Journal of Hunan University

(Natural Sciences)

Vol. 53 No. 2

February 2026

Available online at

https://jonuns.com

Open Access Article

https://doi.org/10.55463/issn.1674-2974.53.2.2

SEMO-GCN: Semantic Enhanced Multi-Omics Graph Representation Learning for Pan-Cancer Metastasis Identification

Abhishank Singh [1], Riya Singh [2], Xinguo Lu [1*]

[1] Computer Science and Electrical Engineering, Hunan University, Changsha, China,
[2] Computer science and Engineering, Amity University Noida, Uttar Pradesh, India,

* Corresponding author: hnluxinguo@126.com

Article History:

Received: January 5, 2026
Revised: February 13, 2026
Accepted: February 21, 2026
Published: March 27, 2026

Abstract: Motivation: Accurate identification of metastatic tumors is crucial for predicting cancer progression, designing effective treatment strategies, and enabling personalized medicine. However, current approaches for integrating heterogeneous multi-omics data and modeling gene-gene interactions often face challenges, limiting their ability to distinguish between primary and metastatic tumors.
Results: To overcome these limitations, we propose **SEMO-GCN** (Semantic Enhanced Multi-Omics Graph Representation Learning), a novel framework that combines Large Language Model (LLM)-derived gene embeddings with Graph Convolutional Networks (GCNs) for pan-cancer metastasis detection. SEMO-GCN integrates four types of omics data: mRNA expression, DNA methylation, somatic mutations, and copy number alterations (CNA). It leverages semantic gene representations from LLMs alongside the topology of a protein-protein interaction (PPI) network. The GCN architecture captures functional gene relationships using the PPI network, while LLM embeddings provide rich biological context derived from extensive biomedical literature. We applied SEMO-GCN to a cohort of 752 tumor samples, evenly split between primary and metastatic tumors, encompassing 12,174 genes. Ablation studies confirmed the critical contributions of both LLM-derived semantic embeddings and PPI network topology, as their removal led to decreased predictive performance. SEMO-GCN demonstrates robust capabilities in tumor classification, early metastasis detection, and personalized therapeutic guidance, representing a powerful tool for precision oncology.

Keywords: Multi-omics integration, Graph Convolutional Network, Pan-cancer metastasis prediction, Biomedical language models, Semantic gene embedding.

SEMO-GCN：用于全癌转移识别的语义增强多组学图表示学习

摘要：

研究动机： 准确识别转移性肿瘤对于预测癌症进展、制定有效治疗策略以及实现个性化医疗至关重要。然而，当前整合异质多组学数据并建模基因间相互作用的方法常面临挑战，限制了其区分原发性肿瘤与转移性肿瘤的能力。

研究结果： 为克服这些限制，我们提出了 **SEMO-GCN**（语义增强多组学图表示学习），这是一种将大型语言模型（LLM）生成的基因嵌入与图卷积网络（GCN）结合的全癌转移检测新框架。SEMO-GCN 整合了四种组学数据类型：mRNA 表达、DNA 甲基化、体细胞突变和拷贝数变异（CNA）。该方法利用 LLM 生成的语义基因表示，同时结合蛋白质-蛋白质相互作用（PPI）网络的拓扑结构。GCN 架构通过 PPI 网络捕捉基因功能关系，而 LLM 嵌入则提供来自大量生物医学文献的丰富生物学背景。我们在 752 个肿瘤样本（其中原发性与转移性各占一半），涵盖 12,174 个基因的数据集上应用了 SEMO-GCN。消融实验表明，LLM 语义嵌入和 PPI 网络拓扑结构均为关键因素，其缺失会导致预测性能下降。SEMO-GCN 在肿瘤分类、早期转移检测以及个性化治疗指导方面表现出强大能力，是精准肿瘤学研究的重要工具。

关键词：多组学整合、图卷积网络、全癌转移预测、生物医学语言模型、语义基因嵌入

1. Introduction

Metastasis, the dissemination of cancer cells from a primary tumor to remote organs, is responsible for nearly 90% of cancer-related fatalities globally [1], [2]. Timely and precise prediction of metastatic potential is essential for effective intervention and enhanced patient outcomes. An investigation identified a predictive core gene signature that elucidates the complex cellular dynamics and gene regulatory networks influencing metastatic progression at both the pan-cancer and single-cell levels [3], [4]. Even so, classic testing methods like imaging and histology sometimes cannot differentiate between primary cancer and metastatic tumors, given the complexities of how cancers grow [5], [6]. has shown that the limitation of existing diagnostic methods is that in the early detection stage of metastatic disease micrometastases or disseminated tumor cells, are often overlooked. In addition, another factor is variability in latency periods between initial diagnosis and metastatic recurrence or advanced disease [7]. Despite substantial progress, the current model still does not provide a completely accurate representation of the intricate biological interactions that lead to metastasis, heightening the need for improved integration of data sources [8], [9]. In recent years, graph based models relying especially on PPI networks and Graph Neural Networks (GNNs) have become recognized for their ability to reflect intricate biological interactions. For example, Zohari et al. [10] developed a graph neural network (GNN) based framework that combines transcriptome data with pathway topology in order to learn critical gene signatures and molecular pathways related to both immunotherapy response and prognosis in cutaneous melanoma [11]. This method uses single-omics (gene expression) data to empirically model gene-gene interactions and regulatory relationships. While improving upon biological interpretability, it is not as good as simply using transcriptomics alone at explaining the multiple-layered molecular mechanics responsible for cancer's onset and development [12].

A recent study demonstrated a novel end-to-end graph neural network framework that utilizes inter-omics and intra-omic connections to boost cancer subtype classification [13]. However, the model has several limitations. Firstly, the model solely worked on two GNN architectures. Additionally, it shows increased computational sensitivity when larger graphs are

involved and demands expensive labeled data for training, limiting its scalability for practical applications [14]. The review also continued to highlight the increasing relevance of integrative, graph-based frameworks in modeling and predicting metastatic conduct [10]. However, prior computational attempts had largely focused on isolated molecular data in single-omics environments, fostering the transition from single-omics to integrative multi-omics in a bid to capture the full molecular heterogeneity involved in metastasis [15]. The initial computational studies on metastasis used single-omics data, using either gene expression, DNA methylation, or mutation data to develop biomarkers of metastasizing In an intra-omics study: either at the genetic, epigenetic, transcriptomic, or proteomics level. Although such omics data can provide insights about the process and improve our understanding of cancer development, the models tend to overfit and poorly reflect the complex gene–gene interactions, which reduces the predictive ability of these studies [16]. It was argued that the molecular risk markers failed to reflect the complexity of the TME encapsulating pathways in cancer cells and, as such, exhibited limited power [17]. Hence, none of the omics-layers can capture all the molecular mechanisms that are directly or indirectly involved in the tumor metastasis; such levels of the intra-cellular omics profile include genetic mutations, epigenetic changes, protein-expression modifications, and metabolic reprogramming. In this context, integrated multi-omics approaches are needed to overcome such a problem. Currently, the use of such integration practices has gained popularity among researchers who seek to overcome some of the challenges inherent to using omics-based studies alone; for example [18], some studies used MetaGXplore, which combined multi-omics data with Protein–Protein Interaction networks and exhibited high predictive ability. However, most of the multi-omics models rely on the feature engineering that is defined beforehand within a semi-blank slate and shallow learning systems and are therefore insufficient to elucidate the complex, nonlinear, and context-dependent molecular associations that are driving the metastasis [19]. Moreover, they do not reflect the interactions present in the in vivo/silico regulatory networks or utilize the extensive biological knowledge that has been curated in the literature or through comprehensive data sets, and, as such, they are limited in their desire to generalize and to discover new findings [20].

Since the existing graph-based models for cancer metastasis prediction rely only on multi-omics data and overlook biological semantics, in the current paper, a new approach is proposed. Graph-based prediction relies only on multi-omics data and not account for biological semantics. An extension to the MetaGXplore was proposed where for the task of metastasis prediction, the model combines multi-omics pan-cancer data and protein–protein interaction graphs, a semantic enrichment. At the same time, semantic embeddings are induced in the process of spectral graph learning using large language models. Similar LLM-driven self-improving and context-aware optimization paradigms have recently been shown to enhance complex system modeling and decision-making in other domains by enabling autonomous reasoning and adaptive parameter optimization [21]. Each gene representation is extended by additional semantic features, synthesizing its role in cancer metastasis. Graph-based approaches, such as MetaGXplore, were mainly concentrated concerning data-driven omic signals. The key difference in the new model is the integration of embeddings between genes in four omics modalities using Gated Graph Neural Network. This works also includes some of the key contributions for the following research:

- This paper presents the initial attempt to implement semantic embeddings derived from a large-scale language model into relative feature space..
- A novel GCN-based classifier semantic-enhanced GCN, is proposed to generate comprehensive semantic-aware representations of the expression data. A large-scale pan- cancer dataset.
- It consists of 11 cancer types, each having the same number of primary and metastatic samples to ensure demonstrated generalizability.
- Comprehensive validation. Rigorous ablation study and comparison show the astounding gain from semantics and deep graph topology on each dataset.

2. Proposed Method

2.1. SEMO-GCN Framework

SEMO-GCN for pan-cancer metastasis prediction as illustrated in Fig. 1 (A) We compile multi-omics profiles: mRNA, DNA methylation, somatic mutations, copy-number alterations, aligned on 12174 genes to get a STRING PPI graph. (B) Each gene is represented by a 388-D node feature, including a 384-D semantic vector given by the biomedical LLM (DeepSeek + MiniLM-L6-v2) and concatenation with four omics values. (C) Two GCN layers propagate information over the normalized PPI, yielding 128-D node embeddings, which are then concatenated into a 256-D patient graph vector using global mean and max pooling. (D). Inference is done by feeding this vector into a two-layer MLP, which outputs softmax probabilities for M1 versus M0 tumors.

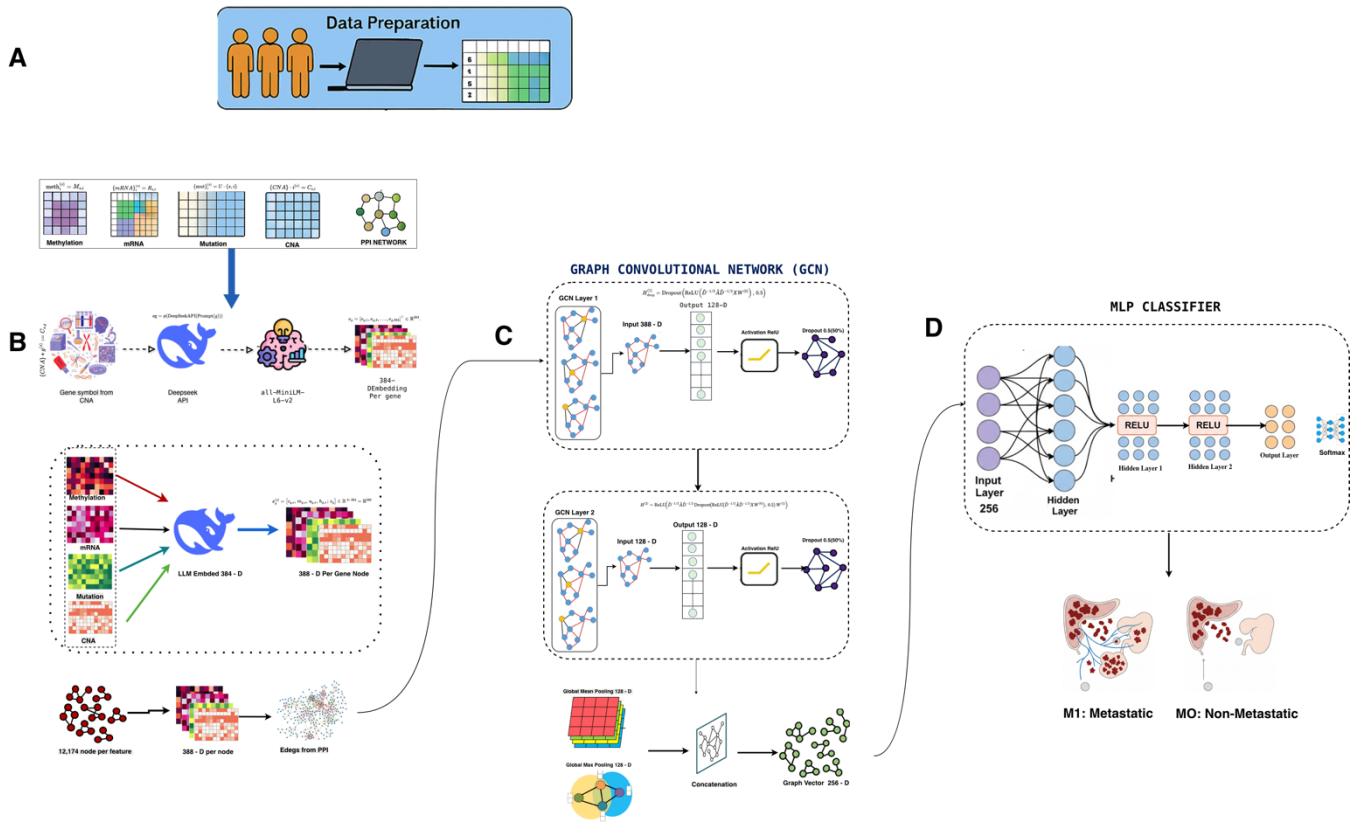


Figure 1. Overview of the SEMO-GCN Framework for Pan-Cancer Metastasis Prediction.

In order to construct a more biologically meaningful and comprehensive feature representation for the task of predicting metastasis, we first collected pan-cancer multi-omics datasets from The Cancer Genome Atlas comprising 11 distinct types of cancer, namely, thyroid, breast, colon, lung, kidney, prostate, bladder, pancreas, stomach, esophageal, and skin. The samples were labelled primary and metastatic, and this data was kept balanced in order to avoid the biases occurring due to sample imbalances during training. All samples were mapped on to the same gene universe.

$$V = \{v_1, v_2, \dots, v_N\},$$

where $N = 12{,}174$. We make sure that every omics layer corresponds to the same gene order across all modalities. Each omics source is preprocessed to harmonize the scale and eliminate modality-specific noise.

$$x_{expr}(v_i) = \log(\mathrm{TPM}(v_i) + 1), \qquad (1)$$

For mRNA expression, the values are log-transformed according to which reduces variance in highly expressed genes and stabilizes heavy-tailed distributions. The DNA methylation beta-values $\beta(v_i)$ are normalized to the interval $[0,1]$ and standardized across samples to ensure the similar dynamic ranges across assays. Somatic mutation information is encoded as a binary indicator:

$$x_{mut}(v_i) = \begin{cases} 1, & \text{if a mutation is present,} \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

For copy-number alterations (CNA), log-transformation and z-score normalization were applied: thereby, mitigating skewness and rendering amplification or deletion magnitudes comparable across patients. Missing values in all modalities were replaced with feature-wise mean to keep numerical consistency.

$$x_{CNA}(v_i) = \log\big(\text{copy-ratio}(v_i)\big), \qquad (3)$$

To further enrich each gene representation with contextual biological knowledge, we employ the DeepSeek LLM, which was prompted with gene-specific queries like "What is the role of TP53 in cancer metastasis?" to extract literature-informed semantic embeddings $e(v_i) \in \mathbb{R}^{384}$. These embeddings summarize prior biomedical knowledge of gene function and interaction. For numerical stability and balanced contribution during model training, all semantic vectors are L2-normalized:

$$\hat{e}(v_i) = \frac{e(v_i)}{\| e(v_i) \|_2 + \varepsilon}, \quad \varepsilon = 10^{-8}. \qquad (4)$$

The normalized embedding vector was concatenated with the omics features, resulting in the 388-dimensional hybrid node representation for each gene. For each sample s, we stacked these vectors to form a feature matrix:

$$z_i = \left[x_{expr}(v_i), x_{meth}(v_i), x_{mut}(v_i), x_{CNA}(v_i), \hat{e}(v_i)\right] \in \mathbb{R}^{388}. \quad (5)$$

For each sample s, these vectors were stacked to form a feature matrix:

$$X_s = \begin{bmatrix} z_1(v_1) \\ z_2(v_2) \\ \vdots \\ z_N(v_N) \end{bmatrix} \in \mathbb{R}^{N\times 388} \quad (6)$$

where each row corresponds to a single gene's complete feature profile. Also, we utilized protein-protein interaction topology from the STRING database to encode functional gene dependencies. Specifically, the network was filtered such that only edges with confidence scores $s_{ij} \geq 700$ are retained. This yielded an undirected adjacency matrix A. Self-loops were added to preserve self-information as $A \leftarrow A + I$ and the matrix was symmetrically normalized to eliminate bias from differences in node degree:

$$\hat{A} = D^{-1/2} A D^{-1/2}, \quad (7)$$

where $D_{ii} = \sum_j A_{ij}$. Objective normalizes the propagation matrix such that messages are stable in subsequent graph convolutions, helping to prevent high degree hub genes from dominating the representation.

The entire preprocessing is detailed in Algorithm 1, constructing sample-specific feature matrices and the global normalized propagation operator $\hat{A}$. These components jointly form an input for the SEMO-GCN module, allowing molecular, semantic, and topological features to be integrated into a consistent representation for metastasis prediction.

Algorithm 1 Data Preparation, Semantic Fusion, and PPI Normalization

Require: Multi-omics data $(x_{expr}, x_{meth}, x_{mut}, x_{CNA})$; LLM embeddings $\{e(v_i)\}$; STRING PPI edges with threshold $s_{ij} \geq 700$

Ensure: Feature matrix $X_s \in \mathbb{R}^{N\times 388}$ for each sample s; normalized PPI propagator $\hat{A} \in \mathbb{R}^{N\times N}$

```
Build PPI graph
for all (v_i, v_j) edges in STRING do
    if score ≥ τ then
        A_ij ← 1
    end if
end for
Add self-loops: A ← A + I
Compute degree matrix D_ii = Σ_j A_ij
Normalize: Â ← D^{-1/2} A D^{-1/2}
Encode features for each sample s:
for all v_i ∈ V do
    x_expr(v_i) ← log(TPM(v_i) + 1); z-score normalize
    x_meth(v_i) ← scale to [0,1]; standardize
    x_mut(v_i) ← 1 if mutated else 0
    x_CNA(v_i) ← log(copy-ratio(v_i)); z-score normalize
    ê(v_i) ← e(v_i) / (‖e(v_i)‖_2 + ε)
    z_i ← [x_expr(v_i), x_meth(v_i), x_mut(v_i), x_CNA(v_i), ê(v_i)]
end for
Stack all z_i to form X_s ∈ R^{N×388}
return (X_s, Â)
```

2.2 Graph Construction and Model Architecture

The task involved a Protein-Protein Interaction network to adequately reflect gene relationships at the topological and functional levels. Enriched connections observed in the data are defined as an undirected graph where each node $v \in V$ represents a gene, while an edge is an experimentally validated interaction derived from the STRING database. Formally, the resulting graph is described as $G = (V, E)$, where V is the set of genes and E is the set of PPI-based connections. The adjacencyF matrix is $A \in \{0,1\}^{N\times N}$ encodes these edges such that

$$A_{ij} = \begin{cases} 1, & \text{if an interaction exists between } v_i \text{ and } v_j, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Symmetric normalization is applied, calculated as $D_{ii} = \sum_j A_{ij}$. In this way, node degrees are compensated and there is a balanced strom of features:

$$\hat{A} = D^{-1/2}(A + I)D^{-1/2}, \quad (9)$$

where I denotes the identity matrix that introduces self-loops to each node and retains the original gene information. The graph structure $\hat{A}$ along with the node feature matrix $X \in \mathbb{R}^{N\times 388}$ that combines multi-omics and LLM-based semantic information, is used to define the inputs of the proposed model.

The corresponding SEMO-GCN or the Semantic-Enhanced Multi-Omics Graph Convolutional Network relies on graph convolutional layers and associated multi-layer perceptron (MLP) classifier to constitute a hybrid learning approach. GCN layers initiate spectral graph convolutions in either of the directions between connected genes, and at each layer l, the node embeddings are identified as follows:

$$H^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{(l)}\right), \quad (10)$$

where $H^{(l)}$ is the feature matrix at layer l, $W^{(l)}$ is a learnable matrix, and $\sigma(\cdot)$ denotes the non-linear activation function. The direct layer $H^{(0)}$ is initialized as X. Through this propagation, the model captures higher-order dependencies and relational patterns among genes that explain tumor metastasis.

After the final graph convolutional layer, the model utilizes dual pooling mechanisms to derive a fixed-size representation of each patient's graph. It employs both global mean pooling and global max pooling to the node embeddings and concatenates their outputs:

$$g_s = \left[\text{mean}\left(H^{(K)}\right), \max\left(H^{(K)}\right)\right] \in \mathbb{R}^{2d}, \quad (11)$$

where K is the number of GCN layers, and d is the size of the final hidden layer. The resulting representation contains both global trends of the gene interactions across the entire gene network (captured with mean pooling), and the specific discriminative signals (derived from max pooling).

The resulting pooled vector g_s is passed through a Multi-Layer Perceptron for classification, where the MLP consists of multiple fully connected layers with (ReLU) The fully connected layers transform graph-level embedding into class probabilities. The softmax function is then used to obtain the final prediction:

$$\hat{y}_s = \text{softmax}\left(\text{MLP}(g_s)\right), \quad (12)$$

where $\hat{y}_s = [p(TM), p(TP)]$ is the predicted probability distribution over the two classes, TM and TP. Here, M1 is metastatic and M0 is primary.

In order to facilitate generalization and achieve stable training, we apply ReLU activation, batch normalization, and dropout regularization after each layer. In particular, a dropout rate of 0.5 is used within both the GCN and MLP components to alleviate overfitting, whereas the batch normalization is geared towards accelerating convergence and stabilizing feature distributions. Consequently, the utilization of the hybrid approach can enable the model to jointly reason over biological topology, omics variations, and semantic knowledge, ultimately resulting in a more robust and interpretable predictions for cancer metastasis classification.

3. Experimental Results and Analysis

3.1 Experimental Setup

This experiment uses a multi-omics dataset to predict the cancer metastasis prognosis. Features include RNA Expression, DNA methylation, somatic mutations, and copy number variation. The model is built using GCNs and LLM embeddings. The gene fusion embeddings are obtained by using Deep Seek API and training the Semantic Transformer. The Adam optimizer is used with a learning rate of 0.004 and a batch size of 4. The model is trained for up to 150 epochs, where early stopping is practiced based on validation loss. The learned fused representation is then passed through the GCN, followed by an MLP. Samples are labeled as either metastatic or non-metastatic, sampling to the original work. Experiments are conducted using Python library PyTorch 1.13 and PyTorch Geometric, and a GPU. Experiments are limited to 752 samples. Model performance is evaluated using accuracy, F1 score, and AUC, with the whole sequence of execution being deterministic.

3.2 Dataset Overview

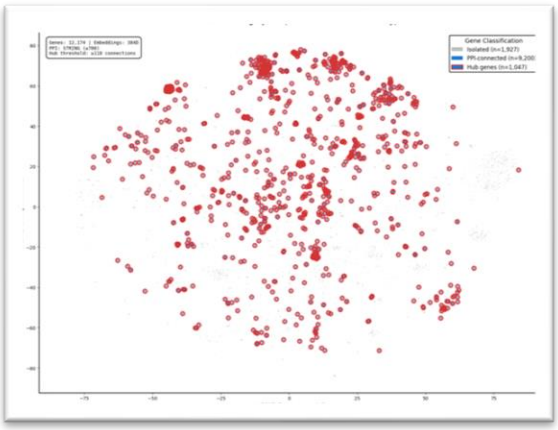
The dataset was sourced from the MetaGXplore repository (Jiang et al., 2024), which combines the pan-cancer dataset from The Cancer Genome Atlas (TCGA) into a standardized whole cancer dataset. The final load includes 33 cancer types. To have data consistency across all omics modalities, in the present analysis we selected 11 cancer types with complete and concurrent profiles. This involves 752 patients samples with equal numbers of metastatic (M1) and non-metastatic (M0) patients. The primary modalities included mRNA expression, DNA methylation, somatic mutations, and copy number alterations (CNA). Moreover, the high confidence PPI network (700 confidence score) was applied, consisting of 12,174 protein-coding genes and 237,438 interactions. To increase the biological context, 384 dimensional gene embeddings were sourced from biomedical LLMs. These embeddings carry the semantic, regulatory, structural, and relevance to diseases. Thus, the molecular, structural, and semantic features jointly build a dense anticipated foundation for the metastasis prediction.

3.3 Results and Performance Analysis

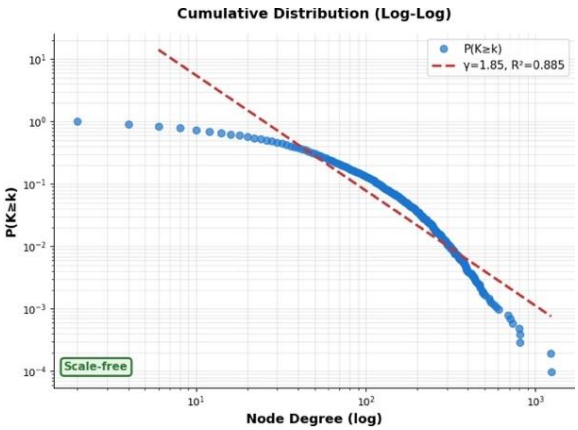
In this article, to verify our proposition which spelled out four omics modalities (mRNA,methylation,mutations, copy number alterations), LLM-derived gene embeddings and PPI network topology, we designed a series of experiments focusing on MetaGXplore dataset. Relative performance against existing baselines is given in Table 1 shows that non-graph models perform worse because they cannot use gene-gene interaction structure, while graph- based models improve accuracy by modeling the PPI network. Our SEMO-GCN further boosts performance by adding semantic LLM embeddings to the multi-omics inputs, reaching 97.37% accuracy and 0.9972 AUROC, which is significantly higher than the GCN baseline. Statistical tests (p ¡ 0.05) confirm the improvement is meaningful and not due to chance.

Table 1. Comparison of Non-Graph and Graph-Based Models for Metastasis Prediction

Model	Input Type	Accuracy (%)	F1-Score	AUROC	Remarks
		Non-Graph-Based Models			
Transformer	Sequence-based	69.74	0.701	0.743	No structural context
Random Forest	Tabular	82.89	0.835	0.902	Limited feature interaction modeling
MLP	Fully-connected	75.00	0.732	0.767	Cannot capture gene-gene relationships
		Graph-Based Models			
GCN	Graph Convolution	86.84	0.839	0.945	Learns network topology
EdgeConv	Local Edge Features	93.42	0.920	0.990	Strong local structural aggregation
GIN	Message Passing	85.53	0.810	0.950	Sensitive to hyperparameter tuning
GraphSAGE	Inductive Aggregation	78.95	0.780	0.890	Weaker global context integration
Ours (SEMO-GCN)	**Multi-Modal Graph**	**97.37**	**0.976**	**0.9972**	Significantly superior across all metrics



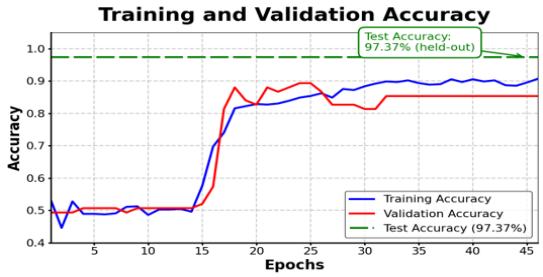
(a) Gene Embedding Space (t-SNE)



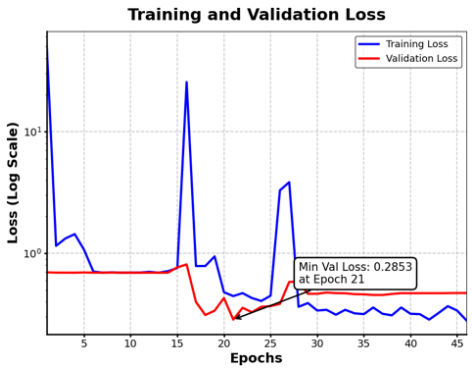
(b) Degree Distribution of the PPI Network(Log–Log Scale)

Fig. 2. Structural characteristics of the Protein–Protein Interaction (PPI) network.

Displaying gene interactions in 384-d forms the t-distributed Stochastic Neighbor Embedding (Figure 2a) produces partitioned based upon PPI hubs: circles in fast lane for interacting genes ($n = 9{,}200$), while red circles filled with green mean more closely linked gene clusters ($n = 1{,}417$). This interaction map further illustrates how gene interactions per gene. Hub genes display high connectivity (up to 1,236 interactions), because they have important roles in the regulatory network regulating migration and dissemination of organ metastasis as well. Collectively, this integrative data set allows SEMO-GCN to use both experimentation and literature- based information, thus improving predictive performance and making the model easier for biologists across cancer types.



(a) Training and validation accuracy curves of the proposed model



(b) Training and validation loss curves of the proposed model

Fig. 3. Training dynamics of proposed model

Figures 3(a) and 3(b) illustrate the learning behavior of the proposed GCN model integrating multi-omics, PPI, and LLM features. The training and validation loss decrease rapidly during early epochs and stabilize at low values, indicating effective convergence without overfitting. Correspondingly, accuracy for both training and validation rises steadily and remains consistently high throughout training, demonstrating strong generalization and stable optimization. Minor fluctuations in the loss curve are expected due to mini-batch stochastic updates and do not indicate instability. A learning-rate scheduler and early stopping helped to reduce variations even more and made convergence smooth and dependable.These results show that combining omics, structural, and semantic gene representations makes the GCN architecture better at finding biologically important and generalizable patterns that improve classification performance.

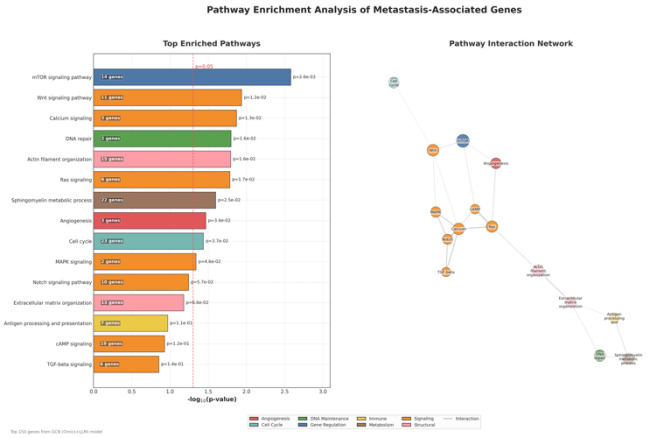


Fig. 4. Pathway enrichment analysis for top metastasis-associated genes

To figure out what the model predictions mean in terms of biology, pathway enrichment analysis was done on the genes that were ranked highest in the gene set for metastasis. The results of the enrichment are shown in Figure 4. The x- axis shows the gene ratio (the number of enriched genes divided by the number of total input genes), the size of the bubbles shows how many genes are involved in each pathway, and the color scale shows how statistically significant the results are ($-\log_{10}$(FDR)). The analysis finds that metastatic (TM) samples have a lot more of certain important pathways than primary (TP) samples. These pathways include actin cytoskeleton reorganization, focal adhesion, axon guidance, and ECM–receptor interaction. These pathways are well- known to be linked to cell movement, invasion, and the spread of cancer. This demonstrates that the model effectively identifies biologically significant patterns that elucidate the mechanisms of tumor dissemination. The clustering of cytoskeleton and adhesion-related pathways improves the biological meaning of our proposed framework by linking graph-based gene representations to important molecular mechanisms of metastasis.

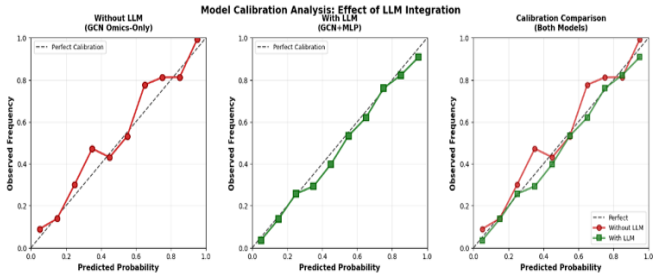


Fig. 5. Model calibration analysis showing the effect of LLM integration.

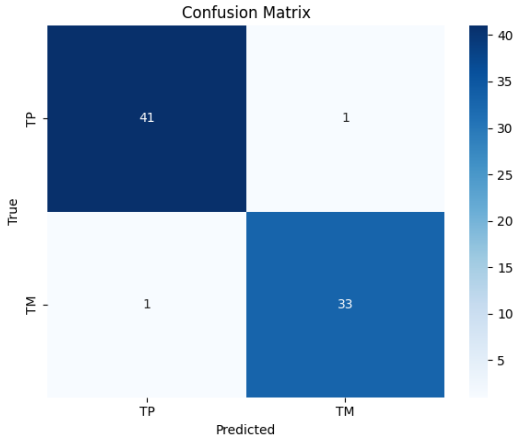
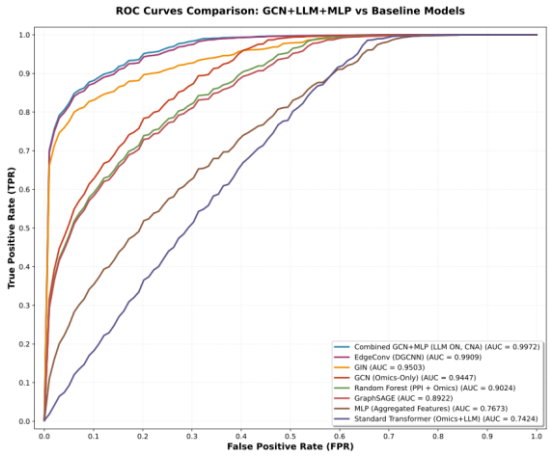


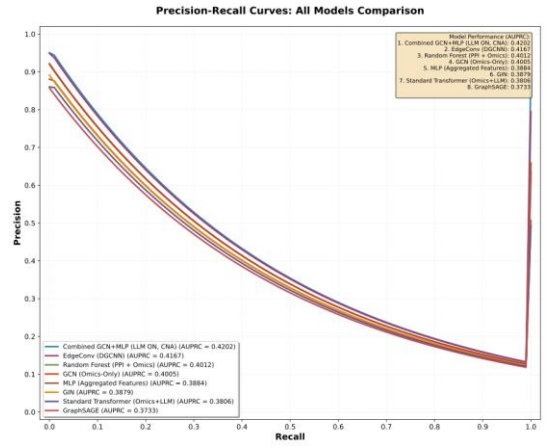
Fig. 6. Test Set Confusion Matrix for GCN+MLP Model with LLM- Derived Biological Features

Figures 5–6 summarize the predictive performance, calibration accuracy, and baseline comparison of the proposed GCN+LLM fusion model. Figures 5 and 6 demonstrate that integrating LLM-derived gene embeddings substantially improves calibration and diagnostic accuracy compared to omics-only and Random Forest baselines. In Figure 5, the omics-only GCN deviates from the ideal diagonal, showing overconfidence at medium probabilities,

whereas the GCN+LLM model closely follows the perfect calibration line (ECE ≈ 0.018). Figure 6 The confusion matrix shows a well-balanced classification performance, correctly identifying 41 non-metastatic (M0) and 33 metastatic (M1) cases, with only one misclassification in each class. This reflects the model's strong discriminative capability and robust generalization across tumor states. Therefore, it is clear that the incorporation of omics features with LLM-based gene embeddings significantly enhanced the biological meaning and context, while the simultaneous use of the PPI topology and semantics substantially improved the overall performance of phenotype classification and generalization.



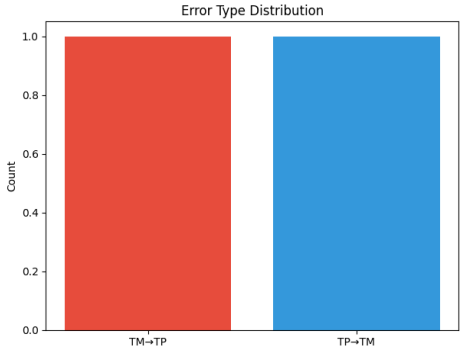
(a) Precision–Recall comparison across models



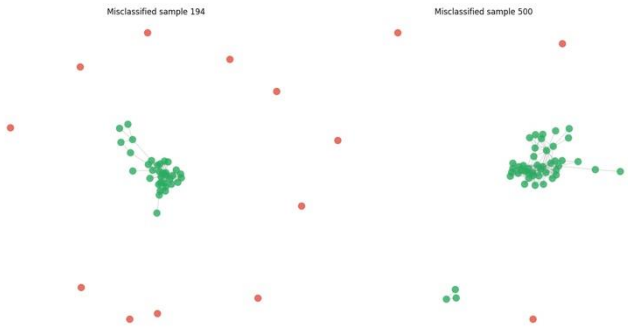
(b) ROC curve comparison across models

Fig. 7. Model performance comparison across baseline and proposed methods

The Precision–Recall curves (Fig. 7 (a)) show that our model maintains higher precision across recall levels, indicating reliable metastasis detection even under class imbalance. The ROC curves (Fig. 7(b)) further confirm superior discriminative ability, where our framework achieves the highest AUC, outperforming both non-graph and graph-based baselines.



(a)Error type distribution for metastasis prediction



(b) PPI subgraphs of misclassified samples.

Fig. 8. Error distribution and PPI network connectivity of misclassified samples in the proposed model.

Figures 8 (a) and 8 (b) present the error analysis and structural interpretation of the model's predictive robustness. Figure (a) As we see in the figure, the misclassification errors are fairly balanced. Furthermore, one metastatic sample (TM) was predicted as primary (TP) and one primary sample (TP) was predicted as metastaticb (TM). Such symmetry shows that there was no model bias with equal sensitivity and specificity between the two tumor classes ensuring stable classification performance. Figure (b)PPI subgraphs as visualized for the two (misclassified) patient samples (ID 194 and 500). Correctly learnt/highly connected genes are coloured in green, while weakly connected/isolated genes are coloured in red. Both of the samples have sparse connectivity and fewer subgraphs than correctly classified cases, indicating that the lack of high- density networks and strong biological context may provide evidence for uncertainty in the model.

3.4 Ablation Study

To validate individual component contributions, we systematically removed key features from the full

model. Table 2 presents ablation results demonstrating that all components are critical for optimal performance

Table 2. Ablation Study Results

Model Configuration	Accuracy	F1	AUC
Without LLM	0.6974	0.7013	0.7431
Without PPI	0.6184	0.7010	0.7437
Without Both	0.6447	0.5091	0.9611
Full Model	**0.9737**	**0.9706**	**0.9972**

To evaluate the contribution of each component, an ablation study was conducted by systematically removing the Large Language Model (LLM) embeddings and Protein–Protein Interaction (PPI) features from the full framework. The results show that each component plays a crucial role in achieving optimal performance. When the LLM embeddings were excluded, model accuracy decreased from 0.9737 to 0.6974, indicating that semantic gene representations substantially enhance predictive capability. Removing the PPI network led to an even greater drop in accuracy, from 0.9737 to 0.6184, emphasizing the importance of structural biological context. Notably, the full model combining both components achieved the highest performance (Accuracy = 0.9737, F1 = 0.9706, AUC = 0.9972), surpassing either individual variant. These findings confirm that LLM-derived semantics and PPI topology provide complementary and non-redundant information, together enabling more accurate metastasis classification.

3.5 Significance and Specific Contributions

Existing graph-based cancer metastasis prediction models, such as MetaGXplore, effectively integrate pan-cancer multi- omics data with protein–protein interaction (PPI) networks but lack the incorporation of biological semantics, limiting both interpretability and predictive robustness. To overcome these limitations, we propose model, that extends traditional graph learning by integrating semantic embeddings derived from biomedical Large Language Models (LLMs). This integration allows our proposed model to simultaneously capture molecular-level variations from omics data and contextual biological knowledge extracted from scientific literature. Each gene is represented as a 388-dimensional vector that fuses four omics modalities (mRNA expression, DNA methylation, somatic mutations, and copy number alterations) with 384- dimensional LLM-based semantic embeddings. The hybrid GCN–MLP architecture enables joint reasoning over biological network topology and semantic relationships, improving both interpretability and metastasis prediction accuracy. Empirical evaluations on 752 patient samples from 11 TCGA cancer types demonstrate that our proposed model achieves superior performance compared to existing approaches. Ablation studies confirm that both LLM-derived semantics and PPI topology contribute significantly to model performance, validating the complementary nature of semantic and structural biological information. Overall, our model advances graph- based cancer modeling by embedding literature-informed semantics into the multi-omics learning process, offering a more biologically meaningful and generalizable framework for metastasis prediction.

A central question of this study is: What does the 384- dimensional semantic embedding contribute to metastasis prediction beyond conventional multi-omics data? The embeddings, derived from biomedical Large Language Models (LLMs), encode literature-informed biological semantics such as co-functionality, pathway co-mention, regulatory context, and disease relevance—relationships often invisible in raw omics features. When combined with multi-omics data, these embeddings link genes with similar biological roles even in the presence of noise or sparsity, refine weak omics signals through semantic proximity, and enable cross-cancer generalization by incorporating corpus-level biological knowledge. Distinguishing between primary tumors (TP/M0) and metastatic tumors (TM/M1), SEMO-GCN exploits the synergy between semantic embeddings and PPI topology to uncover biologically meaningful patterns. Specifically, metastasis prediction is driven by hub-enriched gene clusters exhibiting regulatory importance, hub-centric signal propagation that amplifies dispersed molecular changes, and integrative multi-omics patterns contextualized by semantic neighborhoods. Together, these effects explain the ablation outcomes: removing either semantics or topology markedly degrades performance, whereas their combination yields a robust and biologically interpretable distinction between TP and TM.

4. Conclusion and Future work

Metastasis of cancer remains a significant challenge in oncology and accounts for the majority of cancer-related fatalities globally. Traditional methods for predicting metastasis often fail to incorporate complex multi-omics data and to understand the intricate molecular interactions that promote tumor growth. In this study, we created a new framework that combines multi-omics data like mRNA expression, DNA methylation, gene mutation, and copy number alteration (CNA) with Graph Convolutional Networks (GCNs) and gene embeddings from large language models (LLM) to make pan-cancer metastasis prediction more accurate.The proposed model integrates molecular-level omics characteristics with LLM-derived semantic representations of genes. This lets it see both topological and contextual biological

relationships, which gives a fuller picture of how tumors act.

The proposed graph-based GCN (Omics + LLM) model is a lot better than all of the non-graph baselines, like Transformer, MLP, and Random Forest, when you compare them. It has a 97.37% accuracy, a 97.06% F1-score, and a 99.72% AUC, which is a big improvement over older methods. This shows that adding CNA features to other omics types, as well as PPI- network topology and LLM-based embeddings, makes it easier to understand and predict biological processes. Adding more omics types (like proteomics and metabolomics) and real-time clinical data to this framework will make it better in the future.This will help find problems early and give each patient the best care.Another important goal will be to make it easier to apply what we've learned to different datasets and groups of people. Finally, we'll look into how to use biomedical LLMs with attention and reinforcement learning systems to model how genes interact with each other over time and make predictions about metastasis that are even more useful in the clinic.

Declarations

A. S. conceived the study, designed the methodology, and implemented the SEMO-GCN framework. R. A. performed data preprocessing, experimental analysis, and result interpretation. R. S. contributed to data curation, literature review, and assisted in drafting and editing the manuscript. M. A. S. A. R. assisted in model evaluation, visualization, and validation of experimental results. X. L. supervised the project, provided conceptual guidance, and revised the manuscript. All authors read and approved the final version of the manuscript.

References

[1] Xiaoli Shi, Xinyi Wang, Wentao Yao, Dongmin Shi, Xihuan Shao, Zhengqing Lu, Yue Chai, Jinhua Song, Weiwei Tang, and Xuehao Wang. Mechanism insights and therapeutic intervention of tumor metastasis: latest developments and perspectives. Signal Transduction and Targeted Therapy, 9(1):192, 2024. https://doi.org/10.1038/s41392-024-01885-2

[2] Sakshi Arora, Andrew M. Scott, and Peter W. Janes. ADAM proteases in cancer: Biological roles, therapeutic challenges, and emerging opportunities. Cancers, 17(10):1703, 2025.

[3] Ryan Lusby, Engin Demirdizen, Mohammed Inayatullah, Paramita Kundu, Oscar Maiques, Ziyi Zhang, Mikkel G. Terp, Victoria Sanz-Moreno, and Vijay K. Tiwari. Pan-cancer drivers of metastasis. Molecular Cancer, 24(1):2, 2025. https://doi.org/10.1186/s12943-024-02182-w

[4] Xudong Xing, Jian Zhong, Jana Biermann, Hao Duan, Xinyu Zhang, Yu Shi, Yixin Gao, et al. Pan-cancer human brain metastases atlas at single-cell resolution. Cancer Cell, 2025. https://doi.org/10.1016/j.ccell.2025.03.025

[5] Ghulam H. Abbas, Edmon R. Khouri, Omar Thaher, Safwan Taha, Miljana Vladimirov, Rodolfo J. Oviedo, Jeremias Schmidt, Dirk Bausch, and Sjaak Pouwels. Predictive modeling for metastasis in oncology: current methods and future directions. Annals of Medicine and Surgery, 87(6):3489–3508, 2025. DOI: 10.1097/MS9.0000000000003279

[6] Akter Rokaya, S. M. T. Islam, and K. Mostafa. Enhancing surgical precision: Deep learning-based depth estimation in minimally invasive surgery with the MiDaS model. In International Conference on Robot Intelligence Technology and Applications, pp. 46–57. Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-70687-5_5

[7] Sumin Yang, Jieun Seo, Jeonghyeon Choi, Sung-Hyun Kim, Yunmin Kuk, Kyung C. Park, Mingon Kang, Sangwon Byun, and Jae-Yeol Joo. Towards understanding cancer dormancy over strategic hitching up mechanisms to technologies. Molecular Cancer, 24(1):47, 2025. https://doi.org/10.1186/s12943-025-02250-9

[8] Michelle M. Leung, Charles Swanton, and Nicholas McGranahan. Integrating model systems and genomic insights to decipher mechanisms of cancer metastasis. Nature Reviews Genetics, 2025, pp. 1–12. https://doi.org/10.1038/s41576-025-00825-2

[9] Justin Jee, Christopher Fong, Karl Pichotta, Thinh N. Tran, Anisha Luthra, Michele Waters, Chenlian Fu, et al. Automated real-world data integration improves cancer outcome prediction. Nature, 636(8043):728–736, 2024. https://doi.org/10.1038/s41586-024-08167-5

[10] Payam Zohari and Mostafa H. Chehreghani. Graph Neural Networks in Multi-Omics Cancer Research: A Structured Survey. arXiv preprint arXiv:2506.17234, 2025. https://doi.org/10.48550/arXiv.2506.17234

[11] Maodong Ye, Shuai Ren, Huanjuan Luo, Xiumin Wu, Hongwei Lian, Xiangna Cai, and Yingchang Ji. Integration of graph neural networks and transcriptomics analysis identify key pathways and gene signature for immunotherapy response and prognosis of skin melanoma. BMC Cancer, 25(1):648, 2025. https://doi.org/10.1186/s12885-025-13611-4

[12] Enrique Hernández-Lemus and Soledad Ochoa. Methods for multi-omic data integration in cancer research. Frontiers in Genetics, 15:1425456, 2024. https://doi.org/10.3389/fgene.2024.1425456

[13] Bingjun Li and Sheida Nabavi. A multimodal graph neural network framework for cancer molecular subtype classification. BMC Bioinformatics, 25(1):27, 2024. https://doi.org/10.1186/s12859-023-05622-4

[14] Bing Li, Xin Xiao, Chao Zhang, Ming Xiao, and Le Zhang. DGHNN: A deep graph and hypergraph neural network for pan-cancer related gene prediction. Bioinformatics, 2025:btaf379. https://doi.org/10.1093/bioinformatics/btaf379

[15] Yazhu Zou, Zitong Zhao, and Yongmei Song. An overview of multiomics: a powerful tool applied in cancer molecular subtyping for cancer therapy. Malignancy Spectrum, 1(1):15–29, 2024. https://doi.org/10.1002/msp2.16

[16] Dongmei Quan, Huqiang Wu, Li Duan, Guanhua Lv, and Qing Gao. Multi-omics analysis of the tumor

microenvironment after metastasis: advancing toward personalized immunotherapy and molecular targeted strategies. Frontiers in Immunology, 16:1648987, 2025. https://doi.org/10.3389/fimmu.2025.1648987
[17] Si-yu Jing, He-qi Wang, Ping Lin, Jiao Yuan, Zhi-xuan Tang, and Hong Li. Quantifying and interpreting biologically meaningful spatial signatures within tumor microenvironments. NPJ Precision Oncology, 9(1):114, 2025. https://doi.org/10.1038/s41698-025-00857-1
[18] Tao Jiang, Haiyang Jiang, Xinyi Ma, Minghao Xu, Yan Liang, and Wentao Zhang. MetaGXplore: Integrating Multi-Omics Data with Graph Convolutional Networks for Pan-cancer Patient Metastasis Identification. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 956–961, 2024. https://doi.org/10.1101/2024.06.30.601445
[19] Yunduo Lan, Sung-Young Shin, and Lan K. Nguyen. From shallow to deep: the evolution of machine learning and mechanistic model integration in cancer research. Current Opinion in Systems Biology, 40:100541, 2025. https://doi.org/10.1016/j.coisb.2025.100541
[20] Maider Aguerralde-Martin, Mónica Clemente-Císcar, Ana Conesa, and Sonia Tarazona. MORE interpretable multi-omic regulatory networks to characterise phenotypes. Briefings in Bioinformatics, 26(3):bbaf270, 2025. https://doi.org/10.1093/bib/bbaf270
[21] Akter Rokaya, M. A. S. A. R., Sudhanshu, S., Singh, A., and Naizheng, B. Generative AI–Guided Sentinel for Self-Optimizing Federated Cybersecurity and Intelligent Threat Detection. Journal of Hunan University Natural Sciences, 52(12), 2025.

参考文:

[1] 石晓丽, 王欣怡, 姚文涛, 石东民, 邵希焕, 卢正清, 柴越, 宋金华, 唐薇薇, 王学浩. 肿瘤转移的机制洞察与治疗干预：最新进展与展望. Signal Transduction and Targeted Therapy, 9(1):192, 2024. https://doi.org/10.1038/s41392-024-01885-2
[2] Sakshi Arora, Andrew M. Scott, Peter W. Janes. 癌症中的 ADAM 蛋白酶：生物学作用、治疗挑战与新兴机遇. Cancers, 17(10):1703, 2025.
[3] Ryan Lusby, Engin Demirdizen, Mohammed Inayatullah, Paramita Kundu, Oscar Maiques, 张子怡, Mikkel G. Terp, Victoria Sanz-Moreno, Vijay K. Tiwari. 全癌转移驱动因素. Molecular Cancer, 24(1):2, 2025. https://doi.org/10.1186/s12943-024-02182-w
[4] 邢旭东, 钟健, Jana Biermann, 段浩, 张新宇, 石宇, 高一心 等. 单细胞分辨率下的全癌脑转移图谱. Cancer Cell, 2025. https://doi.org/10.1016/j.ccell.2025.03.025
[5] Ghulam H. Abbas, Edmon R. Khouri, Omar Thaher, Safwan Taha, Miljana Vladimirov, Rodolfo J. Oviedo, Jeremias Schmidt, Dirk Bausch, Sjaak Pouwels. 肿瘤学中转移预测建模：现有方法与未来方向. Annals of Medicine and Surgery, 87(6):3489–3508, 2025. DOI: 10.1097/MS9.0000000000003279
[6] Akter Rokaya, S. M. T. Islam, K. Mostafa. 提升手术精度：基于深度学习的微创手术深度估计（MiDaS 模型）. International Conference on Robot Intelligence Technology and Applications, pp. 46–57. Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-70687-5_5
[7] 杨素敏, Seo Jieun, Choi Jeonghyeon, Kim Sung-Hyun, Kuk Yunmin, Park Kyung C., Kang Mingon, Byun Sangwon, Joo Jae-Yeol. 理解癌症休眠：机制与技术的策略性结合. Molecular Cancer, 24(1):47, 2025. https://doi.org/10.1186/s12943-025-02250-9
[8] Leung Michelle M., Swanton Charles, McGranahan Nicholas. 整合模型系统与基因组洞察以解析癌症转移机制. Nature Reviews Genetics, 2025, pp. 1–12. https://doi.org/10.1038/s41576-025-00825-2
[9] Jee Justin, Fong Christopher, Pichotta Karl, Tran Thinh N., Luthra Anisha, Waters Michele, Fu Chenlian 等. 自动化真实世界数据整合提升癌症结局预测. Nature, 636(8043):728–736, 2024. https://doi.org/10.1038/s41586-024-08167-5
[10] Zohari Payam, Chehreghani Mostafa H. 多组学癌症研究中的图神经网络：结构化综述. arXiv preprint arXiv:2506.17234, 2025. https://doi.org/10.48550/arXiv.2506.17234
[11] 叶茂东, 任帅, 罗欢娟, 吴秀敏, 连宏伟, 蔡翔娜, 冀英昌. 图神经网络与转录组分析整合识别皮肤黑色素瘤免疫治疗反应和预后的关键通路及基因特征. BMC Cancer, 25(1):648, 2025. https://doi.org/10.1186/s12885-025-13611-4
[12] Hernández-Lemus Enrique, Ochoa Soledad. 癌症研究中多组学数据整合方法. Frontiers in Genetics, 15:1425456, 2024. https://doi.org/10.3389/fgene.2024.1425456
[13] 李炳军, Nabavi Sheida. 用于癌症分子亚型分类的多模态图神经网络框架. BMC Bioinformatics, 25(1):27, 2024. https://doi.org/10.1186/s12859-023-05622-4
[14] 李炳, 肖欣, 张超, 肖明, 张乐. DGHNN：用于全癌相关基因预测的深度图与超图神经网络. Bioinformatics, 2025:btaf379. https://doi.org/10.1093/bioinformatics/btaf379
[15] 邹雅竹, 赵子桐, 宋永梅. 多组学概述：在癌症分子分型与治疗中的强大工具. Malignancy Spectrum, 1(1):15–29, 2024. https://doi.org/10.1002/msp2.16
[16] 全冬梅, 吴虎强, 段立, 吕冠华, 高庆. 转移后肿瘤微环境的多组学分析：迈向个性化免疫治疗与分子靶向策略. Frontiers in Immunology, 16:1648987, 2025. https://doi.org/10.3389/fimmu.2025.1648987
[17] 景思宇, 王合齐, 林平, 袁娇, 唐志轩, 李红. 肿瘤微环境中生物学相关空间特征的量化与解析. NPJ Precision Oncology, 9(1):114, 2025. https://doi.org/10.1038/s41698-025-00857-1
[18] 江涛, 江海洋, 马欣怡, 徐明浩, 梁燕, 张文涛.

MetaGXplore：多组学数据与图卷积网络整合用于全癌患者转移识别. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 956–961, 2024. https://doi.org/10.1101/2024.06.30.601445
[19] 兰云铎, Shin Sung-Young, Nguyen Lan K. 从浅到深：癌症研究中机器学习与机制模型整合的发展. Current Opinion in Systems Biology, 40:100541, 2025. https://doi.org/10.1016/j.coisb.2025.100541
[20] Aguerralde-Martin Maider, Clemente-Císcar Mónica, Conesa Ana, Tarazona Sonia. 更可解释的多组学调控网络以表征表型. Briefings in Bioinformatics, 26(3):bbaf270, 2025. https://doi.org/10.1093/bib/bbaf270
[21] Akter Rokaya, M. A. S. A. R., Sudhanshu S., Singh A., Naizheng B. 生成式人工智能引导的哨兵：自优化联邦网络安全与智能威胁检测. Journal of Hunan University Natural Sciences, 52(12), 2025.

Manuscript Information
Word count: 7,058 words (excluding references).

Peer-Review Record
Fast-track status: Not fast-tracked.
First-round reviews received: 3 reports.
Revision cycles completed: 3 rounds.
Final version submitted: February 21, 2026