Path2Space: An AI Approach for Cancer Biomarker Discovery Via Histopathology Inferred Spatial Transcriptomics

Eldad D. Shulman[1,5], Emma M. Campagnolo[1,5], Roshan Lodha[1], Thomas Cantore[1], Tom Hu[2], Maclean Nasrallah[2], Danh-Tai Hoang[3], Kenneth Aldape[4], Eytan Ruppin[1]

[1] Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA.

[2] Division of Neuropathology, Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[3] Biological Data Science Institute, College of Science, Australian National University, Canberra, ACT, Australia.

[4] Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA.

[5] These authors contributed equally: Eldad D. Shulman, Emma M. Campagnolo.

e-mail of corresponding authors: eldad.shulman@nih.gov, eytan.ruppin@nih.gov

Abstract

The rise of spatial transcriptomics (ST) is transforming our understanding of tumor heterogeneity by enabling high-resolution, location-specific mapping of gene expression across the tumor microenvironment. However, the translational potential of spatial transcriptomics is still limited by its high cost, hindering the assembly of large patient cohorts needed for robust biomarker discovery. Here we present *Path2Space*, a deep learning model trained on spatial transcriptomics data to predict spatial gene expression directly from histopathology slides. Studying breast cancer, the model was first trained on one ST breast cancer patients cohort and further independently tested and validated on two other ST cohorts. Path2Space identifies 4,500 genes whose spatial expression is robustly predicted in all three cohorts, markedly more accurate than existing ST predictors. Second, we show that based on the inferred ST, Path2Space accurately infers cell-type abundances in the tumor microenvironment (TME). Thirdly, Applying Path2Space to the TCGA breast cancer cohort we identify three TME spatially grounded subsets of patients that have markedly different survival rates. Fourth, analyzing two large breast cancer patients cohorts treated with trastuzumab, we show that the heterogeneity of the inferred HER2 levels across the tumor microenvironment is predictive of response, concordant with the bystander hypothesis. Finally, we developed a *spatially interpretable* model for predicting trastuzumab response directly from the H&E-slides. This model's performance surpasses that of existing response prediction models that use *measured* bulk multi-omics data. Path2Space is the first approach capable of robustly delineating the tumor microenvironment *on an unprecedented scale* directly from the tumor slides. It heralds the upcoming development of robust spatially grounded biomarkers of patients survival and treatment that promises to transform precision oncology in both the developed and developing worlds.

Introduction

The rise of spatial transcriptomics is transforming our understanding of tumor heterogeneity by enabling high-resolution, location-specific mapping of gene expression across tumors and their microenvironments. In lung adenocarcinoma, it revealed macrophage phenotypic diversity [1,2]; In breast cancer, it illuminated immune-enriched clusters with varying infiltration patterns [3,4] and revealed recurrent neoplastic cell heterogeneity, leading to the identification of nine 'ecotypes' with unique cellular compositions and clinical outcomes[5]; In glioblastoma, it uncovered distinct transcriptional programs of sub-clonal structures[6]. It has also characterized the architecture of invasive tumors in melanoma[7], enabled 3D reconstruction of colorectal cancer[8], and explored intratumoral microbial communities[9]. In HPV-negative oral squamous cell carcinoma, it revealed distinct transcriptional profiles in tumor core and leading edge, with implications for prognosis and drug response across multiple cancer types[10]. Moreover, spatial transcriptomics has elucidated the intricate cross-talk between cancer cells and proximal immune cells in the tumor microenvironment, providing insights into tumor progression and potential therapeutic targets [11]. By providing unprecedented spatial context to gene expression, spatial transcriptomics holds immense potential to unlock novel biomarkers, decipher treatment resistance mechanisms, and guide more precise cancer therapies.

Despite these remarkable advances, the translational potential of spatial transcriptomics remains constrained by its high cost, approximately $35,000 for a typical experiment[12]. This cost limitation hinders the assembly of large patient cohorts, essential for discovering robust biomarkers, developing predictive models, and ensuring the statistical power needed to validate findings across diverse populations. Given that spatial transcriptomics data is often paired with aligned H&E images, predicting spatial gene expression from H&E slides presents an opportunity to extend these insights at scale. Deep learning applied to H&E images has already shown promise in identifying features not apparent to human observers, including genetic mutations[13], bulk mRNA expression[14], and methylation patterns[15].

Previous efforts to predict spatial gene expression from H&E images have shown potential but have been of limited predictive ability, robustly predicting the expression of very few hundreds of genes, at most. Not less importantly, the results reported in these studies have remained confined to spatial transcriptomics datasets, without attempting to study and show their translational clinical applicability for predicting patients survival and treatment response in larger clinical cohorts where only H&E images are available[16–23]. Expanding this capability is essential for uncovering the biological and clinical significance of spatial gene expression while avoiding the prohibitive costs of direct measurements.

To address this challenge, we developed *Path2Space*, a model trained on spatial transcriptomics data that predicts spatial gene expression from histopathology slides with unprecedented accuracy. Given the prevalence of breast cancer and the availability of high-quality data, Path2Space was specifically designed and optimized for this disease. The model demonstrates robust predictive performance on external datasets and can further infer tile-level cell type composition from the predicted spatial gene expression. We then applied Path2Space to the TCGA breast cancer cohort, identifying prognostic metaclusters based on spatial transcriptomics cluster compositions. We further applied Path2Space

to two large cohorts of HER2+ breast cancer patients treated with trastuzumab, showing that our method can uncover spatial biomarkers and predict therapy response at accuracy levels equivalent to those obtained from the measured tumor transcriptomics. Due to the high resolution and cost-effectiveness of our approach, Path2Space forms the basis for similar analyses of large clinical datasets in different indications and treatments, enabling discoveries currently unattainable with the limited scale of existing spatial transcriptomics studies.

Results

Study overview

We developed Path2Space, a deep-learning model to predict spatial gene expression from H&E slide images of breast cancer tumors, aiming to uncover microenvironment characteristics and spatial biomarkers associated with treatment response.

To develop the spatial gene expression predictor, we used a dataset of 25 high-quality slides from 15 triple-negative breast cancer (TNBC) patients[24], each with matched H&E images and gene expression profiles derived from 55 µm diameter spots (**Fig. 1a**, **Methods**). To mitigate technical variability arising from low coverage in spatial transcriptomics data, we implemented a spatial smoothing technique that averages the expression values of each spot with its immediate spot neighbors. This smoothing was applied to both model predictions and original expression measurements (**Supplementary Fig. 1**) and used for model evaluation and downstream analyses. We initially assessed model performance using five-fold cross-validation on the TNBC cohort and further validated it on two independent external spatial transcriptomics datasets: one from a cohort of three patients obtained from 10x Genomics, and another from NCBI, consisting of a single slide from one patient [25].

To estimate cell type fractions at Visium spot resolution using Path2Space-predicted spatial gene expression, we leveraged the PanopTILs dataset[26]. This dataset includes H&E images and cell type annotations for 1,709 regions of interest (ROIs) from 151 TCGA breast cancer patients. External validation was performed using H&E images and cell type annotations from the 10x cohort (**Fig. 1b**).

After testing and validating Path2Space's prediction capabilities for both spatial gene expression and cell-type composition directly from pathology slide images, we turned to study its application to discover spatial biomarkers utilizing pathology slides of large cohorts. We applied Path2Space to the TCGA breast cancer cohort, identifying and characterizing spatial architectures associated with disease-free survival (**Fig. 1c**). We also applied Path2Space to H&E images from two HER2+ breast cancer cohorts treated with chemotherapy and trastuzumab, TransNEO[27] (n=63) and IMPRESS[28] (n=62) to study the tumor microenvironment and infer treatment response biomarkers (**Fig. 1d**).

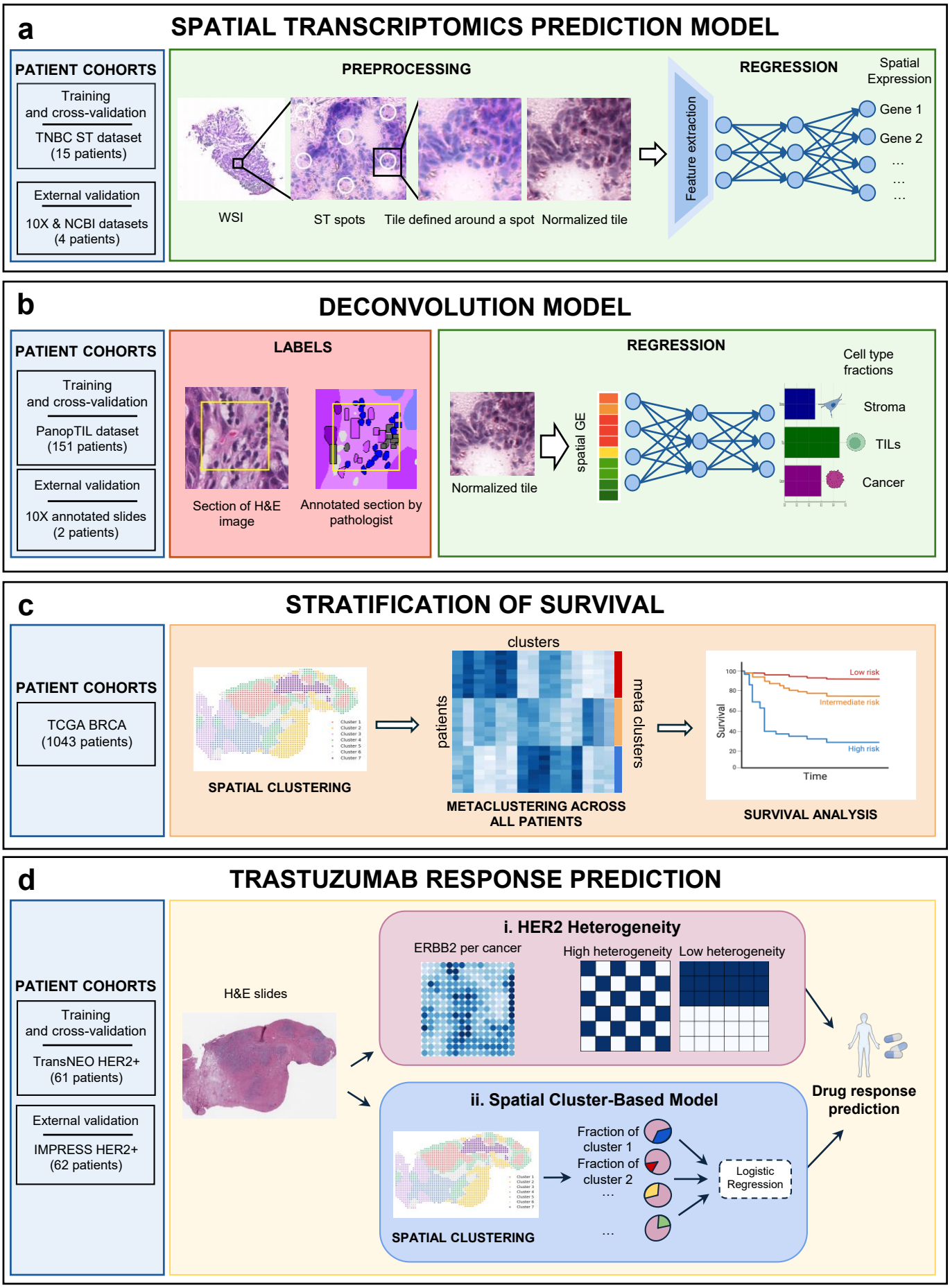

a
SPATIAL TRANSCRIPTOMICS PREDICTION MODEL
PATIENT COHORTS
Training and cross-validation
TNBC ST dataset (15 patients)
External validation
10X & NCBI datasets (4 patients)
PREPROCESSING
WSI
ST spots
Tile defined around a spot
Normalized tile
Feature extraction
REGRESSION
Spatial Expression
Gene 1
Gene 2
b
DECONVOLUTION MODEL
PATIENT COHORTS
Training and cross-validation
PanopTIL dataset (151 patients)
External validation
10X annotated slides (2 patients)
LABELS
Section of H&E image
Annotated section by pathologist
REGRESSION
Normalized tile
spatial GE
Cell type fractions
Stroma
TILs
Cancer
c
STRATIFICATION OF SURVIVAL
PATIENT COHORTS
TCGA BRCA (1043 patients)
SPATIAL CLUSTERING
clusters
patients
meta clusters
METACLUSTERING ACROSS ALL PATIENTS
Low risk
Intermediate risk
High risk
Survival
Time
SURVIVAL ANALYSIS
d
TRASTUZUMAB RESPONSE PREDICTION
PATIENT COHORTS
Training and cross-validation
TransNEO HER2+ (61 patients)
External validation
IMPRESS HER2+ (62 patients)
H&E slides
i. HER2 Heterogeneity
ERBB2 per cancer
High heterogeneity
Low heterogeneity
ii. Spatial Cluster-Based Model
SPATIAL CLUSTERING
Fraction of cluster 1
Fraction of cluster 2
Logistic Regression
Drug response prediction

Fig. 1: Overview of the datasets and computational workflow. A, The Path2Space spatial gene expression model was trained and evaluated on datasets with paired H&E and spatial transcriptomics from the TNBC cohort, with the 10x cohort serving as an external validation set. Whole slide images (WSIs) were partitioned into tiles around each of the spatial transcriptomics spots. These tiles underwent color normalization followed by feature extraction using the CTransPath digital pathology foundational model. The extracted features were then input into a multi-layer perceptron (MLP) to infer gene expression for each tile. **b**, The Path2Space cell type deconvolution model was trained on cell type annotations from the PanopTILs dataset, with the 10x and NCBI cohorts as the external validation sets. Taking the WSI as input, the model uses the inferred spatial gene expression as input to an MLP that estimates cell type fractions per tile. **c**, Path2Space predicted spatial gene expression from TCGA BRCA WSIs, which were used to identify spatial clusters that define tumors spatial profiles. Based on these profiles, an unsupervised algorithm grouped tumors into distinct, prognostic, spatially grounded subtypes. **d**, Two approaches for predicting Trastuzumab response. (i) HER2 Heterogeneity: From the slides, we inferred the HER2 pathway expression per cancer cell and quantified the heterogeneity of its spatial distribution, with high heterogeneity shown to be associated with treatment response. (ii) Spatial Cluster-Based Model: The same approach from panel b was used to define tumor profiles based on their spatial cluster composition. A logistic regression model was trained to predict response to trastuzumab using the tumor spatial profiles, evaluated via cross-validation on the TransNEO dataset, and validated on the independent IMPRESS dataset.

Predicting spatial gene expression from H&E slides

We analyzed a TNBC dataset, consisting of H&E slides and matched gene expression profiles obtained via the Visium platform, to develop a model for predicting gene expression directly from H&E slides. We selected 25 high-quality slides from 15 patients out of 36 based on spot density (**Fig. S1**, **Methods**). These slides contained an average of 1276 ± 444 spots (mean ± SD) and detected 4594 ± 2465 genes per spot (expression values > 0) (**Supplementary Fig. 2a, b**). Due to the high proportion of missing values typical in spatial transcriptomics data, we focused on 14,068 genes detected in ≥5% of spots on each slide.

To evaluate our model, we employed a leave-one-patient-out cross-validation approach. The model was iteratively trained on slides from all patients except one, and then tested on the held-out patient's slides. We assessed gene-wise prediction accuracy by calculating the average Pearson correlation between spatially smoothed (as described in the overview) actual and predicted expression values across all spots in each slide (**Methods**). This yielded a distribution of gene-wise correlations with a median of 0.37 (**Fig. 2a**). Notably, 6,022 out of 14,068 genes (43%) demonstrated a mean correlation above 0.4 (**Fig. 2b**), indicating robust predictive performance for a substantial portion of the transcriptome. We observed a strong correlation between the abundance of specific gene measurements (fewer spots with missing values) and Path2Space's ability to predict gene expression (**Supplementary Fig. 2c**), suggesting that prediction accuracy could further improve in the future with the advent of higher-depth sequencing.

To further validate Path2Space's performance and demonstrate its generalizability, we applied the ST TNBC-trained model to predict spatial gene expression profiles from H&E slide images in two independent cohorts: one from 10x consisting of 4 slides from 3 patients and another from NCBI consisting of 1 slide from 1 patient. The median Pearson correlation between actual and predicted expression values, averaged across all spots per slide and across the cohort, was 0.35 (**Fig. 2a**), with 5,237 genes exceeding a correlation coefficient of 0.4 (**Fig. 2b**). Of note, 5,867 genes exceeded a correlation coefficient of 0.4 in the 10X cohort (**Fig. 2b**).

Remarkably, we observed substantial concordance in highly predicted genes between the TNBC-trained model and the two validation cohorts: 62% (4,311/6,850) of genes achieving a correlation >0.4 in one cohort also surpassed this threshold in the other (**Fig. 2c**). This overlap included clinically relevant genes such as ERBB2 (HER2), MKI67 (Ki-67), and MYC, which are crucial in breast cancer prognosis and treatment decisions (**Fig. 2f-h**). Remarkably, gene set enrichment analysis of the 4,311 concordant genes (correlation >0.4 in all cohorts) revealed significant enrichment (FDR < 0.05) for all 10 of the 10 well known cancer hallmark processes. Notably, pathways related to avoiding immune destruction, evading growth suppressors, and genome instability and mutation showed the strongest enrichment (**Fig. 2d**).

Given that Path2Space was trained exclusively on fresh frozen (FF) slides from the ST TNBC dataset, we assessed its performance on formalin-fixed paraffin-embedded (FFPE) slides included in the two external validation datasets, which included three FF slides and two FFPE slides. Median gene expression correlations ranged from 0.32 to 0.40 for both slide types (**Fig. 2e**), indicating that Path2Space maintained strong predictive performance on FFPE slides, comparable to FF samples. This testifies that Path2Space is well-suited for application to both fresh and archival tissue specimens, important for our translational research goal.

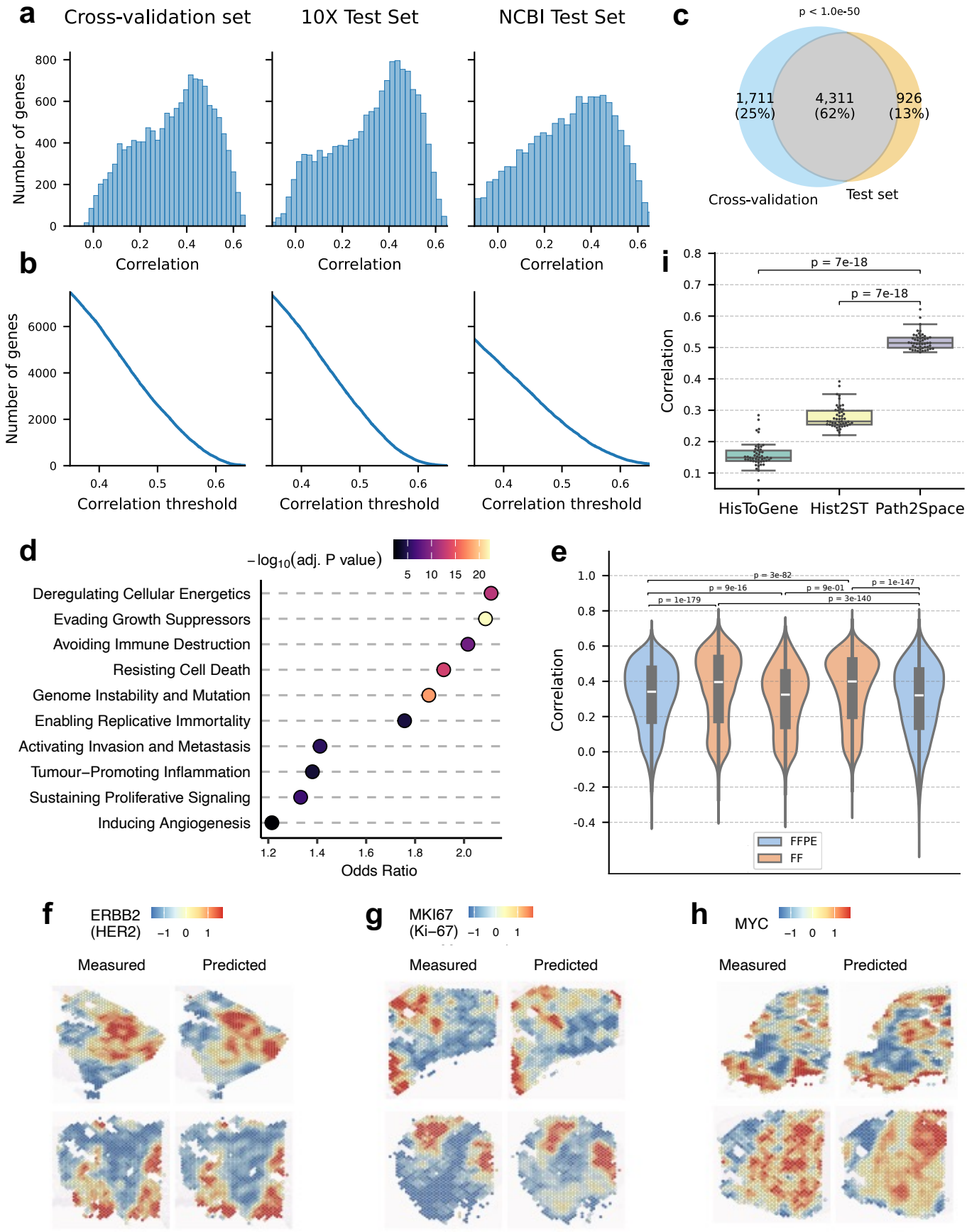

a
Cross-validation set
10X Test Set
NCBI Test Set
Number of genes
Correlation
b
Number of genes
Correlation threshold
c
p < 1.0e-50
1,711 (25%)
4,311 (62%)
926 (13%)
Cross-validation
Test set
i
p = 7e-18
p = 7e-18
Correlation
HisToGene
Hist2ST
Path2Space
d
−log10(adj. P value)
Deregulating Cellular Energetics
Evading Growth Suppressors
Avoiding Immune Destruction
Resisting Cell Death
Genome Instability and Mutation
Enabling Replicative Immortality
Activating Invasion and Metastasis
Tumour−Promoting Inflammation
Sustaining Proliferative Signaling
Inducing Angiogenesis
Odds Ratio
e
p = 3e-82
p = 9e-16
p = 9e-01
p = 1e-147
p = 1e-179
p = 3e-140
Correlation
FFPE
FF
f
ERBB2 (HER2)
Measured
Predicted
g
MKI67 (Ki−67)
Measured
Predicted
h
MYC
Measured
Predicted

Fig. 2: Performance of Path2Space in predicting spatial gene expression. a, Histograms of Pearson correlation coefficients between predicted and actual gene expression values, showing cross-validation results from the TNBC cohort (left) and independent validation in the 10x cohort (middle) and NCBI (right) datasets. **b,** Curves depicting the number of genes surpassing specific correlation thresholds, shown for the TNBC cohort (left), 10x cohort (middle), and NCBI (right). **c,** Venn diagram illustrating the intersection of genes between the two cohorts that exhibit a predicted/measured correlation coefficient exceeding 0.4. The p-value was calculated using the hypergeometric test. **d,** Gene set enrichment analysis of the 4,311 genes predicted with a correlation above 0.4 in both cohorts. **c**, Each row represents a different cancer hallmark, with the horizontal axis displaying the odds ratio. The color scale represents the -log10 adjusted p-value; all gene sets were significant (adjusted p-value ≤ 0.05). **e**, Violin plots showing Pearson correlation coefficients between predicted and actual gene expression values (y-axis) across individual slides (x-axis) from the 10x cohort. Orange violins represent the three fresh frozen slides, while the blue violin shows the distribution for the two FFPE slides. Statistical significance is evaluated using P values from the two-tailed Mann-Whitney U test. **f-h,** Examples of clinically important genes with strong predictions: ERBB2 (HER2) (**f**), MKI67 (Ki-67) (**g**), and MYC (**h**). Spots on the slides are colored according to scaled expression levels. In each panel, the left column shows the measured expression levels, and the right column shows the Path2Space-predicted expression levels for slides in the same row. **i,** Boxplots comparing the performance of Path2Space with Hist2Gene and His2ST, evaluated using leave-one-out cross-validation on a HER2-positive dataset (41 slides). The boxplots display the correlation coefficients between predicted and actual gene expression values for the top 50 genes per slide across each method. Statistical significance is evaluated using P values from the two-tailed Mann-Whitney U test.

Comparison of Path2Space with state-of-the-art methods

Previous efforts to predict spatial gene expression from H&E-stained images have shown promise but remain limited in predictive power. Typically, these methods achieve correlations above 0.4 for only a small subset—dozens to a few hundred—of genes[29], restricting their utility in clinical applications. In a recent comparative review, our previous method, DeepPT[14], demonstrated superior performance in bulk expression inference when applied to spatial transcriptomics data, outperforming five other specialized techniques[30].

To assess Path2Space, we benchmarked it against two state-of-the-art approaches for predicting spatial gene expression from tissue slides: HisToGene[31] and Hist2ST[19]. To ensure a fair comparison, we retrained our model on the same legacy spatial transcriptomics dataset of 36 slides from 8 HER2-positive breast cancer patients[3] used by these methods, employing an identical leave-one-out cross-validation strategy. For the comparison, we evaluated the correlation between our smoothed predictions and the raw, unsmoothed measurements, consistent with the approach used by the other methods. We compared the models using the metric reported in these publications: the mean correlation of the top 50 predicted genes. Path2Space markedly outperformed both models, achieving a median correlation of 0.51, compared to 0.15 and 0.26 for HisToGene and Hist2ST, respectively (**Fig. 2i**).

These findings underscore Path2Space's capability to accurately infer spatial gene expression patterns from H&E images of breast cancer, highlighting its potential for biomarker discovery in large cohorts.

Inferring spatial cell type composition from the inferred spatial transcriptomics in a large cohort of 150 breast cancer patients

Visium's 55 μm spots typically capture signals from multiple cell types, making cellular decomposition crucial for characterizing tumor spatial landscapes. We utilized predicted gene expression in spots to estimate cell type fractions, leveraging the PanopTILs dataset. This dataset comprises 1,709 regions of interest (ROIs) from 151 breast cancer patients in TCGA, annotated via a crowdsourcing process involving medical students and pathologists.

Leveraging the similarity between typical ROI size (50x50 μm) and Visium spot size, we employed Path2Space to infer gene expression for each annotated region. We then developed an MLP regressor using 5-fold cross-validation, grouped by patients, to estimate fractions of the three most prevalent cell types: cancer cells, tumor-infiltrating lymphocytes (TILs), and stromal cells. Model performance was evaluated using Pearson correlation between actual and predicted cell type fractions across all ROIs. Our model achieved correlations of 0.82, 0.74, and 0.58 for cancer cells, TILs, and stromal cells, respectively, with the lower correlation for stromal cells potentially attributable to their lower abundance in the dataset (**Fig. 3a**).

To elucidate our deconvolution model's utilization of H&E-inferred gene expression, we analyzed the effective weights of input genes for each cell type's fraction prediction. Gene set enrichment analysis of these weights revealed significant enrichment (q-value < 0.05) of breast cancer signatures for cancer cell fraction prediction and lymphocyte signatures for TIL prediction (**Fig. 3b,c**). While the stromal signature exhibited the highest enrichment score for stromal prediction, it did not reach statistical significance.

To assess the performance of our Path2Space deconvolution model on independent external validation, we applied it to predict cell type fractions in two slides from the 10x dataset with spot-level pathologist annotations indicating the presence or absence of specific cell types. We obtained ROC AUC scores of 0.9, 0.87, and 0.84 for cancer cells, lymphocytes, and stromal cells, respectively. Remarkably, our H&E-inferred gene expression-based model showed comparable performance to SpaCET[32], which uses the *measured* gene expression as input. While SpaCET outperformed Path2Space for cancer and stromal cells, it demonstrated inferior performance for lymphocytes (**Fig. 3d**). The comparable cell identity decomposition between Path2Space's H&E-based approach and spatial transcriptomics underscores its potential to leverage widely available H&E images, broadening the accessibility of spatial transcriptomics analysis.

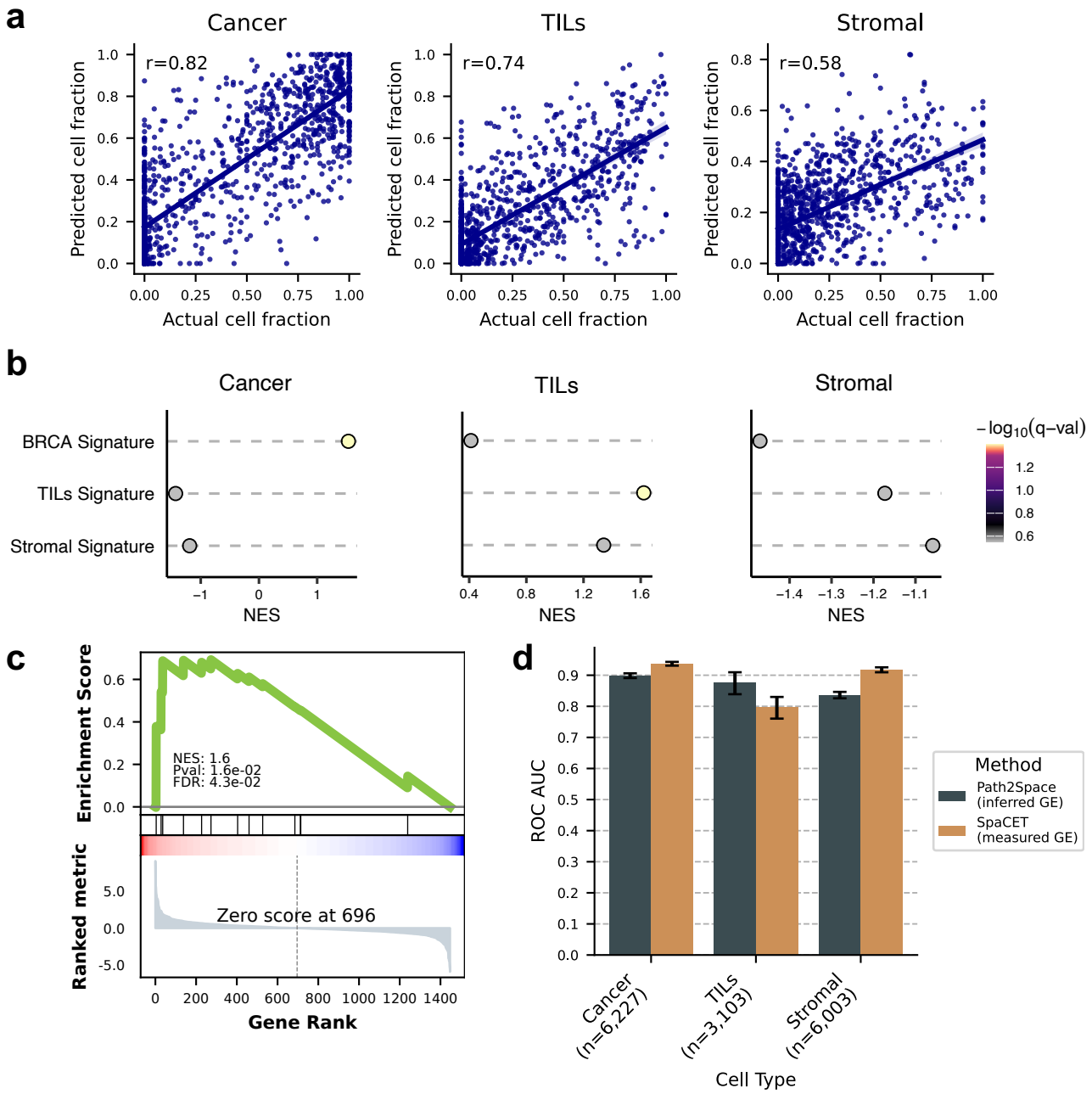


Fig. 3: Path2Space performance in estimating cell type fractions from inferred spatial gene expression. a, Scatter plots showing cross-validation performance in the PanopTIL dataset. Actual values are on the x-axes and predicted values on the y-axes for cancer cells (left), tumor-infiltrating lymphocytes (TILs, center), and stromal cells (right). Pearson correlation coefficients between actual and predicted values are displayed in each plot. **b**, Gene set enrichment analysis (GSEA) of the effective weights of input genes. Each row represents a distinct cell type signature. The x-axis shows the Normalized Enrichment Score (NES). The color scale indicates -log10 adjusted p-value, with gray denoting values below the significance threshold (FDR q-value ≥ 0.05). Analyses for the cancer, TIL, and stromal nodes of the deconvolution models are shown from left to right. c, GSEA plot demonstrating the enrichment of the TIL signature in the model's

TIL node. Genes are ranked according to their effective weights in the TIL node. d, Performance comparison of deconvolution models on the 10x validation set. Path2Space deconvolution is based on inferred gene expression, while SpaCET uses measured gene expression. The y-axis shows the area under the receiver operating characteristic curve (AUC) for cell fraction decompositions by Path2Space (black) and SpaCET (orange). Error bars represent 95% confidence intervals derived from 1,000 bootstrapping iterations. Center points correspond to the computed ROC AUC values for each model.

Analysis of TCGA tumors identifies three subsets of breast cancer patients characterized by different TMEs and associated with markedly distinct survival

Having established Path2Space's ability to accurately infer spatial gene expression and cell types from H&E-stained images, we applied the model to predict the spatial transcriptomics of each patient's slide in the TCGA breast cancer cohort, comprising 1,096 samples from 1,043 patients, to characterize the tumor microenvironment of each sample and identify prognostic biomarkers on a large-scale cohort. To simulate the spatial resolution of Visium training data, we divided each TCGA slide into contiguous "pseudo-spots" covering comparable areas, with an average of 6,131 pseudo-spots per slide, without the typical 45-90 micrometer gaps between Visium spots.

After inferring spatial gene expression for the entire cohort, we employed SpaGCN[33], a graph convolutional network, to group pseudo-spots into spatial domains based on their inferred transcriptomic profiles and spatial organization. This approach captured continuous regions in the tumor microenvironment with similar as possible cell type compositions. To achieve consistent cluster labeling across patients, we applied the Louvain algorithm, grouping these domains into 11 spatial transcriptomic clusters shared across all tumor slides (**Methods**). We refer to these as spatial clusters, with each tumor represented by the proportions of these clusters within its slide.

Hierarchical clustering of the patients tumors, each represented by the composition of the 11 shared spatial clusters in it revealed three distinct spatially grounded subtypes of patients, each subtype characterized by similar spatial cluster compositions (**Fig. 4a,** the clustering was optimized using the elbow method (**Supplementary Fig. 3a**)). Survival analysis of the 853 patients having both survival and slide slides data revealed that these three patients spatial subtypes significantly stratified disease-free survival (Log-rank test, p=0.01) (**Fig. 4b**). Notably, spatial patients subtype 3 (n=233) is significantly associated with poor prognosis, even after adjusting for age and tumor stage (**Supplementary Fig. 3b**). Remarkably, this poor prognosis transcends the patients conventional clinical subtypes classification, and is observed across hormonal-positive (ER+ or PR+; n=385), HER2-positive (n=94), and triple-negative breast cancer (n=103) subtypes (although statistical significance was limited by sample size in the subgroup analyses) (**Fig. 4c**). Finally, a Cox regression analysis, applied separately to each of the original 11 shared spatial clusters identifies two clusters (4 and 5) as being significantly associated with survival ($p < 0.05$) (**Fig. 4d**). Cluster 4 is overrepresented in the poor prognosis subtype 3 and is linked to worse outcomes, became the focus of our subsequent analysis.

Deconvolution of the inferred gene expression revealed that SPACE cluster 4 had a high abundance of cancer cells relative to immune and stromal cells (**Supplementary Fig. 3c**). Differential gene expression analysis identified key breast cancer oncogenes—NOTCH1, TBX3, and PPM1D—as highly overexpressed in this cluster compared to others. Conversely, breast cancer tumor suppressor genes—GATA3, NCOR1, and CDKN1B—were notably downregulated (**Fig. 4e**). Gene set enrichment analysis using the MSigDB Hallmark gene sets[34] comparing this cluster to others showed significant enrichment of proliferation-related pathways and depletion of immune activation signatures (**Fig. 4f**), suggesting a highly aggressive, immunosuppressive TME.

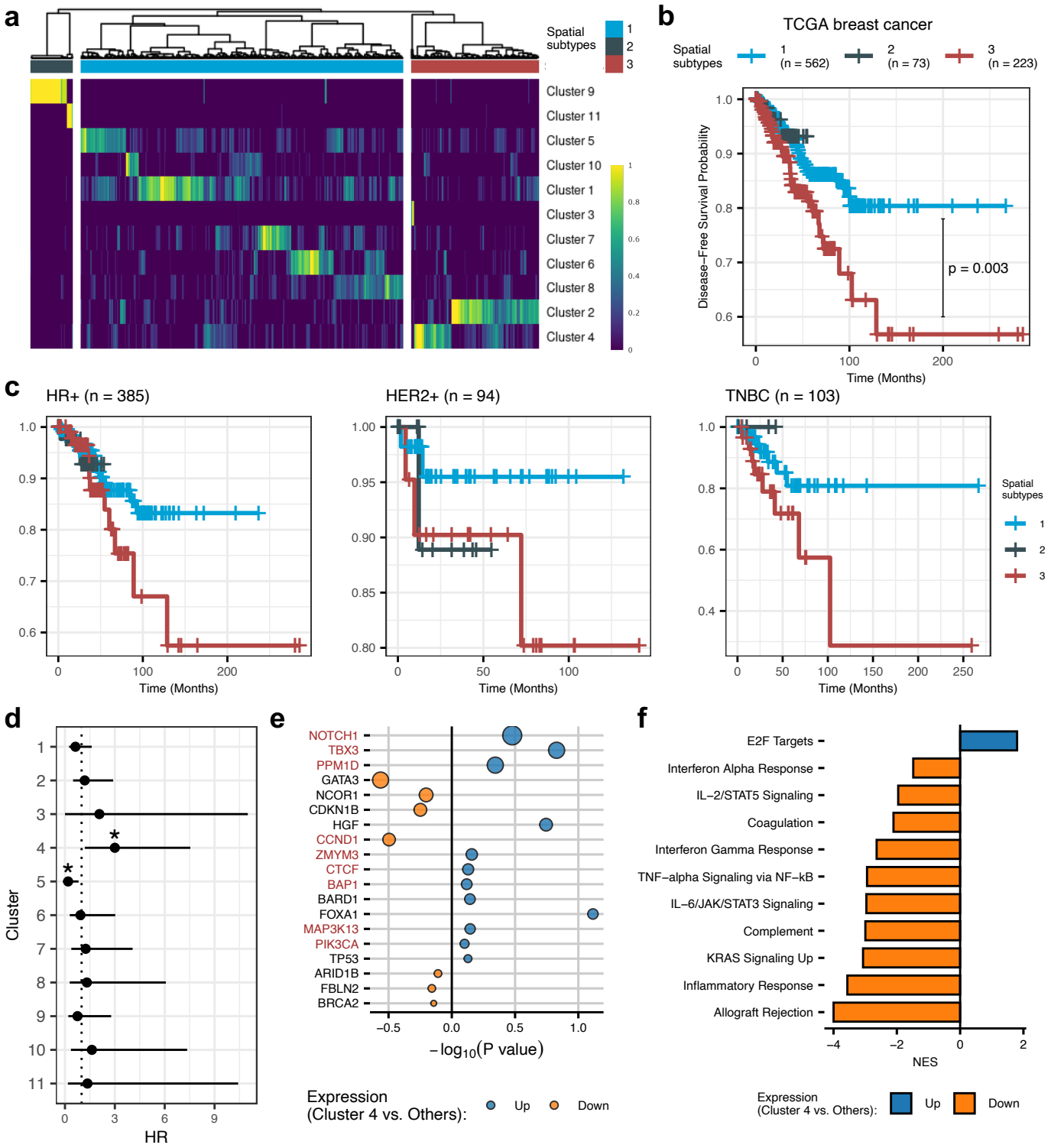


Fig. 4: Identification of prognostic spatial subtypes in the TCGA BRCA cohort. a, Hierarchical clustering of spatial cluster compositions for each tumor slide reveals three distinct spatial subtypes across *n*=858 TCGA breast cancer samples. The heatmap displays patients in rows and clusters in columns, with values corresponding to spatial cluster proportions in each patient. The three spatial subtypes are denoted by the color bar on the left. **b,** Kaplan-Meier

survival analysis shows that the three spatial subtypes significantly stratify patients by their disease-free survival. Patients in spatial subtype 3 (n=233) exhibit significantly poorer prognosis (p=0.01, Log-rank test). **c,** Subgroup survival analysis shows that the poor prognosis of spatial subtype 3 is consistent across the conventional hormone receptor-positive (ER+ or PR+), HER2-positive, and triple-negative breast cancer (TNBC) subtypes, though statistical significance is limited by smaller sample sizes in subgroup analyses. **d,** Cox regression analysis of individual spatial clusters identifies clusters 4 and 5 as significantly associated with survival ($P < 0.05$) across all patients. Cluster 4 is prominent in spatial subtype 3 and linked to worse outcomes. The horizontal axis represents the hazard ratio, with an asterisk (*) indicating statistical significance ($P < 0.05$). **e,** Differential gene expression analysis of spatial cluster 4 reveals key breast cancer oncogenes (black text) and tumor suppressor genes (brown text) that are differentially expressed. Blue dots represent upregulation, and brown dots indicate downregulation. Dot size corresponds to -log10(FDR). **f,** Gene set enrichment analysis (GSEA) reveals significant enrichment of proliferation-related pathways and depletion of immune activation signatures in spatial cluster 4. Blue bars indicate upregulation, and brown bars represent downregulation.

HER2 spatial heterogeneity is a biomarker of trastuzumab response

We next applied Path2Space to large breast cancer cohorts lacking spatial transcriptomics data to detect biomarkers of therapy response based on spatial gene expression. Our focus was to determine the relationship between the spatial distribution of HER2 expression and response to trastuzumab, an antibody targeting HER2 receptors to inhibit their oncogenic functions[35]. We hypothesized that tumors with heterogeneous HER2 expression, where high and low HER2-expressing cancer cells are closely situated in an intermingled fashion, would be more likely to respond to treatment.

To test this hypothesis, we analyzed two breast cancer cohorts treated with chemotherapy (HER2- patients) or a combination of chemotherapy and trastuzumab (HER2+ patients): TransNEO (HER2+ n=61; HER2- n=93) and IMPRESS (HER2+ n=62; HER2- n=64), both providing H&E slide images and response data. Using Path2Space, we inferred spatial gene expression and cell type fractions from pseudo-spots within the slide image. To robustly quantify HER2 expression per cancer cell, we calculated the HER2 pathway activity (normalized enrichment score) for each pseudo-spot and normalized it by the fraction of cancer cells in the same pseudo-spot. HER2 heterogeneity for each slide/patient was scored by computing the global negative Moran's I across the slide (**Methods**), with higher heterogeneity scores reflecting greater co-localization of high and low HER2-expressing cancer cells within the tumor microenvironment. (**Fig. 5a,b**).

Our analysis revealed that a higher HER2 heterogeneity scores were significantly associated with a complete pathological response (pCR) to trastuzumab in both the TransNEO (ROC AUC=0.8, P value=0.006) and IMPRESS (ROC AUC=0.72, P value=0.004) datasets. Reassuringly, in the same cohorts, this heterogeneity measure was not predictive of response for patients treated with chemotherapy without trastuzumab (ROC AUC=0.61, P value= 0.245 and ROC AUC=0.57, P

value=0.348 for TransNEO and IMPRESS respectively) (**Fig. 5c,d**). Furthermore, using these HER2 scores, we achieve a better predictive power of patients treatment response than the actual bulk gene expression of HER2. Analysis of measured bulk RNA-seq in the TransNEO cohort showed that the expression of HER2 (ERBB2) is only weakly associated with response, but to a lower degree than our predictor (ROC AUC= 0.65, P value=0.047). Additionally, the expression of the HER2 pathway was not associated with response (ROC AUC=0.64, P value=0.387) (**Fig. 5e**). Moreover, taking the mean of the predicted HER2 pathway activity per cancer fraction across the whole slide is also not significantly associated with response both in the TransNEO (ROC AUC=0.60, P value=0.216) and IMPRESS (ROC AUC=0.55, P value=0.539) datasets (**Supplementary Fig. 5**). These results demonstrate the considerable added value of spatially resolved gene expression analysis for treatment-specific biomarker discovery.

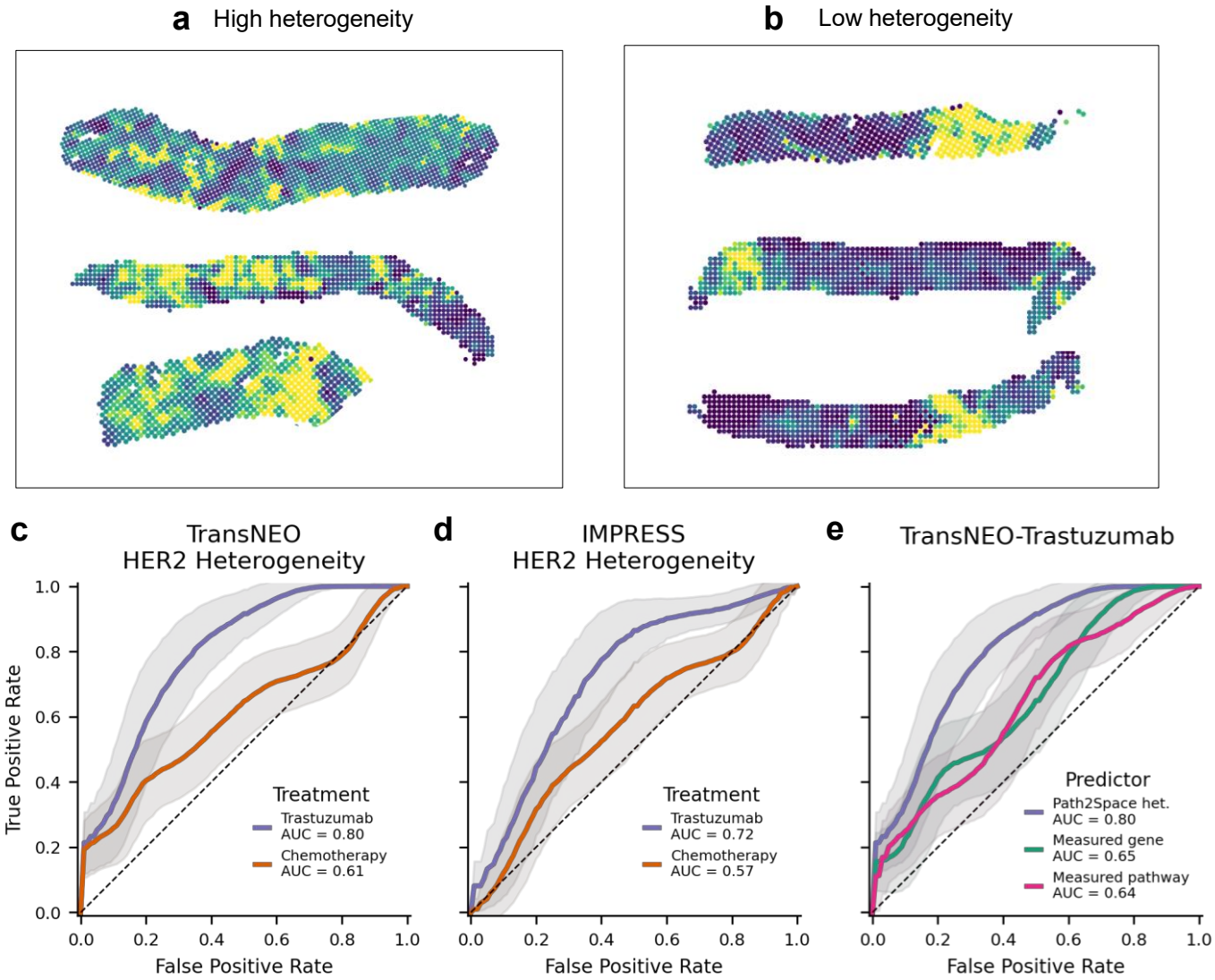


Fig. 5: Association between trastuzumab response and spatial HER2 expression heterogeneity of cancer cells. a-b, Examples of slides from the TransNEO dataset with high (**a**) and low (**b**) heterogeneity indexes. **c-d** Receiver operating characteristic (ROC) curves for HER2 per-cancer heterogeneity score in classifying therapy response in the TransNEO (**c**) and

IMPRESS (**d**) datasets. Purple represents trastuzumab-treated patients, and brown represents chemotherapy-treated patients. **e,** ROC curves for trastuzumab-treated patients in the TransNEO cohort, comparing response classification based on the HER2 per-cancer heterogeneity score inferred by Path2Space, ERBB2 gene expression, and HER2 pathway activity, both measured using bulk RNA-seq. The shaded areas around the curves represent the standard deviation from 1,000 bootstrapping iterations. Area under the curve (AUC) is shown in the legend.

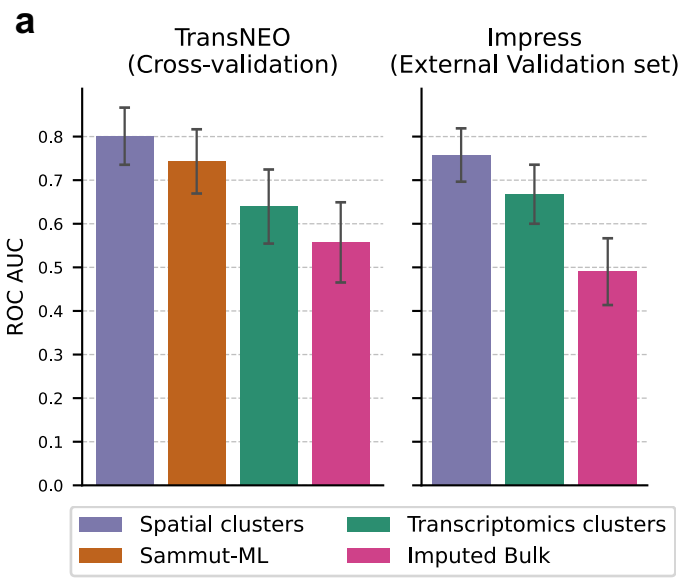
Predicting patient treatment response to trastuzumab via the composition of *spatially grounded clusters* in each tumor sample

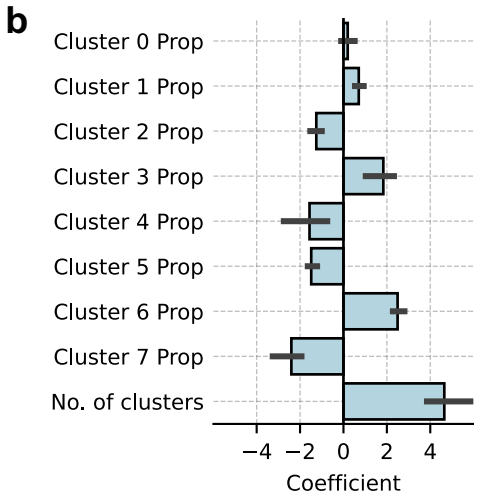
To further identify spatial biomarkers for trastuzumab response, we applied a complementary data-driven approach that models treatment outcomes based on the proportional composition of spatial clusters of cells in the tumor microenvironment. We reasoned that post hoc analysis of the most predictive spatial domains could reveal microenvironmental niches predictive of treatment response, providing further biologically interpretable insights. We used the same approach that we have applied to identify spatial clusters associated with patients survival in the TCGA cohort to infer shared spatial clusters in the TransNEO cohort, identifying eight cross-cohort spatially grounded clusters shared across all slides.

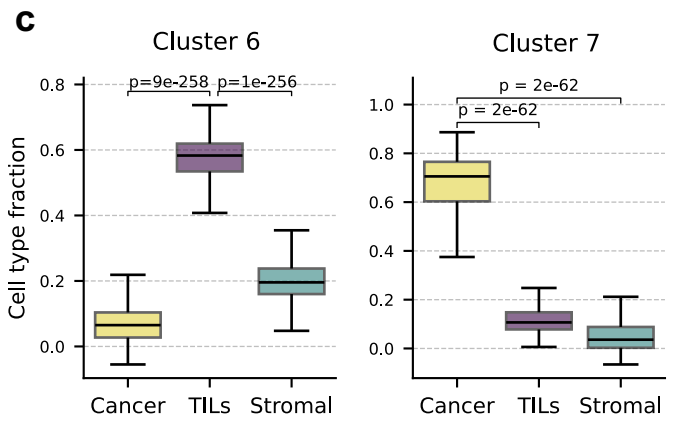
Next, we trained a linear classifier (logistic regression) to predict pathologic complete response (pCR) using the spatially *grounded clusters* proportions in a given input tumor slide as features. Five-fold cross-validation in the TransNEO cohort yielded a ROC AUC of 0.80 and an AUPRC of 0.68 (overall response rate of 0.31). Remarkably, this outperforms the performance of the recently reported Sammut-ML model (ROC AUC=0.74, AUPRC=0.52), which utilizes *measured* clinical, genomic, and bulk RNA sequencing data from an expanded cohort including 94 additional patients receiving only chemotherapy. Models based solely on transcriptomic clusters (ROC AUC=0.64, AUPRC=0.48) or H&E-imputed bulk gene expression (ROC AUC=0.56, AUPRC=0.40) also showed inferior performance (**Fig. 6a, Supplementary Fig. 6**).

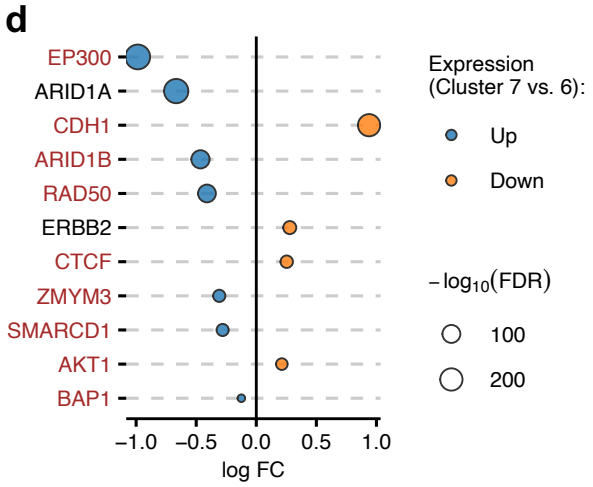
To assess the model's generalizability, we applied our model, trained on the TransNEO cohort, to the independent IMPRESS cohort of breast cancer patients treated with chemotherapy and trastuzumab (n=63). We transferred the cross-cohort spatially grounded cluster labels from TransNEO to each spot in the IMPRESS cohort using Seurat's anchor-based method[36], then assigned spatial clusters to their majority transcriptomic cluster. The trained logistic regression achieved ROC AUC of 0.76 and AUPRC of 0.86 (overall response rate of 0.61) on IMPRESS, surpassing models based on bulk transcriptomic clusters (ROC AUC=0.67, AUPRC=0.80) and H&E-imputed bulk gene expression (ROC AUC=0.41, AUPRC=0.63) (**Fig. 6a, Supplementary Fig. 6**). These results demonstrate the robust performance and generalizability of our spatially-grounded approach across independent cohorts.

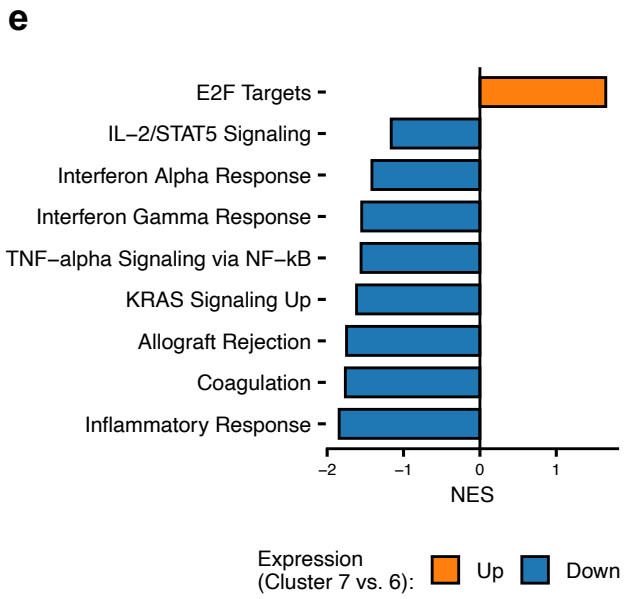
The coefficients from our linear model show that clusters 6 and 7 had the largest absolute values among the individual clusters, with cluster 6 strongly associated with treatment response and cluster 7 linked to non-response (**Fig. 6b**). Path2Space cell fraction estimations show cluster 6 is highly enriched for lymphocytes and cluster 7 for cancer cells in both the TransNEO (**Fig. 6c**) and IMPRESS cohorts (**Supplementary Fig. 6b**). Notably, differential expression analysis comparing clusters 6 and 7 revealed a tendency for tumor suppressors to be upregulated in cluster 6 (7 out of 9 differentially expressed), while oncogenes were exclusively upregulated in cluster 7 (2 out of 2) (**Fig. 6d**). Gene set enrichment analysis using the MSigDB Hallmark gene sets[34] demonstrated significant enrichment of proliferation and immune activation signatures in cluster 6 compared to cluster 7 (**Fig. 6e**). To further investigate this association, we applied SpatialDM[37] to identify differentially regulated ligand-receptor interactions in cluster 6 compared to the other clusters. Of the 517 differentially regulated interactions identified (likelihood ratio test, FDR<0.1, two-sided), 27 were immune-related interactions, notably involving CCL3 and CCL4, which are known to drive anti-cancer effects by promoting cytotoxic TIL infiltration[38] (**Fig. 6f**). Additionally, 97 upregulated interactions in cluster 6 were related to proliferation (**Supplementary Fig. 6c**). In contrast, no interactions were upregulated in cluster 7 compared to the other clusters.


a
TransNEO
(Cross-validation)
Impress
(External Validation set)
ROC AUC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Spatial clusters
Sammut-ML
Transcriptomics clusters
Imputed Bulk


b
Cluster 0 Prop
Cluster 1 Prop
Cluster 2 Prop
Cluster 3 Prop
Cluster 4 Prop
Cluster 5 Prop
Cluster 6 Prop
Cluster 7 Prop
No. of clusters
−4
−2
0
2
4
Coefficient


c
Cluster 6
Cluster 7
Cell type fraction
p=9e-258
p=1e-256
p = 2e-62
p = 2e-62
Cancer
TILs
Stromal


d
EP300
ARID1A
CDH1
ARID1B
RAD50
ERBB2
CTCF
ZMYM3
SMARCD1
AKT1
BAP1
−1.0
−0.5
0.0
0.5
1.0
log FC
Expression
(Cluster 7 vs. 6):
Up
Down
−log10(FDR)
100
200


e
E2F Targets
IL−2/STAT5 Signaling
Interferon Alpha Response
Interferon Gamma Response
TNF−alpha Signaling via NF−kB
KRAS Signaling Up
Allograft Rejection
Coagulation
Inflammatory Response
−2
−1
0
1
NES
Expression
(Cluster 7 vs. 6):
Up
Down

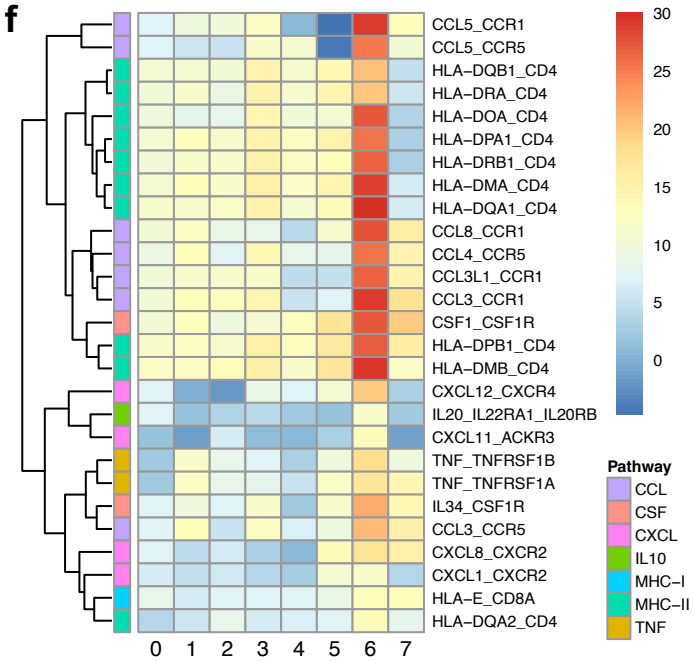

f
CCL5_CCR1
CCL5_CCR5
HLA−DQB1_CD4
HLA−DRA_CD4
HLA−DOA_CD4
HLA−DPA1_CD4
HLA−DRB1_CD4
HLA−DMA_CD4
HLA−DQA1_CD4
CCL8_CCR1
CCL4_CCR5
CCL3L1_CCR1
CCL3_CCR1
CSF1_CSF1R
HLA−DPB1_CD4
HLA−DMB_CD4
CXCL12_CXCR4
IL20_IL22RA1_IL20RB
CXCL11_ACKR3
TNF_TNFRSF1B
TNF_TNFRSF1A
IL34_CSF1R
CCL3_CCR5
CXCL8_CXCR2
CXCL1_CXCR2
HLA−E_CD8A
HLA−DQA2_CD4
0
1
2
3
4
5
6
7
30
25
20
15
10
5
0
Pathway
CCL
CSF
CXCL
IL10
MHC−I
MHC−II
TNF

Fig. 6: Path2Space uncovers links between the spatial organization of the tumor microenvironment and trastuzumab response. a, Area under the receiver operating characteristic curves (AUCs) for classifying trastuzumab response in the TransNEO (cross-validation, left) and IMPRESS (external validation, right) cohorts. The y-axis shows AUCs for models based on Path2Space spatial clusters (purple), Path2Space transcriptomic clusters (green), and bulk gene expression prediction (pink). For TransNEO, the Sammut-ML model (brown) cross-validation performance is also shown. Error bars represent 95% confidence intervals derived from 1,000 bootstrapping iterations. Center points correspond to the computed ROC AUC values for each model. **b,** Feature contributions to the spatial clusters-based response prediction model. Rows represent individual features, and the x-axis shows the mean coefficient across 5 models. Positive coefficients (right) indicate features contributing to response prediction, while negative coefficients (left) indicate features contributing to non-response prediction. Error bars represent standard errors derived from the models. **c,** Box plots showing the fraction of cell types per spot predicted by Path2Space for clusters 6 and 7 in the TransNEO cohort. Statistical significance between groups is evaluated using two-tailed Mann-Whitney U test p-values. Box plot elements: center lines=median; box limits=upper and lower quartiles; whiskers=1.5× interquartile range. **d**, Differential expression of tumor suppressors (brown text) and oncogenes (black text) in cluster 7 versus cluster 6. Brown bars indicate upregulation in cluster 7, while blue bars represent upregulation in cluster 6. Dot size corresponds to -log10(FDR). **e,** Differential enrichment of immune and proliferative Hallmark gene sets from MSigDB in cluster 7 versus cluster 6. Brown bars represent enrichment in cluster 7, while blue bars represent enrichment in cluster 6. **f,** Heatmap showing immune-related ligand-receptor interactions enriched in cluster 6 (likelihood ratio test, FDR < 0.1, two-sided) using SpatialDM. Rows correspond to ligand-receptor interaction names, and columns represent clusters. Cell colors denote the global z-score of the Moran's I statistic for each ligand-receptor pair. The color bar on the left indicates the pathway associated with each interaction.

Discussion

In this study, we introduced Path2Space, a novel deep learning model designed to predict spatial gene expression directly from histopathology slide images. By training Path2Space on spatial transcriptomics (ST) data, we demonstrated its effectiveness in accurately predicting spatial gene expression and inferring spatial cell type composition on an unprecedented level. The model's robustness was validated through cross-validation and external validation, reinforcing the reliability of its H&E-based predictions. We then applied Path2Space to the TCGA breast cancer cohort, identifying three subtypes of breast tumors based on spatial transcriptomics that are significantly associated with disease-free survival. Notably, applying Path2Space to two trastuzumab response cohorts revealed spatial markers significantly associated with treatment outcomes, showcasing its potential for identifying clinically relevant biomarkers and markedly enhancing our ability to predict patients' response speedily and in equitable manner directly from the patients pathology slides.

Our investigation into trastuzumab response employed two complementary approaches, each providing valuable insights. First, we hypothesized that tumors with heterogeneous HER2 expression, where high and low HER2-expressing cancer cells are closely situated, would be more likely to respond to treatment. This hypothesis is based on trastuzumab's ability to modulate immune effector cell cytotoxicity through Fc receptor engagement[39] which may influence the surrounding microenvironment. Our results support this hypothesis, providing the first translational evidence demonstrating that (inferred) HER2 spatial heterogeneity is indeed predictive of response, demonstrating the huge potential value going forward of leveraging large cohorts to uncover spatial biomarkers that bulk gene expression alone may miss.

Next, we developed an interpretable machine learning model to predict treatment response based on a characterization of the spatial organization of the tumor microenvironment, as inferred from the slide predicted spatial gene expression. Performing a further downstream analysis of the most predictive clusters, we were able to learn more on key genes and cell-cell interactions within typical spatial domains that are associated with better or worse outcomes.

While Path2Space shows great promise in predicting spatial gene expression from histopathology slide images, it is not without limitations. Our findings indicate that the correlation between measured ST and predicted gene expression is high for 'only' a subset of approximately 4,500 highly expressed genes (which yet remarkably comprises about two thirds of the genes whose expression is frequently measured in current Visium assays). This puts bounds on the ability of our model to capture the full spectrum of gene expression within the tumor and its microenvironment. Additionally, the deconvolution analysis based on Path2Space inference, which infers spatial cell type composition, is still slightly less accurate compared to deconvolution based on measured ST data, as expected. This discrepancy highlights the continuing importance of conducting spatial transcriptomics experiments for precise cellular mapping.

Nevertheless, bearing those limitations in mind, our results clearly testify that Path2Space's level of accuracy is sufficient for inferring spatial cellular composition from whole slide images (WSI), enabling the analysis of much larger datasets than is currently feasible with ST methods. This scalability opens the door to discoveries that are currently limited by the small size of existing ST datasets, positioning Path2Space as a valuable and cost-effective tool for advancing our understanding of tumors, their microenvironments, and their impact on patient outcomes.

As spatial transcriptomics data continues to grow, with higher read counts and finer spatial resolution—even at a single-cell level[40]—we expect improved predictions of spatial gene expression and more precise inferences of cellular composition. Notably, while trastuzumab response served as a test case, the Path2Space approach is broadly applicable. It can be extended to any treatment where H&E images and response data are available, offering the potential to revolutionize biomarker discovery and treatment response prediction across diverse cancer types and therapies. Furthermore, the model's framework is adaptable to other cancer types, offering the potential for broader insights into different tumor microenvironments. Beyond transcriptomics, Path2Space could be expanded to integrate other spatial omics data, such as methylation and proteomics, as these measurement technologies become available.

Ultimately, Path2Space provides a scalable and cost-effective solution for predicting spatial gene expression, facilitating the analysis of large datasets previously inaccessible to spatial transcriptomics methods. By unlocking the potential of spatial transcriptomics at scale, it can generate unprecedented insights into tumor architecture and its molecular foundations, paving the way for more personalized and effective treatment strategies. By facilitating the analysis of larger patients datasets, Path2Space extends the reach of spatial transcriptomics to advance precision oncology, with the potential to significantly improve patient outcomes.

Methods

Data collection

All datasets used in this study are from publicly available sources. Links to download the datasets can be found in the Data Availability section below.

Spatial transcriptomics data preprocessing

For model training, gene expression values were incremented by 1 and log10-transformed. Due to the high proportion of missing values typical in spatial transcriptomics data, we focused on 14,068 genes that were detected in at least 5% of spots on each slide in the TNBC dataset, ensuring sufficient data for robust training across genes.

Given our smoothing procedure (see below), we restricted training to slides where, on average, each spot had eight neighboring spots within a 200-micrometer distance, the minimum spacing between spots. This filtering resulted in a final set of 25 slides from the TNBC dataset used for model training.

Path2Space computational framework

Spatial gene expression prediction

The architecture consists of three main components (**Fig. 1a**):

1. **Image Processing:** The preprocessing phase begins by dividing whole-slide images into non-overlapping 224x224 pixel tiles centered around each spatial transcriptomics spot. Tiles with more than 50% background content are excluded. This process results in several thousand tiles per slide, depending on the slide's dimensions. To address staining variability, color normalization techniques[14] are applied.

2. **Feature Extraction:** Tile images corresponding to spots are processed using CTransPath[41], a foundational digital pathology model trained through self-supervision on histopathology images. Each tile is represented by a 768-dimensional CTransPath feature vector.

3. **Multilayer Perceptron (MLP) Regression:** This component builds a predictive model linking CTransPath features to whole-genome spot-level gene expression. The model architecture includes three layers: (1) an input layer with 768 nodes, corresponding to the size of the CTransPath feature vector; (2) a hidden layer with 768 nodes; and (3) an output layer with 14,068 nodes, each representing a gene. ReLU activation is applied to each layer to introduce non-linearity and enhance model performance.

Spatially resolved estimation of cell type fractions

We leveraged Path2Space-inferred spot-level gene expression data to estimate cell type composition within each spot. The procedure began within the training fold, where the inferred gene expression values were first normalized to a range of 0–1 using Scikit-learn's[42] MinMaxScaler. Within the same training fold, we then selected the top 1,000 genes with the highest analysis of variance (ANOVA) F-values relative to tumor class using sklearn's SelectKBest and f_classif. A multilayer perceptron (MLP) regression model was then employed to predict the fractions of three cell types: cancer cells, lymphocytes, and stromal cells.

Spatial smoothing

Gene expression values were smoothed by averaging each spot's expression with that of all its neighboring spots within a 200-micrometer radius, the minimum distance between spots in the dataset. On average, this resulted in 8 neighbors per spot being used for smoothing. This procedure was applied to both model predictions and original expression measurements (**Supplementary Fig. 1**) and was used for model evaluation and downstream analyses.

Model training procedure

Spatial gene expression prediction: The models were trained using stochastic gradient descent with mini-batch processing. For optimization, we employed the Adam optimizer with an initial learning rate of 0.0001. The mean squared error (MSE) between predicted and actual values was used as the loss function.

Training of the MLP regressors and classifiers was conducted over 500 epochs. To reduce the risk of overfitting, a dropout rate of 20% was applied to the initial layer. We also implemented early stopping to enhance computational efficiency and further prevent overfitting. The training was stopped if no improvement was observed in the mean correlation between predicted and actual gene expression values on the validation set after 50 epochs.

Spatially resolved estimation of cell type fractions: Using annotations from the PanopTILs dataset, cell type fractions for cancer, lymphocytes, and stromal cells were calculated for each ROI and used as labels. The hyperparameters of the MLP regressors, which infer cell type fractions from the predicted spatial gene expression, were fine-tuned through fivefold cross-validation within the training fold. This optimization process, guided by the regression F-score, was performed using Scikit-learn's GridSearchCV.

Cross-validation approach

Our training and evaluation strategy closely followed the protocol established in our previous work[14,15]. Specifically, we employed a 15×7 nested cross-validation strategy for both the direct and indirect models. The TNBC dataset was initially divided into 15 outer folds, each fold using slides from different patients as the test fold. Each outer fold was then further split into seven inner folds, where slides from a single patient were used for training and validation.

For generating predictions on the TNBC training dataset, we averaged the outputs from the seven models, each trained on a different inner fold, for the patient slides in the respective outer fold. For the external validation datasets and clinical H&E cohorts, predictions from all 15×7=105 models were averaged to produce the final prediction.

For the cell type fraction models, we applied 5-fold cross-validation. When predicting on validation or clinical H&E cohort datasets after training, we averaged the outputs from all five models to obtain the final prediction.

Analysis of predicted spatial gene expression

For downstream analysis of the predicted spatial gene expression, we utilized smoothed values for 4,644 genes that exhibited strong correlations (above 0.4) in both the TNBC and 10x datasets. The data were normalized using Seurat's[36] log-normalization method.

Gene set enrichment analysis was performed using the Python package GSEAPY[43], with the prerank function[44] for normalized enrichment scores and the enrichr function for overrepresentation analysis (**Fig. 1d**).

For clustering, we initially applied SpaGCN[33] to the inferred spatial gene expression data to identify spatial domains within each slide. To ensure consistent cluster labels across tumor slides from different patients, we employed Seurat's FindClusters function to group all spots across the cohort based on their gene expression profiles, resulting in transcriptomic cluster labels consistent across all samples. Each spatial domain was then assigned to the transcriptomic cluster corresponding to the majority of its spots, leading to a consistent classification of eight cross-cohort spatial clusters across all slides.

To transfer these spatial clusters to the independent dataset, we first used SpaGCN to infer per-slide spatial clusters in the new dataset. We then employed Seurat's TransferData function, utilizing an anchor-based approach, to map spatial clusters to the IMPRESS cohort by assigning each spot to its corresponding majority transcriptomic cluster.

Differential expression analysis was conducted using Seurat's FindAllMarkers function with the MAST[45] method. For ligand-receptor interaction analysis, we used the differential analysis pipeline from SpatialDM[37] (z-score approach) while downsampling clusters to maintain equal proportions.

HER2 heterogeneity analysis

To estimate HER2 expression in cancer cells, we used Path2Space-inferred spatial gene expression and deconvolution to compute the normalized enrichment score (NES) for the 'ErbB Signaling Pathway WP673' pathway (WikiPathways library[48]) using single-sample gene set enrichment analysis (ssGSEA) implemented in GSEAPY. The NES values, representing HER2 activity per spot, were further adjusted by dividing them by the estimated cancer fraction at each spot to obtain the HER2 per cancer value.To calculate the global Moran's i for each slide, we used the Moran function from the PySAL-esda (Exploratory Spatial Data Analysis) library[49] .

Trastuzumab response prediction

We then trained a logistic regression model to predict pathologic complete response (pCR) using the proportions of eight spatially grounded clusters in a given tumor slide, along with the number of clusters per slide (i.e., clusters with a proportion > 0) as features. A five-fold cross-validation scheme was employed, where features were scaled using Scikit-learn's MinMaxScaler. Regularization (L1 or L2) was selected for each model through grid search with five-fold cross-validation on the training folds, optimizing for ROC AUC using GridSearchCV. For external validation, the prediction scores from the five models were averaged.

Model evaluation

We evaluated gene-wise prediction accuracy by first calculating the Pearson correlation between spatially smoothed actual and predicted expression values across all spots within each slide. We then averaged these correlations across genes and slides from the same patient. Finally, for each gene, the correlation values were averaged across all patients. Correlations were computed using SciPy[50].

For classification tasks (e.g., **Fig. 4d** and trastuzumab response), we used ROC AUC and AUPRC metrics as implemented in Scikit-learn.

In cross-validation, all metrics were computed by aggregating the results across all folds.

Implementation details

Analysis in this study was mainly performed in Python 3.9.7 with the libraries including Numpy 1.20.3, Pandas 1.3.4, Scikit-learn 1.2.2 and Matplotlib 3.4.3. Image processing including tile partitioning and color normalization was conducted with OpenSlide 1.1.2, OpenCV 4.5.4 and PIL 8.4.0. The feature extraction, feature compression, MLP regression and MLP classification components were implemented using PyTorch 1.12.0. Classification machine-learning packages including logistic regression, were deployed using Scikit-learn 1.2.2.

Data availability

Spatial transcriptomics gene expression and images from the TNBC dataset are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210616. The HER2+ dataset can be accessed at https://doi.org/10.5281/zenodo.4751624, and the 10x dataset is available at https://espace.library.uq.edu.au/view/UQ:4fb74a9. PanopTIL annotations and images are available at https://sites.google.com/view/panoptils/. H&E-stained whole-slide images and patient response data from the TransNEO-breast dataset are available at https://ega-archive.org/studies/EGAS00001004582, and data from the IMPRESS cohort can be accessed at https://tinyurl.com/IMPRESS-DATA.

Code availability

Path2Omics code will be made available for academic research purposes after publication via Zenodo.

Competing interests

E.D.S., E.M.C. and E.R. are listed as inventors on a provisional patent (application no. 63/703,060, United States, 2024) filed based on the methodology outlined in this study. E.R. is a co-founder of Medaware, Metabomed and Pangea Biomed (divested from the latter). E.R. serves as a non-paid scientific consultant to Pangea Biomed under a collaboration agreement between Pangea Biomed and the NCI. The other authors declare no competing interests.

References

1. Wang, Y. *et al.* Multidirectional characterization of cellular composition and spatial architecture in human multiple primary lung cancers. *Cell Death Dis.* **14**, 462 (2023).
2. Wang, Y. *et al.* Spatial transcriptomics delineates molecular features and cellular plasticity in lung adenocarcinoma progression. *Cell Discov* **9**, 96 (2023).
3. Andersson, A. *et al.* Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat. Commun.* **12**, 6012 (2021).
4. Bassiouni, R. *et al.* Spatial Transcriptomic Analysis of a Diverse Patient Cohort Reveals a Conserved Architecture in Triple-Negative Breast Cancer. *Cancer Res.* **83**, 34–48 (2023).
5. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
6. Ravi, V. M. *et al.* Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer Cell* **40**, 639–655.e13 (2022).
7. Karras, P. *et al.* A cellular hierarchy in melanoma uncouples growth and metastasis. *Nature* **610**, 190–198 (2022).
8. Lin, J.-R. *et al.* Multiplexed 3D atlas of state transitions and immune interaction in colorectal cancer. *Cell* **186**, 363–381.e19 (2023).
9. Galeano Niño, J. L. *et al.* Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature* **611**, 810–817 (2022).

10. Arora, R. *et al.* Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nat. Commun.* **14**, 5029 (2023).
11. Huang, S. *et al.* Spatial transcriptomics: a new frontier in cancer research. *Clinical Cancer Bulletin* **3**, 1–26 (2024).
12. Smith, K. D., Prince, D. K., MacDonald, J. W., Bammler, T. K. & Akilesh, S. Challenges and Opportunities for the Clinical Translation of Spatial Transcriptomics Technologies. *Glomerular Dis* **4**, 49–63 (2024).
13. Coudray, N. *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
14. Hoang, D.-T. *et al.* A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nat Cancer* (2024) doi:10.1038/s43018-024-00793-2.
15. Hoang, D.-T. *et al.* Prediction of DNA methylation-based tumor types from histopathology in central nervous system tumors with deep learning. *Nat. Med.* (2024) doi:10.1038/s41591-024-02995-8.
16. He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng* **4**, 827–834 (2020).
17. Xie, R. *et al.* Spatially Resolved Gene Expression Prediction from H&E Histology Images via Bi-modal Contrastive Learning. *arXiv [cs.CV]* (2023).
18. Tan, X. *et al.* STimage:robust, confident and interpretable models for predicting gene markers from cancer histopathological images. *bioRxiv* 2023.05.14.540710 (2023) doi:10.1101/2023.05.14.540710.
19. Zeng, Y. *et al.* Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Brief. Bioinform.* **23**, (2022).
20. Monjo, T., Koido, M., Nagasawa, S., Suzuki, Y. & Kamatani, Y. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Sci. Rep.* **12**, 4133 (2022).
21. Banh, D. & Huang, A. Scalable parametric encoding of multiple modalities. *bioRxiv* 2021.07.09.451779 (2022) doi:10.1101/2021.07.09.451779.
22. Rahaman, M. M., Millar, E. K. A. & Meijering, E. Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning. *Sci. Rep.* **13**, 13604 (2023).
23. Jia, Y., Liu, J., Chen, L., Zhao, T. & Wang, Y. THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. *Brief. Bioinform.* **25**, (2023).
24. Hinshaw, D. C. & Shevde, L. A. The Tumor Microenvironment Innately Modulates Cancer Progression. *Cancer Res.* **79**, 4557–4566 (2019).
25. Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat. Commun.* **14**, 8353 (2023).
26. Liu, S., Amgad, M., Rathore, M. A., Salgado, R. & Cooper, L. A. D. A panoptic segmentation approach for tumor-infiltrating lymphocyte assessment: development of the MuTILs model and PanopTILs dataset. *bioRxiv* (2022) doi:10.1101/2022.01.08.22268814.
27. Sammut, S.-J. *et al.* Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).

28. Huang, Z. *et al.* Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. *NPJ Precis Oncol* **7**, 14 (2023).
29. Jiang, Y., Xie, J., Tan, X., Ye, N. & Nguyen, Q. Generalization of deep learning models for predicting spatial gene expression profiles using histology images: A breast cancer case study. *bioRxiv* 2023.09.20.558624 (2023) doi:10.1101/2023.09.20.558624.
30. Chan, A. S. *et al.* Benchmarking the translational potential of spatial gene expression prediction from histology. *bioRxiv* 2023.12.12.571251 (2023) doi:10.1101/2023.12.12.571251.
31. Pang, M., Su, K. & Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv* 2021.11.28.470212 (2021) doi:10.1101/2021.11.28.470212.
32. Ru, B., Huang, J., Zhang, Y., Aldape, K. & Jiang, P. Estimation of cell lineages in tumors from spatial transcriptomics data. *Nat. Commun.* **14**, 568 (2023).
33. Hu, J. *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
34. Castanza, A. S. *et al.* Extending support for mouse data in the Molecular Signatures Database (MSigDB). *Nat. Methods* **20**, 1619–1620 (2023).
35. Maadi, H., Soheilifar, M. H., Choi, W.-S., Moshtaghian, A. & Wang, Z. Trastuzumab Mechanism of Action; 20 Years of Research to Unravel a Dilemma. *Cancers* **13**, (2021).
36. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
37. Li, Z., Wang, T., Liu, P. & Huang, Y. SpatialDM for rapid identification of spatially co-expressed ligand-receptor and revealing cell-cell communication patterns. *Nat. Commun.* **14**, 3995 (2023).
38. Korbecki, J., Grochans, S., Gutowska, I., Barczak, K. & Baranowska-Bosiacka, I. CC Chemokines in a Tumor: A Review of Pro-Cancer and Anti-Cancer Properties of Receptors CCR5, CCR6, CCR7, CCR8, CCR9, and CCR10 Ligands. *Int. J. Mol. Sci.* **21**, (2020).
39. Clynes, R. A., Towers, T. L., Presta, L. G. & Ravetch, J. V. Inhibitory Fc receptors modulate in vivo cytotoxicity against tumor targets. *Nat. Med.* **6**, 443–446 (2000).
40. Oliveira, M. F. *et al.* Characterization of immune cell populations in the tumor microenvironment of colorectal cancer using high definition spatial profiling. *bioRxiv* 2024.06.04.597233 (2024) doi:10.1101/2024.06.04.597233.
41. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
42. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
43. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, (2023).
44. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
45. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes

and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).