Genomics-guided Representation Learning for Pathologic Pan-cancer Tumor Microenvironment Subtype Prediction

Fangliangzi Meng[1,2], Hongrun Zhang[3], Ruodan Yan[4], Guohui Chuai[1,2✉], Chao Li[4,5✉], and Qi Liu[1,2✉]

[1] Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration (Tongji University), Ministry of Education, Orthopaedic Department of Tongji Hospital, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai, China
[2] National Key Laboratory of Autonomous Intelligent Unmanned Systems, Frontiers Science Center for Intelligent Autonomous Systems, Ministry of Education, Shanghai Research Institute for Intelligent Autonomous Systems, China
[3] Cancer Research UK Cambridge Institute, UK
[4] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
[5] School of Science and Engine, School of Medicien, University of Dundee, UK
qiliu@tongji.edu.cn

Abstract. The characterization of Tumor MicroEnvironment (TME) is challenging due to its complexity and heterogeneity. Relatively consistent TME characteristics embedded within highly specific tissue features, render them difficult to predict. The capability to accurately classify TME subtypes is of critical significance for clinical tumor diagnosis and precision medicine. Based on the observation that tumors with different origins share similar microenvironment patterns, we propose PathoTME, a genomics-guided Siamese representation learning framework employing Whole Slide Image (WSI) for pan-cancer TME subtypes prediction. Specifically, we utilize Siamese network to leverage genomic information as a regularization factor to assist WSI embeddings learning during the training phase. Additionally, we employ Domain Adversarial Neural Network (DANN) to mitigate the impact of tissue type variations. To eliminate domain bias, a dynamic WSI prompt is designed to further unleash the model's capabilities. Our model achieves better performance than other state-of-the-art methods across 23 cancer types on TCGA dataset. Our code is available at https://github.com/Mengflz/PathoTME.

Keywords: Digital Pathology · Domain Adversarial Training · Oncology · Tumor Microenvironment Genomics

1 Introduction

The tumor microenvironment refers to the interactions between tumors and surrounding cells [31], constituting a complex system that plays a crucial role

in patient stratification, tumor prognosis, and therapy response prediction. Besides the binary high/low-level immune infiltration subtypes, Bagaev et al. [2] takes cancer-associated fibroblasts (CAFs) into criterion and proposes four-class TME subtypes: (1) immune-enriched, fibrotic (IE/F); (2) immune-enriched, non-fibrotic (IE); (3) fibrotic (F); and (4) immune-depleted (D). Tumors from the same subtype share similar functional patterns and molecular mechanisms. However, the RNA sequencing data, as a key determinant for distinguishing subtypes, are not included in standard clinical treatment and are hard to acquire. In this case, WSI as a routine clinical diagnostic tool, can shed light on the rapid diagnosis of TME subtypes. However, compared to the straightforward task of predicting tumor molecular subtypes, the features of TME in WSI are relatively disapparent and diverse, not solely based on the quantity or density of immune cells, but also the spatial interaction of various immune and tumor cells. These subtle features, easily overlooked and intermixed with tumor tissue characteristics in WSI, present a substantial challenge.

Deep learning-based WSI analysis has proven promising performance in various clinical applications, including tumor grade diagnosis [29], tumor molecular subtypes classification [4,18,28,30], survival prediction [7,10,11,23,24], gene expression estimation [25], and high/low immune types prediction [14]. However, previous studies are mainly focused on specific tumor types, or classification tasks based on a single biomarker, which result in an inability to overcome tissue specificity and limited ability to capture the comprehensive landscape of tumor patterns. Meanwhile, genomics data reveals high relevance to tumor mechanisms. To contain this, there are some works focusing on the fusion of WSI and genomics [11,24,27,33], and introducing genetic interpretability to WSI images [1]. Fusion methods need both genomics and WSI as inputs in practical operation, which are not always feasible due to the expensive acquisition of genetic information, and therefore are hard to widespread use. Also they cannot be applied to our TME classification task directly. Not only for the reasons previously mentioned but also because genes contain highly related information to TME subtypes, therefore the introduction of them in the evaluation stage is a kind of information leakage. Additionally, some deep learning methods for gene imputation have emerged [15, 25, 32], focusing on gene prediction and spatial expression patterns. However, these methods currently show low correlation in the prediction of individual genes [13]. Since the description of TME requires thousands of genes, the inaccuracy in each one would accumulate, leading to inaccurate TME classification. In this paper, we present the first attempt for pathologic pan-cancer microenvironment classification. Our contributions are as follows:

(1) We propose a gene-guided model (PathoTME) addressing TME subtypes classification based on WSI, which makes the first step to unveil the comprehensive tumor microenvironment landscape among different origins on pan-cancer level.

(2) We introduce Visual Prompt Tuning(VPT) combined with a pretrained feature extractor to eliminate domain bias and further unleash the model's capa-

bilities. We leverage genetic information in the training stage while not requiring genetic information as input during inference. It is noted that gene expression acts as anchors only used in the training phase.

(3) In order to reduce the impact of tissue heterogeneity on performance and variability in staining color, DANN is used to ensure the feasibility of model on pan-cancer datasets and obtain the tissue-conservative patterns. Moreover, cancer types with fewer slides can learn features from those with abundant samples. Finally, PathoTME achieves better performance than other state-of-the-art methods.
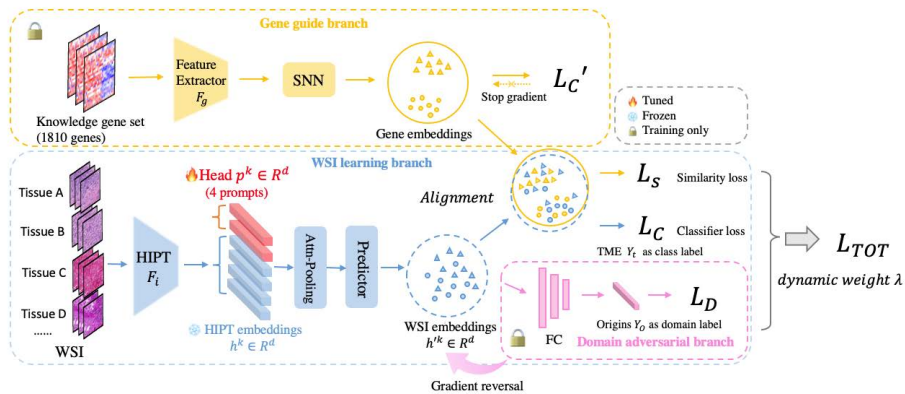

Fig. 1. Architecture of PathoTME.

2   Method

As shown in Fig. 1, our model is composed of three branches, the gene guide branch (yellow), the domain adversarial branch (pink) and the main WSI learning branch (blue). In practical application, only the main WSI branch is active, with the other two branches existing as pre-trained weights. In gene guide branch, genes from a curated knowledge gene sets are collected from previously published works [3,9]. Then we use a single-layer fully connected (FC) with Self-Normalizing Network (SNN) [21] to extract gene embeddings from tabular data. Meanwhile, Multi-instance learning (MIL) is applied as a widely-used method on WSI to aggregate patch embeddings as instances into slide embeddings, on top of that, the attention mechanism can focus on key areas highly related to specific targets. We employ ABMIL [16] to integrate the extracted embeddings with prompts.

At the training stage, gene branch will guide WSI projector to learn discriminative WSI embeddings which close to the paired gene embeddings in the representation space. At the same time, domain adversarial branch will prevent

WSI learning branch from attaining features related to tissue structures irrelevant to TME subtypes. At the inference stage, our network only needs WSI as input, and output the prediction result of TME subtypes.

2.1   Visual prompt tuning

To further fine-tune the feature extractor, we use pretrained 4K HIPT [5] model to extract WSI features, which is frozen during training. Inspired by Visual Prompt Tuning (VPT) [17] strategy, we add four learnable prompts to HIPT embeddings, enabling fine-tuning for our specific task without the extensive retraining or fine-tuning of the entire pretrained model. Embeddings with prompts are input into ABMIL [16] for subsequent training together. It is formulated as:

Let $H = \left\{ h^k \in \mathbb{R}^d \mid k \in \mathbb{N}, 1 \leq k \leq n \right\}$ be the bag of n region embeddings from HIPT; $P = \left\{ p^k \in \mathbb{R}^d \mid k \in \mathbb{N}, 1 \leq k \leq n_p \right\}$ is a collection of n_p learnable prompts. Then the overall embeddings of WSI can be denoted as:

$$H' = [H, P] = \left\{ h'^k \in \mathbb{R}^d \mid k \in \mathbb{N}, 1 \leq k \leq n + n_p \right\}, \tag{1}$$

The ABMIL processing can be modified as:

$$z' = \sum_{k=1}^{K} a_k h'_k, \, a_k = \frac{\exp\left(\mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}'_k)\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}'_j)\right)}. \tag{2}$$

in which $K = n + n_p$, a_k is the attention score.

2.2   Siamese representation similarity

Vanilla MIL network sometimes fails to get optimal performance because of the limited amount of slides. Without efficient learning, it is easy to overlook detailed histopathological features, whereas gene expression includes the majority of information regarding subtle intracellular and extracellular compositions. To generate more effective WSI embedding space, we incorporate genetic information as a regularization factor to guide the learning process of our network by employing a Siamese network architecture [8].

Differing from SimSiam [8], we use WSI x_i and Gene expression x_g as two types of inputs. The whole slide image x_i passes through the image encoder f_i and a MLP head h, whereas the gene expression data g passes through the feature extractor f_g and SNN network to generate the following two embeddings $p_x = h(f_i(x_i))$, $z_g = f_g(x_g)$. We then minimize the negative cosine similarity between the whole slide image and gene expression data embeddings p_x and z_g as follows:

$$D_s(p_x, z_g) = -\frac{p_x \cdot z_g}{\|p_x\|_2 \cdot \|z_g\|_2}, \tag{3}$$

Noteworthily, during the pretraining phase, the gene extractor and SNN network are trained with TME subtypes labeled y_t. Once trained, its weights will be frozen and it will stop gradient descent. We implement it as:

$$L_S = D_s(p_x, stopgrad(z_g)), \tag{4}$$

2.3   Domain adversarial neural network

The similarity loss L_S will bring WSI embeddings closer to paired gene embeddings in the representation space. However, gene expression also has strong tissue-specific characteristics which cannot be avoided by selecting a knowledge gene set. Therefore, WSI embeddings passed through Siamese network contain tissue-specific features and need to be removed while retaining the TME-related features.

Domain Adversarial Neural Networks (DANN) is a representation learning method originally dedicated to domain adaptation [12]. Here, we apply DANN to (1) produce discriminative embeddings for the source domains, which are TME subtypes $Y_t \in \mathbb{R}^4$ in our task, while ensuring that these embeddings are not biased by the differences between domains which are origin types $Y_o \in \mathbb{R}^{23}$, enabling effective domain transfer, (2) remove the bias from stained technology, (3) assist cancer types with fewer samples in efficiently learning features from those with a larger amount of samples.

DANN is composed by three parts: feature extractor $G_f(x;\theta_f)$(in our network is F_c), domain classifier $G_d(x;\theta_d)$ and label predictor $G_y(x;\theta_y)$. Different from regular neural networks, domain classifier G_d has a gradient reverse layer (GRL)R_x, which is a regularizer that is weighted by hyper-parameter λ_p.

$$\lambda_p = \frac{2}{1 + \exp(-\gamma \cdot p)} - 1 \tag{5}$$

γ is a hyperparameter of scheduler, and set 10 in our all experiments, while p is the training progress linearly changing from 0 to 1, i.e., $p = \frac{epoch}{max\ epochs}$.

In total, the loss function of DANN can be written as

$$
\begin{aligned}
L_D(\theta_f, \theta_y, \theta_d) = &\sum_{i=1}^{N} \sum_{d_i=0} L_y\left(G_y\left(G_f\left(x_i;\theta_f\right);\theta_y\right), y_i\right) \\
&+ \sum_{i=1}^{N} L_d\left(G_d\left(R_x\left(G_f\left(x_i;\theta_f\right)\right);\theta_d\right), y_i\right)
\end{aligned}
\tag{6}
$$

Building upon this and incorporating previous negative cosine similarity, we derive our total lossL_{TOT}

$$L_{TOT} = (1 - \lambda_p)L_S + \lambda_p L_D \tag{7}$$

3   Experiments & Results

3.1   Datasets

We use paired diagnostic whole slide images and RNA sequencing data(FPKM-UQ) from 23 TCGA datasets. After removing unpaired cases, 7103 paired samples from 6519 cases remained. These paired samples are randomly stratified

Table 1. Performance of classifying TME subtypes on TCGA Pan-cancer dataset.

Method		TCGA Pan-cancer			
		ROCAUC	PRAUC	Acc	F1 score
Classification (%)	kNN	60.3±0.63	33.0± 0.64	38.4±1.11	34.3±0.99
	ABMIL	63.5±0.69	36.5±0.51	42.1±1.15	35.7±1.29
	ABMIL-gated	62.9±0.99	36.3±0.98	41.6±1.06	35.2±1.45
	TransMIL	63.0±1.14	36.5±1.42	41.5±1.70	31.3±2.94
	DSMIL	61.6±0.85	35.0±1.23	42.5±1.17	32.4±1.37
	CLAM	63.9±0.78	36.9±0.91	42.5±1.62	35.3±1.14
	SupCont	64.4±0.68	37.8±0.82	43.6±0.61	34.1±1.57
	PathoTME(1 stage)	**69.5±0.24**	**42.8±0.41**	**46.6±1.03**	**39.7±0.78**
	PathoTME(2 stages)	67.9±0.36	41.4±0.36	46.2±0.97	**39.7±1.22**
	SNN	94.5±0.28	85.1±0.96	80.0±0.48	77.0±0.33
Ablation study (%)	+Siamese	66.1±0.46	38.8±0.61	45.2±0.93	35.7±1.17
	+DANN	67.3±0.49	41.3±0.87	44.6±1.44	39.2±0.68
	+Siamese+DANN	68.7±0.49	42.1±0.87	46.3±0.66	39.6±0.44
	+Visual prompts	65.8±2.44	38.5±0.64	43.8±1.14	34.1±0.82

into training sets(85%) and test sets(15%) based on the origins. For detailed dataset division and usage, please refer to Fig.S1. The training set is validated using five-fold cross-validation with a fixed seed. Considering potential issues with data imbalance, we include multiple metrics ROCAUC, PRAUC, ACC and F1 score to evaluate the performance of models.

3.2   Implementation details

To accelerate training, we use methods from HIPT [5,6] with pretrained HIPT_4K model to extract WSI region features. The tile size is 4096 pixels and the magnification is 20x. The weights of this network will not be involved in the subsequent learning processes. TME subtypes of each sample is calculated according to methods from [2]. Among all the samples, there are 3133 D and 1912 IE, 1718 F and 1261 IE/F subtypes.

The proposed PathoTME is trained on the training set for most 100 epochs with early stopping strategy. An Adam [20] optimizer is adopted for the training processes of all models with a learning rate of 5e-5 and a weight decay of 1e-5. The dynamics weight λ is set as the previous formula. All experiments are conducted on a server with four NVIDIA GeForce RTX 4090 GPUs. Additionally, our method is implemented on PyTorch with the Python environment.

3.3   Performance evaluation

Pan-cancer performance As shown in Table 1, our model indicates improved ROCAUC, PRAUC, accuracy and F1 score on TME subtypes classification compared with other state-of-the-art MIL methods: ABMIL [16], TransMIL [26],

Table 2. Performance of classifying TME subtypes on each TCGA dataset.

Cancer Type	ABMIL		ABMIL+Siamese		PathoTME	
	ROCAUC	Acc	ROCAUC	Acc	ROCAUC	Acc
ACC	69.7±8.43	43.8±5.85	63.1±6.19	34.4±0.23	**72.7±3.89**	**45.6±4.74**
BLCA	58.2±2.81	34.2±5.53	64.3±3.98	47.6±3.95	**70.5±0.51**	**50.6±3.32**
BRCA	63.7±1.31	40.7±4.03	66.4±1.23	44.9±1.38	**71.2±1.43**	**47.9±2.08**
CESC-SCC	48.2±4.59	32.7±9.19	45.1±7.39	**46.1±1.36**	**51.5±6.16**	34.5±5.91
COREAD	61.5±5.38	35.9±6.22	56.7±10.7	40.0±2.74	**71.0±2.06**	**46.5±6.02**
ESGA-AC	63.7±5.90	37.4±6.06	71.2±2.29	43.7±6.23	**71.6±1.95**	**48.5±4.79**
HNSC	53.8±2.30	34.9±1.59	52.7±4.34	36.5±4.35	**67.2±1.18**	**43.8±4.14**
KIRC	52.8±3.40	34.0±7.90	58.5±2.58	42.2±1.15	**64.4±1.46**	**42.2±2.45**
KIRP	49.8±5.53	37.7±6.86	51.3±4.85	**51.6±1.04**	**60.2±5.00**	47.9±3.53
LUAD	65.4±4.21	39.7±4.34	65.9±2.69	34.4±3.65	**73.1±2.25**	**43.0±5.18**
LUSC	66.4±4.99	46.6±4.33	67.4±4.78	49.0±4.16	**75.6±1.62**	**58.6±2.64**
LICH	67.3±4.55	47.9±7.73	62.8±9.97	43.8±0.84	**71.8±3.59**	**55.1±4.50**
OV	**65.1±10.8**	40.0±13.8	56.6±14.9	34.5±11.9	64.6±7.42	**43.6±7.61**
PRAD	53.8±4.56	33.8±0.90	**55.5±4.48**	**34.4±0.32**	53.6±3.46	33.8±3.28
PAAD	62.7±6.04	38.3±8.01	49.8±5.89	**42.5±12.3**	**67.3±6.55**	41.0±2.35
SKCM	65.4±3.89	43.9±6.02	61.8±0.82	**46.0±2.45**	**69.5±2.44**	44.8±5.38
THCA	65.6±3.76	47.7±2.97	61.1±4.09	43.0±1.23	**71.7±1.61**	**51.0±4.88**
UCEC	**65.4±5.06**	**42.5±9.03**	59.7±7.42	33.3±12.5	63.0±4.97	34.2±3.49

DSMIL [22], CLAM [24].To compare the gene pretraining performance, we include Supervised Contrastive learning model [19], which is trained with gene expression and TME labels, and uses one-layer FC as classifier. We cannot include genomic fusion methods because genes used in the evaluation stage will lead to information leakage. In general, there is an about 5% improvement of ROCAUC, 5% of PRAUC, 4% of accuracy and 5% F1 score across overall 23 datasets. To demonstrate the upper bound of our method's prediction capability, we used an SNN that relies solely on gene expression to showcase the highest limit of TME prediction. However, our model will not use gene data during testing. Two stages mean we train gene guide branch and main branch with similarity loss first, and then remove the gene guide branch and train main branch with DANN. One stage means we train similarity loss, classifier loss and discriminator loss at the same time. One stage performs slightly better, which may be due to the regularization of discriminator playing a role on both two branches and improving the generalization.

Single TCGA dataset performance As shown in Table 2, we include the detailed comparisons on each dataset. For each TCGA dataset, we compare our PathoTME with ABMIL baseline and Siamese-ABMIL. The ABMIL and ABMIL-Siamese methods are trained individually on each dataset, while PathoTME is trained once on TCGA Pan-cancer datasets. Our method indicates a marked promotion over the ABMIL and ABMIL-Siamese in 15 out of 18 datasets on ROCAUC and 13 out of 18 on accuracy. For the rest datasets, we achieve near-optimal results. It indicates that our method is able to achieve similar features

among different datasets, resulting in better performance than learning from a single database. To notice, in some tumor types, Siamese networks might slightly impede performance, likely due to a limited sample size preventing the gene embeddings from learning better representation space. Additionally, because of a large total sample size, PathoTME exhibits significantly lower standard deviations on metrics compared to the other methods, indicating a more stable classification performance. In other methods, we noticed that there are higher variances in some certain types. This might be due to the low number of tumor samples. We performed a correlation analysis between variance and sample size, resulting in $r < -0.6$, indicating a significant negative correlation. We excluded 5 datasets with less than 10 samples in the test set in order to ensure the confidence of the statistical results.

3.4   Ablation analysis

To evaluate the proposed components, we conduct an ablation study on Siamese, DANN and Visual prompts. Table 1 indicates various degrees of different configurations on metrics. Compared with the baseline, which is ABMIL, incorporating Siamese achieves 66.1% ROCAUC, surpassing 3.1% accuracy. The underlying reason may be that Siamese architecture transfers discriminative embedding to WSI space. DANN improves performance to 67.3% on ROCAUC and 44.6% on accuracy, indicating DANN addresses domain adaptation challenges, and helps tumor types with fewer samples in learning features from those with abundant samples. Visual prompts show promotion of 2.4% on ROCAUC and 1.7% on accuracy, indicating VPT can fine-tune the pretrained WSI feature extractor for our task to get better performance.

4   Conclusion & Discussion

Tumor microenvironment subtype prediction based on Whole Slide Images is crucial to precision medicine, which includes patient stratification, prognosis, and the determination of clinical treatment. In this paper, we make the first step to classify TME subtypes. We demonstrate that the proposed PathoTME achieves better results on the task of TME subtype classification. Our method achieves better performance than other state-of-the-art methods, but the overall accuracy on this task is still not satisfactory. There are some possible reasons: (1) Recent research suggested that immune microenvironment may have more refined subtypes and tend to be a continuous spectrum rather than specific categorizations. (2) At WSI feature extraction stage, we use pretrained 4K HIPT model. The extracted features have been significantly compressed to low dimensions, and the regions may be too large to include more details, potentially resulting in the neglect of part important features. (3) There are other distance measures like JSD, MMD and wasserstein distance can be combined in our model. We expect to achieve better performance in this task by more trials.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azher, Z.L., Suvarna, A., Chen, J.Q., Zhang, Z., Christensen, B.C., Salas, L.A., Vaickus, L.J., Levy, J.J.: Assessment of emerging pretraining strategies in interpretable multimodal deep learning for cancer prognostication. BioData Mining **16**(1), 23 (2023)

2. Bagaev, A., Kotlov, N., Nomie, K., Svekolkin, V., Gafurov, A., Isaeva, O., Osokin, N., Kozlov, I., Frenkel, F., Gancharova, O., et al.: Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. Cancer cell **39**(6), 845–865 (2021)

3. Bhattacharya, S., Dunn, P., Thomas, C.G., Smith, B., Schaefer, H., Chen, J., Hu, Z., Zalocusky, K.A., Shankar, R.D., Shen-Orr, S.S., et al.: Immport, toward repurposing of open access immunological assay data for translational and clinical research. Scientific data **5**(1), 1–9 (2018)

4. Chen, C.L., Chen, C.C., Yu, W.H., Chen, S.H., Chang, Y.C., Hsu, T.I., Hsiao, M., Yeh, C.Y., Chen, C.Y.: An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nature communications **12**(1), 1193 (2021)

5. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16144–16155 (June 2022)

6. Chen, R.J., Krishnan, R.G.: Self-supervised vision transformers learn visual concepts in histopathology. Learning Meaningful Representations of Life, NeurIPS 2021 (2021)

7. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Shaban, M., Shady, M., Williams, M., Joo, B., Noor, Z., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell (2022)

8. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)

9. Combes, A.J., Samad, B., Tsui, J., Chew, N.W., Yan, P., Reeder, G.C., Kushnoor, D., Shen, A., Davidson, B., Barczak, A.J., et al.: Discovering dominant tumor immune archetypes in a pan-cancer census. Cell **185**(1), 184–203 (2022)

10. Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Wang, Y., Zhao, S., Yang, H., Landman, B.A., Huo, Y.: Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 626–635. Springer (2022)

11. Ding, K., Zhou, M., Metaxas, D.N., Zhang, S.: Pathology-and-genomics multimodal transformer for survival outcome prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 622–631. Springer (2023)

12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)

13. Gao, R., Yuan, X., Ma, Y., Wei, T., Johnston, L., Shao, Y., Lv, W., Zhu, T., Zhang, Y., Zheng, J., et al.: Harnessing tme depicted by histological images to improve cancer prognosis through a deep learning system. Cell Reports Medicine **5**(5) (2024)

14. Godson, L., Alemi, N., Nsengimana, J., Cook, G.P., Clarke, E.L., Treanor, D., Bishop, D.T., Newton-Bishop, J., Gooya, A., Magee, D.: Immune subtyping of melanoma whole slide images using multiple instance learning. Medical Image Analysis p. 103097 (2024)

15. He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., Zou, J.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nature biomedical engineering **4**(8), 827–834 (2020)

16. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

17. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)

18. Jin, T., Xie, X., Wan, R., Li, Q., Wang, Y.: Gene-induced multimodal pre-training for image-omic classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 508–517. Springer (2023)

19. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning (2021)

20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

21. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks (2017)

22. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)

23. Li, C., Wang, S., Liu, P., Torheim, T., Boonzaier, N.R., van Dijken, B.R., Schönlieb, C.B., Markowetz, F., Price, S.J.: Decoding the interdependence of multiparametric magnetic resonance imaging to reveal patient subgroups correlated with survivals. Neoplasia **21**(5), 442–449 (2019)

24. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)

25. Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., Kamoun, A., Sefta, M., Toldo, S., Zaslavskiy, M., et al.: A deep learning model to predict rna-seq expression of tumours from whole slide images. Nature communications **11**(1), 3877 (2020)

26. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)

27. Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A.J., Gevaert, O.: Multimodal data fusion for cancer biomarker discovery with deep learning. Nature Machine Intelligence **5**(4), 351–362 (2023)

28. Wang, S., Wang, T., Yang, L., Yang, D.M., Fujimoto, J., Yi, F., Luo, X., Yang, Y., Yao, B., Lin, S., et al.: Convpath: A software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. EBioMedicine **50**, 103–110 (2019)

29. Wang, X., Price, S., Li, C.: Multi-task learning of histology and molecular markers for classifying diffuse glioma. arXiv preprint arXiv:2303.14845 (2023)

30. Wei, Y., Chen, X., Zhu, L., Zhang, L., Schönlieb, C.B., Price, S., Li, C.: Multimodal learning for predicting the genotype of glioma. IEEE Transactions on Medical Imaging **42**(11), 3167–3178 (Nov 2023)

31. Whiteside, T.: The tumor microenvironment and its role in promoting tumor growth. Oncogene **27**(45), 5904–5912 (2008)

32. Xie, R., Pang, K., Chung, S., Perciani, C., MacParland, S., Wang, B., Bader, G.: Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. Advances in Neural Information Processing Systems **36** (2024)

33. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21485–21494 (2023)