nature biomedical engineering

Article

https://doi.org/10.1038/s41551-022-00951-w

Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens

Received: 16 May 2022

Accepted: 8 September 2022

Published online: 10 November 2022

Check for updates

Zhenqin Wu[1,2,11], Alexandro E. Trevino[1,11] ✉, Eric Wu[1,3], Kyle Swanson[4], Honesty J. Kim[1], H. Blaize D'Angio[1], Ryan Preska[1], Gregory W. Charville[5], Piero D. Dalerba[6], Ann Marie Egloff[7,8], Ravindra Uppaluri[7,8], Umamaheswar Duvvuri[9], Aaron T. Mayer[1] ✉ & James Zou[1,3,4,10] ✉

Multiplexed immunofluorescence imaging allows the multidimensional molecular profiling of cellular environments at subcellular resolution. However, identifying and characterizing disease-relevant microenvironments from these rich datasets is challenging. Here we show that a graph neural network that leverages spatial protein profiles in tissue specimens to model tumour microenvironments as local subgraphs captures distinctive cellular interactions associated with differential clinical outcomes. We applied this spatial cellular-graph strategy to specimens of human head-and-neck and colorectal cancers assayed with 40-plex immunofluorescence imaging to identify spatial motifs associated with cancer recurrence and with patient survival after treatment. The graph deep learning model was substantially more accurate in predicting patient outcomes than deep learning approaches that model spatial data on the basis of the local composition of cell types, and it generated insights into the effect of the spatial compartmentalization of tumour cells and granulocytes on patient prognosis. Local graphs may also aid in the analysis of disease-relevant motifs in histology samples characterized via spatial transcriptomics and other -omics techniques.

Tumour microenvironments (TMEs) are complex niches characterized by cellular, molecular and genetic heterogeneity. Current research and clinical practice have begun to reflect this complexity, with studies producing unbiased atlases of diseased cells[1,2] and novel therapies increasingly targeted at non-cancer cells, including immune and stromal compartments[3]. Just as the functions of healthy tissues depend on the spatial organization of cells, tumour pathology may depend on the spatial organization of the TME[4].

In situ molecular profiling techniques, including spatial transcriptomic[5–7] and proteomic[8–10] techniques, are increasingly being used for the high-dimensional, high-resolution characterization of TMEs and other tissues. Co-detection by indexing[8] (CODEX) is an in situ molecular profiling technique based on the iterative hybridization and fluorescence imaging of DNA-barcoded antibodies that enables the multiplexed quantification of 40 or more antigens from histological specimens at subcellular resolution.

[1]Enable Medicine, Menlo Park, CA, USA. [2]Department of Chemistry, Stanford University, Stanford, CA, USA. [3]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [4]Department of Computer Science, Stanford University, Stanford, CA, USA. [5]Department of Pathology, Stanford University, Stanford, CA, USA. [6]Department of Pathology and Cell Biology, Columbia University, New York, NY, USA. [7]Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA. [8]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [9]Department of Otolaryngology, University of Pittsburgh, Pittsburgh, PA, USA. [10]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [11]These authors contributed equally: Zhenqin Wu, Alexandro E. Trevino. ✉e-mail: alex@enablemedicine.com; aaron@enablemedicine.com; jamesz@stanford.edu

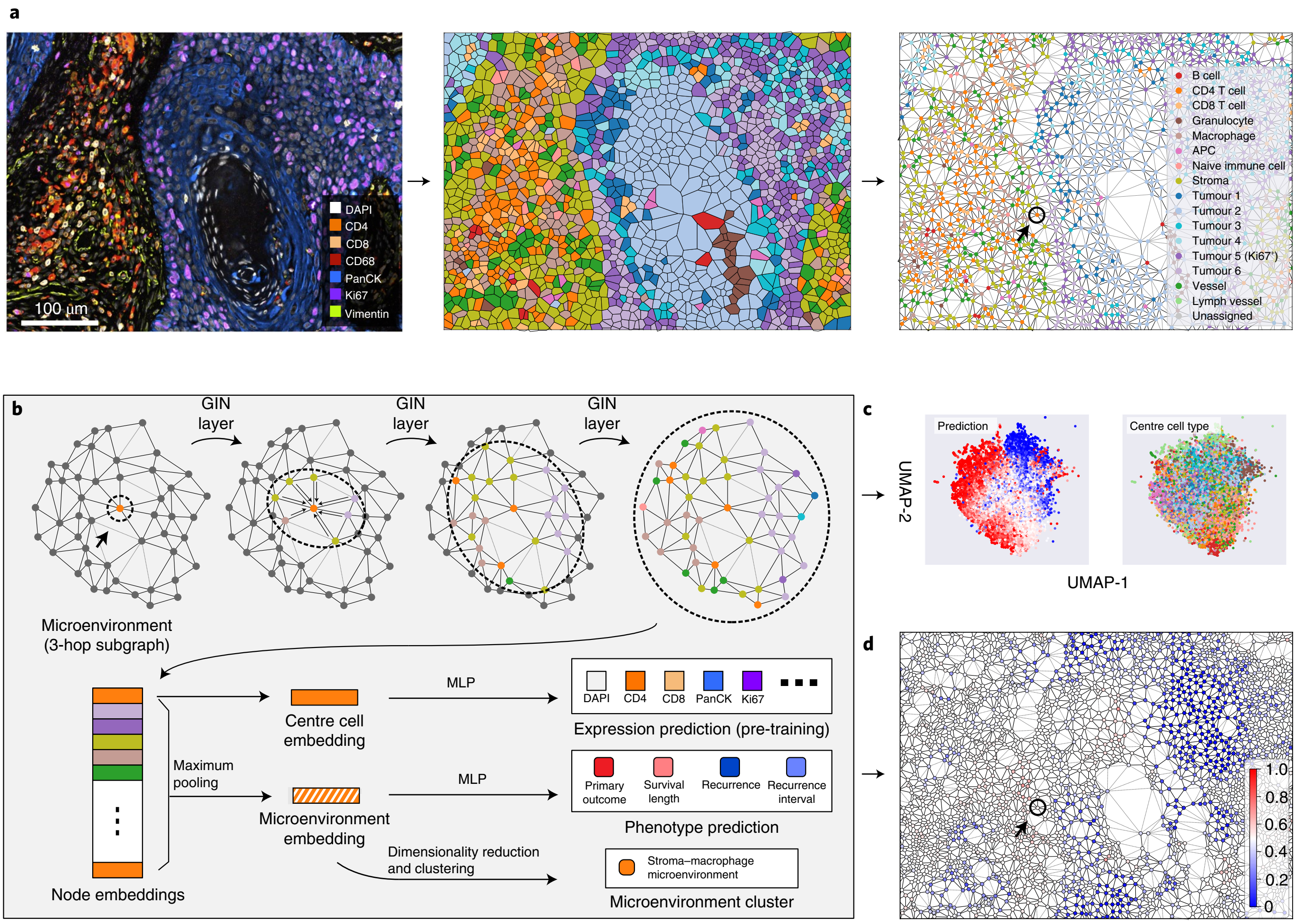


Fig. 1 | Geometric deep learning on multiplexed immunofluorescence imaging. a, Preprocessing of CODEX data. We first transformed the multiplexed fluorescence imaging data (left; DAPI, 4′,6-diamidino-2-phenylindole; PanCK, pan-cytokeratin) into Voronoi diagrams (middle). We then constructed spatial cellular graphs (right; APC, antigen presenting cell) from the Voronoi polygons, where each node is one cell and edges indicate adjacent cells. **b**, Model structure of SPACE-GM. On the query node (marked by the black arrow and circle in panels a, b and d), a three-layer GIN[23] was applied to read the structure of its 3-hop neighbourhood, which we call a microenvironment. Embeddings from the GIN were then used to predict cellular and phenotypic properties. **c**, Dimensionality reduction (uniform manifold approximation and projection[27] (UMAP)) of microenvironment embeddings from all UPMC head-and-neck cancer (HNC) samples generated by SPACE-GM. Each dot represents a microenvironment; colours in the left panel indicate the model's prediction of the primary outcome task, and colours in the right panel indicate the centre cell type. **d**, Predictions of microenvironments aggregated over the whole CODEX sample. Colours of the nodes represent SPACE-GM predicted probabilities for the primary outcome.

Although these spatial technologies capture rich cellular and neighbourhood information, analysis of these spatial data presents new challenges. In particular, how to identify biologically meaningful microenvironments from the rich spatial data and how to characterize the disease relevance of microenvironments are important open questions.

Previous works typically assigned cells to cellular neighbourhoods according to the cell-type compositions of their immediate neighbours[8,11–13]. However, these approaches may miss local spatial relationships between cells. Moreover, as the neighbourhoods are generated in a purely unsupervised fashion, they provide limited insight into which microenvironments are disease relevant. We hypothesize that local spatial arrangements of cells beyond composition could encode rich disease-relevant information.

Here, we present spatial cellular graphical modelling (SPACE-GM), a geometric deep learning framework that employs a graph neural network (GNN) to flexibly model cellular niche structures, or microenvironments, as subgraphs. Each node of the subgraph corresponds to a cell represented by its multiplexed protein levels, and the edges capture neighbour relations. We apply SPACE-GM to three clinically annotated CODEX datasets and show that it identifies disease-relevant microenvironments that accurately predict patient-level phenotypes. We show that SPACE-GM generalizes across studies and disease contexts. Moreover, by analysing the network embeddings, we derive specific insights into how local structural compartmentalization explains patient prognosis and treatment response.

There has been increased interest in applying graph-based deep learning methods to spatial cellular structures in recent literature[14–16]. GNNs[17,18], a class of deep learning methods designed for graph structures, have been applied to a variety of analysis tasks, including cell-type prediction[19], representation learning[20], cellular communication modelling[21] and tissue structure detection[22]. As most of these methods are designed for cellular property modelling, a gap still exists between cellular-level graph analysis and patient-level phenotypes. SPACE-GM bridges this gap by training models using microenvironments as inputs to predict patient phenotypes. Interpretation of SPACE-GM sheds light on how cellular spatial arrangements impact disease and treatment outcomes.

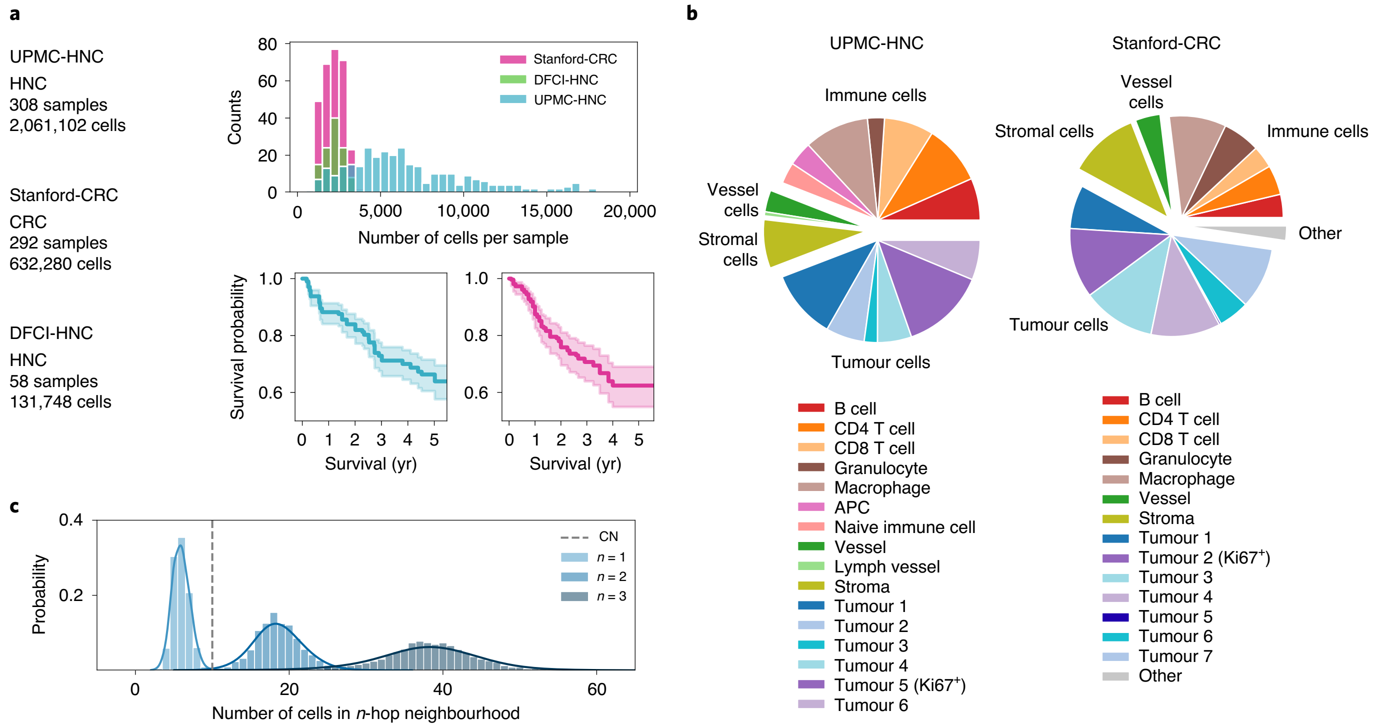


Fig. 2 | CODEX samples and cell characteristics. a, Distributions of the number of cells per sample and overall survival curves for UPMC-HNC (bottom left) and Stanford-CRC datasets (bottom right). Shaded areas in the survival curves represent 95% confidence intervals. **b**, Cell-type compositions of UPMC-HNC and Stanford-CRC datasets. See Supplementary Fig. 1a for the composition of DFCI-HNC dataset. **c**, Histograms of the number of cells in n-hop neighbourhoods ($n = 1, 2, 3$) in the UPMC-HNC dataset. SPACE-GM utilizes a three-layer GIN to read 3-hop neighbourhoods (microenvironments). The dashed vertical line indicates the size (10 cells) of cellular neighbourhood (CN), which is commonly used in previous works[12]. SPACE-GM captures substantially larger neighbourhoods.

Results

To model cellular communities, we first developed a pipeline to segment and classify individual cells from CODEX data (Methods). We then inferred the 2D spatial structure of cells by constructing a Delaunay triangulation and Voronoi diagram of cell centroid coordinates (middle panel of Fig. 1a). Last, we transformed the data into a graph, defining cells as nodes and Delaunay neighbours as edges (right panel of Fig. 1a).

Geometric deep learning models cellular microenvironments

With the graphical representation as input, we propose SPACE-GM as a geometric deep learning tool that reads spatial cellular community structures in TMEs. SPACE-GM employs a graph isomorphism network (GIN)[23] as the backbone and multiple multilayer perceptrons (MLPs) as prediction heads (Methods).

SPACE-GM treats cellular graphs of TMEs as collections of local subgraphs: centre nodes (cells) and their n-hop spatial neighbours (that is, nodes within a graph distance of n edges from the centre node). Empirically, we found that a 3-hop neighbourhood—corresponding to 40 cells on average—was a suitable choice as the size of subgraphs and model performance were balanced (Supplementary Note 1 and Supplementary Fig. 4). We hereafter refer to these 3-hop local subgraphs as 'microenvironments' (Fig. 1b), which may be distinct from the TMEs studied in other settings. Correspondingly, a three-layer GIN is employed in SPACE-GM, which constructs embeddings for centre cells and microenvironments based on node features (for example, one-hot encoded cell types) and edges (graph connectivity). Embeddings can then be passed through the prediction heads to generate estimates for cellular properties (for example, the centre cell expression profile) and patient-level phenotypic properties (for example, the survival outcome of the patient).

In practice, we first pretrained SPACE-GM on cellular property prediction tasks and then finetuned the backbone for patient phenotype predictions, both with microenvironment inputs (bottom row of Fig. 1b). In the text below, 'SPACE-GM' and 'SPACE-GM no-pretraining' represent models trained with or without the pretraining stage, respectively. During inference, we collected predictions from all individual microenvironments from test samples and performed mean aggregation to derive patient-level predictions (Methods and Supplementary Note 2).

Applying SPACE-GM to HNC and colorectal cancer samples

To demonstrate the ability of SPACE-GM to model biologically and clinically relevant signals, we generated three 40-plex CODEX datasets from primary human cancer resections (Methods). Tissues were collected at Stanford University, University of Pittsburgh Medical Center (UPMC) and Dana-Farber Cancer Institute (DFCI). In total, 658 samples were imaged, representing 139 patients with HNC[24] and 110 patients with colorectal cancer (CRC)[25] (Fig. 2a). We refer to these datasets as UPMC-HNC, Stanford-CRC and DFCI-HNC. The samples were annotated with clinical data, including patient survival, disease recurrence and response to therapy (Supplementary Table 1).

The CODEX samples were transformed into graphical representations following the pipeline described above (Fig. 2b and Supplementary Fig. 1a). We extracted microenvironments by enumerating subgraphs of 3-hop neighbours within a distance threshold of 75 µm around each cell. The median microenvironment contained 38 cells, larger than previously proposed cellular neighbourhoods[8,12] (Fig. 2c). SPACE-GM was trained using the pretraining and finetuning approach, in which protein expression of the centre cell was used as the pretraining task and clinical annotations were used as phenotype prediction labels.

SPACE-GM predicts patient phenotypes from cell microenvironments

We applied SPACE-GM to predict survival and recurrence outcomes for patients with UPMC-HNC (Table 1) and Stanford-CRC (Table 2).

Table 1 | Prediction performance on UPMC-HNC and DFCI-HNC tasks

Model	Binary classification (ROC-AUC)			Hazards model (C-index)	Generalization (ROC-AUC)
	Primary outcome	Recurrence	Human papillomavirus (HPV) infection	Survival length	Primary outcome[a]
Linear on composition (sample)	0.783	0.852	0.870	0.696	0.731
MLP on composition (sample)	0.771	0.869	0.879	0.721	0.754
Linear on composition (microenvironment)	0.774	0.823	0.864	0.700	0.799
MLP on composition (microenvironment)	0.814	0.832	0.891	0.751	0.806
SPACE-GM, no-pretraining	0.854	0.882	0.918	0.778	0.853
SPACE-GM	0.867	0.883	0.926	0.799	0.873

Binary classification and hazards model columns report the average performance on the validation folds of UPMC-HNC. The generalization column reports the cross-study (UPMC-HNC to DFCI-HNC) prediction performance. Detailed definitions of primary outcomes are discussed in Supplementary Table 1. [a]Models trained with UPMC-HNC primary outcome tasks were applied to DFCI-HNC samples, and predictions were evaluated for the binary primary tumour response task.

For each prediction task, SPACE-GM was trained with around 70% of the samples and tested on the remaining unseen samples from different coverslips (Methods). SPACE-GM achieved good performance on both datasets, with an area under the receiver operating characteristic curve (ROC-AUC) of above 0.85 on all binary classification tasks and a concordance index (C-index) of around 0.8 on the survival analysis task for UPMC-HNC datasets. Stanford-CRC datasets showed slightly worse performance, probably due to having fewer samples and cells.

For context, we compared SPACE-GM's clinical prediction performance against alternative composition-based methods, applied to either whole samples or subgraphs. As input to these baseline models, we used whole graphs or the same subgraphs of 3-hop neighbourhoods (microenvironments), but featurized as cell-type composition vectors (Methods and Supplementary Fig. 9). Compared with graph representations, composition vectors collapsed spatial structure, losing information about the relative spatial arrangement of cells within a subgraph. Both linear models (logistic regression or proportional hazards regression) and MLPs were trained and evaluated using the same pipeline (Methods).

SPACE-GM consistently outperformed baseline methods on both classification and hazards modelling (time-to-event) tasks (Table 1 and Supplementary Table 2). The uncertainty of the performance metrics was calculated by bootstrapping (Supplementary Note 5 and Supplementary Table 3), which demonstrated the consistent advantages of SPACE-GM. Model pretraining conferred an additional advantage on all prediction tasks. On the most challenging dataset, Stanford-CRC, where composition-based methods generated nearly random predictions, SPACE-GM demonstrated a robust test-set performance (Table 2).

On a representative task—the primary outcome of the UPMC-HNC dataset—we plotted ROC curves for one of the test folds and observed a substantial advantage of SPACE-GM predictions over baseline methods (Fig. 3a). SPACE-GM also achieved better performance than stratifications or predictors with cancer staging and patient demographics features (Supplementary Note 3 and Supplementary Table 5).

It is also worth noting that MLPs based on microenvironment compositions outperformed the same model using whole-graph compositions, which is also reflected in most of the other prediction tasks. This observation indicates that the superiority of geometric deep learning stems, in part, from our microenvironment aggregation strategy. Figure 3b plots the histogram of individual microenvironment predictions before aggregation. In addition to the difference in the distributions of prediction values between positive and negative samples, we also notice that most predictions, regardless of label, are neutral. This suggests that most spatial cellular structures are shared among different patients, but only a small fraction of motifs are highly indicative of patient outcomes. Identifying and characterizing these disease-relevant microenvironments will be highly informative for better diagnostics and therapeutics.

Table 2 | Prediction performance on Stanford-CRC tasks

Model	Binary classification (ROC-AUC)		Hazards model (C-index)	
	Primary outcome	Recurrence	Survival length	Recurrence interval
Linear on composition (sample)	0.563	0.592	0.524	0.537
MLP on composition (sample)	0.551	0.542	0.577	0.570
Linear on composition (microenvironment)	0.576	0.599	0.562	0.569
MLP on composition (microenvironment)	0.547	0.491	0.525	0.519
SPACE-GM no-pretraining	0.684	0.675	0.642	0.669
SPACE-GM	0.739	0.696	0.655	0.713

All columns report the average performance on the validation folds of Stanford-CRC.

Figure 3c plots the survival curves of two patient cohorts in the test data, separated by the median predicted risk from SPACE-GM (trained under the survival length task). We observed a significant difference ($P < 0.001$, log-rank test) in survival probability between the low-risk group and the high-risk group. The survival curves stratified by SPACE-GM risk scores show greater separation than curves stratified using cell-type composition risk scores (Supplementary Fig. 3), suggesting that SPACE-GM captures more granular spatial motifs predictive of patient mortality.

Cross-study generalizability of SPACE-GM

We next sought to evaluate the generalizability of SPACE-GM across datasets. The DFCI-HNC dataset contained CODEX samples collected both before and after neoadjuvant therapy in 29 patients. Although patient survival outcomes in this cohort were not available, samples were annotated on the basis of the degree of pathologic tumour response in surgically resected tumours after neoadjuvant anti-PD1 therapy[26], which served as a proxy for the primary outcome in this experiment.

In the experiment, we trained models on primary outcome labels from the UPMC-HNC dataset with integrated cell types (Supplementary Note 4) and directly applied them to pre-therapy DFCI-HNC samples to predict therapeutic response. Despite potential batch effects resulting from tissue handling, biopsy size (Fig. 2a) and independently generated cell labels (Methods), we found that SPACE-GM generated robust

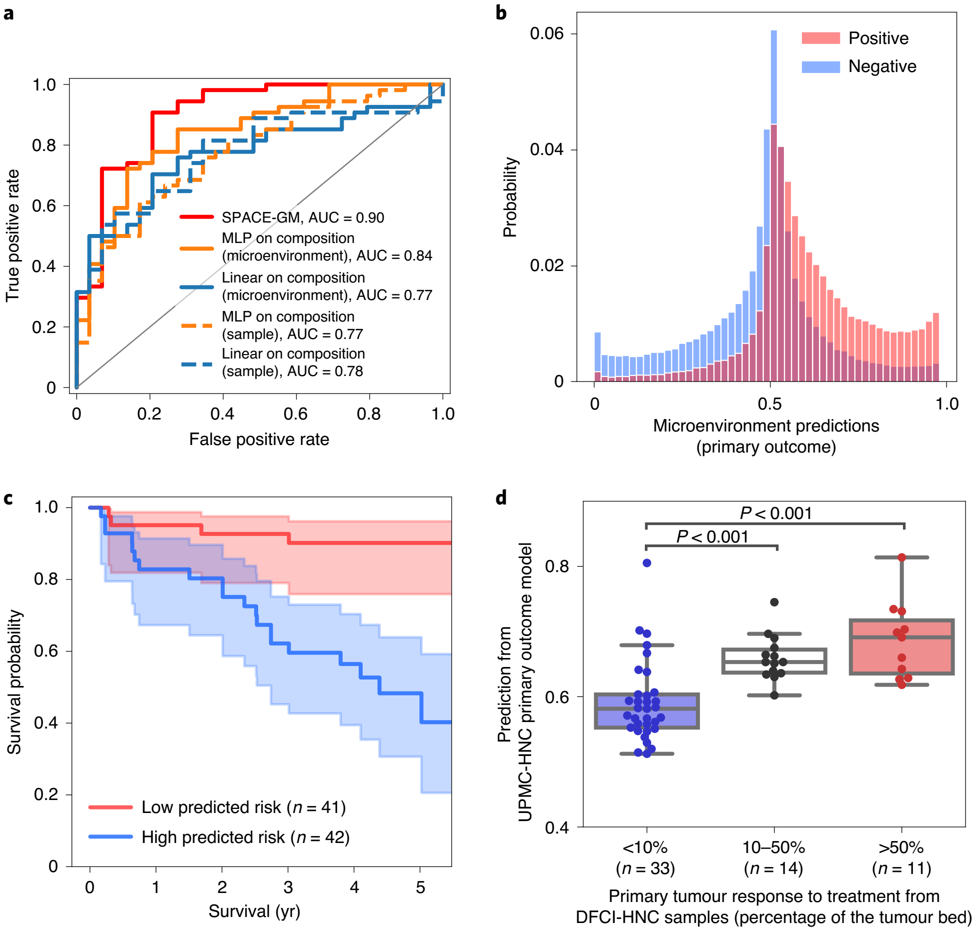


Fig. 3 | Model predictions of primary outcome and survival length in HNC. **a**, ROC curves of different methods on a representative task: the primary outcome prediction for the UPMC-HNC dataset. SPACE-GM outperforms all other methods by a comfortable margin. One of the two validation folds is shown here, and results are representative of the other fold. **b**, Distribution of microenvironment prediction values on the test set from SPACE-GM. Most of the predictions are neutral, with a small portion of positive/negative predictions. **c**, Survival curves of patients in the test set grouped by the SPACE-GM predicted risk. The two groups are separated by the median predicted risk. The survival probability of the low-risk group is significantly higher than that of the high-risk group ($P < 0.001$, log-rank test). **d**, Box plot showing the distributions of SPACE-GM predictions for the DFCI-HNC dataset. Centre lines denote the median, boxes show the quartiles, and whiskers mark the minima and maxima. Models are trained with the primary outcome task of UPMC-HNC. SPACE-GM predicts significantly higher scores ($P < 0.001$, two-sided two-sample t-test) for the high-response group (red box and white box) than for the low-response group (blue box).

estimations, with accuracy surpassing all baseline composition-based models (Table 1).

To examine this result in greater detail, we grouped SPACE-GM predictions by pathologist-annotated pathologic tumour response categories (Fig. 3d) and found that model predictions aligned well with fine-grained pathologic tumour response categories. A two-sample t-test suggested that there are significant differences in predictions between <10% and >10% responder samples ($P < 0.001$, two-sided two-sample t-test). These results demonstrate the robustness and generalizability of SPACE-GM.

Defining disease-relevant cell microenvironments with SPACE-GM

Motivated by the superior performance of SPACE-GM over composition-based baseline methods, we further investigated how characteristics of cellular community structures beyond composition were predictive of clinical outcomes. The prediction accuracy of SPACE-GM suggests that its embedding space, which learns to represent each microenvironment with a numerical vector, is informative of the phenotypes of interest. Visualization with UMAP[27] shows that patient phenotypes were well separated in the embedding space (Fig. 1c and Supplementary Fig. 6).

We used the primary outcome task from the UPMC-HNC dataset as an example to demonstrate how to interpret SPACE-GM embeddings to generate biological hypotheses related to patient outcomes. We first clustered all the microenvironments on the basis of SPACE-GM embeddings (Methods). Clusters showed distinct cell-type enrichment patterns and prediction values (Fig. 4a; see Methods for detailed characterizations of microenvironment clusters). Similar patterns were also observed on unseen test samples (Supplementary Fig. 8).

Four clusters of interest are displayed in Fig. 4b, all of which show significant differences in frequency between positive (no evidence of disease (NED), $n = 54$) and negative (non-NED, $n = 29$) samples, indicating that they possess informative cellular motifs (Fig. 4b, $P = 0.002$ for lymphocyte-rich microenvironments, $P < 0.001$ for the rest, two-sided two-sample t-test).

Voronoi diagrams and raw multiplexed fluorescence images were visualized for example microenvironments in each cluster (middle and bottom rows of Fig. 4b). We named the clusters according to their cell-type compositions and spatial cellular structure patterns. The circle and triangle clusters appeared more in positive outcome samples and were thus associated with a better prognosis. In contrast, the square and star clusters were more enriched in patients with negative outcomes.

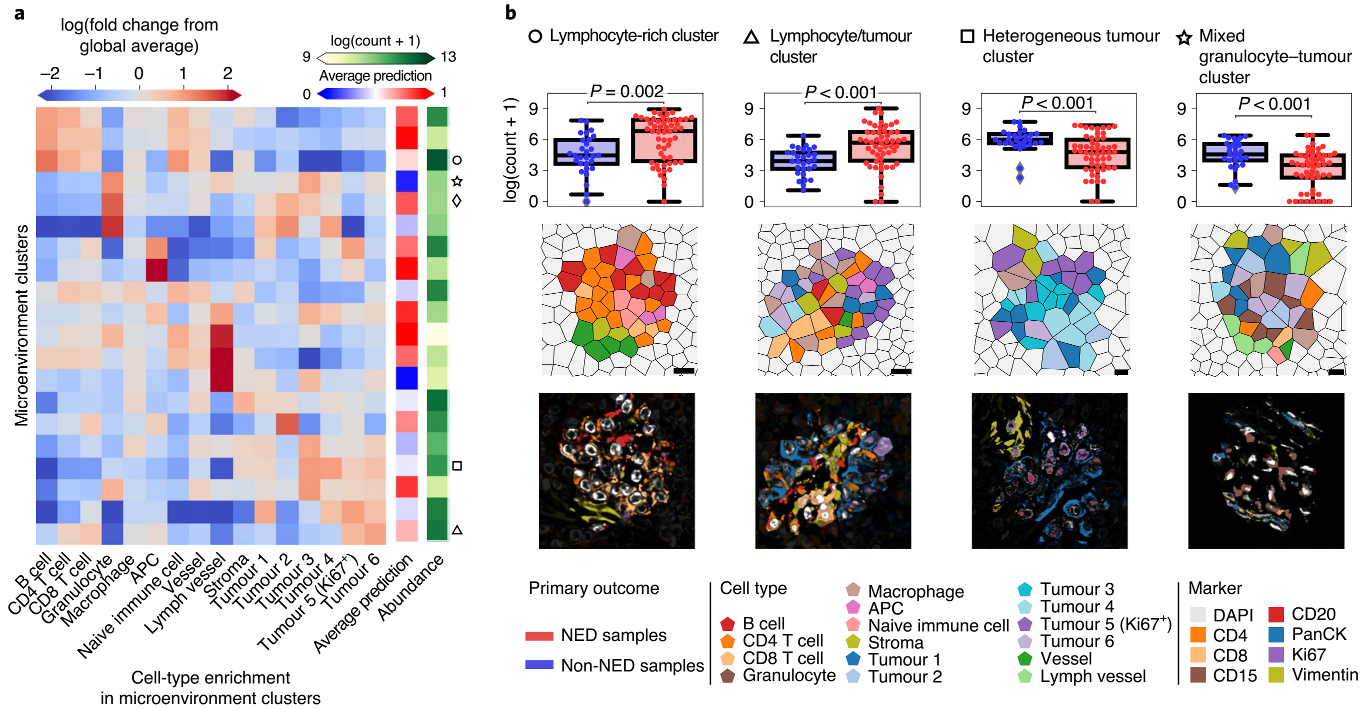


Fig. 4 | Clustering of SPACE-GM embeddings identifies disease-relevant microenvironments. a, A total of 20 clusters are identified by fitting a K-means clustering ($K = 20$) on microenvironment embeddings. The heatmap on the left shows different cell-type enrichment patterns in microenvironment clusters. The column in the middle shows the average prediction values for the primary outcome task from each cluster of microenvironments; note that the colour scale is different from the enrichment matrix. The column on the right shows the abundance of each cluster of microenvironments in the UPMC-HNC dataset. The black shapes on the right indicate the specific cell clusters shown in **b**. Circle, a lymphocyte-rich cluster; star, a tumour cell and myeloid cell (granulocytes and macrophages) cluster characterized by mixed spatial distributions; diamond, a tumour cell and myeloid cell cluster characterized by compartmentalized spatial distributions; square, a heterogeneous tumour cluster characterized by various subtypes of tumour cells; triangle, a mixed lymphocyte/tumour cell cluster. **b**, Top: counts of the appearance of microenvironments in four representative clusters. Centre lines denote the median, boxes show the quartiles, and whiskers mark the minima and maxima. Significant differences are observed between positive (NED) and negative (non-NED) samples ($P = 0.002$ for lymphocyte-rich microenvironments, $P < 0.001$ for the rest, two-sided two-sample t-test). Middle: Voronoi diagrams of sample microenvironments. Bottom: raw CODEX images of sample microenvironments.

In silico permutations suggest disease-relevant spatial motifs in microenvironments

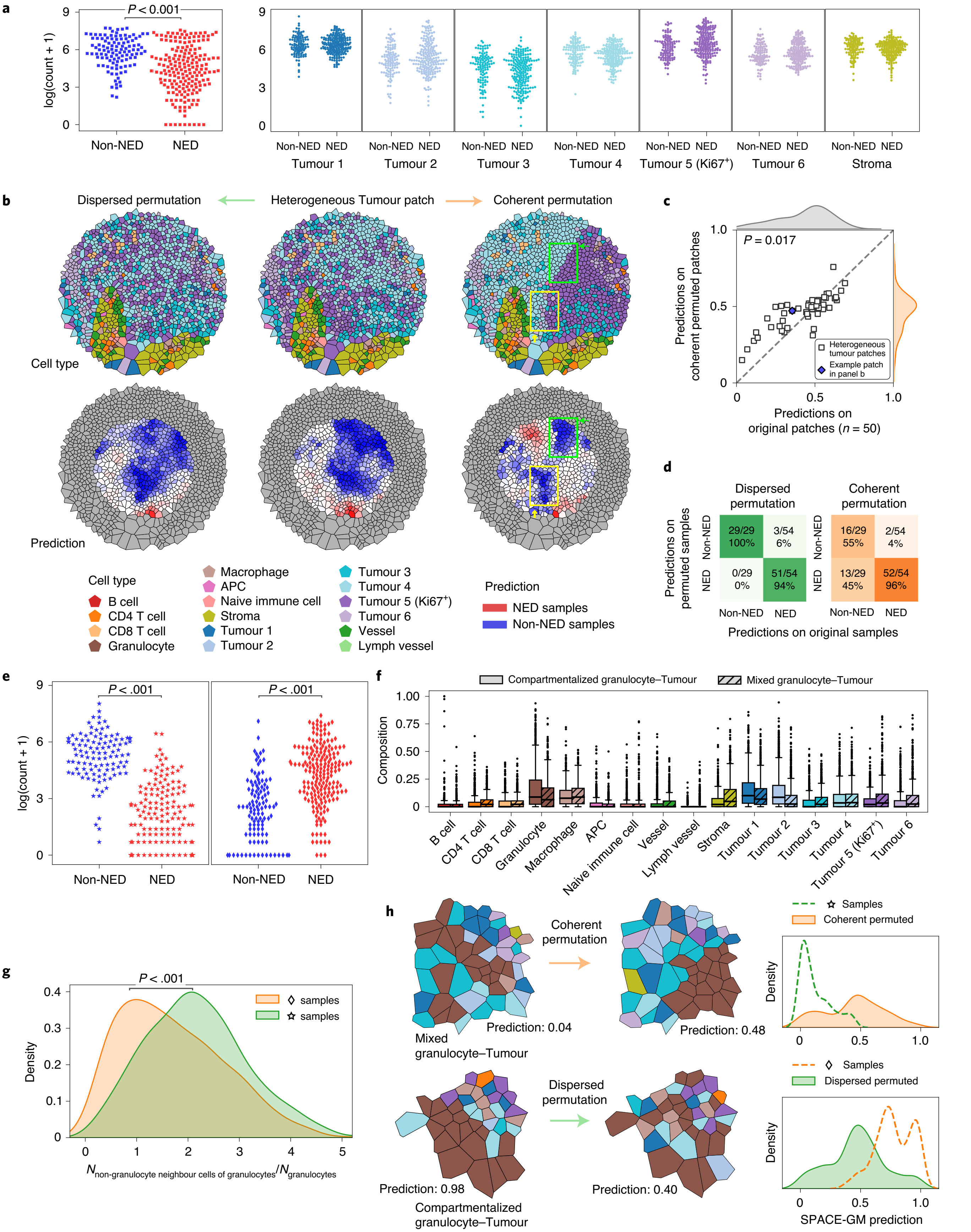
Clusters of disease-relevant microenvironments indicated correlations between microenvironment structures and clinical phenotypes. Within these phenotypes, we were particularly interested in structural characteristics other than composition. For example, the formation of distinctly shaped boundaries between immune and tumour cells could indicate distinct pathways and programs related to tumour progress or immune control.

To directly test how structural arrangements impact the microenvironment function, we computationally permuted nodes in microenvironments and measured SPACE-GM outcome predictions for the permuted graphs. Our permutations were designed to vary the amount of dispersion allowed in the final graph. Thus, a 'dispersed permutation' in our scheme rearranged all cells amongst each other, resulting in highly mixed graphs. A 'coherent permutation' extracted target cell types and placed them into sectors around the centre cell, resulting in cells of the same type appearing together (Methods). In this section, we discuss two representative clusters discovered above and evaluate how permutations towards the two extremes of dispersion levels affect model predictions.

In the heterogeneous tumour cluster, we observed that a group of microenvironments enriched in different tumour subtypes were highly indicative of non-NED outcomes. These tumour subtypes were characterized by different protein expression profiles, as shown in Supplementary Fig. 2b. We compared the appearance of this subtype in all samples against each of the individual tumour subtypes and stromal cells (Fig. 5a). Interestingly, no individual tumour subtype was differentially enriched between positive and negative samples.

To test whether the arrangement of tumour subtypes confers the prediction bias on these samples, we selected regional patches (Methods) and performed the permutations described above. Figure 5b illustrates an example patch (middle column) and the two permuted versions (left and right columns) that mix/separate Tumour 4 and Tumour 5 (Ki67^{+}). No major change was observed after the dispersed permutation as the original patch was highly mixed. In contrast, when subject to coherent permutation, the mean-aggregated prediction increases. Moreover, we observed that the boundaries of different tumour subtypes (yellow and green boxes) overlap with the negative prediction regions, suggesting that the mixing of these tumour subtypes explains the negative prediction in this example.

Results for a diverse set of 50 heterogeneous tumour regional patches confirmed this (Fig. 5c and Methods): SPACE-GM predictions on coherent permuted patches were significantly higher ($P = 0.017$, Wilcoxon signed-rank test). This trend could even be extended to the whole-sample level. On all 83 samples from the test set, we ran coherent permutations and evaluated changes in predictions (Fig. 5d). Of the 29 samples that were predicted to have negative outcomes, 45% had predictions altered after coherent permutation, compared with 0% after dispersed permutation. These results suggest that the spatial mixing of tumour subtypes is a negative predictor of patient outcomes captured by SPACE-GM.



We next evaluated the spatial organization of two SPACE-GM embedding clusters, termed mixed and compartmentalized granulocyte–tumour microenvironments, with similar cell-type compositions (Fig. 5f) but opposite outcome enrichment (Fig. 5e). A visualization of examples from the two clusters (left column of Fig. 5h) suggested that the difference could potentially be attributed to

Fig. 5 | Permutation of nodes in microenvironments helps identify cell–cell interactions that affect predictions. **a**, Swarm plots show counts per sample of heterogeneous TMEs and each individual tumour subtype. Heterogeneous TMEs display a significant difference between NED and non-NED samples ($P < 0.001$, two-sided two-sample t-test). **b**, An example heterogeneous tumour patch (middle column) and its two permuted versions (left and right columns). Areas of negative predictions in the coherent permuted patch overlap with tumour subtype boundaries. **c**, Scatter plot of predictions on 50 heterogeneous tumour patches. Coherent permuted patches have significantly higher predictions than their original patches ($P = 0.017$, Wilcoxon signed-rank test). **d**, Confusion matrices of predicted classes between original samples and permuted samples. **e**, Swarm plots of mixed (left) and compartmentalized (right) granulocyte–tumour microenvironments. **f**, Box plots showing the distributions of cell-type compositions of the two granulocyte–tumour microenvironments. Centre lines denote the median, boxes show the quartiles, and whiskers mark the minima and maxima. **g**, As a heuristic measure, the count of non-granulocyte neighbours of granulocytes divided by the count of granulocytes is calculated for the two microenvironment groups, and distributions show significant differences ($P < 0.001$, two-sided two-sample t-test). **h**, Example microenvironments from the two groups undergo permutations. The distributions of SPACE-GM predictions are shifted due to these permutations.

the compartmentalization of granulocytes, and we verified this by calculating the number of non-granulocyte neighbours for 10,000 granulocytes from these two clusters (Fig. 5g). The distribution showed a significant right shift ($P < 0.001$, two-sided two-sample t-test) for mixed granulocyte/tumour cluster samples, indicating that granulocytes were more dispersed in this cluster.

To further characterize the relationships between the compartmentalization of granulocytes and predictions, we performed the two types of permutation (Methods) on corresponding microenvironments to reverse the spatial organization patterns of granulocytes (middle column of Fig. 5h and Supplementary Fig. 7). Notably, coherent permutation of granulocytes in mixed clusters resulted in improved predicted prognoses, whereas dispersed permutation of compartmentalized granulocyte clusters resulted in microenvironments with poorer prognoses (right column of Fig. 5h).

Discussion

In this work we have presented SPACE-GM, a GNN model that predicts clinical properties of patient tumour samples with multiplexed immunofluorescence inputs. We evaluated the prediction performance on three independently collected patient cohorts and demonstrated the predictive superiority of SPACE-GM over composition-based methods, suggesting that spatial cellular arrangements beyond composition could encode phenotype-related information. We further conducted dimensionality reduction and clustering on the microenvironment embeddings from SPACE-GM to identify discrete groups of disease-relevant microenvironments. Permutation on two example clusters reveals the relationships between the spatial compartmentalization of certain cell types and patient outcomes.

More specifically, SPACE-GM shows that the dispersion of molecularly distinct tumour subtypes can have a negative impact on patient survival outcomes. Of note, similar correlations between cell-type mixing (or tumour-subtype mixing) and poorer outcomes have been observed in the recent literature[28,29]. In another experiment, we verified that the compartmentalization level of granulocytes within the TME is relevant to SPACE-GM predictions of primary outcomes. Interestingly, similar trends have been reported in both HNC[30,31] and other solid cancers[32], in which multiple molecular pathways explaining how neutrophils promote tumour growth[33,34] and metastasis[35,36] are presented.

The results discussed in this work demonstrate the value of modelling patient phenotypes with graph-based microenvironment inputs, from which connections between cellular community structures and patient-level phenotypes can be established and validated. However, we acknowledge that certain caveats still exist in our framework. Inference and analysis of unseen data are still complicated due to the limitations of cell segmentation and classification approaches. The lack of a unified image-cell-type dictionary hinders the generalization of trained models on new datasets or unseen cell types. Comprehensive results from human cell consortia efforts[37,38], as well as computational methods that accommodate unseen cell types[19], could potentially be incorporated to overcome these limitations.

To maximize prediction performance, the dense per-cell predictions by SPACE-GM were mean-aggregated on the basis of the benchmarking results from the basic aggregation methods. As illustrated in Figs. 1d and 3b, highly disease-relevant microenvironments are usually less prevalent, so their spatial distribution in a sample could contain information about tissue architecture at different spatial scales. We anticipate that more sophisticated techniques (for example, hierarchical aggregation on different spatial scales) could be employed in this step to reveal insights into the interplay between tissue-level architectures and patient-level phenotypic properties.

SPACE-GM is a versatile framework for capturing disease-relevant motifs from microenvironments. We applied SPACE-GM to analyse CODEX data in this work. This approach of using local graphs to characterize disease-relevant spatial motifs can be extended to other measurement modalities such as spatial transcriptomics, which would be an interesting direction of future work. Disease-relevant microenvironment embeddings and dense predictions of target phenotypes could be further coupled with downstream analysis (for example, permutation) to reveal relationships between cellular community structure and patient-level phenotypes.

Methods

CODEX data collection

Patient samples and data were obtained using institutional protocols. Tumour tissue samples were prepared, stained and acquired following CODEX User Manual Rev C (https://www.akoyabio.com).

Tissue collection. Tumour microarray cores were collected according to the following guidelines:

- Areas for coring were representative areas of tumours containing neoplastic epithelium as determined by a board-certified anatomic pathologist.
- Core size was 0.6 mm diameter (DFCI-HNC and Stanford-CRC) or 1.0 mm diameter (UPMC-HNC). A minimum of 2 cores was taken from each resection specimen, although some cores may have been filtered by quality control before our analysis.
- Cores were selected to sample tumour centres, not invasive edges.

Coverslip preparation. Coverslips were coated with 0.1% poly-L-lysine solution to enhance the adherence of tissue sections before mounting. The prepared coverslips were washed and stored according to the guidelines in the CODEX User Manual.

Tissue sectioning. Formaldehyde-fixed paraffin-embedded samples were sectioned at a thickness of 3–5 µm on the poly-L-lysine-coated glass coverslips.

Antibody conjugation. Custom conjugated antibodies were prepared using the CODEX Conjugation Kit, which include the following steps:

1. The antibody was partially reduced to expose thiol ends of the antibody heavy chains.
2. The reduced antibody was conjugated with a CODEX barcode.
3. The conjugated antibody was purified.

4. Antibody storage solution was added for antibody stabilization for long-term storage.

Post-conjugated antibodies were validated by sodium dodecyl sulfate–polyacrylamide gel electrophoresis and quality control tissue testing, where immunofluorescence images were stained and acquired following standard CODEX protocols and then evaluated by immunologists.

Staining. CODEX multiplexed immunofluorescence imaging was performed on formaldehyde-fixed paraffin-embedded patient biopsies using the Akoya Biosciences PhenoCycler platform (also known as CODEX), and 5-μm-thick sections were mounted onto poly-L-lysine-treated glass coverslips as tumour microarrays. Samples were pretreated by heating on a 55 °C hot plate for 25 min and cooled for 5 min. Each coverslip was hydrated using an ethanol series: two washes in HistoChoice Clearing Agent, two in 100% ethanol, one wash each in 90%, 70%, 50% and 30% ethanol solutions, and two washes in deionized water (ddH$_2$O). Next, antigen retrieval was performed by immersing coverslips in Tris–EDTA (pH 9.0) and incubating in a pressure cooker for 20 min on the 'high' setting, followed by 7 min to cool. Coverslips were washed twice for 2 min each in ddH$_2$O and then washed in hydration buffer (Akoya Biosciences) twice for 2 min each. Next, coverslips were equilibrated in staining buffer (Akoya Biosciences) for 30 min. The conjugated antibody cocktail solution in staining buffer was added to coverslips in a humidity chamber and incubated for 3 h at room temperature or 16 h at 4 °C. After incubation, the sample coverslips were washed and fixed following the CODEX User Manual.

Data acquisition. Sample coverslips were mounted on a microscope stage. Images were acquired using a Keyence microscope configured to the PhenoCycler Instrument at a ×20 objective. Sample collections were approved by institutional review boards at the UPMC, Stanford Medical Center and DFCI.

Cell segmentation and classification

After image preprocessing, we applied a neural network-based cell segmentation tool, DeepCell[39], on DAPI image channels to identify nuclei, and these nuclear masks were dilated to obtain whole-cell segmented cells. Nuclear segmentation masks were stochastically dilated by flipping pixels with a probability equal to the fraction of positive neighbouring pixels. This dilation was repeated for nine cycles for all CODEX data.

On each CODEX sample, given the segmentation of individual cells, single cell expression was computed for biomarker *j* with the following steps[40]:

- Compute the mean expression value across pixels within the cell segmentation mask. Denote the mean expression value of cell *i* as $x_i^{(j)}$, and denote the array of all expression values $\{x_1^{(j)}, x_2^{(j)}, ...\}$ as $X^{(j)}$.
- Normalize the expression value using quantile normalization and inverse hyperbolic sine transformation:

$$f(x_i^{(j)}) = \text{arcsinh}\left(\frac{x_i^{(j)}}{5Q(0.2; X^{(j)})}\right)$$

where $Q(0.2; X^{(j)})$ represents the 20th quantile of $X^{(j)}$ and arcsinh is the inverse hyperbolic sine function. Denote the array of all normalized expression $\{f(x_1^{(j)}), f(x_2^{(j)}), ...\}$ as $f(X^{(j)})$.

- Calculate the *z*-score of the normalized expression value:

$$z(x_i^{(j)}) = \frac{f(x_i^{(j)}) - \text{MEAN}(f(X^{(j)}))}{\text{SD}(f(X^{(j)}))}$$

To classify cells, we first obtained a cell-by-marker expression matrix filled with preprocessed expression values ($z(x_i^{(j)})$), and then a principal component analysis (PCA) model was applied to extract the top 20 principal components (PCs). We constructed a *k*-nearest neighbour graph (*k* = 30) on the top 20 PCs of the expression matrix and then performed Louvain graph clustering[41] on the result. Clusters were manually annotated according to their cell biomarker expression patterns. This procedure was performed on a subset of 10,000 cells and subsequently used to train a *k*-nearest neighbour algorithm to predict cell types from the normalized expression vector. This algorithm was used to transfer labels to the entire dataset. The average expression of each cell type for all three datasets used in this work is visualized in Supplementary Fig. 2a.

Construction of spatial cellular graphs and microenvironments

For each multiplexed fluorescence image, we identified individual cells using the segmentation and classification pipeline stated above. The set of cells was represented by a set of discrete points located at cellular centroids. The 2D coordinates of these cellular centroids were determined by the segmentation masks of the corresponding cell nuclei.

To capture the spatial neighbourhood relations, we ran a Delaunay triangulation operation on all the cellular centroids. Corresponding Voronoi diagrams could be uniquely determined by connecting the centres of the circumcircles. We employed the function voronoi_regions_from_coords from the package geovoronoi (https://pypi.org/project/geovoronoi/) in this step.

The graphical representation of multiplexed fluorescence images could then be determined by defining cellular centroids as nodes and neighbouring Voronoi polygons (or edges in the Delaunay triangulation) as edges. We further defined two types of edges based on the distance between cellular centroids: edges shorter than 20 μm were treated as neighbouring edges, and edges longer than 20 μm were treated as distant edges. Edge types will be considered in the following neural network forward pass.

The microenvironment represents the local environment around a cell in the entire cellular community. Given the graphical representation derived above, we defined the microenvironment of a query cell as its *n*-hop neighbourhood. Here, the *n*-hop neighbourhood included all cells within a graph distance of *n* edges from the query cell. In this work, we applied *n* = 3 in all datasets and added another constraint of physical distance: the microenvironment of a query cell included all cells that were in its 3-hop neighbourhood and less than 75 μm away from the query cell.

Data split and evaluation metrics

We evaluated SPACE-GM and other baseline methods on the UPMC-HNC and Stanford-CRC datasets. Samples were first split into the training set and test set following the procedure below:

UPMC-HNC, coverslip split. Details of the UPMC-HNC data (Table 1 and Supplementary Tables 3–7) are listed below:

- UPMC-HNC contained 308 distinct samples collected from seven batches/coverslips, and the class balance of clinical annotations was calculated for each coverslip.
- We proposed two validation folds:
  (a) Fold 1.

  - Of the 225 samples in the training set, 64% had positive primary outcomes.
  - Of the 83 samples in the test set, 65% had positive primary outcomes.

  (b) Fold 2.

  - Of the 217 samples in the training set, 65% had positive primary outcomes.

- Of the 91 samples in the test set, 62% had positive primary outcomes.

(c) In each fold, samples from five coverslips were used for training, and samples from the remaining two coverslips were used for testing. The class balance of clinical annotations was kept similar between training and test sets.
(d) Two validation folds had no overlapping test samples.

- Training and evaluation were run independently on the two validation folds, and prediction performance for each task were averaged.

UPMC-HNC, patient cross validation. Details of the data in Supplementary Table 2:

- UPMC-HNC contained 308 distinct samples collected from 81 patients.
- Patients were randomly split into four groups; samples were assigned to their corresponding patient groups.
- Training and evaluation were run following a cross-validation scheme:

(a) In each of the four independent runs, models were trained on samples from three patient groups and evaluated on samples from the remaining group.
(b) Performance was averaged across the four runs.

Stanford-CRC, coverslip split. Details of the data in Table 2:

- Stanford-CRC contained 292 distinct samples from four batches/coverslips, 161 patients, and the class balance of clinical annotations was calculated for each coverslip.
- We proposed two validation folds based on the coverslip:
  (a) Fold 1.

- Of the 229 samples in the training set, 75% had positive primary outcomes.
- Of the 63 samples in the test set, 71% had positive primary outcomes.

(b) Fold 2.

- Of the 220 samples in the training set, 78% had positive primary outcomes.
- Of the 72 samples in the test set, 64% had positive primary outcomes.

(c) In each fold, samples from three coverslips were used for training, and samples from the remaining coverslip were used for testing.
(d) Two validation folds had no overlapping test samples or patients. Note that the two coverslips solely used for training were excluded from testing because of their different size/class balance.

- Training and evaluation were run independently on the two validation folds, and prediction performance for each task was averaged.

In this work, we used clinical annotations as prediction tasks. Annotations were categorized into two forms (Supplementary Table 1): binary classification (for example, primary outcome) and hazards modelling (for example, survival length). On binary classification tasks, we evaluated model performance by calculating ROC-AUC; on hazards modelling tasks, we evaluated performance by calculating the C-index between predicted hazards and observed events (recurrence or death).

SPACE-GM and baseline methods

SPACE-GM. SPACE-GM consists of a GIN[23] backbone and multiple MLP prediction heads.

Inputs of SPACE-GM contain the local spatial graphical structures of microenvironments derived above, as well as the identity and size of each cell in the microenvironments. Specifics of the GIN inputs are detailed in the following sections.

- Node features

- The cell type is formed as a one-hot vector of length $N_{\text{cell type}}$, which is mapped to $N_{\text{cell type}}$ trainable embeddings of length 512 through a lookup table.
- Other features: cell sizes (in pixel) log-transformed and scaled to 0–1 range, a flag indicating if the cell is the centre taking value 0 (if no) or 1 (if yes), are concatenated and transformed to a vector of length 512 through a trainable linear layer. Note that we experimented with explicitly adding normalized expression to node features, but the resulting models have more severe overfitting and lower test-set performance (see discussion below). Expression is hence excluded from node features.
- The two embeddings above are summed and used as the initial node embeddings for GIN. We denote the initial embedding of node v as $h_v^{(0)}$.

- Edge features

- Self-loop edges (connecting the same nodes) are added to input graphs before forward pass.
- Edges are divided into three classes: neighbouring edge, distant edge and self-loop edge. In each GIN layer, an edge is mapped to one of the three edge embeddings of length 512 through a lookup table. We denote the embedding of the edge between node v and node u in kth layer as $e_{vu}^{(k)} = W$, which is dependent on the edge type between v and u. Note that $e_{vu}^{(k)} = e_{uv}^{(k)}$.

-

SPACE-GM employed a three-layer GIN, and in the kth graph convolutional layer.

- Messages are calculated on each edge as follows:

$$m_{vu}^{(k)} = h_u^{(k-1)} + e_{vu}^{(k)}$$

Note that edges in microenvironments are undirected, and messages in both directions are calculated, although they will not necessarily be equal.

- The embedding of node v is updated on the basis of all incoming messages to v:

$$h_v^{(k)} = \text{MLP}^{(k)}\left(\sum_{u\in N(v)} m_{vu}^{(k)}\right)$$

where $N(v)$ is the set of neighbouring nodes of v, the self-loop edge guarantees $v \in N(v)$, and $\text{MLP}^{(k)}$ is the 2-layer MLP of the kth layer.

Embeddings from the last graph convolutional layer are treated as node embeddings, in which the embedding of the centre cell $h_{\text{centre}}^{(3)}$ is used as input for expression prediction (pretraining of SPACE-GM). We aggregated node embeddings to generate the microenvironment embedding:

$$h_G = \text{MAXPOOL}_{v\in G}(h_v^{(3)})$$

where G represents the microenvironment, MAXPOOL is a channel-wise maximum operation (torch.nn.global_max_pool). Microenvironment embeddings are used for sample phenotype predictions.

For expression and phenotype prediction tasks, we employed two separate three-layer MLPs with leaky Rectified Linear Unit (ReLU) activation function, each with N_{tasks} outputs, taking centre cell embedding $h^{(3)}_{\text{centre}}$ and microenvironment embedding h_G as input, respectively. For expression prediction tasks, we minimized the squared L2 norm loss between predictions and labels (torch.nn.MSELoss). For binary classification tasks, we minimized binary cross entropy between sigmoid logit: outputs of the three-layer MLP and class labels (torch.nn.BCEWithLogitsLoss). For hazards modelling tasks, we adapted the Cox partial likelihood for Stochastic Gradient Descent (SGD)[42].

Baseline methods

Baseline methods in this work are constructed on the basis of composition vector inputs. Composition vectors are calculated on whole-sample graphs or microenvironments. In both cases, we count the number of each cell type appearing in the graph/subgraph. The count vector of length $N_{\text{cell type}}$ is then normalized to the frequency vector, denoted as the composition vector.

For example, in a microenvironment k with n_k cells, its composition of cell type j (jth entry of the composition vector) is calculated as follows:

$$c_j^{(k)} = \frac{\sum_{1\leq n\leq n_k} 1\{\text{cell type}(n) = j\}}{n_k}$$

We trained a linear model and a three-layer MLP model on the composition vectors. Logistic regression and Cox regression are used for binary classification and hazards modelling tasks, respectively, as linear models. MLP used the same loss function as SPACE-GM introduced above. See Supplementary Fig. 9 for visualizations of the baseline methods.

Model training and inference

During training (of microenvironment-based methods), we randomly selected microenvironments from all samples in the training set. Their labels came from centre cells (expression prediction) or clinical annotations of the CODEX samples to which they belong (phenotype prediction). Microenvironments were weighted on the basis of their labels to balance the loss of different classes. Adam optimizer[43] was employed to minimize corresponding losses in different tasks.

SPACE-GM was first trained on the expression prediction task. After convergence, we retained the GIN backbone and connected it with the phenotype prediction head (initialized from scratch), and both modules were finetuned on the phenotype task. SPACE-GM no-pretraining had the same model structure as SPACE-GM, but all were initialized from scratch. It was directly trained on the phenotype task until convergence; microenvironment-based MLP models followed the same training pipeline.

During inference, we ran microenvironment-based models (SPACE-GM, MLP and so on) on all microenvironments from the test set, generating dense per-node predictions for test samples. Predictions within the same CODEX sample were mean-aggregated over the entire graph (microenvironment aggregation, Supplementary Note 2 and Supplementary Fig. 5). We evaluated influences from cells located at the edge of the CODEX sample and found the difference to be negligible (Supplementary Note 6 and Supplementary Fig. 10). Aggregated sample-level predictions were evaluated using corresponding metrics.

All models were implemented in Python with scikit-learn[44], pytorch[45] and pytorch geometric[46]. More details about the hyperparameters and implementations can be found in our code repository at https://gitlab.com/enable-medicine-public/space-gm.

Model architecture and input feature choice

SPACE-GM employed a three-layer GIN model with maximum pooling to predict phenotypic properties for microenvironments that are featurized by cell types. Model architecture and input feature choices were chosen on the basis of empirical evidence from benchmarking different variants.

Supplementary Table 6 shows performance on the primary outcome task of UPMC-HNC from different types of GNNs, including GIN, graph convolutional network[47], graph attention network[48] and GraphSAGE[49]. Most variants generate similar or worse performance than GIN.

Supplementary Table 7 shows performance on the primary outcome task of UPMC-HNC from models using different graph pooling methods, including maximum pooling, sum pooling, mean pooling, global attention layer[50] and Set2set module[51] with two iterations. Maximum pooling is overall the best-performing variant.

We also tried training models directly with expression input instead of cell types. For baseline methods, composition vector inputs were replaced by average expression over cells in the microenvironment; for SPACE-GM, we used expression instead of the one-hot encoded cell type as node features. Supplementary Table 4 reports performance for baseline MLP models and SPACE-GM (no-pretraining), both of which show worse performance than cell-type-based models. We speculate that as the train/test split is based on coverslips, the potential batch effect in expression across coverslips may lead to overfitting and worse prediction performance.

Clustering of microenvironment embeddings

To identify clusters of disease-relevant microenvironments, we applied dimensionality reduction and clustering on microenvironment embeddings. Clusters discussed in the main text were generated following the procedure below:

- A total of 100,000 cells and their microenvironments from the training set were randomly sampled and extracted, denoted as the reference dataset.
- SPACE-GM (trained on the UPMC-HNC primary outcome task) was applied to all microenvironments in the reference dataset, and their microenvironment embeddings (h_G) and predictions were collected, denoted as reference embeddings and reference predictions, respectively.
- A PCA model was initialized and fitted to the reference embeddings. We extracted the top 20 PCs, which captured >70% of total variance.
- A UMAP[27] dimensionality reduction model was fitted on the top 20 PCs of reference embeddings, generating 2D visualizations of microenvironment space.
- A K-means clustering was fitted on the top 20 PCs of reference embeddings. We experimented with different settings of K, in which $K = 20$ yields the best result that balances resolution and granularity.

We then applied PCA and K-means models to the test set:

- All cells and microenvironments from test-set samples were extracted.
- SPACE-GM was applied to test-set microenvironments to extract embeddings.
- PCA and K-means models trained with the reference dataset were directly applied to test-set microenvironment embeddings, and their cluster assignments were collected and summarized.

The same dimensionality reduction and clustering pipeline were applied to the composition vectors of microenvironments, which closely resembles the generation of cellular neighbourhood. Further discussion and comparison of the two approaches can be found in Supplementary Note 7 and Extended Data Fig. 1.

For PCA and K-means, we used the Python implementations from scikit-learn[44]: sklearn.decomposition.PCA and sklearn.cluster.KMeans. For UMAP, we used the Python implementation from umap-learn (https://pypi.org/project/umap-learn/).

Characterization of the microenvironment cluster

Microenvironment clusters were characterized by the following factors.

- Composition of cell types
- For a specific microenvironment k with n_k cells, its composition of cell type j was calculated as

$$c_j^{(k)} = \frac{\sum_{1\le n\le n_k} 1\{\text{cell type}(n) = j\}}{n_k}$$

Then, for a cluster of microenvironments $K = \{k_1, k_2, k_3, ..., k_m\}$, we calculated its average composition vector for cell type j as

$$c_j^{(K)} = \frac{1}{m}\sum_{1\le i\le m} c_j^{(k_i)}$$

We further calculated the global average composition for cell type j as

$$\overline{c_j} = \frac{1}{N}\sum_{1\le k\le N} c_j^{(k)}$$

where N is the total number of microenvironments in the entire dataset. By contrasting these values, we can derive the enrichment of cell type j in a specific microenvironment cluster K as

$$r_j^{(K)} = \log\left(\frac{c_j^{(K)}}{\overline{c_j}}\right)$$

These log fold change values are plotted in the left matrix of Fig. 4a.

- Average SPACE-GM predictions
- Note the prediction for a specific microenvironment k as $p^{(k)}$, and then for a cluster of microenvironments $K = \{k_1, k_2, k_3, ..., k_m\}$, calcualte the average prediction as

$$p^{(K)} = \frac{1}{m}\sum_{1\le i\le m} p^{(k_i)}$$

These average prediction values are plotted in the middle column of Fig. 4a.

- Abundance

Among all the N microenvironments, assume there are $N^{(K)}$ microenvironments identified to be in cluster K, and its abundance is calculated as

$$\text{abd}^{(K)} = \log(N^{(K)} + 1)$$

These abundance values are plotted in the right column of Fig. 4a.

In silico permutation of cells in microenvironments

Analysis of SPACE-GM embeddings uncovered groups of microenvironments that are disease relevant; we then applied permutation experiments on microenvironments-of-interest to discover and validate structural motifs that are indicative of phenotypes.

Two general forms of permutations are implemented.

- Dispersed permutation

- A list of target cell types is provided.
- On the microenvironment/patch/sample (jointly denoted as cellular graphs), identify all cells whose cell type appears in the target list and record their spatial location and other cellular features.
- Randomly permute the list of cells identified in the previous step, and then assign the permuted cell types and cellular features back to the original cellular graph.
- Run (trained) SPACE-GM on the permuted cellular graph and perform microenvironment aggregation if needed.

- Coherent permutation

- A list of target cell types and the centre coordinate of the cellular graph are provided.
- On the cellular graph, identify all cells whose cell type appears in the target list and record their spatial location and other cellular features.
- Sort the list of cells identified in the previous step by cell type.
- Calculate polar coordinates for the list of cells using the cellular-graph centre as the pole.
- Assign cell types and cellular features back to the original cellular graph sequentially following the order of the azimuthal angle. The resulting permuted cellular graphs should have cells of the same type appearing in the same sector around the graph centre.

In the experiment on heterogeneous TMEs, we performed permutations on patches and whole samples. Patches were selected following the procedure below:

- For each of the UPMC-HNC samples, we derived the cluster assignment of every individual cell following the dimensionality reduction and clustering pipeline introduced above.
- We iterated through the list of all cells assigned to the heterogeneous TME until we found a cell that satisfied the following criterion. Note that up to one patch can be selected per sample.
  (a) All cells that were within 185 µm of the query cell (denoted as a patch) were isolated.
  (b) More than 40% of the cells in the patch were assigned to heterogeneous TMEs.
- The query cell and its surrounding patch were extracted. For all cells in the patch that were within 110 µm of the centre (to guarantee the completeness of the 3-hop neighbourhood), their microenvironments were extracted and predictions were performed.

A total of 61 patches were selected, from which we picked the 50 patches that had higher entropy (of cell-type frequency vector). Permutations were performed on the following cell types: Tumour 2, Tumour 4, Tumour 5 (Ki67$^+$) and Tumour 6, in which the interaction between Tumour 4 and Tumour 5 (Ki67$^+$) had the biggest influence.

In the experiment on granulocyte–tumour microenvironments, we performed coherent and dispersed permutations on different microenvironment groups to reverse the organization pattern of granulocytes. We did not perform patch-level permutation due to difficulty in finding regional patches rich in either microenvironment. All cell types were included in the target list and permuted in this experiment. Note that in the coherent permutation case, all cells except for tumour cells (Tumour 1 through Tumour 6) were aligned coherently according to the procedure above, and tumour cells were first combined and randomly permuted before aligning to avoid confounding. See Supplementary Fig. 7 for examples.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the results in this study involve patient data and are available from the corresponding author on reasonable request.

Code availability

Commercial software (Akoya Biosciences, Enable Medicine) was used to preprocess images and to classify cells using methods based on published algorithms. Deepcell (https://github.com/vanvalenlab/deepcell-tf) was used to segment cells from multiplexed immunofluorescence images. The codes used for the construction of spatial cellular graphs and for the following analyses are available at https://gitlab.com/enable-medicine-public/space-gm.

References

1. Luca, B. A. et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell* **184**, 5482–5496 (2021).
2. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
3. Bejarano, L., Jordão, M. J. C. & Joyce, J. A. Therapeutic targeting of the tumor microenvironment. *Cancer Discov.* **11**, 933–959 (2021).
4. Lomakin, A. et al. Spatial genomics maps the structure, character and evolution of cancer clones. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.16.439912 (2021).
5. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
6. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
7. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
8. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981 (2018).
9. Angelo, M. et al. Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 (2014).
10. Lin, J.-R., Fallahi-Sichani, M., Chen, J.-Y. & Sorger, P. K. Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging. *Curr. Protoc. Chem. Biol.* **8**, 251–264 (2016).
11. Ali, H. R. et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**, 163–175 (2020).
12. Schürch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**, 1341–1359 (2020); erratum **183**, 838 (2020).
13. Bhate, S. S., Barlow, G. L., Schürch, C. M. & Nolan, G. P. Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors. *Cell Syst.* **13**, 109–130 (2022).
14. Zhou, Y. et al. Cgc-net: cell graph convolutional network for grading of colorectal cancer histology images. In *Proc. IEEE/CVF International Conference on Computer Vision Workshops* 388–398 (IEEE, 2019).
15. Lu, W., Graham, S., Bilal, M., Rajpoot, N. & Minhas, F. Capturing cellular topology in multi-gigapixel pathology images. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 260–261 (IEEE, 2020).
16. Anand, D., Gadiya, S. & Sethi, A. Histographs: graphs in histopathology. In *Medical Imaging 2020: Digital Pathology* Vol. 11320 (eds Tomaszewski, J. E. & Ward, A. D.) 150–155 (SPIE, 2020).
17. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *Proc. 34th International Conference on Machine Learning* Vol. 70 (eds Precup, D. & Teh, Y. W.) 1263–1272 (JMLR.org, 2017).
18. Wu, Z. et al. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4–24 (2021).
19. Brbić, M. et al. Annotation of spatially resolved single-cell data with STELLAR. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.24.469947 (2021).
20. Innocenti, C. et al. An unsupervised graph embeddings approach to multiplex immunofluorescence image exploration. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.09.447654 (2021).
21. Fischer, D. S., Schaar, A. C. & Theis, F. J. Learning cell communication from spatial graphs of cells. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.11.451750 (2021).
22. Kim, J. et al. Unsupervised discovery of tissue architecture in multiplexed imaging. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.15.484534 (2022).
23. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In *Proc.7th International Conference on Learning Representations* (OpenReview, 2019); https://openreview.net/forum?id=ryGs6iA5Km
24. Argiris, A., Karamouzis, M. V., Raben, D. & Ferris, R. L. Head and neck cancer. *Lancet* **371**, 1695–1709 (2008).
25. Dalerba, P. et al. CDX2 as a prognostic biomarker in stage II and stage III colon cancer. *N. Engl. J. Med.* **374**, 211–222 (2016).
26. Uppaluri, R. et al. Neoadjuvant and adjuvant pembrolizumab in resectable locally advanced, human papillomavirus–unrelated head and neck cancer: a multicenter, phase II trial. *Clin. Cancer Res.* **26**, 5140–5152 (2020).
27. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2018).
28. Blise, K. E., Sivagnanam, S., Banik, G. L., Coussens, L. M. & Goecks, J. Single-cell spatial architectures associated with clinical outcome in head and neck squamous cell carcinoma. *NPJ Precis. Oncol.* **6**, 10 (2022).
29. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
30. Trellakis, S. et al. Polymorphonuclear granulocytes in human head and neck cancer: enhanced inflammatory activity, modulation by cancer cells and expansion in advanced disease. *Int. J. Cancer* **129**, 2183–2193 (2011).
31. Lonardi, S. et al. Tumor-associated neutrophils (TANs) in human carcinoma-draining lymph nodes: a novel TAN compartment. *Clin. Transl. Immunol.* **10**, e1252 (2021).
32. Coffelt, S. B., Wellenstein, M. D. & de Visser, K. E. Neutrophils in cancer: neutral no more. *Nat. Rev. Cancer* **16**, 431–446 (2016).
33. Shojaei, F. et al. Bv8 regulates myeloid-cell-dependent tumour angiogenesis. *Nature* **450**, 825–831 (2007).
34. Di Mitri, D. et al. Tumour-infiltrating Gr-1$^+$ myeloid cells antagonize senescence in cancer. *Nature* **515**, 134–137 (2014).
35. Coffelt, S. B. et al. IL-17-producing γδ T cells and neutrophils conspire to promote breast cancer metastasis. *Nature* **522**, 345–348 (2015).
36. Wculek, S. K. & Malanchi, I. Neutrophils support lung colonization of metastasis-initiating breast cancer cells. *Nature* **528**, 413–417 (2015).
37. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
38. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
39. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2021).
40. Hickey, J. W., Tan, Y., Nolan, G. P. & Goltsev, Y. Strategies for accurate cell type identification in CODEX multiplexed imaging data. *Front. Immunol.* **12**, 727626 (2021).
41. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
42. Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-event prediction with neural networks and Cox regression. *J. Mach. Learn. Res.* **20**, 1–30 (2019).

43. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations* (ICLR, 2015); https://www.iclr.cc/archive/www/2015.html
44. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. In *Proc. 33rd International Conference on Neural Information Processing Systems* 8026–8037 (Curran Associates, 2019).
46. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch geometric. Preprint at https://doi.org/10.48550/arXiv.1903.02428 (2019).
47. Kipf, T. N., & Welling, M. Semi-supervised classification with graph convolutional networks. In *Proc. 5th International Conference on Learning Representations* (OpenReview, 2017); https://openreview.net/forum?id=SJU4ayYgl
48. Veličković, P. et al. Graph attention networks. In *Proc. 6th International Conference on Learning Representations* (OpenReview, 2018); https://openreview.net/forum?id=rJXMpikCZ
49. Hamilton, W., Ying, Z., & Leskovec, J. Inductive representation learning on large graphs. In *Proc. 31st International Conference on Neural Information Processing Systems* 1025–1035 (Curran Associates, 2017).
50. Li, Y., Tarlow, D., Brockschmidt, M. & Zemel, R. Gated graph sequence neural networks. In *Proc. 4th International Conference on Learning Representations* (OpenReview, 2016); https://openreview.net/forum?id=HSgW989Kp-q
51. Vinyals, O., Bengio, S. & Kudlur, M. Order matters: sequence to sequence for sets. In *Proc. 4th International Conference on Learning Representations* (ICLR, 2016); https://www.iclr.cc/archive/www/2016.html

Acknowledgements

J.Z. discloses support for the research described in this study from NSF CAREER (grant no. 1942926). K.S. discloses support for the research described in this study from a Knight-Hennessy Fellowship.

Author contributions

Z.W. and A.E.T. designed the model and computational experiments in consultation with E.W. and K.S. Z.W. and A.E.T. analysed the data. Z.W. and A.E.T. wrote the manuscript with input from all authors. G.W.C., P.D.D., A.M.E., R.U. and U.D. provided samples for the experiments. H.J.K., H.B.D. and R.P. performed the experiments and data preprocessing. J.Z. and A.T.M. were responsible for the overall direction and planning of the project.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41551-022-00951-w.

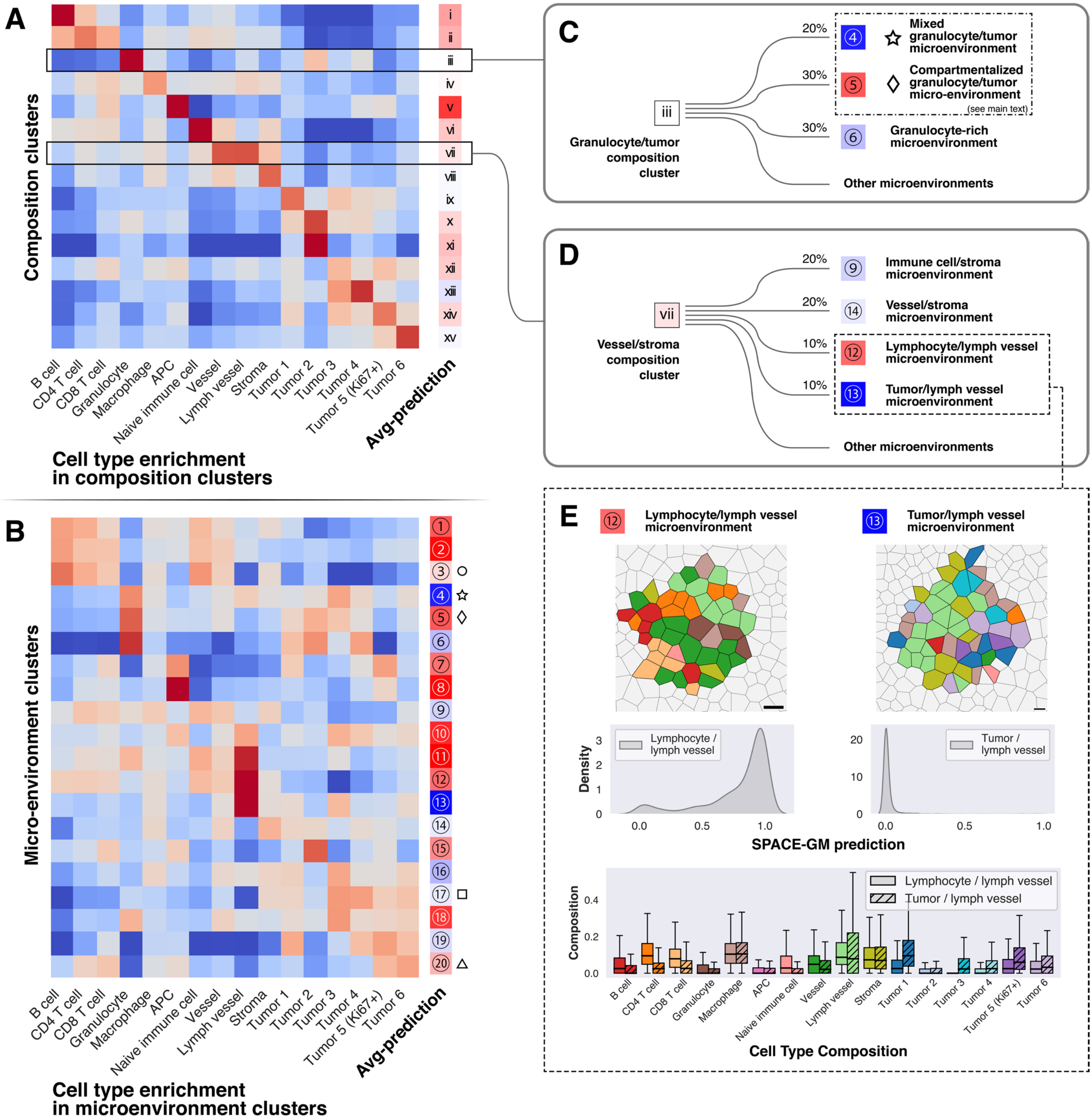
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-022-00951-w.

Correspondence and requests for materials should be addressed to Alexandro E. Trevino, Aaron T. Mayer or James Zou.

Peer review information *Nature Biomedical Engineering* thanks Tae Hyun Hwang, Jonathan Nowak and Jianyu Rao for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Comparison of composition-based clusters and microenvironment clusters. A. and **B.** We generate composition-based clusters with cell type compositions of 3-hop subgraphs (microenvironments) following the same procedure. Note that the average predictions of microenvironment clusters are much more polarized. **C.** The composition cluster highly enriched with granulocytes has neutral average predictions/labels, it could be further dissected into multiple sub-clusters that belong to different microenvironment clusters. We see a pair of granulocyte/tumor microenvironments that have opposite outcome labels, see **Results** for further discussion. **D.** Similarly, the composition cluster enriched with vessel/lymph vessel cells could be dissected into multiple sub-clusters, from which we notice two microenvironment clusters that are both enriched with lymph vessel cells but have different composition and outcome predictions, see Supplementary Note 7 for further discussion. **E.** Comparison of the two microenvironments enriched in lymph vessel cells: the left column shows a microenvironment with more lymphocytes and has overall positive outcomes; the right column shows a contrasting group with more tumor cells and much worse outcome predictions. Observation of tumor cells in close vicinity of lymph vessels indicates potential lymphovascular invasion and will lead to worse prognosis, which aligns with model predictions.

nature portfolio

Corresponding author(s): Alexandro E. Trevino, Aaron T. Mayer and James Zou

Last updated by author(s): Sep 5, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Commercial software (Akoya Biosciences, Enable Medicine) was used to preprocess images and to classify cells using methods based on published algorithms.
Data analysis	Deepcell (https://github.com/vanvalenlab/deepcell-tf) was used to segment cells from multiplexed immunofluorescence images. The codes used for the construction of spatial cellular graphs and for the following analyses are available at https://gitlab.com/enable-medicine-public/space-gm.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data supporting the results in this study involve patient data and are available for research purposes on reasonable request.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	UPMC-HNC: 81 patients, 18 female and 63 male, all diagnosed with head and neck cancer; treatment: surgery. Stanford-CRC: 161 patients, 88 female and 73 male, all diagnosed with colorectal cancer; treatment: surgery. DFCI-HNC: 29 patients, 6 female and 23 male, all diagnosed with head and neck cancer; treatment: neoadjuvant chemotherapy and surgery.
Population characteristics	Banked samples were collected in an unbiased manner from patients with head-and-neck cancer or colorectal cancer from the corresponding institute.
Recruitment	Samples were chosen as an unbiased representation of patients served at each of the three institutes in the study, without regards to demographic or clinical covariates.
Ethics oversight	Sample collections were approved by institutional review boards at the University of Pittsburgh Medical Center, the Stanford Medical Center and the Dana-Farber Cancer Institute.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Samples were chosen as an unbiased representation of patients served at each of the three institutes in the study, without regards to demographic or clinical covariates. No statistics were used in the determination of sample size. We deemed these sample sizes sufficient because they are larger than any other cohort published so far for this data type by an order of magnitude; because data collection is rate-limiting, so larger cohorts would require considerable added expenditure; and because our results reached statistical significance without predetermination of sample size.
Data exclusions	No data were excluded prior to analysis. Some samples were filtered from downstream analyses on the basis of clearly defined assay-quality control metrics.
Replication	We included training and test splits for the data, mirroring experimental covariates, and used samples collected using different methods (for section preparation, and for section size, in particular) and from different institutions, to validate the original cohort.
Randomization	Because the investigators did not manually allocate samples into groups, randomization was not relevant to this study. We controlled for technical effects by using cross-validation across purely technical groups (such as experiment batch).
Blinding	Because the investigators did not manually allocate samples into groups for analysis, blinding was not relevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used	Detailed in Methods.
Validation	The antibodies were purchased from Akoya Biosciences. No additional validations were performed.