulating MAPK signaling pathway," *Exp. Lung Res.*, vol. 47, no. 4, pp. 173–182, Apr. 2021, doi: 10.1080/01902148.2021.1885524.

[53] Z. Dai, B. Xie, B. Yang, X. Chen, C. Hu, and Q. Chen, "KANK3 mediates the p38 MAPK pathway to regulate the proliferation and invasion of lung adenocarcinoma cells," *Tissue Cell*, vol. 80, Feb. 2023, Art. no. 101974, doi: 10.1016/j.tice.2022.101974.

[54] H. Xu, X. Yang, X. Xuan, D. Wu, J. Zhang, X. Xu, Y. Zhao, C. Ma, and D. Li, "STAMBP promotes lung adenocarcinoma metastasis by regulating the EGFR/MAPK signaling pathway," *Neoplasia*, vol. 23, no. 6, pp. 607–623, Jun. 2021.

[55] R. Li, Z. Han, W. Ma, L. Zhang, X. Zhang, Y. Jiang, and W. Dong, "IMP4 silencing inhibits the malignancy of lung adenocarcinoma via ERK pathway," *J. Oncol.*, vol. 2022, pp. 1–15, Oct. 2022.

[56] X. Du, S. Wang, X. Liu, T. He, X. Lin, S. Wu, D. Wang, J. Li, W. Huang, and H. Yang, "MiR-1307–5p targeting TRAF3 upregulates the MAPK/NF-κB pathway and promotes lung adenocarcinoma proliferation," *Cancer Cell Int.*, vol. 20, no. 1, pp. 1–16, Dec. 2020.

[57] A. Swiatkowska, M. Dutkiewicz, P. Machtel, D. M. Janecki, M. Kabacinska, P. Żydowicz-Machtel, and J. Ciesiołka, "Regulation of the p53 expression profile by hnRNP K under stress conditions," *RNA Biol.*, vol. 17, no. 10, pp. 1402–1415, Oct. 2020.

[58] C. Xie, F. Long, L. Li, X. Li, M. Ma, Z. Lu, R. Wu, Y. Zhang, L. Huang, J. Chou, N. Gong, G. Hu, and C. Lin, "PTBP3 modulates P53 expression and promotes colorectal cancer cell proliferation by maintaining UBE4A mRNA stability," *Cell Death Disease*, vol. 13, no. 2, p. 128, Feb. 2022.

[59] N. Hamdane, C. Herdman, J.-C. Mars, V. Stefanovsky, M. G. Tremblay, and T. Moss, "Depletion of the cisplatin targeted HMGB-box factor UBF selectively induces p53-independent apoptotic death in transformed cells," *Oncotarget*, vol. 6, no. 29, pp. 27519–27536, Sep. 2015.

[60] N. Li, C. T.-A. Ngo, O. Aleynikova, N. Beauchemin, and S. Richard, "The p53 status can influence the role of Sam68 in tumorigenesis," *Oncotarget*, vol. 7, no. 44, pp. 71651–71659, Nov. 2016.

JUNXUAN ZHU was born in Suizhou, Hubei, China, in 1998. He received the B.S. degree in process equipment and control engineering from Tianjin University, Tianjin, China, in 2020, where he is currently pursuing the M.S. degree in chemical process machinery. His research interests include computational cancer genomics, multi-omics integration, and machine learning algorithms.

JINHAN ZHANG was born in Tianjin, China, in 1999. She received the B.S. degree in process equipment and control engineering from Tianjin University, Tianjin, in 2021, where she is currently pursuing the M.S. degree in chemical process machinery. Her research interests include bioinformation, machine learning algorithms, and computer vision.

LIYAN WANG was born in Tianjin, China, in 1999. She received the B.S. degree in process equipment and control engineering from Tianjin University, Tianjin, in 2021, where she is currently pursuing the M.S. degree in chemical process machinery. Her research interests include computational cancer genomics, multi-omics integration, and deep learning algorithms.

HAO HUANG was born in Jiujiang, Jiangxi, China, in 1998. He received the B.S. degree in process equipment and control engineering from Tianjin University, Tianjin, China, in 2020, where he is currently pursuing the M.S. degree in chemical process machinery. His research interests include fault diagnosis in chemical engineering and Bayesian structure learning algorithm.

ZHIBO ZHANG was born in Shuozhou, Shanxi, China, in 1981. He received the master's degree in clinical surgery from Shanxi Medical University, Shanxi, in 2009. He is currently with Tianjin Hospital, Tianjin, China. His research interests include the diagnosis and treatment of chronic orthopedic diseases and bone tumor diseases.

KAI SONG was born in Changchun, Jilin, China, in 1975. She received the B.S. and Ph.D. degrees in control science and engineering from Zhejiang University, Zhejiang, in 1998 and 2005, respectively. From 2005 to 2007, she was an Assistant Professor with the Process Equipment and Control Engineering Department, School of Chemical Engineering and Technology, Tianjin University, China. Since 2007, she has been an Associate Professor with the Process Equipment and Control Engineering Department, School of Chemical Engineering and Technology, Tianjin University. From February 2013 to October 2015, she was a Visiting Associate Professor with the Dr. John Minna's Laboratory, Department of Clinical Science, UT Southwestern Medical Center, Dallas, TX, USA. She is in charge of several projects supported by the National Natural Science Foundation of China and the National Key Research and Development Program of China. She is the author of *Introduction of Synthetic Biology* (in Chinese, the first textbook about synthetic biology in China) and more than 80 articles. Since 2015, she started her research and published several articles about bioinformatics in cancer research cooperating with Dr. John Minna and Dr. Adi Gazdar. Her research interests include bioinformatics, synthetic biology, big data, and other applications of machine learning algorithms in biology, cancer, and process control.

XIAOFEI ZHANG was born in Tianjin, China, in 1987. He received the bachelor's degree in clinical medicine from Tianjin Medical University, Tianjin, in 2010, and the M.Res. degree in advanced genomic and proteomic science from the University of Nottingham, Nottinghamshire, U.K., in 2011. He is currently pursuing the M.D. degree in surgery with Tianjin Medical University. His research interests include bioinformatics, genomics, and proteomics.