Pathological Tissue-level Contour Genomic Profile Interpretation of Lung Adenocarcinoma via Spatial and Morphological Features Co-action Graph Neural Network

Yu Yu
School of Biomedical Engineering
Guangzhou Medical University
Guangzhou, China
2022210074@stu.gzhmu.edu.cn

Wen Shi
School of Biomedical Engineering
Guangzhou Medical University
Guangzhou, China
shiwen@gzhmu.edu.cn

Guoxi Xie*
School of Biomedical Engineering
Guangzhou Medical University
Guangzhou, China
guoxixie@gzhmu.edu.cn

Jianing Xi*
School of Biomedical Engineering
Guangzhou Medical University
Guangzhou, China
xjn@gzhmu.edu.cn

2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385514

***Abstract*—The affections from genomics to morphology can prompt the genomic profile interpretation from inexpensive pathological image data rather than highly cost genomic sequencing data. Due to the extremely large size of Whole Slide Image (WSI), directly processing the complete Lung Adenocarcinoma (LUAD) WSI with traditional deep learning methods, will lead to memory overflow. In comparison to complete WSI, the smaller size patches can be easier processed by the existing deep learning based methods. Nevertheless, the split patches severely break the potential relationships between genomic abnormalities and morphological features, and the traditional deep learning methods may be difficult to capture the information of the broken relationship. Fortunately, a recent study has shown that the graph-structure representation can feasibly demonstrate the relationships among both local and remote regions. In consideration of the obstacles of both the break of remote area relationships and the lack of tissue-level contour for genomics-to-morphology associations, we propose Spatial and Morphological Features Co-action Graph Neural Network model (SMCGNN) to achieve the pathological tissue-level contour genomic profile interpretation of LUAD. Our SMCGNN achieves better performance on the genomic profile interpretation task than those of previous researches, yielding a relative performance increment of 9.3%. To the best of our knowledge, our SMCGNN is the first model to interpret the biological tissue-level contour of the genomic abnormality-related morphological regions. In summary, our method can provide pathologists with more fine-grained hints to molecular profile. The interpreted tissue-level regions can be accessed via the link: https://github.com/xianyvxxx/tissue-level-contour-regions-associated-with-genomic-profile.**

Keywords—genomics-to-morphology, WSI, tissue-level region

I. INTRODUCTION

Lung adenocarcinoma (LUAD) is one of the most common subtypes of lung cancer, representing about 40% of all lung cancers [1]. Similar to the causes of most non-small cell lung cancer tumors, LUAD is also driven by genomic changes in genome such as TP53 gene. Nevertheless, how to determine which gene are abnormal in tumor samples still costs a lot [2]. Fortunately, the changes of genome caused by genomic abnormality can further affect the morphological changes at the cellular level, and the information of cellular level can be easily obtained through pathological image. Consequently, the affections from genomics to morphology can prompt the genomic profile interpretation from inexpensive pathological image data rather than highly cost genomic sequencing data. The continuous accumulation of pathological image data offers us the opportunity of finding the specificity of genomics-to-morphology expression in pathological images.

It should be noted that, the relationships between cellular phenotype and genomic abnormality in tumors cannot be easily captured [3]. Therefore, a large number of studies have been proposed to find the connection between genomics and morphology of LUAD tumor. For the completeness of information, most recent pathological image studies adopt Whole Slide Image (WSI) as LUAD pathological data, but WSI also needs more powerful model instead of basic analysis. Thus, with the rapid development of Artificial Intelligence (AI), many studies have adopted deep learning technology as their genomics-to-morphology interpretation model. For example, the traditional deep learning model can capture the features of the WSI through convolutional methods to predict the genomic profile of the input sample [4, 5]. Therefore, many studies have proven the capability of mining genomic aberrations from morphological data of LUAD WSI.

However, as the most common challenge, the ultra large size of WSI images, always occupies tremendous amount of memory when the date are fed into neural network models. Due to the extremely large size of WSI, directly processing the complete LUAD WSI with traditional deep learning methods will lead to memory overflow. To circumvent the overflow, the entire WSI can be split into smaller regions, and many existing researches cut the complete WSI into box-shaped patches as input. In comparison to complete WSI the smaller size patches can be easier processed by the existing deep learning based methods. Nevertheless, the split patches severely break the potential relationships between genomic abnormalities and morphological features, and the traditional deep learning methods may be difficult to capture the information of the broken relationship.

In respect to the issue of the broken relationship, a recent study has found that the oversized LUAD WSI images actually contain redundancy [6]. Actually, the relationships among different areas in the WSI can help the analysis with only a subset of areas rather than the whole areas. Still, the most widely-used WSI methods are based on convolutional neural network, and the sliding kernel cannot represent the remote region relationships instead of local region relationship. Fortunately, a recent study has shown that the graph-structure

This work is supported partially by the Tertiary Education Scientific research project of Guangzhou Municipal Education Bureau (No. 202235388), partially by the Special Foundation in Department of Higher Education of Guangdong (Grant No.2022ZDX2053), partially by the Guangzhou Basic and Applied Basic Research Foundation (No. 2023A04J0386), partially by the Key scientific research project of universities in Guangdong Province (2023KCXTD026), partially by the Guangdong Basic and Applied Basic Research Foundation (No.2022A1515110001), partially by the National Natural Science Foundation of China (Grant Nos. 61901322, and 61974109), and partially by the Plan on enhancing scientific research in GMU.

representation can feasibly demonstrate the relationships among both local and remote regions [5, 7]. For example, a spatially aware graph neural network is proposed to illustrate the molecular profile of pathological image patches, where the graph-based structure can contain the associations of both local and remote patches. However, the recognized patches include multiple sorts of tissues, and the genomic abnormality does not usually associate with the phenotype of many sorts of tissues. For a certain patch of LUAD, the patch regions may include multiple sorts of tissues such as bronchi and lymphatic vessel. With only the box-shaped patches, the researchers still cannot locate the biological tissue-level contour of the genomic abnormality-related morphological regions.

In consideration of the obstacles of both the break of remote area relationships and the lack of tissue-level contour for genomics-to-morphology associations, we propose Spatial and Morphological Features Co-action Graph Neural Network model (SMCGNN) to achieve the pathological tissue-level contour genomic profile interpretation of LUAD. Our model firstly automatically select the key patches from the WSI to tackle the memory overflow issue, then capture the tissue-level contour via region subdivision module, and finally adopt the graph structure to represent both the local and remote associations among the tissue-level regions. When we apply our model on a public LUAD pathological image dataset, our SMCGNN achieves better performance on the genomic profile interpretation task than those of previous research, yielding a relative performance increment of 9.3%. To the best of our knowledge, our SMCGNN is the first model to interpret the biological tissue-level contour of the genomic abnormality-related morphological regions. Consequently, our method can provide pathologists with more fine-grained hints to molecular profile. The main contributions of our work are as follow:

- Establishing a morphological and spatial graph representation to avoid WSI memory overflow by the relationships among both local and remote areas.
- Achieving interpretation of pathological meaningful regions by providing pathologists with tissue-level fine-grained contours.
- Reaching a higher genomic profile prediction by the tissue-level regions when compared with those of the existing methods.

II. MATERIALS AND METHODS

A. Genomic Profile Recognition Datasets

To investigate the associations of genomics-to-morphology expression in pathological images, there are two necessary materials: WSI data and genomic profiles data. The WSI data are acquired from The Cancer Genome Atlas (TCGA) database [8]. TCGA-LUAD comprises 1608 whole-slide digital pathology images from 585 patients. For genomic profile we collect three typical molecular data for interpretation: gene mutations (MUT), copy number alteration (CNA), and protein expression (PEX), where the data are downloaded from cBioPortal database [9]. Through the TCGA barcodes, we can associate patient-specific WSI and

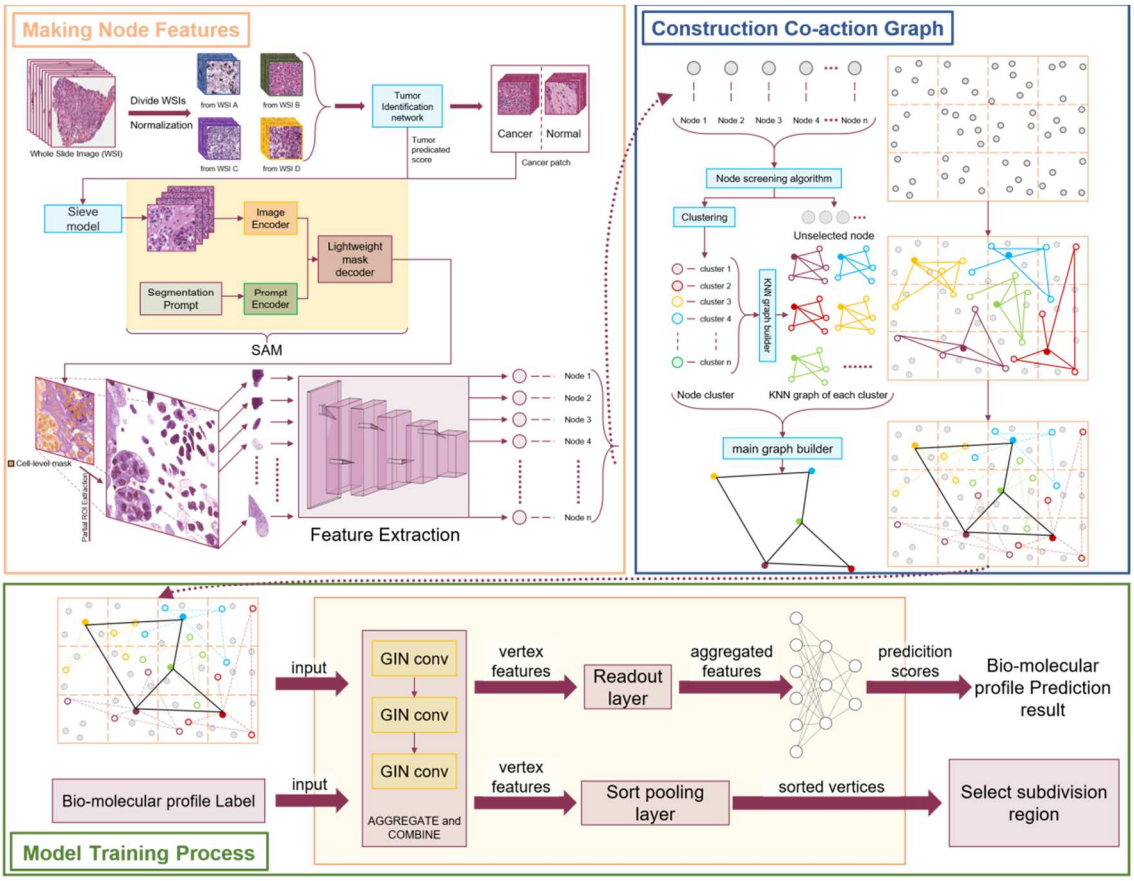

Fig. 1. Framework of Spatial and Morphological Features Co-action GNN. Making node representation part can segment sub-divided regions from WSI and extract node representations from them using a ResNet18 model with the classification layer removed. In the Construction Co-action Graph, we perform Node Select and Construction Cluster Subgraph before constructing the Main Graph. Finally, We input the constructed graph into GIN for training and select the pathological tissue-level contour region through the sort pooling layer.

TABLE I. CASE STATISTICS

Case Name	Data Source	Data Quantity Statistics
WSI data	TCGA: TCGA-LUAD	585 patients
gene mutations	cBioPortal: Lung Adenocarcinoma (TCGA, PanCancer Atlas)	566 patients
copy number variations	cBioPortal: Lung Adenocarcinoma (TCGA, PanCancer Atlas)	511 patients
protein expression	cBioPortal: Lung Adenocarcinoma (TCGA, PanCancer Atlas)	360 patients

genomic data, enabling the interpretation of LUAD genomics-to-morphology relationships (see Table 1 for data details).

To interpret genomic profile that is highly relevant to LUAD, We select typical genes for each profile type using different criteria. For MUT and CNA, we select the most frequently mutated genes [6]. For PEX, we categorize expression data into high-level and low-level expression groups based on the median [6], selecting proteins with the highest high-level expression rates. We focus on the top 10 profiles, such as TP53 and MUC16 (MUT), CDKN2B and C9orf53 (CNA), and C-Raf_pS338 and p53 (PEX) (see Table 2). By focusing on these profiles for WSI interpretation, we concentrate on regions with genomic abnormalities [10-12].

B. Contour Fineness Enhancing via Region Subdivision Module

Patch Division: In order to prevent memory overflow issues that can arise when directly using oversize LUAD WSI images, we cut the complete WSI at the highest resolution into patches of size 512*512, and the WSI contains many empty regions that consume a large amount of memory and do not contribute to the tissue information represente by the WSI. To isolate the parts that contain tissue information from all the patches, we use the maximum between-class variance method to capture the foreground of the patches [13]. Then, to improve the accuracy and robustness of the genomic profile interpretation model, we adopt color normalization on all the patches to eliminate the effects of different staining methods. For avoiding the memory overflow issues, we can reduce the amount of data that are required to be processed simultaneously by dividing WSI into patches.

Patch Selection: Among all the patches cut from WSI, there are only a portion of patches are meaningful for genomic profile interpretation. Therefore, to avoid wasting memory resources from calculating meaningless patches, we first select patches from the Whole Slide Image (WSI) that are relevant to Lung Adenocarcinoma (LUAD) using a fine-tuned ResNet18 model. The trained module shows an accuracy of over 99% on the external dataset and can identify LUAD patches effectively. Afterwards, as for ambiguous patches, We further refine our selection by choosing the top 20% of cancer patches based on the prediction scores of the LUAD patch classification module, effectively reducing redundancy.

Regional Subdivision: Although the LUAD related patches can be efficiently selected, the box-shaped patches still contain more than one tissue types, and genomic abnormalities do not usually correlate with multiple tissue phenotypes simultaneously. Therefore, in order to locate the biological tissue-level contour from box-shaped patches, we also perform regional subdivision on the selected key patches. Inspired by the current state-of-the-art AI progress, we note that the Segment Anything Model (SAM) shows strong image segmentation capabilities [14]. SAM can accurately segment

TABLE II. TOP 10 GENE MUTATION AND COPY NUMBER ALTERATION GENE ALTERATION PERCENTAGE

CNA	alteration percentage	MUT	alteration percentage	PEX	alteration percentage
CDKN2B	16.4%	TP53	52.1%	C-Raf_pS338	17.8%
CDKN2B-AS1	16.4%	TTN	48.1%	RAF1	17.8%
C9orf53	15.7%	MUC16	42.8%	Rad51	17.8%
CLPTM1L	13.3%	CSMD3	39.9%	P53	17.7%
SFTA3	13.3%	RYR2	38.3%	14-3-3_epsilon	17.7%
MIR4457	13.1%	LRP1B	35.5%	Collagen_VI	17.7%
MBIP	13.1%	ZFHX4	32.7%	MAP2K1	17.7%
NKX2-1-AS1	13.1%	USH2A	31.3%	MEK1_pS217_S221	17.7%
NKX2-1	13.1%	KRAS	29.7%	RPS6KB1	17.7%

multiple regions from the input image, which meets our requirement of regional subdivision. The process involves three steps. First, Prompts are provided for each patch to serve as the seeds of segment regions. To ensure that the segmentation regions do not contain more than one type of tissue, cue points covering the entire patch are generated based on the size of the LUAD cell and the pixel spacing for 20x magnification of the Whole Slide Image (WSI). Second, the key patches and their related prompts are encoded into embedding vectors using an image encoder and a prompt encoder. This helps to establish a connection between the key patches and the prompts. Final, the mask decoder is used to generate multiple tissue-level contour region masks for key patches, mapping the embeddings between key patches and prompts to the segment regions. Through the steps above, we can extract more fine-grained tissue-level contour regions from the key patches.

C. Memory Reducing via Region Association Graph Modeling

Local/Remote Region Association: It should be noted that there are both local and remote relationships among the WSI patches. Still, most image storage methods preserve only the local relationships, ignoring the remote ones. Fortunately, according to recent studies [7, 15], the graph structure shows the capability of representing both local and remote association. In a constructed WSI graph structure, we can consider the patches as WSI graph nodes, and the relationships between the patches as WSI graph edges. Since it's challenging to compute the patches directly using the image format, we extract features from the patches to facilitate computation. These features also serve as representations of the relationships among the patches in the WSI graph. For details on feature extraction methods, please see the next subsection Region graph construction. For more details of the methods used to calculate relationships between nodes, please see the below subsection Tissue Hub Selection, Subgraph Clustering, and Tissue-specific Representation.

Tissue-level Contour Region: Segmentation of the tissue-level contour region is difficult due to the complexity of the tissue morphology in the patch. Here, we need to obtain more accurate tissue regions, and to avoid obtaining subdivision regions containing incomplete tissues or multiple tissues at the same time. Therefore, we generate whole, part and subpart three-level masks around the input prompts with a mask decoder in SAM, and rank all masks according to their prediction confidence scores. By this procedure, we can efficiently determines the contours of most tissues and accurately obtain the tissue-level contour regions [14].

Region graph construction: To feasibly calculate the associations between nodes and construct the edges in graph, we should perform feature extraction on the tissue-level

contour regions to obtain a tissue-level representation of the node representation. Here we use pretrained ResNet18 module to extract features from tissue-level contour regions for constructing the WSI graph. After being processed by ResNet18, the tissue-level contour regions can be effectively extracted as tensor-form representations, which are feasible to represent the local and remote region associations by the constructed edges. To find the relationship between the tissue-level contour regions in the patch, we calculate the distance between the tissue-level contour regions by using tensor-form representations. Following the distances between regions, we connect the regions in the patch to each other as a region graph.

Tissue Hub Selection: In a single patch, the number of the tissue-level contour region nodes may not equal the number of tissue types. So, multiple nodes may represent the same tissue in this case. To remove redundant nodes, we introduce the tissue hubs to represent the tissue types, and use the tissue hubs to associate the tissue types in each patch. Specifically, we use a tissue hub selection method based on graph clustering scheme based on K-nearest neighbor. To identify representative tissue-level contour regions in the patch within the feature space, we devised a clustering method based on the region graph from the previous step. This region graph aggregates the features of all tissue-level contour regions in the represented patch and expresses them as nodes. The connections between nodes represent associations between different tissue-level contour areas. Therefore, during clustering, the top five nodes with the highest degrees in the region graph are selected as hubs for the tissue types. To ensure these selected nodes cover sufficient information on the genomic profile under investigation, we also include five randomly selected nodes to supplement the top five nodes with the highest degrees. This selection strategy allows the node number to approximate the tissue type number, thereby eliminating redundant information.

Tissue Subgraph Clusterring: The spatial and morphological features co-action contains two parts. Here we extract the morphological features of the tissue-level contour regions. After we obtain the region graphs above, we perform clustering of the morphological features of each node, classifying nodes with similar features into the same cluster. Then, we construct tissue subgraphs from these clusters, connecting different tissue-level contour regions through morphological association. After that, to facilitate the subsequent computation of the relationship between the tissue subgraphs, we find the five nodes with the highest degree in the tissue subgraph to represent the subgraph. The construction of cluster tissue subgraphs connects different tissue-level contour regions through morphological association, and better connects local and remote areas. Through the above steps, we have implement the morphological features action part of the co-action.

Tissue-specific Representation: Subsequently, for the second action in the co-action, we extract spatial features of tissue-level contour regions for genomic profile pre-interpretation. We build a WSI graph based on the tissue graph to explore the relationship between tissue-level contour regions. The phenotypic profile of a genomic profile on WSI shows a specific spatial distribution. To find regions closely related to the identified genomic profile, we use a graph connection correlated with spatial distribution to link representative nodes in each cluster. We use the patch position on the whole WSI as the basis for connecting edges, considering an edge between two nodes if their distance is less than a fixed threshold (set to 85). The specific calculation method is shown in the following formula:

$$\text{distance} = \sqrt[2]{(n_1^x + n_2^x)^2 + (n_1^y + n_2^y)^2} \tag{1}$$

$$\text{WSI Graph Edge} = \begin{cases} 0, \text{distance} \geq \text{threshold} \\ 1, \text{distance} \leq \text{threshold} \end{cases} \tag{2}$$

Where n_1^x and n_2^x are the upper left corner of WSI is the lateral distance of the origin, n_1^y and n_2^y are fore-and-aft distance. Thus far, we successfully achieve spatial and morphological features co-action through WSI graph.

D. Genomic Profile Prediction Model Training via Graph Isomorphism Networks

Tissue-Association Graph Learning: After representing the relationships between the tissue-level contour regions using a graph structure, we found that traditional neural networks have difficulty processing graph-structured data. Therefore, we employ Graph Isomorphism Networks (GIN) [16] capture all data and labels with better aggregators, such as sum aggregators, and learn more accurate structural information. To achieve interpreting the genomic profile on WSI, firstly we need to update all nodes on the WSI graph in order to capture the structural information and improve the specificity of the WSI graph. The node update method of GIN during training is shown below:

$$h_v^{(k)} = MLP^{(k)}((1 + \epsilon^{(k)}) * h_v^{(k-1)} + \sum_{u \in N(v)} h_u^{(k-1)}) \tag{3}$$

Subsequently, we need to aggregate the information of all nodes to the graph level in order to obtain the overall structure information of the WSI graph. Therefore, we use GlobalAddPooling [16] to aggregate the information of all the nodes into the holistic representation of WSI graph. The holistic representation of WSI graph can be written as:

$$h_G = GlobalAddPooling(\{h_v\}_{v \in G}) \tag{4}$$

where G denotes the input graph, h_v denotes the feature vector of node v, and h_G denotes the holistic representation of WSI graph.

Hyperparameter setting: In order to fairly compare the *models*, we use the same hyperparameters across all models. For all models, we use the Adam optimizer, set the batch size to 32, and set the initial learning rate to 1e-3. We also perform five-fold cross-validation on the TCGA-LUAD dataset to evaluate the generalization ability of the models.

E. Fine Region Prioritizing via Global Sort Processing

After obtaining a trained SMCGNN model, we will explain how to use the model to find regions that are highly correlated with genes. After the WSI graph is input into the SMCGNN model, the node information needs to be aggregated into the holistic representation of WSI graph. Since the holistic representation of WSI graph is difficult to separate out the information on each node, we directly perform the global sort processing after the node updating.

Each kind of data has its own sorting method and in graph structural data, we can sort nodes by their structural roles in

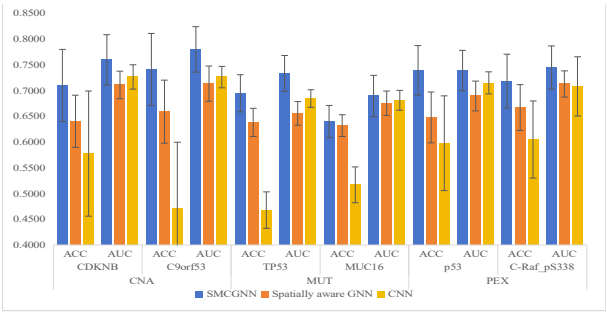

Fig. 2. Comparison of our SMCGNN with traditional deep learning as well as spatially aware GNN on the TCGA-LUAD dataset. The highest point of the histogram is the score for the current metric.

the graph [17]. Here, since the output of GIN after node update is Weisfeiler-Lehman (WL) colors [17] [18], and WL colors can well describe the degree of node's contribution in the graph. Therefore, we use the WL colors of the nodes as structural roles to find the top n points that contribute the most to the WSI graph.

III. RESULT

A. Performance evaluation

To identify regions of genomic abnormality, we assessed the predictive accuracy of each region and selected the most relevant regions to the genomic atlas. We use accuracy (ACC) to measure the fraction of correctly predicted genomic profiles in Whole Slide Images (WSIs), area under the receiver operating characteristic curve (AUC) to assess the prediction ability of each region's genomic abnormality with a larger AUC score indicating better performance, and 95% Confidence Intervals (CIs) calculated by bootstrapping 1,000 times to estimate the uncertainty of the assessed metrics.

B. Competing methods

To evaluate the results of the SMCGNN model genomic profile prediction, we compare the SMCGNN model with traditional deep learning and spatial aware graph neural network through the TCGA-LUAD dataset. In the experiments, we adopt our model to predict the WSI morphological profiles for the LUAD related gene TP53 and MUC16 as the MUT abnormalities, CDKNB and C9orf53 as the CNA abnormalities, and p53 and C-Raf_pS338 as PEX abnormalities.

To compare the prediction capabilities of our propose SMCGNN model and traditional neural network on the genomic profile task, we use the ResNet18 model as a baseline method for genomic profile prediction. ResNet18 model predicts genomic profiles on all patches in the WSI, and the prediction scores of all patches are weighted to determine whether the genomic profiles are expressed on the WSI. Our SMCGNN model showed a significant performance advantage over traditional neural network predictions, with relative increases of 31.2% for average ACC and 4.8% for average AUC (see Figure 2 for details).

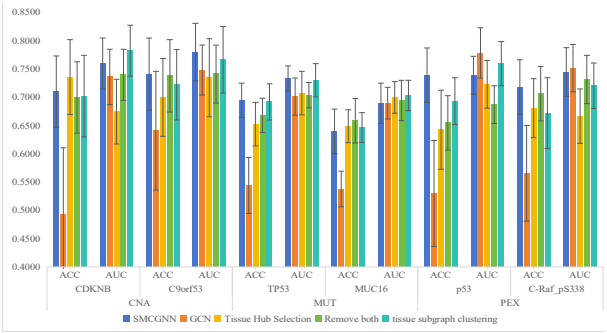

Fig. 3. Comparison of model performance after ablation for each part of the SMCGNN. The highest point of the histogram is the score for the current metric.

We also conducted comparison experiments with graph neural networks to evaluate the impact of using tissue-level contour regions to construct graphs on genome profile interpretation. We chose the spatial aware [6] graph neural network for these experiments, ensuring the same graph data structure. These networks use patches as the minimum scale for constructing graphs, making the information represented by the graph susceptible to other genomic profile phenotypes. Our SMCGNN model outperformed spatially aware graphical neural networks in both average ACC and average AUC metrics, with relative increments of 9.3% and 7.1% for ACC and AUC, respectively (Figure 2). In summary, our proposed SMCGNN model performed better than previous studies in predicting the genomic profile of LUAD.

C. Ablation study

In order to evaluate the contribution of each component in the graph construction process to genome profile prediction, we train the model with the tissue hub selection component removed, the model with the tissue subgraph clustering component removed, and the model with the tissue hub selection and tissue subgraph clustering components removed, respectively, in order to observe the changes in the average ACC and AUC of the ablated model relative to our SMCGNN model in the genomic profile prediction task.

To investigate the representativeness of the hubs for multiple tissue types in the patch, we substituted the tissue hub selection component with random sampling for ablation. Without tissue hub selection, the randomly selected nodes in each patch's tissue-level contour regions could not adequately represent all the tissues contained in the patch. Compared to the ablated model, our SMCGNN model showed relative increments of 4.4% and 5.7% in average ACC and AUC, respectively (Figure 3). In our proposed SMCGNN model, the tissue subgraph clustering component associates tissue-level contour regions with similar morphological features through clustering. To investigate the impact of morphological relationships between each tissue-level contour region on genome profile interpretation, we replaced the tissue subgraph clustering component with random sampling. Compared to the ablated model, the average AUC of the models did not change significantly, but the average ACC improved by 2.8% relative to the post-ablation model (Figure 3).

To explore the interaction between the tissue hub selection and tissue subgraph clustering components, we conducted an additional round of ablation, removing both components previously subjected to ablation. Specifically, we ablated both the tissue hub selection and tissue subgraph clustering components using random sampling. We found that using these two components individually did not always enhance the model's predictive performance. However, their combined use significantly improved the model's performance, suggesting a positive association between the two components in genomic profile prediction. Compared to the model after both components were ablated, our proposed SMCGNN model showed increases of 2.8% and 3.5% in average ACC and AUC, respectively (Figure 3).

After representing the WSI as a graph composed of tissue-level contour region nodes, we needed a more powerful graph learning method to capture the graph's structural information. We used two graph learning methods, GIN and Graph Convolutional Network (GCN), for genomic profile interpretation. Our SMCGNN model, which uses GIN, benefits from the introduction of the unary setup concept in GIN, demonstrating better graph recognition ability. compared to the model using GCN, our SMCGNN model showed advantages in genomic profile interpretation, with improvements of 28.1% and 1.5% in average ACC and AUC, respectively (Figure 3).

D. Tissue-level Region Case Study

By using the tissue-level contour region ranking method described above, we extract the top 20 regions that contribute most to the interpretation of genomic profiles from the correctly predicted WSIs. Consequently, we create a public database containing genomic profiles found from multiple tissue-level contour regions from the WSIs. Our database can be accessed via the link: https://github.com/xianyvxxx/tissue-level-contour-regions-associated-with-genomic-profile. Here Some typical tissue-level regions are shown in Figure 4. We hope that this database will support for genomics-to-morphology research, and help pathologists to identify the phenotypic profiles of genomic abnormality more accurately.

IV. CONCLUSION AND DISCUSSIONS

In this paper, we propose a spatial and morphological feature synergy graph neural network (SMCGNN) to solve the memory overflow problem when processing WSI and successfully obtain pathological tissue-level contour phenotype regions for genomic atlases.The SMCGNN extracts the pathological tissue-level contour regions from WSI and constructs synergy graphs by correlating pathological morphology and spatial features of each region. The genome mapping prediction model is then trained on the synergy map using GIN. SMCGNN outperforms previous models in multiple genome mapping tasks. Due to SMCGNN's small memory requirement and the need for only WSI images as input, this model can provide pathologists with a tissue-level contour of the target genomic profile during digital pathology diagnostics. This can be seamlessly integrated into existing diagnostic processes and potentially open up a market for low-cost genomic examinations. However, there are still some limitations to the model, including insufficient data problem and region label range problem. Several studies illustrate that transfer learning can be effective in solving the problem of insufficient data, which is expected to be used in graph neural networks. Overall, we present an innovative approach to genomic profile interpretation using spatial and morphological features co-action graph neural network. Our SMCGNN shows a promising genomics-to-morphology tissue-level regions interpertation when facing the challenges of memory overflow and remote spatial association.

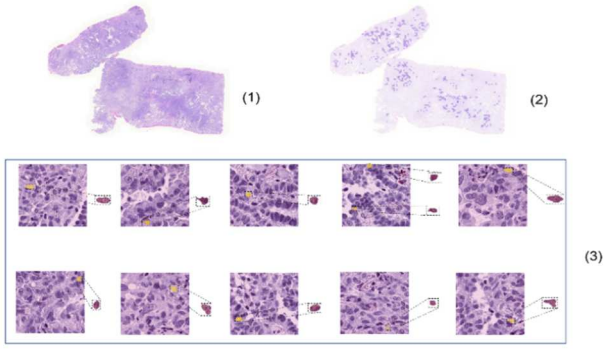

Fig. 4. TP53 mutation prediction the tissue-level contour region (1. the original WSI; 2. the extent of the node region capture by the model detection graph; 3. A portion of the selected tissue-level contour regions in this WSI)

REFERENCESA

[1] D. J. Myers and J. M. Wallen, "Lung adenocarcinoma," in *StatPearls [Internet]*, ed: StatPearls Publishing, 2022.

[2] M. Timilsina, H. Yang, R. Sahay, and D. Rebholz-Schuhmann, "Predicting links between tumor samples and genes using 2-Layered graph based diffusion approach," *BMC bioinformatics,* vol. 20, pp. 1-20, 2019.

[3] M. Fletcher, "The complex relationship between cell transcriptomic state and phenotype," *Nature Genetics,* vol. 55, pp. 1421-1421, 2023.

[4] H. Qu, M. Zhou, Z. Yan, H. Wang, V. K. Rustgi, S. Zhang*, et al.*, "Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning," *NPJ precision oncology,* vol. 5, p. 87, 2021.

[5] L. Chen, L. Yu, and L. Gao, "Potent antibiotic design via guided search from antibacterial activity evaluations," *Bioinformatics,* vol. 39, p. btad059, 2023.

[6] K. Ding, M. Zhou, H. Wang, S. Zhang, and D. N. Metaxas, "Spatially aware graph neural networks and cross-level molecular profile prediction in colon cancer histopathology: a retrospective multi-cohort study," *The Lancet Digital Health,* vol. 4, pp. e787-e795, 2022.

[7] J. Xi, L. Ye, Q. Huang, and X. Li, "Tolerating data missing in breast cancer diagnosis from clinical ultrasound reports via knowledge graph inference," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3756-3764.

[8] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe*, et al.*, "Toward a shared vision for cancer genomic data," *New England Journal of Medicine,* vol. 375, pp. 1109-1112, 2016.

[9] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy*, et al.*, "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer discovery,* vol. 2, pp. 401-404, 2012.

[10] J. Xi, X. Yuan, M. Wang, A. Li, X. Li, and Q. Huang, "Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication," *Bioinformatics,* vol. 36, pp. 1855-1863, 2020.

[11] J. Xi, A. Li, and M. Wang, "HetRCNA: a novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 17, pp. 422-434, 2018.

[12] C. Ao, X. Ye, T. Sakurai, Q. Zou, and L. Yu, "m5U-SVM: identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation," *BMC biology,* vol. 21, p. 93, 2023.

[13] N. Ostu, "A threshold selection method from gray-level histograms," *IEEE Trans SMC,* vol. 9, p. 62, 1979.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson*, et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643,* 2023.

[15] J. Xi, Z. Miao, L. Liu, X. Yang, W. Zhang, Q. Huang*, et al.*, "Knowledge tensor embedding framework with association enhancement for breast ultrasound diagnosis of limited labeled samples," *Neurocomputing,* vol. 468, pp. 60-70, 2022.

[16] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," *arXiv preprint arXiv:1810.00826,* 2018.

[17] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[18] A. Leman and B. Weisfeiler, "A reduction of a graph to a canonical form and an algebra arising during this reduction," *Nauchno-Technicheskaya Informatsiya,* vol. 2, pp. 12-16, 1968.