OXFORD

Briefings in Bioinformatics, 2026, **27(1)**, bbaf735

https://doi.org/10.1093/bib/bbaf735

Problem Solving Protocol

MoAGNN: a multi-omics hierarchical graph neural network for subtype classification and prognosis prediction in lung adenocarcinoma

Cheng-Pei Lin[1,‡], Yann-Jen Ho[1,‡], Yen-Peng Chiu[2], Yun Tang[1,3], You Sheng Paik[1], Guan-Ting Chen[1], Wei-Chih Huang[1,3,*], Tzong-Yi Lee[1,3,4,*]

[1]Institute of Bioinformatics and Systems Biology, College of Engineering Bioscience, National Yang Ming Chiao Tung University, No. 75, Boai Street, East Dist., Hsinchu 300, Taiwan

[2]Institute of Data Science and Engineering, College of Computer Science, National Yang Ming Chiao Tung University, No. 1001, Daxue Road, East Dist., Hsinchu 300, Taiwan

[3]Molecular Bioinformatics Center, College of Engineering Bioscience, National Yang Ming Chiao Tung University, No. 75, Boai Street, East Dist., Hsinchu 300, Taiwan

[4]Center for Intelligent Drug Systems and Smart Bio-devices (IDS_2B), National Yang Ming Chiao Tung University, No. 75, Boai Street, East Dist., Hsinchu 300, Taiwan

*Corresponding authors. Wei-Chih Huang, Institute of Bioinformatics and Systems Biology, College of Engineering Bioscience, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan. E-mail: brucehuang@nycu.edu.tw; Tzong-Yi Lee, Institute of Bioinformatics and Systems Biology, College of Engineering Bioscience, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan. E-mail: leetzongyi@nycu.edu.tw

‡Cheng-Pei Lin and Yann-Jen Ho contributed equally.

Abstract

Lung adenocarcinoma (LUAD), the most common subtype of nonsmall cell lung cancer, exhibits substantial molecular heterogeneity, complicating subtype classification, progression assessment, and treatment decision-making. Advances in high-throughput sequencing enable multi-omics analysis to reveal cancer mechanisms and biomarkers, yet the high dimensionality, heterogeneity, and interrelationships of omics layers such as transcriptome, microRNA expression, methylome, and copy number variation remain challenging to integrate through conventional methods. Most existing graph-based approaches represent patients as nodes, obscuring gene-level regulatory dynamics and limiting biological interpretability. To address this, we propose the Multi-omics Hierarchical Graph Neural Network (MoAGNN), a novel architecture that represents genes as nodes, integrates four omics, and leverages graph convolution with self-attention–based graph pooling to identify informative molecular nodes, thereby enhancing predictive performance and interpretability for LUAD subtype classification, tumor staging, and prognosis prediction. Multi-omics datasets from The Cancer Genome Atlas (TCGA) were used and results showed that MoAGNN achieved a test accuracy of 0.89 for LUAD subtype classification, outperforming conventional models (Random Forest, Support Vector Machine and Multi-Layer Perceptron) as well as state-of-the-art graph-based models MoGCN, a multi-omics integration model based on graph convolutional network, and MOGLAM, an end-to-end interpretable multi-omics integration method. Furthermore, we validated the generalizability of this framework on the GSE81089 dataset, demonstrating its potential applicability to clinically relevant risk assessment. Subsequent functional enrichment and survival analyses validated the biological relevance of the key genes identified by MoAGNN, supporting their potential roles in LUAD progression, and suggesting the broader applicability of this framework in multi-omics cancer research.

Keywords: lung adenocarcinoma; multi-omics; graph neural network; self-attention; deep learning

Introduction

Lung cancer is one of the most commonly diagnosed cancers and remains the leading cause of cancer-related mortality worldwide [1]. Despite medical advances, lung cancer continues to pose a significant health burden owing to its high incidence and poor prognosis. Lung adenocarcinoma (LUAD), the most frequent subtype of nonsmall cell lung cancer, exhibits strong molecular and clinical heterogeneity, which complicates both diagnosis and treatment [2]. Transcriptomic profiling from The Cancer Genome Atlas (TCGA) has defined three reproducible LUAD subtypes: terminal respiratory unit (TRU), proximal-inflammatory (PI), and proximal-proliferative (PP). TRU tumors exhibit stable transcriptional programs and favorable outcomes, whereas PI and PP show increasing inflammatory and proliferative activity, reflecting the progression from TRU to PP, highlighting the challenge of accurate subtype classification and prognosis [3].

With the development of high-throughput sequencing technologies, multi-omics profiling has emerged as an effective approach for cancer research [4, 5]. Omics refers to the large-scale analysis of biological molecules across multiple layers, including the transcriptome, microRNA (miRNA) expression, methylome (MT), and copy number variation (CNV). Studies have shown that combining transcriptomic, epigenomic, and genomic data improves cancer subtype classification [5], prognostication [6], and biomarker discovery [7]. Nevertheless, multi-omics data are inherently high dimensional and heterogeneous, necessitating

Received: September 17, 2025. **Revised:** December 01, 2025. **Accepted:** December 15, 2025

advanced computational strategies for effective integration and interpretation. This highlights the need for novel approaches capable of modeling complex molecular interactions while maintaining biological interpretability.

The pronounced molecular heterogeneity of LUAD poses significant challenges for accurate subtype classification, tumor staging, and survival risk stratification. Recent advances in deep learning have begun to address these challenges. With convolutional neural networks, SurvCNN employed convolutional neural networks for survival prediction [8] and graph neural networks (GNNs) have been increasingly adopted to model gene interactions within graph structures [9]. For instance, graph convolutional network-based frameworks have been applied to integrate multi-omics data for NSCLC subtype classification and biomarker identification [10]. Among the GNN-based methods, MoGCN [11] and MOGLAM [9] have demonstrated promising performance in multi-omics integration and cancer subtype prediction. However, they rely on static sample-level graphs that overlook gene-level regulation, lack hierarchical modeling, and fail to capture disease-specific molecular dynamics. These limitations underscore the need for frameworks that model gene-level interactions and provide clinically interpretable insight into LUAD heterogeneity.

In this study, we propose a gene-centric heterogeneous graph neural network, the Multi-modal Hierarchical Graph Neural Network (MoAGNN), which integrates transcriptome, miRNA, methylation, and CNV features within a biologically informed framework. By incorporating regulatory directions and employing hierarchical pooling with attention mechanisms, MoAGNN aims to improve LUAD subtype classification, stage prediction, and survival risk assessment, while enhancing biological interpretability. Importantly, this framework directly addresses the limitations of existing methods by moving beyond static sample-level representations to provide an interpretable biologically meaningful model of multi-omics regulatory dynamics.

Materials and Methods

Overview of MoAGNN

This study employed a multistep analytical framework to investigate LUAD from clinical and molecular perspectives (Fig. 1). Four omics datasets (transcriptome, miRNA, MT, and CNV) were obtained from TCGA and preprocessed. Differential expression analysis of transcriptome data identified genes for network construction, with gene–gene co-expression estimated using WGCNA and additional omics features incorporated to form a heterogeneous graph. On this basis, we developed a Multi-omics Hierarchical Graph Neural Network (MoAGNN) that applies graph convolution, hierarchical pooling, and attention mechanisms for predictive modeling. The framework was evaluated on three tasks, subtype classification, stage classification, and survival prediction, with enrichment and survival analyses performed to assess the biological relevance of MoAGNN-identified genes.

Data source and data preprocessing

LUAD data were obtained from TCGA using the R package 'TCGAbiolinks' (version 2.34.0) [12]. Three independent cohorts were established, each tailored to a specific predictive task: subtype classification, stage classification, and survival prediction. The LUAD subtype cohort comprised 185 patients, annotated as TRU (n = 69), PI (n = 68), and PP (n = 48). The LUAD stage cohort included 450 patients, categorized into early-stage (stage I–II, n = 358) and late-stage (stage III–IV, n = 92). The LUAD survival cohort included 441 patients, stratified into low-risk (n = 256) and high-risk (n = 185) groups using two years as the cutoff, determined by rounding the median overall survival, OS (1.78 years) to the nearest integer (Supplementary Fig. S1). In addition, n = 59 normal lung tissue samples were included for comparison.

For data preprocessing, transcriptome and miRNA profiles were normalized using the TCGAanalyze_Normalization function in the EDASeq R package (version 2.40.0). MT data, CpG islands within 1500 bp upstream of transcription start sites were extracted. β-values of CpG sites were transformed into M-values using the lumi R package (version 2.57), and averaged per gene to represent methylation levels [13, 14]. CNV data were processed using GISTIC 2.0 with the default extreme method [15]. Quality control was then applied to remove features or samples with excessive missing or zero values. Specifically, transcriptome, miRNA, and CNV features with more than 20% missing or zero values across samples were excluded, and samples with more than 20% missing or zero values across these omics types were discarded. For the MT dataset, any feature containing zero or missing values was removed to preserve integrity. Finally, the dataset was partitioned into 80% for training and 20% for testing using stratified random sampling to preserve class balance. Within the training set, 5-fold cross-validation was performed for model selection and to mitigate overfitting.

External validation was performed using the GEO cohort GSE81089, which provides RNA-seq data with available stage information. The dataset included 197 samples with stage information, comprising 161 early-stage and 36 late-stage cases. We reused the transcriptome-derived gene nodes and their original WGCNA edge weights from the modeling framework, without repeating feature selection or network reconstruction. Using this fixed gene set and network topology, we reconstructed predictive models for stage classification to enable evaluation on this external dataset.

Omics-specific feature selection strategy

Differential feature selection was performed independently for each omics dataset. For the transcriptome, genes with counts per million $\geq$0.5 in at least two samples were retained, and differentially expressed genes (DEGs) were identified using DESeq2 (version 1.46.0) [16]. Thresholds were set to $|\log_2 FC| > 2$ with adjusted P-value (padj) < 0.05 for subtype classification. For stage and survival, applying $|\log_2 FC| > 2$ yielded fewer than 10 DEGs; therefore, the cutoff was relaxed to $|\log_2 FC| > 0.5$ with padj < 0.05 to ensure sufficient features for downstream modeling. For miRNA, differential expression analysis was conducted with limma (version 3.62.2), using a uniform criterion of $|\log_2 FC| > 0.1$ and padj < 0.1 across the three tasks.

For MT and CNV, one-way ANOVA was used to assess group differences for each task. MT and CNV features were selected using $P < .001$ for all three tasks, except that survival CNV used $P < .01$, as no CNV features met the 0.001 criterion. These task-specific thresholds were chosen to balance statistical rigor with preserving a sufficient number of features for downstream model training. The filtered features were subsequently used as inputs to conventional machine-learning models, serving as baselines for comparison with the proposed graph-based framework.

Construction of graphs with intra- and inter-omics associations

To capture both gene–gene interactions and cross-omics relationships, the graph $G = (V, E)$ was constructed (Fig. 2). The node set V represents genes, and the edge set E represents their interactions.

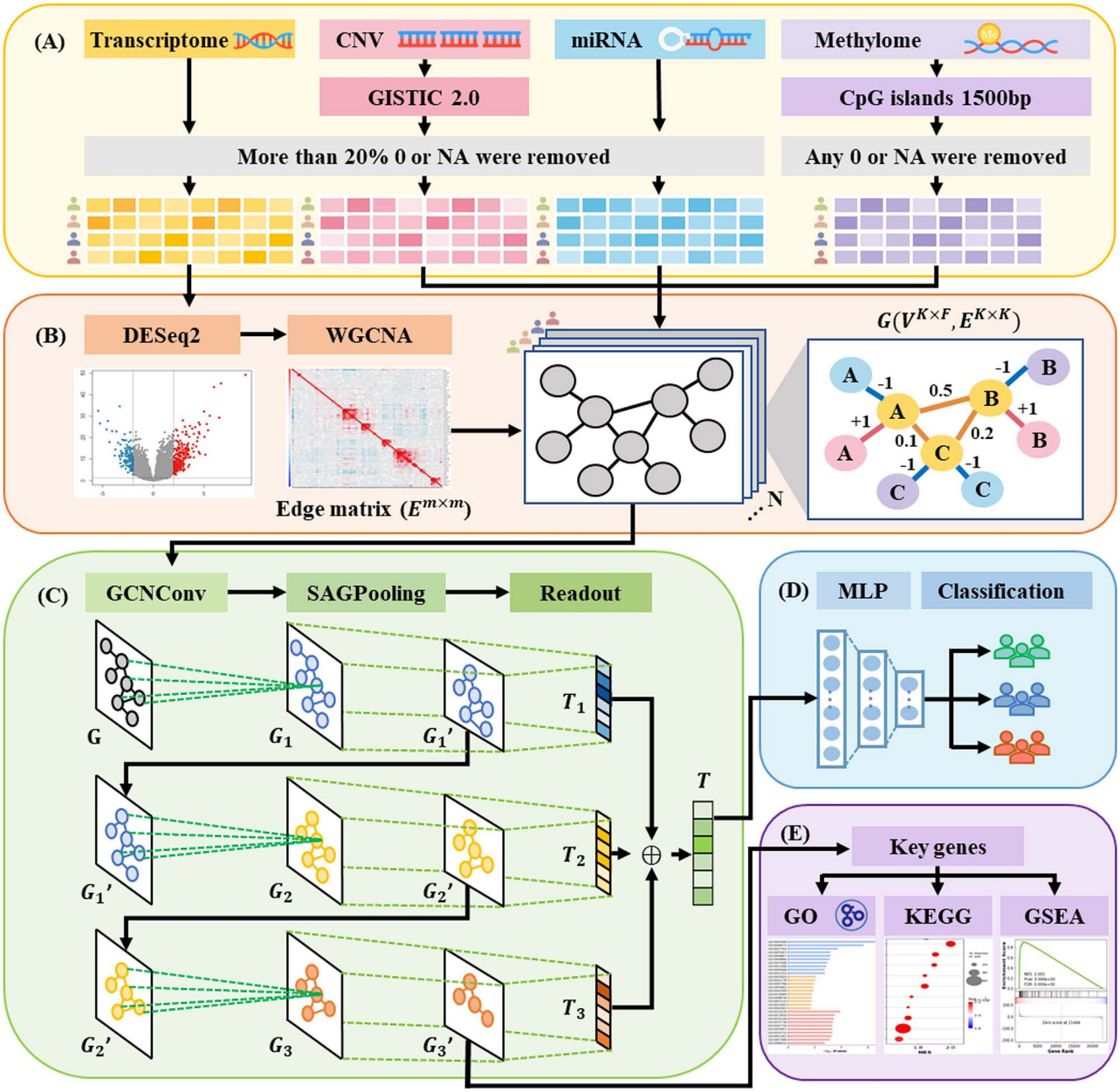

Figure 1. Workflow of MoAGNN. (A) Multi-omics datasets were collected from TCGA and preprocessed. (B) Graph nodes defined from DEGs, while weighted edges represent gene–gene co-expression inferred by WGCNA. Additional omics layers were incorporated by linking genes with inter-omics edges that reflect known regulatory effects. (C) Resulting heterogeneous graph was processed by our MoAGNN model. (D) Learned graph representations were passed to a multi-layer perceptron (MLP) for predictive tasks. (E) Functional enrichment of key genes identified.

The node feature matrix $X \in \mathbb{R}^{K\times F}$represents the F-dimensional features of K nodes, while the adjacency matrix $A \in \mathbb{R}^{K\times K}$describes their connectivity.

The transcriptome-based graph $G_{\text{transcriptome}} = (V_{\text{transcriptome}}, E_{\text{transcriptome}})$was constructed using m DEGs as nodes. Each node was assigned its expression level as a one-dimensional feature, while edges were defined by co-expression relationships inferred with WGCNA (version 1.73). Pairwise correlations were transformed into a weighted adjacency matrix using a soft-thresholding power β, selected with the *pickSoftThreshold* function to approximate scale-free topology. The adjacency matrix thus quantifies the strength of gene co-expression, with higher values representing stronger similarity [17].

To construct a binary edge matrix, candidate thresholds were selected from empirical percentiles (10th–90th) of the adjacency distribution, and the optimal threshold was determined based on downstream model performance. This setting kept the graph compact while limiting spurious edge formation. For multi-omics integration, inter-omics edges were established between transcriptome and miRNA (E_{mir}), MT (E_{MT}), and CNV (E_{CNV}) for the same genes. Following the prior knowledge-guided strategy reported by Yan *et al.* [18], MT–mRNA and CNV–mRNA edges were assigned weights of −1 and +1, respectively. Given that both miRNA and methylation typically repress gene expression, the weights of E_{mir} and E_{MT} were set to −1, while the weight of E_{CNV} was set to +1 to reflect its positive association with transcript levels.

The final heterogeneous graph can be expressed as:

$$V = \{V_{\text{transcriptome}}, V_{miRNA}, V_{MT}, V_{CNV}\}$$
$$E = \{E_{\text{transcriptome}}, E_{miRNA}, E_{MT}, E_{CNV}\} \quad (1)$$

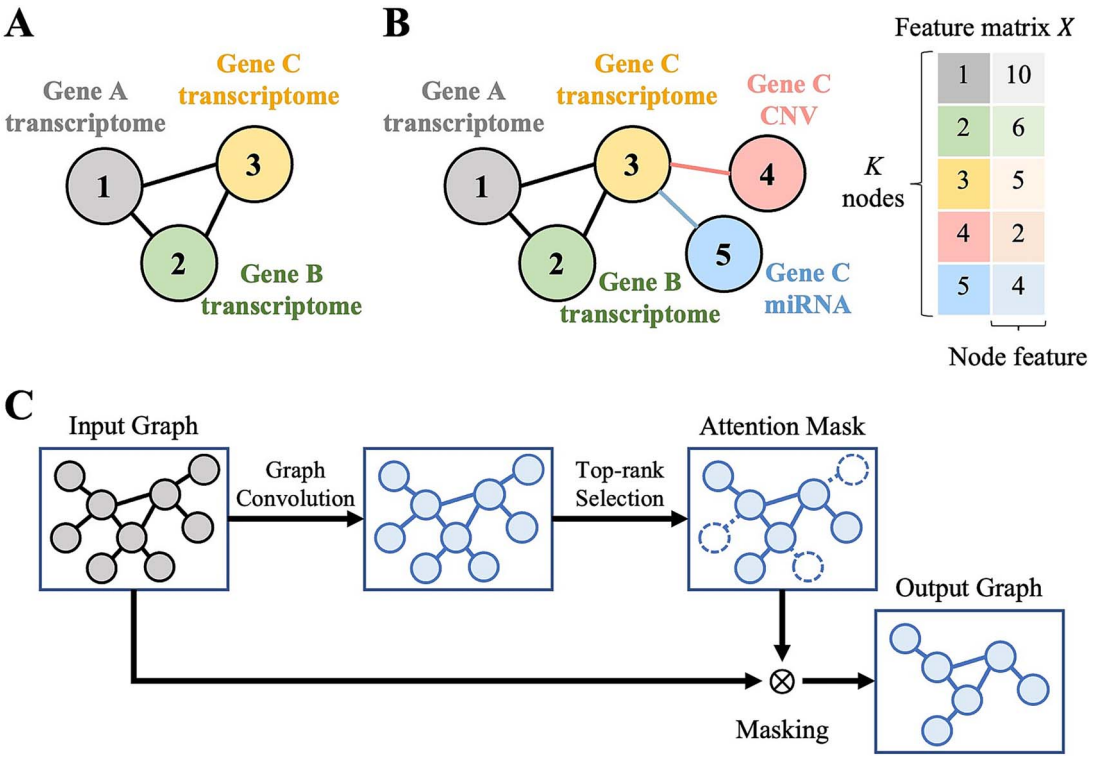

Figure 2. Graph structure and self-attention graph pooling. (A) A transcriptome-based graph with DEGs as nodes and connecting them through WGCNA-derived co-expression edges. (B) A multi-omics heterogeneous graph generated by integrating intra-omics co-expression with inter-omics associations between the four omics. (C) Input graph was processed with a self-attention pooling layer, where node importance is scored and low-scoring nodes (dotted circles) are pruned.

This design allowed us to jointly model intra- and inter-omics relationships within a unified graph structure.

MoAGNN model construction

Graph convolutional layers, as proposed by Kipf and Welling [19], were employed to learn node representations. At each layer, node features are updated as:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}\Theta\right) \quad (2)$$

where $\tilde{A}$ is the adjacency matrix with self-loops, $\tilde{D}$ is the corresponding degree matrix, and Θ denotes learnable parameters, and σ used in this work is ReLU.

To reduce graph complexity and enhance interpretability, we applied self-attention graph pooling (SAGPooling) [20]. Self-attention scores were computed as:

$$S = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta_{att}\right) \quad (3)$$

where X, is the node feature matrix and Θ_{att} is the learnable parameter. Nodes were ranked by their scores, and the top $\lceil P\ K \rceil$ nodes were retained according to the pooling ratio:

$$idx = \text{top-rank}\,(S, \lceil P\ K \rceil) \quad (4)$$

This mechanism enables the model to focus on informative genes while pruning less relevant ones.

The readout function aggregates features from each pooling layer by concatenating global mean pooling and global max pooling:

$$T_i = concat\left(\frac{1}{k_i}\sum_{n=1}^{k_i} x_n, \max_{n=1}^{k_i} x_n\right) \quad (5)$$

where k_i is the number of retained nodes and x_n represents the feature vector of the n-th node. This design captures both global trends and dominant signals contributing to classification.

The final MoAGNN model concatenates hierarchical features from three pooling layers into a comprehensive graph representation, which is then processed by a three-layer Multi-Layer Perceptron (MLP) with ReLU activations to produce the final predictions. The model is trained using the Adam optimizer (learning rate 0.0005, batch size 24) with dropout (0.2) and early stopping (patience 50) over a maximum of 1000 epochs.

MoAGNN model evaluation and functional analysis

To evaluate performance, we compared MoAGNN against five baselines, including Random Forest (RF) [21], Support Vector Machine (SVM) [22], MLP [23], and two state-of-the-art GNN-based models, MOGLAM [9] and MoGCN [11]. Model performance was assessed using standard classification metrics, including Accuracy (ACC), Precision, Recall, F1-score, and Matthews Correlation Coefficient (MCC) [24].

Functional enrichment analysis was performed using GSEApy [25] (version 1.1.4) with Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology (GO) categories, including Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) [26]. The 2023 releases of GO and KEGG gene sets

Table 1. Number of selected features after data preprocessing and filtering across LUAD tasks

Task	Transcriptome	miRNA	MT	CNV
Original Features	60,660	41,586	1457	19,492
Subtype	738	215	5766	3281
Stage	1052	3	140	42
Survival	401	6	44	130

were used to identify significantly enriched biological functions and pathways.

Results

Differential and statistical feature selection across omics

Feature selection was carried out separately for each omics type across the three LUAD tasks. Transcriptome and miRNA features were defined from the intersection of DEGs or differentially expressed miRNAs across pairwise comparisons, whereas methylation and CNV features were selected using ANOVA. As summarized in Table 1, subtype analysis preserved substantially more features than stage or survival, reflecting stronger molecular signals in LUAD subtypes.

For subtype classification, transcriptome analysis identified 738 DEGs across the three pairwise subtype comparisons (Fig 3A–C). Heatmaps for each comparison (Fig 3D–F) demonstrated clear clustering patterns, and PCA further highlighted the discriminative power of these genes in separating the LUAD subtypes (Fig 3G–J). For stage classification, transcriptome data yielded 1052 DEGs between early- and late-stage patients, whereas the miRNA differential expression analysis identified one up-regulated and two down-regulated features (Supplementary Fig. S2). For survival prediction, 401 transcriptomic DEGs were identified between high- and low-risk groups, along with two up-regulated and four down-regulated miRNAs (Supplementary Fig. S3). However, heatmaps and PCA plots based on stage- and survival-associated features showed limited separation, indicating that these tasks had weaker discriminative power compared to subtype classification.

Predictive performance of MoAGNN with different hyperparameters

MoAGNN includes two key hyperparameters, the edge threshold (*thresh*) for graph construction and the pooling ratio (*P*) for hierarchical pooling. The number of graph nodes constructed for each task and omics type is summarized in Supplementary Table S1. Candidate thresholds were then defined from DEG correlation percentiles, and increasing thresh progressively reduced the number of edges (Fig. 4A). As an example, detailed edge statistics for the subtype task are shown in Supplementary Table S2. The best classification performance for subtype was achieved when the threshold was set to the 75th percentile (4.22E-12), as shown in Fig. 4B. Based on this optimal threshold, pooling ratios ranging from 0.05 to 1.0 were evaluated, and the best performance was obtained at $P = .35$ (Fig. 4C). For the stage and survival tasks, parameters were selected using the same procedure (Supplementary Tables S3 and S4). With these optimized settings, MoAGNN was evaluated across omics combinations (Fig. 4D). The transcriptome contributed most strongly to classification accuracy, while the miRNA layer exhibited only a limited effect.

Comparative evaluation of MoAGNN and baseline models

As shown in Table 2, MoAGNN outperformed other approaches in LUAD subtype prediction, achieving improvements across both accuracy-based and correlation-based metrics (Fig. 4E). For stage classification, however, its performance did not exceed that of baseline methods (Supplementary Table S5; Fig. 4F). This limitation may reflect the relatively modest transcriptomic differences between early- and late-stage LUAD, which restricted the availability of informative DEGs for network construction. In addition, the WGCNA-derived co-expression networks for the stage task were comparatively sparse, with fewer strong connections than in the subtype task and weaker correlations than in the survival task, further reducing graph expressiveness and limiting predictive capacity.

For survival prediction, r results are shown in Fig. 4G and detailed in Supplementary Table S6. MoAGNN achieved performance comparable to other models and notably reached the highest precision (83.72%), underscoring its strong ability to correctly identify high-risk patients. Taken together, these findings demonstrate that MoAGNN excels in subtype prediction and provides clinically meaningful advantages in survival risk assessment, although its performance in stage classification is hindered by weaker transcriptomic signals and less informative network structures.

External validation of stage predictive models

External validation using the GSE81089 cohort was performed to assess whether the transcriptome-derived biomarkers and their associated network structure remained effective in the independent dataset. Given that transcriptome features contributed most substantially to the predictive performance in the TCGA analyses, focusing on the RNA-derived gene set enabled a direct evaluation of whether the key biomarkers identified in the primary analysis retained their discriminative ability across cohorts. This validation was conducted by retaining the TCGA-derived gene set together with their WGCNA-defined gene–gene connectivity and applying them directly to the GSE81089 dataset. Using 5-fold cross-validation, the final performance metrics on GSE81089 were as follows: accuracy 0.775, precision 0.485, recall 0.733, F1 score 0.579, and MCC 0.451. The ability to maintain stable predictive performance under an entirely new set of samples and clinical labels indicates that both the selected biomarkers and their underlying network relationships were correctly captured, demonstrating that the model can generalize effectively across independent cohorts. Together, these findings confirm that the transcriptome-derived biomarkers and the WGCNA-informed network structure preserve their predictive capacity across datasets, underscoring

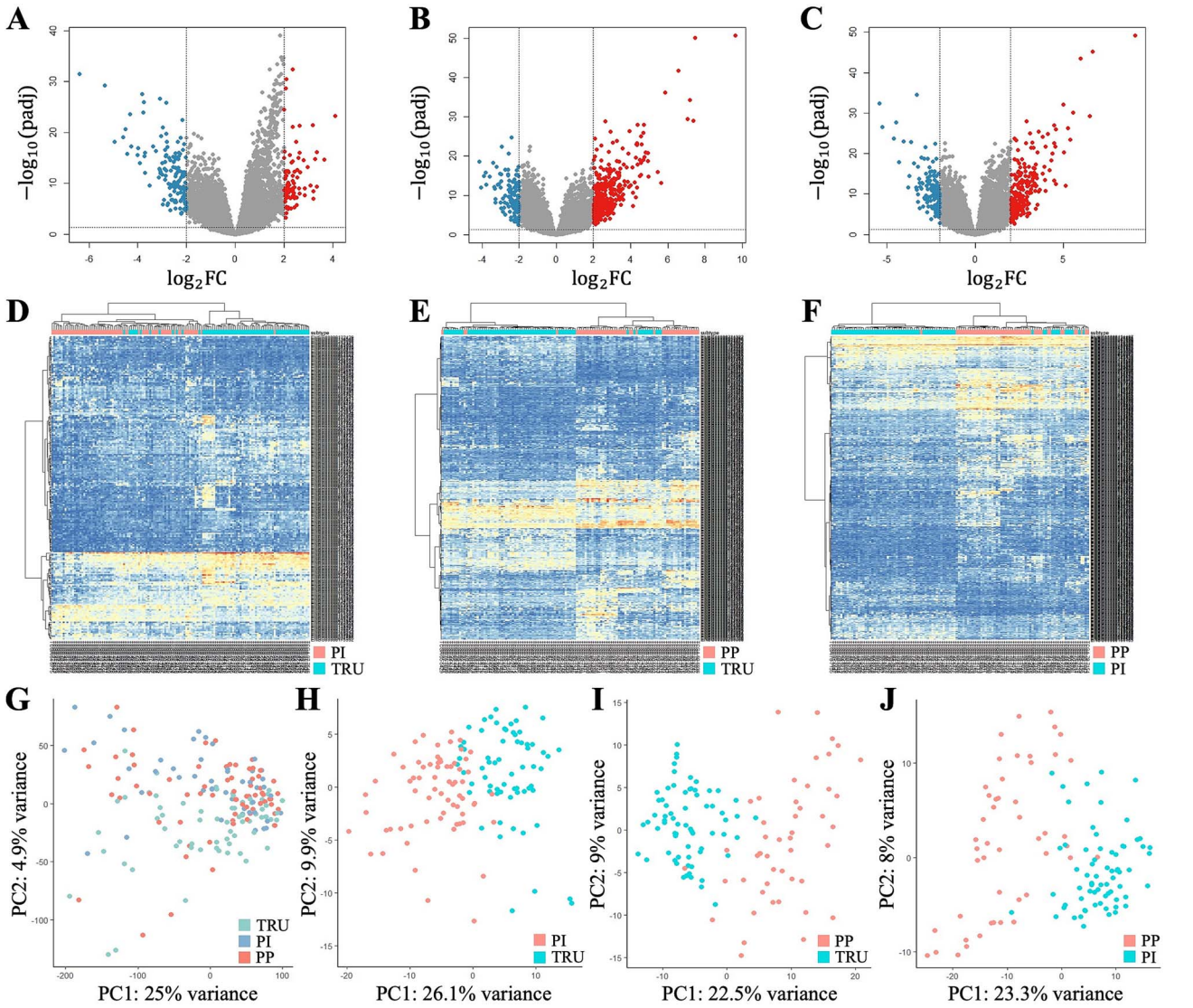

Figure 3. Differential expression analysis and PCA of transcriptome data. (A–C) Volcano plots of pairwise subtype comparisons: TRU versus PI (81 up-regulated, 138 down-regulated), TRU versus PP (218 up-regulated, 128 down-regulated), and PI versus PP (327 up-regulated, 120 down-regulated). (D–F) Heatmaps showing clustering of DEGs. Up-regulated genes are shown in red, down-regulated in blue. (G) PCA using all genes; (H–J) PCA using DEGs from each pairwise comparison.

Table 2. Performance comparison of MoAGNN and baseline models on LUAD subtype

	MoAGNN	RF	SVM	MLP	MOGLAM	MoGCN
Accuracy	0.891	0.837	0.864	0.783	0.811	0.841
Precision	0.907	0.874	0.854	0.786	0.815	0.852
Recall	0.893	0.817	0.854	0.783	0.811	0.841
F1-score	0.897	0.829	0.854	0.778	0.762	0.842
MCC	0.838	0.759	0.793	0.676	0.679	0.763

the generalizability and robustness of the proposed modeling framework.

Expression and functional characterization of key genes in LUAD

A total of 74 key genes were prioritized by the SAGPooling layer of MoAGNN (Supplementary Table S7). Among the subtype-enriched genes, FGB and BMP6 were markedly up-regulated in the PP subtype (Fig. 5A and B). FGB, a coagulation-pathway component, has been linked to tumor progression, metastasis, and immune evasion through remodeling of the tumor microenvironment; higher FGB expression correlates with poorer performance status and more advanced stage [27]. Although BMP6 is generally down-regulated in NSCLC and regarded as tumor-suppressive, its PP-restricted elevation here may reflect subtype-specific activation of the TGF-β/BMP axis associated with EMT and aggressive behavior [28]. In contrast, RSPO2 decreased progressively from TRU to PP (Fig. 5C), consistent with reports that RSPO2

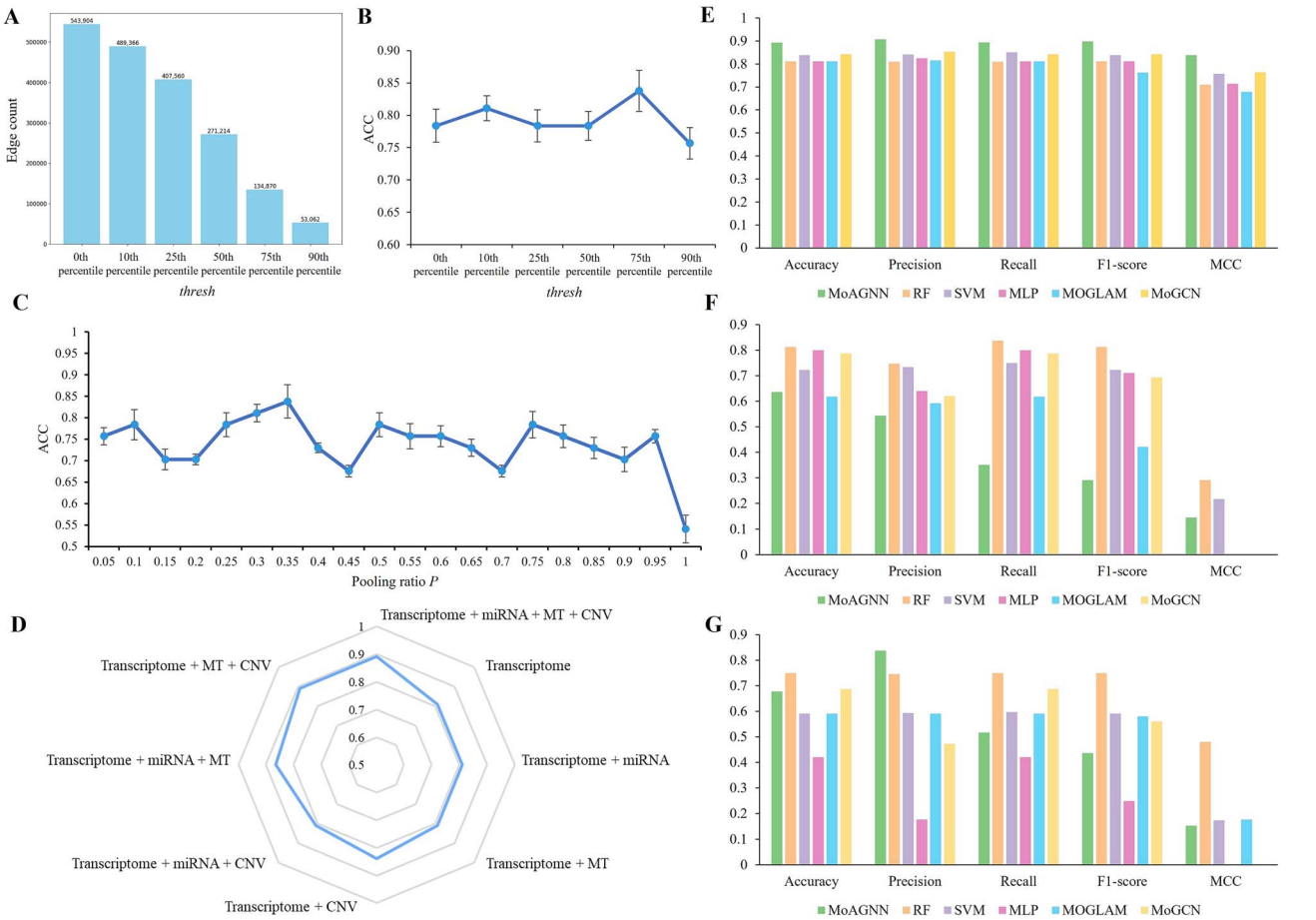

Figure 4. Model optimization and performance evaluation. (A) Edge counts in transcriptome graphs across different threshold settings. (B) Classification accuracy across varying edge thresholds. (C) Model performance across *P* using the optimal threshold (75th percentile). (D) Radar plot comparing model accuracy across different omics combinations, highlighting the dominant contribution of transcriptome, MT, and CNV features. (E) Subtype classification, (F) stage classification, and (G) survival prediction. Bar plots show model performance across accuracy, precision, recall, F1-score, and MCC. Baseline models include RF, SVM, MLP, MOGLAM, and MoGCN.

loss promotes EMT and invasion in LUAD [29]. Among the putative novel genes, AVPR1B and CD1E decreased from TRU to PI to PP (Fig. 5D and E). The decline of AVPR1B indicates attenuation of vasopressin-mediated GPCR signaling via this receptor in higher-risk subtypes [30], whereas CD1E downregulation impairing lipid antigen presentation and M1 macrophage polarization supports an immune-cold microenvironment [31]. Conversely, DRP2 increased along the same trajectory (Fig. 5F); higher DRP2 associated with poorer survival, and functional evidence links DRP2 to EMT-driven invasiveness, highlighting its therapeutic relevance [32]. Collectively, these subtype-resolved patterns indicate that both established and emerging genes contribute to the molecular stratification of LUAD.

To further investigate the functional context of these key genes, we performed GO and KEGG enrichment analyses. In the BP category, regulation of SMAD protein signal transduction (GO:0060390), plasminogen activation (GO:0031639), and ornithine metabolic process (GO:0006591) were significantly enriched, implicating these genes in TGF-β regulation, extracellular matrix (ECM) remodeling, and amino acid–derived metabolic processes. In the CC category, serine-type endopeptidase complex (GO:1905370) was enriched, suggesting potential roles in protease-mediated activation of latent signaling molecules and ECM turnover, whereas in the MF category, transition metal ion transmembrane transporter activity (GO:0046915) was identified, reflecting potential links to redox balance and enzyme cofactor availability (Fig. 5G). KEGG pathway analysis further highlighted enrichment in cancer-associated signaling processes, including cell adhesion molecules and the TGF-β signaling pathway (Fig. 5H). Together, these enrichments point toward coordinated regulation of growth factor signaling, ECM remodeling, immune modulation, and tumor metabolism, consistent with molecular programs that underlie LUAD subtype differentiation and progression. Notably, the concordance between GO and KEGG analyses highlights a shared role of these genes in TGF-β signaling and ECM dynamics, implicating integrated transcriptional, proteolytic, and metabolic programs in LUAD progression.

Pathway enrichment analysis of LUAD subtype via GSEA

To further explore the functional differences between LUAD subtypes and normal lung tissue, we performed GSEA using DEGs identified from each subtype-to-normal comparison. In the comparison of TRU subtype and normal samples, GSEA revealed broad downregulation of metabolism- and signaling-related pathways (Fig. 6A). Pathways with negative enrichment scores such as AGE-RAGE signaling, regulation of lipolysis in adipocytes, and cholesterol metabolism indicated distinct metabolic features in the TRU subtype. For PI subtype versus normal samples, pathways related

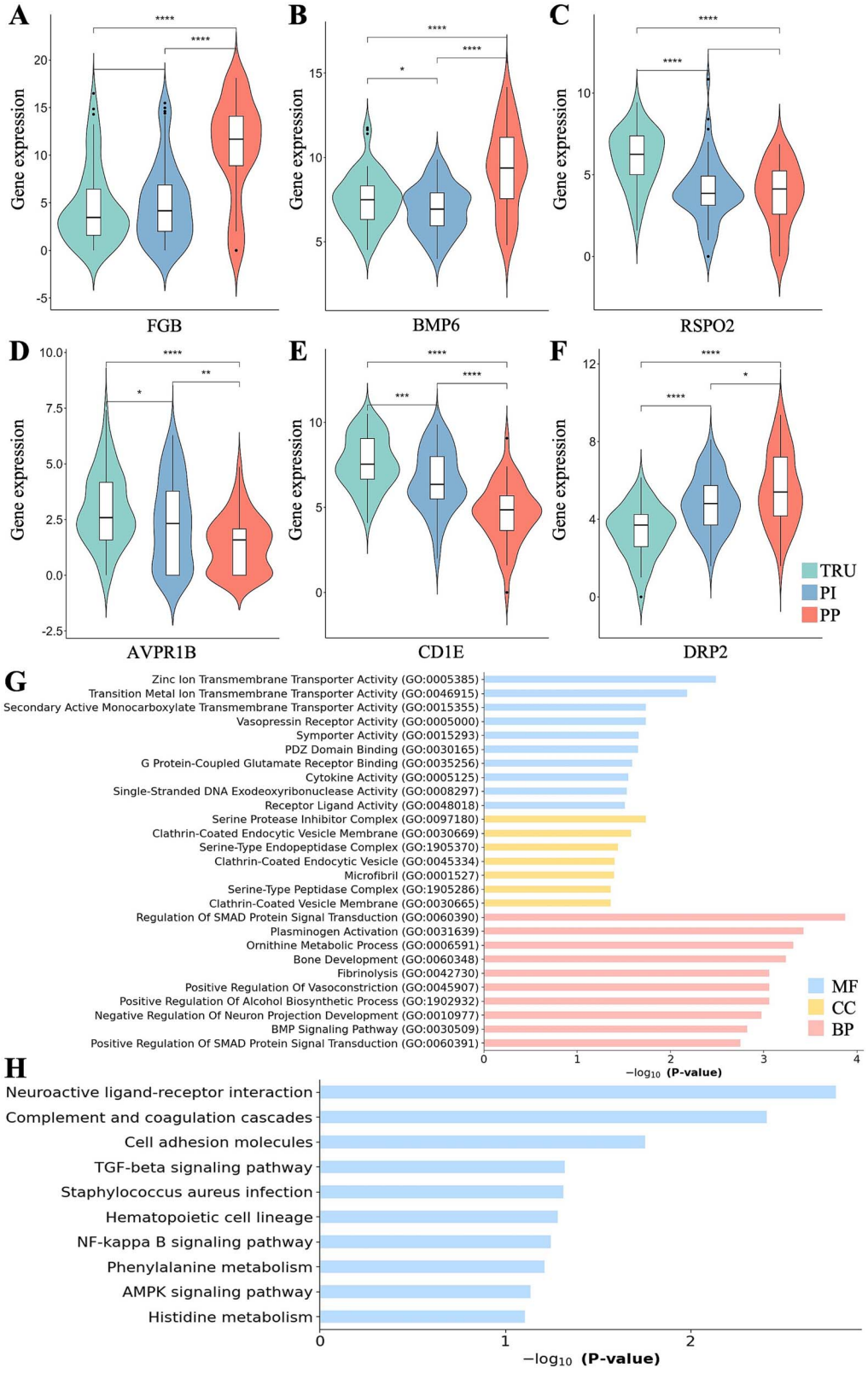

Figure 5. Key gene expression patterns and functional enrichment analyses. Expression profiles of representative genes identified by MoAGNN across LUAD subtypes: (A) *FGB*, (B) *BMP6*, (C) *RSPO2*, (D) *AVPR1B*, (E) *CD1E*, and (F) *DRP2*, (G) GO enrichment analysis of key genes, showing the top 10 terms, (H) KEGG pathway enrichment analysis of key genes, with the top 10 pathways ranked by -$\log_{10}$ (*P*-value). *$P < .05$, **$P < .01$, ***$P < .001$, ****$P < .0001$.

to cell cycle and oocyte meiosis were up-regulated, while cGMP-PKG signaling pathway was down-regulated (Fig. 6B). In PP subtype versus normal samples (Fig. 6C), down-regulated pathways were predominantly associated with immune and inflammatory responses, including viral myocarditis, *Staphylococcus aureus* infection, and Th17 cell differentiation.

Following the individual GSEA analyses for each LUAD subtype, we generated an integrated heatmap (Fig. 6D) to compare the normalized enrichment scores (NESs) of the top 10 pathways with the highest absolute NES values across TRU, PI, and PP. The heatmap revealed a clear separation between proliferative and immune-metabolic functions. Proliferative pathways such as cell cycle and oocyte meiosis were prominently enriched in PI, with PP also showing moderate activation consistent with elevated mitotic activity in more aggressive tumors [33]. Cellular senescence was likewise increased in PI, potentially reflecting stress-induced arrest mechanisms or subtype-specific aging responses [34]. In contrast, immune and metabolic pathways including Th17 cell differentiation, allograft rejection, *S. aureus* infection, AGE-RAGE signaling, regulation of lipolysis, and PPAR signaling were strongly suppressed in PP, indicative of an immunosuppressive and metabolically repressed tumor microenvironment characterized by impaired T-cell signaling and reduced antigenic activity [35, 36]. TRU exhibited attenuated proliferative activity and a more balanced immune-metabolic profile, aligning with its terminal-respiratory-unit–like phenotype. Collectively, these results highlight marked functional divergence across LUAD subtypes: TRU displays relative metabolic stability, PI is defined by proliferative and senescence-related signatures, and PP exhibits a dual pattern of proliferative activation and pronounced immune-metabolic suppression.

Survival analysis to assess the prognostic significance of key genes

To evaluate the prognostic relevance of the key genes identified by MoAGNN, Kaplan–Meier survival analysis was performed on a cohort of 431 LUAD patients. For each gene, patients were stratified into high- and low-expression groups, and overall survival (OS) was compared using the log-rank test. Out of the 74 key genes analyzed, 45 showed statistically significant associations with OS ($P < .05$) (Supplementary Fig. S4). The top three genes with the lowest *P*-values were Ras Homolog Family Member V (*RHOV*, $P = 9.77e\text{-}05$), Tetraspanin 11 (*TSPAN11*, $P = 1.00e\text{-}04$), and Nuclear Receptor Subfamily 1 Group I Member 2 (*NR1I2*, $P = 2.70e\text{-}04$), as shown in Fig 6E–G. Patients in the low-expression group of *RHOV* exhibited significantly better OS, suggesting that *RHOV* may function as a risk factor. In contrast, patients in the high-expression groups of *TSPAN11* and *NR1I2* had significantly improved OS, indicating their potential protective roles in LUAD prognosis.

Discussion and Conclusions

This study proposed MoAGNN, a hierarchical graph neural network that integrates transcriptome, miRNA, MT, and CNV data to classify LUAD molecular subtypes, tumor stages, and survival risks. By embedding prior biological knowledge and co-expression structures, MoAGNN constructs disease-specific heterogeneous graphs to model both intra- and inter-omics relationships. This framework enables biologically informed learning and improved interpretability compared to conventional deep learning approaches.

MoAGNN achieved its best performance in LUAD subtype classification, which is consistent with the biology of a gene-driven endpoint because PCA and differential expression analyses demonstrated clear molecular separation among subtypes. In addition, the resulting co-expression graphs were well structured, enabling SAGPooling to capture subtype-specific hubs with known or plausible roles in LUAD pathogenesis. By contrast, prediction of tumor stage and survival risk was weaker. DEG-based PCA indicated substantial transcriptomic overlap between early-stage and late-stage disease and between high-risk and low-risk groups, suggesting limited separability at the expression level. Moreover, at matched percentile thresholds across the three tasks, the stage-specific graphs exhibited the lowest edge density and modularity, indicating weaker, less coherent gene–gene structure. This pattern implies that pathological stage is only weakly coupled to the baseline transcriptomic state, which explains the relatively limited performance of gene-only models for stage prediction. Conceptually, this gap reflects the underlying biology of each endpoint. Subtype is largely defined by gene-level alterations, whereas stage is influenced by multiple nonmolecular factors such as age, smoking history, tumor size, comorbidities, and sampling heterogeneity. Survival lies even further downstream and is additionally shaped by treatment regimens, care pathways, and post-diagnosis events. These factors reduce the mutual information between omics and labels and thus limit the performance of a purely omics-based model for stage/survival. As a next step, integrating clinical features could complement omics and improve predictive performance for these endpoints.

SAGPooling enabled MoAGNN to identify subtype-specific genes with known or putative relevance to LUAD, with several biomarkers showing expression patterns consistent with prior studies, while others displayed subtype-dependent deviations likely reflecting underlying molecular heterogeneity. Beyond established markers, the model also highlighted potentially novel contributors to LUAD biology, indicating its capacity for biomarker discovery. Pathway-level analyses further emphasized distinct functional programs among subtypes: TRU exhibited relatively preserved metabolic activity, PI showed strong proliferative and senescence-related signaling, and PP demonstrated combined proliferative activation and suppression of immune–metabolic pathways. These coherent patterns underscore MoAGNN's ability to recover meaningful functional heterogeneity and support its utility for subtype-aware interpretation and therapeutic stratification.

The prognostic associations observed in our analysis further illustrate MoAGNN's capacity to identify clinically meaningful biomarkers. RHOV, a Rho GTPase implicated in epithelial–mesenchymal transition and metastatic progression, emerged as a strong indicator of poor survival, consistent with its reported oncogenic activity [37, 38]. In contrast, TSPAN11 showed a favorable association with overall survival, aligning with its proposed roles in immune modulation and cell–cell interactions [39, 40]. Rather than reiterating individual gene effects, these patterns collectively demonstrate that MoAGNN captures biologically coherent risk signals and can extend beyond subtype delineation to support biomarker prioritization and survival stratification in LUAD.

This study initially explores applying biologically informed graph neural networks to multi-omics integration in LUAD. MoAGNN shows potential in stratifying molecular subtypes and identifying genes associated with tumor progression, some of which align with previous findings, while others may offer novel insights. Although its performance in stage and survival prediction was limited, likely due to data complexity and

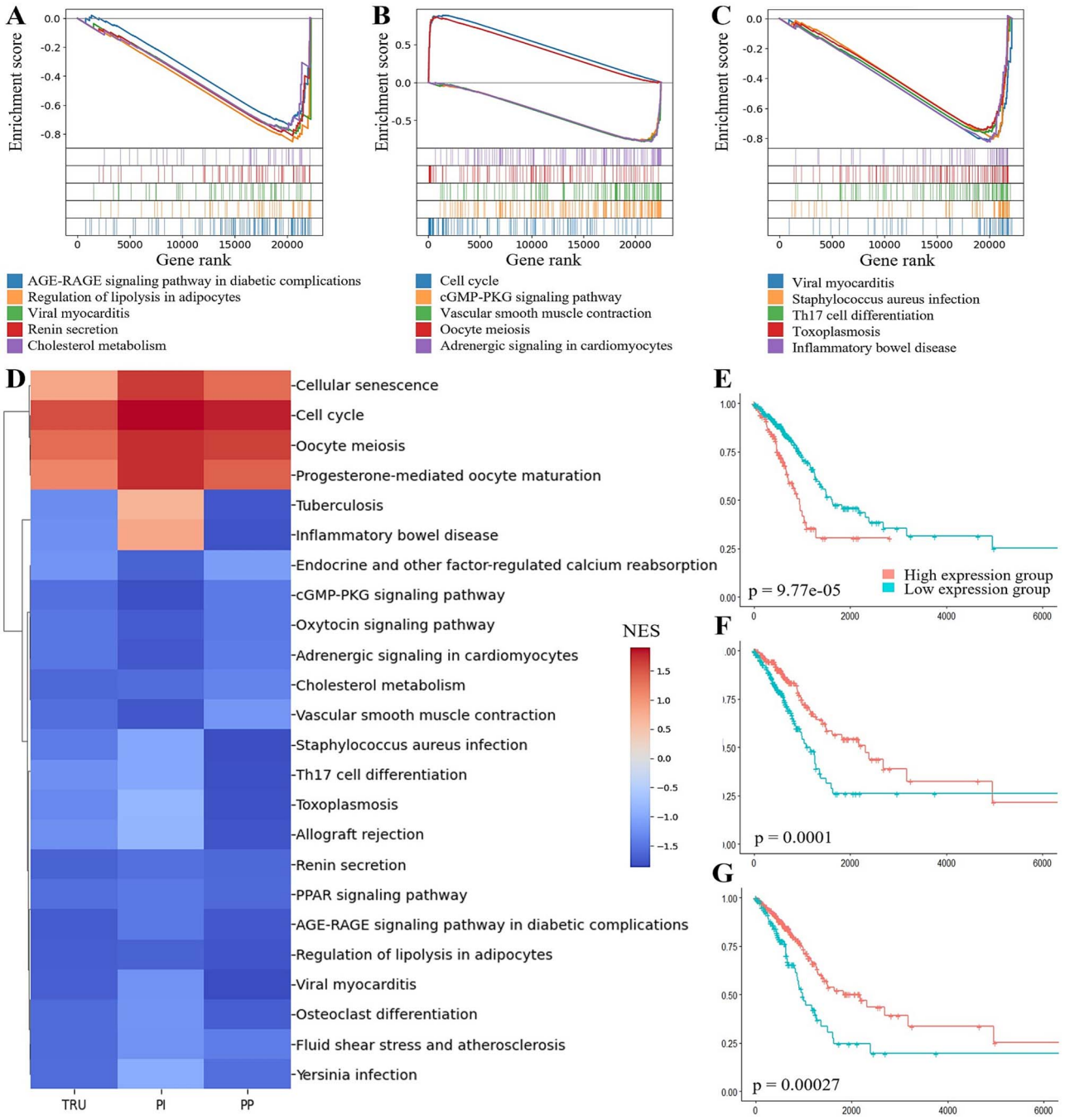

Figure 6. Pathway enrichment and survival analysis of LUAD subtypes. (A–C) GSEA enrichment plots comparing each LUAD subtype with normal samples: (A) TRU versus normal, (B) PI versus normal, and (C) PP versus normal; (D) heatmap showing the top 10 enriched pathways across LUAD subtypes; (E–G) Kaplan–Meier survival curves for LUAD patients stratified by expression levels of key genes: (E) *RHOV*, (F) *TSPAN11*, and (G) *NR1I2*.

heterogeneity, our findings suggest that incorporating biological priors into graph learning frameworks may aid in extracting more interpretable and clinically relevant features. Further work is needed to validate these findings and explore integrating additional data types, such as clinical variables and other omics, to enhance model performance and robustness.

Key Points

- MoAGNN is a hierarchical graph neural network that integrates transcriptome, miRNA, methylation, and CNV data to predict LUAD subtype, stage, and survival, with strong performance in subtype classification and improved interpretability.
- MoAGNN effectively balances predictive performance and interpretability by identifying biologically informative genes across omics layers.
- Kaplan–Meier survival analyses confirm the prognostic value of MoAGNN-identified genes, supporting their potential as biomarkers for risk stratification and therapeutic targeting in LUAD.

- GO and KEGG enrichment analyses of selected genes revealed strong involvement in TGF-β signaling regulation, extracellular matrix remodeling, and cell adhesion molecules, highlighting biological processes critical to LUAD progression.
- GSEA highlights a trajectory from metabolic stability in TRU, to proliferation in PI, and immune-metabolic suppression in PP, reflecting progressive reprogramming in LUAD.

Acknowledgements

The authors sincerely appreciate the library of National Yang Ming Chiao Tung University for providing effective database service and financially supporting this research.

Author contributions

C-P. L. (Conceptualization), C-P. L., Y-J. H. and T-Y. L. (Methodology), T-Y. L. (Resources), C-P. L. and Y-J. H. (Data curation), , C-P. L., Y-J. H. and Y-P. C. (Model construction and Validation), Y. T., Y. S. P. and G-T. C. (Investigation), C-P. L. (Writing—original draft preparation), Y-J. H., Y. T., and W-C. H. (Writing—review and editing), C-P. L. and Y-J. H. (Visualization), W-C. H. and T-Y. L (Supervision). All authors have read and agreed to the published version of the manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Funding

This work was financially supported by the Center for Intelligent Drug Systems and Smart Biodevices (IDS^{2}B) and Cancer and Immunology Research Center from The Featured Areas Research Center Program (114W031101) of the Higher Education Sprout Project and Yushan Young Fellow Program (114C51N039) by the Ministry of Education (MOE) and National Science and Technology Council (NSTC 114–2321-B-A49–009-, 114–2634-F-039-001, 114–2740-B-400-005, 114–2640-B-A49–001 and 113–2221-E-A49–160-MY3) in Taiwan and The National Health Research Institutes (NHRI-EX11411320BI, EX11511320BI) in Taiwan.

Data availability

Source code is available at https://github.com/leebiomicslab/MoAGNN.git

References

1. Siegel RL, Kratzer TB, Giaquinto AN, *et al.* Cancer statistics, 2025. *CA Cancer J Clin* 2025;**75**:10–45. https://doi.org/10.3322/caac.21871.
2. Myers DJ. *Lung Adenocarcinoma*. StatPearls, 2023.
3. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;**511**:543–50. https://doi.org/10.1038/nature13385.
4. Wilkerson MD, Yin X, Walter V, *et al.* Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PLoS One* 2012;**7**:e36530. https://doi.org/10.1371/journal.pone.0036530.
5. Wei J, Wang X, Guo H, *et al.* Subclassification of lung adenocarcinoma through comprehensive multi-omics data to benefit survival outcomes. *Comput Biol Chem* 2024;**112**:108150. https://doi.org/10.1016/j.compbiolchem.2024.108150.
6. Lee TY, Huang KY, Chuang CH, *et al.* Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. *Comput Biol Chem* 2020;**87**:107277. https://doi.org/10.1016/j.compbiolchem.2020.107277.
7. Dhillon A, Singh A, Bhalla VK. Biomarker identification and cancer survival prediction using random spatial local best cat swarm and Bayesian optimized DNN. *Appl Soft Comput* 2023;**146**:110649. https://doi.org/10.1016/j.asoc.2023.110649.
8. Kalakoti Y, Yadav S, Sundar D. SurvCNN: a discrete time-to-event cancer survival estimation framework using image representations of omics data. *Cancer* 2021;**13**:3106. https://doi.org/10.3390/cancers13133106.
9. Ouyang D, Liang Y, Li L, *et al.* Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. *Comput Biol Med* 2023;**164**:107303. https://doi.org/10.1016/j.compbiomed.2023.107303.
10. Park MK, Lim JM, Jeong J, *et al.* Deep-learning algorithm and concomitant biomarker identification for NSCLC prediction using multi-omics data integration. *Biomolecules* 2022;**12**:12. https://doi.org/10.3390/biom12121839.
11. Li X, Ma J, Leng L, *et al.* MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front Genet* 2022;**13**:806842. https://doi.org/10.3389/fgene.2022.806842.
12. Colaprico A, Silva TC, Olsen C, *et al.* TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016;**44**:e71. https://doi.org/10.1093/nar/gkv1507.
13. Ching T, Ha J, Song MA, *et al.* Genome-scale hypomethylation in the cord blood DNAs associated with early onset preeclampsia. *Clin Epigenetics* 2015;**7**:21. https://doi.org/10.1186/s13148-015-0052-x.
14. Ching T, Song MA, Tiirikainen M, *et al.* Genome-wide hypermethylation coupled with promoter hypomethylation in the chorioamniotic membranes of early onset pre-eclampsia. *Mol Hum Reprod* 2014;**20**:885–904. https://doi.org/10.1093/molehr/gau046.
15. Mermel CH, Schumacher SE, Hill B, *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;**12**:R41. https://doi.org/10.1186/gb-2011-12-4-r41.
16. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8.
17. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559. https://doi.org/10.1186/1471-2105-9-559.
18. Yan H, Weng D, Li D, *et al.* Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration. *Brief Bioinform* 2024;**25**:25. https://doi.org/10.1093/bib/bbae184.
19. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint.arXiv:1609.02907 2016. https://doi.org/10.48550/arXiv.1609.02907.

20. Lee J, Lee I, Kang J. Self-attention graph pooling. In: Chaudhuri K, Salakhutdinov R, (eds.), *Proceedings of the 36th International Conference on Machine Learning (ICML)* Brookline, MA, USA: PMLR; 2019, 3734–3743.
21. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324.
22. Hearst MA, Dumais ST, Osuna E, *et al.* Support vector machines. *IEEE Intel Syst Appl* 1998;**13**:18–28. https://doi.org/10.1109/5254.708428.
23. Bradley JB. Neural networks - a comprehensive foundation – Haykin, S. *Inf Process Manag* 1995;**31**:786–6. https://doi.org/10.1016/0306-4573(95)90003-9.
24. Singh P, Singh N, Singh KK, *et al.* Chapter 5 – diagnosing of disease using machine learning. In: Singh KK, Elhoseny M, Singh A, *et al.* (eds.), *Machine Learning and the Internet of Medical Things in Healthcare* London, UK: Academic Press, Elsevier, 2021, 89–111.
25. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics* 2023;**39**:btac757. https://doi.org/10.1093/bioinformatics/btac757.
26. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. *The gene ontology consortium*. *Nat Genet* 2000;**25**:25–9.
27. Xue J, Han W, Zhao W, *et al.* Fibrinogen beta chain as a prognostic biomarker and therapeutic target in tobacco-associated lung adenocarcinoma. *Discov Oncol* 2025;**16**:2014. https://doi.org/10.1007/s12672-025-03780-w.
28. Xu Z, Chen C. Abnormal expression and prognostic significance of bone morphogenetic proteins and their receptors in lung adenocarcinoma. *Biomed Res Int* 2021;**2021**:6663990. https://doi.org/10.1155/2021/6663990.
29. Raslan AA, Oh YJ, Jin YR, *et al.* R-Spondin2, a positive canonical WNT signaling regulator, controls the expansion and differentiation of distal lung epithelial stem/progenitor cells in mice. *Int J Mol Sci* 2022;**23**:3089. https://doi.org/10.3390/ijms23063089.
30. Pagani JH, Zhao M, Cui Z, *et al.* Role of the vasopressin 1b receptor in rodent aggressive behavior and synaptic plasticity in hippocampal area CA2. *Mol Psychiatry* 2015;**20**:490–9. https://doi.org/10.1038/mp.2014.47.
31. Zhang Q, Feng J, Liu K, *et al.* STK11 mutation impacts CD1E expression to regulate the differentiation of macrophages in lung adenocarcinoma. *Immun Inflam Dis* 2023;**11**:e958. https://doi.org/10.1002/iid3.958.
32. Chen Z, Shi H, Hu W, *et al.* DRP2 promotes EMT and serves as a potential therapeutic target for LUAD treatment. *Sci Rep* 2025;**15**:16590. https://doi.org/10.1038/s41598-025-01611-0.
33. Eymin B, Gazzeri S. Role of cell cycle regulators in lung carcinogenesis. *Cell Adhes Migr* 2010;**4**:114–23. https://doi.org/10.4161/cam.4.1.10977.
34. Lee S, Schmitt CA. The dynamic nature of senescence in cancer. *Nat Cell Biol* 2019;**21**:94–101. https://doi.org/10.1038/s41556-018-0249-2.
35. Wang S, An J, Hu X, *et al.* Single-cell RNA sequencing reveals immune microenvironment of small cell lung cancer-associated malignant pleural effusion. *Thorac Cancer* 2024;**15**:98–103. https://doi.org/10.1111/1759-7714.15145.
36. Zhou T, Yang P, Tang S, *et al.* Classification of lung adenocarcinoma based on immune checkpoint and screening of related genes. *J Oncol* 2021;**2021**:1–12. https://doi.org/10.1155/2021/5512325.
37. Zhang D, Jiang Q, Ge X, *et al.* RHOV promotes lung adenocarcinoma cell growth and metastasis through JNK/c-Jun pathway. *Int J Biol Sci* 2021;**17**:2622–32. https://doi.org/10.7150/ijbs.59939.
38. Chen H, Xia R, Jiang L, *et al.* Overexpression of RhoV promotes the progression and EGFR-TKI resistance of lung adenocarcinoma. *Front Oncol* 2021;**11**:619013. https://doi.org/10.3389/fonc.2021.619013.
39. Zhang H, Song Q, Shang K, *et al.* Tspan protein family: focusing on the occurrence, progression, and treatment of cancer. *Cell Death Dis* 2024;**10**:187. https://doi.org/10.1038/s41420-024-01961-0.
40. The Human Protein Atlas. TSPAN11. Uhlén M, (ed.), Stockholm, Sweden: SciLifeLab, 2025.

Briefings in Bioinformatics, 2026, **27(1)**, bbaf735
https://doi.org/10.1093/bib/bbaf735
Problem Solving Protocol