BIBS seminar

Molecular Subtyping of Cancer Based on Robust Graph Neural Network and Multi-Omics Data Integration

2023. 9. 7

Sangyeon Shin

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 13 May 2022
doi: 10.3389/fgene.2022.884028

Check for updates

Molecular Subtyping of Cancer Based on Robust Graph Neural Network and Multi-Omics Data Integration

Chaoyi Yin[1†], Yangkun Cao[1†], Peishuo Sun[1], Hengyuan Zhang[1], Zhi Li[2], Ying Xu[3] and Huiyan Sun[1]**

[1]School of Artificial Intelligence, Jilin University, Changchun, China, [2]Department of Medical Oncology, the First Hospital of China Medical University, Shenyang, China, [3]Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA, United States

Contents

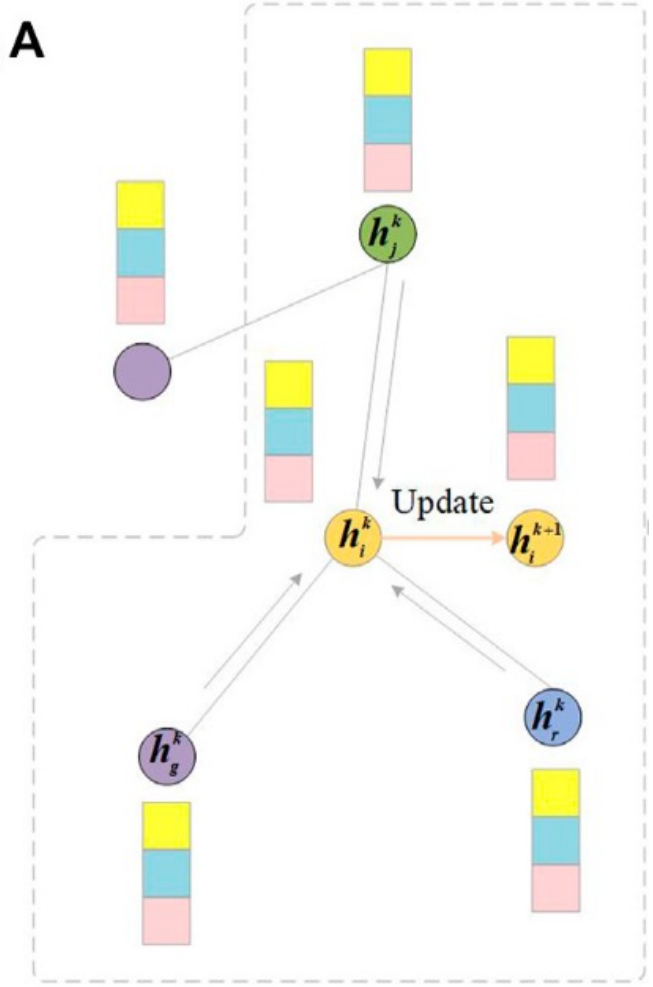
Introduction

Heterogeneity among cancers

- Cancer is a complex and highly individualized disease with diverse subtypes.
- **Molecular heterogeneity exists among different subtypes** of the same cancer type.
- As cancer patients of distinct molecular subtypes usually **respond differently** to same treatment.
- So accurate subtype classification can not only assist **precision diagnosis** but also facilitate **effective targeted treatment**.

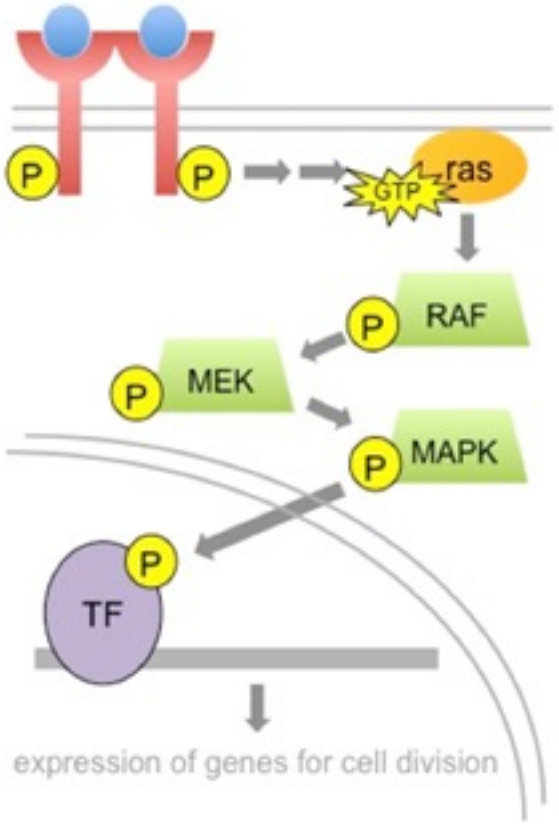
GNNs for a node classification task

- With the strong representation ability of graph-structured data, **graph neural networks (GNNs)** have achieved great success and are gradually used in **a node classification task**.
- It provides one way to obtain new representations of nodes by combining the connectivity and features of its local neighborhood.
- Although GNN are powerful, they are vulnerable when the skeleton of the graph and nodes' feature are **mixed with noise.**
- So a **robust** GNN model is necessary for accurately and stably predicting cancer subtypes.

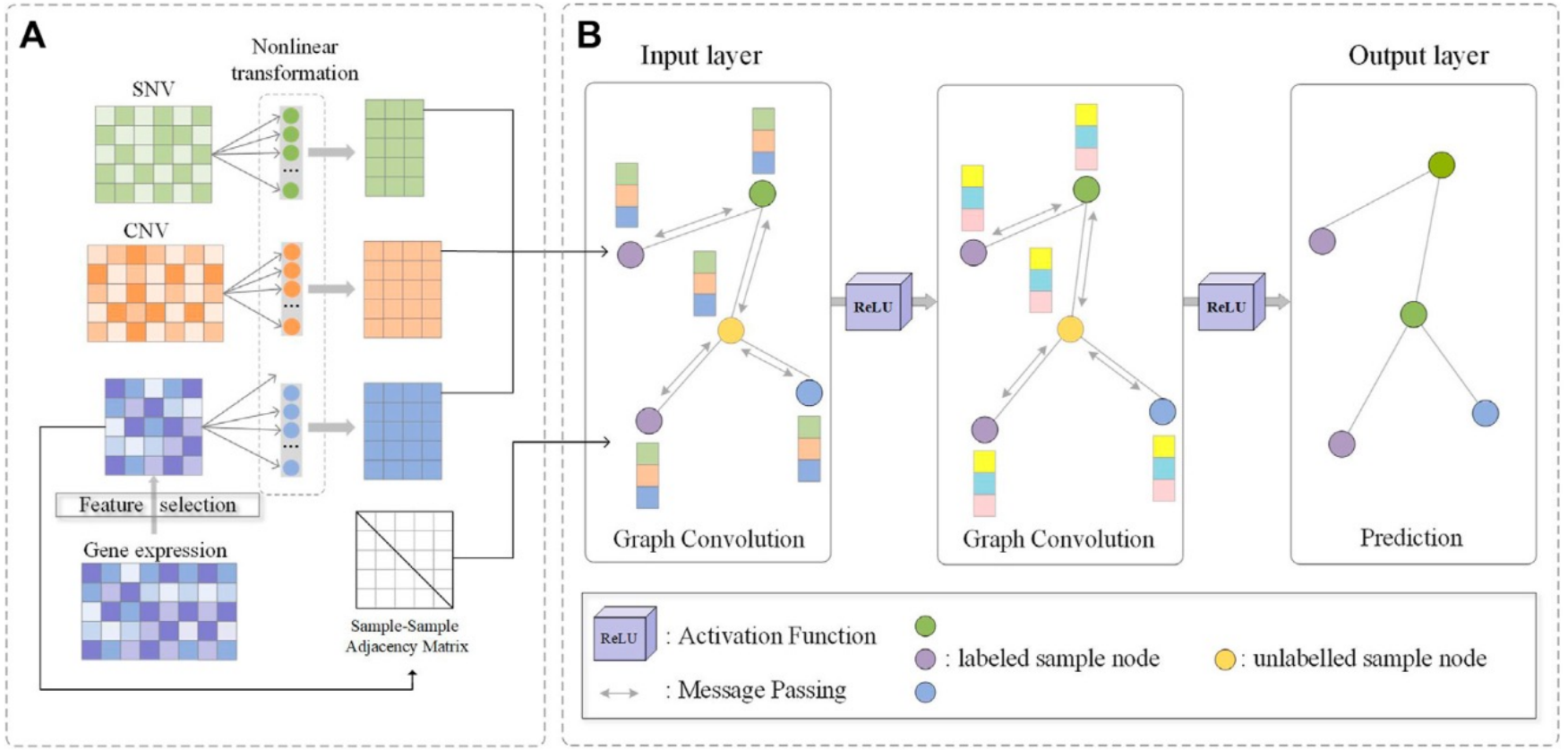


Multi-Omics data integration for cancer

- It is well known that abnormal behaviors of cancer cells are the result of a series of **gene mutations, gene copy number variation, and gene transcription level changes** in key regulatory pathways.
- **Integrating multiple types of omics data** can provide a view to better understand the **interrelationships** of the involved biomolecules and their functions.
- Multi-omics data integration improves the prediction accuracy of patients' clinical outcome.



M-GCN



- Using the Graph
- Removing noise
- Multi-Omics
- Predicting cancer subtypes

Materials and Methods

Materials

- TCGA, The Cancer Genome Atlas database
  - Breast Cancer (BRCA, n = 518, Vuong et al., 2014)
    - Estrogen receptor positive (ER+)
    - Human epidermal growth factor receptor 2 positive (HER2+)
    - Triple-negative breast cancer (TNBC)
  - Stomach adenocarcinoma (STAD, n = 221, Bass et al., 2014)
    - Chromosomal instability (CIN)
    - Epstein-Barr virus (EBV)
    - Microsatellite instability (MSI)
    - Genomically stable (GS)
- Consist of Gene expression, SNV(single nucleotide variants), CNV(copy number variation)

Preprocessing

TABLE 1 | Dataset attributes.

Cancer	#Subtype	#Samples of each subtype	#CNV features	#SNV features	# Gene expression features
BRCA	ER+	386	74	62	124
	HER2+	35			
	TNBC	97			
STAD	CIN	107	169	166	128
	EBV	23			
	MSI	46			
	GS	45			

Gene expression

✓ Filtering: expression value more than 10(BRCA) and 3(STAD)

✓ Normalize with FPKM(fragments per kilobase of exon per million fragments) and log2 transformation

Preprocessing

TABLE 1 | Dataset attributes.

Cancer	#Subtype	#Samples of each subtype	#CNV features	#SNV features	# Gene expression features
BRCA	ER+	386	74	62	124
	HER2+	35			
	TNBC	97			
STAD	CIN	107	169	166	128
	EBV	23			
	MSI	46			
	GS	45			

SNV(Single Nucleotide Variant)

✓ Filtering: Mutation frequency more than 0.03(BRCA) and 0.1(STAD)

CNV(Copy Number Variation)

✓ Filtering: genes having significant amplifications or deletions rates

Methods

- ✓ Feature Selection
- ✓ Sample-Sample Similarity Graph Construction
- ✓ GCN Model Integrating Multi-Omics Data for Sample Classification
  - Multi-Omics Data Features Transformation
  - Neighbor Importance Estimation
  - Layer-Wise Graph Memory
  - Node Aggregation with Multi-View Representations Based on GCN
  - Loss and Optimization
- ✓ Evaluation Metrics

Feature Selection

✓ Low noise for constructing a purified sample-sample similarity graph and effective message passing

✓ **HSIC Lasso**(Yamada et al., 2014)

- supervised non-linear feature selection method

$$\min_{\gamma \in \mathbb{R}^d} \frac{1}{2} \left\| \bar{L} - \sum_{l=1}^{d} \gamma_l \bar{K}^{(l)} \right\|_{Frob}^2 + \lambda \|\gamma\|_1, \text{s.t. } \gamma_1, \gamma_2, \ldots, \gamma_d \geq 0, \quad (1)$$

Regression coefficient

Frobenius norm

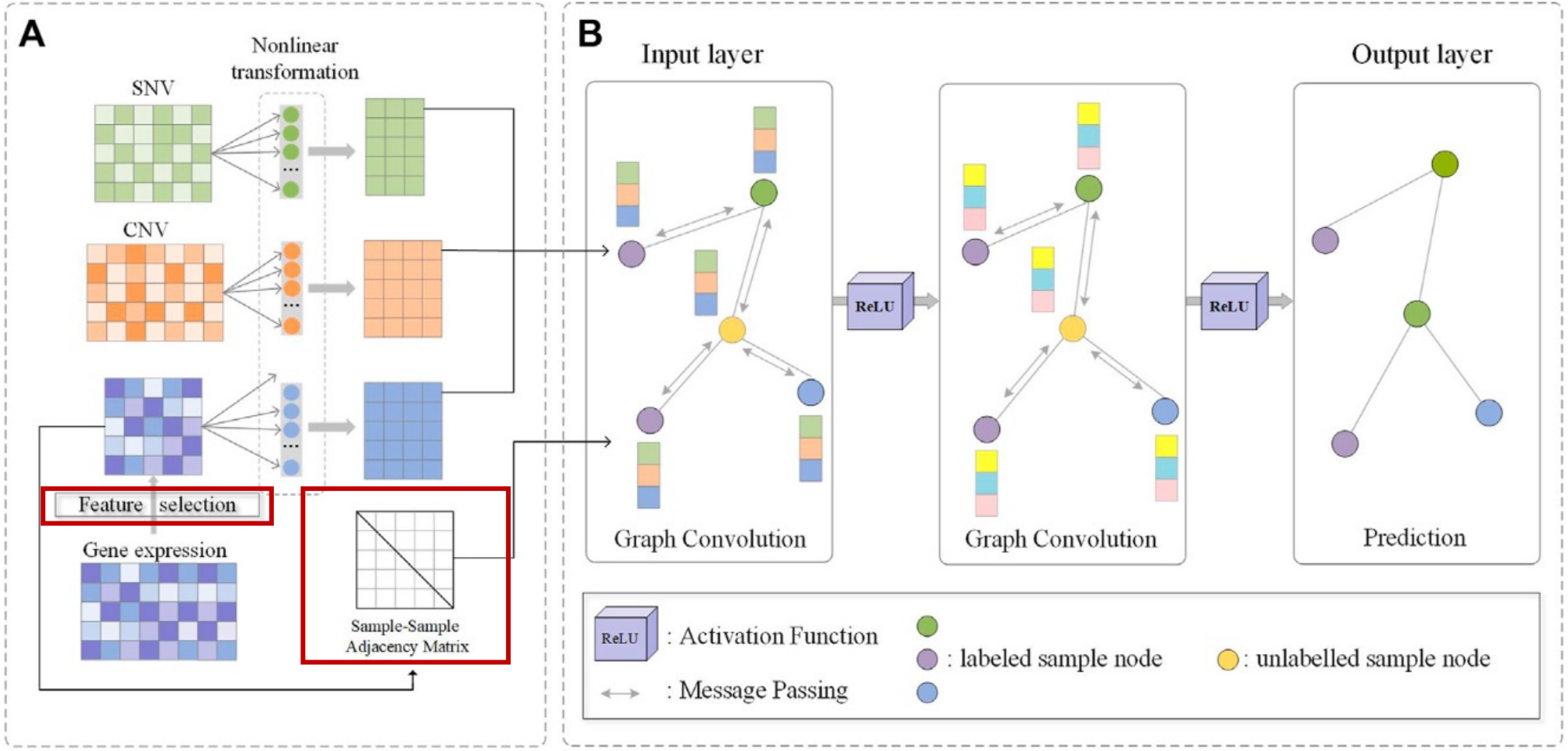
Regression coefficient of *l*-th feature

λ: regularization parameter

$$\bar{L} = \Gamma L \Gamma$$
$$\Gamma = I_n - \frac{1}{n} 1_n 1_n^T$$
$$\bar{K}^{(l)} = \Gamma K^{(l)} \Gamma$$

Centered Gram matrix, Kernel Function

Feature Selection



Sample-Sample Similarity Graph Construction

✓ Using Transcriptomic data and spearman's correlation

✓ Generate adjacency matrix $A \in \mathbb{R}^{n \times n}$

$$A_{ij} = \begin{cases} 1, & \rho_{ij} \geq r, p \leq 0.05 \\ 0, & others \end{cases}, \qquad (2)$$

r: threshold of correlation coefficient
p: p value of the correlation

✓ The node's feature matrix X

$$X = [X_m, X_c, X_e] \begin{cases} X_m \in \mathbb{R}^{n \times f_1} & \text{SNV feature matrix} \\ X_c \in \mathbb{R}^{n \times f_2} & \text{CNV feature matrix} \\ X_e \in \mathbb{R}^{n \times f_3} & \text{Gene expression feature matrix} \end{cases}$$

Multi-Omics Data Features Transformation

- ✓ Non-linear transformations to improve samples' feature representations
- ✓ ReLU activation function

ReLU

$$H^0 = [\sigma(X_m W_m), \sigma(X_c W_c), \sigma(X_e W_e)], \quad (3)$$

Latent feature matrix

Learnable non-linear transformation

- $H^0 \in \mathbb{R}^{n \times (f_1' + f_2' + f_3')}$

Final output multi-view representations of samples

GNNGUARD

✓ When a new sample is concatenated, some noises could be introduced.

✓ To mitigate the impact of noises, GNNGUARD was used.

- It improves robustness of GCN models by **detecting fake edges** of graph structure and **removes their weights** in message passing of GCN.

- Implemented by '**neighbor importance estimation**' and '**layer-wise graph memory**'

Neighbor Importance Estimation

- ✓ For quantify the relevance between node *i* and node *j*
- ✓ Evaluate the importance weight of each edge based on similarity

$$s_{ij}^k = \left(h_i^k \odot h_j^k\right) / \left(\| h_i^k \|_2 \| h_j^k \|_2 \right), \quad (4)$$

Cosine similarity

$$\alpha_{ij}^k = \begin{cases} s_{ij}^k / \Sigma_{j \in N_i^*} s_{ij}^k \times \hat{N}_i^k / \left(\hat{N}_i^k + 1\right) & if\ i \neq j \\ 1 / \left(\hat{N}_i^k + 1\right) & if\ i = j \end{cases}, \quad (5)$$

Normalized by neighbor node $\left(\hat{N}_i^k = \Sigma_{j \in N_i^*} \| s_{ij}^k \|_0\right)$

$$1_{P_0}\left(\sigma\left(c_{ij}^k W\right)\right) = \begin{cases} 0 & if\ \sigma\left(c_{ij}^k W_n\right) < P_0 \\ 1 & otherwise \end{cases}, \quad (6)$$

Edge pruning

$$\hat{\alpha}_{ij}^k = \alpha_{ij}^k 1_{P_0}\left(\sigma\left(c_{ij}^k W_n\right)\right). \quad (7)$$

Edge pruning

Layer-Wise Graph Memory

✓ Because the weighted graph changes in each layer, for a stable training to keep partial memory

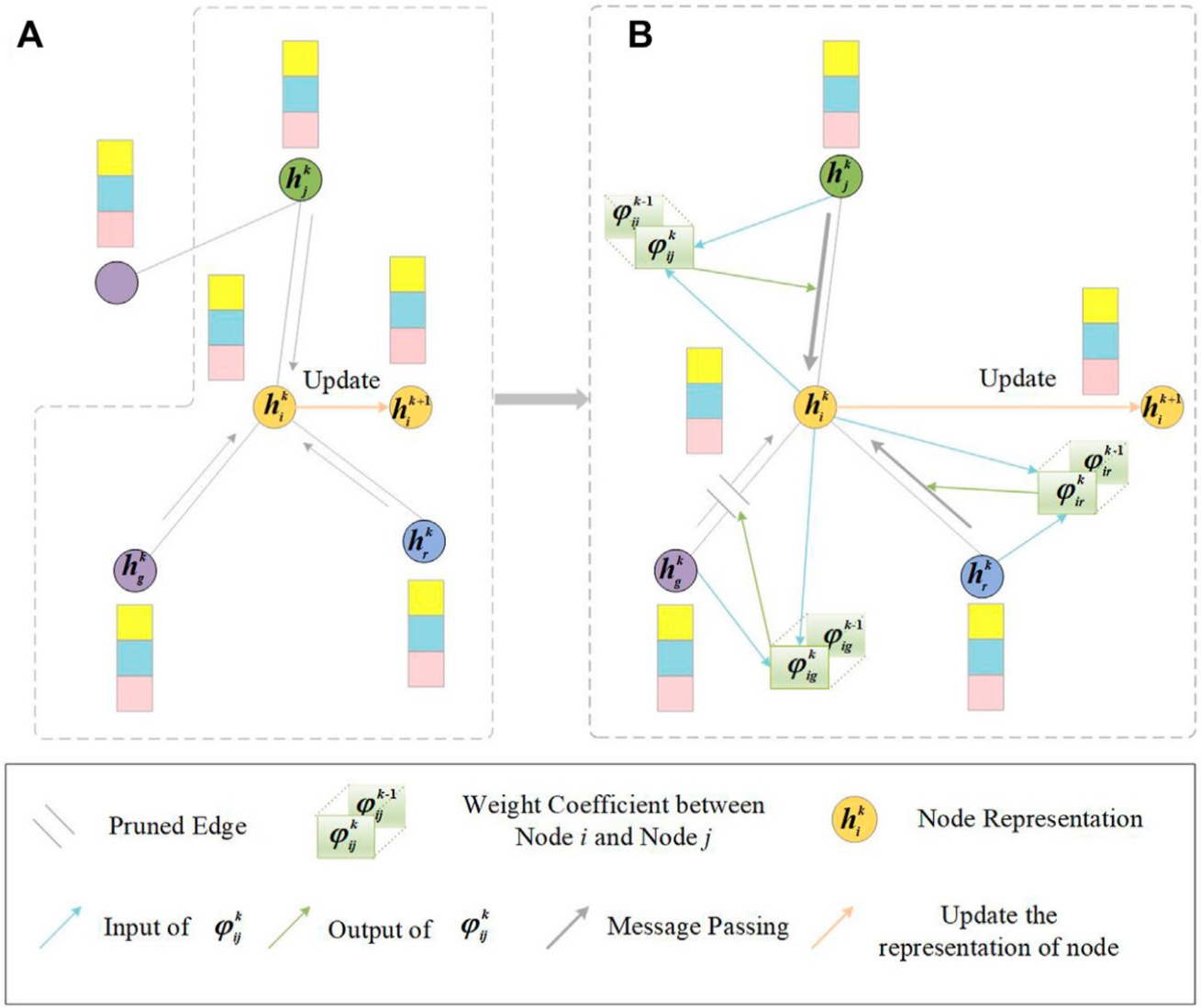
Weight for edge e

$$\varphi_{ij}^{k} = \beta\varphi_{ij}^{k-1} + (1-\beta)\hat{\alpha}_{ij}^{k}, \quad \beta \in [0, 1] \qquad (8)$$

✓ Beta is a learnable parameter

✓ The M-GCN model can learn more robust and informative representations

GNNGUARD


A
B
Update
Update
Pruned Edge
Weight Coefficient between Node i and Node j
Node Representation
Input of
Output of
Message Passing
Update the representation of node

Node Aggregation Based on GCN

✓ Learn comprehensive representations of sample nodes and multi-omics data

$$H^{k+1} = \sigma\left(\hat{A}^{k} H^{k} W^{k}\right), \qquad (9)$$

$\hat{A}^{k} = \tilde{D}^{k-\frac{1}{2}} \tilde{A}^{k} \tilde{D}^{k-\frac{1}{2}}$ **Normalization with division by Degree matrix D**

$\tilde{D}_{ii}^{k} = \sum_{j} \tilde{A}_{ij}^{k}$ **Degree matrix, sum of the edges that is connected to node *i***

✓ Neighbor information with adjacency matrix and weights of the edges

✓ Update the feature vector in the current layer based on previous layer

$$P = \mathrm{softmax}\left(H^{k}\right), \qquad P \in \mathbb{R}^{n \times V} \qquad (10)$$

✓ Calculate the probability of which molecular subtyping of each sample belongs to

Loss and Optimization

- ✓ Cross-entropy is used as the loss function

$$\mathcal{L} = -\frac{1}{n}\Sigma_i \Sigma_{v=1}^{V} y_{iv} \log(\boldsymbol{P}_{iv}), \tag{11}$$

- ✓ V is the number of molecular subtypes
- ✓ y is the ground truth label of i-th sample

Results

Subtype Classification Performance

✔ Compare the performance of M-GCN with six methods on STAD and BRCA

- K-nearest neighbor classifier (KNN)
- Random Forest (RF)
- Support vector machine (SVM)
- Gaussian naïve Bayse (GNB)
- DeepCC: neural network-based method with transcriptomic data
- Li's method: GCN-based methods which integrates CNV and gene expression

Figure 3.

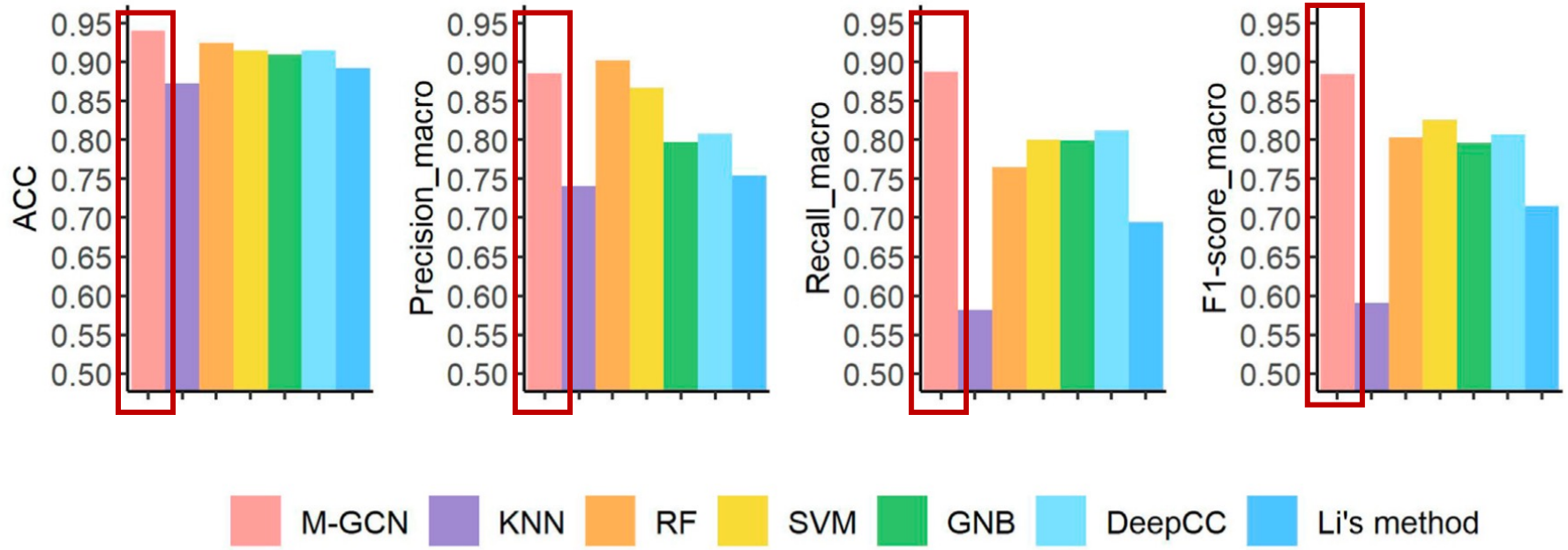


FIGURE 3 | Prediction performance under four evaluation metrics of seven methods in the BRCA dataset. Pink bar represents the final performance of M-GCN; purple bar, orange bar, yellow bar, and green bar refer to the performance of KNN, RF, SVM, and GNB, respectively. Light blue bar is for the performance of DeepCC, and dark blue bar is for the performance of Li's method.

✓ Compare the performance of M-GCN with six methods on BRCA

Table 2.

TABLE 2 | Classification results of M-GCN on each subtype of BRCA.

	Ratio predicted as ER+ (%)	Ratio predicted as HER2+ (%)	Ratio predicted as TNBC (%)
ER+	**95.9**	0.51	3.59
HER2+	0	**80**	20
TNBC	7	3	**90**

The meaning of the bold values provided in Tables 2 and 3 is "the highest prediction ratio in each subtype".

Figure 4.

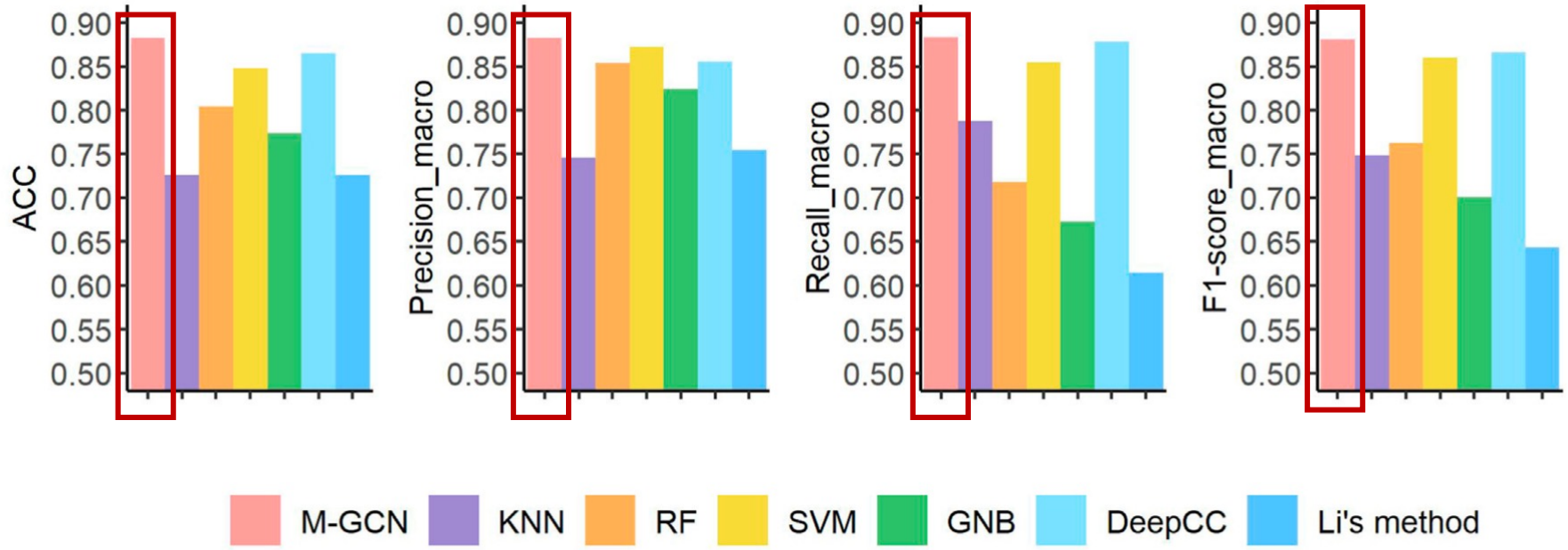


FIGURE 4 | Prediction performance under four evaluation metrics of seven methods in the STAD dataset. Pink bar represents the final performance of M-GCN; purple bar, orange bar, yellow bar, and green bar refer to the performance of KNN, RF, SVM, and GNB, respectively. Light blue bar is for the performance of DeepCC, and dark blue bar is for the performance of Li's method.

✓ Compare the performance of M-GCN with six methods on STAD

Table 3.

TABLE 3 | Classification results of M-GCN on each subtype of STAD.

	Ratio predicted as CIN (%)	Ratio predicted as EBV (%)	Ratio predicted as MSI (%)	Ratio predicted as GS (%)
CIN	**93.64**	0	2.72	3.64
EBV	0	**100**	0	0
MSI	6	0	**90**	4
GS	20	4	6	**70**

Figure 5. Contribution of Each Element in STAD

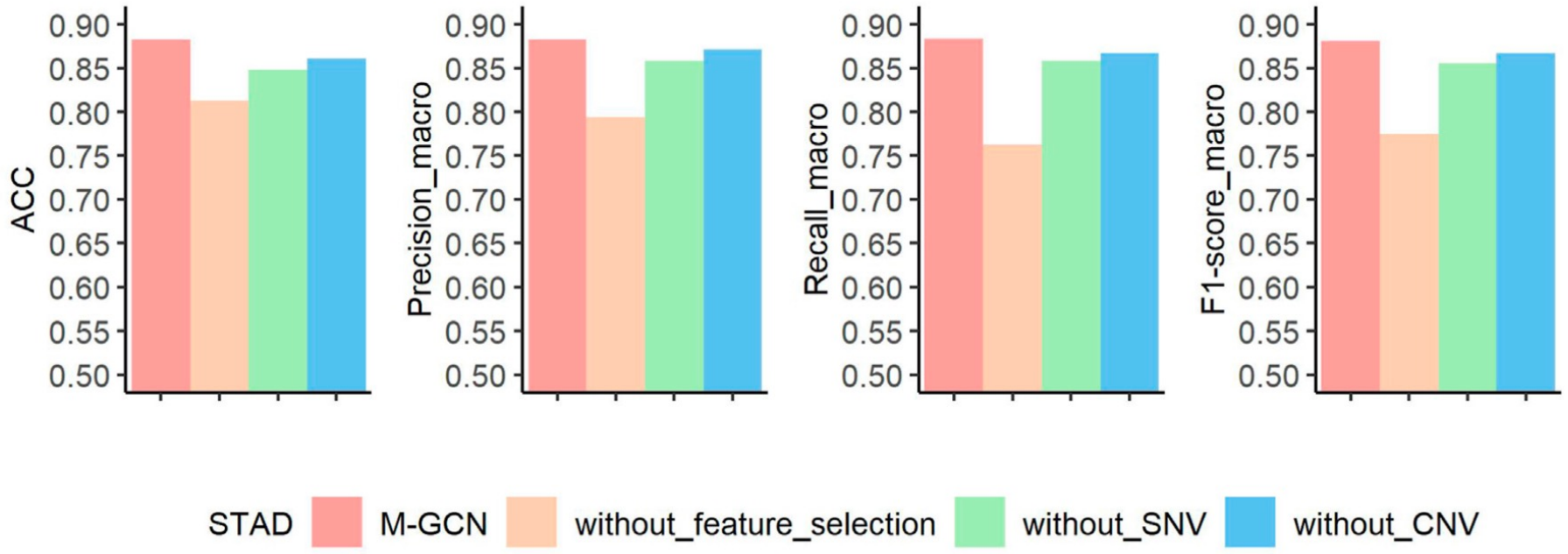


FIGURE 5 | Results of ablation experiment of M-GCN in STAD. Pink bar represents the final performance of M-GCN, orange bar represents the performance of M-GCN without feature selection, green bar is for the performance of M-GCN without SNV features, and blue bar is for the performance of M-GCN without CNV features.

✓ Conduct three ablation experiments

- feature selection step - SNV data - CNV data

Figure 6. Contribution of Each Element in BRCA

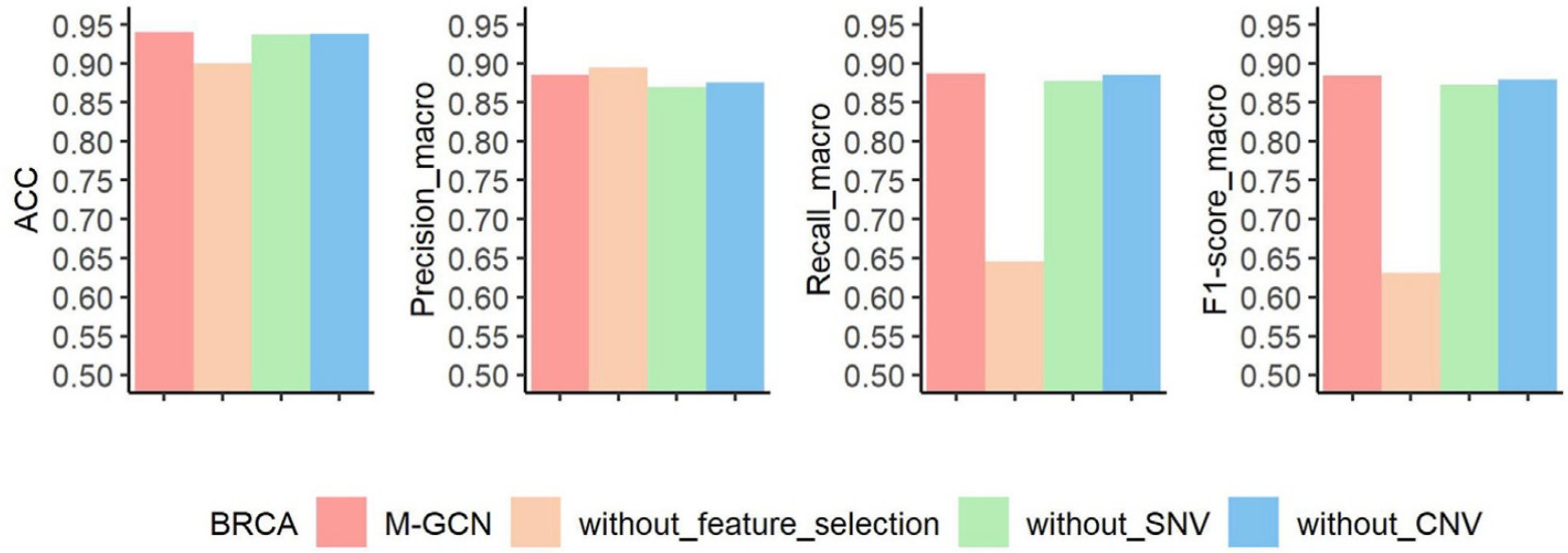


FIGURE 6 | Results of ablation experiment of M-GCN in BRCA. Pink bar represents the final performance of M-GCN, orange bar represents the performance of M-GCN without feature selection, green bar is for the performance of M-GCN without SNV features, and blue bar is for the performance of M-GCN without CNV features.

✓ Conduct three ablation experiments

- feature selection step - SNV data - CNV data

Biomarkers of Each Subtype

- ✓ Identify ten genes with highest z-score of each subtype
  - z-score normalization on the expression matrix
  - mean value of each gene in every subtype
- ✓ Perform biological process(BP) and KEGG pathway enrichment analysis

TABLE 4 | Specific biomarkers of each BRCA subtype and their enrichment pathways. The listed biomarkers rank in descending order from high to low specific score.

Molecular subtypes	Biomarker	Pathway and *p*-value
ER+	ESR1 AGR3 GATA3 PCSK6 FLJ45983 BCAS1 PMAIP1 GPR77 SCGB2A2 C10orf82	Response to estradiol (*p*-value = 1.09E-02)

Biomarkers of Each Subtype(BRCA)

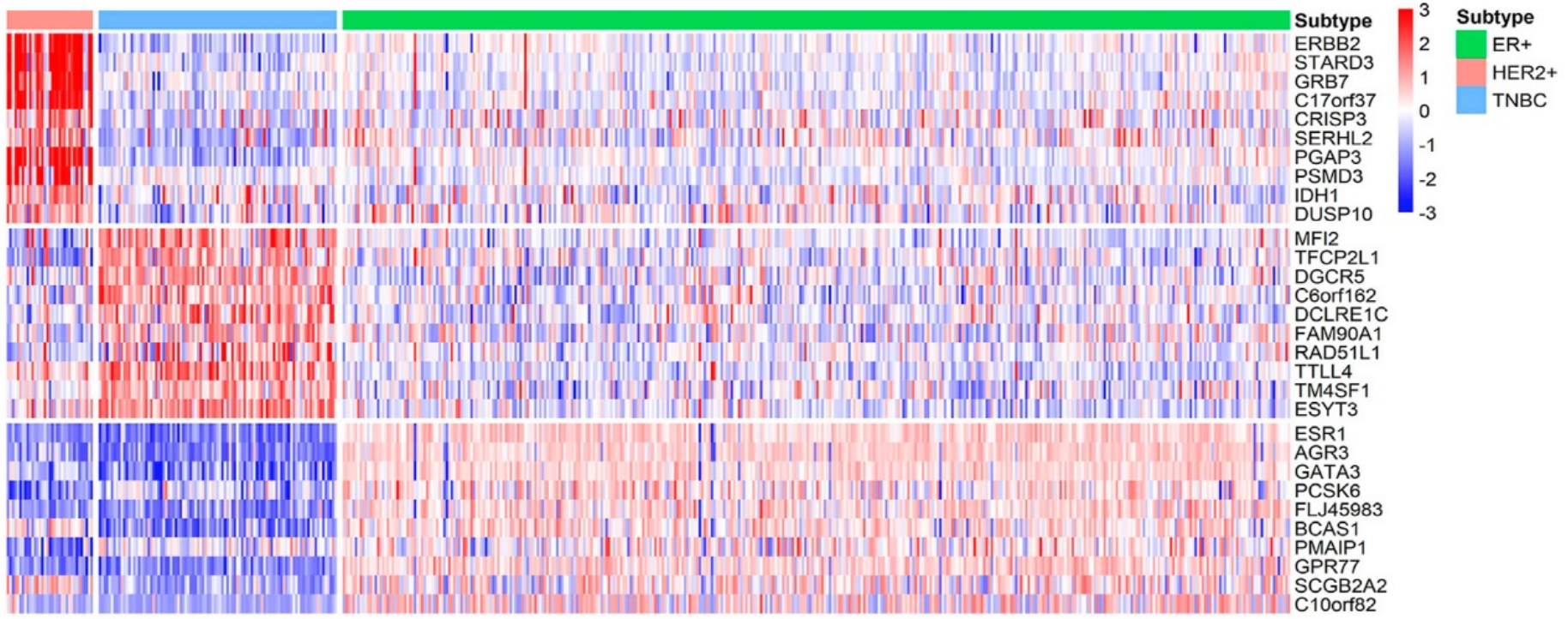


FIGURE 7 | Heatmap of the z-score normalized gene expression of the molecular subtype-specific biomarker genes in BRCA. Green bar, pink bar, and blue bar at the top represent ER+, HER2+, and TNBC subtype, respectively.

Biomarkers of Each Subtype(STAD)

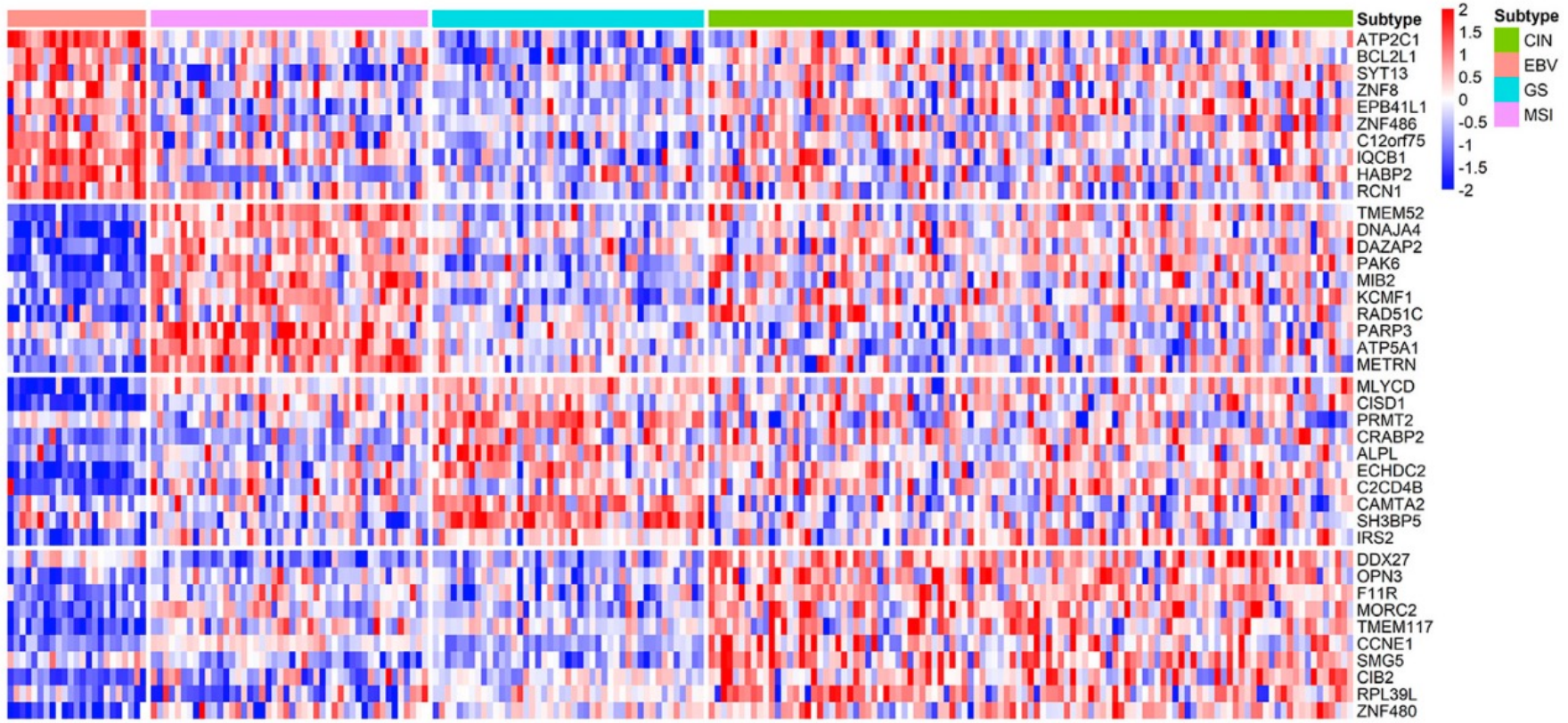


FIGURE 8 | Heatmap of the z-score normalized gene expression of the molecular subtype-specific biomarker genes in STAD. Green bar, pink bar, blue bar, and purple bar at the top represent CIN, EBV, GS, and MSI subtype, respectively.

Discussion

Discussion

✓ They propose a new framework **M-GCN** for molecular subtyping of cancer, which is empowered by integrated multi-omics data and a robust graph convolutional network.

✓ M-GCN first learns subtype-related features by **HSIC Lasso** to denoise data and construct a relatively pure sample–sample similarity graph.

✓ M-GCN assigns higher weights to similar nodes and utilizes layer-wise graph memory to limit the network to improve the **robustness** of the model based on **GNNGUARD**.

Thanks!