Bi-level Graph Learning Unveils Prognosis-Relevant Tumor Microenvironment Patterns from Breast Multiplexed Digital Pathology

Zhenzhen Wang[1,2], Cesar A. Santa-Maria[3,4], Aleksander S. Popel[1], Jeremias Sulam[1,2]

[1]Department of Biomedical Engineering, Johns Hopkins University.
[2]Mathematical Institute for Data Science, Johns Hopkins University.
[3]Department of Oncology, Johns Hopkins University.
[4]Sidney Kimmel Comprehensive Cancer Center.

Contributing authors: zwang218@jhu.edu; csantam2@jhu.edu; apopel@jhu.edu; jsulam1@jhu.edu;

Abstract

The tumor microenvironment is widely recognized for its central role in driving cancer progression and influencing prognostic outcomes. Despite extensive research efforts dedicated to characterizing this complex and heterogeneous environment, considerable challenges persist. In this study, we introduce a data-driven approach for identifying patterns of cell organizations in the tumor microenvironment that are associated with patient prognoses. Our methodology relies on the construction of a bi-level graph model: (i) a cellular graph, which models the intricate tumor microenvironment, and (ii) a population graph that captures inter-patient similarities, given their respective cellular graphs, by means of a soft Weisfeiler-Lehman subtree kernel. This systematic integration of information across different scales enables us to identify patient subgroups exhibiting unique prognoses while unveiling tumor microenvironment patterns that characterize them. We demonstrate our approach in a cohort of breast cancer patients, where the identified tumor microenvironment patterns result in a risk stratification system that provides complementary, new information with respect to alternative standards. Our results, which are validated in a completely independent cohort, allow for new insights into the prognostic implications of the breast tumor microenvironment, and this methodology could be applied to other cancer types more generally.

Keywords: tumor microenvironment, prognosis, spatial analysis, graph learning, graph kernel, breast cancer.

1 Introduction

The tumor microenvironment (TME) is a complex ecosystem, comprising proliferating tumor cells, tumor stroma, immune cells, blood and lymphatic vessels [1]. There is accumulating evidence underscoring the pivotal role of the TME in driving tumor progression [2], contributing to treatment resistance

[3], and influencing patient prognosis [4]. Recent technological advancements in spatial multiplex proteomics, such as multiplexed immunohistochemistry (mIHC) [5], co-detection by indexing (CODEX) [6], multiplexed ion beam imaging (MIBI) [7], and imaging mass cytometry (IMC) [8], have enabled the simultaneous assessment of a wide spectrum of proteins in tissue specimens. These state-of-the-art technologies provide in-depth phenotypic profiling of hundreds of thousands of cells, allowing for a comprehensive exploration of the complexity and heterogeneity of TMEs at the single-cell level [9].

Increasing attention has been given to studying the TME at varying scales [10, 11], including the analysis of cell type compositions in various cancers [12–14], the quantification of spatial distances between pairs of cell phenotypes [15], and the characterization of cellular neighborhoods involving more than two cell types [16–18]. While these works have unveiled numerous unique patterns across a spectrum of cancers, associating them with clinical implications such as treatment response and patient prognoses remains an important and difficult challenge. In most cases, this question still relies on the formulation of explicit hypotheses regarding the relationship between the TME and diseases, grounded in domain expertise and prior knowledge [12, 19, 20]. However, this mostly hypothesis-driven approach may naturally constrain the exploration of novel relationships and patterns.

On the other hand, there is a growing interest in harnessing graph-based learning techniques to analyze the association between the TME and disease without the need for explicit hypotheses in a data-driven manner [21, 22]. Numerous recent studies have reported promising results by leveraging graph neural networks (GNNs) to model the TME and predict the presence [23, 24], grade [25, 26], stage [27], subtype [28, 29], and prognosis [30–34] of diverse types of cancers. Yet, the limited size, biological heterogeneity, and differences in staining and imaging protocols of clinical datasets constrain the quality of modern machine learning models by impeding their generalization capacity, particularly in cross-study scenarios that are often overlooked [35, 36]. Additionally, the interpretability of deep learning models, in general, remains a challenge [37, 38]. Many of these studies do not discuss explanation in their works [25, 26, 28, 29, 32, 33], some employ gradient-based [23, 27, 34] and permutation-based [24] methods to generate an importance heatmap as the models' explanation, and others conduct post-hoc analyses on learned graph features to obtain some clinical implications from their models [30, 31]. These explanations, although providing some insights into learned information, might not be clear, intuitive, or meaningful, especially for clinical experts and cancer researchers.

This study circumvents most of these limitations by proposing a novel data-driven and unsupervised learning approach to stratify patients and unveil TME patterns relevant to prognosis. We refer to this method as BiGraph, which entails the construction of two interconnected graphs: a patient-specific *cellular graph*, and a subsequent *population graph* given the patients' characteristics captured by the former (see Fig. 1). The cellular graph models the cell spatial distribution of the TME for individual patients, capturing detailed information about the proximity, organization, and phenotypic information of cells extracted from tissue microarrays. On the other hand, the population graph captures similarities among all patients given their TME patterns at a macro, population level, with strong connections indicating high inter-patient similarities. A novel graph kernel function, referred to as the *Soft*-WL subtree kernel, serves as the bridge between these two hierarchies of graphs, measuring the similarity between pairs of cellular graphs. Such a Soft-WL kernel is a relaxation of the well-known Weisfeiler-Lehman (WL) subtree kernel [39], which involves the identification of heterogeneous TME patterns from patients' cellular graphs and the comparison of the abundance of these patterns across patients (see Fig. 2). The combination of these graphs at different scales facilitates the identification of patient subgroups with similar TME patterns through community detection methods. In turn, the distinct survival outcomes observed among different patient subgroups provide valuable insights into the underlying associations between TME patterns and patients' prognoses.

While the developed methodology is general, we focus our study on breast cancer. The analysis results demonstrate that BiGraph successfully provides a new risk stratification for breast cancer patients, which is complementary to the existing clinical subtyping system [40]; in other words, the found biomarkers provide insights that enhance the stratification provided by standard biomarkers of breast cancer, such as hormone receptor (HR) and human epidermal growth factor receptor 2 (HER2).

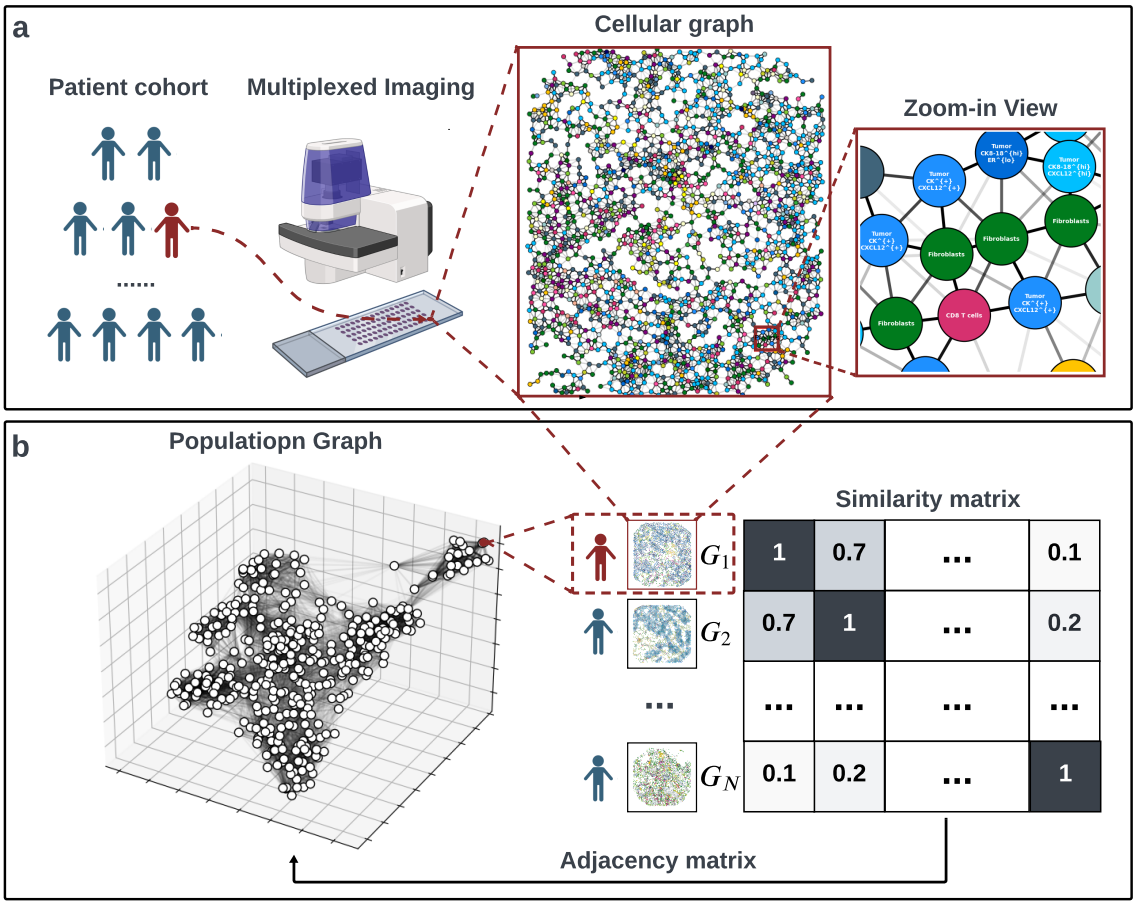


Fig. 1: **An overview of the BiGraph method. (a)** The cellular graph is constructed to model patients' tumor microenvironment (TME) based on their multiplexed images. Each node represents a cell, and cells are connected via edges with varying weights inversely correlated with inter-cellular distance. **(b)** The population graph takes the inter-patient similarity matrix as its adjacency matrix, where each node is a patient, and edge weight represents the inter-patient similarity.

Furthermore, the patient subgroups are characterized by specific, quantifiable TME patterns that lead to the discovery of prognosis-relevant TME patterns. To underscore the robustness of our findings, we validate them in both an inner validation set (comprising a randomly partitioned hold-out set from the same data source) as well as an external validation set that has been independently collected and curated. These experiments showcase the generalizability of our results across diverse datasets, even in the presence of variations in antigen panels and cell phenotyping systems.

2 Results

2.1 BiGraph: Interconnected cellular graph and population graph capture multi-scale information

To tackle the identification of prognosis-relevant TME patterns, we develop an unsupervised bi-level graph learning method dubbed BiGraph. BiGraph is designed to be applicable to single-cell imaging data obtained through multiplexed imaging technologies. It takes as input tabular data containing the

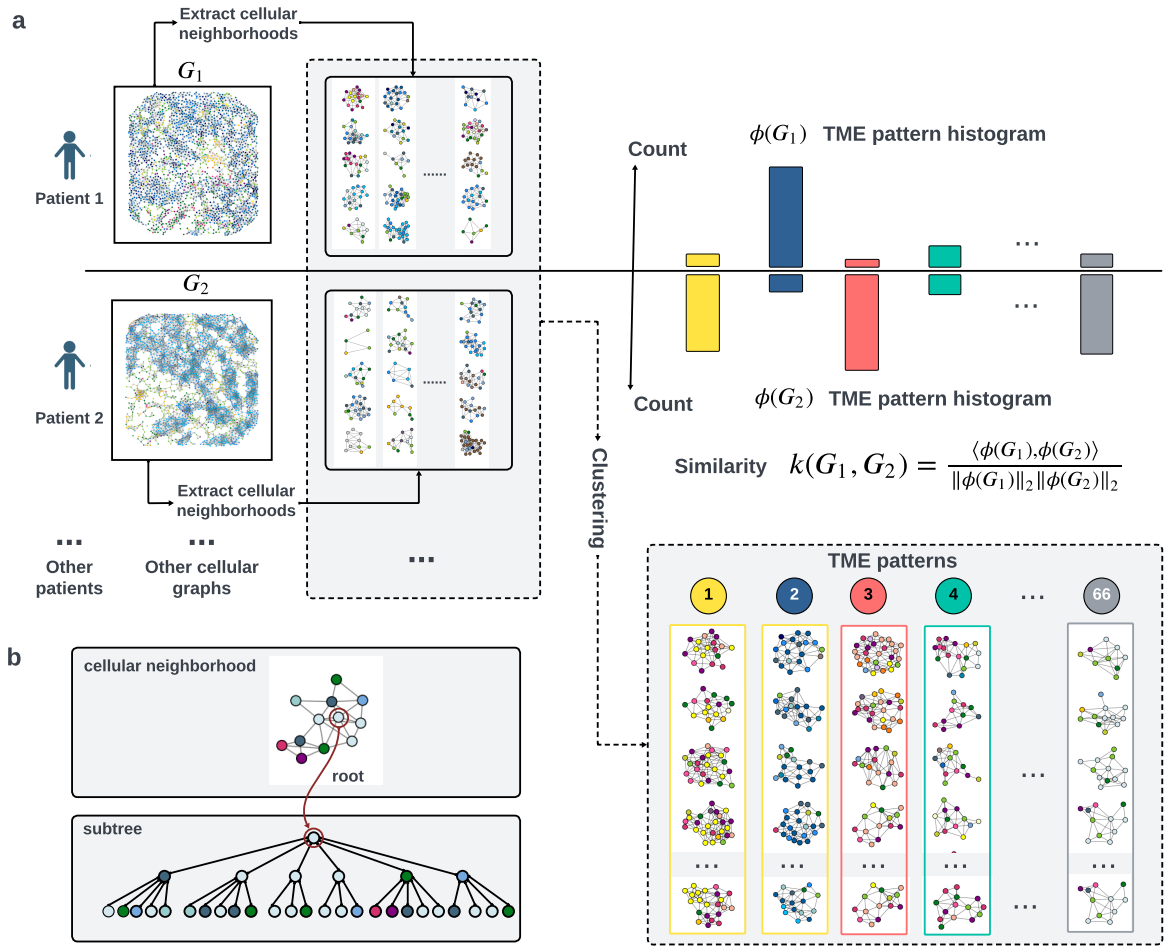


Fig. 2: **The Soft-WL subtree kernel. (a)** Schematic diagram of the Soft-WL subtree kernel method. Numerous subtrees – each of which represents a cellular neighborhood – are extracted from cellular graphs. Similar subtrees are clustered together to form TME patterns. A TME pattern histogram is calculated for each cellular graph by counting the abundance of each TME pattern. The similarity between two cellular graphs is the normalized inner product of their TME pattern histograms. **(b)** The correspondence between a cellular neighborhood and a subtree. Given a cellular neighborhood with a designated root node, a subtree can be derived by iteratively identifying the neighborhoods of the root.

spatial coordinates and phenotypes of cells and outputs a risk stratification of patients as well as the prognosis-relevant patterns that characterize certain subgroups.

The primary innovation of BiGraph relies on exploiting relations across levels of graphs, namely, a cellular graph and a population graph, by means of a graph kernel method. The cellular graph models the TME of each individual patient, with each node in it corresponding to a cell and characterized by its spatial coordinates and phenotype label. Nodes (i.e., cells) are connected through edges representing inter-cellular interactions (see Fig. 1.a). Unlike conventional approaches that employ fixed distance thresholds to determine cell connectivity [15, 16, 30], we construct a complete cellular graph where all possible pairs of cells are potentially connected, with the strength of interaction decreasing as the distance between cells increases (see Section 5.2).

On the other hand, inspired by the PhenoGraph algorithm [41] that encodes phenotypic similarities of cells in a graph, our population graph models inter-patient similarities based on their TME patterns.

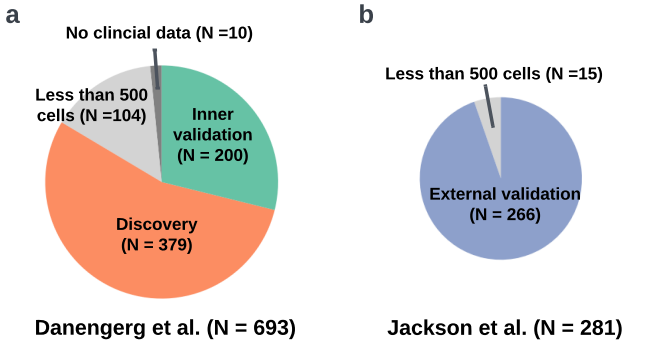
Each node of the population graph represents a patient, and every patient is connected to potentially all other patients through edges with varying weights (see Fig. 1.b, Section 5.4). Each edge weight represents the similarity between the corresponding two patients, each characterized by their specific cellular graph. This similarity between patient's cellular graphs is measured by a novel relaxation of the popular Weisfeiler-Lehman subtree kernel [39], namely the soft-WL subtree kernel. Such a kernel generates numerous *subtrees* from each cellular graph, each rooted in a single cell. A *feature embedding* is calculated for each subtree through an iterative graph convolution process. The subtree feature embedding encodes the unique composition and spatial organization of cells within a local cellular neighborhood surrounding the subtree root (see Section 5.3.1). Subtrees with similar embeddings are clustered to form *TME patterns* (see Section 5.3.2), and inter-patient similarity is established by comparing the abundance of TME patterns across patients (see Section 5.3.3). These notions of similarities between patients allow us to obtain clusters or sub-communities of patients. We employ the popular Louvain method [42] on the population graph to detect these communities, each representing a patient subgroup with common TME patterns (see Section 5.4). Importantly, the entire process is unsupervised: no information about the survival of patients is used to obtain the population graph. Survival analysis is subsequently applied to each patient subgroup to assess their relative risks. Furthermore, to better understand the distinct survival outcomes of these patient subgroups, the most characteristic TME patterns in each subgroup are identified, and their prognostic impacts are investigated (see Section 5.6).

To center our study of BiGraph in the context of breast cancer, we analyze a patient cohort curated by Danenberg et al. [16] with 693 breast cancer patients, each accompanied by their 37-dimensional imaging mass cytometry (IMC) images of tissue microarray cores (TMA) and clinical data. This publicly available dataset provides well-established cell segmentation and phenotyping results, delineating 32 major cell phenotypes, as illustrated in Extended Data Fig. 1.a. From this dataset, 10 patients with no clinical data and 104 patients with less than 500 cells identified in their corresponding TMA images are excluded. The remaining 579 patients are randomly partitioned into two subsets containing 379 patients and 200 patients, respectively. We refer the larger subset with 379 patients to the *discovery set* and develop our BiGraph method based on it. The smaller subset with 200 patients is held out for validation purposes, referred to as the *inner validation set* (see Fig. 3.a). An external breast cancer patient cohort, independently curated by Jackson et al [17], is used as an *external validation set*. This dataset contains 281 patients, each accompanied by their 35-dimensional IMC images of TMAs and clinical data. The cell segmentation and phenotyping results are also publicly available (see Extended Data Fig. 1.b). Following the same patient exclusion criteria, 266 patients with more than 500 cells are included in the study (see Fig. 3.b). Noticeably, the external validation set employs a distinct antigen panel and phenotyping system from the discovery set – which is a common challenge in current biomarker discovery research – making this validation task challenging (see Fig. 3.e and 3.f). Fortunately, about 50% of the antigens in the external validation set's antigen panel are shared with that in the discovery set, as shown in the shaded regions of Fig. 3.c. We use the shared antigens to match these two cell phenotyping systems and then conduct validation, as we detail in Section 5.7.1.

2.2 Cellular graph analysis reveals heterogeneous TME patterns

The discovery set encompasses 379 breast cancer patients with their corresponding TMA images. For each patient, a cellular graph is constructed to model the spatial distribution of different types of cells within the TME. The Soft-WL subtree kernel generates numerous subtrees rooted at every single cell from each cellular graph. Each rooted subtree has a 32-dimensional feature embedding derived from an iterative graph convolution process, which encodes the composition of distinct cell phenotypes and their spatial arrangement in a local cellular neighborhood surrounding the subtree root (see Section 5.3.1).

The feature embeddings of subtrees allow us to cluster them, in an unsupervised fashion (see Section 5.3.2), culminating in a total of 66 clusters. Each of these clusters is considered a TME pattern, with unique composition and spatial organizations of cells. The proportion of each of them in every patient



c

Clinical Variables		Discovery	Inner Validation	External validation
Grade	1	35 (9%)	16 (8%)	35 (13%)
	2	147 (39%)	57 (28%)	113 (42%)
	3	190 (50%)	123 (62%)	118 (44%)
	Unknown	7 (2%)	4 (2%)	0 (0%)
Stage	I	92 (24%)	42 (21%)	N/A
	II	190 (50%)	101 (51%)	
	III	27 (7%)	26 (13%)	
	IV	4 (1%)	3 (1%)	
	Unknown	66 (17%)	28 (14%)	
Clinical Subtypes	HR+/HER2-	243 (64%)	116 (58%)	173 (65%)
	HR+/HER2+	16 (4%)	15 (7.5%)	25 (9%)
	HR-/HER2+	16 (4%)	14 (7%)	22 (8%)
	TNBC	50 (13%)	31 (15.5%)	45 (17%)
	Unknown	54 (14%)	24 (12%)	1 (0%)
Age	<50	64 (17%)	43 (22%)	54 (20%)
	50-70	171 (45%)	91 (45%)	119 (45%)
	>70	144 (38%)	66 (33%)	93 (35%)
lymphatic invasion	Yes	175 (46%)	101 (50%)	108 (41%)
	No	136 (36%)	73 (36%)	30 (30%)
	Unknown	68 (18%)	26 (13%)	30 (32%)
5-year survival rate	Overall	73%	79%	78%
	Disease specific	80%	83%	N/A

d

Danengerg et al.	Jackson et al.
Pan-CK	Pan-CK
CK8-18	CK8-18
CK5	CK5
ER	ER
HER2	HER2
CD31-vWF	CD31-vWF
SMA	SMA
CD45	CD45
CD3	CD3
CD20	CD20
CD68	CD68
Ki-67	Ki-67
c-Caspase3	c-Caspase3
Histone H3	Histone H3
DNA	DNA
CD15	CK7
CD57	CK14
CXCL12	CK19
FSP1	E/P-Cadherin
Caveolin-1	CD44
Podoplanin	PR
PDGFRB	EGFR
HLA-ABC	CD31
CD4	Vimentin
CD8	Fibronectin
B2M	CAIX
CD38	p53
FOXP3	cMyc
CD16	GATA3
ICOS	Twist
CD163	Slug
CD11c	p-HH3
PD-1	p-S6
OX40	p-mTOR
GITR	
HLA-DR	

Fig. 3: **An overview of datasets. (a)** Composition and partition for the 693 patients curated by Danenberg et al [16]. **(b)** Composition and partition for the 281 patients curated by Jackson et al [17]. **(c)** Clinical information of the discovery, inner validation, and external validation sets. **(d)** The antigens used in the datasets curated by Danenberg et al. and Jackson et al. Shaded regions indicate the shared antigens.

is calculated, revealing a high inter-patient heterogeneity of the prevalence (Fig. 4.a). A *signature* is assigned to each TME pattern to describe its unique composition and spatial organization of cells, which is formally defined as the average feature embedding of all the subtrees belonging to the designated TME pattern. The signature map reveals that the TME patterns exhibit heterogeneous features, and many of these involve intricate interactions among multiple cell phenotypes (Fig. 4.b). Based on the predominant cell phenotype, the TME patterns are grouped into four main categories: *tumor niches* that are mostly composed of tumor cells; *immune niches* with immune cells dominating; *stromal niches* that consist of almost all stromal cells; and *interface niches*, where tumor, immune, and stromal cells are interacting (Fig. 4.b).

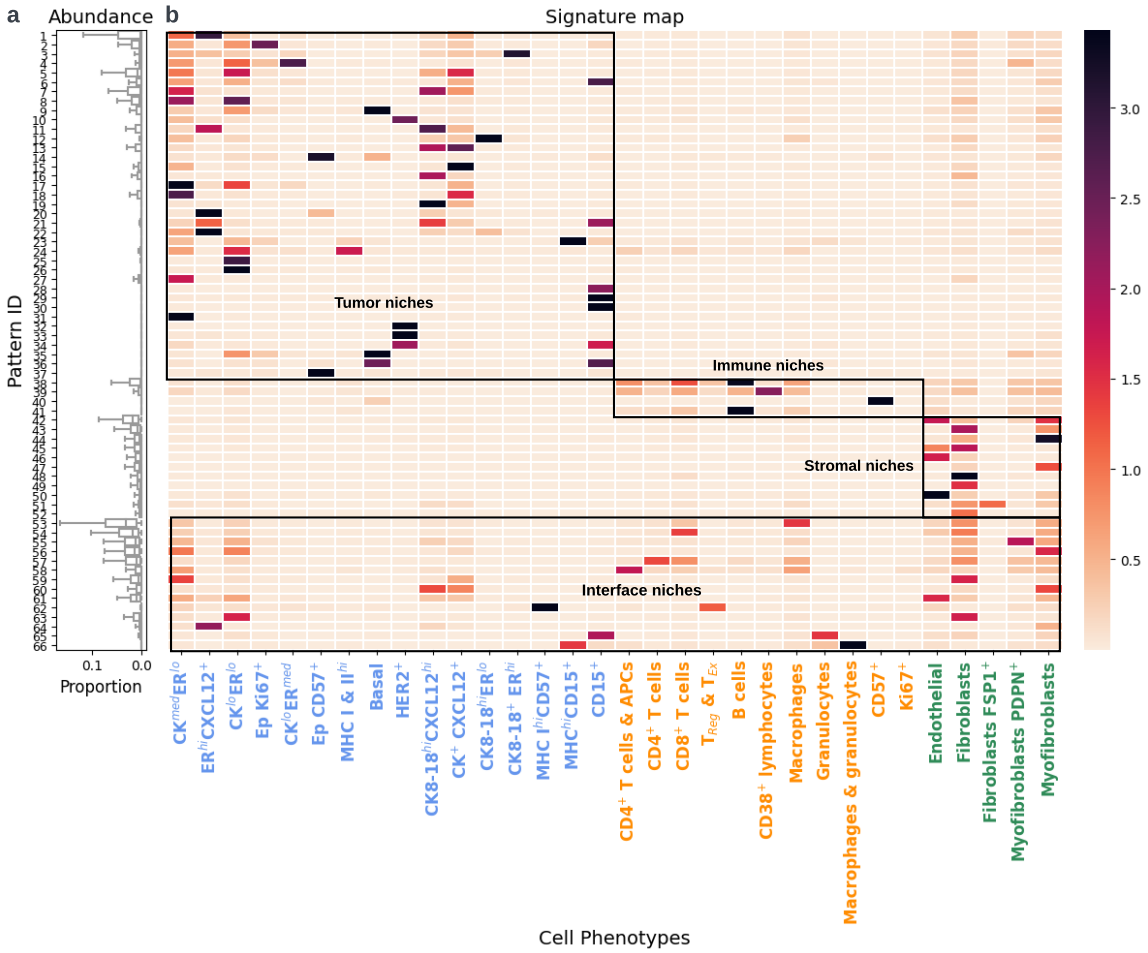


Fig. 4: **Abundance and Signature map of TME patterns. (a)** Boxplot of the proportion of each TME pattern in each patient across the discovery set. **(b)** Signature map of these TME patterns; x-axis: cell phenotype; y-axis: pattern ID. x-axis text color indicates the cell type category: blue: tumor cells; orange: immune cells; green: stromal cells. Based on the dominated cell type, the 66 TME patterns are roughly classified into four classes: tumor niche, immune niche, stromal niche, and interface niche, as the bounding boxes indicate.

2.3 Abundance of TME patterns provides inter-patient similarity measurement

Given the TME patterns extracted from each patient, we need a metric that precisely quantifies their similarities based on these patterns. Kernel functions are a natural choice. Specifically, the Soft-WL subtree kernel quantifies inter-patient similarity by comparing the abundance of these patterns in their respective TMEs. The similarity score between any pair of patients is determined by a normalized inner product of their histograms, describing the proportion of the 66 TME patterns (see Section 5.3.3, Equation (5)). The similarity score for any patient pair falls within the range $[0, 1]$, with two identical graphs achieving a maximum similarity score of 1.

To illustrate the computation of inter-patient similarity, Fig. 5.a presents a randomly selected "template patient" from the discovery set (for illustration purposes) and depicts its cellular graph. The similarities between this template patient and all other patients are calculated and visualized,

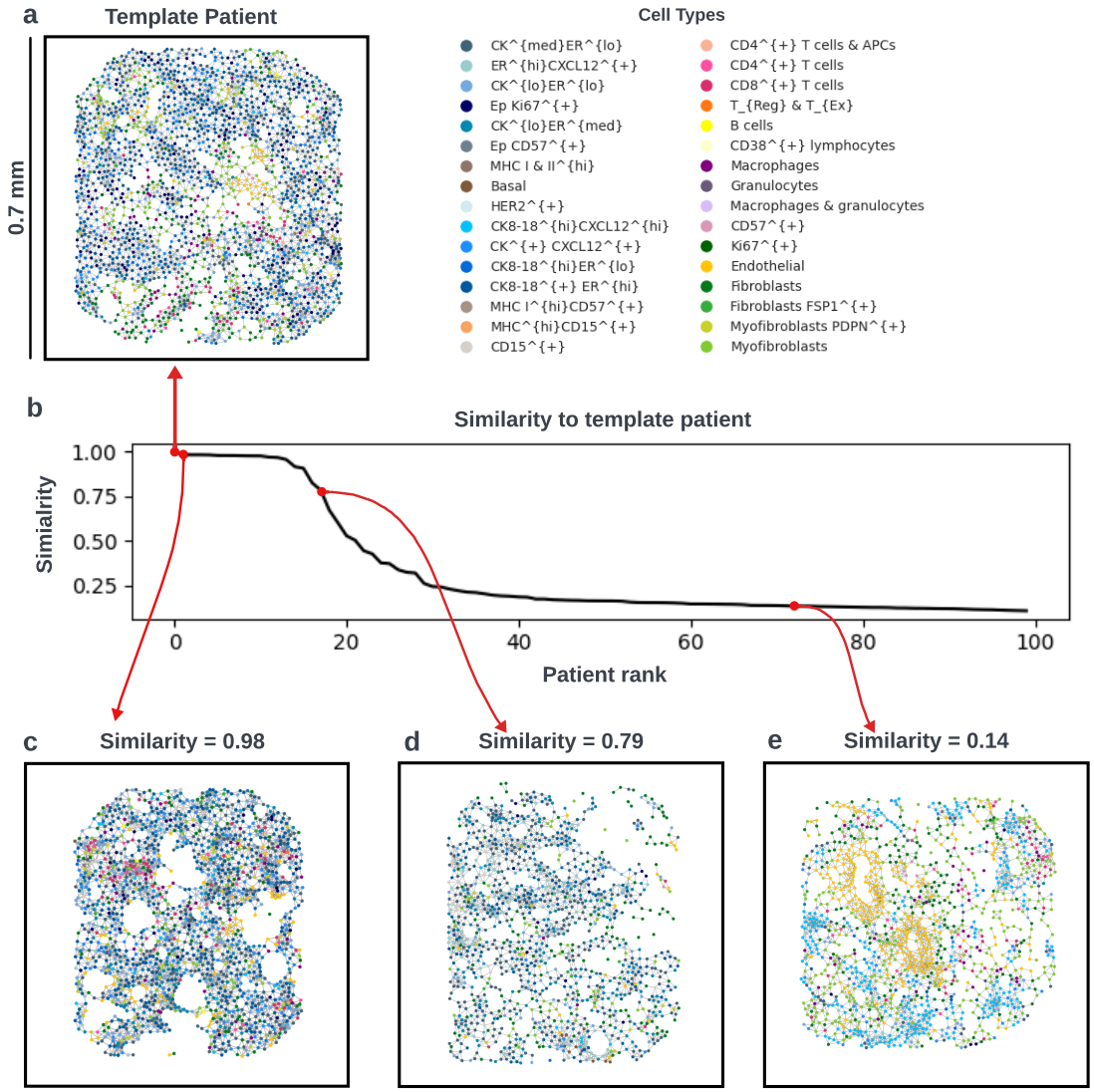


Fig. 5: **Demonstration of the inter-patient similarities measured by graph kernel methods** **(a)** The cellular graph of a selected patient, referred to as the "template patient". **(b)** The similarity scores of other patients compared to the template patient (x-axis: patient ranking; y-axis: similarity score). Patients are sorted based on their similarity score, and only the top 100 patients are shown. **(c)-(e)** Cellular graphs of patients with varying similarity scores to the template patient.

with the top 100 patients' similarities shown in Fig. 5.b. To provide a more in-depth understanding of such comparison, we highlight three representative patients with varying degrees of similarity (see Fig. 5.c-e). The cellular graph of this template patient contains diverse tumor cells, with the predominant presence of cytokeratin-positive cells, including $CK8/18^{+}ER^{high}$ cells, $CK8/18^{high}CXCL12^{high}$ cells, $CK^{+}CXCL12^{+}$ cells, and $CK^{med}ER^{low}$ cells. Proliferative (i.e. $Ki67^{+}$) tumor cells are also scattered throughout. Stromal, such as fibroblasts, myofibroblasts, and endothelial, intermingle with tumor cells, and a clear tumor-stromal boundary is difficult to discern. Immune cells, including $CD4^{+}$ T cells, $CD8^{+}$ T cells, and macrophages, are relatively rare and dispersed. Note how the patient with the

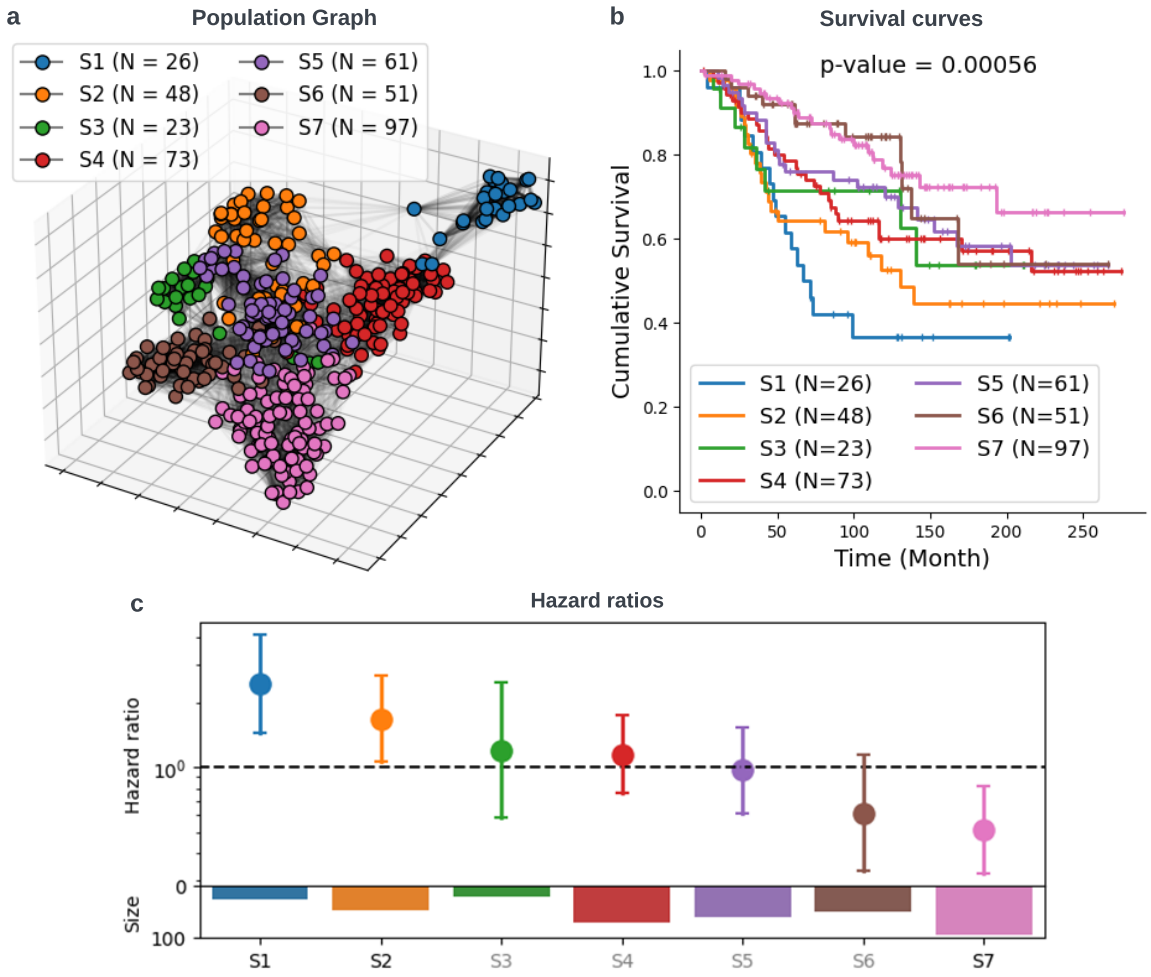


Fig. 6: **Population graph facilitating risk stratification. (a)** Three-dimensional visualization of the population graph and the results of community detection. **b** Kaplan-Meier (K-M) survival plots depicting seven distinct patient subgroups. A multivariate log-rank test is conducted to compare the survival curves, with the resulting p-value indicated in the title. **c** Relative hazard ratios (with 95% confidence intervals) of the seven patient subgroups, estimated using a Cox proportional hazard model. Text color (grey vs. black) distinguishes statistically significant and non-significant associations (black: statistically significant). The size (i.e., number of patients) of each subgroup is presented in the corresponding barplot.

highest similarity (similarity score of 0.98) exhibits nearly identical patterns to the template patient (see Fig. 5.c). Another patient with a similarity score of 0.79 (see Fig. 5.d) also features a substantial number of interactions among various cytokeratin-positive cells, including $CK8/18^{+}ER^{high}$ cells, $CK8/18^{high}CXCL12^{high}$ cells, and $CK^{+}CXCL12^{+}$ cells. However, distinct from the template patient, this patient includes many $CK^{low}ER^{med}$ cells and $ER^{high}CXCL12^{+}$ cells, resulting in the emergence of new patterns. Additionally, this patient exhibits a relatively immune-cold TME, with only a few immune cells infiltrating the tumor cells. In contrast, the last example patient demonstrates a substantially lower similarity score of 0.14, indicating significantly different patterns (Fig. 5.e). This patient presents with a vascular tumor, characterized by an abundance of endothelial cells. The tumor cells are relatively homotypic, dominated by $CK8/18^{high}CXCL12^{high}$ cells. This diversity in TME patterns

Table 1: Average inter-patient similarities of clinical subtypes and BiGraph-derived subtypes.

Clinical		BiGraph	
HER2+	0.43	S1	0.75
TNBC	0.15	S2	0.26
HR+/HER2-	0.20	S3	0.73
		S4	0.45
		S5	0.32
		S6	0.61
		S7	0.37
Entire cohort		**0.18**	

illustrates the capability of the Soft-WL subtree kernel to discern nuanced differences among patients' TMEs.

2.4 Population graph analysis provides data-driven risk stratification

Once we have a precise metric to quantify the similarity between any pair of patients, we proceed to construct a population graph that captures the inter-patient similarities across the entire set. In this graph, each node corresponds to an individual patient, and edges connecting nodes are assigned weights representing their respective similarity scores (see Section 5.4). A community detection algorithm is employed to reveal patient subgroups in an unsupervised manner without utilizing the patients' survival information (see Section 5.4). The survival outcomes of these patient subgroups are analyzed and compared with the standard clinical subtyping system. We elaborate on these results in the following subsections.

2.4.1 Patient subgroups exhibit significantly distinct survival outcomes

To identify patient subgroups characterized by similar TME patterns, we employ the widely used Louvain community detection algorithm [42] on the population graph. This process yields seven distinct communities, each representing a subgroup of patients exhibiting high intra-group similarities (see Section 5.4). These identified patient subgroups are denoted as S1 to S7. Notably, the nodes in the population graph lack spatial coordinates; hence, for visualization purposes, the Fruchterman-Reingold force-directed algorithm [43] is employed to depict a 3D representation of the population graph. This representation and the results of community detection (i.e., the resulting patient subgroups) are depicted in Fig. 6.a.

Disease-specific survival functions of the seven patient subgroups are estimated using the Kaplan-Meier (K-M) estimator [44] and shown in Fig. 6.b. A multivariate log-rank test [45] is conducted to compare their survival outcomes, revealing a statistically significant difference ($p = 0.00056$) among the groups. Hazard ratios for these patient subgroups are calculated using the Cox proportional regression model [46], and are presented in Fig. 6.c. The results reveal that S1 exhibits statistically significantly worse survival compared to other patients (hazard ratio (hr) = 2.43, 95% confidence interval (CI): 1.43-4.11, $p = 0.0009$). S2 also demonstrates significantly worse survival (hr = 1.67, 95% CI: 1.06-2.65, $p = 0.03$). Conversely, S7 exhibits statistically significantly better survival (hr = 0.51, 95% CI: 0.32-0.82, $p = 0.005$).

2.4.2 BiGraph enhances risk stratification capacity of clinical subtypes

Breast cancer is conventionally classified into 4 clinical subtypes based on the expression of hormone receptor (HR) and human epidermal growth factor receptor 2 (HER2) – HR+/HER2-, HR-/HER2+, HR+/HER2+, and triple-negative breast cancer (TNBC) [40]. Moreover, HR-/HER2+ and HR+/HER2+ are commonly merged as HER2+ in clinical consideration since they are both likely to respond to anti-HER2 treatment. It is important to understand how the BiGraph-derived groups

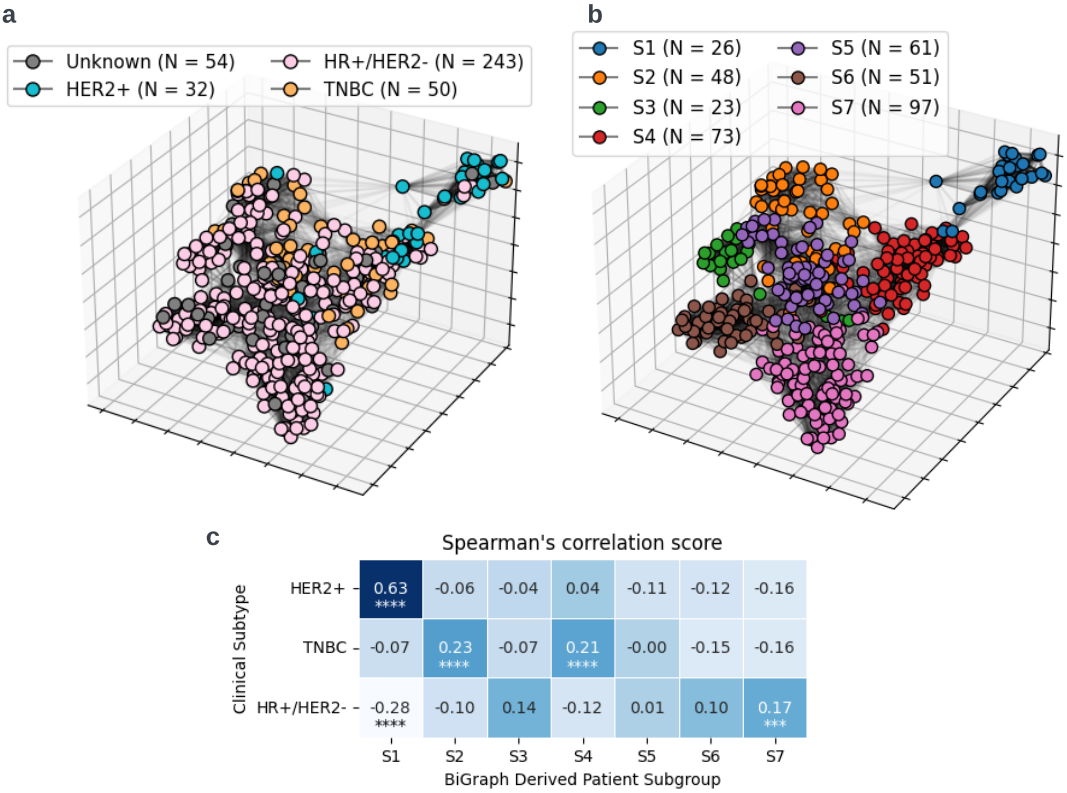


Fig. 7: **Correlation between clinical subtypes and BiGraph-derived subtypes. (a)** Population Graph with nodes painted by clinical subtypes. **(b)** Population Graph with nodes painted by BiGraph-derived subtypes. **(c)** Spearman's correlation scores between clinical subtypes and BiGraph derived subtypes. ****: p-value < 1e-4; ***: p-value < 1e-3.

relate to this standard stratification. To compare the clinical subtyping system and the BiGraph-derived subtyping system, we calculate the average inter-patient similarity for each subset of patients within both frameworks using the Soft-WL subtree kernel, revealing that the clinical subtypes generally have very low intra-subtype similarities (see Table 1). Notably, TNBC exhibits exceptional heterogeneity, with an averaged inter-patient similarity as low as 0.15, even lower than the similarity of 0.18 calculated for the entire heterogeneous breast cancer cohort. In contrast, HER2+ breast cancer demonstrates comparatively greater homogeneity. These results suggest that clinical subtyping of breast cancer inadequately captures the intricate inter-patient heterogeneity when considering a boarder range of antigens and cell phenotypes. Furthermore, Spearman's correlation scores [47] are used to assess the alignment between two stratification systems (clinical subtyping and BiGraph), revealing that they are generally not aligned with each other well (see Fig. 7). Yet, certain correlations do exist. Specifically, S1 is positively correlated with HER2+ patients, and negatively correlated with HR+/HER2-. S7, on the other hand, exhibits a positive correlation with HR+/HER2-. Additionally, S2 and S4 are positively correlated with TNBC.

A crucial observation is that the intersection sets between BiGraph-derived subtypes and clinical subtypes reveal unique survival patterns in some cases. For instance, HER2+ patients (N = 32) exhibit worse survival compared to others in the discovery set. Further stratification using S1 identifies a subset (N = 19) with even worse survival, and a complement subset (N = 13) in the HER2+ patients with

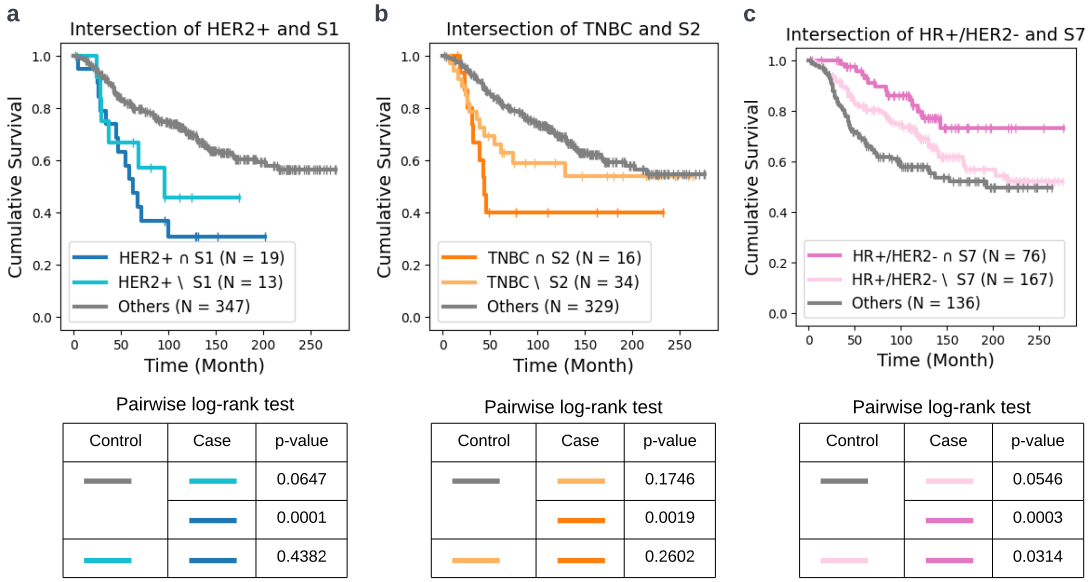


Fig. 8: **BiGraph enhances the risk stratification of clinical subtype (a)** K-M survival plots of the intersection of HER2+ and S1 patients; the complement of S1 in HER2+ patients; and all other patients. **(b)** K-M survival plots of the intersection of TNBC and S2 patients; the complement of S2 in TNBC patients; and all other patients. **(c)** K-M survival plots of the intersection of HR+/HER2- and S7 patients; the complement of S7 in HR+/HER2- patients; and all other patients.The pairwise log-rank tests are used to compare the survivals of three groups of patients in each case.

slightly better survival (Fig. 8.a). Similarly, S2 divides TNBC patients (N = 50) into a subset (N = 16) with exceptionally worse survival and a complement set (N = 34) (Fig. 8.b). Furthermore, HR+/HER2- patients (N = 243) display better survival compared to others, while its intersection with S7 patients (N = 76) shows even better survival (Fig. 8.c). These results suggest that the BiGraph-derived subtyping system provides complementary information for the clinical subtyping system, enhancing its risk stratification capacity beyond that achieved by the clinical subtyping system alone.

2.5 BiGraph unveils prognosis-relevant TME patterns

An important feature of BiGraph – distinguishing it from alternative, more opaque data-driven methods – is its explainability. The construction of the population graph relies on comparing the histogram of TME patterns in patients via the soft WL subtree kernel, which allows us to directly identify the most characteristic patterns for each patient subgroup. Furthermore, the prognostic impacts of these characteristic patterns can be systematically investigated.

2.5.1 Patient subgroups are characterized by unique TME patterns

To identify characteristic patterns of different survival profiles, the expression of TME patterns within and outside a patient subgroup are compared using the Hodges–Lehmann statistic [48]. A TME pattern is deemed "characteristic" in a patient subgroup if its Hodges–Lehmann statistic surpasses 50% of the maximum value (see Section 5.5, Extended Data Fig. 4). Results indicate that distinct patient subgroups exhibit characteristic patterns with diverse cellular compositions and organizations (see Fig. 9). Representative examples of these characteristic patterns are exhibited in Fig. 10. Although subtrees show the hierarchical structure of cells in an abstract manner, they do not intuitively convey

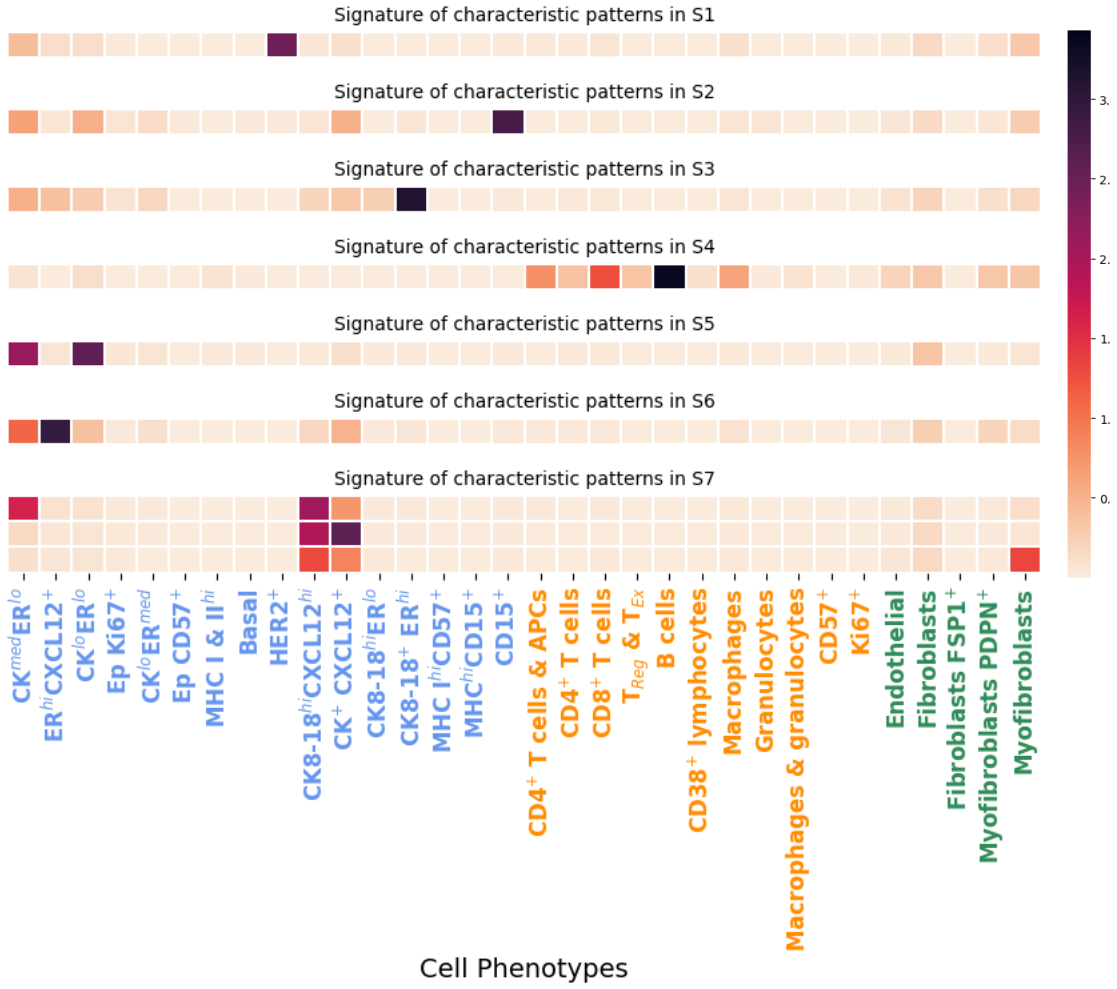


Fig. 9: **Signature of characteristic patterns in each patient subgroup.** Each row shows the signature (i.e., averaged feature embedding) of a characteristic pattern. The x-axis indicates cell phenotype and x-axis text color indicates the cell type category: blue: tumor cells; orange: immune cells; green: stromal cells.

the spatial cellular organization. For interpretation purposes, a corresponding cellular neighborhood centered at the subtree root is displayed in Fig. 10, where the neighborhood boundary is determined by cells' contributions to the subtree feature embedding (see Section 5.3.1, Equation (3)).

Patient subgroup S1, exhibiting the worst survival outcome, is distinguished by "HER2$^+$ tumor niche" patterns, showing enrichment of HER2$^+$ tumor cells surrounded by fibroblasts and myofibroblasts. In S2, a patient subgroup exhibiting a statistically significantly worse survival outcome, "CD15$^+$ tumor niche" patterns are most characteristic, featuring enriched CD15$^+$ tumor cells intermixed with some other tumor cells. S3, a relatively small and homogeneous patient subgroup, is characterized by an "enriched CK8-18$^+$ER^{high} tumor niche". The most distinctive pattern within S4 closely resembles the well-known tertiary lymphoid structure (TLS) [49], and is termed a "TLS-like niche". This pattern encompasses a variety of immune cells, including B cells (predominantly represented), CD4$^+$ T cells, CD8$^+$ T cells, regulatory T cells (Treg), macrophages, and CD38$^+$ lymphocytes. The most characteristic pattern in S5 is a "CK$^+$ER^{low} tumor niche", exhibiting a mixture of CK^{med}ER^{low} and

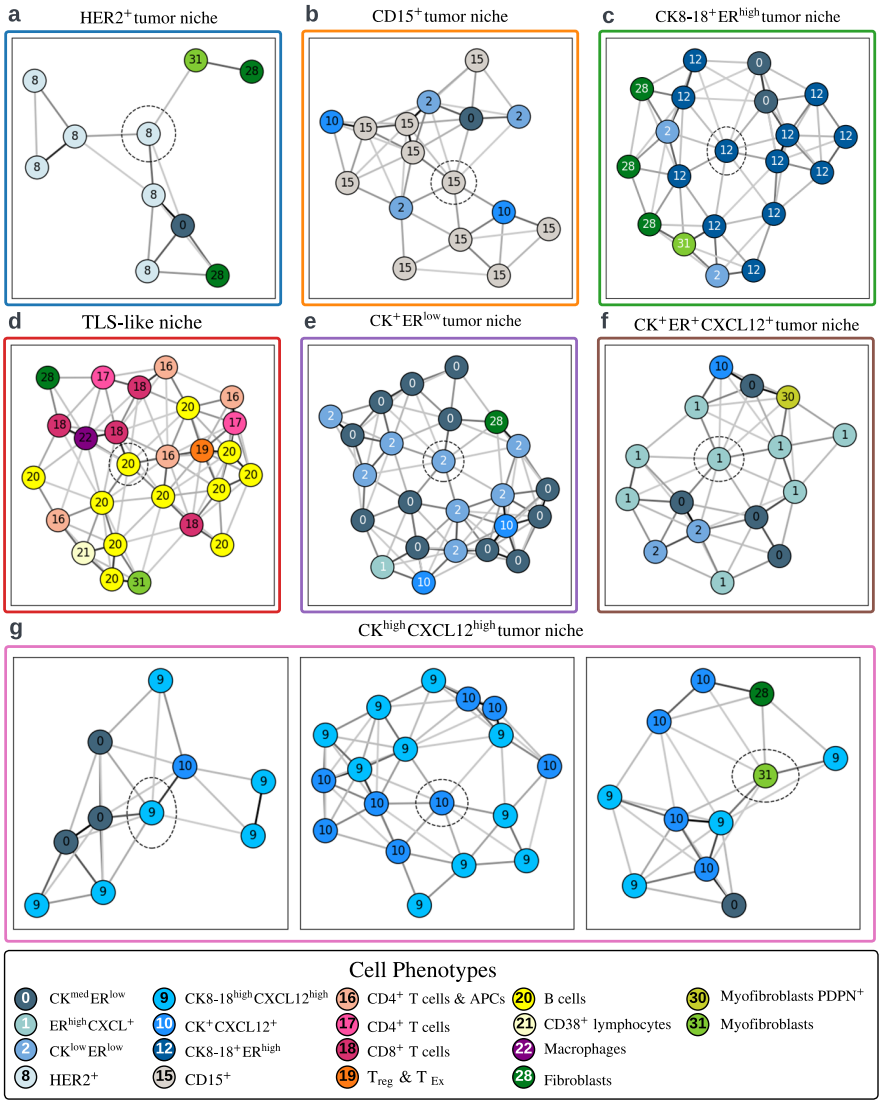


Fig. 10: **Representative examples of characteristic patterns in each patient subgroup.** To intuitively show the spatial organization of cells, the corresponding cellular neighborhoods of subtrees are shown, where dashed circles indicate the subtree root. **(a)** HER2$^+$ tumor niche, characteristic in S1; **(b)** CD15$^+$ tumor niche, characteristic in S2; **(c)** CK8-18$^+$ER^{high} tumor niche, characteristic in S3; **(d)**: TLS-like niche, characteristic in S4; **(e)** CK$^+$ER^{low} tumor niche, characteristic in S5; **(f)** CK$^+$ER$^+$CXCL12$^+$ tumor niche, characteristic in S6; **(g)** CK^{high}CXCL12^{high} tumor niche, characteristic in S7.

CK^{low}ER^{low} tumor cells. S6 is characterized by a "CK$^+$ER$^+$CXCL12$^+$ tumor niche" pattern, featuring the co-localization of CK^{med}ER^{low} and ER^{high}CXCL12$^+$ tumor cells. S7, the patient subgroup exhibiting the best survival outcome, has technically three characteristic patterns, while they exhibit a common feature: a mixture of CK8-18^{high}CXCL12^{high} cells and CK$^+$CXCL12$^+$ cells. Therefore, these three patterns are consolidated into one, named a "CK^{high}CXCL12^{high} tumor niche".

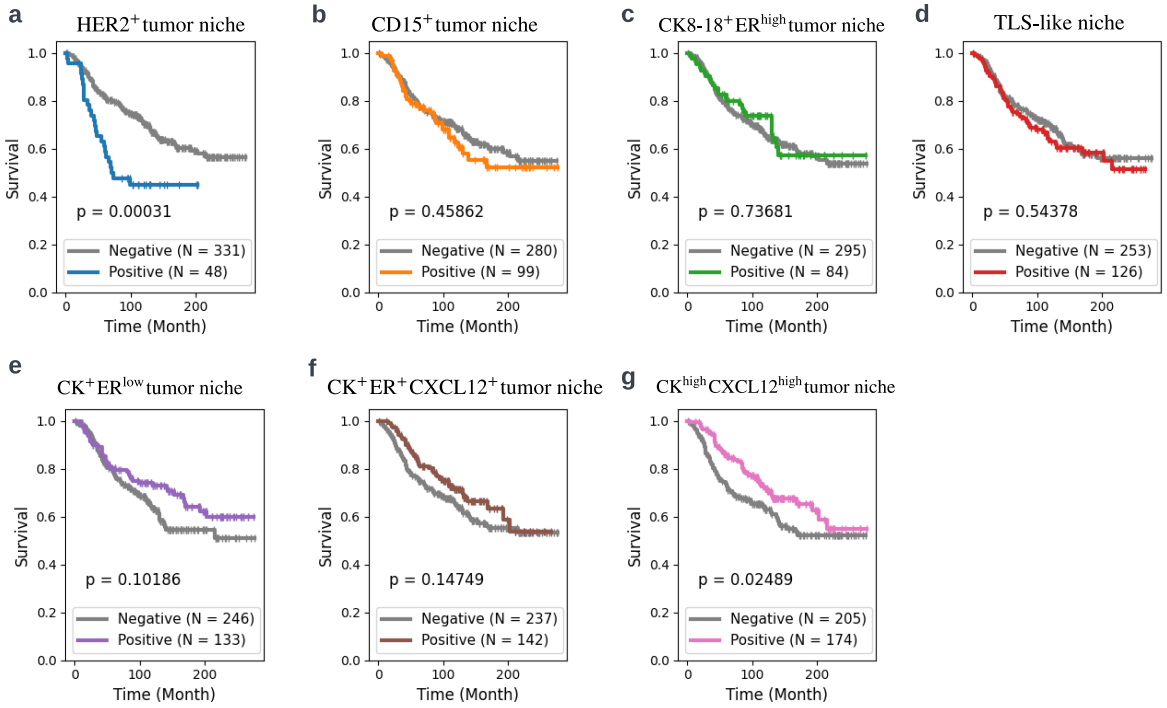


Fig. 11: **Prognostic impacts of the characteristic TME patterns**. Patients are categorized into "positive" and "negative" groups based on the proportion of a designated TME pattern with 1% as the threshold. Their survivals are compared by log-rank test, and the resulting p-values are indicated in each corresponding figures.**(a)** HER2$^+$ tumor niche, characteristic in S1; **(b)** CD15$^+$ tumor niche, characteristic in S2; **(b)** CK8-18$^+$ER^{high} tumor niche, characteristic in S3; **(d)**: TLS-like niche, characteristic in S4; **(e)** CK$^+$ER^{low} tumor niche, characteristic in S5; **(f)** CK$^+$ER$^+$CXCL12$^+$ tumor niche, characteristic in S6; **(g)** CK^{high}CXCL12^{high} tumor niche, characteristic in S7.

2.5.2 Prognostic impacts of the characteristic TME patterns

The results above showcase the most characteristic TME patterns in each subgroup, but these do not necessarily imply associations with prognosis. This naturally leads to the exploration of potential associations between these characteristic TME patterns and patient survival outcomes, which we explore next. The prognostic impacts of a specific TME pattern are assessed by categorizing patients into "positive" and "negative" groups based on the proportion of that pattern in a patient. A 1% threshold is employed for patient stratification, the same as the cut-off point used in the definition of hormone receptor (HR) status [50]. For each characteristic pattern, survival outcomes of positive and negative patients are compared using the log-rank test (Fig. 11). Results indicate that the HER2$^+$ tumor niche, characteristic of S1, is associated with worse survival ($p = 0.00031$). In contrast, the CK^{high}CXCL12^{high} tumor niche, characteristic of S7, is associated with better survival ($p = 0.025$). For patterns characteristic in S5 and S6, positive patients exhibit slightly better survival outcomes than negative patients, although the differences are not statistically significant. These results are broadly consistent with the risk stratification provided by the seven patient subgroups – S1 patients display significantly worse survival outcomes, and the most characteristic TME pattern proves to be a negative prognostic factor. Conversely, S7 patients exhibit significantly better survival outcomes, with the most characteristic TME pattern emerging as a positive prognostic factor.

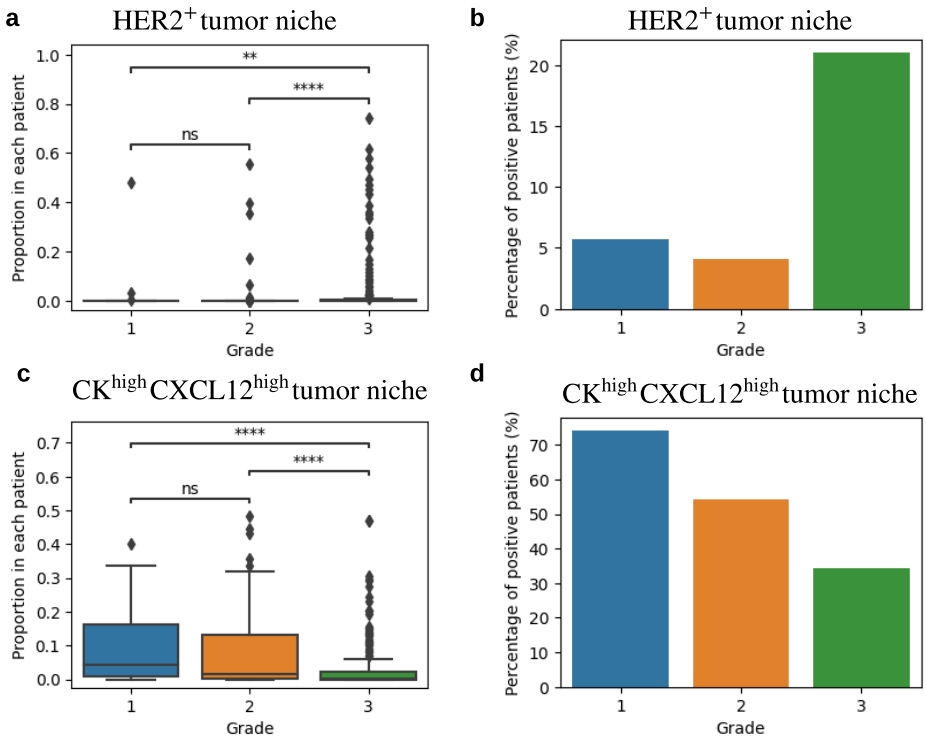


Fig. 12: **Correlation between TME patterns HER$^+$ and CK^{high}CXCL12^{high} and tumor grade. (a)** The proportion of HER2$^+$ tumor niche per patient for tumor grade 1, 2, and 3; **(b)** The ratio of patients presenting HER2$^+$ tumor niche (i.e., proportion $> 1\%$) in grade 1, 2, and 3 groups; **(c)** The proportion of CK^{high}CXCL12^{high} tumor niche per patient for tumor grade 1, 2, and 3; **(d)** The ratio of patients presenting CK^{high}CXCL12^{high} tumor niche in grade 1, 2, and 3 groups.

2.5.3 Characteristic patterns are correlated with tumor grade

To better understand the prognostic impacts of these two prognosis-relevant TME patterns (HER$^+$ and CK^{high}CXCL12^{high} tumor niches), we investigate their correlation with clinical information, such as tumor grade, stage, and metastasis. Results suggest that the expression of these TME patterns is significantly associated with tumor grade. The expression of HER2$^+$ tumor niche per patient is significantly higher in grade 3 (G3) patients compared to others (see Fig. 12.a). Furthermore, 21% of grade 3 patients have HER2$^+$ tumor niche present (i.e., proportion $> 1\%$) in their TMEs. This ratio is largely higher than that in grade 1 (G1) (6%) and grade 2 (G2) patients (4%) (see Fig. 12.b). In contrast, the expression of CK^{high}CXCL12^{high} tumor niche is significantly lower in grade 3 patients compared to others (see Fig. 12.c). The ratio of patients presenting CK^{high}CXCL12^{high} tumor niche in the G3 group is 34%, which is also lower than that in G1 patients (74%) and G2 patients (54%) (see Fig. 12.d). On the other hand, there is no significant correlation found between these patterns and tumor stage or metastasis (see Extended Data Fig. 5, Extended Data Fig. 6).

2.6 Validation and generalization

As detailed in section 2.1, a subset comprising 200 patients is reserved from the 579 patients curated by Danenberg et al. [16], denoted as the "inner validation set". Additionally, an external breast cancer patient cohort curated by Jackson et al. [17], referred to as the "external validation set", is utilized to evaluate BiGraph's cross-study generalization capacity. Importantly, both sets allow us to explore generalization under different settings: the former provides generalization results to data collected

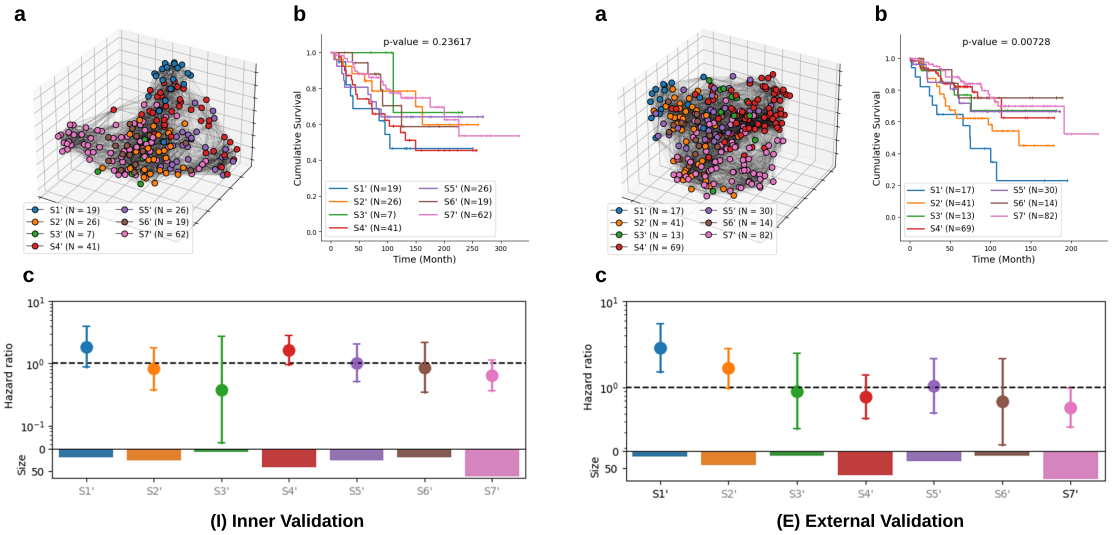


Fig. 13: **Validation of the risk stratification results. (I)** Results in inner validation set (N=200). **(E)** Results in external validation set (N=266). **(a)** Population graph with nodes painted by estimated patient subgroup labels; **(b)** K-M estimation of the survival functions of seven mapped patient subgroups. A multivariate log-rank test is used to compare them, and the p-value is indicated in the title.; **(c)** Hazard ratios of the seven mapped patient subgroups. The text color (grey vs. black) distinguishes between statistically significant and non-significant associations (black: statistically significant). The size (i.e., the number of patients) of each subgroup is presented in the corresponding barplot.

under the same protocol as those employed in the discovery set, whereas the latter involves data collected independently: the external validation set employs a distinct antigen panel and phenotype system compared to the discovery set. This latter setting adds complexity to the validation process, while on the other hand provides a stronger demonstration of the applicability of BiGraph in real-world scenarios – as standardized antigen panels are not yet common in practice. To address this challenge, we employ a mapping approach to align the cell phenotyping in the external validation set with that of the discovery set. This alignment is achieved by identifying shared antigens between both sets, as indicated by shaded regions in Fig. 3. To be more precise, every cell in the external validation set is mapped to the closest cell phenotype cluster in the discovery set, where the cell-to-cell phenotypic distance is given by the Euclidean distance of their expressions of the shared antigens (see Section 5.7.1, and the cell phenotyping alignment results are included in Extended Data Fig. 7). With the cell phenotyping system of the external validation set aligned, we validate the major results found in the discovery set in both inner- and external validation sets.

2.6.1 Validation of the patient risk stratification

As detailed in Section 2.4.1, the community detection algorithm identifies seven patient subgroups from the population graph of the discovery set with significantly distinct survivals. In the validation stage, every unseen new patient (with their corresponding cellular graph) is mapped to one of the seven patient subgroups using a k-nearest neighbor (k-NN) rule [51]. The similarity between a new patient and patients in the discovery set is measured by the Soft-WL subtree kernel function defined on the discovery set (see Section 5.7.3). In this way, seven mapped patient subgroups can be identified in the inner- and external-validation sets respectively, denoted as $S1', S2'$, etc., where these "prime" counterparts denote "validation".

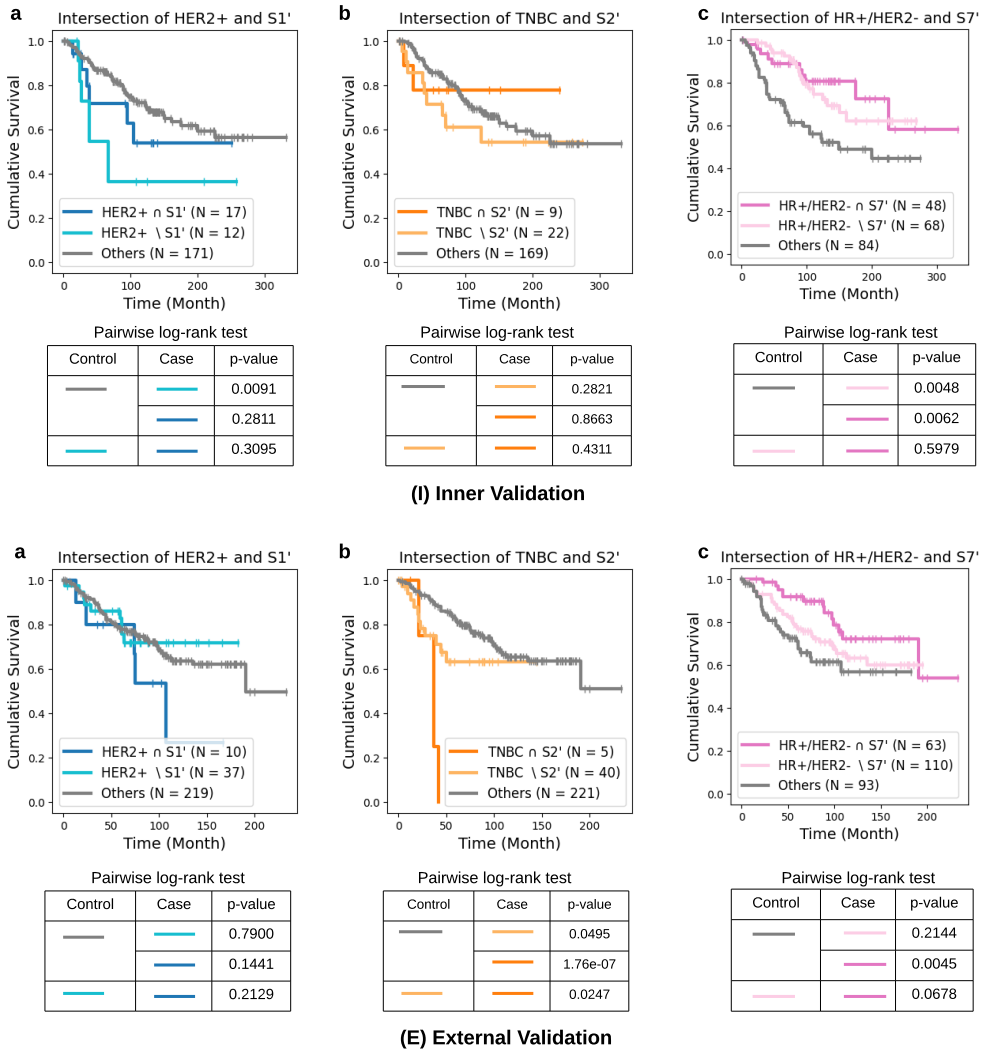


Fig. 14: **Validation of the enhancement of risk stratification that BiGraph provides to clinical subtyping. (I)** Results in the inner validation set (N=200). **(E)** Results in external validation set (N=266). **(a)** Survival plots of the intersection of HER2+ and S1′ patients; the complement of S1′ in HER2+ patients; and all other patients. **(b)** Survival plots of the intersection of TNBC and S2′ patients; the complement of S2′ in TNBC patients; and all other patients. **(c)** Survival plots of the intersection of HR+/HER2- and S7′ patients; the complement of S7′ in HR+/HER2- patients; and all other patients.The pairwise log-rank tests are used to compare the survivals of three groups of patients.

We analyze the survival outcomes of the seven mapped patient subgroups in both inner validation (Fig. 13. I) and external validation sets (Fig. 13. E). The multivariate log-rank test is used to compare K-M survival plots of the seven patient subgroups. The survival outcomes of the seven mapped patient subgroups do not have a statistically significant difference ($p = 0.24$, Fig. 13. I.b) in the inner-validation set. On the other hand, there is a statistically significant distinction among their survival outcomes in the external validation set ($p = 0.007$, Fig. 13. E.b). By comparing the hazard ratios of the seven mapped subgroups with the subgroups found in the discovery set (Fig. 6.c), we observe a

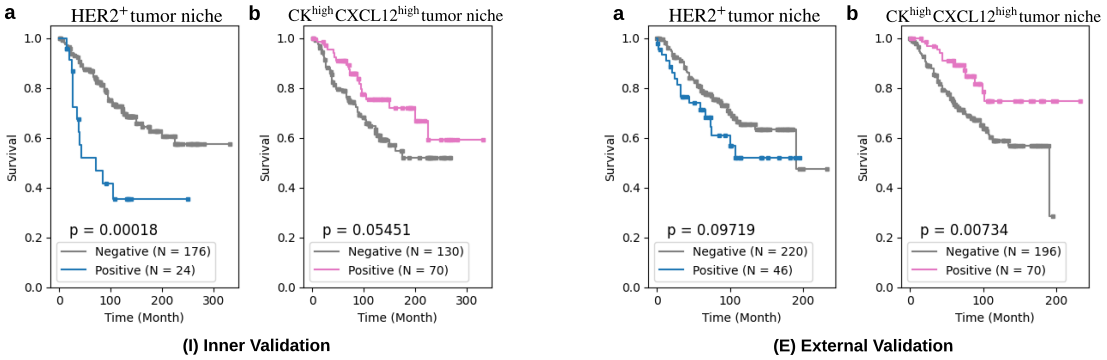


Fig. 15: **Validation of the prognostic impacts of TME patterns. (I)** Results in inner validation set (N=200). **(E)** Results in external validation set (N=266). **(a)** K-M survival plots of patients with (proportion $\geq 1\%$) and without (proportion $< 1\%$) the TME pattern, HER2$^+$ Tumor niche; **(b)** K-M survival plots of patients with (proportion $\geq 1\%$) and without (proportion $< 1\%$) the TME pattern, CK^{high}CXCL12^{high} tumor niche;

consistent trend from S1 to S7, which provides support for the reproducibility of our findings. Furthermore, our validation results confirm that S1$'$ in the external validation set exhibits significantly worse survival ($p<0.05$), consistent with the prognosis of S1 in the discovery set. Similarly, S7$'$ in the external validation set demonstrates better survival outcomes ($p<0.05$), mirroring the survival pattern of S7 observed in the discovery set. In the inner validation set, S1$'$ has a comparatively high hazard ratio, and S7$'$ has a comparatively low hazard ratio, while both lack statistical significance.

Moreover and importantly, we find that the intersection set between a clinical subtype and a BiGraph-derived subgroup exhibits exceptionally better (or worse) survivals in some cases (Fig. 7). We validate these results on the two validation sets. While the intersection of HER2+ patients and S1$'$ patients (Fig. 14. I.a, Fig. 14. E.a,), and the intersection of TNBC patients and S2$'$ patients (Fig. 14. I.b, Fig. 14. E.b,), show contradictory results in the validation sets, the intersection between HR+/HER2- patients and S7 patients (Fig. 14. I.c, Fig. 14. E.c,) indeed has an exceptionally better survival compared with the complement of S7 in HR+/HER2- patients and patients with other clinical subtypes, validating the findings in the discovery set.

2.6.2 Validation of the prognostic quality of TME patterns

As detailed in Section 2.5, we find two prognostic-relevant TME patterns in the discovery set: the HER2$^+$ tumor niche, associated with worse survival; and the CK^{high}CXCL12^{high} tumor niche, associated with better survival. We validate the prognostic impacts of these two patterns on both the inner validation set and external validation sets. To estimate the expression of the two TME patterns of interest in unseen patients, we map the subtrees derived from their cellular graphs to the closest TME pattern (see Section 5.7.2). Patients in the validation sets are categorized as "positive" or "negative" based on the expression of the studied TME pattern, with a 1 % proportion as the cut-off point as before. Results show that HER2$^+$ tumor niche positive expressed patients indeed exhibit worse survival outcomes than negatives, in both the inner validation set ($p = 0.0002$) and external validation set ($p = 0.097$), with the distinction being statistically significant only in the inner validation set (Fig. 15. I.a, and Fig. 15. E.a). On the other hand, CK^{high}CXCL12^{high} tumor niche positive patients demonstrate better survival outcomes than the remaining in both inner validation set ($p = 0.055$) and external validation set ($p = 0.007$), with the distinction being statistically significant only in the external validation set (Fig. 15. I.a, and Fig. 15. E.a). Additionally, the relationship between these two prognosis-relevant TME patterns and tumor grade is validated. HER2$^+$ tumor niche is statistically

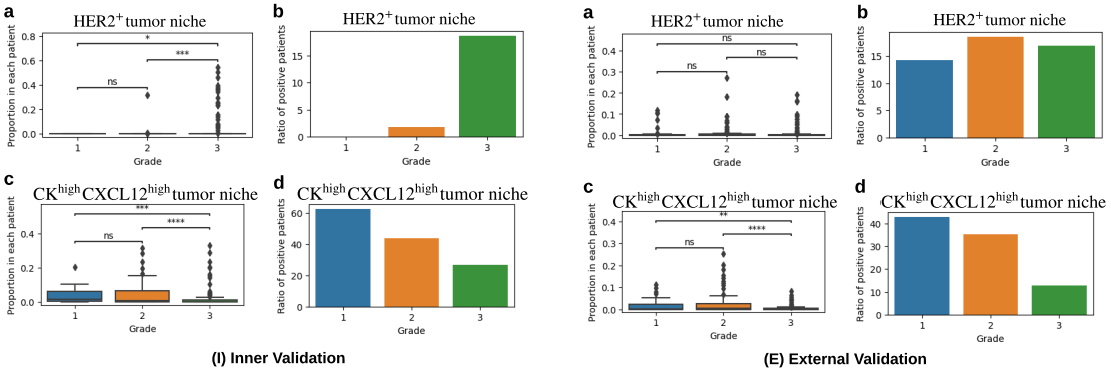


Fig. 16: **Validation of relationship between TME patterns and tumor grade. (I)** Results in inner validation set (N=200). **(E)** Results in external validation set (N=266). **(a)** The proportion of HER2$^+$ tumor niche per patient for tumor grade 1, 2, and 3; **(b)** The ratio of patients presenting HER2$^+$ tumor niche (i.e., proportion $> 1\%$) patient in grade 1, 2, and 3 groups; **(c)** The proportion of CK^{high}CXCL12^{high} tumor niche per patient for tumor grade 1, 2, and 3; **(d)** The ratio of patients presenting CK^{high}CXCL12^{high} tumor niche in grade 1, 2, and 3 groups.

significantly over-expressed in G3 patients in the inner validation set (Fig. 16.I. a), consistent with the results on the discovery set. However, there are no significant differences in the external validation set (Fig. 16.E. a). On the other hand, both inner validation set and external validation set results confirm that CK^{high}CXCL12^{high} tumor niche is statistically significantly under-expressed in G3 patients (Fig. 15. I.c, and Fig. 15. E.c), supporting our findings.

2.7 Comparison with alternative methods

The primary insight behind BiGraph is to establish a link between patient-specific cellular features (based on the spatial distribution of cell types) and the population-level distribution by utilizing a graph kernel function. In the previous discussions, we introduced the Soft-WL subtree kernel, which measures inter-patient similarity by assessing the abundance of TME patterns representing local structures within TMEs. This naturally prompts the question: are there alternative approaches to measuring inter-patient similarity giving similar or novel discoveries? We explore the following alternative methods: (i) using the abundance of cell types, which employs information on the concentration of cell types but disregards their spatial organization; (ii) using pairwise proximity of cell types, which only considers pairwise inter-cellular spatial relationships; and (iii) using the conventional WL subtree kernel, which employs the counting of subtrees as a metric similar to our approach while conducting a more stringent comparison of graphs. All subsequent processes in the analysis, including the construction of a population graph, community detection, and survival analysis, are executed in a manner consistent with previous sections.

2.7.1 Abundance of cell types

In this approach, each patient is characterized by a histogram of cells, and inter-patient similarity is determined by the normalized inner product of corresponding histograms. This approach disregards spatial information and inter-cellular interactions. The community detection algorithm identifies seven patient subgroups in the population graph, denoted as S1 to S7. Notably, only S1 exhibits significantly worse survival (see Fig. 17.a). The S1 subgroup identified here and that found by the Soft-WL subtree kernel reveals a large overlap, with an intersection over union (IoU) score of 0.85. Furthermore, S1 identified here is characterized by an over-presentation of HER2$^+$ cells, aligning with the HER2$^+$ tumor

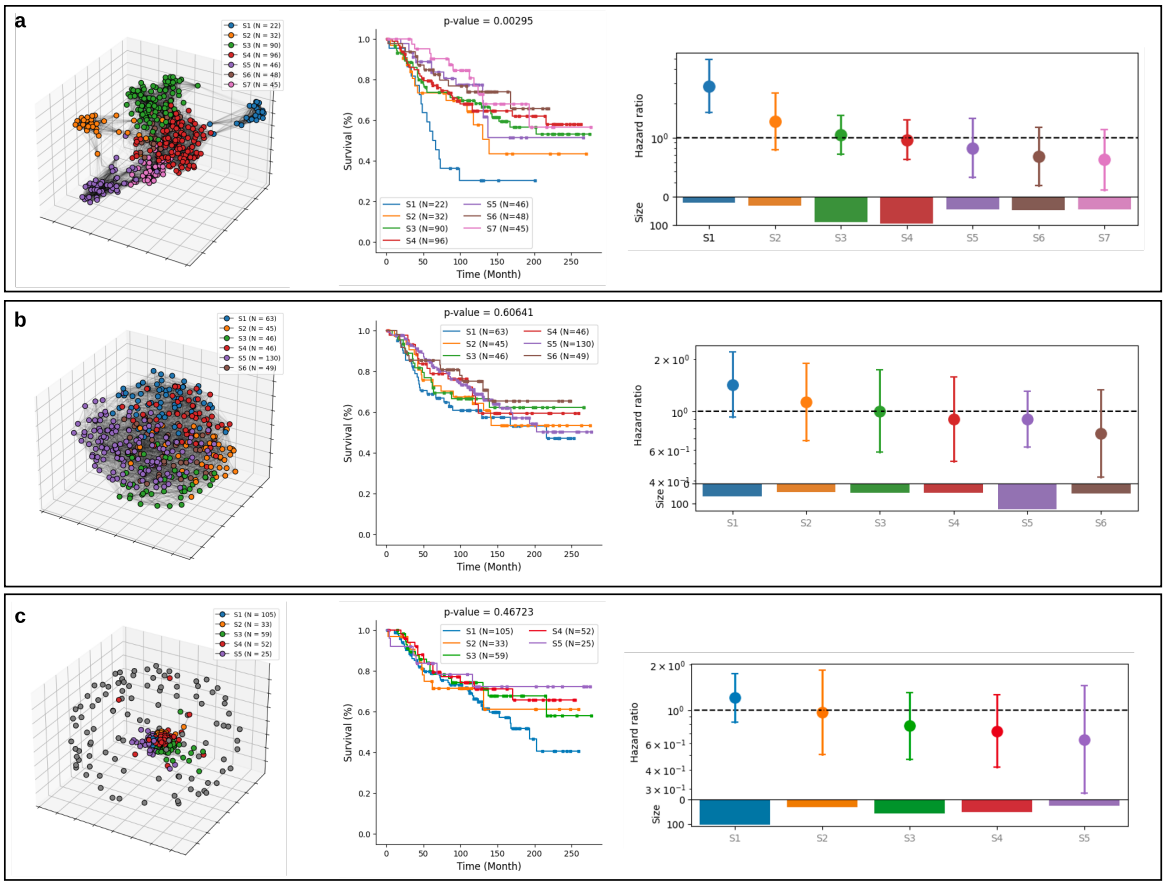


Fig. 17: **Validation of the risk stratification results. (a)** Risk stratification results from characterizing patients by the abundance of cell types on the discovery set. From left to right, the figure shows a population graph with nodes painted by patient subgroup labels; K-M survival plots of patient subgroups; and hazard ratios estimated by the Cox proportional model. **(b)** Risk stratification results from characterizing patients by the pairwise distance of cell types on the discovery set, whose layout is the same as (a). **(c)** Risk stratification results of WL subtree kernel on the discovery set, which layout is same as (a).

niche, characteristic in S1 found by Soft-WL subtree kernel. However, the S7 subgroup identified by the Soft-WL subtree kernel, characterized by TME patterns involving multiple types of cells, does not align with any of the patient subgroups identified using the abundance of cell types.

2.7.2 Pairwise proximity of cell phenotypes

In this approach, each patient is characterized by the average proximity score – which increases as the spatial distance decreases – between any possible pairs of cell phenotypes (see Equation (6)). While this method considers spatial information to some extent, it is limited to pairwise relationships. The resulting population graph does not demonstrate any clear clusters, and the community detection algorithm identifies six patient subgroups, none of which exhibit distinct survival outcomes (see Fig. 17.b).

2.7.3 WL subtree kernel

The Soft-WL subtree kernel, presented in this work, measures inter-patient similarity by identifying identical TME patterns across patients, each corresponding to a cluster of similar subtrees. In contrast, the classic WL subtree kernel [39] iteratively identifies *isomorphic* subtrees across graphs (see Section 5.8). We explore two formulations of the WL subtree kernel: one that considers isomorphic subtrees found in all previous iterations and accumulates the similarity (see Equation (8)), and another that only considers isomorphic subtrees found at the last iteration (see Equation (9)). Results indicate that the former formulation yields almost the same results as characterizing patients by the abundance of cell types (see Extended Data Fig. 8). This is attributed to the stringent criteria of isomorphic comparison. As iteration progresses, the depth of subtrees to check increases, and isomorphic subtrees become more rare. Thus, iteration zero (where a subtree is a single node) dominates the similarity score. This phenomenon also poses challenges to the latter formulation by leading to the *diagonal dominance* problem [52], where a cellular graph is only identical to itself, not similar to any others. The population graph derived shows a substantial number of isolated nodes, representing patients dissimilar to any others. The community detection algorithm identifies five patient subgroups, none with distinct survival outcomes (see Fig. 17.c). These results suggest that the stringent comparison employed in the conventional WL subtree kernel may hinder the discovery of important patterns (e.g., patient subgroups) in a cohort. Additionally, the feature space of the WL subtree kernel has an extremely high dimension (see Extended Data Fig. 9), making the interpretation (i.e., identification of characteristic subtrees) prohibitive.

3 Discussion

The TME represents a fundamental piece in the study of cancer biology, and several recent studies indicate its promise for discovering biomarkers for prognosis and even treatment resistance [3, 4, 19, 20, 53–55]. However, the number of broadly applicable biomarkers from pathology that provide strong prognosis information, while generalizing to different studies, remains limited. Additionally, while some useful metrics, such as tumor-infiltrating lymphocytes (TILs) score [56, 57], tumor-immune mixing score [15], presence of tertiary lymphatic structure (TLS) [49, 58], etc, have been derived from pathology using traditional methods of segmentation tools (e.g. [59]) and spatial statistics [15], driven mostly by hypotheses relying on domain expertise, the discovery of data-driven biomarkers for prognosis has remained unexplored.

Modern deep learning methods, including graph neural networks, constitute an appealing alternative for data-driven biomarkers for prognosis [21, 22]. Yet, their employment in real-world clinical settings remains constrained by the lack of rigorous external validation [35, 36] and limited explainability [25, 26, 28, 29, 32, 33]. Instead, this study introduces a data-driven methodology centered on breast cancer, that relies on two simple observations: *i)* representative phenotypic patterns of the TME can be adaptively learned by studying the underlying patient-specific cellular graph constructed from multiplexed tissue image data, and *ii)* the relative abundance of such patterns in different patients can be employed to provide a measure of their similarity, from which a population-level graph can be constructed. This bi-level process allows us to obtain – automatically and in an unsupervised manner – different patient subgroups, with potentially distinct prognoses, either better or worse. Compared with other data-driven methods, such as graph-based deep learning, a significant strength of our methodology is the complete transparency of the features that provide the risk stratification and characterize better and worse survival. These features are dubbed as *TME patterns*, and they represent similar subtrees in the TME; each corresponding to a local cellular neighborhood with unique cellular composition and spatial organization. As a result, we can easily unveil the most over-expressed structures that characterize better (and worse) survival, which elucidates the underlying associations between TME patterns and patients' prognosis, as demonstrated in the Results section. Relying on more complex models, such as those based on graph neural networks, would have made it significantly more challenging – if not altogether impossible – to characterize such biomarkers clearly and easily.

Another contribution of this work is the formalization of the Soft-WL subtree kernel, a novel relaxation of the WL subtree kernel [39]. This kernel method efficiently captures and compares local structures in complex cellular graphs with numerous cells and various phenotypes. It surpasses other simpler methods, such as quantifying the abundance of different types of cells, by considering the spatial co-localization and interactions of cells. On the other hand, it addresses the diagonal dominance [52] issue of the conventional WL subtree kernel, providing a smoother comparison of subtrees. A limitation of this method is the over-smoothing issue [60, 61] that is shared by many other graph learning methods. As the graph convolution progresses and subtree depth increases, subtrees tend to become increasingly similar to each other, preventing us from capturing higher-order structures in the graphs. We believe that the BiGraph could be improved further, for example by replacing the Soft-WL subtree kernel with more advanced graph kernel methods.

In this study, we provide a data-driven risk stratification for a cohort of breast cancer patients – featuring seven subgroups of breast with varying prognostic patterns. This risk-stratification scheme is furthermore proved to be robust in both an inner validation set (a subset of patients randomly selected from the original cohort); and an external validation set (a completely independent dataset collected with different acquisition protocols and antigen panels). Patients in both validation sets studied (internal and external) mirror the prognosis of those with similar TMEs in the discovery set. Furthermore, and importantly, by comparing this new risk stratification system with the standard clinical subtyping system based on the expression of HR and HER2, we found that our system provides supplementary information that enhances the risk stratification capacity of the conventional guidelines.

By analyzing the TME patterns that characterize each patient subgroup, we observe that HER2$^+$ tumor niches are shown to be strongly associated with worse survival, which is also validated in validation sets. As shown in many previous works, HER2 is an epidermal growth factor receptor, and the over-expression has been associated with a more aggressive breast cancer phenotype with decreased survival [62–65]. Our study further is consistent with this observation, but it is found in an entirely data-driven manner. On the other hand, we found that tumor niches expressing high levels of cytokeratins and CXCL12 are associated with better outcomes. This finding is particularly noteworthy as the prognostic impact of the interplay between these proteins, especially in breast cancer, is underexplored. While some literature discusses their individual impacts, controversial evidence on their prognostic impact exists: some demonstrate its association with the down-regulation of genes linked to metastasis inhibition, promoting invasion and migration of breast cancer cells [66, 67], while others suggest that higher expression of CXCL12 is associated with better survival, suppressing tumor growth and invasion in both breast [68] and pancreatic cancer [69]. The prognostic impact of cytokeratin is even more complicated due to the diverse range of subtypes. CK5/6, CK14, and CK20 are shown to be positively associated with high tumor grade of breast cancer [70, 71], and CK17 is also linked to poor clinical outcomes in breast cancer patients [72]. Conversely, down-regulation of CK8/18 has been reported to promote breast cancer progression, implying a favorable prognostic impact of CK8/18 [73–75]. Our study provides a new insight: that the co-expression of high levels of cytokeratin and CXCL12 is associated with better breast cancer prognosis. Importantly, this observation is found in an automatic, data-driven manner. The intricate mechanism underlying this association warrants further investigation.

4 Conclusion

BiGraph employs a bi-level graph learning scheme to systematically integrate information across different scales: from local cell distribution to patient populations. It allows for data-driven and explainable risk stratification and biomarker discoveries even in small and heterogeneous cohorts. While the results are centered on breast cancer, this methodology holds the potential to enhance our understanding of other diseases and facilitate the development of personalized care.

5 Methods

This section introduces the details of the methodology developed and applied in our study.

5.1 Preliminaries

We first present some key notation and definitions concerning graphs to facilitate the introduction of other methodologies.

A *graph* $G = (\mathbb{V}, \mathbb{E})$ is defined by a tuple of a set of nodes $\mathbb{V}$ and a set of edges $\mathbb{E} \subseteq \{\{u, v\} \subseteq \mathbb{V} \mid u \neq v\}$. The set of nodes and the set of edges for a graph G can be written as $V(G)$ and $E(G)$, respectively. For each node v, the set of nodes with an edge connected to a node v is defined as its neighborhood, denoted as $N(v)$. The structure of a graph can be fully characterized by its adjacency matrix, denoted as A. An adjacency matrix with only binary entries represents a binary graph. Conversely, an adjacency matrix with continuous scalar entries represents a weighted graph, where A_{uv} indicates the weight of the edge connecting nodes u and v. Two graphs, G_1 and G_2, are deemed *isomorphic* if there exist a bijection $\varphi : V(G_1) \to V(G_2)$, such that $(\varphi(u), \varphi(v)) \in E(G_2)$ for any $(u, v) \in E(G_1)$.

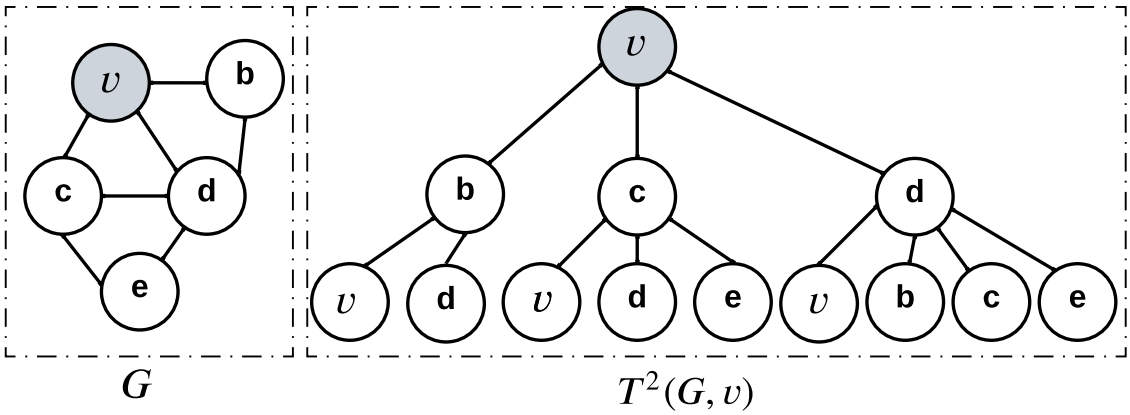


Fig. 18: **An illustration of a subtree rooted at node v (shaded) with a depth of 2.**

For a graph G, a *subtree* can be derived given a root node v, with the neighbors $N(v)$ serving as its children. This subtree can be expanded by iteratively adding the neighbors of the children, akin to increasing the depth of a subtree. A subtree rooted at v with a depth of h is denoted $T^h(G, v)$ (see Fig. 18). Two subtrees are considered *isomorphic* if they exhibit identical node sets at every level.

5.2 Construction of cellular graphs from TMA slides

The construction of the cellular graph involves representing the tumor microenvironment (TME) based on spatial cell coordinates and cell phenotypes. Each node in the cellular graph represents a single cell. The cell phenotype serves as the label of each corresponding node in the graph, determined by clustering multi-channel protein expression data. In this study, we use the published phenotype results of the discovery set [16] and the validation set [17], respectively. Edges in the cellular graph represent inter-cellular connections. Previous approaches typically employed a fixed distance threshold (d_0) to create a binary cellular graph, where two cells are connected only if the distance between them is smaller than d_0. However, there is no consensus on the appropriate value for d_0 and various choices have been employed, such as 4 μm [30], 8 μm [16], and 32 μm [15]. Additionally, some of the works count the distance between cell centroids, while others consider the minimal distance. To provide a more comprehensive representation of inter-cellular interaction, our approach constructs a weighted cellular

graph, wherein all possible pairs of cells are connected and the strength of the connection decreases with increasing inter-cellular distance. The edge weight (i.e., w_{uv}) is calculated via a Gaussian kernel given by

$$w_{uv} = \exp(-\alpha \cdot d^2), \tag{1}$$

where d represents the spatial (Euclidean) distance between centers of cells u and v in micrometers (μm), and α is a parameter set to 0.01 to calibrate the edge weights. The choice of α determines how fast these weights decrease with distance. As a concrete example, cells with a distance smaller than 3 μm have strong connections with edge weights higher than 0.9, while cells with a distance larger than 22 μm have weak connections with edge weights smaller than 0.01. In cases where a patient has multiple tissue microarray (TMA) cores, a separate cellular graph is constructed based on each core image, and then these individual graphs are combined as disconnected components to form a larger cellular graph representing the entire TME for that patient. This approach makes it possible to accommodate patients with multiple images while preserving the individual characteristics of each cellular graph and maintaining the entire methodology unaltered.

5.3 Soft-WL subtree Kernel

In this section, we delve into the Soft-WL subtree kernel employed in our work, which is a relaxation of the well-known Weisfeiler-Lehman (WL) kernel [39]. The WL subtree kernel, as detailed in Section 5.8, relies on isomorphic subtrees as a similarity measure across graphs, and is only applicable to binary graphs. The Soft-WL subtree kernel is designed to handle weighted and complete cellular graphs and provides a smoother comparison between subtrees. Fig. 19 shows a toy example of the Soft-WL subtree kernel. It comprises three main steps: (1) generation of subtrees (see Fig. 19.a); (2) identification of TME patterns by clustering subtrees (see Fig. 19.b); and (3) measurement of inter-patient similarity based on the abundance of subtrees (see Fig. 19.c). We detail each of them below.

5.3.1 Generation of subtrees

Given that the cellular graph constructed is weighted and complete, with each cell connected with all others with varying weights, the resulting subtrees are typically large and weighted, leading to a rare occurrence of isomorphic subtrees. Instead, a feature embedding is computed for each subtree through a graph convolution process. Formally, we consider a cellular graph $G = (\mathbb{V}, \mathbb{E})$ with a weighted adjacency matrix A. Noticeably, self-loops are added to each node, and thus the diagonal entries of A are all 1. Each node, v, is initially assigned an attribute, which is the one-hot encoding of its cell phenotype, denoted as $x_v^{(0)} \in \mathbb{R}^m$, where m is the number of unique cell phenotypes. The node attribute is then iteratively updated as follows:

$$x_v^h = x_v^{h-1} + \sum_{u \in \mathbb{V} \setminus \{v\}} x_u^{h-1} A_{uv}, \tag{2}$$

where h is the iteration index. Intuitively, this graph convolution process enables the updated node feature to incorporate not only its phenotype information but also that of its neighbors. Thus, the updated node feature, x_v^h, is considered the feature embedding of the subtree rooted at v with a depth of h, denoted as $T^h(G, v)$. In this way, this feature embedding encodes the structure of a local cellular neighborhood surrounding v. Note that the cellular neighborhood does not have a clear boundary, but one can decide on a boundary by applying a threshold τ to the impact of neighborhood nodes on this feature embedding. To be more specific, Equation (2) can be rewritten as

$$x_v^h = x_v^0 (A^h)_{vv} + \sum_{u \in \mathbb{V} \setminus \{v\}} x_u^0 (A^h)_{uv}, \tag{3}$$

where A^h denotes the hth power of the adjacency matrix A, and x_v^h is A_{uv}^h denotes the influence of a node u to x_v^h. Then, any node u with $(A^h)_{uv} > \tau$ is considered part of the cellular neighborhood. This hyper-parameter, τ, is user-defined, and we set it as 0.01 in the experiments. In this way, N subtrees

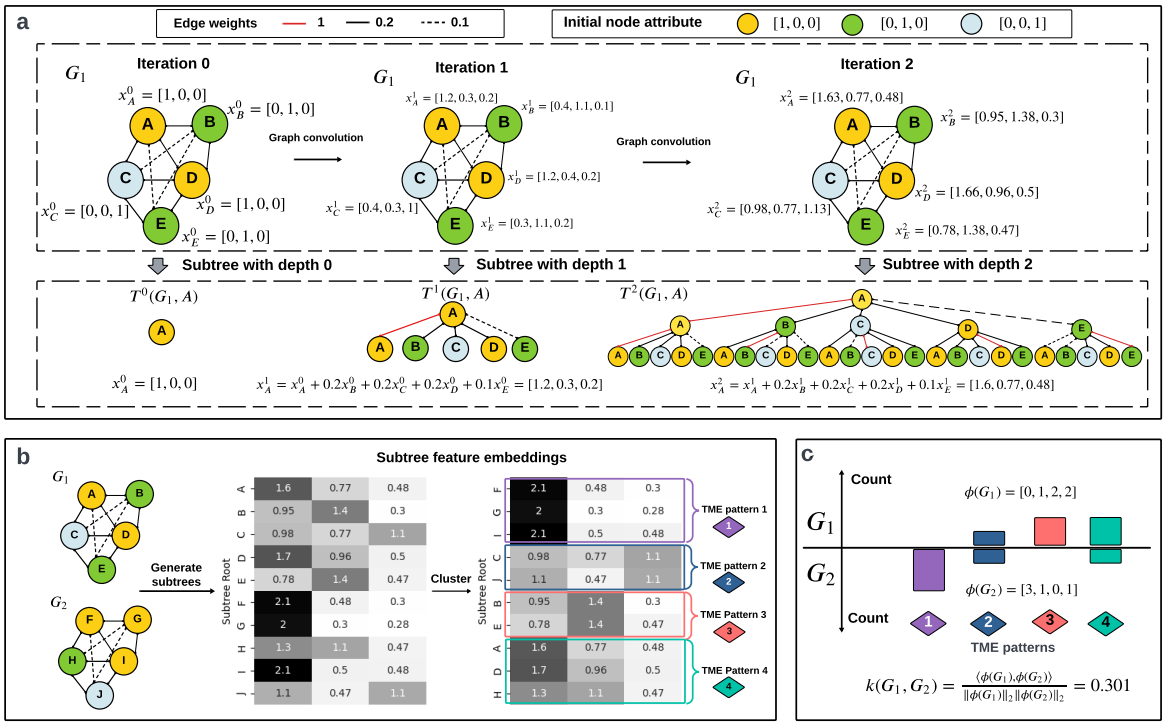


Fig. 19: **A toy example of the Soft-WL subtree kernel applied to two small cellular graphs. (a)** The process of generating subtrees from the small cellular graph G_1 with five nodes is illustrated. Nodes are initially assigned attributes representing the cell type (denoted by node color), and these attributes are iteratively updated through a graph convolution process. At each iteration, the updated node attribute serves as the feature embedding of a subtree rooted at that node with a designated depth. Subtrees rooted at node A with depths 0, 1, and 2 are displayed. **(b)** Clustering of the ten subtrees generated from two cellular graphs, G_1 and G_2, each rooted at a single node with a depth of 2. The resulting four clusters correspond to four TME patterns. **c** The similarity score between G_1 and G_2, denoted as $k(G_1, G_2)$, determined by the normalized inner product of their histograms of the four TME patterns.

can be generated from a cellular graph with N cells, each rooted at a single cell, and having a feature embedding derived from the graph convolution process above. The feature embedding of each subtree represents the structure of a local cellular neighborhood surrounding each cell, which is the root of the tree.

5.3.2 Identification of TME patterns

Unlike the traditional WL kernel [39], which focuses on the distribution of subtrees as a similarity measure, the Soft-WL subtree kernel clusters similar subtrees $T^h(G, v)$ based on their feature embeddings $x_v^{(h)}$. Moreover, instead of comparing subtrees embedding themselves, we will compare the distribution of *cell patterns* across patients as characterized by their embeddings.

To this end, we employ a clustering algorithm across subtrees from the entire discovery set. Although the clustering method can be general, we employ the PhenoGraph algorithm [41] since it has been shown to be less sensitive to hyperparameters in the context of single-cell data. The official Python implementation from PhenoGraph (https://github.com/jacoblevine/PhenoGraph) is used. Each resulting cluster is considered a *TME pattern*, which thus characterizes similar subtrees (i.e. those having similar feature embeddings). Furthermore, each TME pattern is assigned a signature, denoted `sign`,

given by the cluster centroid. More precisely, for each ith out of K clusters,

$$\texttt{sign}(i) = \frac{\sum_{l(v)=i} x_v^h}{|\{T^h(v) \in \mathbb{T} \mid l(v) = i\}|}, \tag{4}$$

where x_v^h is the subtree feature embedding of node (cell) v, $\mathbb{T}$ denotes the set of subtrees derived from the entire discovery set, and $l : V \to \{1, \ldots, K\}$ denotes a function mapping a subtree to its assigned cluster identity.

5.3.3 Inter-patient similarity measurement

The Soft-WL Subtree kernel calculates the inter-patient similarity score by comparing the abundance of these TME patterns across patients. Let $\phi(G_i)$ be the histogram of TME patterns in the graph of each subject G_i. This similarity is then given by

$$k(G_i, G_j) = \frac{\langle \phi(G_i), \phi(G_j) \rangle}{\|\phi(G_i)\|_2 \|\phi(G_j)\|_2}, \tag{5}$$

where $\| \cdot \|_2$ denotes the ℓ_2 (Euclidean) norm. The normalization ensures that the similarity score for any pairs of patients falls within the range of 0 to 1, with identical graphs achieving a maximal similarity of 1. The hyperparameter, h, controls the depth of subtrees and indirectly decides the size of the corresponding local cellular neighborhood. We set it as $h = 2$ in the experiments. As detailed in Section 5.3.1, the cellular neighborhood boundary is decided by the contributions of surrounding cells to the subtree embedding. The cellular neighborhoods contain a median of 16 cells, with a median radius (defined by the maximum distance between a subtree root and the cells within that cellular neighborhood) of 30 um.

5.4 Population graph and community detection

In the previous section, patient-specific graphs were constructed to characterize the local statistics of cell distribution in terms of recurrent patterns. In contrast, the population graph that we describe now is constructed to encode the similarities *across* patients. Each node of the population graph represents a patient, and every patient is connected to potentially all other patients through edges with scalar weights. Such weights represent the similarity between any two patients measured by the Soft-WL subtree method, given by Equation (5). Thus, the patient graph can be represented by the *Kernel matrix* $K \in \mathbb{R}^{N \times N}$ for the N patients $\{G_1, G_2, ..., G_N\}$ in the discovery set, and each entry K_{ij} is given by $K_{ij} = k(G_i, G_j)$ (Equation (5)).

The population graph is preprocessed in the following two steps before proceeding to detect sub-communities within this graph (as per instructions in [41]): (1) identifying the $k^\star$-nearest neighbors ($k^\star$-NN); (2) replacing the edge weight between two patients by the intersection-over-union of their respective $k^\star$-NN sets. The Louvain community detection method [42] (Python package "networkx" [76]) is then employed to detect communities within the population graph, where each community represents a patient subgroup characterized by high intra-group similarities in their TMEs. We set the hyperparameter $k^\star = 30$ in the experiments. Although this hyperparameter impacts the partition granularity, the patient subgrouping results show substantial consistency across various choices of $k^\star$ (see Extended Data Fig. 2).

5.5 Identification of characteristic patterns

The pattern histogram is normalized on a per-patient basis by dividing the occurrence of each pattern by the sum of occurrences of all patterns within that patient, yielding the proportion of each TME pattern in individual patients (see Extended Data Fig. 3). Consequently, the Hodges–Lehmann statistic [48] is employed to compare the distribution of a specific TME pattern's proportions among patients

inside and outside a patient subgroup (Extended Data Fig.4). A higher Hodges–Lehmann statistic indicates the increased expression of the studied pattern in the patient subgroup. A TME pattern is deemed "characteristic" if its Hodges–Lehmann statistic surpasses 50% of the maximum value within that patient subgroup, as highlighted in colored boxplots in Extended Data Fig. 4.

5.6 Prognosis analysis

The Kaplan-Meier (K-M) estimator [44] is used to generate survival plots for each patient subgroup. For the discovery set and inner validation set, the disease-specific survivals are considered. On the other hand, the overall survival is considered for the external validation set, since the disease-specific survival data is not available. The multivariate log-rank test [45] is used to compare their survivals. A Cox proportional hazard regression model [46] is used to calculate hazard ratios for each patient subgroup, with patients outside each patient subgroup serving as the baseline. In assessing the prognostic impact of a specific TME pattern, patients are stratified into positive and negative groups based on the expression of the target TME pattern, where a proportion of 1% is used as the threshold for categorization. The survival outcomes of positive and negative patients are compared using the pairwise log-rank test. All of these prognosis-related statistical methods are conducted with the Python implementation from "lifelines" [77].

5.7 Validation and generalization

The robustness and generalization of the major findings in the discovery set are rigorously assessed in both an inner validation set and an external validation set. The validation process encompasses mapping at the cellular, subtree, and patient levels. We detail each of them in the following.

5.7.1 Cell phenotyping system mapping

Cell phenotype mapping is only needed for the external validation set since it employs a different cell phenotyping system from that in the discovery set. Moreover, the external validation set also employs a different antigen marker panel, with approximately 50% of the antigens shared between two sets. Each cell, irrespective of its origin in the discovery set or the external validation set, is characterized by a continuous feature vector representing the expression profiling of the *shared* antigens. For a cell from the external validation, mapping is achieved by assigning it to the phenotypically closest cell phenotype (i.e., a cluster of cells) defined within the discovery set. The phenotypic distance between a cell and a cell cluster is given by the Euclidean distance between the cell's feature vector and the cluster centroid.

5.7.2 TME pattern mapping

For each new patient in the validation set with its corresponding cellular graph, subtree generation and feature computation are carried out in an analogous manner to that described in Section 5.3.1. Instead of re-conducting clustering on these subtrees from the validation set to find new patterns, they are mapped to the closest TME pattern (i.e., a cluster of subtrees) previously defined within the discovery set (i.e. allowing us to assign a new patient to a predefined subgroup). The distance between a subtree and a TME pattern is given by the Euclidean distance between the subtree's feature vector and the pattern's signature (i.e. cluster centroid).

5.7.3 Patient subgroup mapping

Once TME patterns are mapped, the similarity between patients across sets can be calculated using Equation (5). Given a new patient in the validation set, its top three most similar patients in the discovery set are identified. The new patient is then mapped to the patient subgroup via weighted majority voting among the three nearest neighbors. In this way, a corresponding patient subgroup, denoted as Si', can be identified for every group Si, both in the inner validation and external validation

sets. Survival analysis is performed on them to discern whether a mapped patient subgroup mirrors the survival outcome of its phenotypically similar patient subgroup defined within the discovery set.

5.8 Comparative methods

This section outlines alternative methods for measuring inter-patient similarity scores considering Equation (5) as a general formula, but where $\phi(G)$ varies for each method.

- Abundance of cell types: In this method, $\phi(G)$ represents the histogram of cell phenotypes within a patient (i.e., corresponding cellular graphs).
- Pairwise proximity of cell phenotypes: Given a dataset with n unique cell phenotypes, $\phi(G)$ is a feature vector with $\frac{n(n+1)}{2}$ entries, each corresponding to the average proximity score between a pair of cell phenotypes given by:

$$\text{proximity}(i,j) = \frac{\sum_{\{(u,v)\in E(G)|p(u)=i,p(v)=j\}} w_{uv}}{|\{(u,v)\in E(G) \mid p(u)=i, p(v)=j\}|}, \tag{6}$$

  where i and j are two cell phenotypes (where i may equal to j), $E(G)$ denotes the set of edges, w_{uv} is the proximity score between cell u and v, akin to the edge weight given by Equation (1), and p is a function mapping each node (i.e., cell) to the corresponding cell phenotype.
- WL-subtree kernel [39]: This kernel measures the similarity between two graphs by counting the occurrence of isomorphic subtrees in them. Notably, the WL subtree kernel only takes binary graphs; therefore: weighted cellular graphs are converted to binary graphs by applying a threshold of 0.01 to the edge weights. This threshold connects two cells if the distance between their centroids is less than 21 um. The identification of isomorphic subtrees involves a *color refinement* procedure, where node colors are iteratively updated based on their neighbors. To be more specific, each node is initially assigned a color based on its cell phenotype, denoted as $\sigma^0(v)$. The node color can be either a discrete alphabet or an integer. The node color is subsequently updated as follows:

$$\sigma^h(v) = \texttt{relabel}(\sigma^{h-1}(v), \texttt{sort}(\{\{\sigma^{h-1}(u) \mid u \in N(v)\}\})), \tag{7}$$

  where $\sigma^h(v)$ is the node color at iteration h, and $N(v)$ represents the neighbors of v. The function `sort` collects and sorts the multiset of colors of $N(v)$, and the function `relabel` maps the pair $(\sigma^{h-1}(v), \texttt{sort}(\sigma^{h-1}(u) \mid u \in N(v)))$ to a unique color not used in previous iterations. Two nodes with the same color after iteration h imply the isomorphism of the two subtrees rooted at these nodes with depth h. Notably, the color refinement is conducted universally for all the graphs in the dataset rather than individually, ensuring a consistence color system across all graphs. In this case, the feature $\phi(G)$ represents the histogram of unique colors, and two slightly different formulations are considered. The first formulation considers the occurrence of colors from all previous iterations, given by

$$\phi(G) = [c(G,\sigma_{01}), ..., c(G,\sigma_{0|\Sigma_0|}), ..., c(G,\sigma_{h1}), ..., c(G,\sigma_{h|\Sigma_h|})], \tag{8}$$

  where σ_{hi} is a unique color appearing at the h-th iteration of color refinement, $\Sigma_h = \{\sigma_{h1}, ..., \sigma_{h|\Sigma_h|}\}$ denotes the set of unique colors at the h-th iteration, and $c(G,\sigma_{hi})$ denotes the occurrence of color σ_{hi} in G. The second formulation only considers the colors from the last iteration of color refinement, given by

$$\phi(G) = [c(G,\sigma_{h1}), ..., c(G,\sigma_{h|\Sigma_h|})], \tag{9}$$

  The hyperparameter h controls the depth of subtrees, similar to its counterpart in the Soft-WL subtree kernel. We set it to 2 in the experiments, consistent with the setting used for the Soft-WL subtree kernel. The WL subtree kernel is implemented with the Python package "Grakel" [78].

5.9 Writing

The large language model (ChatGPT 3.5) was sporadically used to improve conciseness in the writing of this paper.

Acknowledgement

This study was supported by the National Institutes of Health grant R01CA138264.

Data availability

The multiplexed image data and corresponding clinical data used in this work are publically available at https://doi.org/10.5281/zenodo.5850952 (Danenberg et al [16]) and https://doi.org/10.5281/zenodo.3518284 (Jackson et al [17]).

Code availability

The code used for implementing the BiGraph method and all the analysis in this paper is available at https://github.com/Sulam-Group/BiGraph4TME.git

References

[1] Soysal, S. D., Tzankov, A. & Muenst, S. E. Role of the tumor microenvironment in breast cancer. *Pathobiology* **82**, 142–152 (2015).

[2] Hinshaw, D. C. & Shevde, L. A. The tumor microenvironment innately modulates cancer progression. *Cancer research* **79**, 4557–4566 (2019).

[3] Khalaf, K. *et al.* Aspects of the tumor microenvironment involved in immune resistance and drug resistance. *Frontiers in immunology* **12**, 656364 (2021).

[4] Wu, X., Deng, Z. & Zhao, Q. Immunotherapy improves disease prognosis by affecting the tumor microenvironment: A bibliometric study. *Frontiers in Immunology* **13**, 967076 (2022).

[5] Glass, G., Papin, J. A. & Mandell, J. W. Simple: a sequential immunoperoxidase labeling and erasing method. *Journal of Histochemistry & Cytochemistry* **57**, 899–905 (2009).

[6] Black, S. *et al.* Codex multiplexed tissue imaging with dna-conjugated antibodies. *Nature protocols* **16**, 3802–3835 (2021).

[7] Keren, L. *et al.* Mibi-tof: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science advances* **5**, eaax5851 (2019).

[8] Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* **11**, 417–422 (2014).

[9] van Dam, S., Baars, M. J. & Vercoulen, Y. Multiplex tissue imaging: spatial revelations in the tumor microenvironment. *Cancers* **14**, 3170 (2022).

[10] Hu, B., Sajid, M., Lv, R., Liu, L. & Sun, C. A review of spatial profiling technologies for characterizing the tumor microenvironment in immuno-oncology. *Frontiers in Immunology* **13**, 996721 (2022).

[11] Elhanani, O., Ben-Uri, R. & Keren, L. Spatial profiling technologies illuminate the tumor microenvironment. *Cancer cell* (2023).

[12] Wagner, J. *et al.* A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* **177**, 1330–1345 (2019).

[13] Sorin, M. *et al.* Single-cell spatial landscapes of the lung tumour immune microenvironment. *Nature* **614**, 548–554 (2023).

[14] Massalha, H. *et al.* A single cell atlas of the human liver tumor microenvironment. *Molecular systems biology* **16**, e9682 (2020).

[15] Keren, L. *et al.* A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **174**, 1373–1387 (2018).

[16] Danenberg, E. *et al.* Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature genetics* **54**, 660–669 (2022).

[17] Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).

[18] Feng, Y. *et al.* Spatial analysis with spiat and spasim to characterize and simulate tissue microenvironments. *Nature Communications* **14**, 2697 (2023).

[19] Granier, C. *et al.* Tim-3 expression on tumor-infiltrating pd-1+ cd8+ t cells correlates with poor clinical outcome in renal cell carcinoma. *Cancer research* **77**, 1075–1082 (2017).

[20] Mi, H., Ho, W. J., Yarchoan, M. & Popel, A. S. Multi-scale spatial analysis of the tumor microenvironment reveals features of cabozantinib and nivolumab efficacy in hepatocellular carcinoma. *Frontiers in immunology* **13**, 892250 (2022).

[21] Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C. & Petersson, L. A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics* **95**, 102027 (2022).

[22] Gogoshin, G. & Rodin, A. S. Graph neural networks in cancer and oncology research: Emerging and future trends. *Cancers* **15**, 5858 (2023).

[23] Li, B., Nelson, M. S., Savari, O., Loeffler, A. G. & Eliceiri, K. W. Differentiation of pancreatic ductal adenocarcinoma and chronic pancreatitis using graph neural networks on histopathology and collagen fiber features. *Journal of Pathology Informatics* **13**, 100158 (2022).

[24] Graham, S. *et al.* Screening of normal endoscopic large bowel biopsies with interpretable graph learning: a retrospective study. *Gut* **72**, 1709–1721 (2023).

[25] Wang, H. *et al.* Ccf-gnn: A unified model aggregating appearance, microenvironment, and topology for pathology image classification. *IEEE Transactions on Medical Imaging* (2023).

[26] Abbas, S. F., Le Vuong, T. T., Kim, K., Song, B. & Kwak, J. T. Multi-cell type and multi-level graph aggregation network for cancer grading in pathology images. *Medical Image Analysis* **90**, 102936 (2023).

[27] Martin, N. G. *et al.* A graph based neural network approach to immune profiling of multiplexed tissue samples. *Annu Int Conf IEEE Eng Med Biol Soc.* 3063–3067 (2022).

[28] Pati, P. *et al.* Hierarchical graph representations in digital pathology. *Medical image analysis* **75**, 102264 (2022).

[29] Lu, W. *et al.* Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. *Medical Image Analysis* **80**, 102486 (2022).

[30] Wu, Z. *et al.* Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering* **6**, 1435–1448 (2022).

[31] Jiménez-Sánchez, D. *et al.* Naronet: Discovery of tumor microenvironment elements from highly multiplexed images. *Medical image analysis* **78**, 102384 (2022).

[32] Wang, Y. *et al.* Cell graph neural networks enable the precise prediction of patient survival in gastric cancer. *NPJ precision oncology* **6**, 45 (2022).

[33] Fu, X., Patrick, E., Yang, J. Y., Feng, D. D. & Kim, J. Deep multimodal graph-based network for survival prediction from highly multiplexed images and patient variables. *Computers in Biology and Medicine* **154**, 106576 (2023).

[34] Lee, Y. *et al.* Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering* 1–15 (2022).

[35] Kleppe, A. *et al.* Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer* **21**, 199–211 (2021).

[36] Echle, A. *et al.* Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer* **124**, 686–696 (2021).

[37] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

[38] Teneggi, J., Luster, A. & Sulam, J. Fast hierarchical games for image explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**, 4494–4503 (2022).

[39] Shervashidze, N., Schweitzer, P., Van Leeuwen, E. J., Mehlhorn, K. & Borgwardt, K. M. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research* **12** (2011).

[40] Szymiczek, A., Lone, A. & Akbari, M. R. Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. *Clinical genetics* **99**, 613–637 (2021).

[41] Levine, J. H. *et al.* Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).

[42] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

[43] Fruchterman, T. M. & Reingold, E. M. Graph drawing by force-directed placement. *Software: Practice and experience* **21**, 1129–1164 (1991).

[44] Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**, 457–481 (1958).

[45] Mantel, N. *et al.* Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* **50**, 163–170 (1966).

[46] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).

[47] De Winter, J. C., Gosling, S. D. & Potter, J. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods* **21**, 273 (2016).

[48] Rosenkranz, G. K. A note on the hodges–lehmann estimator. *Pharmaceutical statistics* **9**, 162–167 (2010).

[49] Schumacher, T. N. & Thommen, D. S. Tertiary lymphoid structures in cancer. *Science* **375**, eabf9419 (2022).

[50] Hammond, M. E. H. *et al.* American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). *Archives of pathology & laboratory medicine* **134**, e48–e72 (2010).

[51] Peterson, L. E. K-nearest neighbor. *Scholarpedia* **4**, 1883 (2009).

[52] Kriege, N. M., Johansson, F. D. & Morris, C. A survey on graph kernels. *Applied Network Science* **5**, 1–42 (2020).

[53] Whiteside, T. The tumor microenvironment and its role in promoting tumor growth. *Oncogene* **27**, 5904–5912 (2008).

[54] Yang, Y. *et al.* Tumor-infiltrating cytotoxic t cells and tumor-associated macrophages correlate with the outcomes of neoadjuvant chemoradiotherapy for locally advanced rectal cancer. *Frontiers in oncology* **11**, 743540 (2021).

[55] Mi, H. *et al.* Quantitative spatial profiling of immune populations in pancreatic ductal adenocarcinoma reveals tumor microenvironment heterogeneity and prognostic biomarkers. *Cancer research* **82**, 4359–4372 (2022).

[56] Locy, H. *et al.* Assessing tumor-infiltrating lymphocytes in breast cancer: a proposal for combining immunohistochemistry and gene expression analysis to refine scoring. *Frontiers in Immunology* **13**, 794175 (2022).

[57] Zhang, D. *et al.* Scoring system for tumor-infiltrating lymphocytes and its prognostic value for gastric cancer. *Frontiers in immunology* **10**, 71 (2019).

[58] Sautès-Fridman, C., Petitprez, F., Calderaro, J. & Fridman, W. H. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nature Reviews Cancer* **19**, 307–325 (2019).

[59] Wang, Z. *et al.* Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images. *IEEE transactions on medical imaging* **41**, 3952–3968 (2022).

[60] Rusch, T. K., Bronstein, M. M. & Mishra, S. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993* (2023).

[61] Zhang, X., Xu, Y., He, W., Guo, W. & Cui, L. A comprehensive review of the oversmoothing in graph neural networks. *CCF Conference on Computer Supported Cooperative Work and Social Computing* 451–465 (2023).

[62] Iqbal, N. & Iqbal, N. Human epidermal growth factor receptor 2 (her2) in cancers: overexpression and therapeutic implications. *Molecular biology international* **2014** (2014).

[63] Krishnamurti, U. & Silverman, J. F. Her2 in breast cancer: a review and update. *Advances in anatomic pathology* **21**, 100–107 (2014).

[64] Loibl, S. & Gianni, L. Her2-positive breast cancer. *The Lancet* **389**, 2415–2429 (2017).

[65] Yarden, Y. Biology of her2 and its importance in breast cancer. *Oncology* **61**, 1–13 (2001).

[66] Zielińska, K. A. & Katanaev, V. L. The signaling duo cxcl12 and cxcr4: Chemokine fuel for breast cancer tumorigenesis. *Cancers* **12**, 3071 (2020).

[67] Kang, H. *et al.* Stromal cell derived factor-1: its influence on invasiveness and migration of breast cancer cells in vitro, and its association with prognosis and survival in human breast cancer. *Breast Cancer Research* **7**, 1–9 (2005).

[68] Lv, Z.-D. *et al.* Cxcl12 chemokine expression suppresses human breast cancer growth and metastasis in vitro and in vivo. *International Journal of Clinical and Experimental Pathology* **7**, 6671 (2014).

[69] Roy, I. *et al.* Cxcl12 chemokine expression suppresses human pancreatic cancer growth and metastasis. *PloS one* **9**, e90400 (2014).

[70] Shao, M.-M. *et al.* Keratin expression in breast cancers. *Virchows Archiv* **461**, 313–322 (2012).

[71] McGinn, O. *et al.* Cytokeratin 5 alters β-catenin dynamics in breast cancer cells. *Oncogene* **39**, 2478–2492 (2020).

[72] Van De Rijn, M. *et al.* Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome. *The American journal of pathology* **161**, 1991–1996 (2002).

[73] Woelfle, U., Sauter, G., Santjer, S., Brakenhoff, R. & Pantel, K. Down-regulated expression of cytokeratin 18 promotes progression of human breast cancer. *Clinical cancer research* **10**, 2670–2674 (2004).

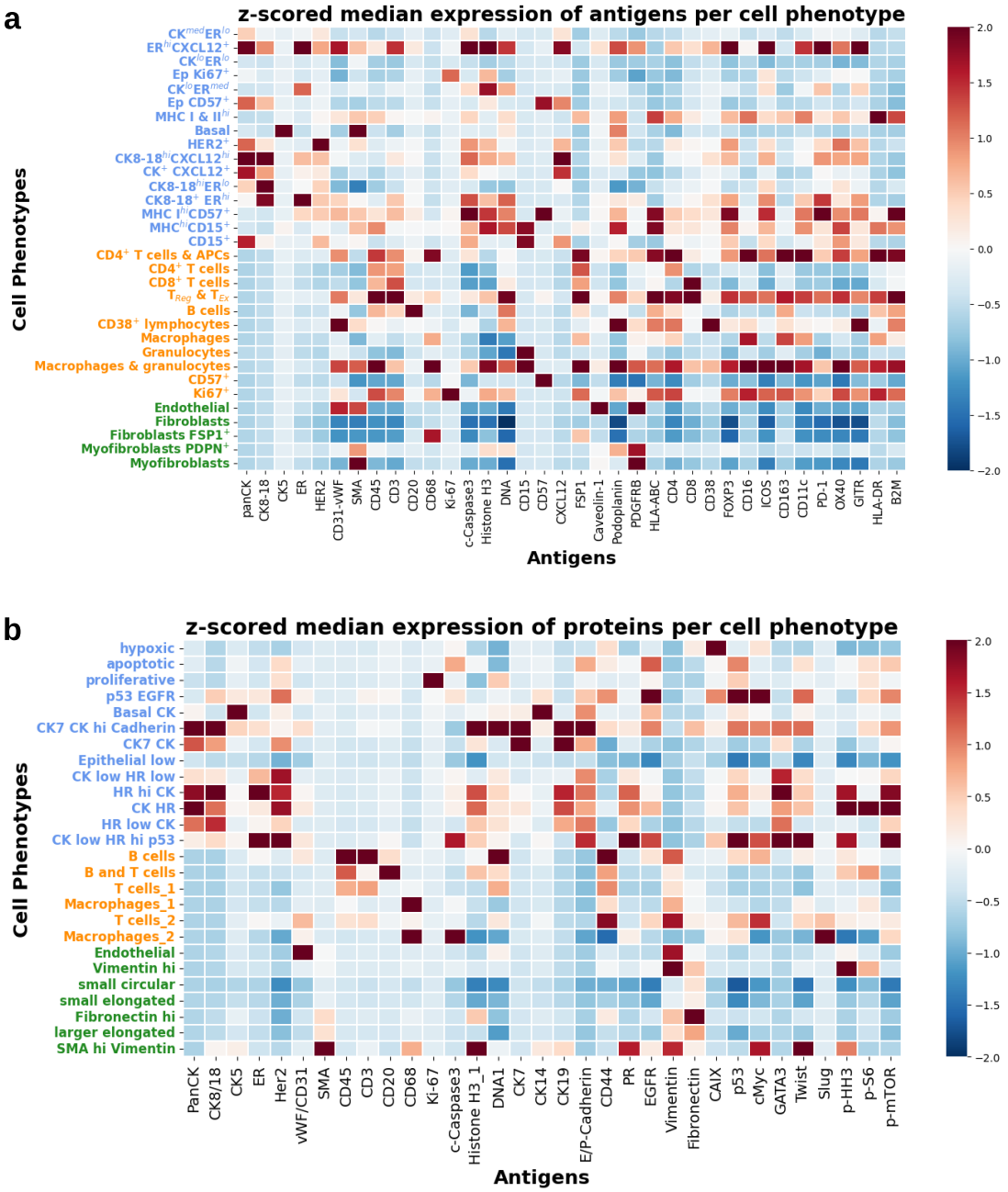
[74] Aiad, H. A. *et al.* Relationship of ck8/18 expression pattern to breast cancer immunohistochemical subtyping in egyptian patients. *ecancermedicalscience* **8** (2014).

[75] Menz, A. *et al.* Diagnostic and prognostic impact of cytokeratin 18 expression in human tumors: a tissue microarray study on 11,952 tumors. *Molecular Medicine* **27**, 1–16 (2021).

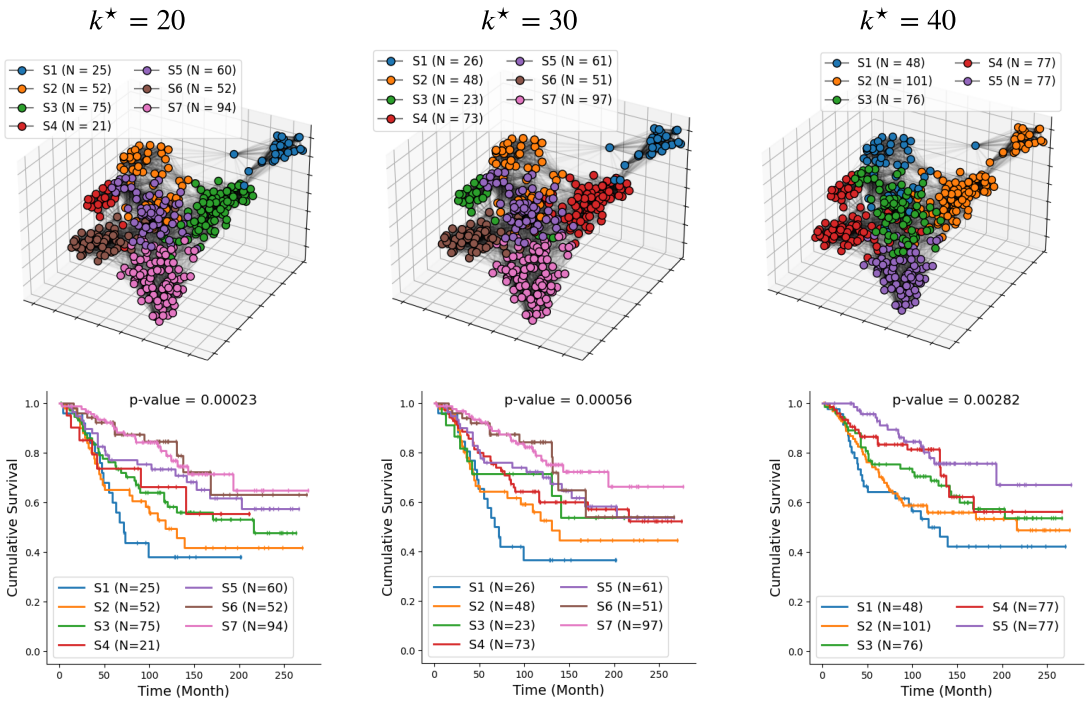
[76] Hagberg, A., Swart, P. & S Chult, D. Exploring network structure, dynamics, and function using networkx. Tech. Rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008).

[77] Davidson-Pilon, C. lifelines: survival analysis in python. *Journal of Open Source Software* **4**, 1317 (2019). URL https://doi.org/10.21105/joss.01317.

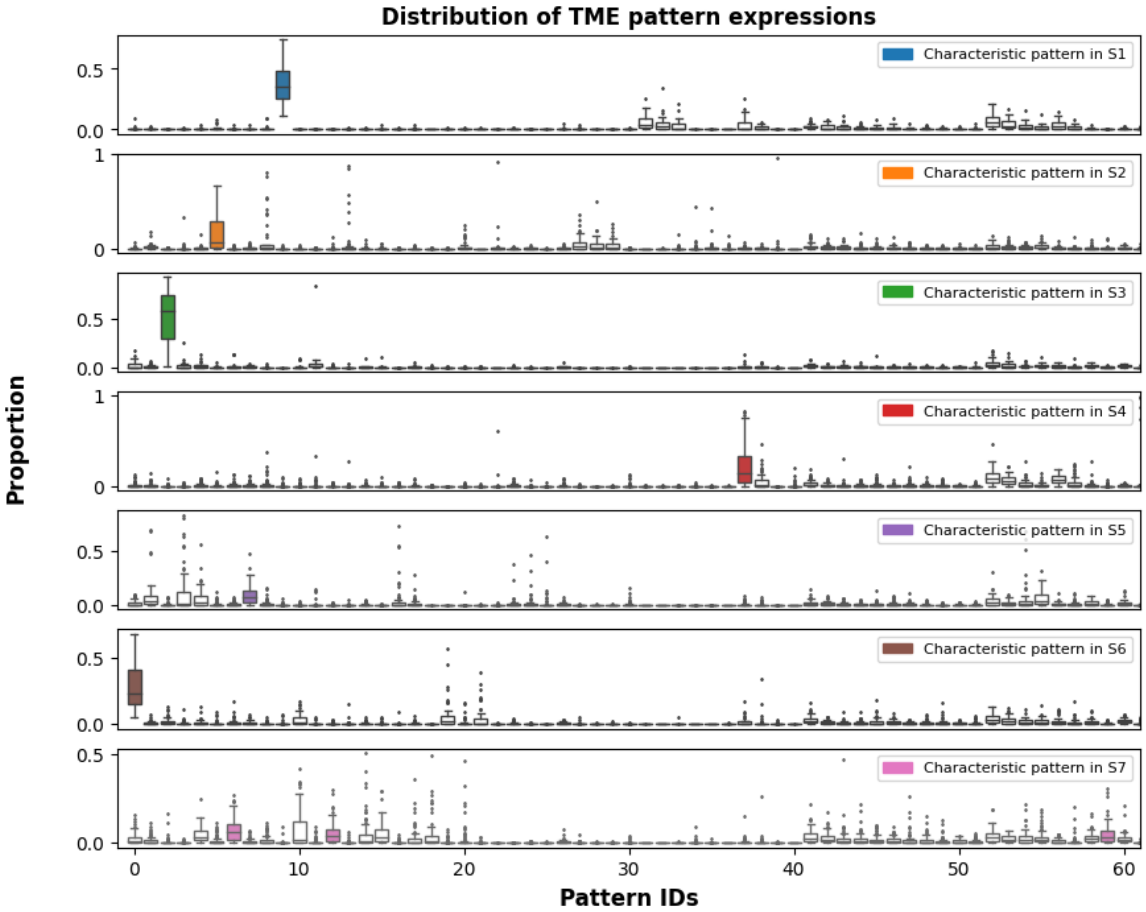
[78] Siglidis, G. *et al.* Grakel: A graph kernel library in python. *Journal of Machine Learning Research* **21**, 1–5 (2020).



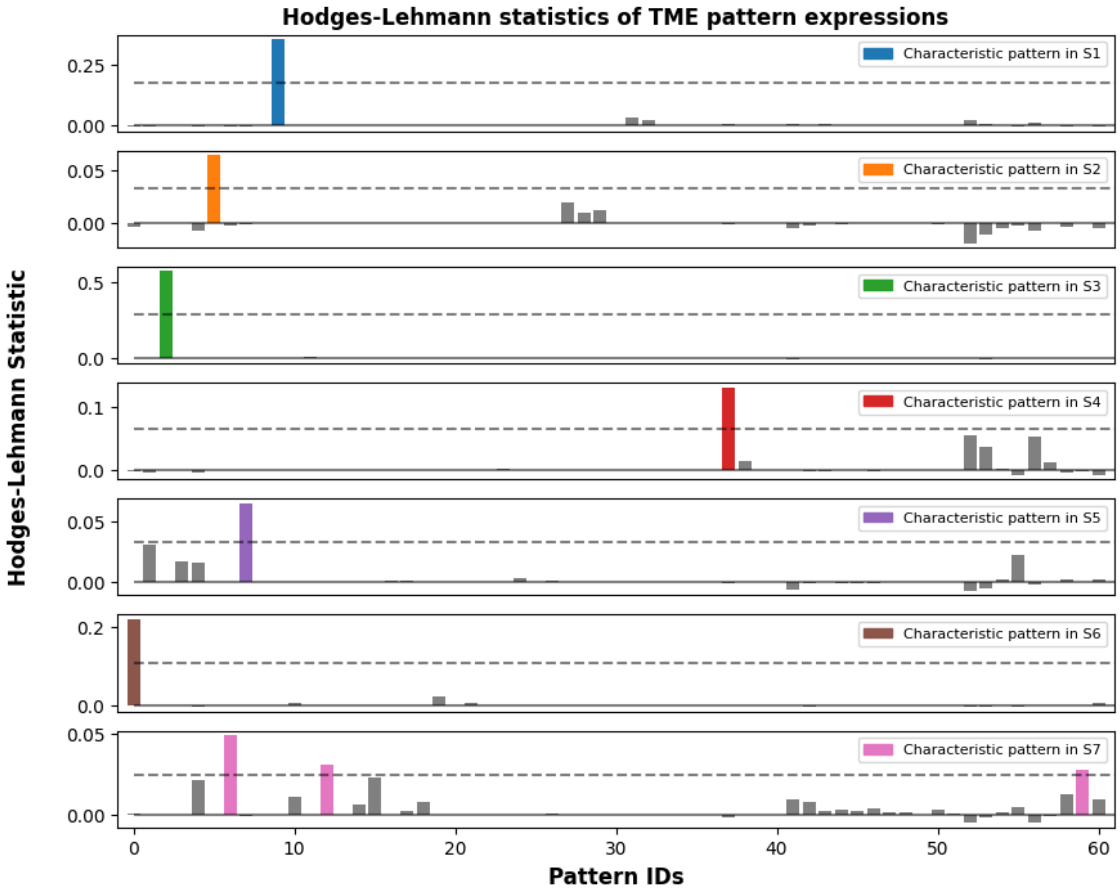
Extended Data Fig. 1: **Cell phenotyping system of the datasets.** **(a)** The median antigen profiling of different cell phenotypes in the discovery set and inner validation set curated by Danenberg et al. [16]. **(b)** The median antigen profiling of different cell phenotypes in the discovery set and inner validation set curated by Jackson et al. [17]. In **(a)** and **(b)**, the y-axis label colors indicate the cell categories: blue: tumor cells; orange: immune cells; green: stromal cells.



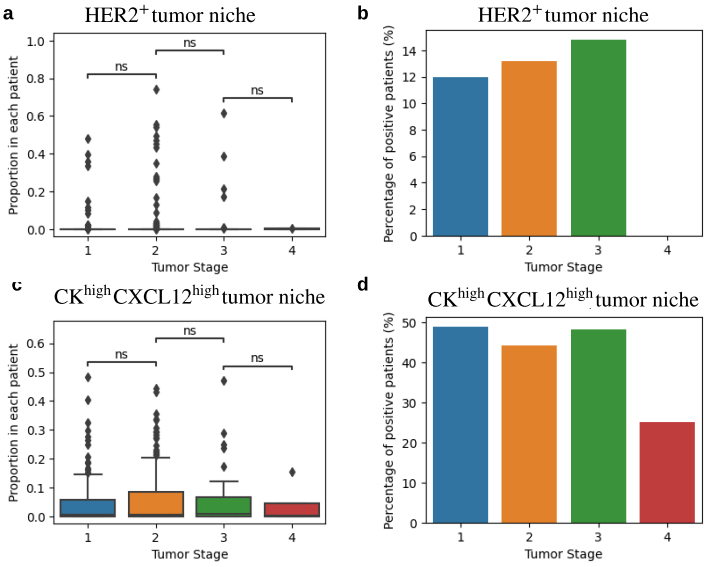
Extended Data Fig. 2: **Detected patient subgroups with different choices of hyperparameter $k^\star$.** The first row displays population graphs with nodes painted by patient subgroup ids, and the second row shows the survival plots of patient subgroups. Multi-variate log-rank test is used to compare their survivals, and the p-values are shown on the titles.



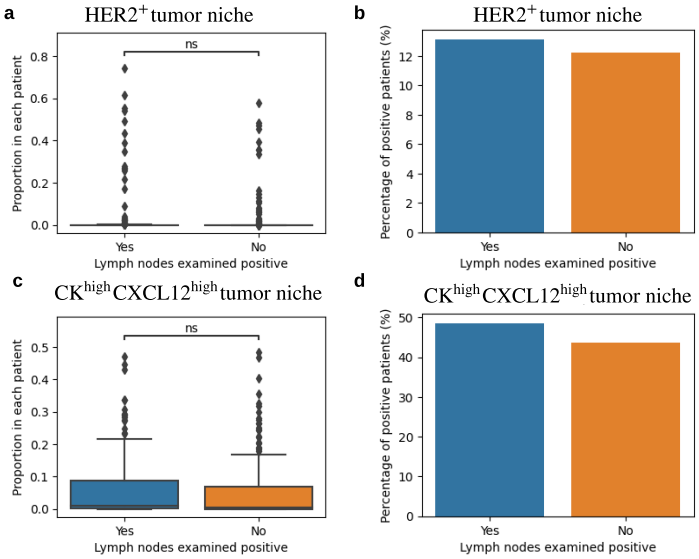
Extended Data Fig. 3: **Distribution of TME pattern expressions in each of the seven patient subgroups** The proportion of each specific TME pattern within an individual patient is calculated by normalizing the corresponding TME pattern histogram. The distribution of these proportions within each patient subgroup is visually represented using box plots. Characteristic patterns are highlighted by colored box plots.



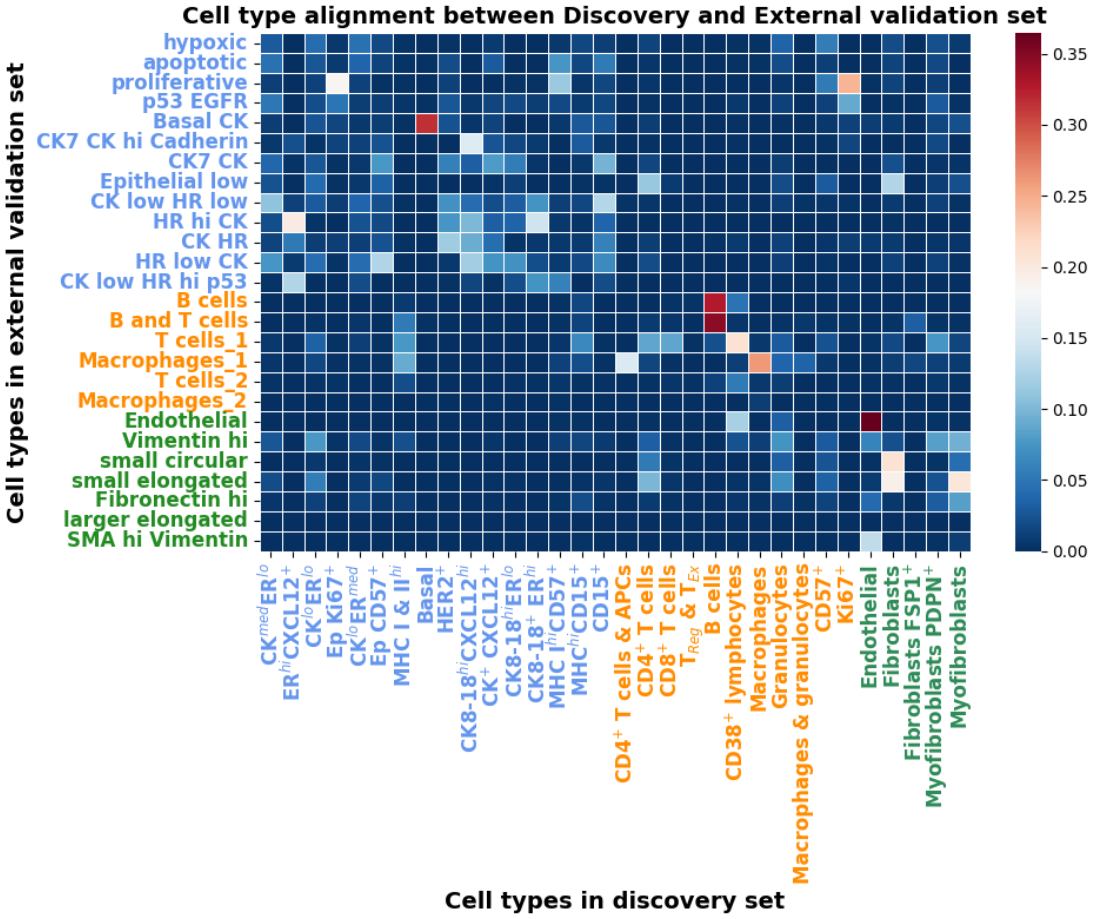
Extended Data Fig. 4: **Hodge-lehmann statistics of TME pattern expressions in each of the seven patient subgroups.** For each specific TME pattern within a designated patient subgroup, the distribution of its proportions among patients inside and outside that patient subgroup is compared using the Hodges–Lehmann statistic. A larger Hodges–Lehmann statistic infers an over-expression of that particular pattern in that patient subgroup. A TME pattern is deemed "characteristic" in a patient subgroup if the Hodges–Lehmann statistic exceeds 50% of the maximized value within that patient subgroup, highlighted by colored bar plots.



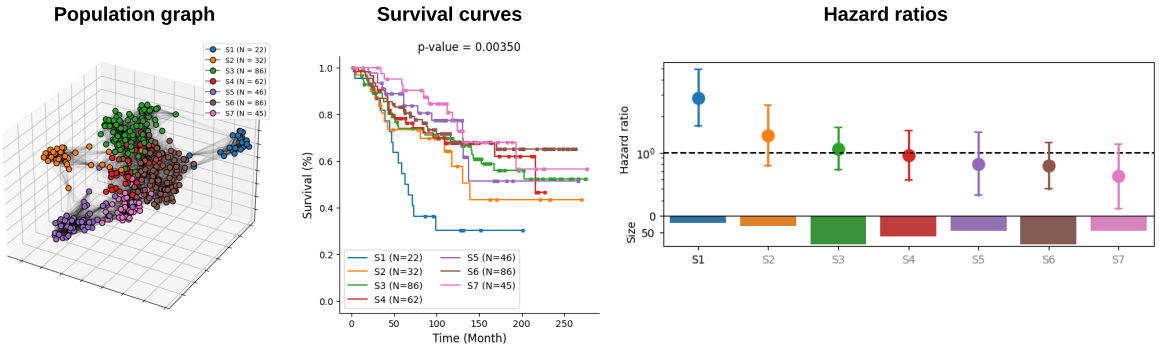
Extended Data Fig. 5: **Correlation between certain TME patterns and tumor stage. (a)** The proportion of HER2$^+$ tumor niche per patient for tumor stage 1, 2, 3, and 4; **(b)** The ratio of patients presenting HER2$^+$ tumor niche (i.e., proportion $> 1\%$) patient in tumor stage 1, 2, 3, and 4 groups; **(c)** The proportion of CK^{high}CXCL12^{high} tumor niche per patient for tumor stage 1, 2, 3, and 4; **(d)** The ratio of patients presenting CK^{high}CXCL12^{high} tumor niche in tumor stage 1, 2, 3, and 4 groups.



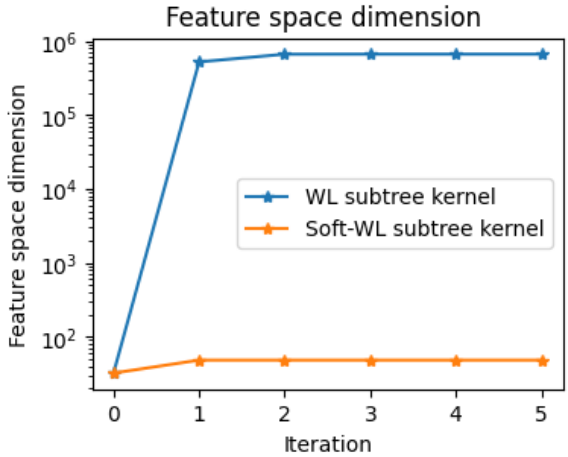
Extended Data Fig. 6: **Correlation between certain TME patterns and metastasis. (a)** The proportion of HER2$^+$ tumor niche per patient for patients with and without metastasis, where metastasis is considered happening if at least one lymph node is examined positive; **(b)** The ratio of patients presenting HER2$^+$ tumor niche (i.e., proportion $> 1\%$) patients with and without metastasis; **(c)** The proportion of CK^{high}CXCL12^{high} tumor niche per patient for patients with and without metastasis; **(d)** The ratio of patients presenting CK^{high}CXCL12^{high} tumor niche in metastatic and primary tumor groups.



Extended Data Fig. 7: **Cell phenotyping system alignment between discovery set and external validation set.** Each cell within the external validation set is associated with two distinct phenotype labels: firstly, the original label as designated in the external validation set phenotyping system as documented by Jackson et al [17]; and secondly, the label assigned within the phenotyping system of the discovery set, determined through cell phenotype mapping. The Intersection over Union (IoU) is computed for each pair of cell phenotypes derived from these two distinct phenotyping systems, and the results are visually represented in the heatmap.



Extended Data Fig. 8: **Risk stratification results from WL subtree kernel (accumulating similarities from all previous iterations).** From left to right, the figure shows a population graph with nodes painted by patient subgroup labels; K-M survival plots of patient subgroups; and hazard ratios estimated by the Cox proportional model.



Extended Data Fig. 9: **Feature space dimension of WL and Soft-WL subtree kernels** To the WL subtree kernel, the feature space dimension refers to the number of unique colors of each iteration of color refinement. To soft-WL subtree kernel, the feature space dimension refers refers to the number of TME patterns at each iteration of graph convolution.