Modeling Whole-Slide Images as Dynamic Tumor Microenvironment Fields

Lei Wu[1*], Jiashuai Liu[1*], Di Zhang[1], Zhangpeng Gong[1], Yingkang Zhan[1], Yi Niu[1], Jiusong Ge[1], Chunze Yang[1], Kai Yi[3], Mireia Crispin-Ortuzar[3], Chen Li[2,1†], Zeyu Gao[3†]

[1]School of Computer Science and Technology, Xi'an Jiaotong University
[2]Institute of Intelligent Medicine, Chinese Academy of Medical Sciences
[3]University of Cambridge

Abstract

Due to the gigapixel-scale nature of whole-slide images (WSIs), weakly supervised WSI analysis is commonly formulated as a multiple instance learning (MIL) problem, where patch-level features are aggregated into slide-level representations. However, diagnostic and prognostic evidence often arises from spatially coherent tumor microenvironment regions and their interactions, rather than isolated patches alone. Existing patch-level or static region-based methods usually overlook how tissue regions should be adaptively formed and subsequently evolved through microenvironment interactions across heterogeneous boundaries. In this paper, we propose **Concept-Guided Tumor Microenvironment Evolution** (**TMEvolve**), a reaction-diffusion-inspired framework that models WSIs as latent tumor microenvironment fields over discrete patch graphs. TMEvolve instantiates this view as a learnable graph-discretized evolution process over patch neighborhoods. It first forms adaptive soft tissue regions as coherent microenvironment units, then performs pseudo-time evolution through two complementary local dynamics: intra-region diffusion, which stabilizes latent states within coherent tissue compartments, and concept-guided boundary flux, which propagates visual feature signals and language-derived concept signals across heterogeneous region interfaces. The evolved microenvironment regions are finally aggregated for slide-level prediction. We evaluate TMEvolve on six datasets across three weakly supervised WSI tasks: survival prediction, gene expression prediction, and histological subtype classification. TMEvolve consistently improves over representative MIL methods, pathology foundation models, and concept-guided baselines. Ablation studies and visualizations further support the effectiveness and interpretability of TMEvolve, highlighting the value of dynamic region modeling and boundary interaction. Our implementation is publicly available at `https://github.com/leowu-ai/TMEvolve`.

1 Introduction

Digitized whole-slide images provide gigapixel-scale histopathological morphology for computational pathology, supporting cancer diagnosis, prognosis, molecular prediction, and subtype classification [30, 6, 9, 10, 46]. Since fine-grained annotations are expensive, weakly supervised multiple instance learning (MIL) is widely used, where each slide is represented as a bag of patches and trained with slide-level labels [23, 30, 28, 38]. Recent pathology vision-language methods further enhance WSI representation learning by introducing semantic priors [21, 22, 31, 49, 27]. However,

* These authors contributed equally.
† Co-corresponding authors: `cli@pumc.edu.cn, zg323@cam.ac.uk`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

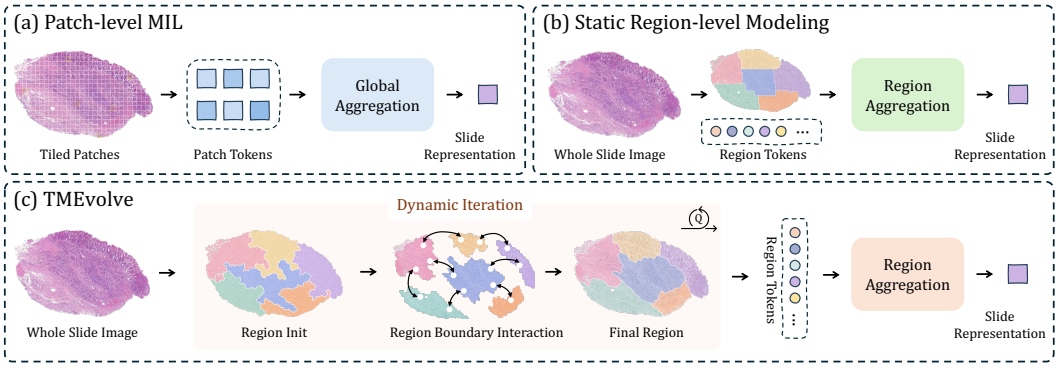

Figure 1: Overview of WSI modeling paradigms. Patch-level MIL ignores region structure, while static region modeling relies on rigid regions. TMEvolve models adaptive tissue regions and their concept-guided inter-region evolution.

most existing WSI-level models still lack an explicit mechanism to organize patches into tissue-level structures and model interactions among microenvironment regions.

Conventional MIL directly aggregates patch tokens into a slide representation through attention pooling [23], Transformer aggregation [38], or other fusion mechanisms [28] (Figure 1(a)). Although effective, this patch-to-slide paradigm treats a WSI as a static collection of weakly related instances and overlooks region-level tissue organization. In routine diagnosis, pathologists often reason in a region-centric manner by organizing morphology into spatially coherent tissue compartments, such as tumor nests, stromal areas, and immune-enriched microenvironments. Static region modeling partially addresses this issue by aggregating local or pre-defined regions (Figure 1(b)) [8, 7, 50, 24, 3]. However, fixed grids, heuristic partitions, or static neighborhoods may cut across natural tissue boundaries and mix heterogeneous microenvironment components. Moreover, they usually treat regions as static aggregation units, without modeling how tissue states become coherent within regions or how heterogeneous regions influence each other across boundaries.

These limitations motivate a dynamic view of WSI representation learning. A WSI can be regarded as a spatial tissue field whose latent state varies over tissue coordinates. Within relatively homogeneous regions, neighboring patches tend to share coherent morphological and semantic states, suggesting a diffusion prior for local coordination [7, 24]. Across heterogeneous interfaces, different microenvironment states may interact directionally, requiring concept-guided modeling of cross-region flux and reaction [11, 35, 15, 14, 1, 25]. This motivates a framework inspired by reaction–diffusion dynamics, where local states are stabilized within coherent tissue regions and semantic interactions are modeled across heterogeneous boundaries.

In this paper, we propose **Concept-Guided Tumor Microenvironment Evolution** (**TMEvolve**), a dynamic system for WSI representation learning informed by pathological concepts (Figure 1(c)). Rather than solving a reaction–diffusion partial differential equation (PDE), TMEvolve instantiates this view as a learnable graph-discretized latent evolution framework over patch neighborhoods, motivated by studies linking graph neural message passing to diffusion or PDE-inspired dynamics [4, 13, 5, 42]. TMEvolve first forms adaptive soft tissue regions from morphological similarity and spatial continuity. It anchors these regions with language-encoded pathological concepts and evolves them in pseudo-time through intra-region diffusion and concept-guided boundary flux. The evolved region representations are finally aggregated for slide-level prediction.

We evaluate TMEvolve on survival prediction, gene expression prediction, and subtype classification. Across six cohorts, TMEvolve achieves strong performance compared with representative MIL methods, pathology foundation models, and survival-oriented baselines. Ablation studies validate the key design choices, while visualization results show that TMEvolve progressively refines tissue regions and provides interpretable evidence for slide-level prediction.

Our contributions are summarized as follows:

(1) We propose TMEvolve, an evolution-aware MIL framework that models WSIs as dynamic tumor microenvironment fields composed of interacting tissue regions. (2) We introduce dynamic

region partition and intra-region coordination to construct spatially coherent and concept-bearing microenvironment units. (3) We design a concept-guided inter-region evolution mechanism to regulate boundary interactions between heterogeneous tissue regions. (4) TMEvolve achieves strong performance across survival prediction, gene expression prediction, and subtype classification, with ablation and visualization results supporting its effectiveness and interpretability.

2 Related Work

Weakly supervised MIL baselines for WSI prediction. Weakly supervised multiple instance learning has become a standard paradigm for WSI analysis, where each slide is represented as a bag of patch instances and trained with slide-level labels. Representative MIL methods mainly differ in how patch-level information is aggregated into slide-level representations. ABMIL [23] introduces attention-based pooling for bag-level prediction, DSMIL [28] uses a dual-stream formulation to identify critical instances, TransMIL [38] models long-range patch correlations with Transformer aggregation, and RRT-MIL [39] improves slide representations through feature re-embedding. These methods provide strong general baselines across survival prediction, gene expression prediction, and subtype classification. However, most of them primarily focus on patch aggregation, without explicitly modeling how heterogeneous tissue states interact during representation learning.

Pathology foundation and vision-language models. Recent pathology foundation models provide stronger transferable representations for whole-slide analysis. Models such as CHIEF [41], Feather [37], and TITAN [12] have shown strong performance in diagnosis, prognosis, and general WSI representation learning. Vision-language and concept-guided methods further introduce semantic priors into pathology modeling [18, 44, 19]. PLIP [21], Quilt [22], and CONCH [31] align pathology images with language representations, while CITE [48], QPMIL [20], ConcepPath [49], and GECKO [27] use textual concepts or vision-language cues for interpretable classification or molecular prediction. These methods are closely related to our comparisons because they use stronger pretrained representations or semantic guidance to improve downstream WSI prediction. Nevertheless, they mainly use language or concepts for global alignment, query-based aggregation, or concept-level prediction, whereas TMEvolve uses pathological concepts to guide interactions among tissue states during pseudo-time evolution.

Spatial and microenvironment-aware WSI modeling. Beyond global slide-level aggregation, several studies have explored spatially aware WSI representations [17, 45]. Hierarchical, graph-based, or clustering-guided models, such as HIPT [8], Patch-GCN [7], graph-based WSI modeling [50], and clustering-guided graph MIL [24], incorporate local tissue context or spatial relationships into slide representation learning. These methods show the value of spatial organization, but they usually use spatial structure as a fixed context for feature aggregation. From a biomedical perspective, tumor progression is shaped by the organization and interaction of tumor microenvironment components, including immune infiltration [16, 14, 1], tumor invasion [15], immune exclusion, and stromal remodeling [25, 26]. Motivated by this view, TMEvolve models WSIs as dynamic tumor microenvironment fields, where concept-bearing tissue states are refined through intra-region coordination and concept-guided boundary interaction.

3 Method

In this section, we present **TMEvolve**, an evolution-aware MIL framework that models a WSI as an adaptive tumor microenvironment field. Instead of aggregating static patch instances, TMEvolve formulates WSI representation learning as a graph-discretized latent reaction–diffusion process [42], where within-region diffusion stabilizes coherent tissue compartments, boundary anisotropic flux models heterogeneous interface interactions, and concept-conditioned reaction incorporates language-encoded pathological semantics. As shown in Figure 2, the following subsections describe the latent field formulation, adaptive region diffusion, concept-conditioned reaction, boundary flux, pseudo-time evolution, and task-specific aggregation.

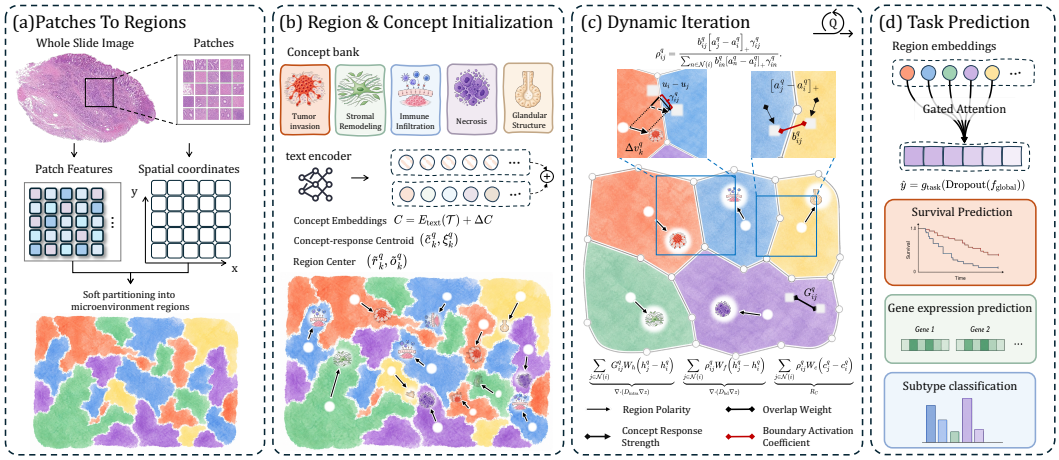

Figure 2: Overview of TMEvolve. (a) TMEvolve first encodes WSI patches and spatial coordinates, and softly partitions them into adaptive regions. (b) Region states are initialized with language-encoded concepts, concept-response centroids, and region centers. (c) Dynamic interaction refines region representations through intra-region diffusion, concept-conditioned reaction, and anisotropic boundary flux. (d) The evolved region embeddings are aggregated by gated attention for survival prediction, gene expression prediction, and subtype classification.

3.1 WSI as a Latent Reaction–Diffusion Microenvironment Field

We first introduce the latent field formulation underlying TMEvolve. In a classical reaction–diffusion system, the evolution of a spatial state field $z(u, q) \in \mathbb{R}^d$ is written as [32]

$$\frac{\partial z(u,q)}{\partial q} = \nabla \cdot \big(D(u,q)\nabla z(u,q)\big) + R(z,u,q), \tag{1}$$

where u denotes the spatial tissue coordinate, q denotes the pseudo-time variable for iterative latent-state evolution, $D(u, q)$ controls spatial diffusion, and $R(z, u, q)$ denotes reaction effects. For WSI representation learning, this formulation naturally matches the heterogeneous tissue field, where local consistency is needed within coherent regions and interactions are expected across heterogeneous interfaces. We therefore decompose the diffusion operator into a within-region component D_{intra} and a boundary anisotropic component D_{bd}, and introduce a concept-conditioned semantic source term $R_{\mathcal{C}}$:

$$\frac{\partial z(u,q)}{\partial q} = \nabla \cdot \big(D_{\text{intra}}(u,q)\nabla z(u,q)\big) + \nabla \cdot \big(D_{\text{bd}}(u,q)\nabla z(u,q)\big) + R_{\mathcal{C}}(z,u,q). \tag{2}$$

Here, the first term stabilizes representations inside coherent microenvironment regions, the second term models anisotropic flux across heterogeneous tissue interfaces, and the third term injects pathological concept semantics into latent state evolution. The pseudo-time variable q denotes iterative representation refinement rather than chronological disease progression.

Since WSIs are observed as discrete patch sets, we implement this formulation on a patch neighborhood graph. Given initial patch features $H^0 = [h_i^0]_{i=1}^N \in \mathbb{R}^{N\times d}$ and normalized coordinates $U = [u_i]_{i=1}^N \in \mathbb{R}^{N\times 2}$, the continuous state $z(u_i, q)$ is represented by h_i^q. A detailed graph-discretized derivation from Eq. (2) to the patch-level evolution updates is provided in Appendix A.

3.2 Adaptive Region Field and Within-region Diffusion

We discretize the within-region diffusion term $\nabla \cdot (D_{\text{intra}}\nabla z)$ by constructing an adaptive region field over WSI patches. The region field estimates where local information exchange should be encouraged: patches within coherent tissue compartments are allowed to diffuse, while propagation across heterogeneous or uncertain boundaries is suppressed.

At initialization, we obtain spatially dispersed region seeds by farthest point sampling. Each region k is associated with a region feature prototype r_k^q and a spatial center o_k^q. At pseudo-time step q, the

adaptive region field is represented by a soft assignment matrix $S^q \in \mathbb{R}^{N\times K}$, where $S^q_{i,k}$ denotes the membership of patch i to region k. To construct adaptive regions that are morphologically consistent and spatially continuous, the soft assignment is defined as:

$$E^q_{i,k} = \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}\left(\left\|\frac{h^q_j}{\|h^q_j\|_2} - \frac{r^q_k}{\|r^q_k\|_2}\right\|_2^2 + \|u_j - o^q_k\|_2^2\right), \quad S^q_{i,k} = \frac{\exp(-E^q_{i,k})}{\sum_{k'=1}^K \exp(-E^q_{i,k'})}. \tag{3}$$

where $\mathcal{N}(i)$ denotes the spatial neighborhood of patch i. The adaptive region field induces a state-dependent conductivity on the patch graph. For each neighboring patch pair (i, j), we compute their normalized membership-overlap weight G^q_{ij} and perform within-region diffusion as

$$G^q_{ij} = \frac{\sum_{k=1}^K S^q_{i,k}S^q_{j,k}}{\sum_{n\in\mathcal{N}(i)}\sum_{k=1}^K S^q_{i,k}S^q_{n,k}}, \quad \widetilde{h}^q_i = h^q_i + \sum_{j\in\mathcal{N}(i)} G^q_{ij}W_h\left(h^q_j - h^q_i\right). \tag{4}$$

Here, G^q_{ij} serves as the state-dependent conductivity for $\nabla\cdot(D_{\text{intra}}\nabla z)$, assigning larger weights to neighboring patches with similar region memberships and suppressing cross-boundary smoothing.

3.3 Concept-guided Boundary Reaction–Flux Evolution

After within-region diffusion, we further model interactions across heterogeneous microenvironment interfaces. Instead of treating boundary flux and concept reaction as two independent modules, TMEvolve couples them through a shared anisotropic boundary conductivity. This conductivity is derived from region heterogeneity, concept-response contrast, and region-level polarity.

For a given cancer type or downstream task, we collect task-relevant pathological descriptions $\mathcal{T} = \{\tau_1, \ldots, \tau_M\}$, such as invasive growth, immune infiltration, necrosis, and stromal remodeling. A text encoder maps these descriptions into concept embeddings, which are adapted by a learnable residual:

$$C = E_{\text{text}}(\mathcal{T}) + \Delta C = [c_1, \ldots, c_M]^\top \in \mathbb{R}^{M\times d}. \tag{5}$$

Given the soft assignment matrix S^q and the locally coordinated patch representations $\widetilde{H}^q$, we summarize each region by its semantic state $\widetilde{r}^q_k = \sum_{i=1}^N S^q_{i,k}\widetilde{h}^q_i / \sum_{i=1}^N S^q_{i,k}$ and geometric center $\widetilde{o}^q_k = \sum_{i=1}^N S^q_{i,k}u_i / \sum_{i=1}^N S^q_{i,k}$. Each region is then anchored to its dominant pathological concept by $\theta^q_k = \arg\max_m(\widetilde{r}^{q\top}_k c_m/\sqrt{d})$, and we set $\widetilde{c}^q_k = c_{\theta^q_k}$. The assigned concept $\widetilde{c}^q_k$ defines the region-level semantic identity and provides the semantic reference for locating concept-responsive patches within the region.

Region polarity. To capture where the dominant concept is spatially concentrated inside region k, we compute the concept-response centroid and its deviation from the geometric region center as

$$\eta^q_{i,k} = \frac{S^q_{i,k}\exp\left(\widetilde{h}^{q\top}_i\widetilde{c}^q_k/\sqrt{d}\right)}{\sum_{\ell=1}^N S^q_{\ell,k}\exp\left(\widetilde{h}^{q\top}_\ell\widetilde{c}^q_k/\sqrt{d}\right)}, \qquad \xi^q_k = \sum_{i=1}^N \eta^q_{i,k}u_i, \qquad \Delta v^q_k = \xi^q_k - \widetilde{o}^q_k. \tag{6}$$

Here, $\eta^q_{i,k}$ combines region membership and concept similarity, ξ^q_k denotes the concept-response centroid, and Δv^q_k captures its spatial bias relative to the geometric region center, providing the directional cue for anisotropic boundary flux.

Concept response strength. We further assign the region-level concept back to patch-level states. For each patch i, we identify its dominant region by $\beta^q_i = \arg\max_k S^q_{i,k}$ and assign the corresponding concept $c^q_i = \widetilde{c}^q_{\beta^q_i}$. The patch-level concept response strength is defined as $a^q_i = \widetilde{h}^{q\top}_i c^q_i/\sqrt{d}$. The response strength a^q_i is used to measure concept-response contrast between neighboring patches, which later regulates boundary interactions across heterogeneous region interfaces.

Many pathology-relevant signals arise around tumor–stroma, tumor–immune, glandular–stromal, and necrosis-related boundaries. Therefore, based on the region polarity Δv^q_k and patch-level concept response strength a^q_i defined above, we model cross-region interactions as anisotropic boundary

reaction–flux. The boundary conductivity is jointly determined by **boundary heterogeneity**, **concept-response contrast**, and **directional compatibility**.

Boundary heterogeneity. For each graph edge (i, j), we define the boundary activation coefficient as $b_{ij}^q = 1 - \sum_{k=1}^{K} S_{i,k}^q S_{j,k}^q$. It measures the cross-region heterogeneity from soft assignments, where a larger b_{ij}^q indicates that the edge is more likely to lie on a heterogeneous microenvironment boundary.

Directional compatibility. To regulate the direction of boundary flux, we use the normalized polarity of the source patch region. For patch j, we identify its dominant region by $\epsilon_j^q = \arg\max_k S_{j,k}^q$ and define $\widehat{\Delta v}_j^q = \Delta v_{\epsilon_j^q}^q / \|\Delta v_{\epsilon_j^q}^q\|_2$. The directional compatibility coefficient is defined as $\gamma_{ij}^q = [(u_i - u_j)^\top \widehat{\Delta v}_j^q]_+$, where $[x]_+ = \max(x, 0)$. A larger γ_{ij}^q indicates that the direction from source patch j to target patch i is more consistent with the concept-response polarity of the source region.

Boundary conductivity. Together with the concept response strength a_i^q, we define the anisotropic boundary conductivity as

$$\rho_{ij}^q = \frac{b_{ij}^q \left[a_j^q - a_i^q\right]_+ \gamma_{ij}^q}{\sum_{n \in \mathcal{N}(i)} b_{in}^q \left[a_n^q - a_i^q\right]_+ \gamma_{in}^q}. \tag{7}$$

Here, b_{ij}^q captures **boundary heterogeneity**, $\left[a_j^q - a_i^q\right]_+$ captures positive **concept-response contrast**, and γ_{ij}^q imposes anisotropic **directional compatibility**. Thus, ρ_{ij}^q serves as the graph-discretized conductivity for boundary-localized reaction–flux.

Reaction–flux update. With ρ_{ij}^q, we instantiate the boundary anisotropic diffusion term $\nabla \cdot (D_{\mathrm{bd}} \nabla z)$ and the concept-conditioned semantic source term $R_{\mathcal{C}}$ on heterogeneous tissue interfaces. Since the locally coordinated representation $\widetilde{h}_i^q$ already contains the within-region diffusion update, the subsequent boundary-localized evolution is written as

$$h_i^{q+1} = \widetilde{h}_i^q + \underbrace{\sum_{j \in \mathcal{N}(i)} \rho_{ij}^q W_f \left(h_j^q - h_i^q\right)}_{\nabla \cdot (D_{\mathrm{bd}} \nabla z)} + \underbrace{\sum_{j \in \mathcal{N}(i)} \rho_{ij}^q W_c \left(c_j^q - c_i^q\right)}_{R_{\mathcal{C}}}. \tag{8}$$

The first boundary term models feature-level anisotropic flux across heterogeneous interfaces, and the second term instantiates $R_{\mathcal{C}}$ as a boundary-localized semantic reaction. Equivalently, by expanding $\widetilde{h}_i^q$, this update gives a graph-discretized realization of Eq. (2), where the diffusion conductivities, concept responses, and region assignments are dynamically recomputed along pseudo-time.

Region refresh. After each pseudo-time step, the updated patch states induce a new adaptive region field. Specifically, we recompute the patch-to-region soft assignment S^{q+1} using the updated patch states and update the region prototype and spatial center as

$$r_k^{q+1} = \frac{\sum_{i=1}^{N} S_{i,k}^{q+1} h_i^{q+1}}{\sum_{i=1}^{N} S_{i,k}^{q+1}}, \quad o_k^{q+1} = \frac{\sum_{i=1}^{N} S_{i,k}^{q+1} u_i}{\sum_{i=1}^{N} S_{i,k}^{q+1}}. \tag{9}$$

Low-mass regions are also merged as described in the adaptive region construction step. Since G_{ij}^q, ρ_{ij}^q, c_i^q, and S^q are recomputed from the current states, TMEvolve performs state-dependent latent field evolution rather than static graph message passing. After Q evolution steps, we obtain the evolved patch states H^Q, the final soft assignment matrix S^Q, and the evolved region states $R^Q = \{r_k^Q\}_{k=1}^K$ from the final region update.

3.4 Evolution-aware Aggregation and Task Prediction

We aggregate the evolved region states for downstream prediction [23]. Each r_k^Q encodes region-level morphology together with the accumulated effects of within-region diffusion, boundary anisotropic flux, and concept-conditioned reaction. The slide-level representation and prediction are computed by

$$f_{\mathrm{global}} = \sum_{k=1}^{K} \beta_k r_k^Q, \quad \hat{y} = g_{\mathrm{task}} \left(\mathrm{Dropout}(f_{\mathrm{global}})\right), \tag{10}$$

where β_k is the attention weight of region k, and $g_{\mathrm{task}}(\cdot)$ is the task-specific prediction head for survival prediction, subtype classification, or gene expression prediction.

Table 1: Main results on survival prediction. Results are C-index ($\times 100$), reported as mean $\pm$ std. The best result in each cohort is shown in bold and the second-best result is underlined.

Cohort	ABMIL	DSMIL	TransMIL	RRT-MIL	CHIEF	Feather	TITAN	VLSA	ProtoSurv	Ours
LUAD	63.9±5.3	64.5±3.3	66.8±4.8	66.1±5.2	63.1±3.0	63.5±1.4	66.8±4.0	64.5±4.8	66.5±5.7	**69.5±4.2**
GBC	75.0±4.8	73.8±5.0	73.3±3.6	74.2±4.3	73.8±5.3	75.2±2.6	75.4±3.5	76.1±4.4	76.2±1.8	**77.4±3.8**
BLCA	61.9±6.4	55.5±7.1	60.2±3.9	61.6±4.1	61.2±6.5	58.6±3.8	63.0±2.0	60.9±3.6	62.6±7.0	**65.9±5.7**
BRCA	66.7±4.2	60.6±2.7	66.1±3.8	66.0±3.6	66.8±3.1	62.7±1.1	67.4±3.7	69.3±6.4	**70.4±1.1**	70.3±3.1

Table 2: Main results on fine-grained and coarse-grained subtype classification. Results are reported as mean $\pm$ std. AUC and ACC are reported in percentage (%). The best result in each metric is shown in bold and the second-best result is underlined.

Dataset	Level	Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	CHIEF	Feather	TITAN	CITE	QPMIL	Ours
EBRAINS	Fine	AUC	97.9±0.2	97.9±0.3	98.0±0.3	97.7±0.6	97.2±0.4	97.7±0.4	98.2±0.4	98.0±0.5	97.7±0.2	**98.8±0.2**
		ACC	75.9±1.1	76.8±1.5	76.5±1.2	76.3±2.2	71.5±1.5	76.4±2.3	**79.2±1.8**	76.2±2.9	75.7±2.2	79.0±1.3
	Coarse	AUC	99.4±0.2	99.5±0.1	99.5±0.1	99.5±0.2	98.5±0.3	98.9±0.2	99.5±0.1	98.8±0.4	99.6±0.1	**99.6±0.2**
		ACC	92.9±1.2	93.1±0.5	93.1±0.8	92.8±1.4	87.4±1.2	91.7±1.0	**93.8±1.0**	92.0±0.8	93.5±0.4	93.7±0.7
BRACS	Fine	AUC	85.7±0.7	85.3±2.0	85.4±1.3	85.9±0.9	86.3±1.5	88.0±2.3	85.5±3.2	86.6±2.3	87.3±1.6	**89.4±0.6**
		ACC	60.7±2.0	59.4±3.9	58.6±1.7	58.6±1.6	62.7±3.9	62.2±3.3	61.4±5.0	61.8±3.9	61.5±2.0	**65.4±1.0**
	Coarse	AUC	92.0±2.3	90.2±3.5	90.8±3.6	90.9±2.6	93.1±1.3	91.9±1.3	**93.5±1.7**	89.8±2.2	92.3±1.3	92.6±0.7
		ACC	82.5±1.4	81.8±3.6	80.7±2.8	79.0±4.5	81.8±2.1	81.4±2.2	82.5±2.4	81.6±1.8	83.3±2.8	**83.5±1.8**

To avoid degenerate region usage, we regularize the final region mass $\pi_k^Q = \frac{1}{N}\sum_{i=1}^{N} S_{i,k}^Q$ and define the overall objective as

$$\mathcal{L}_{\text{div}} = \sum_{k=1}^{K}\left(\pi_k^Q - \frac{1}{K}\right)^2, \quad \mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}. \tag{11}$$

Here, λ_{div} controls the strength of the diversity regularization, and $\mathcal{L}_{\text{task}}$ is instantiated according to the downstream task.

4 Experiments

4.1 Experimental Setup

Datasets and tasks. We evaluate TMEvolve on six datasets across survival prediction, gene expression prediction, and subtype classification. For public TCGA cohorts, we use LUAD [33], BLCA [33], and BRCA [40]. Survival prediction is evaluated on these three TCGA cohorts and an additional private gallbladder cancer (GBC) cohort, which will be made publicly available in the future. For gene expression prediction, we use LUAD, BLCA, and BRCA to predict eight clinically meaningful genes per cohort. For subtype classification, we evaluate on BRACS [2] and EBRAINS [36] under both fine-grained and coarse-grained settings.

Implementation details. We use CONCH v1.5 [31] as the patch feature extractor and Qwen3-Embedding-8B [47] as the text feature extractor. Pathological concepts are generated with GPT-5 [34] and reviewed by pathology experts. We set the diversity regularization weight to $\lambda_{\text{div}} = 0.1$. More dataset details and hyperparameter settings are provided in Appendix B, computational cost in Appendix C, and full concept banks in Appendix D.

Evaluation metrics. We report the concordance index (C-index) for survival prediction, Pearson correlation coefficient for gene expression prediction, and AUC and accuracy (ACC) for subtype classification under both fine-grained and coarse-grained settings.

4.2 Comparison with State-of-the-art Methods

Comparison methods. We compare TMEvolve with representative WSI-level methods across survival prediction, subtype classification, and gene expression prediction. The compared methods include general MIL baselines such as ABMIL [23], DSMIL [28], TransMIL [38], and RRT-MIL [39];

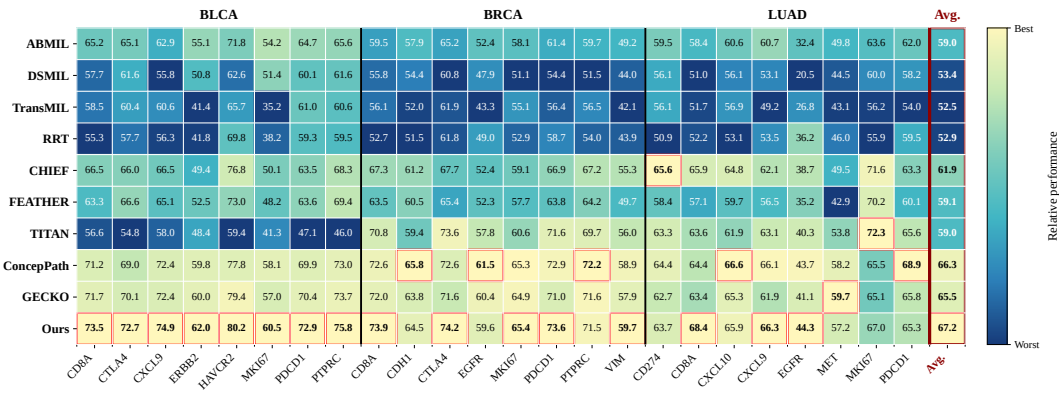

Figure 3: Gene-level prediction performance across TCGA datasets using Pearson correlation multiplied by 100. Columns denote gene prediction tasks, rows denote compared methods, and color intensity indicates relative performance within each task, with brighter colors representing better results. Red boxes highlight the best result in each column. The average column summarizes overall performance across all gene-level tasks.

Table 3: Ablation study of TMEvolve on survival prediction, gene expression prediction, and BRACS subtype classification. Results are reported as mean $\pm$ std. The best result in each column is shown in bold.

Variant	Survival		Gene Prediction		Subtype Classification	
	BLCA C-index	LUAD C-index	BLCA Pearson	LUAD Pearson	BRACS AUC	BRACS ACC
w/o Intra-region Diffusion	0.6524±0.0549	0.6812±0.0474	0.7058±0.0323	0.6088±0.0284	0.8883±0.0142	0.6431±0.0218
Random Concepts	0.6477±0.0570	0.6728±0.0621	0.7016±0.0297	0.6044±0.0307	0.8873±0.0119	0.6344±0.0095
w/o Pseudo-time Evolution	0.6421±0.0468	0.6746±0.0540	0.7022±0.0267	0.6061±0.0298	0.8885±0.0085	0.6470±0.0243
w/o $\mathcal{L}_{\text{div}}$	0.6538±0.0460	0.6817±0.0471	0.7104±0.0263	0.6125±0.0361	0.8894±0.0084	0.6513±0.0196
Full TMEvolve	**0.6589±0.0570**	**0.6952±0.0418**	**0.7158±0.0326**	**0.6226±0.0288**	**0.8940±0.0057**	**0.6544±0.0095**

pathology foundation or WSI-level models such as CHIEF [41], Feather [37], and TITAN [12]; survival-oriented methods such as VLSA [29] and ProtoSurv [43]; subtype-oriented concept models such as CITE [48] and QPMIL [20]; and gene-oriented concept-guided methods such as ConcepPath [49] and GECKO [27] when applicable.

Main results. For *survival prediction*, Table 1 shows that TMEvolve achieves the best performance on three out of four cohorts and remains highly competitive on BRCA. Compared with representative MIL, foundation-model, and survival-oriented baselines, TMEvolve shows consistent advantages on LUAD, GBC, and BLCA, suggesting that heterogeneous tissue regions and their interactions provide informative prognostic evidence. For *subtype classification*, Table 2 shows that TMEvolve performs strongly under both fine-grained and coarse-grained settings, achieving the best or competitive results on EBRAINS and BRACS across AUC and ACC. For *gene expression prediction*, Figure 3 presents gene-level Pearson correlation results across TCGA cohorts. TMEvolve achieves strong and consistent performance on clinically relevant genes, especially immune- and proliferation-related genes, suggesting that dynamic microenvironment modeling helps capture weak and spatially distributed morphology-related molecular signals. Additional detailed gene-level results are provided in Appendix E.

4.3 Ablation Study and Hyperparameter Sensitivity

Main module ablation. As shown in Table 3, removing *intra-region diffusion*, *pseudo-time evolution*, or the diversity loss $\mathcal{L}_{\text{div}}$ consistently degrades performance, demonstrating the importance of local diffusion, iterative refinement, and balanced region usage for stable microenvironment modeling. Replacing *language-encoded concepts* with random embeddings also weakens the results, especially for gene expression prediction, suggesting that language-derived semantic priors help capture molecularly related morphology. Overall, the full TMEvolve performs best across tasks, supporting the joint effectiveness of region stabilization, concept guidance, and pseudo-time evolution.

Table 4: Hyperparameter sensitivity analysis of TMEvolve with respect to the number of pseudo-time evolution steps Q and the number of adaptive regions K. Results are reported as mean $\pm$ std. The best result in each column is shown in bold.

Setting	Survival		Gene Prediction		Subtype Classification	
	BLCA C-index	LUAD C-index	BLCA Pearson	LUAD Pearson	BRACS AUC	BRACS ACC
$Q = 1$	0.6491±0.0562	0.6791±0.0483	0.7123±0.0273	0.6076±0.0330	0.8876±0.0066	0.6472±0.0312
$Q = 2$	0.6485±0.0328	0.6803±0.0532	0.7111±0.0270	0.6174±0.0354	**0.8940±0.0057**	**0.6544±0.0095**
$Q = 3$	**0.6589±0.0570**	**0.6952±0.0418**	**0.7158±0.0326**	**0.6226±0.0288**	0.8830±0.0124	0.6386±0.0322
$Q = 4$	0.6523±0.0337	0.6894±0.0419	0.7130±0.0271	0.6091±0.0388	0.8785±0.0127	0.6349±0.0341
$K = 2$	0.6417±0.0553	0.6614±0.0408	0.7038±0.0302	0.6118±0.0279	0.8894±0.0123	0.6433±0.0208
$K = 4$	0.6444±0.0362	0.6878±0.0449	0.7084±0.0277	0.6135±0.0317	**0.8940±0.0057**	**0.6544±0.0095**
$K = 6$	0.6429±0.0476	**0.6952±0.0418**	0.7107±0.0280	**0.6226±0.0288**	0.8892±0.0109	0.6489±0.0284
$K = 8$	**0.6589±0.0570**	0.6850±0.0464	**0.7158±0.0326**	0.6220±0.0260	0.8876±0.0081	0.6488±0.0363
$K = 10$	0.6414±0.0523	0.6830±0.0603	0.7087±0.0322	0.6201±0.0239	0.8867±0.0090	0.6415±0.0199

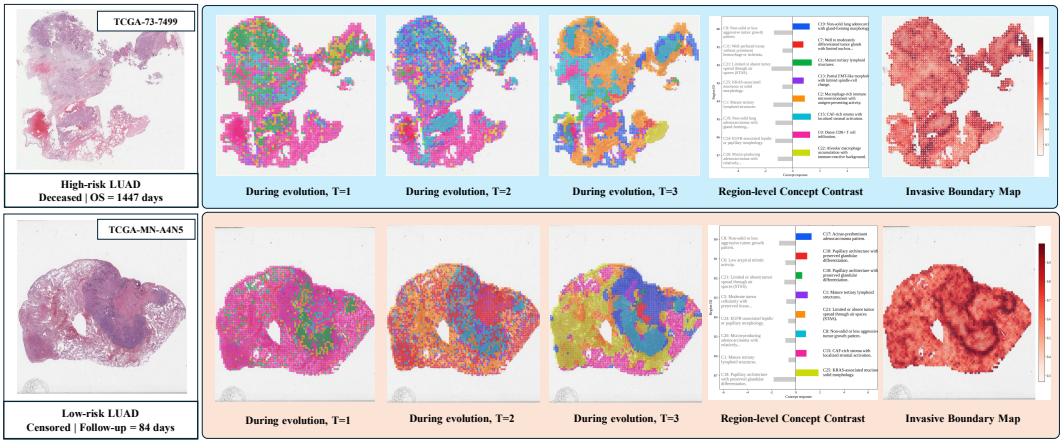

Figure 4: Microenvironment evolution in LUAD. For each case, we show the original WSI thumbnail, region partitions during evolution, the region-level concept contrast, and the boundary response map. The evolution process shows how TMEvolve progressively refines patch assignments into spatially coherent and concept-aligned microenvironment regions. The concept contrast and boundary response map provide semantic and interaction-level evidence for slide-level risk prediction.

Sensitivity analysis. As shown in Table 4, subtype classification works well with fewer pseudo-time steps and coarser regions, whereas survival and gene expression prediction benefit from more refined microenvironment modeling due to weaker and more spatially distributed signals. Moderate settings, such as $Q = 3$ and $K = 6$ or $K = 8$, generally perform better, while further increasing Q or K brings no consistent gain and may introduce redundant interactions or over-smoothing.

4.4 Interpretability Analysis

Beyond quantitative evaluation, we examine whether TMEvolve provides interpretable evidence for slide-level prediction. Figure 4 visualizes representative LUAD cases with different risk outcomes. Rather than relying only on isolated high-attention patches, TMEvolve reveals how tissue organization is refined during pseudo-time evolution, where scattered assignments are gradually consolidated into spatially coherent and concept-aligned regions. The region-level concept contrast links these regions to clinically meaningful semantics, while the boundary response map localizes heterogeneous tissue interfaces involved in cross-region interaction. In high-risk cases, adverse tumor-related concept responses are more prominent, whereas low-risk cases show relatively stronger immune- or glandular-related patterns. Together, these visualizations provide complementary evidence for interpreting slide-level risk prediction, with additional cases provided in Appendix F.

5 Conclusion

In this paper, we present TMEvolve, an evolution-aware MIL framework that models WSIs as dynamic tumor microenvironment fields. By forming adaptive concept-bearing regions and evolving them across heterogeneous tissue boundaries, TMEvolve captures region-level microenvironment interactions beyond static patch aggregation. Experiments on survival prediction, gene expression prediction, and subtype classification demonstrate its effectiveness, while ablation and visualization analyses validate the contributions of intra-region diffusion, concept guidance, pseudo-time evolution, and interpretable boundary-level evidence. These results highlight dynamic microenvironment modeling as a promising direction for weakly supervised computational pathology.

Acknowledgments

This work was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2025ZD0544802), the Shaanxi Province Key R&D Program (2025SF-YBXM-363 and 2024SF-GJHX-32), the Key Research and Development Program of Ningxia Hui Autonomous Region (2023BEG02023), the "Research on Key Technologies for Full-Chain Intelligent Pathological Diagnosis" project of the First Affiliated Hospital of Xi'an Jiaotong University (HX202440), the China Postdoctoral Science Foundation (2026M791685), the National Natural Science Foundation of China (62506291), the XJTU Research Fund for AI Science (2025YXYC004), the Xi'an Jiaotong University Kunpeng & Ascend Center of Cultivation, the Fundamental Research Funds for the Central Universities (Nos. xzy012026032 and xzy012026033), the Royal Society (RGS\R2\252688), and GE HealthCare.

Competing Interests. The authors declare no competing interests.

References

[1] Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C Obenauf, Helen Angell, Tessa Fredriksen, Lucie Lafontaine, Anne Berger, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, 39(4):782–795, 2013.

[2] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022.

[3] Josef Cersovsky, Sadegh Mohammadi, Dagmar Kainmueller, and Johannes Hoehne. Towards hierarchical regional transformer-based multiple instance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3952–3960, October 2023.

[4] Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In *International conference on machine learning*, pages 1407–1418. PMLR, 2021.

[5] Benjamin Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Xiaowen Dong, and Michael Bronstein. Beltrami flow and neural diffusion on graphs. *Advances in Neural Information Processing Systems*, 34:1594–1609, 2021.

[6] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE transactions on medical imaging*, 41(4):757–770, 2020.

[7] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021.

[8] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16144–16155, 2022.

[9] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer cell*, 40(8):865–878, 2022.

[10] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

[11] Karin E De Visser and Johanna A Joyce. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer cell*, 41(3):374–403, 2023.

[12] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, pages 1–13, 2025.

[13] Moshe Eliasof, Eldad Haber, and Eran Treister. Pde-gcn: Novel architectures for graph neural networks motivated by partial differential equations. *Advances in neural information processing systems*, 34:3836–3849, 2021.

[14] Wolf Herman Fridman, Franck Pagès, Catherine Sautès-Fridman, and Jérôme Galon. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, 12 (4):298–306, 2012.

[15] Peter Friedl and Stephanie Alexander. Cancer invasion and the microenvironment: plasticity and reciprocity. *Cell*, 147(5):992–1009, 2011.

[16] Jérôme Galon, Anne Costes, Fatima Sanchez-Cabo, Amos Kirilovsky, Bernhard Mlecnik, Christine Lagorce-Pagès, Marie Tosolini, Matthieu Camus, Anne Berger, Philippe Wind, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964, 2006.

[17] Zeyu Gao, Anyu Mao, Yuxing Dong, Hannah Clayton, Jialun Wu, Jiashuai Liu, ChunBao Wang, Kai He, Tieliang Gong, Chen Li, et al. Smmile enables accurate spatial quantification in digital pathology using multiple-instance learning. *Nature cancer*, 6(12):2025–2041, 2025.

[18] Zeyu Gao, Kai He, Weiheng Su, Xiaobo Pang, Ines P Machado, Mercedes Jimenez-Linan, Brian Rous, Chunbao Wang, Chengzu Li, William McGough, et al. Alpaca: Adapting llama for pathology context analysis to enable slide-level question answering. *Nature Communications*, 2026.

[19] Jiusong Ge, Yingkang Zhan, Wenjie Zhao, Di Zhang, Ke Wang, Jiashuai Liu, Chunze Yang, Chengzu Li, Jian Zhang, Yuxin Dong, et al. Thinking in scales: Accelerating gigapixel pathology image analysis via adaptive continuous reasoning. *arXiv preprint arXiv:2605.19491*, 2026.

[20] Jiaxiang Gou, Luping Ji, Pei Liu, and Mao Ye. Queryable prototype multiple instance learning with vision-language models for incremental whole slide image classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 3158–3166, 2025.

[21] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

[22] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36: 37995–38017, 2023.

[23] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018.

[24] Guillaume Jaume, Pushpak Pati, Valentin Anklin, Antonio Foncubierta, and Maria Gabrani. Histocartography: A toolkit for graph analytics in digital pathology. In *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 156 of *Proceedings of Machine Learning Research*, pages 117–128. PMLR, 27 Sep 2021.

[25] Johanna A Joyce and Douglas T Fearon. T cell exclusion, immune privilege, and the tumor microenvironment. *Science*, 348(6230):74–80, 2015.

[26] Raghu Kalluri. The biology and function of fibroblasts in cancer. *Nature Reviews Cancer*, 16 (9):582–598, 2016.

[27] Saarthak Kapse, Pushpak Pati, Srikar Yellapragada, Srijan Das, Rajarsi R Gupta, Joel Saltz, Dimitris Samaras, and Prateek Prasanna. Gecko: Gigapixel vision-concept contrastive pretraining in histopathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20020–20030, 2025.

[28] Bin Li, Yin Li, and Kevin W Eliceiri. 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*, pages 14318–14328, 2021.

[29] Pei Liu, Luping Ji, Jiaxiang Gou, Bo Fu, and Mao Ye. Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. *arXiv preprint arXiv:2409.09369*, 2024.

[30] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[31] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

[32] James D. Murray. *Mathematical Biology I: An Introduction*, volume 17 of *Interdisciplinary Applied Mathematics*. Springer, New York, 3 edition, 2002.

[33] Cancer Genome Atlas Research Network et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543, 2014.

[34] OpenAI. Introducing gpt-5, 2025.

[35] Daniela F Quail and Johanna A Joyce. Microenvironmental regulation of tumor progression and metastasis. *Nature medicine*, 19(11):1423–1437, 2013.

[36] Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, Baran Atli, Clemens Christian Vogel, Petra A Mercea, Romana Prihoda, Ellen Gelpi, Christine Haberler, Romana Höftberger, Johannes A Hainfellner, et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1):55, 2022.

[37] Daniel Shao, Richard J Chen, Andrew H Song, Joel Runevic, Ming Y Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer? *arXiv preprint arXiv:2506.09022*, 2025.

[38] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[39] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11343–11352, 2024.

[40] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[41] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

[42] Yuelin Wang, Kai Yi, Xinliang Liu, Yu Guang Wang, and Shi Jin. Acmp: Allen-cahn message passing with attractive and repulsive forces for graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2022.

[43] Junxian Wu, Xinyi Ke, Xiaoming Jiang, Huanwen Wu, Youyong Kong, and Lizhi Shao. Leveraging tumor heterogeneity: Heterogeneous graph representation learning for cancer survival prediction in whole slide images. *Advances in Neural Information Processing Systems*, 37: 64312–64337, 2024.

[44] Chunze Yang, Qidong Liu, Wenjie Zhao, Yue Tang, Jiusong Ge, Di Zhang, Jiashuai Liu, Lei Wu, Junbo Lu, Ni Zhang, et al. Pathnavigate: A training-free pathology agent with surprise-guided scan and shared slide memory for whole-slide image vqa. *arXiv preprint arXiv:2605.23559*, 2026.

[45] Di Zhang, Zhangpeng Gong, Xiaobo Pang, Jiashuai Liu, Junbo Lu, Hao Cui, Jiusong Ge, Zhi Zeng, Kai Yi, Yinghua Li, et al. Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis. *arXiv preprint arXiv:2602.21637*, 2026.

[46] Di Zhang, Jiashuai Liu, Youyuan Ma, Jiusong Ge, Zhi Zeng, Wenfang Sun, Qidong Liu, Kai He, Yefeng Zheng, Weimiao Yu, et al. Learnable frozen feature augmentation for few-shot biomarker prediction from pathology whole-slide images. *Bioinformatics*, 42(8):btag597, 2026.

[47] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[48] Yunkun Zhang, Jin Gao, Mu Zhou, Xiaosong Wang, Yu Qiao, Shaoting Zhang, and Dequan Wang. Text-guided foundation model adaptation for pathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 272–282. Springer, 2023.

[49] Weiqin Zhao, Ziyu Guo, Yinshuang Fan, Yuming Jiang, Maximus CF Yeung, and Lequan Yu. Aligning knowledge concepts to whole slide images for precise histopathology image analysis. *npj Digital Medicine*, 7(1):383, 2024.

[50] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.

Supplementary Material

Modeling Whole-Slide Images as Dynamic Tumor Microenvironment Fields

Additional method details, extended experiments, and discussions for TMEvolve

Contents of the Supplement

Section	Content	Page
A	**Graph discretization of the latent reaction–diffusion formulation**	15
B	**Additional experimental details**	15
B.1	Dataset details	15
B.2	Hyperparameter settings	16
C	**Computational cost analysis**	16
D	**Pathological concept construction**	17
E	**Additional gene expression prediction results**	17
F	**Additional visualization results**	23
F.1	Five-fold Kaplan–Meier survival curves	23
F.2	Additional interpretability visualization	24
G	**Limitations**	24
H	**Broader impact**	25

A Graph Discretization of the Latent Reaction–Diffusion Formulation

We provide a graph-discretized derivation of the latent reaction–diffusion formulation used in TMEvolve. The goal is not to solve a continuous biological PDE, but to instantiate its structural form as a learnable latent evolution process over WSI patch neighborhoods.

Given patch locations $\{u_i\}_{i=1}^N$, we construct a spatial neighborhood graph with node set $\mathcal{V}$ and neighborhood $\mathcal{N}(i)$ for each patch i. The continuous latent state $z(u_i, q)$ at location u_i is represented by the patch feature h_i^q. For a graph edge (i, j), the spatial gradient of the latent field can be approximated by the edge-wise state difference

$$\nabla z(u_i, q) \implies h_j^q - h_i^q. \tag{12}$$

A state-dependent diffusion coefficient $D(u, q)$ is then represented by a graph conductivity κ_{ij}^q, which controls the strength of information exchange along edge (i, j). Accordingly, the diffusion flux divergence at node i is discretized as

$$\nabla \cdot (D\nabla z)(u_i, q) \implies \sum_{j\in\mathcal{N}(i)} \kappa_{ij}^q \big(h_j^q - h_i^q\big). \tag{13}$$

This discretization can also be interpreted as the negative gradient of a graph Dirichlet energy:

$$\mathcal{E}_\kappa(H^q) = \frac{1}{2}\sum_{i=1}^N \sum_{j\in\mathcal{N}(i)} \kappa_{ij}^q \left\|h_i^q - h_j^q\right\|_2^2. \tag{14}$$

Taking a gradient descent step on this energy yields

$$-\frac{\partial \mathcal{E}_\kappa}{\partial h_i^q} \propto \sum_{j\in\mathcal{N}(i)} \kappa_{ij}^q \big(h_j^q - h_i^q\big), \tag{15}$$

which corresponds to a graph diffusion update that encourages neighboring states with large conductivity to become more coherent.

In TMEvolve, the generic conductivity κ_{ij}^q is instantiated in two complementary ways. For within-region diffusion, κ_{ij}^q is implemented as the region-membership overlap weight G_{ij}^q, which assigns large conductivity to neighboring patches with similar adaptive region memberships. For boundary evolution, κ_{ij}^q is implemented as the anisotropic boundary conductivity ρ_{ij}^q, which is activated by cross-region heterogeneity, concept-response contrast, and directional compatibility. Thus, the graph-discretized form of the latent reaction–diffusion equation becomes

$$h_i^{q+1} = h_i^q + \sum_{j\in\mathcal{N}(i)} G_{ij}^q W_h\big(h_j^q - h_i^q\big) + \sum_{j\in\mathcal{N}(i)} \rho_{ij}^q W_f\big(h_j^q - h_i^q\big) + \sum_{j\in\mathcal{N}(i)} \rho_{ij}^q W_c\big(c_j^q - c_i^q\big). \tag{16}$$

The first summation discretizes within-region diffusion, the second summation discretizes boundary-localized anisotropic flux, and the third summation implements a concept-conditioned semantic reaction source. All conductivities and concept assignments are recomputed at each pseudo-time step, making the evolution process state-dependent rather than a fixed graph smoothing operation.

B Additional Experimental Details

B.1 Dataset Details

Table 5 summarizes the datasets used in our experiments. The evaluation covers three types of weakly supervised WSI tasks, including survival prediction, gene expression prediction, and subtype classification. For survival prediction, we use three TCGA cohorts, including LUAD, BLCA, and BRCA, together with a private GBC cohort. For gene expression prediction, we use TCGA LUAD, BLCA, and BRCA with matched RNA-seq profiles, where 25 genes are processed for each cohort [33, 40]. For subtype classification, we evaluate TMEvolve on BRACS and EBRAINS under fine-grained and coarse-grained label settings [2, 36].

Table 5: Summary of datasets used in our experiments.

Dataset	Source	Task	Label / Target	Dataset size / processed samples
LUAD	TCGA	Survival prediction	Overall survival	478 cases / 541 diagnostic slides
BLCA	TCGA	Survival prediction	Overall survival	386 cases / 457 diagnostic slide files
BRCA	TCGA	Survival prediction	Overall survival	1,098 cases / 1,133 FFPE slides
GBC	Private cohort	Survival prediction	Overall survival	357 cases
LUAD	TCGA	Gene expression prediction	25 processed genes	628 clinical samples; 623 matched with RNA-seq
BLCA	TCGA	Gene expression prediction	25 processed genes	435 clinical samples; 430 matched with RNA-seq
BRCA	TCGA	Gene expression prediction	25 processed genes	1,222 clinical samples; 1,220 matched with RNA-seq
BRACS	Public dataset	Subtype classification	7 fine / 3 coarse classes	547 WSIs
EBRAINS	Digital Brain Tumour Atlas	Subtype classification	30 fine / 12 coarse classes	2,318 WSIs

B.2 Hyperparameter Settings

Table 6 summarizes the main implementation settings. All experiments use five-fold cross-validation. We use standard K-fold splitting for survival and gene expression prediction, and stratified K-fold splitting for subtype classification. The dataset- and task-specific choices of the number of regions K and evolution steps Q are summarized in Table 7.

Table 6: Main hyperparameter settings of TMEvolve.

Setting	Value
Patch encoder	CONCH v1.5
Text encoder	Qwen3-Embedding-8B
Hidden dimension d	512
Optimizer	Adam, lr $= 1 \times 10^{-4}$
Batch size	1 slide
Cross-validation	5-fold
Regions K	Dataset/task-specific
Evolution steps Q	Dataset/task-specific
Neighborhood size	8
Dropout	0.3
λ_{div}	0.1

Table 7: Dataset- and task-specific settings of the number of regions K and evolution steps Q.

Survival Prediction

Dataset	K	Q
LUAD	6	3
GBC	4	3
BLCA	8	3
BRCA	6	2

Subtype Classification

Dataset	Level	K	Q
BRACS	Fine	4	2
BRACS	Coarse	4	2
EBRAINS	Fine	4	2
EBRAINS	Coarse	4	2

Gene Expression Prediction

Dataset	K	Q
LUAD	6	3
BLCA	8	3
BRCA	8	3
–	–	–

C Computational Cost Analysis

We compare the computational cost of TMEvolve with representative WSI-level baselines on LUAD survival prediction. All methods in Table 8 are evaluated using pre-extracted patch-level WSI features with batch size 1 on a server with eight NVIDIA RTX 4090 GPUs, excluding the offline feature extraction cost. We focus on methods that process patch-level features during inference, including MIL baselines, survival-oriented models, and concept-guided WSI-level methods. Foundation-model baselines that directly operate on pre-computed slide-level embeddings are not included in this runtime comparison, since their inference mainly applies a lightweight prediction head to already aggregated slide representations.

As shown in Table 8, TMEvolve contains 5.18M trainable parameters and requires 6.48 ms per slide. Its parameter size is comparable to structured or concept-guided models such as ConcepPath [49] and

Table 8: Computational cost comparison on LUAD survival prediction. Runtime is measured using pre-extracted patch-level WSI features with batch size 1, excluding offline feature extraction.

Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	VLSA	GECKO	ConcepPath	ProtoSurv	CITE	QPMIL	TMEvolve
Trainable Params (M)	1.84	1.91	3.59	3.66	2.77	3.31	4.87	7.03	5.80	2.98	5.18
Time / slide (ms)	0.59	0.94	4.80	4.25	1.51	1.05	1.33	3.94	5.08	1.14	6.48

Table 9: Shared pathological concepts used across datasets.

Category	Concept descriptions
Immune	Dense $CD8^+$ T cell infiltration; mature tertiary lymphoid structures; macrophage-rich immune microenvironment with antigen-presenting activity; plasma cell-rich reactive stroma; immune-permissive tumor interface with lymphocyte access.
Tumor	Moderate tumor cellularity with preserved tissue organization; low atypical mitotic activity; well to moderately differentiated tumor glands with limited nuclear pleomorphism; non-solid or less aggressive tumor growth pattern.
Vascular	Limited tumor necrosis with mostly viable tissue; organized microvasculature and stromal remodeling; well-perfused tissue without prominent hemorrhage or ischemia.
Invasion	Focal tumor budding or limited local invasion; partial EMT-like morphology with limited spindle-cell change; mild desmoplastic stromal reaction; CAF-rich stroma with localized stromal activation.

CITE [48], while its runtime remains within the millisecond range. The additional cost mainly comes from dynamic region assignment, local coordination, and concept-guided boundary interaction over patch-level features. Overall, TMEvolve provides a reasonable trade-off between computational cost and explicit microenvironment modeling capacity.

Compared with lightweight MIL baselines, the additional runtime of TMEvolve mainly comes from its iterative evolution module, including region assignment, local coordination, and boundary interaction over patch-level features. Nevertheless, the model remains efficient for WSI-level prediction while enabling more explicit microenvironment modeling.

D Pathological Concept Construction

TMEvolve uses language-encoded pathological concepts as semantic priors to guide region-level microenvironment modeling [21, 31, 49, 27]. Instead of relying only on patch-level visual features, we introduce concept descriptions that summarize high-level tissue states related to tumor morphology, immune activity, stromal response, invasion, necrosis, vascular remodeling, and tissue organization [11, 35, 15, 14, 1, 25, 26]. These concepts are encoded by the text encoder and used as semantic guidance for region routing and concept-guided boundary interaction.

We organize the concept set into two parts: a shared concept bank and dataset-specific concept banks. As shown in Tables 9 and 10, the shared concept bank contains general histopathological and tumor microenvironment patterns that are broadly applicable across cancer types, while dataset-specific concept banks capture cohort-dependent pathological characteristics. The final concept set for each dataset is obtained by combining the shared concepts with its corresponding dataset-specific concepts. In the random concept ablation, we replace the language-encoded concept embeddings with random concept probes of the same dimensionality, thereby removing the pathological semantic prior while keeping the parameter budget comparable.

E Additional Gene Expression Prediction Results

To provide a more comprehensive evaluation of molecular prediction, we report full gene-level results on TCGA-LUAD, TCGA-BLCA, and TCGA-BRCA in Tables 11–13. The genes highlighted in the main paper were selected a priori based on clinical relevance and biological interpretability rather than model performance. For BLCA, the selected genes include immune and checkpoint-related markers such as CD8A, CTLA4, CXCL9, HAVCR2, PDCD1, and PTPRC, together with MKI67 for

Table 10: Dataset-specific pathological concepts used in TMEvolve.

Dataset	Dataset-specific concept descriptions
LUAD	Lepidic growth pattern; acinar-predominant adenocarcinoma pattern; papillary architecture with preserved glandular differentiation; non-solid lung adenocarcinoma with gland-forming morphology; mucin-producing adenocarcinoma with relatively differentiated morphology; limited or absent tumor spread through air spaces (STAS); alveolar macrophage accumulation with immune-reactive background; collapsed alveoli with mild fibrotic stroma; EGFR-associated lepidic or papillary morphology; KRAS-associated mucinous or solid morphology.
BRCA	Well-differentiated tubular carcinoma morphology; benign fibrocystic changes and organized lobules; low mitotic rate with uniform epithelial cells; invasive ductal carcinoma in stroma; tubule formation and glandular differentiation; high nuclear grade and pleomorphism; DCIS with comedo/cribriform patterns; HER2-enriched high-grade morphology; TNBC-like morphology with pushing borders; HR^+ morphology with lower grade; tumor-infiltrating lymphocytes (TILs); stromal fibrosis and desmoplasia; necrotic foci in high-grade carcinoma.
BLCA	Low-grade non-invasive papillary tumors; thickened but organized non-invasive urothelium; absence of muscularis propria invasion; papillary urothelial carcinoma; muscle-invasive urothelial carcinoma; high-grade urothelial carcinoma; CIS-like flat urothelial atypia; squamous differentiation with keratinization; nested or micropapillary variant morphology; inflammation and lymphocytic infiltration; muscle-invasive front with desmoplasia; necrotic and hemorrhagic aggressive regions; basal-like urothelial carcinoma.
GBC	Well-differentiated glands confined to the mucosa; reactive biliary epithelium without true dysplasia; intact muscle layer without invasive nests; adenocarcinoma with irregular glands; biliary-type infiltrative adenocarcinoma; poorly differentiated with solid growth; mucinous differentiation with mucin pools; perineural or lymphovascular invasion; dense desmoplastic reaction; chronic cholecystitis-like background; tumor necrosis and stromal remodeling; muscle or connective tissue invasion; immune infiltration at invasive margin.
BRACS	Normal breast tissue with organized terminal duct lobular units (TDLU); benign lesions such as usual ductal hyperplasia or fibroadenoma; atypical ductal hyperplasia with partial lumen involvement; flat epithelial atypia with uniform round nuclei; ductal carcinoma in situ with intact myoepithelial layer; invasive breast carcinoma with stromal infiltration; well-differentiated morphology with minimal nuclear atypia; high-grade invasive growth with significant pleomorphism and loss of architecture.
EBRAINS	Normal brain parenchyma with preserved neurons and glial cells; low-grade astrocytoma with mild hypercellularity and fibrillary background; oligodendroglioma morphology with classic uniform round nuclei and clear halos; high-grade glioma with prominent microvascular proliferation; glioblastoma morphology featuring pseudopalisading necrosis; dense infiltration of atypical glial cells with marked nuclear pleomorphism; reactive gliosis with evenly distributed gemistocytic astrocytes; tumor margins showing diffuse infiltration into normal white matter.

proliferation and ERBB2 for clinically relevant receptor signaling. For BRCA, we include immune-related genes CD8A, CTLA4, PDCD1, and PTPRC, the proliferation marker MKI67, epithelial and mesenchymal state markers CDH1 and VIM, and EGFR as a receptor signaling gene. For LUAD, we focus on immune and inflammatory signaling genes CD274, CD8A, CXCL10, CXCL9, and PDCD1, together with MKI67, EGFR, and MET to reflect proliferation and oncogenic signaling. These representative genes assess whether WSI morphology can capture immune activity, proliferation, phenotypic states, and clinically relevant signaling programs across cancer types.

Unlike the main paper, which focuses on these representative genes for compact presentation, the appendix includes all reported genes and all available evaluation metrics. We compare TMEvolve with representative MIL baselines, pathology foundation models, and concept-guided gene prediction methods [23, 28, 38, 39, 41, 37, 12, 49, 27]. Pearson correlation and Spearman correlation measure continuous gene expression prediction, while AUC evaluates the ability to distinguish high- and low-expression groups after median-based binarization. All values are multiplied by 100 for compact presentation.

Table 11: Full gene-level prediction results on TCGA-LUAD. Each cell reports mean±std over five folds. The best and second-best results for each gene-metric pair are shown in bold and underlined, respectively.

Gene	Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	TITAN	CHIEF	Feather	GECKO	ConcepPath	Ours
ACTA2	Pearson	44.8±6.6	21.2±11.4	36.0±9.9	41.1±6.9	**65.8±5.3**	<u>63.2±6.6</u>	54.7±5.2	45.9±10.0	55.1±8.8	55.7±10.9
	Spearman	41.5±9.7	18.2±13.0	32.7±12.4	38.6±6.7	**63.1±4.6**	<u>61.9±7.6</u>	52.6±5.7	45.1±11.3	54.2±9.8	53.4±14.2
	AUC	71.1±3.7	57.7±7.3	64.2±4.8	68.3±3.4	78.6±1.5	**79.9±4.2**	75.4±4.4	72.3±6.0	<u>78.8±5.3</u>	76.7±7.3

Continued on next page

Table 11: Full gene-level prediction results on TCGA-LUAD (continued).

Gene	Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	TITAN	CHIEF	Feather	GECKO	ConcepPath	Ours
BRAF	Pearson	35.7±8.3	26.5±13.5	29.9±8.9	34.6±10.3	42.6±2.6	44.9±7.9	41.4±4.1	44.1±12.1	43.7±10.9	**47.0±9.8**
	Spearman	35.7±5.7	24.4±12.1	29.4±6.9	33.0±6.2	38.3±3.0	42.3±6.0	36.2±3.3	40.8±9.3	40.9±9.0	**44.9±5.7**
	AUC	67.0±1.8	62.3±6.2	64.2±3.0	66.0±3.5	67.0±3.3	70.0±4.5	65.4±2.2	68.6±3.8	68.7±2.8	**71.8±2.3**
CD274	Pearson	59.5±5.4	56.1±5.3	56.1±8.1	50.9±10.0	63.4±6.5	**65.6±4.1**	58.4±6.7	62.7±7.0	64.4±6.9	63.7±6.7
	Spearman	57.1±5.5	54.2±5.5	51.0±10.5	49.0±10.3	60.2±5.6	**62.9±4.8**	57.0±6.9	60.5±6.9	62.2±5.7	60.9±4.3
	AUC	76.9±3.3	77.0±2.5	74.4±4.5	72.4±5.4	79.0±3.3	**81.4±3.5**	77.9±3.9	79.6±3.9	80.1±3.4	79.3±3.1
CD4	Pearson	57.8±7.2	50.4±8.9	45.2±12.7	48.9±8.8	64.9±7.1	**66.2±4.1**	61.8±3.6	61.3±4.0	62.8±4.4	65.5±3.1
	Spearman	53.2±6.9	43.3±14.6	40.2±10.2	45.6±11.3	60.5±9.1	**64.4±2.1**	58.8±2.4	61.3±5.6	61.0±5.1	62.3±6.7
	AUC	74.0±3.0	69.4±7.7	67.5±5.6	70.6±5.4	79.6±3.1	**80.6±1.3**	76.8±2.3	78.9±3.9	78.7±3.9	79.4±3.7
CD8A	Pearson	58.4±2.3	51.0±4.4	51.7±6.4	52.2±5.2	63.6±5.9	65.9±4.2	57.2±5.6	63.4±5.8	64.4±6.1	**68.4±5.0**
	Spearman	58.7±4.8	51.4±4.6	51.8±7.3	52.3±8.5	63.5±8.1	66.0±3.6	56.2±5.5	62.6±4.1	63.8±5.1	**66.2±3.6**
	AUC	77.3±2.7	73.7±5.9	73.3±4.7	73.9±4.8	80.5±5.9	80.5±2.5	77.5±2.8	78.8±4.6	79.3±5.3	**80.7±3.3**
CDH1	Pearson	33.3±10.0	30.0±9.7	29.7±12.5	25.7±11.1	35.8±4.5	**46.3±4.4**	34.1±8.3	36.9±10.9	39.3±15.1	43.5±3.3
	Spearman	26.1±6.8	23.7±4.9	25.5±9.7	19.3±7.5	35.0±3.6	**40.4±5.0**	30.0±8.7	34.1±7.2	34.4±11.4	38.4±4.1
	AUC	61.2±4.6	60.6±4.7	61.6±4.8	57.1±5.7	67.0±3.3	**68.4±4.9**	62.5±3.1	66.4±4.9	66.0±5.3	68.3±6.6
CTLA4	Pearson	58.9±4.8	52.3±2.3	53.5±7.5	57.0±7.0	**67.9±1.8**	67.3±1.8	63.4±3.5	62.5±4.8	65.6±4.6	65.4±2.4
	Spearman	56.3±5.8	50.5±5.4	52.7±7.1	54.1±5.6	66.3±2.2	65.8±2.2	60.1±5.3	61.7±5.4	**66.8±3.2**	64.6±3.3
	AUC	74.9±4.9	73.4±4.2	74.2±4.6	73.9±4.0	80.6±1.6	80.2±1.8	78.2±3.7	78.5±5.2	**81.6±3.8**	79.2±4.8
CXCL10	Pearson	60.6±4.4	56.1±8.3	56.9±6.0	53.1±6.9	61.9±5.3	64.8±5.1	59.7±5.1	65.3±4.2	**66.6±7.4**	65.9±4.5
	Spearman	60.5±6.0	55.3±8.1	57.0±6.5	52.2±8.9	62.3±4.6	65.0±5.1	59.2±4.4	63.9±4.6	**67.3±10.5**	65.5±5.7
	AUC	79.3±4.7	77.5±6.0	77.2±3.9	75.6±4.4	81.3±3.1	81.8±1.8	78.4±3.8	82.5±2.5	**83.3±7.1**	83.1±3.2
CXCL9	Pearson	60.7±5.4	53.1±4.4	49.2±12.7	53.5±10.4	63.2±4.3	62.1±4.5	56.5±4.7	61.9±5.3	66.1±3.7	**66.3±3.3**
	Spearman	59.5±5.0	51.9±5.9	49.7±12.6	54.6±11.3	61.4±5.9	63.0±4.5	56.5±5.7	61.9±5.5	**65.7±6.0**	64.7±5.7
	AUC	79.0±3.7	77.5±4.1	74.4±8.0	77.8±7.4	81.7±5.8	82.9±3.8	79.7±3.2	82.3±3.9	**83.6±4.8**	82.8±4.7
EGFR	Pearson	32.4±6.8	20.5±6.4	26.8±8.3	36.2±5.7	40.3±9.0	38.7±11.0	35.2±10.4	41.1±9.6	43.7±10.0	**44.3±8.6**
	Spearman	35.9±9.6	26.8±5.8	26.2±10.7	38.0±7.3	41.8±7.8	42.5±7.9	35.3±6.9	41.4±10.9	42.3±9.4	**42.8±9.1**
	AUC	68.3±3.4	63.5±6.6	63.1±7.3	67.7±3.4	71.1±4.7	**71.6±3.4**	66.8±3.1	71.0±4.7	71.1±4.4	69.7±3.9
ERBB2	Pearson	35.9±11.8	37.7±9.7	28.6±7.7	23.7±9.9	41.3±4.1	39.4±6.5	43.4±7.7	41.9±6.0	**49.0±12.8**	48.5±7.8
	Spearman	31.1±8.8	34.6±5.8	24.7±4.1	20.2±7.9	35.8±4.4	33.8±4.7	36.7±8.0	36.8±6.3	**42.8±10.2**	41.7±4.7
	AUC	65.3±5.0	65.1±4.9	63.0±2.1	58.5±2.1	66.8±3.9	66.2±4.2	67.3±4.3	67.0±5.2	**69.8±5.5**	69.0±2.2
FOXP3	Pearson	50.4±9.4	48.1±9.1	43.8±9.2	45.8±8.3	**58.8±5.6**	57.1±4.1	53.5±6.2	56.4±7.6	58.7±7.1	58.5±6.4
	Spearman	46.6±8.2	43.4±12.5	39.3±8.1	40.7±9.3	**57.5±7.1**	55.9±4.9	53.0±5.1	55.4±10.2	57.5±10.6	57.2±9.2
	AUC	72.0±5.7	71.3±5.8	67.2±3.9	68.3±5.7	77.9±3.5	76.8±3.1	75.9±2.8	77.1±4.2	**78.5±5.1**	78.5±4.7
HAVCR2	Pearson	61.9±4.9	56.5±5.9	58.3±5.7	59.6±3.3	66.1±4.2	67.5±3.8	63.5±2.4	64.0±5.0	67.9±7.0	**68.1±4.7**
	Spearman	56.7±7.8	53.5±7.8	55.1±5.6	55.2±3.6	60.4±3.6	64.9±4.4	59.3±2.5	62.4±5.9	**65.5±3.1**	64.6±3.9
	AUC	75.6±5.6	75.1±4.9	75.9±3.4	74.9±4.2	77.3±2.6	79.1±2.9	75.5±4.1	79.0±6.6	**80.9±3.3**	80.6±4.2
HIF1A	Pearson	51.0±6.1	48.9±11.3	47.0±6.7	38.7±13.9	52.5±6.5	49.1±4.6	46.6±5.1	53.1±7.6	55.8±5.9	**58.4±4.3**
	Spearman	50.0±7.4	47.8±14.0	46.4±10.0	37.8±14.1	50.5±8.0	47.6±7.1	45.5±7.2	50.8±8.7	52.4±8.3	**55.2±5.9**
	AUC	77.1±5.5	75.4±8.5	74.5±5.3	69.7±8.5	75.8±5.8	74.4±3.4	73.4±5.3	77.6±5.5	77.7±5.6	**78.5±3.3**
KEAP1	Pearson	4.3±13.6	-1.8±12.6	13.4±12.0	4.4±6.7	24.0±11.1	**25.3±14.3**	18.6±8.9	20.8±8.7	22.0±5.4	24.5±6.1
	Spearman	8.7±13.9	0.7±9.8	10.2±6.8	3.6±5.7	22.1±12.0	**26.7±11.2**	16.7±9.7	21.0±11.5	23.2±10.5	26.4±5.7
	AUC	55.9±6.7	52.4±5.4	53.0±2.7	51.6±4.4	60.6±7.4	63.0±5.9	58.0±3.8	61.1±4.8	61.9±4.8	**63.2±1.9**
KRAS	Pearson	17.4±12.4	13.6±7.8	14.5±11.0	9.0±13.1	**31.3±1.0**	30.1±6.7	27.6±4.5	26.0±10.0	29.9±11.5	31.1±13.1
	Spearman	15.2±10.3	13.1±7.2	9.7±10.4	7.1±14.1	**31.5±5.6**	30.2±3.7	25.3±4.4	24.5±7.9	22.4±11.2	24.8±13.3
	AUC	56.5±2.5	57.9±6.5	54.9±4.1	52.7±9.1	**63.2±4.6**	63.1±2.8	61.7±3.1	61.8±5.0	60.2±5.3	62.8±7.8
LAG3	Pearson	57.8±5.7	53.2±6.5	52.0±8.8	48.0±8.5	62.0±5.8	62.5±5.2	55.3±6.6	62.8±4.2	65.0±7.8	**65.0±5.5**
	Spearman	56.4±7.3	50.8±8.5	49.7±8.2	47.3±9.8	60.6±5.9	62.4±6.2	53.9±6.9	61.6±6.6	64.1±10.2	**64.5±6.5**
	AUC	76.7±4.9	74.1±6.3	73.2±4.7	72.5±7.2	80.8±4.4	80.8±3.9	74.9±3.7	81.6±4.9	81.1±7.2	**81.7±5.1**
MET	Pearson	49.8±2.2	44.5±8.0	43.1±6.7	46.0±7.1	53.8±5.1	49.5±4.4	42.9±7.2	**59.7±4.4**	58.2±2.3	57.2±5.6
	Spearman	48.6±4.1	42.3±9.4	41.0±8.1	41.4±7.2	50.4±4.6	47.7±5.6	35.7±8.8	**57.1±6.1**	54.4±5.8	54.0±7.6
	AUC	73.2±2.8	71.2±6.3	68.7±5.1	68.1±2.9	74.7±3.5	72.1±2.7	65.4±4.9	**77.7±3.0**	76.5±4.3	75.4±3.3

Continued on next page

Table 11: Full gene-level prediction results on TCGA-LUAD (continued).

Gene	Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	TITAN	CHIEF	Feather	GECKO	ConcepPath	Ours
MKI67	Pearson	63.6±1.8	59.9±3.3	56.3±3.2	55.9±3.0	**72.3±2.7**	71.6±2.5	70.2±4.7	65.1±4.7	65.5±6.3	67.0±4.0
	Spearman	63.3±2.2	60.6±4.6	54.4±4.1	54.9±2.3	**72.2±2.7**	71.5±3.2	70.2±4.6	66.0±5.1	65.7±5.5	67.1±3.7
	AUC	81.8±2.7	80.1±4.9	76.0±3.0	78.0±2.7	87.0±3.0	85.9±2.3	**87.2±4.7**	82.6±4.8	82.3±4.7	83.1±4.5
PDCD1	Pearson	61.9±6.1	58.2±7.2	54.0±10.2	59.5±3.5	65.6±4.9	63.3±3.5	60.1±4.3	65.8±4.5	**68.9±7.1**	65.3±6.7
	Spearman	61.5±7.4	54.6±11.2	53.3±11.3	55.7±6.3	64.7±6.4	62.0±4.6	58.1±4.1	64.6±5.4	**67.6±8.6**	63.8±8.4
	AUC	80.0±5.1	76.4±5.4	75.6±6.0	76.8±6.0	82.0±4.6	80.0±4.1	77.5±2.8	**82.7±2.3**	82.7±4.7	81.8±5.1
PTPRC	Pearson	57.7±6.9	49.3±6.0	49.2±15.8	51.6±13.8	65.5±7.2	**66.4±4.0**	63.3±3.3	59.7±1.7	62.1±4.1	64.1±1.5
	Spearman	57.1±6.5	45.7±10.3	49.7±12.9	50.7±14.2	63.4±5.8	**66.1±2.8**	62.6±3.3	59.6±2.6	62.6±4.1	63.0±2.9
	AUC	79.0±3.8	72.2±6.6	74.5±5.6	75.0±7.7	**80.8±5.6**	80.4±4.1	79.4±3.8	80.0±2.5	80.7±2.4	80.6±3.4
STK11	Pearson	25.7±8.2	22.9±10.4	23.3±5.3	17.4±15.1	39.9±7.8	37.2±6.2	37.0±3.9	38.1±5.0	39.7±6.4	**42.0±9.3**
	Spearman	22.5±9.1	18.9±8.4	19.2±6.8	14.3±15.7	**36.7±8.7**	35.2±8.6	33.0±8.7	33.2±6.5	35.6±5.2	35.9±10.4
	AUC	58.8±4.8	58.7±6.3	58.6±5.3	54.8±9.9	**68.3±7.1**	67.8±7.3	64.8±6.5	65.8±6.5	66.7±4.3	66.0±5.6
TP53	Pearson	21.6±10.4	21.2±12.0	24.8±11.4	16.6±13.1	17.1±10.1	27.8±7.7	19.5±13.7	25.8±6.4	**30.7±8.7**	23.4±8.8
	Spearman	19.4±11.5	18.6±13.1	22.4±8.5	15.7±14.1	19.1±9.4	**29.6±9.3**	19.2±12.8	23.4±7.8	27.0±11.7	23.0±12.1
	AUC	61.3±6.2	60.8±7.6	62.5±4.7	59.9±5.7	61.8±5.1	**67.4±6.3**	61.4±7.2	63.6±2.9	64.3±4.9	63.3±6.8
VIM	Pearson	50.2±9.5	38.6±6.6	41.4±3.9	40.7±11.7	56.8±5.1	57.9±4.3	47.8±8.0	56.5±6.4	59.8±7.3	**61.5±6.0**
	Spearman	48.2±11.6	34.6±7.3	41.1±6.6	38.3±10.5	55.2±5.8	57.1±5.7	47.8±5.5	55.2±6.9	58.5±7.2	**59.0±9.0**
	AUC	73.0±7.4	64.5±3.9	68.5±4.7	66.0±4.7	76.4±2.9	77.7±3.4	72.6±3.7	76.4±5.0	78.4±5.3	**79.0±6.7**

Table 12: Full gene-level prediction results on TCGA-BLCA. Each cell reports mean±std over five folds. The best and second-best results for each gene-metric pair are shown in bold and underlined, respectively.

Gene	Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	TITAN	CHIEF	Feather	GECKO	ConcepPath	Ours
ACTA2	Pearson	59.0±6.5	50.1±3.0	52.1±6.9	57.2±10.1	46.0±14.3	72.1±2.3	67.5±2.6	71.2±1.6	**73.5±2.2**	71.7±1.4
	Spearman	59.2±8.0	49.4±4.4	53.4±5.2	57.4±11.1	45.1±15.8	72.4±3.8	67.8±2.6	72.0±2.4	**72.6±3.9**	71.9±1.9
	AUC	77.6±8.1	73.0±3.2	74.5±5.1	76.7±10.4	72.9±12.6	84.8±4.7	83.1±3.3	84.9±3.9	**85.1±3.2**	85.1±3.0
BRAF	Pearson	28.8±14.4	33.8±16.3	32.7±13.3	26.2±11.4	24.9±18.2	45.6±5.7	43.6±9.0	47.9±6.3	47.8±7.2	**49.2±6.2**
	Spearman	26.7±8.6	27.1±13.0	28.7±11.3	22.1±11.4	23.0±16.9	41.8±6.0	38.0±10.1	43.3±6.1	43.5±5.6	**46.3±5.4**
	AUC	64.5±5.0	60.2±5.5	63.2±6.9	61.5±3.8	60.1±9.3	69.6±4.9	68.5±5.1	71.8±3.6	72.3±2.1	**73.3±3.9**
CD274	Pearson	60.3±9.3	53.7±6.3	54.1±5.4	41.2±13.7	37.5±14.5	64.4±6.3	57.9±5.9	64.9±7.4	66.0±7.4	**66.6±4.8**
	Spearman	61.4±8.8	52.2±9.3	54.8±5.9	40.0±17.9	37.0±19.0	64.0±4.7	56.3±4.9	64.2±8.0	65.2±8.2	**66.0±5.4**
	AUC	81.8±6.2	77.2±5.5	78.7±3.4	70.4±8.9	65.5±12.4	82.3±3.8	78.9±2.0	82.1±6.0	83.1±6.5	**83.4±4.6**
CD4	Pearson	60.8±5.9	52.2±6.9	53.7±8.1	54.1±5.4	44.1±12.8	62.8±1.4	62.3±6.4	69.7±5.8	**71.0±8.2**	70.9±6.6
	Spearman	61.4±3.6	50.3±10.2	54.7±6.2	55.6±7.6	45.4±12.5	62.6±2.2	62.3±7.5	69.6±6.8	70.5±8.3	**70.6±7.1**
	AUC	79.8±1.8	75.8±6.2	78.4±1.7	78.1±3.0	73.0±7.7	81.7±2.6	82.1±3.8	85.3±4.7	**86.0±4.5**	85.8±4.3
CD8A	Pearson	65.2±8.7	57.7±8.2	58.5±8.7	55.3±7.7	56.6±4.2	66.5±6.4	63.3±5.4	71.7±3.2	71.2±3.3	**73.5±5.6**
	Spearman	62.4±8.9	53.2±9.4	54.5±9.9	52.0±8.5	54.0±1.9	63.4±5.2	59.8±5.1	70.4±2.1	69.0±3.2	**71.6±4.6**
	AUC	79.7±4.0	75.1±4.4	74.4±5.2	73.4±5.7	77.7±3.5	80.1±4.3	77.7±3.4	**83.3±2.4**	81.3±1.6	83.0±3.2
CDH1	Pearson	51.9±6.9	33.2±9.5	32.3±17.2	41.4±17.5	11.0±20.5	52.6±4.5	50.2±11.8	58.0±8.4	54.4±7.6	**61.0±6.9**
	Spearman	32.4±14.9	25.4±16.8	23.1±10.5	21.8±18.6	4.1±17.6	38.5±10.6	38.4±7.2	**41.9±11.0**	39.5±10.8	39.5±12.3
	AUC	63.3±9.9	62.5±11.7	60.4±6.0	59.5±10.3	49.2±9.1	**68.4±6.8**	66.8±4.7	67.8±7.4	67.6±8.1	65.8±6.7
CTLA4	Pearson	65.1±4.1	61.6±6.4	60.4±5.2	57.7±6.8	54.8±3.2	66.0±7.1	66.6±6.1	70.1±5.3	68.9±2.4	**72.7±4.9**
	Spearman	65.6±4.6	60.5±8.4	58.5±5.7	58.1±6.7	55.4±6.9	63.8±6.4	65.4±8.5	68.6±6.2	66.7±3.4	**70.6±6.6**
	AUC	82.1±2.6	79.1±3.3	77.1±7.0	78.7±5.6	78.2±4.8	80.1±5.0	81.1±6.2	**83.5±5.9**	82.2±4.4	83.5±5.9
CXCL10	Pearson	63.8±2.8	59.8±6.3	61.4±2.1	55.6±5.1	44.8±18.4	66.1±7.1	65.5±5.7	70.3±3.2	69.3±3.3	**71.6±4.8**
	Spearman	62.5±4.4	56.2±9.2	59.6±4.0	53.3±5.8	44.5±23.8	65.4±8.3	64.8±7.7	70.3±5.0	68.8±4.1	**71.1±5.6**
	AUC	80.8±5.1	75.3±6.0	80.1±3.2	76.0±3.2	68.4±12.4	82.6±4.9	80.3±5.2	**85.2±5.0**	83.2±4.4	85.1±4.8
CXCL9	Pearson	62.9±3.4	55.8±8.0	60.6±2.5	56.3±9.8	58.0±7.8	66.5±7.9	65.1±6.4	72.4±2.4	72.4±1.6	**74.9±1.5**
	Spearman	62.5±3.1	54.1±9.3	59.5±2.2	54.4±9.7	53.2±10.2	66.2±7.1	64.3±8.3	72.6±3.8	71.6±1.7	**74.0±2.5**
	AUC	80.1±3.6	75.1±4.4	76.6±4.4	75.4±5.4	74.1±8.2	83.3±3.5	82.3±6.0	85.6±4.8	85.7±2.5	**86.8±2.6**

Continued on next page

Table 12: Full gene-level prediction results on TCGA-BLCA (continued).

Gene	Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	TITAN	CHIEF	Feather	GECKO	ConcepPath	Ours
EGFR	Pearson	40.9±8.0	36.8±7.1	33.7±14.4	35.9±4.9	**54.2±17.8**	46.1±9.0	40.5±6.3	45.5±8.1	43.5±9.7	46.1±6.2
	Spearman	44.7±7.3	41.5±4.7	33.2±14.1	37.9±6.6	**51.8±18.0**	49.0±8.6	41.5±6.1	47.3±6.2	45.0±6.6	48.1±5.7
	AUC	72.6±2.7	70.5±1.6	67.0±5.3	68.3±3.3	**76.9±11.8**	74.5±3.7	71.2±2.6	74.3±4.0	73.0±1.1	74.0±3.7
ERBB2	Pearson	55.1±2.2	50.8±8.0	41.4±8.3	41.8±7.9	48.4±18.2	49.4±5.4	52.5±7.7	60.0±4.7	59.8±7.7	**62.0±6.2**
	Spearman	53.8±7.9	52.2±8.0	42.3±11.7	46.0±9.8	48.3±17.3	55.3±3.0	54.7±6.5	61.7±5.8	61.5±5.5	**62.9±7.7**
	AUC	75.8±8.1	77.5±4.4	70.8±7.0	74.4±6.3	71.7±10.8	78.0±1.4	76.6±4.3	80.4±4.0	**81.1±3.0**	80.4±4.2
FOXP3	Pearson	54.2±7.3	55.8±7.3	54.7±6.8	49.0±8.0	43.7±17.2	59.5±3.3	58.3±9.0	62.7±3.5	63.5±3.0	**64.1±4.8**
	Spearman	54.0±9.0	55.6±8.4	54.0±7.1	48.5±9.6	41.7±23.1	58.9±4.6	57.4±9.6	62.4±3.4	63.4±3.4	**63.9±5.0**
	AUC	77.7±4.2	78.3±3.4	79.0±5.0	75.2±5.8	72.6±17.5	79.6±2.8	78.7±5.0	80.9±3.3	**81.4±2.6**	81.0±2.6
HAVCR2	Pearson	71.8±2.4	62.6±5.2	65.7±7.9	69.8±5.4	59.4±13.8	76.8±3.4	73.0±4.7	79.4±3.6	77.8±3.1	**80.2±3.1**
	Spearman	72.8±1.0	60.7±5.3	66.6±7.2	67.9±6.0	54.7±14.7	77.4±2.4	73.3±6.1	80.3±4.3	78.5±3.6	**80.3±4.4**
	AUC	86.8±2.3	79.8±2.2	83.9±5.2	83.8±5.8	78.5±8.4	89.3±1.4	86.2±3.7	**89.5±2.8**	88.4±1.9	89.3±3.1
HIF1A	Pearson	42.2±11.6	35.0±5.9	37.1±11.4	39.9±10.0	39.9±8.2	45.3±5.7	42.7±9.2	48.4±10.0	**49.4±7.4**	47.8±10.0
	Spearman	42.2±14.2	37.7±6.2	37.5±14.7	40.9±9.3	40.0±6.0	44.3±7.6	42.7±10.1	**48.9±10.7**	48.6±7.9	48.3±12.0
	AUC	71.9±9.9	69.0±3.6	68.9±9.1	71.9±6.0	69.8±7.1	70.1±3.8	71.7±6.6	74.3±3.3	74.3±3.9	**74.4±5.4**
KEAP1	Pearson	34.1±10.4	29.0±11.4	24.4±13.7	14.4±7.0	28.2±20.5	26.1±5.1	26.1±3.7	32.9±8.9	31.5±5.6	**34.5±4.9**
	Spearman	33.1±7.1	25.6±12.3	24.2±8.4	15.2±5.4	25.8±22.5	26.2±5.1	25.4±7.6	34.7±7.2	33.4±8.3	**36.3±3.8**
	AUC	65.4±2.8	59.7±7.1	61.8±3.5	58.2±4.1	66.4±6.5	62.0±3.6	61.3±5.1	**68.7±5.6**	66.2±7.5	67.9±4.2
KRAS	Pearson	20.4±8.0	14.5±9.3	15.4±3.8	19.6±10.3	9.2±19.9	27.8±6.7	22.1±10.7	34.7±7.8	**35.2±10.8**	31.9±8.6
	Spearman	20.3±8.0	15.3±8.6	17.8±3.1	20.4±7.7	11.7±24.0	33.3±7.0	23.6±8.3	**38.3±8.6**	38.1±11.8	34.8±9.7
	AUC	59.2±6.9	57.1±5.1	60.8±2.2	60.2±3.3	55.2±17.6	66.3±4.2	60.0±5.1	**69.1±6.1**	68.7±6.2	66.4±5.0
LAG3	Pearson	70.4±2.1	64.9±5.9	62.2±7.1	60.1±6.0	46.6±11.1	69.5±6.5	70.5±3.8	76.0±1.1	73.7±2.7	**77.5±4.1**
	Spearman	67.5±2.2	62.5±7.8	59.6±4.6	56.1±7.4	45.2±13.4	69.5±7.0	68.9±5.0	75.0±2.2	71.5±4.5	**76.0±4.5**
	AUC	82.7±3.0	81.1±3.8	79.3±4.6	74.9±3.6	73.0±7.0	84.9±4.9	85.3±1.8	**88.0±2.6**	84.9±2.7	87.6±1.9
MET	Pearson	33.9±7.6	30.0±6.9	31.7±11.8	22.1±13.6	36.7±14.9	41.5±6.7	36.5±6.6	**46.2±7.5**	45.1±2.9	44.6±6.5
	Spearman	37.8±12.6	29.4±8.4	31.9±14.0	24.7±17.0	38.4±16.3	41.9±9.9	34.9±6.4	**46.5±4.7**	45.7±2.6	45.9±3.7
	AUC	71.8±7.8	67.4±5.9	69.2±7.9	63.3±9.5	69.1±10.1	71.1±3.8	70.0±4.1	**76.5±3.6**	74.2±1.5	75.0±3.1
MKI67	Pearson	54.2±3.3	51.4±6.0	35.2±6.7	38.2±8.1	41.3±11.8	50.1±9.7	48.2±4.2	57.0±4.5	58.1±5.3	**60.5±4.1**
	Spearman	44.9±9.1	42.8±8.3	29.2±4.9	33.3±11.5	39.7±8.7	47.3±10.7	43.4±9.9	52.5±6.5	53.6±9.7	**53.7±8.6**
	AUC	69.6±5.8	68.5±4.1	62.1±4.0	63.9±5.6	65.0±10.8	71.5±6.7	68.9±5.2	72.6±5.0	72.7±6.8	**73.1±6.5**
PDCD1	Pearson	64.7±4.7	60.1±5.6	61.0±3.5	59.3±5.4	47.1±11.5	63.5±5.3	63.7±6.7	70.4±5.0	69.9±2.8	**72.9±6.2**
	Spearman	63.5±3.6	58.5±6.1	59.3±5.7	57.2±4.6	50.9±11.8	63.3±5.4	62.7±8.1	69.1±4.5	69.0±2.1	**72.0±6.1**
	AUC	82.2±2.9	78.7±3.9	79.9±4.3	79.6±1.9	75.2±7.3	80.5±4.4	80.2±5.1	84.6±4.7	83.1±3.1	**85.0±4.4**
PTPRC	Pearson	65.7±3.0	61.6±7.7	60.6±2.9	59.5±4.7	46.0±10.9	68.3±4.5	69.4±7.8	73.7±5.5	73.0±5.3	**75.8±5.3**
	Spearman	66.9±5.3	61.4±9.3	63.2±2.5	60.8±5.7	44.3±10.4	69.0±5.9	70.2±8.6	73.9±5.8	73.5±5.5	**76.3±5.2**
	AUC	81.4±4.8	78.6±6.2	81.2±3.4	80.3±3.8	70.8±10.7	85.5±4.4	84.8±5.8	85.7±4.9	85.2±5.3	**87.5±5.0**
STK11	Pearson	27.7±8.3	20.1±4.5	26.0±6.9	5.9±15.7	31.2±9.9	23.4±12.1	20.5±9.1	28.1±11.0	30.1±4.9	**33.5±7.2**
	Spearman	31.2±11.8	20.6±6.0	25.3±14.9	10.9±8.8	**37.8±13.6**	24.0±13.7	24.0±10.4	30.5±10.4	32.3±5.9	36.3±7.9
	AUC	66.1±8.9	60.5±3.9	62.1±9.4	56.0±1.6	68.0±9.0	60.4±6.3	61.8±5.0	66.2±8.0	65.8±5.6	**69.8±7.3**
TP53	Pearson	6.7±7.1	7.8±10.0	5.1±14.2	11.0±10.7	17.9±34.1	13.0±5.4	**23.6±10.2**	16.8±13.1	19.0±8.6	22.1±12.8
	Spearman	7.7±1.4	7.9±6.5	3.9±15.5	12.3±7.4	14.5±31.3	16.3±5.2	**23.9±5.1**	20.1±9.2	19.5±5.9	23.3±9.7
	AUC	53.5±4.1	53.0±3.6	52.8±8.0	57.3±4.4	57.8±17.6	58.8±3.0	61.7±4.0	61.2±3.9	60.1±2.0	**62.7±3.8**
VIM	Pearson	59.9±8.4	54.7±6.2	52.0±7.6	53.1±7.4	33.7±20.7	67.9±5.9	64.0±2.2	**68.0±6.1**	67.4±5.9	67.5±4.9
	Spearman	61.9±6.9	55.5±6.2	54.8±7.0	52.4±10.5	35.6±21.2	**68.9±6.7**	65.5±3.7	67.8±8.0	68.1±7.0	68.2±6.6
	AUC	82.8±2.3	78.4±6.2	78.8±2.5	75.5±6.8	70.4±12.6	85.7±4.0	84.9±2.3	85.5±4.1	**85.7±3.4**	85.7±3.2

Table 13: Full gene-level prediction results on TCGA-BRCA. Each cell reports mean±std over five folds. The best and second-best results for each gene-metric pair are shown in bold and underlined, respectively.

Gene	Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	TITAN	CHIEF	Feather	GECKO	ConcepPath	Ours
ACTA2	Pearson	42.2±5.3	40.6±10.8	40.2±8.6	39.9±8.1	56.3±6.6	54.3±5.5	50.3±5.2	58.0±4.8	**59.2±7.0**	58.1±6.4
	Spearman	40.2±3.5	37.7±11.2	39.4±7.4	37.5±9.0	54.5±6.2	54.4±3.9	48.7±6.3	58.4±4.0	**58.7±7.2**	57.6±7.0
	AUC	70.1±2.3	68.8±6.2	70.0±3.9	68.9±5.1	77.2±3.3	77.4±3.6	73.8±3.8	79.5±3.0	**79.7±3.7**	79.3±4.3
BRAF	Pearson	20.5±5.1	15.6±3.9	18.8±5.1	21.2±3.3	29.9±8.0	29.9±8.8	30.5±5.3	34.0±7.7	**36.9±6.4**	35.8±6.5
	Spearman	19.7±4.9	13.7±4.7	17.5±7.7	20.7±3.5	28.5±7.5	30.0±9.5	29.8±5.6	33.9±8.4	**36.9±5.5**	35.1±6.5
	AUC	59.8±2.6	56.0±2.6	58.5±4.7	61.7±1.6	64.7±4.6	66.1±4.2	64.8±3.8	66.4±3.7	**68.9±4.1**	68.5±2.6
CD274	Pearson	45.1±7.5	41.9±5.8	41.9±7.6	41.4±8.3	58.3±4.8	54.2±3.6	54.5±6.5	59.5±2.1	**60.5±3.1**	59.0±2.5
	Spearman	44.7±4.3	41.3±5.6	40.2±8.5	41.1±5.7	56.1±5.5	51.5±5.1	52.7±5.8	57.9±3.6	**58.1±3.8**	57.2±2.6
	AUC	71.3±2.9	70.0±3.9	69.3±6.1	70.4±3.2	77.0±3.5	74.0±3.6	75.6±2.9	**77.8±3.0**	77.6±2.8	76.9±1.7
CD4	Pearson	58.3±2.3	51.6±6.0	53.3±4.6	53.9±5.2	67.0±3.5	65.9±2.9	62.9±3.8	70.4±3.6	**71.0±2.9**	70.2±3.8
	Spearman	56.8±4.4	51.2±9.7	50.9±4.3	51.0±6.2	65.8±4.7	64.6±3.5	60.9±5.1	69.2±4.1	**70.0±4.3**	69.7±4.8
	AUC	77.7±4.8	75.3±7.1	75.5±3.1	73.9±3.1	82.4±4.8	81.5±3.3	79.6±4.3	83.7±3.9	84.2±4.6	**84.6±4.3**
CD8A	Pearson	59.4±5.5	55.8±8.7	56.1±6.2	52.7±10.4	70.7±4.0	67.3±2.5	63.6±5.1	72.0±4.3	72.6±3.7	**73.9±3.0**
	Spearman	58.5±6.6	54.4±10.2	54.3±8.5	50.2±12.0	69.1±4.9	65.8±3.1	60.5±4.7	70.0±4.2	70.4±4.2	**71.8±3.1**
	AUC	78.9±3.1	77.0±4.6	76.7±4.7	74.4±5.8	83.9±2.1	81.5±1.8	79.9±1.0	84.3±3.0	84.9±2.2	**85.5±1.8**
CDH1	Pearson	57.9±6.4	54.4±7.1	52.0±5.8	51.5±5.8	59.4±8.4	61.2±4.6	60.5±6.8	63.8±5.3	**65.7±4.7**	64.5±5.5
	Spearman	47.7±3.9	44.3±6.3	45.1±5.9	42.1±6.6	51.1±9.0	52.6±5.9	50.9±6.5	55.0±4.0	**56.5±7.5**	56.4±6.9
	AUC	70.7±2.5	69.0±3.9	69.8±4.1	69.0±4.6	73.7±5.7	74.6±4.2	73.1±4.6	75.5±3.7	**76.5±5.4**	76.3±5.2
CTLA4	Pearson	65.2±3.0	60.8±1.0	61.9±4.2	61.8±2.7	73.6±1.5	67.7±3.1	65.4±2.6	71.6±2.7	72.6±2.8	**74.2±1.5**
	Spearman	64.0±6.9	59.4±4.3	59.1±4.8	60.5±5.9	72.3±3.3	65.8±4.2	62.1±3.3	69.3±3.9	70.7±3.4	**73.0±2.7**
	AUC	81.4±4.8	78.8±3.5	78.1±2.9	79.0±4.5	86.1±3.1	83.1±2.6	80.8±3.0	84.0±2.9	84.8±2.7	**86.4±2.3**
CXCL10	Pearson	61.1±2.4	57.4±4.6	56.8±2.2	56.4±4.9	70.5±2.1	65.9±4.9	64.4±2.3	69.9±3.5	70.4±3.6	**70.9±2.3**
	Spearman	58.7±3.5	56.2±7.5	55.0±1.6	54.6±6.9	70.4±2.8	64.3±6.0	62.1±4.1	69.6±4.6	69.6±4.5	**70.8±2.9**
	AUC	77.7±2.7	77.0±4.2	75.4±1.1	75.6±5.6	84.5±1.4	80.8±2.2	79.7±2.2	84.3±2.1	84.4±2.6	**84.7±2.0**
CXCL9	Pearson	62.0±5.1	58.0±7.0	61.3±6.5	61.4±9.0	70.2±3.2	66.2±3.5	62.8±2.3	70.6±3.8	71.5±2.7	**72.0±2.7**
	Spearman	61.0±6.2	57.1±9.9	59.8±7.6	60.5±10.1	69.2±3.8	64.0±3.4	60.3±2.2	69.0±4.1	70.0±3.0	**70.8±2.4**
	AUC	80.4±3.1	77.3±5.5	79.0±2.8	79.5±6.0	84.4±2.5	81.6±1.5	79.8±2.1	84.5±2.0	84.8±1.5	**85.4±1.5**
EGFR	Pearson	52.4±6.5	47.9±9.9	43.3±9.6	49.0±6.3	57.8±4.4	52.4±7.4	52.3±2.9	60.4±5.4	**61.5±4.3**	59.5±6.5
	Spearman	54.0±6.3	50.4±9.5	44.5±8.7	50.1±7.0	57.9±4.3	50.5±6.9	51.6±2.7	61.2±3.6	**61.7±4.2**	60.3±6.0
	AUC	77.6±4.3	76.2±4.6	71.7±4.6	75.2±5.2	79.5±3.5	75.1±3.9	76.8±2.6	81.5±2.8	**81.5±3.1**	80.8±3.7
ERBB2	Pearson	26.5±7.9	25.9±8.4	20.1±7.6	22.7±8.0	35.6±7.4	36.2±7.1	31.4±9.5	38.2±6.0	**42.1±4.8**	41.2±6.5
	Spearman	18.2±8.1	19.8±10.8	13.9±8.8	17.4±12.4	32.1±6.4	30.4±6.8	29.5±5.6	33.3±7.0	**37.2±4.0**	35.6±3.5
	AUC	56.0±4.6	57.1±7.1	54.3±4.8	55.7±7.0	62.3±3.5	62.4±5.0	62.2±1.9	64.6±2.7	**66.4±1.0**	65.0±0.7
FOXP3	Pearson	59.9±2.9	54.3±5.0	57.3±5.7	57.1±6.6	69.8±2.9	63.3±4.1	61.3±3.9	68.1±1.9	69.0±5.3	**70.2±2.4**
	Spearman	59.7±3.6	54.4±7.6	56.8±7.3	55.3±8.0	69.3±2.2	62.7±4.6	59.5±4.2	67.9±2.2	69.1±4.7	**70.3±2.8**
	AUC	80.9±2.4	77.1±5.6	79.7±4.3	77.8±5.4	85.1±2.1	82.4±3.1	80.0±3.6	85.4±1.6	85.6±3.5	**86.6±2.4**
HAVCR2	Pearson	53.4±1.5	48.4±4.3	45.1±3.8	49.6±4.2	61.6±5.1	61.7±6.8	58.3±4.6	**66.4±4.0**	65.3±5.4	65.7±4.8
	Spearman	51.1±2.3	45.9±5.0	41.8±4.1	47.5±2.4	61.2±5.6	60.4±7.1	57.6±6.3	66.0±4.5	65.0±5.7	**66.1±4.9**
	AUC	74.2±2.3	72.8±3.0	69.8±3.3	73.3±2.5	80.1±3.3	81.0±4.3	78.8±4.6	82.4±2.6	81.9±3.6	**82.6±3.4**
HIF1A	Pearson	38.9±6.9	39.2±7.9	35.5±12.3	38.0±7.4	44.5±4.8	45.4±4.9	41.7±7.2	45.3±6.7	46.4±5.1	**47.6±7.1**
	Spearman	38.0±5.4	38.5±6.8	35.3±11.7	37.3±7.5	43.9±5.8	44.0±5.3	40.7±8.1	44.4±6.3	44.5±5.7	**44.9±7.1**
	AUC	67.7±2.5	68.0±2.5	66.9±5.9	67.9±4.1	71.9±4.7	71.8±2.7	70.4±4.3	72.3±3.7	72.2±3.9	**72.5±4.2**
KEAP1	Pearson	20.3±3.8	22.7±3.0	18.8±8.3	19.9±2.9	22.6±8.4	27.3±5.3	21.2±5.0	31.5±7.2	**37.7±6.6**	33.2±8.8
	Spearman	19.6±4.0	21.8±3.1	17.6±7.1	20.0±3.7	24.3±7.8	26.4±4.8	22.9±6.1	30.1±6.2	**35.1±6.4**	29.8±8.0
	AUC	61.4±3.2	62.1±2.2	60.0±4.8	62.0±2.4	62.4±3.8	64.0±2.4	62.0±3.1	65.2±3.7	**66.4±5.3**	64.4±3.9
KRAS	Pearson	16.5±5.4	14.5±4.8	19.4±6.4	18.0±4.1	26.0±4.3	27.6±2.7	18.9±5.5	27.8±2.8	**30.3±2.8**	30.1±1.2
	Spearman	17.5±6.1	11.9±4.3	17.2±5.8	16.4±4.9	26.7±4.2	27.9±4.8	17.2±2.8	28.9±4.5	**30.5±4.4**	29.1±2.8
	AUC	58.7±3.8	55.9±3.0	57.4±3.1	58.0±3.8	62.5±2.4	63.2±3.2	56.5±1.6	64.1±3.3	**64.2±3.5**	63.5±2.1
LAG3	Pearson	60.2±2.6	57.9±3.6	54.5±6.1	56.9±2.3	66.5±2.3	64.3±2.5	61.4±1.7	68.0±0.9	67.9±2.2	**70.1±0.6**
	Spearman	55.4±5.0	54.0±6.7	47.8±6.1	51.9±4.1	62.8±4.2	60.2±2.3	56.1±3.2	63.9±3.5	63.0±3.1	**66.1±3.1**
	AUC	77.0±4.0	76.1±4.9	73.2±3.1	75.9±1.9	80.6±4.4	78.4±1.1	76.8±2.6	81.9±2.7	81.1±3.1	**82.1±3.1**

Continued on next page

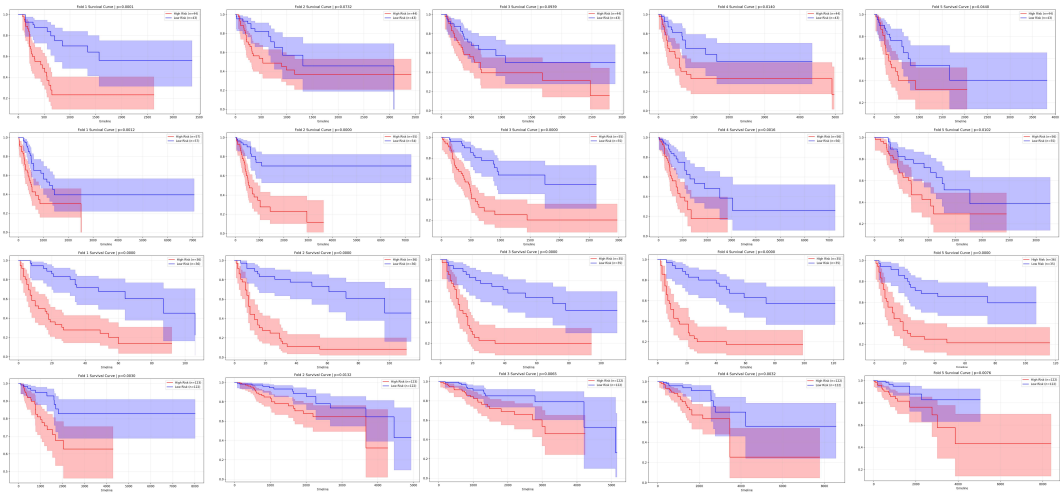

Figure 5: Five-fold Kaplan–Meier survival curves for survival prediction. Rows from top to bottom correspond to BLCA, LUAD, GBC, and BRCA, respectively. For each dataset, the five columns show the Kaplan–Meier curves from the five cross-validation folds. Patients are stratified into high-risk and low-risk groups according to the predicted slide-level risk scores.

Table 13: Full gene-level prediction results on TCGA-BRCA (continued).

Gene	Metric	ABMIL	DSMIL	TransMIL	RRT-MIL	TITAN	CHIEF	Feather	GECKO	ConcepPath	Ours
	Pearson	47.6±5.0	41.4±8.2	42.9±7.2	45.9±4.2	54.3±4.6	50.9±5.5	51.5±5.5	57.1±3.1	**59.4±3.2**	59.0±3.8
MET	Spearman	48.2±4.9	42.1±7.8	43.0±6.4	46.9±4.6	55.1±2.8	50.6±3.9	51.6±4.5	57.8±2.3	**60.1±3.0**	59.5±3.2
	AUC	74.1±4.2	70.7±4.5	71.1±3.3	72.9±4.8	77.4±2.0	74.7±2.5	75.9±2.3	79.5±1.9	80.5±2.6	**80.7±2.7**
	Pearson	58.1±3.1	51.2±9.0	55.1±5.7	52.9±7.4	60.6±4.0	59.1±4.6	57.7±0.8	64.9±2.0	65.3±2.5	**65.4±1.4**
MKI67	Spearman	60.3±2.0	52.7±7.1	55.8±6.2	54.0±8.3	62.2±3.3	60.1±4.2	58.4±1.8	65.7±1.6	66.0±2.9	**66.4±1.6**
	AUC	80.3±1.3	76.8±2.2	77.5±4.1	76.4±6.1	81.7±2.7	80.8±3.0	78.6±3.7	82.7±2.0	**83.0±2.9**	83.0±2.6
	Pearson	61.4±3.4	54.4±11.6	56.4±5.7	58.7±4.6	71.6±2.3	66.9±3.7	63.9±3.5	71.0±3.2	72.9±3.8	**73.6±2.3**
PDCD1	Spearman	58.8±5.7	52.0±13.6	52.3±8.1	55.1±7.0	69.7±2.3	64.6±4.5	60.4±3.9	68.6±3.7	71.1±4.8	**71.8±2.9**
	AUC	78.8±3.4	75.2±6.6	75.2±4.5	76.6±4.7	85.0±1.2	81.2±2.7	79.9±1.8	84.5±0.8	**86.0±1.4**	85.2±1.2
	Pearson	59.7±4.3	51.5±11.1	56.5±4.9	54.0±6.5	69.7±3.5	67.2±2.8	64.2±4.5	71.6±3.4	**72.2±2.1**	71.5±3.5
PTPRC	Spearman	58.6±6.1	51.2±12.8	54.2±6.1	53.0±6.8	68.6±4.7	65.6±3.6	62.7±4.8	71.3±4.3	**71.7±2.8**	71.3±4.3
	AUC	78.4±3.6	75.5±6.3	76.2±3.7	75.6±3.8	82.3±3.6	81.1±2.1	80.7±3.0	**84.9±3.1**	84.7±2.5	84.0±2.5
	Pearson	31.4±7.4	26.6±6.9	27.4±2.9	24.0±2.6	35.6±6.1	37.3±3.1	33.9±7.7	41.7±2.9	**46.6±5.4**	43.6±3.5
STK11	Spearman	26.5±7.7	24.6±8.0	25.8±6.6	23.6±3.7	32.3±5.3	35.0±3.9	30.7±5.2	40.0±4.5	**43.2±3.9**	39.9±3.4
	AUC	62.1±4.3	61.7±4.8	61.4±3.4	61.9±3.7	64.7±3.8	67.4±1.1	64.6±3.3	69.8±3.4	**71.6±2.2**	70.3±2.4
	Pearson	8.1±5.2	5.3±4.8	6.6±6.8	3.9±7.0	11.4±8.5	9.9±5.5	9.7±4.8	10.8±7.9	11.5±7.3	**11.8±5.7**
TP53	Spearman	3.1±6.9	5.2±5.3	6.5±5.6	6.1±4.8	9.0±7.1	7.3±3.8	7.5±6.2	8.9±4.3	**11.2±5.3**	9.8±5.4
	AUC	50.3±4.0	52.8±2.5	53.7±3.5	54.6±3.3	54.5±3.2	54.2±1.3	55.3±2.0	54.4±3.9	**56.3±2.8**	54.7±1.7
	Pearson	49.2±4.9	44.0±11.4	42.1±7.2	43.9±7.4	56.0±6.5	55.3±5.0	49.7±1.5	57.9±3.9	58.9±4.6	**59.6±4.2**
VIM	Spearman	48.0±3.1	43.3±9.5	40.3±6.6	42.3±9.0	54.4±6.6	54.4±5.3	47.0±2.4	56.9±4.6	58.1±5.2	**58.9±4.6**
	AUC	73.7±2.5	71.7±4.9	70.2±3.3	70.9±4.3	77.0±1.8	77.3±2.9	72.5±1.5	78.3±2.5	79.1±2.8	**79.9±1.7**

F Additional Visualization Results

F.1 Five-fold Kaplan–Meier Survival Curves

To further evaluate the survival stratification ability of TMEvolve, we provide the Kaplan–Meier survival curves for all five cross-validation folds. For each fold, patients are divided into high-risk and low-risk groups according to the predicted risk scores. As shown in Figure 5, TMEvolve generally separates the two risk groups across different cohorts, indicating that the learned slide-level risk representations are consistently associated with patient survival outcomes.

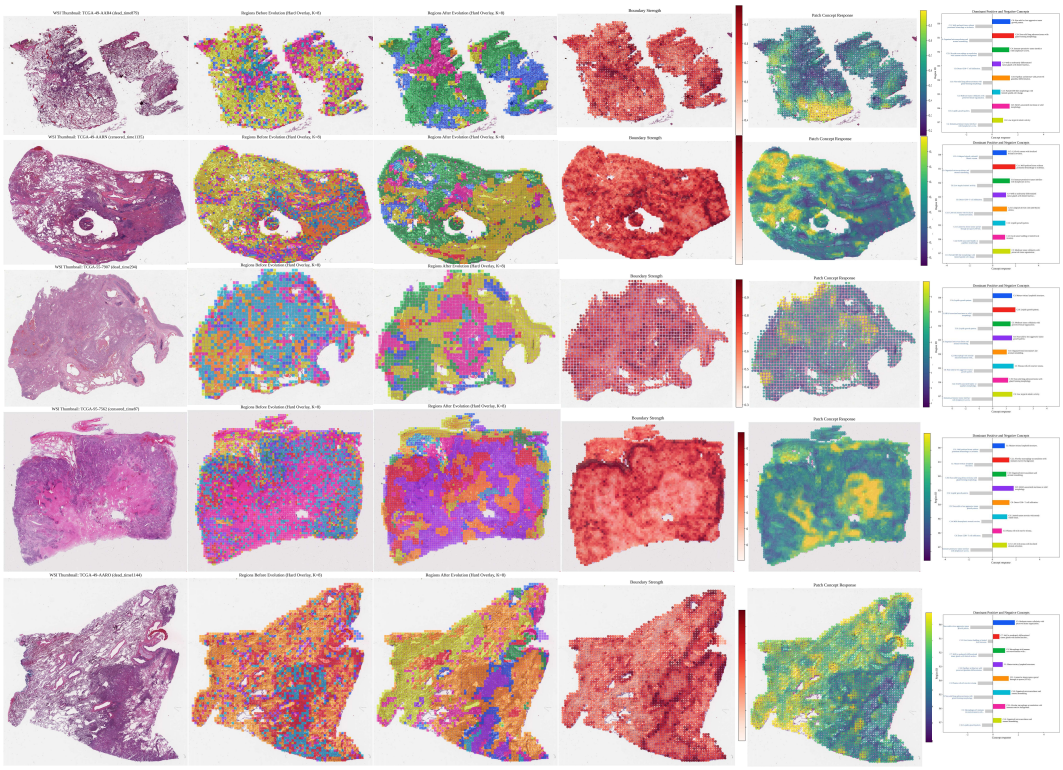

Figure 6: Additional interpretability visualization of TMEvolve on representative TCGA-LUAD WSI cases. From left to right, we show the original WSI thumbnail, the region partition before evolution, the region partition after evolution, the boundary strength map, the patch-level concept response map, and the dominant concept response summary. Compared with the initial partitions, the post-evolution regions become more spatially coherent, while the boundary and concept response maps highlight heterogeneous tissue interfaces and concept-related microenvironmental patterns in LUAD.

F.2 Additional Interpretability Visualization

To provide more qualitative evidence for the interpretability of TMEvolve, we present additional visualization results on representative WSI cases in Figure 6. For each case, we visualize the original WSI thumbnail, the region partition before evolution, the region partition after evolution, the boundary strength map, the patch-level concept response map, and the dominant concept response summary. These visualizations show how TMEvolve organizes patch-level features into coherent tissue regions and further refines the region structure through concept-guided microenvironment evolution.

G Limitations

TMEvolve is designed as a representation-learning framework for modeling region-level microenvironment organization, rather than a mechanistic simulation of biological tumor progression. The pseudo-time evolution process should therefore be interpreted as iterative feature refinement over heterogeneous tissue regions, not as true temporal dynamics of the tumor microenvironment. In addition, although the pathological concept bank is reviewed by experts for histopathological relevance and visual observability in H&E WSIs, its coverage may still be limited for rare or dataset-specific patterns.

Another limitation is that region partitioning is learned from slide-level supervision without pixel-level tissue annotations. Thus, the learned regions are model-derived microenvironment organizations rather than exact pathological segmentation masks. While our visualizations suggest improved spatial coherence and conceptual interpretability, further validation using expert annotations, spatial

molecular profiles, or multi-scale region graphs would strengthen the biological interpretation and extend TMEvolve to capture longer-range tissue interactions.

H Broader Impact

TMEvolve may have positive societal impacts by improving weakly supervised computational pathology models for prognosis, molecular prediction, and subtype classification, and by providing region- and boundary-level interpretability that may support pathology research and clinical decision support. However, the method also carries potential negative impacts if deployed without sufficient validation. Prediction errors, dataset bias, domain shifts across hospitals or scanners, and over-interpretation of model-derived regions may lead to misleading conclusions in clinical settings. Therefore, TMEvolve should be used as a research and decision-support tool rather than a standalone diagnostic system, and any clinical deployment would require rigorous external validation, expert oversight, privacy protection, and compliance with institutional and ethical regulations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the paper's main contributions and scope. They state that TMEvolve models WSIs through adaptive region formation, intra-region diffusion, concept-guided boundary evolution, and evolution-aware aggregation, and the claimed effectiveness is evaluated across survival prediction, gene expression prediction, and subtype classification.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated limitations section. It discusses that TMEvolve is a representation refinement framework rather than a direct simulation of real biological tumor progression, that it relies on a predefined pathological concept bank, that the learned regions should be interpreted as model-derived tissue organizations rather than exact segmentation masks, and that local neighborhood-based evolution may not fully capture long-range tissue dependencies.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results such as theorems, propositions, or lemmas. The mathematical formulations define the proposed graph-discretized latent evolution process, and a detailed derivation from the latent reaction–diffusion formulation to the patch-level update is provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the key information needed to reproduce the main experimental results, including dataset sources and sizes, task definitions, evaluation metrics, feature and text encoders, training settings, cross-validation protocol, task-specific hyperparameters, and implementation details. Additional experimental details are provided in the appendix, and code is included in the supplementary material for review.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is included in the supplementary material with a README and environment file for review and reproduction. Public datasets used in the paper are cited and can be accessed from their original sources, while the private GBC cohort will be released upon acceptance after de-identification and institutional approval.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the main experimental settings needed to understand the reported results, including dataset/task definitions, evaluation metrics, five-fold cross-validation protocol, feature and text encoders, optimizer, learning rate, batch size, dropout, neighborhood size, regularization weight, and task-specific choices of the number of regions and evolution steps. Additional dataset details, hyperparameter settings, and implementation details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports the main quantitative results as mean and standard deviation across cross-validation folds. Error bars are provided for survival prediction, subtype classification, ablation studies, and hyperparameter sensitivity analyses, which helps characterize the variability and stability of the reported performance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides the main computational resource information needed to understand and reproduce the experiments. It reports that experiments are conducted with pre-extracted patch-level WSI features using batch size 1 on a server with eight NVIDIA RTX 4090 GPUs, and includes a computational cost analysis with trainable parameters and inference time per slide. These details characterize the compute setting and execution cost of the proposed method and compared baselines.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. The experiments are conducted on publicly available datasets and a private clinical cohort used under appropriate ethical, privacy, and institutional data-use constraints. No personally identifiable information is disclosed, and the private cohort is not openly released due to patient privacy and ethical restrictions. The paper also discusses reproducibility, limitations, and potential societal considerations related to computational pathology.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both potential positive and negative societal impacts. Positive impacts include improving weakly supervised computational pathology for prognosis, molecular prediction, and subtype classification, as well as providing interpretable region- and boundary-level evidence. Potential negative impacts include prediction errors, dataset bias, domain shift, and over-interpretation of model-derived regions, especially if the method is deployed without sufficient external validation and expert oversight.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data or models with a high risk of misuse, such as pre-trained language models, image generators, or scraped datasets. The proposed method is a computational pathology model for weakly supervised WSI analysis. For sensitive clinical data, the private cohort is not openly released due to ethical, privacy, and institutional data-use constraints, and only code and releasable data will be made available.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original creators and owners of the datasets, pretrained models, and compared methods used in the paper are properly credited through citations. Public assets are used according to their corresponding licenses and terms of use, and the private clinical cohort is used under institutional, ethical, and privacy constraints.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets introduced by the paper include the TMEvolve implementation, task-specific concept banks, and the private GBC cohort. The implementation is included in the supplementary material with documentation, the concept construction is described in the appendix, and the private cohort will be released upon acceptance after de-identification and institutional approval.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or direct interaction with human subjects. The experiments are conducted on publicly available datasets and a retrospective private clinical cohort under appropriate ethical, privacy, and institutional data-use constraints, and no participant instructions, screenshots, or compensation procedures are applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study uses publicly available de-identified datasets and a retrospective private clinical cohort. The private cohort was used under institutional ethical approval and privacy-preserving data-use constraints. Patient-identifiable information was removed before analysis, and the study involves no direct intervention or interaction with participants, resulting in minimal risk to study participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describes the use of LLMs in the core methodology. Specifically, GPT-5 is used to generate task-relevant pathological concept descriptions, which are then reviewed by pathology experts and encoded by a text encoder to guide concept-based microenvironment modeling. The use of LLMs is therefore disclosed because it contributes to the construction of the concept bank rather than being limited to writing, editing, or formatting.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.