A multimodal graph neural network framework for cancer molecular subtype classification

Bingjun Li[1] and Sheida Nabavi[1*]

*Correspondence:
sheida.nabavi@uconn.edu

[1] Department of Computer Science and Engineering, University of Connecticut, Storrs, USA

Abstract

Background: The recent development of high-throughput sequencing has created a large collection of multi-omics data, which enables researchers to better investigate cancer molecular profiles and cancer taxonomy based on molecular subtypes. Integrating multi-omics data has been proven to be effective for building more precise classification models. Most current multi-omics integrative models use either an early fusion in the form of concatenation or late fusion with a separate feature extractor for each omic, which are mainly based on deep neural networks. Due to the nature of biological systems, graphs are a better structural representation of bio-medical data. Although few graph neural network (GNN) based multi-omics integrative methods have been proposed, they suffer from three common disadvantages. One is most of them use only one type of connection, either inter-omics or intra-omic connection; second, they only consider one kind of GNN layer, either graph convolution network (GCN) or graph attention network (GAT); and third, most of these methods have not been tested on a more complex classification task, such as cancer molecular subtypes.

Results: In this study, we propose a novel end-to-end multi-omics GNN framework for accurate and robust cancer subtype classification. The proposed model utilizes multi-omics data in the form of heterogeneous multi-layer graphs, which combine both inter-omics and intra-omic connections from established biological knowledge. The proposed model incorporates learned graph features and global genome features for accurate classification. We tested the proposed model on the Cancer Genome Atlas (TCGA) Pan-cancer dataset and TCGA breast invasive carcinoma (BRCA) dataset for molecular subtype and cancer subtype classification, respectively. The proposed model shows superior performance compared to four current state-of-the-art baseline models in terms of accuracy, F1 score, precision, and recall. The comparative analysis of GAT-based models and GCN-based models reveals that GAT-based models are preferred for smaller graphs with less information and GCN-based models are preferred for larger graphs with extra information.

Keywords: Graph attention network, Multi-omics integration, Cancer subtype, Molecular subtype

Background

The fast-growing high-throughput sequencing technology has made DNA and RNA sequencing more efficient and accessible, resulting in a large collection of multi-omics data which makes molecular profiling possible. Due to the heterogeneity in cancer and the complexity of the biological processes, employing multi-omics sequencing data are crucial to more accurate cancer classification and tumor profiling. Many researchers have proposed methods that incorporate multi-omics data for either cancer type classification or cell type clustering [1–11]. These methods show that utilizing multi-omics data improves performance, and provides a better understanding of the key pathophysiological pathways across different molecular layers [12]. A typical multi-omics data generated from DNA and RNA sequencing usually consists of mRNA expression, microRNA (miRNA) expression, copy number variation (CNV), and DNA methylation [13]. The difference in data distributions across each omic, and the complex inter-omics and intra-omic connections (certain omic can act as a promotor or suppressor to genes) add more challenges to developing an integrative multi-omics classification method for cancer molecular subtypes.

Recent studies have shown that cancer taxonomy based on molecular subtypes can be crucial for precision oncology [13, 14]. An accurate cancer molecular subtype classifier is crucial for early-stage diagnosis, prognosis, and drug development. Traditional cancer taxonomy is based on its tissue origin. In 2014, The Cancer Genome Atlas (TCGA) Research Network proposed a new clustering method for cancers based on their integrated molecular subtypes that share mutations, copy-number alterations, pathway commonalities, and micro-environment characteristics instead of their tissue of origin [13]. They found 11 subtypes from 12 cancer types. In 2018, they applied the new taxonomy method to 33 cancer types and found 28 molecular subtypes [15]. The new cancer taxonomy provides a better insight into the heterogeneous nature of cancer.

With the recent development in deep learning models, data-driven models benefit from the powerful feature extraction capability of deep learning networks in many fields [16–19]. Most multi-omics integrative models employ an early fusion approach that aggregates multi-omics data (mainly by concatenation) and then applies a deep neural network as a feature extractor; or a late fusion approach that first extracts features from each omic by deep neural networks and then aggregates extracted features as inputs to the classification network. For efficient implementation of multi-omics integrative models, convolutional neural networks (CNNs) are widely used [20].

Traditional deep neural networks are based on the assumption that the inner structure of the data is in Euclidean space [21]. Because of the complex interactions across many biological processes, such data structure is not a proper representation of bio-medical data, and researchers proposed graph-based data structures to tackle this limitation. In 2016, a graph convolution network (GCN), ChebNet, was proposed [16]. It uses the Chebyshev polynomial as the localized learning filter to extract the graph feature representation. In 2017, Petar Velickovic et al. proposed a graph attention network (GAT) that overcomes GCN's disadvantage of dependence on the Laplacian eigenbasis [22]. GAT uses masked self-attention layers to enable nodes to attend over their neighborhoods' features [22]. With the recent growing interest in the graph neural network, many graph-based classification methods have been proposed in the bio-medical field.

To utilize the power of graph-structured data, Ramirez et al. proposed a GCN method to use intra-omic connections, protein-protein interaction networks, and gene co-expression networks. The model achieves a 94.71% classification accuracy for 33 cancer types and normal tissue on TCGA data [23]. To use the intra-omic connection across multiple omics, Wang et al. proposed MOGONET, a late-fusion GCN-based method that integrates multi-omics data for bio-medical data classification. And they achieve 80.61% accuracy on breast cancer subtype classification with BRCA dataset [5]. To compensate for the limitation of GCN, that it only extracts local representation on the graph, Li et al. proposed a parallel-structured GCN-based method that utilizes a gene-based prior knowledge graph for cancer molecular subtype classification [1]. There are also other ways to structure the graph. Wang et al. proposed a GCN-based method that uses a KNN-generated cell-cell similarity graph for single-cell sequencing data classification [24].

Since the introduction of GAT in 2017, it has gained more and more interest. Shanthamallu et al. proposed a GAT-based method, GrAMME, with two variations that use a supra-graph approach and late-fusion approach to extract features from a multi-layer graph with intra-omic connections only for classification in social science and political science datasets [25]. On the other hand, Kaczmarek et al. proposed a multi-omics graph transformer to utilize an inter-omics connection only graph, the miRNA-gene target network, for cancer classification on 12 cancer types from the TCGA data [7].

There are three common disadvantages of these approaches. First, most of them consider only one kind of connections in their model, either inter-omics or intra-omic connections. They do not aim to utilize both inter-omics and intra-omic connections for more effective feature extraction. Second, they only consider one kind of GNN models, either GCN or GAT. We find that GAT and GCN have their strength in different scenarios as shown in our experiments. Different graph layers are preferred for different tasks even with datasets in a similar domain. Third, most of these methods have not been tested on a more complex classification task. They are used for classification based on the cell-of-origin taxonomy such as cancer type classification and have not been applied to a more complex classification task such as cancer molecular subtype classification, which is more useful for diagnosis, prognosis, and treatment. Inspired by our previous work on the cancer molecular subtype classification based solely on intra-omic connections, we aim to develop a multi-omics integrative framework that exploits the powerful data aggregation property of GCN or GAT models (depending on the situation) and utilizes both the intra-omic network and the inter-omics network for more precise classification.

Our goal is to build an accurate, robust, and efficient multi-omics integrative predictive model to classify these cancer molecular subtypes. In this work, we propose a general framework that can be used with any graph neural networks as the feature extractor, incorporate both gene-based and non-gene-based prior biological knowledge (primarily miRNA), and learn a knowledge graph consisting of both intra-omic and inter-omics connections. We apply the proposed model to classify cancer molecular subtypes and breast cancer molecular subtypes. We choose breast cancer as it is one of the most common and lethal cancers with a large number of samples in TCGA. It can be categorized into four major molecular subtypes based on the gene expression of the cancer cells, and

breast cancer subtypes have significant impacts on the patient's survival rates [26]. Our experimental results show the proposed method outperforms both the graph-based and CNN-based state-of-the-art methods.

Our contributions in this study are (i) a novel generalized GNN-based multi-omics integrative framework for cancer molecular subtype classification, (ii) a supra-graph approach that can incorporate both intra-omic and inter-omics prior biological knowledge in the form of graphs, (iii) a representation of multi-omics data in the form of heterogeneous multi-layer graph, and (iv) a comparative analysis of GCN and GAT based models at different combinations of omics and different graph structures.

Method and materials

The overview of the proposed framework structure is shown in Fig. 1. The input data for the proposed framework is shown as a graph structure on the leftmost side. The data consists of three omics, mRNA expression (orange boxes), copy number variation (CNV) (yellow boxes), and miRNA (green boxes). The details of the network structure are discussed in the following Network Section. The proposed framework consists of 4 major modules: Module (1) a linear dimension-increase neural network, Module (2) a graph neural network (GNN), Module (3) a decoder, and Module (4) a shallow parallel network. Any kind of graph neural network can be used in Module 2. In this study, we focus on graph convolutional network (GCN) and graph attention network (GAT), which are two major kinds of GNN. Experiments about the effect of the decoder and the shallow parallel network modules are discussed in our ablation study.

Network

We build a heterogeneous multi-layer graph based on the prior biological knowledge, i.e. gene-gene interaction (GGI) network from BioGrid and miRNA-gene target network from miRDB [27, 28]. Inspired by the meta-path and supra-graph approach for the multi-layered network models [25, 29], we build a supra-graph with miRNA-miRNA meta-paths. A miRNA-miRNA meta-path is defined as if two miRNA nodes are connected to the same gene node from the GGI network and miRNA-gene network. An example of how we construct the supra-graph is shown in Fig. 2. Meta-paths are shown as dotted lines in the figure.

The adjacency matrix of the supra-graph is an $(N + M) \times (N + M)$ matrix, where N is the number of genes and M is the number of miRNA. Every node in the graph is assumed to be self-connected, thus the diagonal elements of the adjacency matrix in the study are 1. The adjacency matrix of the supra-graph is shown in Eq. (1).

$$\mathbf{A}_{Supra} = \begin{bmatrix} \mathbf{A}_{gene-gene} & \mathbf{A}_{gene-mi} \\ \mathbf{A}_{gene-mi}^{T} & \mathbf{A}_{mi-mi}, \end{bmatrix}, \tag{1}$$

where $\mathbf{A}_{gene-gene} \in \mathbb{R}^{N \times N}$, $\mathbf{A}_{gene-mi} \in \mathbb{R}^{N \times M}$, and $\mathbf{A}_{mi-mi} \in \mathbb{R}^{M \times M}$.

We also construct four different kinds of graphs other than supra-graph in our ablation study and apply them to five input combinations of omics: mRNA, miRNA, mRNA + miRNA, mRNA+CNV, mRNA + MiRNA + CNV, to test the effect of the different graphs on the model performance. The four different graphs are defined as follows.
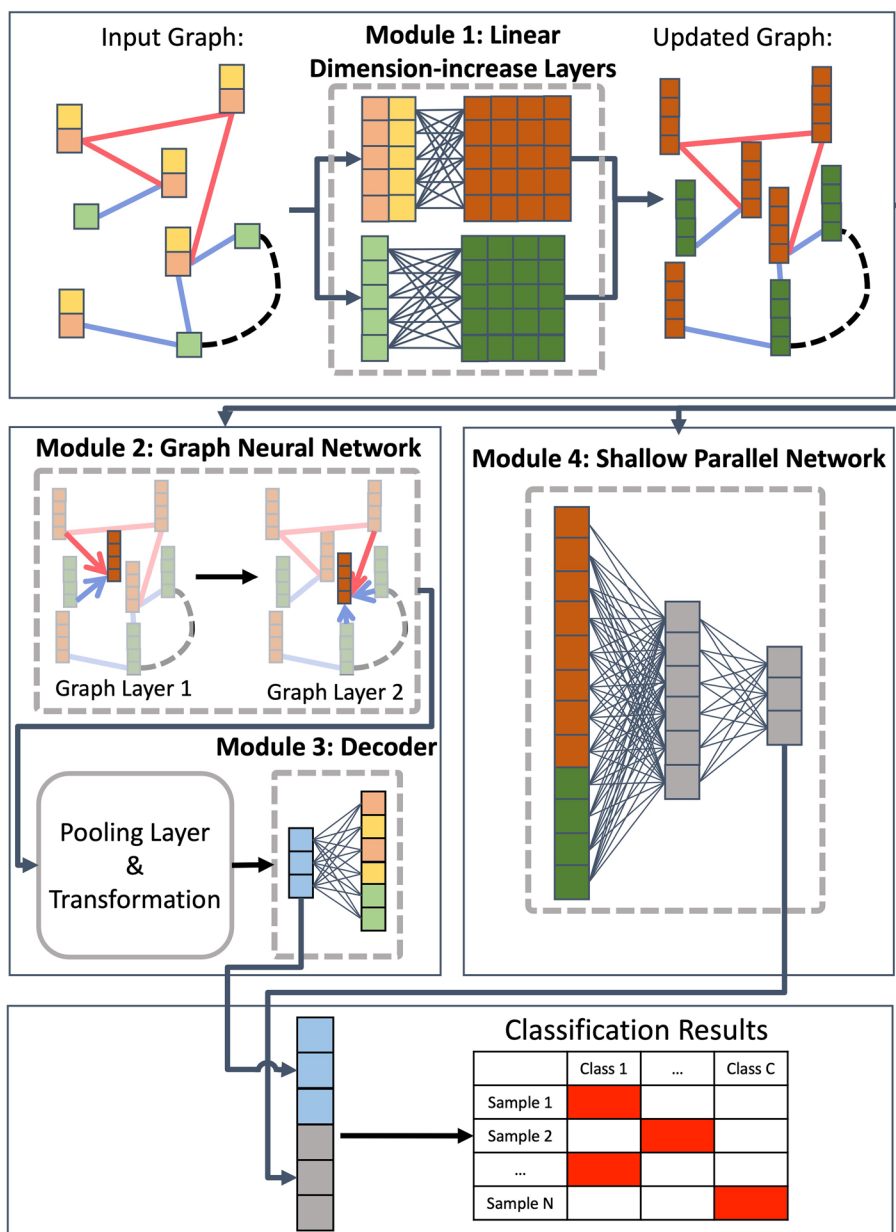
Fig. 1 The overall structure of the proposed model has four major modules shown as dotted grey rectangles. The input graph consists of inter-omics (red edges), intra-omic (blue edges) edges and miRNA-miRNA meta-path (black dashed edges), and three omics data, mRNA (orange boxes), CNV (yellow boxes), and miRNA (green boxes) is shown as the leftmost side. Module 1 consists of two parallel linear dimension-increase layers for gene-based nodes and miRNA-based nodes. The upgraded graph shown in the middle is obtained by feeding the node attributes from the input graph through module 1, where the dark orange boxes are the updated gene-based node attributes and the dark green boxes are the updated miRNA-based node attributes. Module 2 consists of two graph neural network layers, which can be any graph neural networks. The output of module 2 is then fed into a max pooling layer and then a transformation layer to obtain the learned graph representation (blue boxes). Module 3 consists of a decoder to reconstruct the graph representation back to the input graph node attributes. Module 4 consists of a shallow fully connected network that takes the updated node attributes as the input. The output of the parallel network (grey cubes) is then concatenated with the learned graph representation, and passes through a classification layer for the classification task
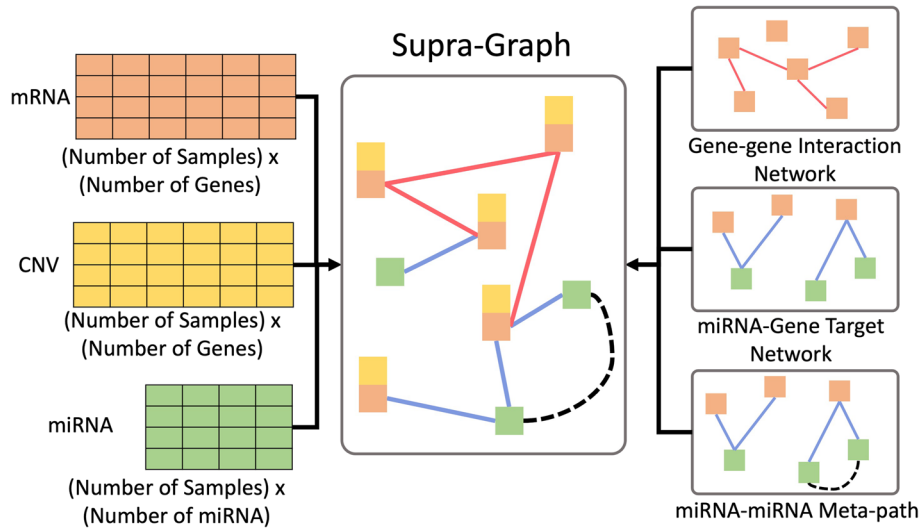
Fig. 2 The overall graph, supra-graph, is constructed from three different omic data on the left-hand side and two prior knowledge graphs on the right-hand side. mRNA (orange table) and CNV (yellow table) data are considered gene-based, which have the same dimension. miRNA (green table) data has the same number of rows but different feature lengths for each sample

Only Gene-based Nodes When the input combination of omics is mRNA or mRNA+mRNA+CNV ($M = 0$), the graph is built with the GGI network, $\mathbf{A} = \mathbf{A}_{gene-gene} \in \mathbb{R}^{N \times N}$.

Only miRNA-based Nodes When the input combination of omics is miRNA ($N = 0$), the graph is built with only miRNA meta-path network, $\mathbf{A} = \mathbf{A}_{mi-mi} \in \mathbb{R}^{M \times M}$.

Only Intra-class Edges The graph only contains GGI network and miRNA meta-path network.

$$\mathbf{A}_{Supra} = \begin{bmatrix} \mathbf{A}_{gene-gene} & \mathbf{0}_{N,M} \\ \mathbf{0}_{M,N} & \mathbf{A}_{mi-mi} \end{bmatrix} \in \mathbb{R}^{(N+M) \times (N+M)}. \tag{2}$$

Only Inter-class Edges The graph only contains miRNA-gene target network.

$$\mathbf{A}_{Supra} = \begin{bmatrix} \mathbf{I}_{N,N} & \mathbf{A}_{gene-mi} \\ \mathbf{A}_{gene-mi}^{T} & \mathbf{I}_{M,M} \end{bmatrix} \in \mathbb{R}^{(N+M) \times (N+M)}. \tag{3}$$

The input graph is denoted as a tuple $\mathcal{G} = (V, E, \mathbf{X}_V)$, where V is the set of nodes, E is the set of edges, and $\boldsymbol{x}_V$ is the node attributes. The prior knowledge is incorporated into the model through the supra-graph defined above. In the supra-graph, nodes consist of both gene-based nodes and miRNA-based nodes, and edges are assigned by the adjacency matrix. Each gene-based node has a node attribute of a vector consisting of both gene expression and CNV data, $\boldsymbol{x}_{v \in V_{gene}} \in \mathbb{R}^2$. Each miRNA-based node has a node attribute as a scalar, $x_{v \in V_{\text{miRNA}}} \in \mathbb{R}$. The gene-based nodes and miRNA-based nodes are fed through a linear dimension-increase layer, denoted as Module 1 in Fig. 1 to achieve the same node attribute dimension, $\mathbf{X}'_V \in \mathbb{R}^{(N+M) \times F}$, where F is the increased node attribute dimension.

Graph neural network: convolution-based

As mentioned before, any graph neural network can be used in the GNN module. We use ChebNet [16] to implement the GCN in this study. The supra-graph adjacency matrix introduced in the previous network section is first Laplacian normalized to $\mathbf{L}$ as expressed in Eq. (4).

$$\mathbf{L} = \mathbf{I} + \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{1/2}, \tag{4}$$

where $\mathbf{I} \in \mathbb{R}^{(N+M)\times(N+M)}$ is an identity matrix, and the degree matrix $\mathbf{D} \in \mathbb{R}^{(N+M)\times(N+M)}$ is a diagonal matrix. The eigen decomposition form of $\mathbf{L}$ can be obtained as

$$\mathbf{L} = \mathbf{U}\Lambda\mathbf{U}^T, \tag{5}$$

where $\mathbf{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n)$ is a matrix of n orthonormal eigenvectors of $\mathbf{L}$, therefore $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. And $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is the eigenvalue matrix [16].

After transforming the graph on the Fourier domain, the learning filter can be approximated by a Kth-order Chebshev polynomial. The convolution on the graph by such localized learning filter, $h(\Lambda)$ can be expressed in Eq. (6).

$$y = \mathbf{U}h(\Lambda)\mathbf{U}^T\mathbf{X}_j = \mathbf{U}\sum_{k=1}^{K-1}\beta_k T_k(\tilde{\Lambda})\mathbf{U}^T\mathbf{X}_j = \sum_{k=0}^{K-1}\beta_k T_k(\tilde{\mathbf{L}}\mathbf{X}_j), \tag{6}$$

where $\mathbf{X}_j \in \mathbb{R}^{(N+M)\times F}$ is the features of j-th sample, $\tilde{\mathbf{L}} = 2\mathbf{L}/\lambda_{max} - \mathbf{I}$, and $T_k(\tilde{\mathbf{L}}) = 2\tilde{\mathbf{L}}T_{k-1}(\tilde{\mathbf{L}}) - T_{k-2}(\tilde{\mathbf{L}})$ with $T_0(\tilde{\mathbf{L}}) = \mathbf{I}$ and $T_1(\tilde{\mathbf{L}}) = \tilde{\mathbf{L}}$. K is a hyper-parameter, where $K = 5$ in our study. A max-pooling layer with $p = 8$ is used to reduce the number of nodes and one layer of fully connected network is used to transform the learned local feature representation to a vector of length 64 for each sample, $\boldsymbol{\theta}_1 \in \mathbb{R}^{64}$.

Graph-neural network: attention-based

GAT aims to solve the problem of GCN's dependence on Laplacian eigenbasis of the graph adjacency matrix [22]. The updated node attributes are first passed through a linear transformation by a learnable weight, denoted as $\mathbf{W} \in \mathbb{R}^{F'\times F}$, where F is the updated node attribute dimension and F' is the intended output dimension for this GAT layer. Then, the self-attention coefficients for each node can be calculated as Eq. (7).

$$e_{ij} = a(\mathbf{W}\boldsymbol{x}_i, \mathbf{W}\boldsymbol{x}_j), \tag{7}$$

where e_{ij} represents the importance of node j to node i and $\boldsymbol{x}_i, \boldsymbol{x}_j$ are the node attributes for node i, j. Such attention score is only calculated for $j \in NB(i)$, where $NB(i)$ is all the first-order neighbor nodes around node i. The method normalizes the attention score by a softmax layer of e_{ij} and uses LeakyReLU as the activation function as express in Eq. (8).

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\vec{\boldsymbol{a}}^T[\mathbf{W}\boldsymbol{x}_i||\mathbf{W}\boldsymbol{x}_j]))}{\sum_{k\in NB(i)}\exp(\text{LeakyReLU}(\vec{\boldsymbol{a}}^T[\mathbf{W}\boldsymbol{x}_i||\mathbf{W}\boldsymbol{x}_k]))} \tag{8}$$

The output for each node can be expressed as Eq. (9).

$$\boldsymbol{x}_i' = \sigma(\sum_{j \in NB(i)} \alpha_{ij}\mathbf{W}\boldsymbol{x}_j). \tag{9}$$

A multi-head attention mechanism is used to stabilize the attention score. In our study, the number of heads is 8. Similar to the GCN-based GNN module, the output is then passed through a max-pooling layer and a transformation layer to obtain the local graph representation, $\boldsymbol{\theta}_1 \in \mathbb{R}^{64}$.

Decoder and shallow parallel network

As shown in Fig. 1, the decoder is a two-layer fully connected network that is used to reconstruct the node attributes on the input graph. To compensate the localization property of either GCN or GAT layer in the GNN module, we use a parallel shallow fully connected network. Since the prior knowledge graphs have many limitation [1], we may neglect some global patterns in the data when extracting features based on the graph structure only. A shallow two-layer fully connected network is able to learn the global features of the data while ignoring the actual inner structure of the data. These two modules help the framework to better extract the overall sample feature representation. The effect of including vs. excluding these two modules is discussed in detail in the Ablation Study Section.

The input of the parallel network is the updated node attributes, $\mathbf{X}_V' \in \mathbb{R}^{(N+M) \times F}$ and the output global representation of the sample, $\boldsymbol{\theta}_1$ is in the same dimension as the local feature representation from the GNN module, $\boldsymbol{\theta}_2 \in \mathbb{R}^{64}$. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are then concatenated and passed through a classification layer for prediction.

Loss function

In the proposed framework, we define the loss function L as a linear combination of three loss functions in Eq. (10).

$$L = \lambda_1 L_{ent} + \lambda_2 L_{recon} + \lambda_3 L_{reg}, \tag{10}$$

where λ_1, λ_2 and λ_3 are linear weights, L_{ent} is the standard cross-entropy loss for the classification results, L_{recon} is the mean squared error for the reconstruction loss when the decoder is included, and L_{reg} is the squared l^2 norm of the model parameters to penalize the number of parameters to avoid overfitting. L_{recon} is defined as

$$L_{recon} = \sum_{j}(\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j)^2, \tag{11}$$

where $\boldsymbol{x}_j$ is the flattened feature vector of j-th sample and $\hat{\boldsymbol{x}}_j$ is the corresponding reconstructed vector. We denote $\mathbf{W}_{all}$ as the vector consists of all parameters in the model and the L_{reg} is defined as

$$L_{reg} = \sum_{w \in \mathbf{W}_{all}} w^2. \tag{12}$$

Results and discussion

We apply the proposed model to two different classification problems. The first is cancer molecular subtype classification on the TCGA Pan-cancer dataset and the second is breast cancer subtype classification on the TCGA breast invasive carcinoma (BRCA) dataset [15, 30].

Data and experiment settings

The TCGA Pan-cancer RNA-seq data, CNV data, miRNA data, and molecular subtype labels are obtained from the University of California Santa Cruz's Xena website [31]. We only keep samples that have all three omics data and molecular subtype labels, and collect 9,027 samples in total. We use 17,946 genes that are common in both the gene expression data and the CNV data, and 743 miRNAs. The total number of molecular subtypes is 27 and there is a clear imbalance among these 27 classes as shown in Fig. 3. All samples from class 24 are excluded from the study due to the lack of miRNA data. For BRCA subtype classification, there are 981 samples in total with 4 subtypes as shown in Table 1. For the experiments on both datasets, 80% of the data is used for training, 10% is used for validation, and 10% is used for testing. All classes are present in the test set.

All expression values are normalized within their own omics. We select the top 700 genes ranked by gene expression variances across the samples, and the top 100 miRNAs
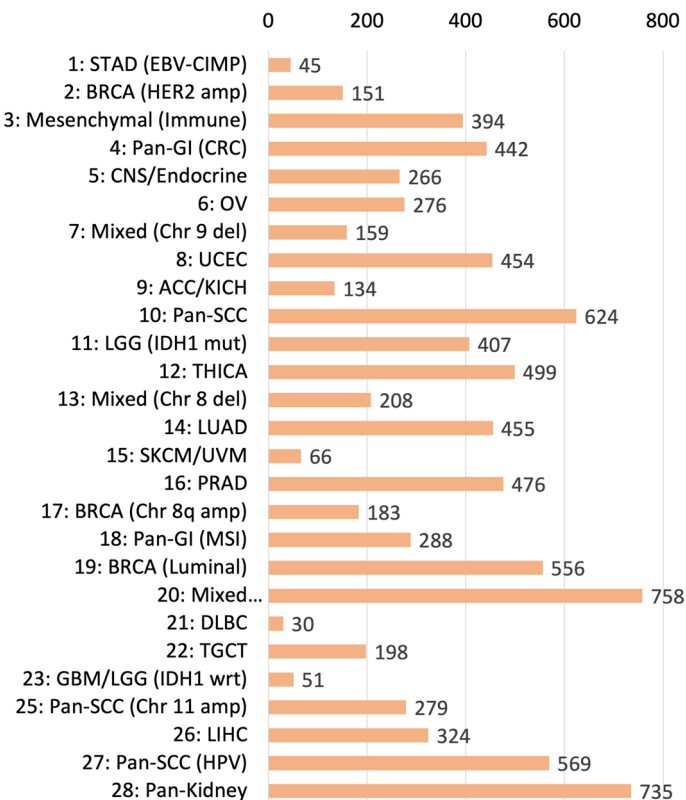

Fig. 3 The number of cases in each molecular subtypes is shown. All samples from class 24 are excluded due to lack of miRNA data

Table 1 Number of cases in each BRCA subtype

BRCA subtypes	Counts
LumA	529
LumB	197
Basal	175
Her2	80

Table 2 Configurations of baseline models on omics, graph structure, gnn layers, and regularizaiton modules

Model	Omics			Graph		GNN Layer		Module	
	mRNA	CNV	miRNA	Intra-omic	Inter-omic	GCN	GAT	Decoder	Parallel
GCN (Original) [23]	✓	–	–	✓	–	✓	–	–	–
GCN (Modified)	✓	✓	✓	✓	–	✓	–	–	–
Multi-omics GCN (Original) [1]	✓	✓	–	✓	–	✓	–	✓	✓
Multi-omics GCN (Modified)	✓	✓	✓	✓	–	✓	–	✓	✓
GrAMME (Modified) [25]	✓	✓	✓	–	✓	–	✓	–	–
Multi-omics GAT (Original) [7]	✓	–	✓	✓	–	–	✓	–	–
Multi-omics GAT (Modified)	✓	✓	✓	✓	–	–	✓	–	–

by miRNA expression variance. Results are averaged from five individual trials. The details of the model structure and hyperparameters are disclosed in the appendix. The model is implemented using Pytorch Geometric Library.

Baseline models

We selected four state-of-the-art models [1, 7, 23, 25] as baseline models to evaluate the performance of the proposed approach. These four baseline models are implemented within the proposed framework in two forms, one is with the original structure, and the other is with some modifications to accommodate the multi-omics data. The details of all graph-based baseline implementation configurations are shown in Table 2. We also included a fully-connected neural network (FC-NN) as a Euclidean-based baseline model. Conventional machine learning methods, such as Random Forest and SVM are not included in the scope of this study because they do not scale well to the multi-omics data as mentioned in our previous work [1].

Fully-connected neural network (FC-NN)

The FC-NN is one of the widely used deep learning model for data in Euclidean space. The implemented structure is the same as the parallel structure. The input data is passed through a dimension-increase layer and then flattened. The flattened data is passed through three hidden layers and a softmax layer for classification.

GCN models by Ramirez et al.

The GCN model on cancer type classification is designed for gene expression data with intra-omic connections only [23]. The implementation of the original structure and the modified structure is a GCN model with no regularization modules.

Multi-omics GCN models by Li et al.

The multi-omics GCN model on cancer molecular subtype classification is designed for gene expression and CNV data with intra-omic connections only [1]. The implementation of both structures is a GCN model with a decoder and a parallel structure as shown in Table 2.

GrAMME

Since GrAMME is not designed for cancer type classification [25], we modified the original structure for multi-omics data. GrAMME is designed for a GAT model with intra-omic connections only. The implementation is a GAT model with no regularization modules.

Multi-omics GAT by Kaczmarek et al.

The multi-omics graph transformer on 12 cancer type classification is designed for gene expression and miRNA data with inter-omics connections only [7]. As shown in Table 2, the main difference between multi-omics GAT and GrAMME is the construction of the graph.

Performance on classification

For both classification tasks, the results of the proposed model and the baseline models are shown in Table 3. The proposed model with GAT layers outperforms all the baseline models for both tasks in all four metrics and the proposed model with GCN layers achieves third for the pan-cancer classification, and second for the breast cancer subtype

Table 3 Results of the proposed and baseline models with 700 genes for molecular subtype classification on the TCGA pan-cancer dataset and cancer subtype classificaiton on the TCGA BRCA dataset

Model	Pan-cancer		BRCA	
	Accu.[a]	F1	Accu.[a]	F1
Proposed w/ GAT	**83.9% ± 1.4%**	**0.84 ± 0.01**	**86.4% ± 1.9%**	**0.87 ± 0.02**
Proposed w/ GCN	81.2% ± 0.6%	0.81 ± 0.01	83.8% ± 0.9%	0.84 ± 0.01
FC-NN	78.4% ± 0.8%	0.75 ± 0.02	80.8% ± 1.1%	0.80 ± 0.02
GCN (Original) [23]	77.6% ± 0.9%	0.76 ± 0.02	82.8% ± 1.2%	0.84 ± 0.01
GCN (Modified)	78.5% ± 1.2%	0.77 ± 0.02	81.8% ± 1.4%	0.82 ± 0.01
Multi-omics GCN (Original) [1]	78.6% ± 0.9%	0.78 ± 0.01	81.8% ± 1.1%	0.82 ± 0.01
Multi-omics GCN (Modified)	80.2% ± 0.8%	0.79 ± 0.01	82.8% ± 0.9%	0.83 ± 0.01
GrAMME (Modified) [25]	81.4% ± 1.3%	0.81 ± 0.03	82.8% ± 1.6%	0.84 ± 0.03
Multi-omics GAT (Original) [7]	76.3% ± 1.2%	0.76 ± 0.02	81.8% ± 1.3%	0.82 ± 0.02
Multi-omics GAT (Modified)	79.7% ± 1.3%	0.79 ± 0.02	82.8% ± 1.4%	0.84 ± 0.02

The bold font indicates the highest values and the values after ± sign are the standard deviations.

[a] Accu. stands for Accuracy

classification. For the task of pan-cancer molecular subtype classification, the additional omic data in the modified structure improve the model performance in all three cases of the baseline model with the original structure vs. the baseline model with the modified structure. For the same task, the multi-omics GCN model with the decoder and parallel structure shows superior performance among all the baseline models that utilize GCN layers. And GrAMME, which utilizes intra-omic connections, performs better than GAT models that utilize inter-omics connections. GrAMME is the best-performing one among the baseline models for the pan-cancer task. Overall, we see the proposed model achieves the best performance for the classification task on the complex pan-cancer molecular subtype classification in all four metrics and we can conclude that more omics improve the performance of models, and the models with more restriction modules or GAT layers have better performance.

For breast cancer subtype classification, the overall trend is slightly different from that in the previous task. In most cases of including more omics, the performance of the models shows little or no improvement. We believe it is due to the nature of breast cancer taxonomy. The subtype is based on the expression level of multiple proteins. Thus, it makes the breast cancer subtype to be more closely related to the gene expression omic than the pan-cancer molecular subtype does. Such characteristic of the breast cancer subtype makes the model only using gene expression data perform very well such as the original GCN model. However, the proposed model still outperforms any baseline models by a large margin in all four metrics.

Ablation study

We conduct an ablation study to evaluate the effects of different numbers of genes, different training set splits, different combinations of modules within the model, and different combination of omics and graphs on the performance of the proposed model.

Different numbers of genes

We trained the proposed model and all baseline models at the 300 and 500 genes for pan-cancer molecular subtype classification and 300, 500, 1000, 2000, and 5000 genes for breast cancer subtype classification. The limitation of the test scope on pan-cancer classification is due to the computation constraints caused by its large number of samples. As shown in Table 4, increasing the number of gene nodes improves the performance of all models. FC-NN model demonstrates great improvement in performance as the number of genes increases. And the proposed model with the GAT layer outperforms the baseline models at both numbers of genes.

The accuracy and F1 scores of the proposed model and the baseline models for BRCA subtype classification are shown in Fig. 4. The proposed model with GAT performs best when the number of genes is smaller than 1000 and the proposed model with GCN performs best when the number of genes is larger than 1000. The proposed GAT-based model yields the best result with an accuracy of 88.9% and an F1 score of 0.89 when using 700 genes; and the proposed GCN-based model yields the best result with an accuracy of 90.1% and an F1 score of 0.90 when using 5000 genes. The detailed results are shown in the supplementary file (Additional file 1). The performance of the proposed model with GAT deteriorates beyond 1000 genes, but the performance of the proposed

Table 4 Results of the proposed model and baseline models with 300 and 500 genes for molecular subtype classification using the TCGA pan-cancer dataset

Model	300		500	
	Accu.[a]	F1	Accu.	F1
Proposed w/ GAT	**77.6% ± 1.6%**	**0.76 ± 0.02**	**81.6% ± 1.2%**	**0.80 ± 0.01**
Proposed w/ GCN	75.8% ± 1.1%	0.74 ± 0.02	80.0% ± 1.2%	0.79 ± 0.02
FC-NN	65.9% ± 1.3%	0.59 ± 0.04	77.5% ± 1.4%	0.74 ± 0.02
GCN (Original)	74.5% ± 1.6%	0.72 ± 0.05	76.1% ± 1.3%	0.73 ± 0.03
GCN (Modified)	75.5% ± 1.4%	0.72 ± 0.03	77.9% ± 1.1%	0.77 ± 0.02
Multi-omics GCN (Original)	76.4% ± 1.3%	**0.76 ± 0.03**	77.4% ± 1.3%	0.77 ± 0.03
Multi-omics GCN (Modified)	77.4% ± 1.3%	**0.76 ± 0.02**	78.2% ± 1.2%	0.75 ± 0.02
GrAMME (Modified)	77.4% ± 1.5%	**0.76 ± 0.02**	79.6% ± 1.4%	0.79 ± 0.02
Multi-omics GAT (Original)	73.4% ± 1.8%	0.71 ± 0.04	75.1% ± 1.5%	0.74 ± 0.04
Multi-omics GAT (Modified)	75.8% ± 1.5%	0.74 ± 0.04	77.4% ± 1.3%	0.74 ± 0.02

The bold font indicates the highest values and the values after ± *sign are the standard deviations*

[a] Accu. stands for Accuracy

model with GCN continues to rise as the number of genes grows beyond 1000 genes. All GAT-based baseline models show similar deterioration around 1000 genes. We think the high computation cost of the GAT-based model can cause it to perform worse on a large graph than on a small graph. Overall, we can conclude that the proposed model with GCN layers scales better than that with GAT layers at a large number of genes.

In the process of testing the models on a large graph, we also find that a GAT-based model is more stable on a smaller learning rate compared to a GCN-based model. We believe it is caused by GAT's high computation costs since a high learning rate may cause the model to be stuck in a local optimum.

Overall, we see the proposed model achieves the best performance and scales well with a larger number of genes. We can also conclude that more genes and more omics mostly improve the performance of models, the models with more modules have better performance, and GAT-based models perform better with smaller graphs while GCN-based models scale better at larger graphs.

Different training set split

To examine the performance of the proposed model on a complex dataset with a smaller training set, we tested the model on the Pan-cancer dataset using three different training set splits. This approach was taken to mimic situations where only a smaller labeled dataset is available in the real world. The training set splits were set at 70%, 60%, and 50%, with corresponding testing set splits of 20%, 30%, and 40%. Throughout these tests, the validation set split was consistently kept at 10%.

As shown in Table 5, the proposed model with the GAT layer exhibits a slight performance deterioration at 70% and 60% training set splits. However, it displays a more pronounced decline in classification accuracy at 50%. In contrast, the proposed model with the GCN layer demonstrates consistent and robust performance across all three training-validation-testing splits. However, its classification accuracy is lower than that of the model with the GAT layer at 70% and 60% training set splits. Therefore, we can conclude that the proposed model with the GAT layer achieves
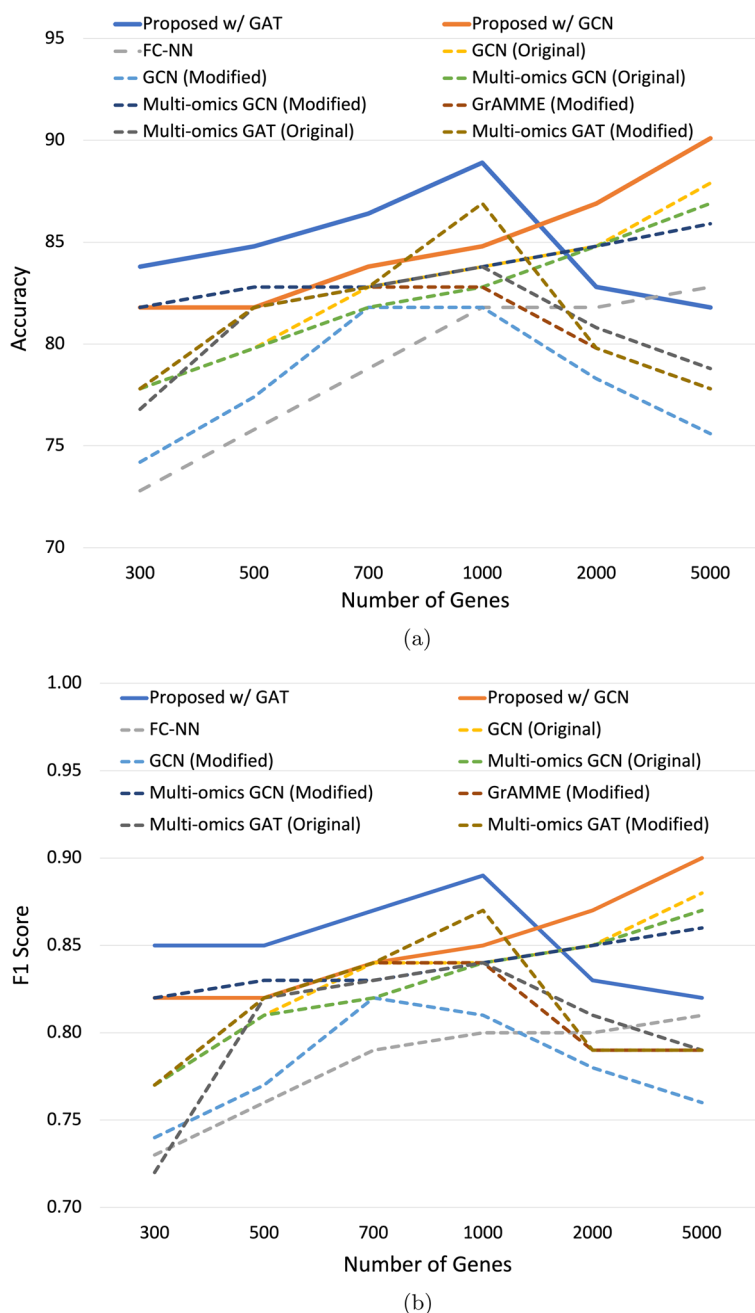
(a)

(b)

Fig. 4 Performance of the Proposed Models and Baseline Models with Different Numbers of Genes on BRCA Dataset. (**a**) The accuracy of the proposed model with GAT (blue solid line) or GCN (orange solid line) and baseline models (dashed line) are plotted against different numbers of genes (300, 500, 700, 1000, 2000, and 5000) for BRCA subtype classification. (**b**) The F1 scores of the proposed model with GAT (blue solid line) or GCN (orange solid line) and baseline models (dashed line) are plotted against different numbers of genes (300, 500, 700, 1000, 2000, and 5000) for BRCA subtype classification

superior performance compared to the model with the GCN layer when the training set is relatively small. However, the model with the GCN layer outperforms at a very small training set (50%). Overall, the proposed model with the GCN layer offers more robust classification performance with smaller training sets.

Table 5 Proposed Model with Different Training-validation-testing Split

Model	Training set ratio					
	70%		60%		50%	
	Accu.[a]	F1	Accu.[a]	F1	Accu.[a]	F1
Proposed w/ GAT	82.5% ± 1.5%	0.82 ± 0.02	79.9% ± 4.0%	0.78 ± 0.06	74.2% ± 7.5%	0.71 ± 0.10
Proposed w/ GCN	77.9% ± 1.2%	0.76 ± 0.02	76.7% ± 0.4%	0.75 ± 0.01	77.3% ± 2.5%	0.76 ± 0.03

The values after ± sign are the standard deviations

[a] Accu. stands for Accuracy

Different combinations of modules

To examine the effect of different modules within the proposed model, we test three different variants of the proposed model for the Pan-cancer molecular subtype classification. All variants of the proposed model are trained with all three omics data at 300, 500, and 700 genes. The proposed model without the decoder acts as a parallel structured GNN model, the proposed model without the parallel structure acts as a graph autoencoder model, and the proposed model without both the decoder and the parallel structure acts as a graph-classification GNN model.

As shown in Table 6, models without the parallel structure perform poorly compared to those without the decoder at any number of genes in general. It shows that the parallel structure plays an important role in feature extraction, which also demonstrates the benefit of including both local features and global features. When the graph size is small (300 genes), the model without the decoder and the parallel structure performs more poorly compared to those with either component. However, when the graph size is large enough (500 genes and 700 genes), the model without the decoder and the parallel structure performs relatively the same compared to those with either of the component. We believe the extra information in the large graph compensates for the loss in performance caused by the exclusion of either the decoder or the parallel structure.

Table 6 Results of the Variants of the Proposed Model for Molecular Subtype Classification Using the TCGA Pan-cancer Dataset

GNN layers (Module)	300		500		700	
	Accu.[a]	F1	Accu.[a]	F1	Accu.[a]	F1
GAT (No Decoder)	**76.3% ± 1.6%**	**0.76 ± 0.03**	**78.2% ± 1.2%**	**0.77 ± 0.01**	**80.2% ± 1.2%**	**0.79 ± 0.01**
GCN (No Decoder)	75.3% ± 1.2%	0.74 ± 0.02	76.8% ± 0.8%	0.75 ± 0.01	79.3% ± 0.8%	0.78 ± 0.01
GAT (No Parallel)	75.4% ± 1.8%	0.73 ± 0.03	76.1% ± 1.7%	0.73 ± 0.02	79.8% ± 1.3%	0.78 ± 0.02
GCN (No Parallel)	73.5% ± 1.2%	0.72 ± 0.02	75.4% ± 1.2%	0.73 ± 0.01	76.7% ± 0.8%	0.75 ± 0.01
GAT (No Decoder & Parallel)	74.9% ± 1.4%	0.73 ± 0.02	76.4% ± 0.9%	0.74 ± 0.01	80.1% ± 0.8%	0.79 ± 0.01
GCN (No Decoder & Parallel)	73.1% ± 1.2%	0.73 ± 0.02	75.6% ± 0.8%	0.73 ± 0.01	77.3% ± 0.02%	0.76 ± 0.01

The bold font indicates the highest values and the values after ± sign are the standard deviations.

[a] Accu. stands for Accuracy

Table 7 Results of the proposed model on different combinations of omics and networks at 500 genes using the TCGA pan-cancer dataset

Data	Network	GAT		GCN	
		Accu.[e]	F1	Accu.[e]	F1
mRNA[a]	Intra-omic[c]	77.0% ± 1.9%	0.75 ± 0.03	76.1% ± 0.9%	0.73 ± 0.01
miRNA[b]	Intra-omic[d]	74.0% ± 0.4%	0.70 ± 0.01	68.2% ± 4.1%	0.63 ± 0.04
mRNA+CNV[a]	Intra-omic[c]	79.1% ± 1.4%	0.77 ± 0.03	77.1% ± 0.7%	0.76 ± 0.01
mRNA+miRNA	Inter-omic	76.1% ± 1.6%	0.73 ± 0.03	75.4% ± 0.7%	0.73 ± 0.01
	Intra-omic	77.3% ± 1.6%	0.75 ± 0.03	76.8% ± 0.7%	0.74 ± 0.01
mRNA+CNV+miRNA	Inter-omic	80.3% ± 1.6%	0.80 ± 0.02	77.4% ± 0.6%	0.74 ± 0.01
	Intra-omic	**80.5% ± 1.2%**	**0.80 ± 0.02**	**78.2% ± 0.6%**	**0.75 ± 0.01**

The bold font indicates the highest values and the values after ± sign are the standard deviations.

[a] Data contains no miRNA-based nodes, so only 500 gene nodes in the graph

[b] Data contains no gene-based nodes, so only 100 miRNA nodes in the graph

[c] The graph contains only gene-gene connections

[d] The graph contains only miRNA-miRNA meta-path connections

[e] Accu. stands for accuracy

Different combination of omics and graphs

To test the effect of different choices of omics and different graphs, we generate five different combinations of omics. The five combinations of omics are mRNA, miRNA, mRNA + CNV, mRNA + miRNA, and mRNA + CNV + miRNA. For mRNA + miRNA and mRNA + CNV + miRNA, two different variants of graphs are also tested. All models are conducted for Pan-cancer molecular subtype classification, and trained with 500 genes except for only miRNA omic, which contains only 100 miRNA nodes. As shown in Table 7, the best-performing setting is mRNA + CNV + miRNA with intra-omic edges for both GAT-based and GCN-based models. The worst-performing setting is miRNA, which has the smallest graph size and information. Models on mRNA + CNV perform better than those on mRNA + miRNA, but adding miRNA to mRNA + CNV (mRNA + CNV + miRNA setting) still improves the model performance. Models with intra-omic graph performs slightly better than models with inter-omics graph. The performance difference across different settings is the same for both GAT-based and GCN-based models.

Conclusion

In this study, we propose a novel end-to-end multi-omics GNN framework for accurate and robust cancer subtype classification. The proposed model utilizes multi-omics data in the form of a heterogeneous multi-layer graph, which is the supra-graph built from GGI network, miRNA-gene target network, and miRNA meta-path. While GNNs have been previously employed for genomics data analysis, our model's novelty lies in the utilization of a heterogeneous multi-layer multiomics supra-graph. The supra-graph not only incorporates inter-omics and intra-omic connections from established biological knowledge but also integrates genomics, transcriptomics, and epigenomics data into a single graph, providing a novel advancement in cancer subtype classification. The proposed model outperforms all four baseline models for cancer molecular subtype classification. We do a thorough comparative analysis of GAT

and GCN-based models at different numbers of gene settings, different combinations of omics, and different graphs.

Comparing the proposed model to the baseline models, it achieves the best performance for cancer molecular subtype classification and BRCA subtype classification. The proposed model with GAT layers performs better than that with GCN layers at smaller-size graphs (smaller than 1000 genes). However, the performance of the GAT-based model deteriorates as the size of the graph grows beyond a certain threshold. On the other hand, the performance of the GCN-based model continues to improve as the size of the graph grows. Therefore, we can conclude that a GAT-based model is more suitable on a smaller graph, where it has a higher feature extraction ability and its computation cost isn't that high yet.

By studying the effect of different modules within the proposed model and different combinations of omics, we find the addition of a decoder and the parallel structure, and including other omics improves the performance of the proposed model. The benefit of using parallel structure outweighs that of decoder, especially on smaller-size graphs, and the benefit of adding CNV is higher than that of adding miRNA. We also find that using a graph with only intra-omic edges yields a better performance than using a graph with only inter-omics edges, which agrees with the results from the previous study [7].

The proposed model also has some limitations. We investigate only two well-established and widely adopted GNN models. New models are emerging with the recent blooming of studies in GNN models. As the size of the graph grows or more omics are added, GAT-based models become more sensitive to parameters and take a much longer time to train. It is our future research direction to overcome such limitations. The proposed model for cancer subtype classification depends on labeled data, which is costly to annotate and difficult to obtain in the real world. Exploring unsupervised learning for cancer subtype detection is also a direction we aim to pursue in our future research.

In summary, incorporating gene-based and non-gene-based omic data in the form of a supra-graph with inter-omics and intra-omic connections improves the cancer subtype classification. The GAT-based model is preferable on smaller graphs than the GCN-based model. GCN-based model is preferable when dealing with large and complex graphs.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05622-4.

> **Additional file 1**. Detailed Results and Model Settings.

Availability of data and materials
TCGA Pan-cancer dataset and TCGA BRCA dataset are both obtained from Xena database (https://xenabrowser.net), the detailed link for TCGA Pan-cancer dataset is (https://xenabrowser.net/datapages/?cohort=TCGA

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References

1. Li B, Wang T, Nabavi S. Cancer molecular subtype classification by graph convolutional networks on multi-omics data. In: Proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics, BCB 2021 2021, vol. 1. https://doi.org/10.1145/3459930.3469542
2. Zhang X, Zhang J, Sun K, Yang X, Dai C, Guo Y. Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In: Proceedings—2019 IEEE international conference on bioinformatics and biomedicine, BIBM 2019, 2019; pp. 765–769 https://doi.org/10.1109/BIBM47256.2019.8983228
3. Yang B, Zhang Y, Pang S, Shang X, Zhao X, Han M. Integrating multi-omic data with deep subspace fusion clustering for cancer subtype prediction. IEEE/ACM Trans Comput Biol Bioinform. 2019;18(1):216–26. https://doi.org/10.1109/tcbb.2019.2951413.
4. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. Moli: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35:501–9. https://doi.org/10.1093/bioinformatics/btz318.
5. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, Huang K. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12:3445. https://doi.org/10.1038/s41467-021-23774-w.
6. Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using multi-view factorization autoencoder (mae). BMC Genomics. 2019;20:1–11. https://doi.org/10.1186/s12864-019-6285-x.
7. Kaczmarek E, Jamzad A, Imtiaz T, Nanayakkara J, Renwick N, Mousavi P. Multi-omic graph transformers for cancer classification and interpretation. Pac Symp Biocomput. 2022;27:373–84.
8. Lotfollahi M, Litinetskaya A, Theis FJ. Multigrate : single-cell multi-omic data integration, 1–5 2022; https://doi.org/10.1101/2022.03.16.484643
9. Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, Zhang J, Salama P, Rizkalla M, Han Z, Huang K. Salmon: survival analysis learning with multi-omics neural networks on breast cancer. Front Genet. 2019;10:1–13. https://doi.org/10.3389/fgene.2019.00166.
10. Bai J, Li B, Nabavi, S. Semi-supervised classification of disease prognosis using cr images with clinical data structured graph. In: Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics, 2022; pp. 1–9
11. Chai H, Zhou X, Zhang Z, Rao J, Zhao H, Yang Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Comput Biol Med. 2021;134:104481.
12. Heo YJ, Hwa C, Lee GH, Park JM, An JY. Integrative multi-omics approaches in cancer research: from biological networks to clinical subtypes. Mol Cells. 2021;44:433–43. https://doi.org/10.14348/molcells.2021.0042.
13. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44.
14. Mateo J, Steuten L, Aftimos P, André F, Davies M, Garralda E, Geissler J, Husereau D, Martinez-Lopez I, Normanno N, et al. Delivering precision oncology to patients with cancer. Nat Med. 2022;28(4):658–65.
15. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173(2):291–304.
16. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. Adv Neural Infn Process Syst. 2016;29:3844–52.
17. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.
18. He S, Pepin L, Wang G, Zhang D, Miao F. Data-driven distributionally robust electric vehicle balancing for mobility-on-demand systems under demand and supply uncertainties. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS), 2020; IEEE, pp. 2165–2172
19. Wang T, Li B, Nabavi S. Single-cell RNA sequencing data clustering using graph convolutional networks. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021; IEEE, pp. 2163–2170
20. Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Front Oncol. 2020;10:1030.
21. Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst. 2020;32(1):4–24.
22. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv, 2017; 1–12
23. Ramirez R, Chiu Y-C, Hererra A, Mostavi M, Ramirez J, Chen Y, Huang Y, Jin Y-F. Classification of cancer types using graph convolutional neural networks. Front Phys. 2020;8:203.

24. Wang T, Bai J, Nabavi S. Single-cell classification using graph convolutional networks. BMC Bioinformat. 2021;22(1):1–23.
25. Shanthamallu US, Thiagarajan JJ, Song H, Spanias A. Gramme: semisupervised learning using multilayered graph attention models. IEEE Trans Neural Netw Learn Syst. 2020;31:3977–88. https://doi.org/10.1109/TNNLS.2019.2948797.
26. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. Breast cancer subtypes based on er/pr and her2 expression: comparison of clinicopathologic features and survival. Clin Med Res. 2009;7(1–2):4–13.
27. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, Dolma S, Coulombe-Huntington J, Chatr-Aryamontri A, Dolinski K, Tyers M. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30(1):187–200.
28. Chen Y, Wang X. mirdb: an online database for prediction of functional microrna targets. Nucleic Acids Res. 2020;48(D1):127–31.
29. Lee B, Zhang S, Poleksic A, Xie L. Heterogeneous multi-layered network model for omics data integration and analysis. Front Genet. 2020;10:1–11. https://doi.org/10.3389/fgene.2019.01381.
30. 13, B..W.H..H.M.S.C.L...P.P.J..K.R., data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, G., for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, I., et al.: Comprehensive molecular portraits of human breast tumours. Nature 490(7418), 61–70 (2012)
31. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, Banerjee A, Luo Y, Rogers D, Brooks AN, et al. Visualizing and interpreting cancer genomics data via the xena platform. Nat Biotechnol. 2020;38(6):675–8.

Publisher's Note